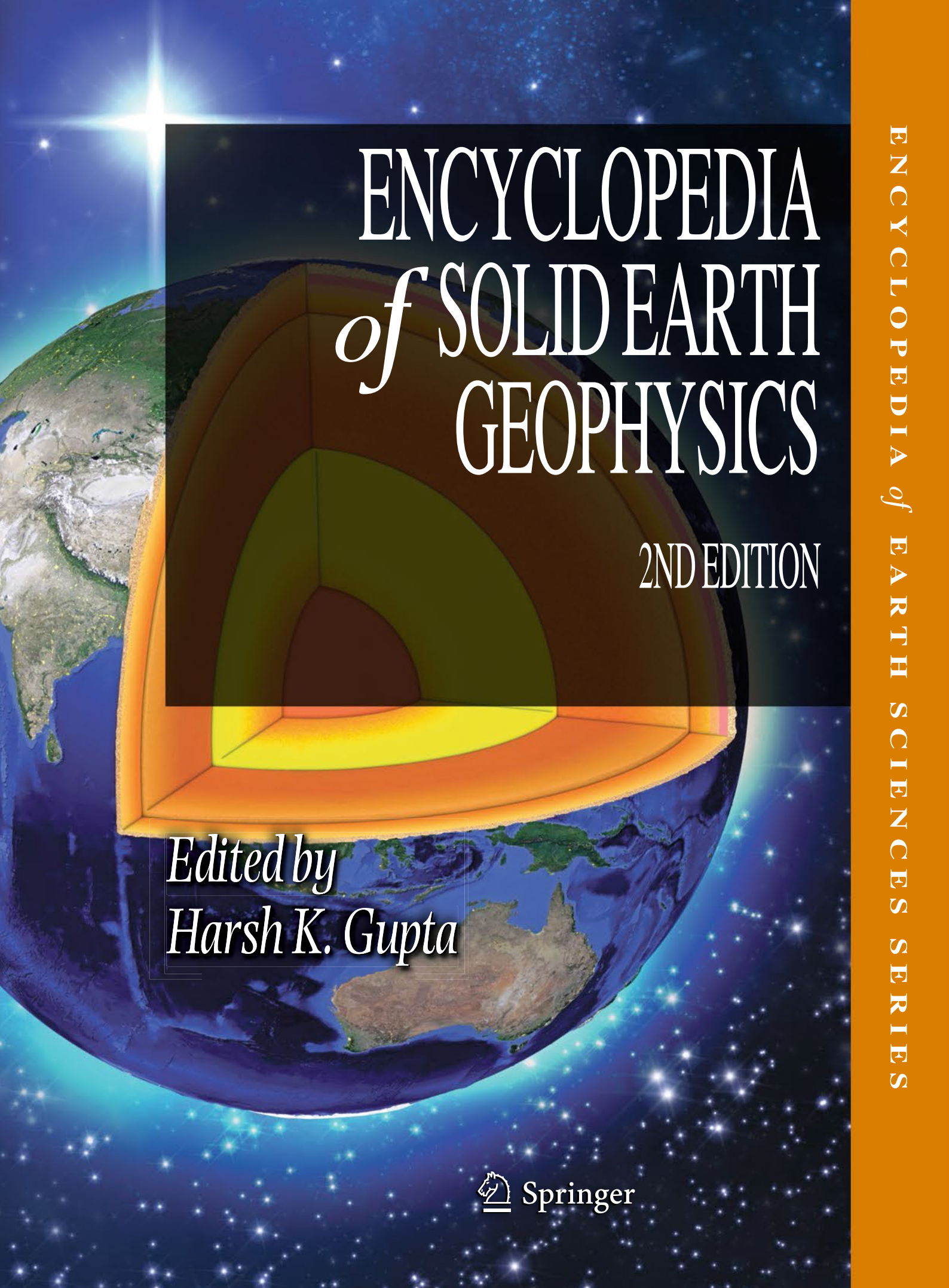




ENCYCLOPEDIA *of* SOLID EARTH GEOPHYSICS

2ND EDITION

Edited by
Harsh K. Gupta

ENCYCLOPEDIA of EARTH SCIENCES SERIES



Springer

ENCYCLOPEDIA *of*
SOLID EARTH GEOPHYSICS

Encyclopedia of Earth Sciences Series

ENCYCLOPEDIA OF SOLID EARTH GEOPHYSICS

Volume Editor

Harsh K. Gupta
National Geophysical Research Institute
Council of Scientific and Industrial Research
Hyderabad, India

Editorial Board

Kusumita Arora
National Geophysical Research Institute
Council of Scientific and Industrial Research (CSIR)
Hyderabad, India

Anny Cazenave
Laboratoire d'Etudes en Géophysique et Océanographie Spatiales (LEGOS)
CNES & Observatoire Midi-Pyrénées
Toulouse, France

Eric Robert Engdahl
Denver, CO, USA

Rainer Kind
Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences
Potsdam, Germany

Ajay Manglik
National Geophysical Research Institute
Council of Scientific and Industrial Research (CSIR)
Hyderabad, India

Sukanta Roy
Ministry of Earth Sciences, Govt. of India
Borehole Geophysics Research Laboratory
Karad, India

Kalachand Sain
Wadia Institute of Himalayan Geology
Uttarakhand, India

Seiya Uyeda
Tokyo, Japan

Aims of the Series

The Encyclopedia of Earth Sciences Series provides comprehensive and authoritative coverage of all the main areas in the Earth Sciences. Each volume comprises a focused and carefully chosen collection of contributions from leading names in the subject, with copious illustrations and reference lists. These books represent one of the world's leading resources for the Earth Sciences community. Previous volumes are being updated and new works published so that the volumes will continue to be essential reading for all professional earth scientists, geologists, geophysicists, climatologists, and oceanographers as well as for teachers and students. See the back of this volume for a current list of titles in the Encyclopedia of Earth Sciences Series.

Go to <http://www.springerlink.com/reference-works/> to visit the "Earth Sciences Series" online.

About the Series Editor

Professor Charles W. Finkl has edited and/or contributed to more than eight volumes in the Encyclopedia of Earth Sciences Series. He has been the Executive Director of the Coastal Education and Research Foundation and Editor-in-Chief of the international Journal of Coastal Research for the past 35 years. He is also the Series Editor of the Coastal Research Library (Springer). In addition to these duties, he is Distinguished University Professor Emeritus at Florida Atlantic University (FAU) (Boca Raton, Florida). He is a graduate of Oregon State University (Corvallis) and the University of Western Australia (Perth). Work experience includes the International Nickel Company of Australia (Perth), Coastal Planning & Engineering (Boca Raton, Florida), and Technos Geophysical Consulting (Miami, Florida). He has published numerous peer-reviewed technical research papers and edited or co-edited and contributed to many books. Dr. Finkl is a Certified Professional Geological Scientist (Arvada, Colorado), a Certified Professional Soil Scientist (Madison, Wisconsin), a Certified Wetland Scientist (Lawrence, Kansas), and a Chartered Marine Scientist (London). Academically, he served as a Demonstrator at the University of Western Australia, Courtesy Professor at Florida International University (Miami), Program Professor and Director of the Institute of Coastal and Marine Studies at Nova Southeastern University (Port Everglades, Florida) and Full Professor at FAU. During his career, he acquired field experience in Australia; the Bahamas; Puerto Rico, Jamaica; Brazil; Papua New Guinea and other SW Pacific islands; southern Africa; Western Europe; and the Pacific Northwest, Midwest, and Southeast USA. Dr. Finkl is a member of several professional societies including the Geological Society of America; Soil Science Society of America; Institute of Marine Engineering, Science and Technology; and the Society of Wetland Specialists. He is a recipient of the International Beach Advocacy Award (Florida Shore & Beach Preservation Association), Certificate of George V. Chilingar Medal of Honor (Russian Academy of Natural Sciences), and Lifetime Commitment to Coastal Science Award (International Coastal Symposium).

Founding Series Editor

Professor Rhodes W. Fairbridge (deceased) has edited more than 24 Encyclopedias in the Earth Sciences Series. During his career he has worked as a petroleum geologist in the Middle East, been a WW II intelligence officer in the SW Pacific and led expeditions to the Sahara, Arctic Canada, Arctic Scandinavia, Brazil and New Guinea. He was Emeritus Professor of Geology at Columbia University and was affiliated with the Goddard Institute for Space Studies.

ENCYCLOPEDIA OF EARTH SCIENCES SERIES

ENCYCLOPEDIA *of* SOLID EARTH GEOPHYSICS

Second Edition

edited by

HARSH K. GUPTA

*National Geophysical Research Institute
Council of Scientific and Industrial Research
Hyderabad, India*

 Springer

Editor

Harsh K. Gupta
National Geophysical Research Institute
Council of Scientific and Industrial Research
Hyderabad, India

ISBN: 978-3-030-58630-0

This publication is available also as:

Electronic publication under ISBN 978-3-030-58631-7 and

Print and electronic bundle under ISBN 978-3-030-58632-4

Cover illustration: C.B. 3D ILLUSTRATION & ANIMATION | illustration by Christoph Burgstedt

Every effort has been made to contact the copyright holders of the figures and tables which have been reproduced from other sources. Anyone who has not been properly credited is requested to contact the publishers, so that due acknowledgement may be made in subsequent editions.

All rights reserved for the contributions: *Earthquake Sounds; Electrical Resistivity Surveys and Data Interpretation; Gravity, Data to Anomalies; Gravity, Global Models; Instrumentation, Electrical Resistivity; Sentinel Satellites Initiate New Era in Earth Observatories*

© Springer Nature Switzerland AG 2021

1st edition published by: © Springer Science + Business Media B.V. 2011

The original Encyclopedia of Solid Earth Geophysics was compiled by David E. James, and was first published in the Encyclopedia of Earth Sciences Series in 1989.

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Switzerland AG.

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

About the Editor	xiii	Continental Rifts	64
Contributors	xv	<i>A. M. Celâl Şengör</i>	
Preface to the Second Edition	xxxi	Core Dynamo	78
Preface to the First Edition	xxxiii	<i>Ulrich R. Christensen</i>	
Acknowledgments	xxxv	Core-Mantle Coupling	86
Absolute Age Determinations: Radiometric	1	<i>Paul H. Roberts and Jonathan M. Aurnou</i>	
<i>Richard W. Carlson</i>		Crustal Reflectivity (Oceanic) and Magma Chamber	101
Anthropogenic Seismicity Related to Exploitation of Georesources	8	<i>Satish C. Singh</i>	
<i>Stanislaw Lasocki and Beata Orlecka-Sikora</i>		Curie Temperature	112
Archaeomagnetism	15	<i>Vincenzo Pasquale</i>	
<i>Donald H. Tarling</i>		Deep Scientific Drilling	115
Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India	19	<i>Ulrich Harms and Harold Tobin</i>	
<i>Harsh K. Gupta</i>		Deep Seismic Reflection and Refraction Profiling	127
Biogeophysics	33	<i>Kabir Roy Chowdhury</i>	
<i>Lee Slater and Estella Atekwana</i>		Differential Rotation of the Earth's Inner Core	144
Body Waves	37	<i>Xiaodong Song</i>	
<i>Mahmoud Mohamed Selim Saleh</i>		Earth Rotation	149
Borehole Seismic Networks and Arrays	44	<i>Harald Schuh and Sigrid Böhm</i>	
<i>Marco Bohnhoff and Peter Malin</i>		Earth Tides	156
Characteristic Earthquakes and Seismic Gaps	53	<i>John M. Wahr</i>	
<i>David D. Jackson and Yan Y. Kagan</i>		Earth, Density Distribution	160
Continental Crustal Structure	56	<i>Frank D. Stacey and Paul M. Davis</i>	
<i>Rolf Meissner and Hartmut Kern</i>		Earth's Structure, Core	164
Continental Drift	63	<i>Lianxing Wen</i>	
<i>Alan G. Smith</i>		Earth's Structure, Global	166
		<i>Jean-Paul Montagner</i>	

vi	CONTENTS	
Earth's Structure, Lower Mantle <i>Edward J. Garnero, Allen K. McNamara and James A. Tyburczy</i>	176	Earthquakes, Energy <i>Domenico Di Giacomo and Peter Borman</i> 288
Earth's Structure, Upper Mantle <i>Guust Nolet</i>	183	Earthquakes, Intensity <i>Gottfried Grünthal and Roger M. W. Musson</i> 292
Earthquake Lights <i>John S. Derr, France St-Laurent, Friedemann T. Freund and Robert Thériault</i>	189	Earthquakes, Intraplate <i>Prantik Mandal</i> 299
Earthquake Precursors and Prediction <i>Toshiyasu Nagao, Masashi Kamogawa and Seiya Uyeda</i>	193	Earthquakes, Location Techniques <i>Clifford H. Thurber</i> 301
Earthquake Prediction, M8 Algorithm <i>Alik Ismail-Zadeh and Vladimir Kossobokov</i>	204	Earthquakes, PAGER <i>David J. Wald, Kishor S. Jaiswal, Kristin D. Marano and Michael Hearne</i> 308
Earthquake Rupture: The Inverse Problem <i>Shamita Das</i>	208	Earthquakes, ShakeCast <i>Kuo-wan Lin, David J. Wald and Daniel Slosky</i> 312
Earthquake Sounds <i>Andrew J. Michael</i>	220	Earthquakes, ShakeMap <i>David J. Wald, C. Bruce Worden, Eric M. Thompson and Michael Hearne</i> 316
Earthquake Source Theory <i>Raul Madariaga</i>	224	Earthquakes, Strong-Ground Motion <i>Giuliano F. Panza, Cristina La Mura, Fabio Romanelli and Franco Vaccari</i> 321
Earthquake, Aftershocks <i>Mian Liu and Seth Stein</i>	229	Earthquakes, Volcanogenic <i>J. W. Neuberg</i> 329
Earthquake, Archaeoseismology <i>Klaus-G. Hinzen</i>	232	Electrical Properties of Rocks <i>Takashi Yoshino</i> 339
Earthquake, Focal Mechanism <i>Emile A. Okal</i>	236	Electrical Resistivity Surveys and Data Interpretation <i>M. H. Loke, D. F. Rucker, J. E. Chambers, P. B. Wilkinson and O. Kuras</i> 344
Earthquake, Foreshocks <i>Mian Liu</i>	241	Electromagnetic Methods, Imaging Magma Bodies <i>Kathryn A. Whaler</i> 350
Earthquake, Magnitude <i>Peter Borman</i>	243	Electromagnetic Pulsations and Magnetic Storms <i>G. S. Lakhina and B. T. Tsurutani</i> 354
Earthquakes and Crustal Deformation <i>Robert McCaffrey</i>	254	Electronic Geophysical Year <i>William K. Peterson, Daniel N. Baker, C. E. Barton, Peter Fox, Mark A. Parsons and Emily A. CoBabe-Ammann</i> 359
Earthquakes in the Himalaya <i>Dibyashakti Panda, Bhaskar Kundu and Vineet K. Gahalaut</i>	262	Energy Budget of the Earth <i>Jean-Claude Mareschal and Claude Jaupart</i> 361
Earthquakes on Oceanic Transform Faults, Unexpectedly Large <i>Shamita Das</i>	274	Energy Partitioning of Seismic Waves <i>Kalachand Sain</i> 368
Earthquakes, Did You Feel It? <i>David J. Wald, Vincent Quitoriano and James W. Dewey</i>	278	Equatorial Electrojet <i>Archana Bhattacharyya</i> 372
Earthquakes, Early and Strong Motion Warning <i>Richard M. Allen</i>	282	Fault Zone Guided Waves <i>Peter Eric Malin</i> 375

	CONTENTS	vii
Fiber Optic Distributed Strain Sensing for Seismic Applications <i>Thomas Reinsch, Philippe Jousset and Charlotte M. Krawczyk</i>	379	Geomagnetic Field, IGRF <i>Aude Chambodut</i> 500
Fiber-Optic Sensing in Geophysics, Temperature Measurements <i>Jan Henninges and Ali Masoudi</i>	384	Geomagnetic Field, Measurement Techniques <i>Mioara Manda and Anca Isac</i> 502
Floating Seismographs (MERMAIDS) <i>Yann Hello and Guust Nolet</i>	395	Geomagnetic Field, Polarity Reversals <i>Carlo Laj</i> 507
Fractal Scaling of Earthquakes <i>Simanchal Padhy and Vijay P. Dimri</i>	399	Geomagnetic Field, Secular Variation <i>Monika Korte</i> 514
Fractals and Chaos <i>Ravi P. Srivastava, Nimisha Vedanti and Vijay P. Dimri</i>	405	Geomagnetic Field, Theory <i>Friedrich H. Busse</i> 515
Free Oscillations of the Earth <i>Sarva Jit Singh and Sunita Rani</i>	412	Geomagnetically Induced Currents <i>G. S. Lakhina, R. Hajra and B. T. Tsurutani</i> 523
Geodesy, Figure of the Earth <i>Kusumita Arora</i>	423	Geophysical Well Logging <i>Miroslav Kobr</i> 527
Geodesy, Ground Positioning, and Levelling <i>Stelios P. Mertikas</i>	427	Geothermal Heat Pumps <i>Ladislaus Rybach</i> 537
Geodesy, Networks, and Reference Systems <i>Hayo Hase</i>	434	Geothermal Record of Climate Change <i>Michael G. Davis, David S. Chapman and Robert N. Harris</i> 541
Geodesy, Physical <i>V. Chakravarthi</i>	442	GPS, Data Acquisition, and Analysis <i>Carine Bruyninx, Wim Aerts and Juliette Legrand</i> 547
Geodetic Pendulums, Horizontal Ultra Broad Band <i>Carla Braitenberg</i>	447	GPS, Tectonic Geodesy <i>Jeffrey T. Freymueller</i> 558
Geodynamics <i>Alessandro M. Forte</i>	452	Gravimeters <i>Andrew Hugill</i> 578
Geoelectromagnetism <i>Antal Adam and Laszlo Szarka</i>	454	Gravity Anomalies, Interpretation <i>Mikhail K. Kaban</i> 585
Geoid <i>Paramesh Banerjee</i>	465	Gravity Data, Advanced Processing <i>Christopher J. Swain and Jonathan F. Kirby</i> 591
Geoid Determination, Computational Methods <i>Michael G. Sideris</i>	470	Gravity Data, Regional-Residual Separation <i>Kumarendra Mallick, Vasanthi Anthwar and Krishna Kant Sharma</i> 596
Geoid Determination, Theory and Principles <i>Michael G. Sideris</i>	476	Gravity Field of the Earth <i>Christopher Jekeli</i> 608
Geoid Undulation, Interpretation <i>Petr Vaniček</i>	482	Gravity Field, Temporal Variations from Space Techniques <i>Anny Cazenave, J. L. Chen and G. Ramillien</i> 621
Geomagnetic Excursions <i>Martha Schwartz</i>	486	Gravity Field, Time Variations from Surface Measurements <i>Virendra M. Tiwari and Jacques Hinderer</i> 626
Geomagnetic Field, Global Pattern <i>Susan Macmillan</i>	488	Gravity Measurements, Absolute <i>Mark A. Zumberge</i> 633

viii	CONTENTS	
Gravity Method, Airborne <i>Uwe Meyer</i>	637	International Gravity Formula <i>Hans-Jürgen Götze</i> 789
Gravity Method, Principles <i>Hans-Jürgen Götze</i>	640	International Polar Year 2007–2008 <i>David J. Carlson</i> 790
Gravity Method, Satellite <i>G. Balmino</i>	645	International Year of Planet Earth <i>Eduardo F. J. de Mulder and Wolfgang Eder</i> 792
Gravity Method, Surface <i>Dinesh Chandra Mishra and Virendra M. Tiwari</i>	656	Inverse Theory, Artificial Neural Networks <i>William A. Sandham and David J. Hamilton</i> 796
Gravity Modeling, Theory and Computation <i>Jean-Pierre Barriot and Lydie Sichoix</i>	662	Inverse Theory, Global Optimization <i>Mrinal K. Sen and Paul L. Stoffa</i> 807
Gravity, Data to Anomalies <i>Ron Hackney</i>	668	Inverse Theory, Linear <i>Pravin K. Gupta</i> 814
Gravity, Global Models <i>Nikolaos K. Pavlis</i>	677	Inverse Theory, Monte Carlo Method <i>Malcolm Sambridge and Kerry Gallagher</i> 821
Gravity, Gradiometry <i>Christopher Jekeli</i>	692	Inverse Theory, Singular Value Decomposition <i>Ajay Manglik</i> 827
Great Earthquakes <i>Roger M. W. Musson</i>	708	Isostasy <i>A. B. Watts</i> 831
Heat and Ground Water Flow <i>Vincenzo Pasquale, Massimo Verdoya and Paolo Chiozzi</i>	717	Isostasy, Thermal <i>Derrick Hasterok and David S. Chapman</i> 847
Heat Flow Determinations, Continental <i>John H. Sass and Graeme Beardsmore</i>	722	KTB Depth Laboratory: A Window into the Upper Crust 855 <i>Ulrich Harms and Jochem Kück</i>
Heat Flow, Continental <i>Paul Morgan</i>	727	Legal Continental Shelf: Geology, Geophysics, and Tectonics 861 <i>Elana Geddis, Ray Wood and Vaughan Stagpoole</i>
Heat Flow, Seafloor: Methods and Observations <i>Earl E. Davis and Andrew T. Fisher</i>	736	Lithosphere, Continental <i>David E. James</i> 866
Height Systems, Vertical Datums and Their Unification <i>Michael G. Sideris</i>	749	Lithosphere, Continental: Thermal Structure <i>Claude Jaupart and Jean-Claude Mareschal</i> 872
High-Frequency Seismology <i>Simanchal Padhy</i>	757	Lithosphere, Mechanical Properties <i>Evgueni Burov</i> 884
Impact Craters on Earth <i>Richard A. F. Grieve and Gordon R. Osinski</i>	769	Lithosphere, Oceanic <i>James S. McClain</i> 893
Instrumentation, Electrical Resistivity <i>M. H. Loke, O. Kuras, J. E. Chambers, D. F. Rucker and P. B. Wilkinson</i>	776	Lithosphere, Oceanic: Thermal Structure <i>Earl E. Davis and David S. Chapman</i> 906
Instrumentation, EM <i>Steven Constable</i>	782	Lithospheric Magnetic Anomalies from Satellite Data 915 <i>Stavros Kotsiaros</i>
International Geophysical Year <i>Ralph W. Baird</i>	786	Magnetic Anisotropy <i>Leonardo Sagnotti</i> 923

	CONTENTS	ix
Magnetic Anomalies: Interpretation <i>E. Thébault</i>	935 Mapping of Magnetic Variations from Space <i>Nandini Nagarajan</i>	1115
Magnetic Anomaly Map, Global <i>Kumar Hemant Singh</i>	943 Microcontinents <i>Carmen Gaina and Joanne Whittaker</i>	1120
Magnetic Data Enhancements and Depth Estimation <i>Clive Foss</i>	957 Numerical Methods for Flow in Fractured Porous Media <i>Luca Formaggia, Alessio Fumagalli and Anna Scotti</i>	1125
Magnetic Domains <i>Susan L. Halgedahl</i>	971 Numerical Methods, Boundary Element <i>Michele Cooke</i>	1130
Magnetic Gradiometry <i>Harald von der Osten-Woldenburg</i>	983 Numerical Methods, Domain Decomposition	1133
Magnetic Gradiometry in Archaeo-geophysics <i>Harald von der Osten-Woldenburg</i>	987 Alfio Quarteroni and Luca Formaggia	1137
Magnetic Methods, Airborne <i>Mike Dentith</i>	993 Numerical Methods, Finite Difference <i>Johan O. A. Robertsson and Joakim O. Blanch</i>	1145
Magnetic Methods, Principles <i>Kusumita Arora</i>	999 Numerical Methods, Finite Element <i>J. N. Reddy</i>	1149
Magnetic Methods, Surface <i>Nandini Nagarajan</i>	1007 Numerical Methods, Multigrid <i>Wim A. Mulder</i>	1155
Magnetic Modeling, Theory, and Computation <i>Mioara Manda, Carmen Gaina and Vincent Lesur</i>	1015 Ocean Bottom Seismics <i>Ingo A. Pecher, Jörg Bialas and Ernst R. Flueh</i>	1162
Magnetometers <i>Ivan Hrvoic</i>	1029 Ocean, Spreading Center <i>K. S. Krishna</i>	1166
Magnetotelluric Data Processing <i>Gary Egbert</i>	1036 Orogenic Belts <i>A. M. Celâl Şengör</i>	1187
Magnetotelluric Interpretation <i>John F. Hermance</i>	1042 Paleomagnetic Field Intensity <i>Andrew Biggin, Greig A. Paterson, Neil Suttie and John Shaw</i>	1193
Magnetotellurics, Crustal Imaging <i>Prasanta K. Patro</i>	1050 Paleomagnetism, Magnetostratigraphy <i>Donald R. Prothero</i>	1202
Magnetovariation Studies <i>Nandini Nagarajan</i>	1056 Paleomagnetism, Measurement Techniques and Instrumentation <i>T. Radhakrishna, J. D. A. Piper and Asanulla R. Mohamed</i>	1215
Mantle Convection <i>David Bercovici and Elvira Mulyukova</i>	1059 Paleomagnetism, Polar Wander <i>Jean Besse, Vincent Courtillot and Marianne Greff</i>	1225
Mantle D'' Layer <i>Thorne Lay</i>	1079 Paleomagnetism, Principles <i>William Lowrie</i>	1235
Mantle Plume: Spreading Ridge Interactions <i>K. S. Krishna and M. Ismaiel</i>	1085 Paleoseismology <i>Shinji Toda</i>	1248
Mantle Plumes <i>Cinzia G. Farnetani and Albrecht W. Hofmann</i>	1094 Plate Motions in Time: Inferences on Driving and Resisting Forces <i>Giampiero Iaffaldano and Hans-Peter Bunge</i>	
Mantle Viscosity <i>W. R. Peltier</i>	1107	

x	CONTENTS	
Plate Tectonics, Precambrian <i>Y. J. Bhaskar Rao, T. Vijaya Kumar and E. V. S. S. K. Babu</i>	1256	Seismic Discontinuities in the Transition Zone <i>Lev P. Vinnik</i> 1390
Plate-Driving Forces <i>Alessandro M. Forte</i>	1267	Seismic Hazard 1394 <i>Andrzej Kijko</i>
Plates and Paleoreconstructions <i>Alan G. Smith</i>	1274	Seismic Imaging, Overview 1407 <i>Gerard T. Schuster</i>
Poroelasticity <i>Ran Bachrach</i>	1280	Seismic Instrumentation 1419 <i>Duncan Carr Agnew</i>
Propagation of Elastic Waves: Fundamentals <i>Francisco J. Sánchez-Sesma and Ursula Iturrarán-Viveros</i>	1283	Seismic Microzonation 1425 <i>Fumio Yamazaki and Yoshihisa Maruyama</i>
Radioactivity in Earth's Core <i>V. Rama Murthy</i>	1293	Seismic Monitoring of Nuclear Explosions 1429 <i>Paul G. Richards, Zhongliang Wu, Won-Young Kim and David P. Schaff</i>
Radiogenic Heat Production in the Continental Crust <i>Claude Jaupart and Jean-Claude Mareschal</i>	1298	Seismic Noise 1442 <i>Dhananjay Kumar and Imtiaz Ahmed</i>
Radiogenic Heat Production of Rocks <i>Christoph Clauser</i>	1304	Seismic Phase Nomenclature: The IASPEI Standard 1447 <i>Johannes Schweitzer, Dmitry A. Storchak and Peter Borman</i>
Recovery of Source Magnetization Direction from Magnetic Field Data <i>Clive Foss</i>	1310	Seismic Properties of Rocks 1459 <i>Nikolas I. Christensen</i>
Remanent Magnetism <i>Laurie Brown and Suzanne McEnroe</i>	1319	Seismic Quiescence and Activation 1464 <i>Gennady Sobolev</i>
Remote Sensing and GIS Techniques for Tectonic Studies <i>Semere Solomon and Woldai Ghebreab</i>	1325	Seismic Ray Theory 1472 <i>Vlastislav Červený and Ivan Pšenčík</i>
Remote Sensing, Applications to Geophysics <i>Hojjatollah Ranjbar</i>	1330	Seismic Reservoir Characterization 1487 <i>Dhananjay Kumar</i>
SAR Interferometry <i>Masato Furuya</i>	1335	Seismic Seiches 1492 <i>Art McGarr</i>
Satellite Altimetry <i>Stelios P. Mertikas and Constantine Kokolakis</i>	1343	Seismic Signals in Well Water Observations 1493 <i>R. K. Chadha</i>
Seafloor Spreading <i>Richard N. Hey</i>	1349	Seismic Stereotomography 1498 <i>Gilles Lambaré and Thibaut Allemand</i>
Sedimentary Basins <i>Magdalena Scheck-Wenderoth</i>	1353	Seismic Structure at Mid-Ocean Ridges 1502 <i>Donald W. Forsyth</i>
Seismic Data Acquisition and Processing <i>Kabir Roy Chowdhury</i>	1365	Seismic Tomography 1507 <i>Guust Nolet</i>
Seismic Diffraction <i>Enru Liu</i>	1385	Seismic Velocity and Temperature Relationships 1511 <i>Kalachand Sain</i>
		Seismic Viscoelastic Attenuation 1512 <i>Vernon F. Cormier</i>

CONTENTS		xi
Seismic Wave Propagation in Real Media: Numerical Modeling Approaches <i>Ursula Iturrarán-Viveros and Francisco J. Sánchez-Sesma</i>	1525	Seismology, Rotational <i>William H. K. Lee</i> 1673
Seismic Waves, Scattering <i>Ludovic Margerin</i>	1537	Sentinel Satellites Initiate New Era in Earth Observatories 1685 <i>Randy Showstack</i>
Seismic Zonation <i>Yanxiang Yu, Mengtan Gao and Guangyin Xu</i>	1550	Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting 1687 <i>Stuart Crampin</i>
Seismic, Ambient Noise Correlation <i>Michel Campillo, Philippe Roux and Nikolai M. Shapiro</i>	1557	Single and Multichannel Seismics 1697 <i>Tamás Tóth</i>
Seismic, Artificial Intelligence to Neural Intelligence for Advanced Interpretation <i>Kalachand Sain and Priyadarshi Chinmoy Kumar</i>	1562	Slow Earthquake 1705 <i>Teruyuki Kato</i>
Seismic, Migration <i>Samuel H. Gray and Nittala Satyavani</i>	1567	Spherical Harmonic Analysis Applied to Potential Fields 1714 <i>Nikolaos K. Pavlis</i>
Seismic, Receiver Function Technique <i>Rainer Kind and Xiaohui Yuan</i>	1580	Statistical Seismology 1724 <i>David A. Rhoades, Annemarie Christophersen and Sebastian Hainzl</i>
Seismic, Reflectivity Method <i>Mrinal K. Sen</i>	1592	Subduction Zones 1728 <i>Geoffrey A. Abers</i>
Seismic, Super-Virtual Refraction Interferometry <i>Sherif M. Hanafy</i>	1602	Surface Waves 1738 <i>Barbara Romanowicz</i>
Seismic, Velocity, and Density Relationships <i>Kalachand Sain</i>	1607	T Waves 1751 <i>Emile A. Okal</i>
Seismic, Waveform Modeling and Tomography <i>Yanghua Wang and Ying Rao</i>	1608	Tectonic Geomorphology 1753 <i>Sampat Kumar Tandon and Vinee Srivastava</i>
Seismicity, Intraplate <i>Paul Bodin</i>	1621	Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat 1760 <i>Christoph Clauser</i>
Seismicity, Subduction Zone <i>Akira Hasegawa</i>	1625	Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity 1769 <i>Christoph Clauser</i>
Seismogram Interpretation <i>Ota Kulhanek and Leif Persson</i>	1635	Time-Dependent Seismic Hazard Assessment 1787 <i>Zhongliang Wu</i>
Seismological Networks <i>István Bondár and Eric Robert Engdahl</i>	1644	Time-Reversal in Seismology 1790 <i>Carène Larmat and Clarence S. Clay</i>
Seismology and Environment <i>Jean-Paul Montagner, Anne Mangeney and Eléonore Stutzmann</i>	1655	TOPO-EUROPE: From the Deep Earth to the Surface of Continental Europe and Its Margins 1794 <i>Sierd Cloetingh and TOPO-EUROPE Working Group</i>
Seismology, Global Earthquake Model <i>John F. Schneider, Jephraim Oro, Anselm Smolka, Peter Suhadolc and Zhongliang Wu</i>	1661	Total Electron Content in Seismo-ionospheric Studies 1802 <i>J. K. Catherine and R. Rajesh</i>
Seismology, Monitoring of CTBT <i>Zhongliang Wu and Paul G. Richards</i>	1669	

xii	CONTENTS	
Tracking Earth's Water in Motion from Satellite Gravity Observations <i>J. L. Chen</i>	1813	Unified Scaling Law for Earthquakes that Generalizes the Fundamental Gutenberg-Richter Relationship <i>Vladimir Kossobokov</i> 1893
Trans-European Suture Zone <i>H. Thybo and I. M. Artemieva</i>	1819	Vertical Seismic Profiling <i>James W. Rector III and Maria-Daphne Mangriotis</i> 1897
Traveltime Tomography Using Controlled-Source Seismic Data <i>Colin A. Zelt</i>	1828	Very Long Baseline Interferometry <i>Helmut Wiesemeyer and Axel Nothnagel</i> 1902
Tsunami <i>Steven N. Ward</i>	1848	Wavelet Analysis <i>Mikhail Kulesh</i> 1909
Tsunami Watch and Warning Centers <i>Shailesh Nayak, Srinivasa Kumar Tummala and E. Pattabhi Rama Rao</i>	1868	Author Index 1917 Subject Index 1921
Tsunami: Bay of Bengal <i>Vineet K. Gahalaut</i>	1880	
Uncertainties in Geomodelling Related to Geophysical Inversion <i>Hans-Jürgen Götze and Florian Wellmann</i>	1885	

About the Editor

Harsh K. Gupta is currently a member of the Atomic Energy Regulatory Board (India) and president of the Geological Society of India. He has previously been a member of the National Disaster Management Authority (India); secretary to the Government of India, Department of Ocean Development; director of the National Geophysical Research Institute, Hyderabad; vice chancellor of Cochin University of Science and Technology; and professor at the University of Texas at Dallas, USA. He is internationally respected for his work on artificial water reservoir-triggered earthquakes and developing criteria to distinguish them from normal earthquakes. As leader of the third Indian scientific expedition to Antarctica during 1983–1984, he succeeded in setting up India's first permanent Antarctic research station "Dakshin Gangotri" in just one Antarctic summer, which is still a record.

Prof. Gupta chaired the Steering Committee of the Global Seismic Hazard Program. After the disastrous 2004 Sumatra earthquake, he spearheaded the establishment of the Indian Tsunami Early Warning System. He has published over 200 articles in reviewed journals, authored 5 books that have been published by Elsevier and Springer. Prof. Gupta is a recipient of the Shanti Swarup Bhatnagar Prize, Waldo E. Smith Medal of American Geophysical Union, USSR Academy of Sciences "100 years of International Geophysics Memorial Medal," the Axford Gold Medal of the Asia Oceania Geosciences Society (AOGS), and the National Mineral Award of Excellence, among many honors. He is a fellow of the Indian National Science Academy, the World Academy of Sciences, and American Geophysical Union. Prof. Gupta has been president of the IUGG and the AOGS and is the founder president of the Asian Seismological Commission.

Contributors

Geoffrey A. Abers
Lamont-Doherty Earth Observatory
Columbia University
Palisades, NY, USA

Antal Adam
Geodetic and Geophysical Research, Institute of the
Hungarian Academy of Sciences
Sopron, Hungary

Wim Aerts
KU Leuven, Electrical Engineering (ESAT) TC
Diepenbeek Campus
Diepenbeek, Belgium

Duncan Carr Agnew
Institute of Geophysics and Planetary Physics, Scripps
Institution of Oceanography
University of California San Diego
La Jolla, CA, USA

Imtiaz Ahmed
BP
Houston, TX, USA

Thibaut Allemand
CGG
Massy, France

Richard M. Allen
Berkeley Seismological Laboratory, Department of Earth
and Planetary Science
University of California
Berkeley, CA, USA

Vasanthi Anthwar
National Geophysical Research Institute
Hyderabad, India

Kusumita Arora
National Geophysical Research Institute
Council of Scientific and Industrial Research (CSIR)
Hyderabad, India

I. M. Artemieva
Marine Geodynamics
GEOMAR Helmholtz Center for Ocean Research
Kiel, Germany
and
Department of Geophysics
Stanford University
Palo Alto, CA, USA

Estella Atekwana
College of Earth, Ocean, and Environment
University of Delaware
Newark, DE, USA

Jonathan M. Aurnou
Department of Earth and Space Sciences
University of California
Los Angeles, CA, USA

E. V. S. S. K. Babu
Geochronology Division
CSIR-National Geophysical Research Institute
Hyderabad, India

Ran Bachrach
Geophysics and Planetary Sciences Department
Tel Aviv University
Tel Aviv, Israel

Ralph W. Baird
Baird Petrophysical International
Houston, TX, USA

Daniel N. Baker
Laboratory for Atmospheric and Space Physics
University of Colorado
Boulder, CO, USA

G. Balmino
Centre National d'Etudes Spatiales
Toulouse, France

Paramesh Banerjee
Earth Observatory of Singapore
Prayukti Geomatics Pte Ltd
Singapore, Singapore

Jean-Pierre Barriot
Observatoire Géodésique de Tahiti
University of French Polynesia
Faaa-Tahiti, French Polynesia

C. E. Barton
Australian National University
Canberra, ACT, Australia

Graeme Beardmore
University of Melbourne
Parkville, VIC, Australia
and
Hot Dry Rocks Pty Ltd
Windsor, VIC, Australia

David Bercovici
Department of Earth and Planetary Sciences
Yale University
New Haven, CT, USA

Jean Besse
Laboratoire de Paléomagnétisme
Institut de Physique du Globe de Paris
Paris, France

Archana Bhattacharyya
Indian Institute of Geomagnetism
Navi Mumbai, India

Jörg Bialas
Leibniz-Institut für Meereswissenschaften
University of Kiel
Kiel, Germany

Andrew Biggin
Geomagnetism Lab, Department of Earth, Ocean and
Ecological Sciences
University of Liverpool
Liverpool, UK

Joakim O. Blanch
BHP
Houston, TX, USA

Paul Bodin
Pacific Northwest Seismic Network (PNSN)
University of Washington
Seattle, WA, USA

Sigrid Böhm
Research Division Higher Geodesy, Department of
Geodesy and Geoinformation
TU Wien
Vienna, Austria

Marco Bohnhoff
GFZ German Research Center for Geosciences
Geomechanics and Scientific Drilling
Potsdam, Germany

István Bondár
Research Centre for Astronomy and Earth Sciences
Geodetic and Geophysical Institute
Kövesligethy Radó Seismological Observatory
Budapest, Hungary

Peter Borman
(Deceased)

Carla Braitenberg
Department of Mathematics and Geosciences
University of Trieste
Trieste, Italy

Laurie Brown
Department of Geosciences
University of Massachusetts
Amherst, MA, USA

Carine Bruyninx
Department of Reference Systems and Planetology
Royal Observatory of Belgium
Brussels, Belgium

Hans-Peter Bunge
Geophysics Section, Department of Earth and
Environmental Sciences
Ludwig Maximilians University of Munich
Munich, Germany

Evgueni Burov
(Deceased)

Friedrich H. Busse
Institute of Physics, University of Bayreuth
Bayreuth, Germany

Michel Campillo
Observatoire de Grenoble
Université Joseph Fourier and CNRS
Grenoble, France

David J. Carlson
British Antarctic Survey
IPY International Programme Office
Cambridge, UK

Richard W. Carlson
Department of Terrestrial Magnetism
Carnegie Institution of Washington
Washington, DC, USA

J. K. Catherine
CSIR-NGRI
Hyderabad, India

Anny Cazenave
LEGOS, Observatoire Midi Pyrénées
Toulouse, France

Vlastislav Červený
Department of Geophysics, Faculty of Mathematics and
Physics
Charles University
Praha, Czech Republic

R. K. Chadha
CSIR:National Geophysical Research Institute
Hyderabad, India

V. Chakravarthi
Centre for Earth, Ocean and Atmospheric Sciences
University of Hyderabad
Hyderabad, India

J. E. Chambers
Geophysical Tomography Team
British Geological Survey
Keyworth, Nottingham, UK

Aude Chambodut
Institut de Physique du Globe de Strasbourg
Université de Strasbourg/Ecole et Observatoire des
Sciences de la Terre, CNRS
Strasbourg Cédex, France

David S. Chapman
Department of Geology and Geophysics
The University of Utah
Salt Lake City, UT, USA

J. L. Chen
Center for Space Research
University of Texas
Austin, TX, USA

Paolo Chiozzi
Department of Earth, Environment and Life Sciences
University of Genoa
Genoa, Italy

Nikolas I. Christensen
Department of Earth and Ocean Sciences
University of British Columbia
Vancouver, BC, Canada

Ulrich R. Christensen
Max-Planck-Institut für Sonnensystemforschung
Göttingen, Germany

Annemarie Christophersen
GNS Science
Lower Hutt, New Zealand

Christoph Clauser
Institute for Applied Geophysics and Geothermal Energy
RWTH Aachen University
Aachen, Germany

Clarence S. Clay
(Deceased)

Sierd Cloetingh
Tectonics Group, Department of Earth Sciences
Utrecht University
Utrecht, The Netherlands

Emily A. CoBabe-Ammann
Research and Innovations Office
University of Colorado
Boulder, CO, USA

Steven Constable
Institute of Geophysics and Planetary Physics
Scripps Institution of Oceanography
La Jolla, CA, USA

Michele Cooke
Geosciences Department
University of Massachusetts at Amherst
Amherst, MA, USA

Vernon F. Cormier
Physics Department
University of Connecticut
Storrs, CT, USA

Vincent Courtillot
Institut de Physique du Globe de Paris
Paris, France

Stuart Crampin
British Geological Survey
Murchison House
Edinburgh, UK

Shamita Das
Department of Earth Sciences
University of Oxford
Oxford, UK

Earl E. Davis
Pacific Geoscience Centre
Geological Survey of Canada
Sidney, BC, Canada

Michael G. Davis
Department of Physical Sciences
Arkansas Tech University
Russellville, AR, USA

Paul M. Davis
Earth and Space Sciences
UCLA
Los Angeles, CA, USA

Eduardo F. J. de Mulder
Earth Science Matters Foundation
Haarlem, The Netherlands

Mike Dentith
School of Earth Sciences (M004)
The University of Western Australia
Crawley, WA, Australia

John S. Derr
Tijeras, NM, USA

James W. Dewey
U.S. Geological Survey, Denver Federal Center
Lakewood, CO, USA

Domenico Di Giacomo
Department 2: Physics of the Earth
GFZ German Research Center for Geosciences
Potsdam, Germany

Vijay P. Dimri
CSIR-National Geophysical Research Institute
Hyderabad, India

Wolfgang Eder
GeoCentre, Geobiology
University of Göttingen
Göttingen, Germany

Gary Egbert
College of Oceanic and Atmospheric Sciences
Oregon State University
Corvallis, OR, USA

Eric Robert Engdahl
Center for Imaging the Earth's Interior, Department of
Physics
University of Colorado at Boulder
Boulder, CO, USA

Cinzia G. Farnetani
Institut de Physique du Globe de Paris
Paris, France

Andrew T. Fisher
Department of Earth and Planetary Sciences
University of California at Santa Cruz
Santa Cruz, CA, USA

Ernst R. Flueh
Leibniz-Institut für Meereswissenschaften
University of Kiel
Kiel, Germany

Luca Formaggia
MOX, Department of Mathematics
Politecnico di Milano
Milan, Italy

Donald W. Forsyth
Department of Geological Sciences
Brown University
Providence, RI, USA

Alessandro M. Forte
Department of Geological Sciences
University of Florida
Gainesville, FL, USA

Clive Foss
Mineral Resources
CSIRO
Lindfield, NSW, Australia

Peter Fox
Rensselaer Polytechnic Institute
Troy, NY, USA

Friedemann T. Freund
Code SCR
NASA Ames Research Center/San Jose State University
Moffett Field, CA, USA

Jeffrey T. Freymueller
Department of Earth and Environmental Sciences
Michigan State University
East Lansing, MI, USA

Alessio Fumagalli
Department of Mathematics
Politecnico di Milano
Milan, Italy

Masato Furuya
Department of Natural History Sciences
Hokkaido University
Sapporo, Japan

Vineet K. Gahalaut
National Geophysical Research Institute (CSIR)
Hyderabad, India

Carmen Gaina
Department of Geosciences, CEED – Centre for Earth
Evolution and Dynamics
University of Oslo
Oslo, Norway

Kerry Gallagher
UMR 6118 – Géosciences Rennes
Geosciences, Université de Rennes 1
Rennes Cedex, France

Mengtian Gao
Institute of Geophysics
China Earthquake Administration
Beijing, China

Edward J. Garnero
School of Earth and Space Exploration
Arizona State University
Tempe, AZ, USA

Elana Geddis
Harbour Chambers
Wellington, New Zealand

Woldai Ghebreab
Department of Geology and Environmental Science
University of Akron
Akron, OH, USA

Hans-Jürgen Götze
Institute of Geosciences, Working group Satellite- and
Aerogeophysics
Christian-Albrechts-Universität zu Kiel
Kiel, Germany

Samuel H. Gray
CGGVeritas
Calgary, AB, Canada

Marianne Greff
Institut de Physique du Globe de Paris
Paris, France

Richard A. F. Grieve
Department of Earth Sciences
University of Western Ontario
London, ON, Canada
and
Institute for Earth and Space Exploration
University of Western Ontario
London, ON, Canada

Gottfried Grünthal
Helmholtz Centre Potsdam
German Research Centre for Geosciences GFZ
Potsdam, Germany

Harsh K. Gupta
National Geophysical Research Institute
Council of Scientific and Industrial Research
Hyderabad, India

Pravin K. Gupta
Department of Earth Sciences
Indian Institute of Technology Roorkee
Roorkee, Uttarakhand, India

Ron Hackney
Petroleum and Marine Division
Geoscience Australia
Canberra, ACT, Australia

Sebastian Hainzl
GFZ German Research Centre for Geosciences
Potsdam, Germany

R. Hajra
Indian Institute of Technology Indore
Simrol, Indore, India

Susan L. Halgedahl
Department of Geology and Geophysics
University of Utah
Salt Lake City, UT, USA

David J. Hamilton
Consultraining
Edinburgh, UK

Sherif M. Hanafy
College of Petroleum Engineering and Geosciences (CPG)
King Fahd University of Petroleum and Minerals
(KFUPM)
Dhahran, Saudi Arabia

Ulrich Harms
Scientific Drilling
GFZ German Research Centre for Geosciences
Potsdam, Germany

Robert N. Harris
College of Earth, Ocean, and Atmospheric Sciences
Oregon State University
Corvallis, OR, USA

Hayo Hase
Geodätisches Observatorium Wettzell
Bundesamt für Kartographie und Geodäsie
Bad Kötzing, Germany

Akira Hasegawa
RCPEV, Graduate School of Science
Tohoku University
Sendai, Japan

Derrick Hasterok
Earth Sciences Department
University of Adelaide
Adelaide, SA, Australia

Michael Hearne
U.S. Geological Survey
Denver Federal Center
Lakewood, CO, USA

Yann Hello
IRD Geoazur and Université Côte d'AZUR
Sophia Antipolis, France

Jan Henniges
Geoenergy
GFZ German Research Centre for Geosciences
Potsdam, Germany

John F. Hermance
Department of Geological Sciences
Brown University
Providence, RI, USA

Richard N. Hey
Hawaii Institute of Geophysics and Planetology, School of
Ocean and Earth Science and Technology
University of Hawaii
Honolulu, HI, USA

Jacques Hinderer
Institut de Physique du Globe de Strasbourg, CNRS
Université de Strasbourg
Strasbourg, France

Klaus-G. Hinzen
Department of Geosciences
University of Cologne
Cologne, Germany

Albrecht W. Hofmann
Lamont-Doherty Earth Observatory
Columbia University
Palisades, NY, USA
and
Max-Planck-Institut für Chemie
Mainz, Germany

Ivan Hrvoic
GEM Systems, Inc
Markham, ON, Canada

Andrew Hugill
Scintrex Ltd
Concord, ON, Canada

Giampiero Iaffaldano
Department of Geosciences and Natural Resource
Management
University of Copenhagen
Copenhagen, Denmark

Anca Isac
Geological Institute of Romania
Surlari National Geomagnetic Observatory
Bucharest, Romania

M. Ismaiel
Centre for Earth, Ocean and Atmospheric Sciences
University of Hyderabad
Hyderabad, India

Alik Ismail-Zadeh
Institute of Applied Geosciences
Karlsruhe Institute of Technology
Karlsruhe, Germany
and

Institute of Earthquake Prediction Theory and
Mathematical Geophysics
Russian Academy of Sciences (IEPT RAS)
Moscow, Russia

Ursula Iturrarán-Viveros
Facultad de Ciencias, Departamento de Matemáticas
Universidad Nacional Autónoma de México, Ciudad
Universitaria
México City, Mexico

David D. Jackson
Department of Earth and Space Sciences
University of California Los Angeles
Los Angeles, CA, USA

Kishor S. Jaiswal
U.S. Geological Survey
Denver Federal Center
Lakewood, CO, USA

David E. James
Department of Terrestrial Magnetism
Carnegie Institution of Washington
Washington, DC, USA

Claude Jaupart
Université de Paris Institut de Physique du Globe
Paris, France

Christopher Jekeli
Division of Geodetic Science, School of Earth Sciences
Ohio State University
Columbus, OH, USA

Philippe Jousset
GFZ German Research Centre for Geosciences
Potsdam, Germany

Mikhail K. Kaban
Department 1: Geodesy and Remote Sensing, Sektion 1.3:
Earth-System Modelling
Deutsches GeoForschungsZentrum Potsdam – GFZ
Potsdam, Germany

Yan Y. Kagan
Department of Earth and Space Sciences
University of California Los Angeles
Los Angeles, CA, USA

Masashi Kamogawa
Division for Earthquake Prediction Research
Global Center for Asian and Regional Research,
University of Shizuoka
Shizuoka, Japan

Teruyuki Kato
Earthquake Research Institute
The University of Tokyo
Bunkyo-ku, Tokyo, Japan

Hartmut Kern
Kiel, Germany

Andrzej Kijko
Natural Hazard Centre, Africa
University of Pretoria
Pretoria, Republic of South Africa

Won-Young Kim
Lamont-Doherty Earth Observatory
Columbia University
Palisades, NY, USA

Rainer Kind
Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences
Potsdam, Germany
and
Freie Universität
Fachrichtung Geophysik
Berlin, Germany

Jonathan F. Kirby
School of Earth and Planetary Sciences
Curtin University
Perth, WA, Australia

Miroslav Kobl
Institute of Hydrogeology, Engineering Geology and
Applied Geophysics
Charles University in Prague
Prague, Czech Republic

Constantine Kokolakis
Geodesy and Geomatics Engineering Laboratory, School
of Mineral Resources Engineering
Technical University of Crete
Chania, Greece

Monika Korte
Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences
Potsdam, Germany

Vladimir Kossobokov
Institute of Earthquake Prediction Theory and
Mathematical Geophysics
Russian Academy of Sciences (IEPT RAS)
Moscow, Russia

Stavros Kotsiaros
Planetary Magnetospheres Lab
NASA Goddard Space Flight Center
Greenbelt, MD, USA
and
University of Maryland College Park
College Park, MD, USA
and
Technical University of Denmark (DTU)
Kongens Lyngby, Denmark

Charlotte M. Krawczyk
Institute for Applied Geosciences
Technical University Berlin
Berlin, Germany
and
GFZ German Research Centre for Geosciences
Potsdam, Germany

K. S. Krishna
Centre for Earth, Ocean and Atmospheric Sciences
University of Hyderabad
Hyderabad, India

Jochem Kück
Scientific Drilling
GFZ German Research Centre for Geosciences
Potsdam, Germany

Mikhail Kulesh
Information Technology
DIS AG
Mahlow, Germany

Ota Kulhanek
Department of Earth Sciences, Section of Seismology
Uppsala University
Uppsala, Sweden

Dhananjay Kumar
Chevron
Houston, TX, USA
and
BP
Houston, TX, USA

Priyadarshi Chinmoy Kumar
Wadia Institute of Himalayan Geology
Dehradun, India

T. Vijaya Kumar
Geochronology Division
CSIR-National Geophysical Research Institute
Hyderabad, India

Bhaskar Kundu
Department of Earth and Atmospheric Sciences
NIT Rourkela
Rourkela, India

O. Kuras
Geophysical Tomography Team
British Geological Survey
Keyworth, Nottingham, UK

Cristina La Mura
Department of Geosciences
University of Trieste
Trieste, Italy

Carlo Laj
Laboratoire des Sciences du Climat
Unité mixte CEA-CNRS-UVSQ
Gif-sur-Yvette, France

G. S. Lakhina
Indian Institute of Geomagnetism
New Panvel (W), Navi Mumbai, India

Gilles Lambaré
CGG
Massy, France

Carène Larmat
Geophysics Group, EES-17, MS D542
Los Alamos National Laboratory
Los Alamos, NM, USA

Stanislaw Lasocki
Department of Seismology
Institute of Geophysics Polish Academy of Sciences
Warsaw, Poland

Thorne Lay
Earth and Planetary Sciences Department
University of California
Santa Cruz, CA, USA

William H. K. Lee
U.S. Geological Survey
Menlo Park, CA, USA

Juliette Legrand
Department of Reference Systems and Planetology
Royal Observatory of Belgium
Brussels, Belgium

Vincent Lesur
Université de Paris, Institut de physique du globe de Paris,
CNRS
Paris, France

Kuo-wan Lin
U.S. Geological Survey
Denver Federal Center
Lakewood, CO, USA

Enru Liu
China University of Mining and Technology
Xuzhou, China

Mian Liu
Department of Geological Sciences
University of Missouri
Columbia, MO, USA

M. H. Loke
Geotomo Software Sdn. Bhd.
Penang, Gelugor, Malaysia

William Lowrie
Institute of Geophysics
Swiss Federal Institute of Technology
Zürich, Switzerland

Susan Macmillan
Geomagnetism Team
British Geological Survey
Edinburgh, UK

Raul Madariaga
Ecole Normale Supérieure, Université PSL
Paris, France

Peter Malin
GFZ German Research Center for Geosciences
Geomechanics and Scientific Drilling
Potsdam, Germany

Peter Eric Malin
Earth and Ocean Sciences
Duke University
Durham, NC, USA

Kumarendra Mallick
National Geophysical Research Institute
Hyderabad, India

Prantik Mandal
Earthquake Seismology
CSIR-National Geophysical Research Institute
Hyderabad, Telangana, India

Mioara Manda
CNES – Centre National d'Etudes Spatiales
Paris, France

Anne Mangeney
Seismological Laboratory, Institut de Physique du Globe
Université de Paris, UMR CNRS/7154
Paris, France

Ajay Manglik
National Geophysical Research Institute (Council of
Scientific and Industrial Research)
Hyderabad, India

Maria-Daphne Mangriotis
School of Geosciences and Grant Institute of Earth Science
University of Edinburgh
Edinburgh, UK
and
Institute of GeoEnergy Engineering
Heriot-Watt University
Edinburgh, UK

Kristin D. Marano
U.S. Geological Survey
Denver Federal Center
Lakewood, CO, USA

Jean-Claude Mareschal
Centre GEOTOP-UQAM
University of Québec
Montréal, QC, Canada

Ludovic Margerin
Institut de Recherche en Astrophysique et Planétologie
Observatoire Midi-Pyrénées/C.N.R.S.
Toulouse, France

Yoshihisa Maruyama
Graduate School of Engineering
Chiba University
Inage-ku, Chiba, Japan

Ali Masoudi
ORC
University of Southampton
Southampton, UK

Robert McCaffrey
Department of Geology
Portland State University
Portland, OR, USA

James S. McClain
Department of Earth and Planetary Sciences
University of California, Davis
Davis, CA, USA

Suzanne McEnroe
Norwegian Geological Survey
Trondheim, Norway

Art McGarr
Earthquake Science Center
U.S. Geological Survey
Menlo Park, CA, USA

Allen K. McNamara
Department of Earth and Environmental Sciences
Michigan State University
East Lansing, MI, USA

Rolf Meissner
(Deceased)

Stelios P. Mertikas
Geodesy and Geomatics Engineering Laboratory, School
of Mineral Resources Engineering
Technical University of Crete
Chania, Greece

Uwe Meyer
Sub-Department Geophysical Reconnaissance –
Resources and Near Surface Processes
Federal Agency for Geosciences and Mineral Resources
Hannover, Germany

Andrew J. Michael
USGS
Moffett Field, CA, USA

Dinesh Chandra Mishra
(Deceased)

Asanulla R. Mohamed
National Centre for Earth Science Studies
Trivandrum, India

Jean-Paul Montagner
Seismological Laboratory, Institut de Physique du Globe
University Paris-Diderot, UMR CNRS/7154
Paris, France

Paul Morgan
Colorado Geological Survey
Denver, CO, USA

Wim A. Mulder
Faculty of Civil Engineering and Geosciences, Department
of Geoscience and Engineering
Delft University of Technology
Delft, The Netherlands
and
Shell Global Solutions International BV
Amsterdam, The Netherlands

Elvira Mulyukova
Department of Earth and Planetary Sciences
Yale University
New Haven, CT, USA

Roger M. W. Musson
British Geological Survey
Edinburgh, UK
and
School of Geosciences
University of Edinburgh
Edinburgh, UK

Toshiyasu Nagao
Institute of Oceanic Research and Development
Tokai University
Shizuoka, Japan

Nandini Nagarajan
National Geophysical Research Institute
Hyderabad, India

Shailesh Nayak
National Institute of Advanced Studies (NIAS)
Bengaluru, India

J. W. Neuberg
Institute of Geophysics and Tectonics, School of Earth and
Environment
University of Leeds
Leeds, UK

Guust Nolet
IRD Geoazur and Université Côte d'Azur
Sophia Antipolis, France

Axel Nothnagel
VLBI Research Group, Institute of Geodesy and
Geoinformation
University of Bonn
Bonn, Germany

Emile A. Okal
Department of Earth and Planetary Sciences
Northwestern University
Evanston, IL, USA

Beata Orlecka-Sikora
Department of Seismology
Institute of Geophysics Polish Academy of Sciences
Warsaw, Poland

Jephraim Oro
GEM Foundation
Pavia, Italy

Gordon R. Osinski
Department of Earth Sciences
University of Western Ontario
London, ON, Canada
and
Institute for Earth and Space Exploration
University of Western Ontario
London, ON, Canada

Simanchal Padhy
Seismological Observatory
CSIR-National Geophysical Research Institute
Hyderabad, India

Dibyashakti Panda
Department of Earth and Atmospheric Sciences
NIT Rourkela
Rourkela, India

Giuliano F. Panza
Department of Geosciences
University of Trieste
Trieste, Italy
and
Earth System Physics Section/Sand Group
The Abdus Salam International Centre for Theoretical
Physics
Trieste, Italy

Mark A. Parsons
Rensselaer Polytechnic Institute
Troy, NY, USA

Vincenzo Pasquale
Department of Earth, Environment and Life Sciences
University of Genoa
Genoa, Italy

Greig A. Paterson
Geomagnetism Lab, Department of Earth, Ocean and
Ecological Sciences
University of Liverpool
Liverpool, UK

Prasanta K. Patro
CSIR-National Geophysical Research Institute
Hyderabad, India

Nikolaos K. Pavlis
Geodesy and Geophysics Basic and Applied Research
National Geospatial-Intelligence Agency (NGA)
Reston, VA, USA

Ingo A. Pecher
GNS Science
Lower Hutt, New Zealand
and
Institute of Petroleum Engineering
Heriot-Watt University
Edinburgh, UK

W. R. Peltier
Department of Physics
University of Toronto
Toronto, ON, Canada

Leif Persson
Department of Underwater Research
FOI
Stockholm, Sweden

William K. Peterson
Laboratory for Atmospheric and Space Physics
University of Colorado
Boulder, CO, USA

J. D. A. Piper
Geomagnetism Laboratory, Department of Earth and
Ocean Sciences
University of Liverpool
Liverpool, UK

Donald R. Prothero
Department of Geology
Occidental College
Los Angeles, CA, USA

Ivan Pšenčík
Department of Geophysics, Faculty of Mathematics and
Physics
Charles University
Praha, Czech Republic
and
Institute of Geophysics
Academy of Sciences of Czech Republic
Praha, Czech Republic

Alfio Quarteroni
MOX, Department of Mathematics
Politecnico di Milano
Milan, Italy
and
CMCS-MATHICSE, EPFL
Lausanne, Switzerland

Vincent Quitoriano
U.S. Geological Survey, Denver Federal Center
Lakewood, CO, USA

T. Radhakrishna
National Centre for Earth Science Studies
Trivandrum, India
and
GITAM University
Nagadenehalli, Bengaluru, India

R. Rajesh
CSIR-NGRI
Hyderabad, India

V. Rama Murthy
(Deceased)

G. Ramillien
GET, Observatoire Midi Pyrénées
Toulouse, France

Sunita Rani
Department of Mathematics
Guru Jambheshwar University of Science and Technology
Hisar, India

Hojjatollah Ranjbar
Department of Mining Engineering
Shahid Bahonar University of Kerman
Kerman, Iran

E. Pattabhi Rama Rao
Earth System Science Organisation–Indian National
Centre for Ocean Information Services (ESSO–INCOIS)
Hyderabad, India

Ying Rao
College of Geophysics
China University of Petroleum (Beijing)
Beijing, China

Y. J. Bhaskar Rao
Geochronology Division
CSIR-National Geophysical Research Institute
Hyderabad, India

James W. Rector III
Department of Civil and Environmental Engineering
University of California at Berkeley
Berkeley, CA, USA

J. N. Reddy
Department of Mechanical Engineering
Texas A&M University
College Station, TX, USA

Thomas Reinsch
GFZ German Research Centre for Geosciences
Potsdam, Germany

David A. Rhoades
GNS Science
Lower Hutt, New Zealand

Paul G. Richards
Lamont-Doherty Earth Observatory
Columbia University
Palisades, NY, USA

Paul H. Roberts
Institute of Geophysics and Planetary Physics
University of California
Los Angeles, CA, USA

Johan O. A. Robertsson
Institute of Geophysics, Department of Earth Sciences
ETH-Zürich
Zürich, Switzerland

Fabio Romanelli
Department of Geosciences
University of Trieste
Trieste, Italy

Barbara Romanowicz
Berkeley Seismological Laboratory
Berkeley, CA, USA
and
Collège de France
Paris, France

Philippe Roux
Observatoire de Grenoble
Université Joseph Fourier and CNRS
Grenoble, France

Kabir Roy Chowdhury
Department of Earth Sciences
Utrecht University
Utrecht, The Netherlands

D. F. Rucker
hydroGEOPHYSICS, Inc.
Tucson, AZ, USA

Ladislaus Rybach
Institute of Geophysics
ETH Zurich
Zurich, Switzerland

Leonardo Sagnotti
Istituto Nazionale di Geofisica e Vulcanologia
Rome, Italy

Kalachand Sain
Wadia Institute of Himalayan Geology
Dehradun, India

Malcolm Sambridge
Seismology and Mathematical Geophysics
Research School of Earth Sciences, The Australian
National University
Canberra, ACT, Australia

Francisco J. Sánchez-Sesma
Instituto de Ingeniería
Universidad Nacional Autónoma de México, Ciudad
Universitaria
México City, Mexico

William A. Sandham
Scotsig
Glasgow, UK

John H. Sass
(Deceased)

Nittala Satyavani
Seismic Group
National Geophysical Research Institute
Hyderabad, India

David P. Schaff
Lamont-Doherty Earth Observatory
Columbia University
Palisades, NY, USA

Magdalena Scheck-Wenderoth
Helmholtz Centre Potsdam, GFZ German Research Centre
for Geosciences
Potsdam, Germany
and
RWTH Aachen University
Aachen, Germany

John F. Schneider
GEM Foundation
Pavia, Italy

Harald Schuh
Department 1 “Geodesy” at Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences
Potsdam, Germany

Gerard T. Schuster
Division of Environmental and Earth Sciences
King Abdullah University of Science and Technology
Thule, Saudi Arabia

Martha Schwartz
Department of Earth Sciences
University of Southern California
Los Angeles, CA, USA

Johannes Schweitzer
NORSAR
Kjeller, Norway

Anna Scotti
Department of Mathematics
Politecnico di Milano
Milan, Italy

Mahmoud Mohamed Selim Saleh
Department of Nature and Applied Sciences
Al-Aflaj Community College, AL-Kharj University
Al-Aflaj, Riyadh, Saudi Arabia

Mrinal K. Sen
Department of Geological Sciences, UT Institute for
Geophysics, Jackson School of Geosciences
The University of Texas at Austin
Austin, TX, USA

A. M. Celâl Şengör
Faculty of Mines, Department of Geology and Eurasia
Institute of Earth Sciences
Istanbul Technical University
Istanbul, Turkey

Nikolai M. Shapiro
Institut de Physique du Globe de Paris
Paris, France

Krishna Kant Sharma
Department of Applied Geology
University of Madras
Chennai, India

John Shaw
Geomagnetism Lab, Department of Earth, Ocean and
Ecological Sciences
University of Liverpool
Liverpool, UK

Randy Showstack
American Geophysical Union
Washington, DC, USA
and
Independent Journalist
Washington, DC, USA

Lydie Sichoix
Observatoire Géodésique de Tahiti
University of French Polynesia
Faaa-Tahiti, French Polynesia

Michael G. Sideris
Department of Geomatics Engineering
University of Calgary
Calgary, AB, Canada

Kumar Hemant Singh
Department of Earth Sciences
Indian Institute of Technology Bombay
Mumbai, India

Sarva Jit Singh
Indian National Science Academy
New Delhi, India

Satish C. Singh
Laboratoire de Géoscience Marines
Institut de Physique du Globe de Paris
Paris, France

Lee Slater
Department of Earth and Environmental Sciences
Rutgers University Newark
Newark, NJ, USA

Daniel Slosky
U.S. Geological Survey
Denver Federal Center
Lakewood, CO, USA

Alan G. Smith
(Deceased)

Anselm Smolka
GEM Foundation
Pavia, Italy

Gennady Sobolev
Institute of Physics of the Earth
Russian Academy of Sciences
Moscow, Russia

Semere Solomon
Det Norske Veritas
DNV Research and Innovation
Høvik, Norway

Xiaodong Song
Department of Geology
University of Illinois at Urbana-Champaign
Urbana, IL, USA

Ravi P. Srivastava
Equinor ASA
Bergen, Norway

Vineet Srivastava
Department of Earth and Environmental Sciences
IISER Bhopal
Bhopal, India

Frank D. Stacey
Division of Exploration and Mining
CSIRO
Kenmore, Australia

Vaughan Stagpoole
GNS Science
Lower Hutt, New Zealand

Seth Stein
Department of Earth and Planetary Sciences
Northwestern University
Evanston, IL, USA

France St-Laurent
LaSalle, QC, Canada

Paul L. Stoffa
Department of Geological Sciences, UT Institute for
Geophysics, Jackson School of Geosciences
The University of Texas at Austin
Austin, TX, USA

Dmitry A. Storchak
International Seismological Centre (ISC)
Thatcham, UK

Eléonore Stutzmann
Seismological Laboratory, Institut de Physique du Globe
Université de Paris, UMR CNRS/7154
Paris, France

Peter Suhadolc
Department of Geosciences
University of Trieste
Trieste, Italy

Neil Suttie
School of Archaeological and Forensic Sciences
University of Bradford
Bradford, UK

Christopher J. Swain
School of Earth and Planetary Sciences
Curtin University
Perth, WA, Australia

Laszlo Szarka
Geodetic and Geophysical Research, Institute of the
Hungarian Academy of Sciences
Sopron, Hungary

Sampat Kumar Tandon
Department of Earth and Environmental Sciences
IISER Bhopal
Bhopal, India

Donald H. Tarling
School of Earth, Ocean and Environmental Sciences
University of Plymouth
Plymouth, UK

E. Thébault
Laboratoire Magma et Volcans, Université Clermont
Auvergne (UCA), UMR CNRS 6524, Campus des
Cézeaux, 6, Avenue Blaise Pascal
63178 AUBIERE, FRANCE

Robert Thériault
Québec Ministry of Energy and Natural Resources
Québec, QC, Canada

Eric M. Thompson
U.S. Geological Survey
Denver Federal Center
Lakewood, CO, USA

Clifford H. Thurber
Department of Geoscience
University of Wisconsin-Madison
Madison, WI, USA

H. Thybo
Eurasia Institute of Earth Sciences
Istanbul Technical University
Istanbul, Turkey
and
Centre for Earth Evolution and Dynamics (CEED)
University of Oslo
Oslo, Norway

Virendra M. Tiwari
Gravity and Magnetic Studies Group
CSIR- National Geophysical Research Institute (CSIR)
Hyderabad, India

Harold Tobin
Department of Earth and Space Sciences
University of Washington
Seattle, WA, USA

Shinji Toda
Disaster Prevention Research Institute
Kyoto University
Uji, Kyoto, Japan

Tamás Tóth
Geomega Limited
Budapest, Hungary

B. T. Tsurutani
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA, USA

Srinivasa Kumar Tummala
Earth System Science Organisation–Indian National
Centre for Ocean Information Services (ESSO–INCOIS)
Hyderabad, India
and
ICG/IOTWMS Secretariat
UNESCO IOC Perth Regional Programme Office
West Perth, WA, Australia

James A. Tyburczy
School of Earth and Space Exploration
Arizona State University
Tempe, AZ, USA

Seiya Uyeda
Japan Academy
Tokyo, Japan

Franco Vaccari
Department of Geosciences
University of Trieste
Trieste, Italy

Petr Vaníček
Department of Geodesy and Geomatics Engineering
University of New Brunswick
Fredericton, NB, Canada

Nimisha Vedanti
CSIR-National Geophysical Research Institute
Hyderabad, India

Massimo Verdoya
Department of Earth, Environment and Life Sciences
University of Genoa
Genoa, Italy

Lev P. Vinnik
Institute of Physics of the Earth
Moscow, Russia

Harald von der Osten-Woldenburg
National Heritage Department
Regional Government of Baden-Wuerttemberg
Esslingen am Neckar, Germany

John M. Wahr
(Deceased)

David J. Wald
U.S. Geological Survey
Denver Federal Center
Lakewood, CO, USA

Yanghua Wang
Resource Geophysics Academy, Department of Earth
Science and Engineering
Imperial College London
London, UK

Steven N. Ward
Institute of Geophysics and Planetary Physics
University of California at Santa Cruz
Santa Cruz, CA, USA

A. B. Watts
Department of Earth Sciences
University of Oxford
Oxford, UK

Florian Wellmann
Computational Geoscience and Reservoir Engineering
RWTH Aachen University
Aachen, Germany

Lianxing Wen
Department of Geosciences
State University of New York at Stony Brook
Stony Brook, NY, USA

Kathryn A. Whaler
School of GeoSciences
University of Edinburgh
Edinburgh, UK

Joanne Whittaker
Institute for Marine and Antarctic Studies (IMAS)
University of Tasmania
Hobart, TAS, Australia

Helmut Wiesemeyer
Millimeter and Submillimeter Astronomy
Max-Planck-Institute for Radio Astronomy
Bonn, Germany

P. B. Wilkinson
Geophysical Tomography Team
British Geological Survey
Keyworth, Nottingham, UK

Ray Wood
CRP-OCS Ltd
Haumoana, New Zealand

C. Bruce Worden
U.S. Geological Survey
Denver Federal Center
Lakewood, CO, USA

Zhongliang Wu
Institute of Earthquake Forecasting
China Earthquake Administration
Beijing, China

Guangyin Xu
Institute of Geophysics
China Earthquake Administration
Beijing, China

Fumio Yamazaki
National Research Institute for Earth Science and Disaster
Resilience
Tsukuba, Ibaraki, Japan

Takashi Yoshino
Institute for Planetary Materials
Okayama University
Misasa, Japan

Yanxiang Yu
Institute of Geophysics
China Earthquake Administration
Beijing, China

Xiaohui Yuan
Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences
Potsdam, Germany

Colin A. Zelt
Department of Earth Science
Rice University
Houston, TX, USA

Mark A. Zumberge
Institute of Geophysics and Planetary Physics, Scripps
Institution of Oceanography
University of California, San Diego
La Jolla, CA, USA

Preface to the Second Edition

The oldest surviving encyclopedia is *Naturalis Historia* by Pliny the Elder, a Roman statesman in the first century AD. It has 37 chapters covering the history, geography, geology, and various other aspects of the world. In today's digital world of easy access to information, encyclopedias continue to be a valued source of reliable content contributed by experts. Subject encyclopedias contain in-depth entries offering essential and up-to-date information on the respective field of study, reflecting the current state of knowledge in the discipline as well as current trends. Encyclopedias published in electronic format are much easier to use than printed versions, given their search functions.

All information about the Earth's interior comes from field observations and measurements made within the top few kilometers of the surface, from laboratory experiments and from the powers of human deduction, relying on complex numerical modeling. Solid Earth geophysics encompasses all these endeavors and aspires to define and quantify the internal structure and processes of the Earth in terms of physical principles, corresponding mathematical formulations, and computational procedures. The role of solid Earth geophysics has gained prominence with increasing recognition of the fact that understanding Earth processes is central to the continued well-being of humanity. Apart from the exploration of natural resources, geophysical investigations provide crucial insights regarding the mutual relationships between climate and tectonics and on the effects of global change in terms of a broad range of natural hazards. The field continues to grow rapidly, both in fundamental and applied aspects, parallel to advances in allied fields of science and technology.

The second edition of the *Encyclopedia of Solid Earth Geophysics (ESEG)* serves as a comprehensive compendium of information on important topics in solid earth geophysics and provides systematic and up-to-date

coverage of central concepts and key topics of interest, while making no claims to represent every aspect of this multi-disciplinary and multi-faceted field. The earliest version of the *Encyclopedia of Solid Earth Geophysics*, edited by Prof. David E. James, was published by Van Nostrand Reinhold publishing company in 1989. The extraordinary growth and diversification of the field over the next 20 years called for a major revision. A completely new work, namely *Encyclopedia of Solid Earth Geophysics*, published by Springer in 2011, was edited by Harsh K. Gupta. It brought together over 200 entries covering established and new concepts in geophysics across sub-fields such as gravity, geodesy, geoelectricity, geomagnetism, seismology, seismics, deep earth interior and processes, plate tectonics, geothermics, and computational methods in a consistent format. Less than 10 years later, the need for an updated second edition was driven by the desire to expand the scope of the entries, which would include the new knowledge derived through novel observations and techniques. Consequently, this edition includes 256 entries written by 347 authors, and includes 44 new topics, such as Fiber Optics in Geophysics, Tracking Earth's Water in Motion, A New Era of Earth Observatories and Others Based on Satellite Observations, Deep Continental Boreholes Like the KTB, Seismic Arrays in Boreholes, Environmental Seismology, High-Frequency Seismology, Floating Seismographs, and several other new topics that have established themselves in recent years.

The entries, written by leading experts, are intended to provide a holistic treatment of solid Earth geophysics and guide researchers to more detailed sources of knowledge should they require them. A basic understanding of solid Earth geophysics is essential for professionals in many allied disciplines, such as civil engineering, environmental sciences, and the mining, exploration and software Industries, and for NGOs working on large-scale social agendas,

etc. This encyclopedia offers readers an authoritative and up-to-date reference source of extraordinary scope, drawing its unique strength from the contributions prepared by experts from around the globe.

2021 is the diamond jubilee year of the creation of the National Geophysical Research Institute, Council of Scientific & Industrial Research (CSIR-NGRI), where I spent

a major part of my academic life. It is a pleasure to commemorate the diamond jubilee of CSIR-NGRI with the publication of the second edition of ESEG.

Harsh K. Gupta

Preface to the First Edition

All information about the Earth's interior comes from field observations and measurements made within the top few kilometers of the surface, from laboratory experiments and from the powers of human deduction, relying on complex numerical modeling. Solid Earth Geophysics encompasses all these endeavors and aspires to define and quantify the internal structure and processes of the Earth in terms of the principles of physics, corresponding mathematical formulations and computational procedures. The role of Solid Earth Geophysics has gained prominence with increasing recognition of the fact that knowledge and understanding of Earth processes are central to the continued well being of the global community. Apart from persistent search for natural resources, this research line is linked to basic investigations regarding the mutual relationships between climate and tectonics and on the effects of global change in terms of a wide spectrum of natural hazards. Consequently, the pursuit of this science has seen spectacular progress all over the world in recent decades, both in fundamental and applied aspects, necessarily aided by advancements in allied fields of science and technology.

The *Encyclopedia of Solid Earth Geophysics* aims to serve as a comprehensive compendium of information on important topics of Solid Earth Geophysics and provide a systematic and up-to-date coverage of its important aspects including primary concepts as well as key topics of interest. It, however, does not claim to chronicle each and every niche area that in reality is a part of this multidisciplinary and multi-faceted science. Neither does it attempt to describe the basic physics of matter and energy systems, which comprise the underlying tenets of geophysical research. The first edition of this Encyclopedia, edited by Prof. David James, was published in 1989 by the Van Nostrand Reinhold publishing company. The

extraordinary growth and diversification of this science over the last twenty years called for a complete revision. This is realized by identifying the necessary topics and bringing together over 200 articles covering established and new concepts of Geophysics across the sub-disciplines such as Gravity, Geodesy, Geoelectricity, Geomagnetism, Seismology, Seismics, Deep Earth Interior and Processes, Plate Tectonics, Geothermics, Computational Methods, etc. in a consistent format. Exceptional Exploration Geophysics and Geotechnical Engineering topics are included for the sake of completeness. Topics pertaining to near Earth environs, other than the classical 'Solid Earth', are not within the scope of this volume as it is felt that the growth of knowledge in these fields justify a dedicated volume to cover them.

Articles written by leading experts intend to provide a holistic treatment of Solid Earth Geophysics and guide researchers to more detailed sources of knowledge should they require them. As basic understanding and application of Solid Earth Geophysics is essential for professionals of many allied disciplines such as Civil Engineering; Environmental Sciences; Mining, Exploration and software industries; NGOs working on large scale social agenda; etc., it would be useful to them to have access to a ready and up-to-date source of knowledge on key topics of Solid Earth Geophysics. Hopefully, this Encyclopedia would prove to be an authoritative and current reference source with extraordinary width of scope, drawing its unique strength from the expert contributions of editors and authors across the globe.

I am grateful to Anny Cazenave, Kusumita Arora, Bob Engdahl, Seiya Uyeda, Rainer Kind, Ajay Manglik, Kalachand Sain and Sukanta Roy, members of the Editorial Board for their constant advice and guidance in developing

the framework of this Encyclopedia and help with the editorial work. I am equally grateful to all the authors who readily agreed to contribute and honoured the guidelines and time schedule.

Petra van Steenberg, Sylvia Blago, Simone Giesler and D. Nishantini from Springer were very co-operative. It has been a pleasure working with Springer.

Ms. M. Uma Anuradha provided extraordinary assistance in the preparation of this volume. My wife Manju and daughters Nidhi & Benu supported me through the entire project.

Harsh K. Gupta
Editor-in-Chief

Acknowledgments

It has been a great pleasure working on the second edition of the *Encyclopedia of Solid Earth Geophysics*. My thanks go to Springer, particularly Petra Van Steenbergen and Sylvia Blago, for requesting it less than 10 years after the release of the first edition in 2011. I am also grateful to Anny Cazenave, Kusumita Arora, Bob Engdahl, Seiya Uyeda, Rainer Kind, Ajay Manglik, Kalachand Sain, and Sukanta Roy, the members of the editorial board, for their constant advice and guidance in developing the framework for this encyclopedia and support with the editorial work. I am equally indebted to the 347 authors who readily agreed to contribute and adhered to the guidelines and schedule.

The manuscripts went through several iterations during their development, and at each stage, they were carefully reviewed by prominent experts in their respective fields,

numbering over 200 from all over the world. Please accept my heartfelt thanks.

Petra van Steenbergen, Sylvia Blago, and Johanna Klute from Springer were very co-operative throughout this project, while Santhiya Rajarathinam helped with printing the manuscripts. It has been a pleasure working with Springer on this endeavor.

M. Uma Anuradha provided extraordinary assistance in the preparation of this volume.

I gratefully acknowledge the support provided by the CSIR-National Geophysical Research Institute, Hyderabad, and the National Academy of Sciences, India.

Last but certainly not least, my wife Manju and daughters Nidhi and Benu gave their unflagging encouragement throughout the entire project.

Harsh K. Gupta

Absolute Age Determinations: Radiometric

Richard W. Carlson

Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington, DC, USA

Definition

Radiometric dating uses the decay of naturally occurring radioactive elements to determine the absolute age of geologic events.

Introduction

Time is an essential datum in geology. Some geological processes, particularly those responsible for generating natural hazards, occur over time intervals that can be measured with conventional clocks. Most geologic processes, however, occur on timescales beyond human experience – thousands, millions, and even billions of years (see ► [“Continental Drift”](#), ► [“Continental Rifts”](#), ► [“Geodynamics”](#), ► [“Lithosphere, Continental”](#), ► [“Paleomagnetism, Magnetostratigraphy”](#), ► [“Paleoseismology”](#), and ► [“Plate Tectonics, Precambrian”](#)). For these, the chronometer of choice is radiometric dating, where the decay of naturally occurring radioactive elements is translated into time (Dickin 2005; Faure and Mensing 2005). In its application to geochronology, the most important aspect of radioactive decay is that it occurs at a known, constant rate, independent of environmental factors such as changing temperature and pressure, at least within the ranges of these parameters found outside the interior of stars.

Only 10 years after the discovery of radioactivity, the renowned physicist Ernest Rutherford measured the amount of helium in a uranium ore and derived an age of 500 million years

for the ore (Rutherford 1906). By 1913, radioactive dating of rocks had made important inroads into geologic investigations of Earth history (Holmes 1913). The discovery of isotopes, also in 1913, and the improvement of the instruments used to measure their abundance (mass spectrometers) over the next decades, allowed radioactive dating to be applied to an increasingly broad range of natural materials, and hence, geologic processes. This, in turn, facilitated the transition of isotopic measurements from physics to geology departments and established radiometric geochronology as a tool of widespread use in the geosciences in the years following World War II.

Radioactivity and the Systematics of Its Use as a Chronometer

Radiometric dating is based on the principle of nuclear transmutation, common to alchemy. Some elements have isotopes whose nuclei contain an unstable number of neutrons and protons. The instability is most often remedied by the ejection of material from the nucleus. Alpha decay involves the ejection of two protons and two neutrons (a 4-helium (^4He) nucleus or alpha particle). Beta decay occurs through loss of an electron from the nucleus, turning a neutron into a proton. Another form of radioactive decay involves the capture of an electron from those surrounding the nucleus, turning a proton into a neutron. All forms of radioactive decay also release energy, and are thus a heat source in Earth’s interior (see ► [“Radiogenic Heat Production of Rocks”](#) and ► [“Energy Budget of the Earth”](#)). The result of radioactive decay is that an isotope of one element is transformed into an isotope of another element. While the path from lead to gold does not occur naturally, if one waits long enough, radioactive decay will eventually transform all uranium and thorium into lead.

Table 1 lists the major radioactive nuclides that have been used for dating various geologic events. Naturally occurring radioactive isotopes are produced through four mechanisms:

Absolute Age Determinations: Radiometric, Table 1 Radioactive elements commonly used for absolute age determinations

Parent isotope	Production mechanism	Daughter isotope	Half-Life (Million years)
7-Beryllium (^7Be)	Cosmogenic	7-Lithium (^7Li)	53 days
210-Lead (^{210}Pb)	Uranium decay	210-Bismuth (^{210}Bi)	22.3 years
226-Radium (^{226}Ra)	Uranium decay	222-Radon (^{222}Rn), ^4He	1,622 years
14-Carbon (^{14}C)	Cosmogenic, bomb	14-Nitrogen (^{14}N)	5,730 years
231-Protactinium (^{231}Pa)	Uranium decay	227-Thorium (^{227}Th), ^4He	0.033
234-Uranium (^{234}U)	Uranium decay	230-Thorium (^{230}Th), ^4He	0.25
36-Chlorine (^{36}Cl)	Cosmogenic, bomb	36-Argon (^{36}Ar)	0.31
26-Aluminum (^{26}Al)	Cosmogenic, stellar	26-Magnesium (^{26}Mg)	0.73
230-Thorium (^{230}Th)	Uranium decay	226-Radium (^{226}Ra), ^4He	0.75
60-Iron (^{60}Fe)	Stellar	60-Nickel (^{60}Ni)	1.5
10-Beryllium (^{10}Be)	Cosmogenic, stellar	10-Boron (^{10}B)	1.6
53-Manganese (^{53}Mn)	Cosmogenic, stellar	53-Chromium (^{53}Cr)	3.7
107-Palladium (^{107}Pd)	Stellar	107-Silver (^{107}Ag)	6.5
182-Hafnium (^{182}Hf)	Stellar	182-Tungsten (^{182}W)	9
129-Iodine (^{129}I)	Stellar, cosmogenic	129-Xenon (^{129}Xe)	15.7
244-Plutonium (^{244}Pu)	Stellar	Various fission products	80
146-Samarium (^{146}Sm)	Stellar	142-Neodymium (^{142}Nd)	103
235-Uranium (^{235}U)	Stellar	207-Lead (^{207}Pb), ^4He	704
40-Potassium (^{40}K)	Stellar	40-Argon (^{40}Ar) 40-calcium (^{40}Ca)	1,270
238-Uranium (^{238}U)	Stellar	206-Lead (^{206}Pb), ^4He	4,469
232-Thorium (^{232}Th)	Stellar	208-Lead (^{208}Pb), ^4He	14,010
176-Lutetium (^{176}Lu)	Stellar	176-Hafnium (^{176}Hf)	35,700
187-Rhenium (^{187}Re)	Stellar	187-Osmium (^{187}Os)	41,600
87-Rubidium (^{87}Rb)	Stellar	87-Strontium (^{87}Sr)	48,800
147-Samarium (^{147}Sm)	Stellar	143-Neodymium (^{143}Nd), ^4He	106,000
190-Platinum (^{190}Pt)	Stellar	186-Osmium (^{186}Os)	450,000

1. Stellar nucleosynthesis where the very high temperatures and pressures present in stellar interiors fuse nuclei, creating new elements (Truran and Heger 2005).
2. Some radioactive elements decay into other radioactive elements. For example, 238-uranium requires eight alpha decays and six beta decays before it reaches stable 206-lead. Along the way, the decay steps include some isotopes with long enough decay lives to be useful in geochronology. These include ^{234}U , ^{230}Th , ^{226}Ra , ^{222}Rn , and ^{210}Pb .
3. High-energy cosmic ray particles collide with atoms in Earth's atmosphere or surface rocks with enough energy to cause nuclear reactions. For example, 14-carbon is created when 14-nitrogen in the atmosphere captures a neutron released by a cosmic ray interaction with some other atom. Other cosmic-ray-produced nuclides include ^{10}Be and ^{26}Al that are made by spallation, which occurs when an energetic cosmic ray proton simply breaks off a fragment from the nucleus of an atom with which it collides. Radioactive isotopes produced in these reactions are known as cosmogenic isotopes.
4. Above-ground detonations of nuclear bombs introduced into the environment substantial quantities of a number of radioactive species that have been used to investigate a number of atmosphere – Earth's surface exchange processes, and for tracing ocean water circulation.

The probability that a radioactive isotope will decay over any time interval is described by the equation:

$$-dN/dt = \lambda N \quad (1)$$

where N is the number of atoms of the radioactive species (parent isotope), t is time, and λ is the rate constant of the decay. Rate constants often are converted to “half-life”; the time needed for half of the radioactive species to decay. The half-life of a radioactive species is equal to $\ln(2)/\lambda$. Integrating Eq. (1) gives:

$$N = N_0 e^{-\lambda t} \quad (2)$$

where N_0 is the number of atoms present when $t = 0$. This equation can be used to determine ages if one has independent information on the initial abundance of the parent isotope. For example, in ^{14}C dating, if one assumes that the ratio of ^{14}C to stable ^{12}C is constant in the atmosphere, then one need only measure the present day $^{14}\text{C}/^{12}\text{C}$ ratio in a material that obtained its carbon from the atmosphere and use Eq. (2) in order to determine the material's age. The assumption of a constant atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio is now known to be invalid because the production rate of ^{14}C in the atmosphere depends

on issues as diverse as sunspot activity and the varying strength of Earth's magnetic field (see ► [“Paleomagnetic Field Intensity”](#)). Various schemes have been used to correct ^{14}C chronology for variations in the atmospheric ^{14}C production rate including comparison with the carbon in tree rings dated simply by counting the annual growth rings (Friedrich et al. 1999) or, for longer time intervals, to growth rings in corals dated with ^{230}Th and ^{234}U (Fairbanks et al. 2005). Although the variations in ^{14}C production rate complicate ^{14}C dating, they potentially provide information on the variability of Earth's magnetic field, or sunspot activity, and how these parameters may affect Earth's climate (Bard and Frank 2006).

Because the initial abundance of the parent isotope generally is not known, most radioactive dating schemes measure the increase in the abundance of the decay product (daughter). If no atoms are lost, the abundance of the daughter isotope (D) increases with time at the rate described by:

$$D = N_0 - N = N(e^{\lambda t} - 1) \quad (3)$$

Thus, in an ideal situation where, for example, a mineral forms that contains radioactive ^{40}K , but no ^{40}Ar , the age of the mineral can be determined simply by measuring the abundance of ^{40}K (N) and ^{40}Ar (D) in the mineral after some decay interval has passed. In nature, perfect separation of parent from daughter isotope is rare, so if some atoms of the decay product are present initially (D0), then:

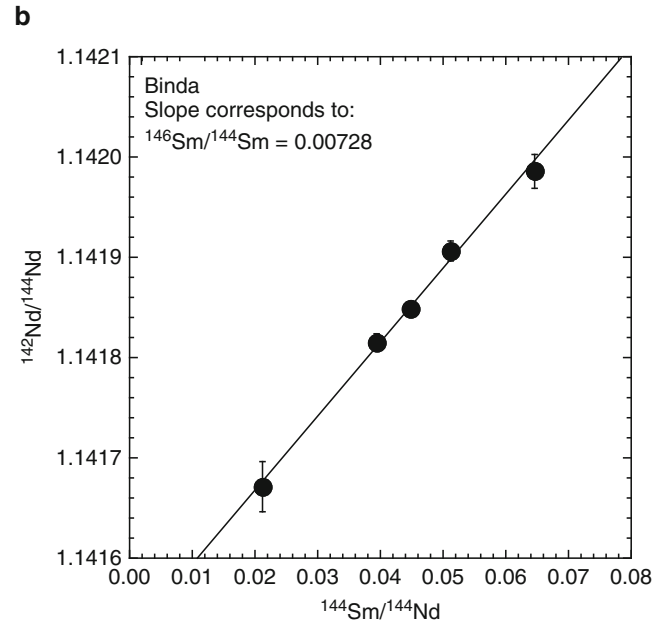
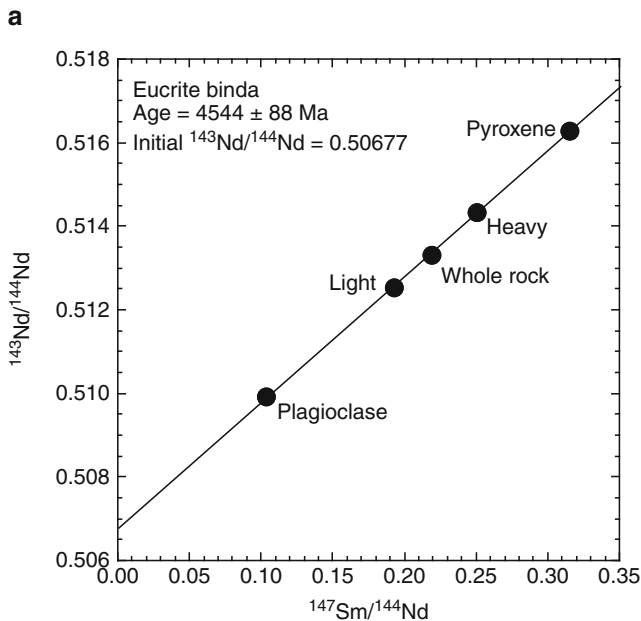
$$D = D_0 + N(e^{\lambda t} - 1) \quad (4)$$

Because mass spectrometers can measure isotopic ratios of some elements to precisions of 0.0005% whereas elemental abundances can only be determined to about 0.1–1% precision, Eq. 4, using the ^{238}U - ^{206}Pb decay scheme as an example, can be expressed as:

$$\left(^{206}\text{Pb}/^{204}\text{Pb}\right)_m = \left(^{206}\text{Pb}/^{204}\text{Pb}\right)_0 + \left(^{238}\text{U}/^{204}\text{Pb}\right)_m (e^{\lambda t} - 1) \quad (5)$$

where “m” stands for the measured ratio and λ is the decay constant of ^{238}U . Pb-204 is a stable isotope of lead. If two or more samples with different U/Pb ratios are formed at the same time with the same initial Pb isotope composition ($(^{206}\text{Pb}/^{204}\text{Pb})_0$), after some time has passed, plotting the measured $^{206}\text{Pb}/^{204}\text{Pb}$ for each sample against its $^{238}\text{U}/^{204}\text{Pb}$ ratio will produce a line, called an isochron, whose slope and y-intercept correspond to $(e^{\lambda t} - 1)$, and the initial $^{206}\text{Pb}/^{204}\text{Pb}$ ratio, respectively (Fig. 1a). Thus, the slope of an isochron provides the time when the samples on the line had the same Pb isotope composition. The U–Pb system is unique in that it contains two isotopes of U that decay into two isotopes of Pb.

Thus, an analogous Eq. (5) can be written for the decay of ^{235}U to ^{207}Pb . More importantly, the two equations can be combined to give:



Absolute Age Determinations: Radiometric, Fig. 1 Isochron diagrams for the long-lived ^{147}Sm - ^{143}Nd and short-lived, now extinct, ^{146}Sm - ^{142}Nd systems. The data shown are for minerals separated from the Binda eucrite, a type of meteorite that is a volcanic rock erupted on some small planetesimal. The meteorite consists primarily of the minerals plagioclase and pyroxene, which have different ratios of Sm to Nd

when they form, but the same initial Nd isotopic composition. As Sm decays, it changes the isotopic composition of Nd. Using an equation similar to 5, the slope on the ^{147}Sm - ^{143}Nd diagram directly provides the age of the sample, but on the ^{146}Sm - ^{142}Nd diagram the slope provides the initial abundance of ^{146}Sm (actually the $^{146}\text{Sm}/^{144}\text{Sm}$ ratio as illustrated in an equation analogous to Eq. 8). Data from Boyet et al. (2010)

$$\frac{\left(\frac{{}^{207}\text{Pb}}{{}^{204}\text{Pb}}\right)_m - \left(\frac{{}^{207}\text{Pb}}{{}^{204}\text{Pb}}\right)_0}{\left(\frac{{}^{206}\text{Pb}}{{}^{204}\text{Pb}}\right)_m - \left(\frac{{}^{206}\text{Pb}}{{}^{204}\text{Pb}}\right)_0} = \frac{\left(\frac{{}^{235}\text{U}}{{}^{238}\text{U}}\right)(e^{\lambda_5 t} - 1)}{(e^{\lambda_8 t} - 1)} \quad (6)$$

where λ_5 and λ_8 are the decay constants for ${}^{235}\text{U}$ and ${}^{238}\text{U}$, respectively. Because the ${}^{235}\text{U}/{}^{238}\text{U}$ ratio is nearly constant in nature at any given time, this equation allows an age to be determined from measurements of lead isotopic composition alone, without the need for concentration determinations of U or Pb. Such ages, called Pb–Pb ages, are obtained by plotting $\left(\frac{{}^{207}\text{Pb}}{{}^{204}\text{Pb}}\right)_m$ versus $\left(\frac{{}^{206}\text{Pb}}{{}^{204}\text{Pb}}\right)_m$, which will produce a line for two or more samples whose slope is equal to the right hand side of Eq. (6). This equation can then be solved numerically to determine “ t ”, the time when the samples formed.

A very active application of radiometric chronology uses the variety of short-lived (half-lives of 10^5 – 10^8 year) radionuclides that were present when the solar system formed, but have long since decayed away (Carlson and Boyet 2009). For these systems, the “ t ” in Eq. (2) cannot be referenced to today because $e^{-\lambda t}$ would always be close to zero. Instead, the “ t ” for these extinct systems must be a time interval referenced to a time when the parent isotope still existed. In this case, using the decay of ${}^{26}\text{Al}$ to ${}^{26}\text{Mg}$ as an example, Eq. (5) can be rewritten as:

$$\left(\frac{{}^{26}\text{Mg}}{{}^{24}\text{Mg}}\right)_{\Delta t} = \left(\frac{{}^{26}\text{Mg}}{{}^{24}\text{Mg}}\right)_0 + \left(\frac{{}^{26}\text{Al}}{{}^{27}\text{Al}}\right)_0 \left(\frac{{}^{27}\text{Al}}{{}^{24}\text{Mg}}\right)_m e^{-\lambda \Delta t} \quad (7)$$

Plotting the two measurable parameters of this equation (${}^{27}\text{Al}/{}^{24}\text{Mg}$, ${}^{26}\text{Mg}/{}^{24}\text{Mg}$) for two or more samples will produce a line whose slope gives not the age, but the $\left(\frac{{}^{26}\text{Al}}{{}^{27}\text{Al}}\right)_0$ at the time when the samples formed (Fig. 1b). The time interval between the reference time and the time of sample formation is then:

$$\Delta t = -\left(\ln \left[\left(\frac{{}^{26}\text{Al}}{{}^{27}\text{Al}}\right)_{\Delta t} / \left(\frac{{}^{26}\text{Al}}{{}^{27}\text{Al}}\right)_0 \right]\right) / \lambda \quad (8)$$

To convert these relative ages into absolute ages requires cross-calibration with a chronometer that provides absolute ages. For example, a certain type of mineral inclusion within the Allende meteorite provided a Pb–Pb age of 4567.6 ± 0.4 Ma and an Al–Mg isochron whose slope corresponds to ${}^{26}\text{Al}/{}^{27}\text{Al} = 4.96 (\pm 0.25) \times 10^{-5}$ (Jacobsen et al. 2008). This provides a “pinning point” for the ${}^{26}\text{Al}/{}^{27}\text{Al}$ ratio present in the solar system at a known time. Thus, if analyses of some other sample produced an Al–Mg isochron whose slope corresponded to, for example, ${}^{26}\text{Al}/{}^{27}\text{Al} = 2.5 \times 10^{-5}$, this sample would be one half-life of ${}^{26}\text{Al}$, or 750,000 years,

younger than the other. Extinct radioactive dating systems thus provide not absolute, but relative, ages. If these relative ages are anchored to an absolute age, as done in the example above, then the extinct system can be used to indicate that the hypothetical sample mentioned above is 4566.8 million years old. A good example of the use of extinct radionuclides is the attempt to date the oldest volcanism in the solar system. Several types of meteorites are interpreted to be lavas produced on small planetesimals. One of these, the D’Orbigny meteorite, has provided Al–Mg, Mn–Cr, Hf–W, and Pb–Pb ages of ± 0.5 Ma, 4564.0 ± 0.6 Ma, 4562.8 ± 1.5 Ma, and 4564.42 ± 0.12 Ma, respectively. These results offer the promise of temporal precisions of 0.002% for the earliest events in solar system history.

The Application of Radiometric Dating

The relatively simple equations given in the previous section provide the basis by which measurements of radioactive decay products can be translated into absolute ages. Applying these chronometers to natural processes requires matching the distinct properties of the many radioactive systems listed in Table 1 to the problem being investigated. Systems with short half-lives can be used to determine precise ages of geologic events only a few to a few thousand years old. Many of these short-lived nuclides are created either in the atmosphere or upper meter of Earth’s surface, which makes them particularly amenable to dating a variety of near-surface processes. The concentrations of ${}^{10}\text{Be}$ or ${}^{26}\text{Al}$ in quartz in rocks or sediments can be used to date the time when the rock was exposed to cosmic rays, which can provide information on both uplift and erosion rates (von Blanckenburg 2005) (see ► “Paleoseismology”). The rates of water movement from surface through subsurface aquifers can be traced with a number of radiometric chronometers (Bethke and Johnson 2008). The presence of ${}^{10}\text{Be}$ in the lavas erupted at convergent margins tracks the subduction of surficial sediments to the 100 km + depths of magma generation (Morris 1991) (see ► “Subduction Zones”). Carbon-14 is used to date a huge variety of events from the eruption of young volcanoes, movement along faults (see ► “Earthquakes and Crustal Deformation”, ► “Earthquake, Archaeoseismology” and ► “Paleoseismology”), the rates of ocean water circulation and ocean uptake of atmospheric CO_2 , and the death of a living organism with its applications in archeology and paleoecology (Broecker 2005). The longer-lived radionuclides can be applied to a wide variety of issues concerning the evolution of the solid earth, from dating the formation of crustal rocks (see also ► “Paleomagnetism, Magnetostratigraphy”), to determining the time of continental assembly and breakup (see ► “Continental Drift” and ► “Plates and Paleoreconstructions”), to defining the age of the Earth.

The key point in the application and proper interpretation of radiometric dating is understanding exactly what event is being dated. What “starts” a radiometric clock is some event that changes the composition of a sample and fixes this new compositional state until the sample is dated in the laboratory. For example, ^{14}C often is used to determine the age of young volcanic eruptions. This system works in this application not by dating the rock itself, but by dating charcoal formed when the lava flowed over, and burnt, a bush or tree. What the ^{14}C age actually dates is the time when the plant stopped exchanging carbon with Earth’s atmosphere. In most cases, this happened because the plant was killed by the volcanic eruption, in which case, the ^{14}C age accurately dates the eruption. However, if a 100-year-old lava flow were to flow over a 10,000-year-old bristlecone pine tree trunk lying long dead on the ground, the ^{14}C date of the charcoal would be 10,000 years, not 100 years.

The precision of a radiometric age depends on the amount radioactive decay has changed the isotopic composition of the parent or daughter element. This depends primarily on the time passed relative to the half-life of the system, and in systems where measurement of the daughter element is involved, on the degree to which parent and daughter element are fractionated from one another. Thus, one would use ^{14}C with its 5,730 year half-life to date an organic remain from ancient Rome, but it would be useless to date a million year old limestone because the ^{14}C “clock” stopped measuring time when the ^{14}C decayed away after a few half-lives into the history of the limestone. Similarly, ^{147}Sm , with a 106 billion year half-life, is used to date a variety of materials ranging in age from millions to billions of years old, but would be useless to date very young processes because only 7×10^{-15} g of ^{143}Nd is produced every thousand years per ppm of Sm, and this amount simply is too small to measure accurately.

Besides choosing the proper radiometric system for the application, understanding whether a radiometric age is providing an accurate age for the event of interest demands an understanding of the processes that set, and reset, a radiometric clock. For Eq. (5) to provide an accurate age demands that the samples used to construct the isochron: (1) all formed at exactly the same time, (2) all had exactly the same initial daughter isotope composition, (3) experienced no change in parent/daughter element ratio between formation of the sample and its measurement in the laboratory, and (4) nothing other than the decay of the parent modified the daughter isotopic composition. In nature, all of these requirements have the potential to fail, but, in some cases, these “failures” can still provide useful information.

A good candidate to meet the requirements above is a volcanic rock. Under conditions ideal for an accurate age, a well-mixed magma will crystallize, over a short time interval after eruption, a variety of minerals with different parent/

daughter elemental ratios, but the same initial daughter isotopic composition that they inherit from the magma. Over time, if the minerals experience no chemical exchange with their surroundings, the radioactive decay products will build up allowing their measurement to provide the accurate age of eruption of this volcanic rock. The Apollo 15 basalt 15,386 provides evidence that when these criteria are met, ages from many radioactive systems can agree. For this sample, the age (in billions of years) obtained from different radiometric systems are: K-Ar 3.89 ± 0.04 , Rb-Sr 3.88 ± 0.01 , Sm-Nd 3.85 ± 0.08 , and the U-Pb age for zircon from a compositionally similar rock is 3.89.

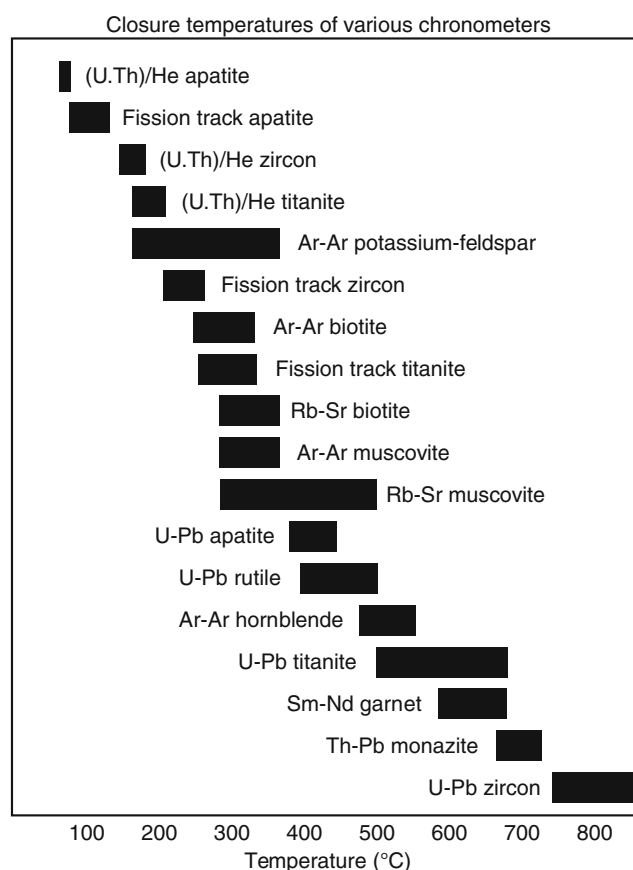
For volcanic rocks this old, the assumption that the minerals all formed at the same time, at least relative to the ancient age of the rock, likely will be accurate. When one moves to younger volcanic systems, and the use of chronometers with better temporal resolution, the issue of the duration of crystallization becomes important. Several cases have now been presented where different minerals in the same volcanic rock give ages that range from the age of eruption to significantly older. One example is the Coso Volcanic field in California where zircons in one of the rhyolites range in age from the same as a K-Ar date on sanidine from the rhyolite, interpreted as the eruption age, to as much as 200,000 years older (Miller and Wooden 2004). This result suggests that emplacement and crystallization of distinct magma batches in the crustal magma chamber that eventually fed this eruption is a prolonged process occurring over many thousands of years. Similar results have now been reported in many young volcanic rocks (Cooper and Reid 2008). Although the age range of the minerals mentioned above is interpreted as reflecting a prolonged magmatic event, another way that an old crystal can get into a young magma is if a crystal from an older rock underlying the volcano is picked up by the magma during its ascent. Finding such “xenocrystic” material, particularly within more viscous, lower melting temperature, silica-rich magmas, is so common that it can be used to detect and even determine the age of older crust buried beneath a volcano, in some cases enabling the identification of older basement where none was previously known to exist (Smyth et al. 2007).

This brings up what is possibly the most insidious cause of an inaccurate radiometric age: the generally poorly-known temperature history of the sample. Because the rate of chemical diffusion within solids increases with temperature, elevated temperatures facilitate chemical changes. Heating a mineral can cause gases, such as He and Ar that are present in a crystal lattice as a result of their production through radioactive decay, to diffuse out of the crystal and escape along grain boundaries. At a temperature of 1,000 °C, diffusion will cause a Sr atom in a mineral to move about 2 mm in a million years. If a small grain of a high Rb/Sr ratio mineral, such as mica, is next to a grain of a low Rb/Sr ratio, and high

Sr content, mineral like plagioclase, the ingrowth of radiogenic ^{87}Sr in the mica will be diluted by diffusional exchange with the unradiogenic Sr in the plagioclase. Diffusion thus makes all radioactive chronometers “thermochronometers” that date the time when the material of interest reached a low-enough temperature to stop diffusional resetting of the radiogenic ingrowth. Because rates of diffusion are a function both of the element that is diffusing and the structure of the crystal through which it is diffusing, different radiometric systems, and different minerals, can have different sensitivities to temperature excursions. This gives rise to the concept of “closure” or “blocking” temperature, above which the diffusion rate is too high to allow for the buildup of the radioactive decay product. In other words, the radiometric clock does not start until the mineral falls below its closure temperature. In the Coso example above, the K–Ar age obtained for the sanidine is interpreted as the eruption age because even if this mineral formed well prior to eruption, the 200–400 °C closure temperature of the K–Ar system in sanidine (Fig. 2) is so low that the radiogenic ^{40}Ar being produced in the sanidine was diffusing out as fast as it was being produced because of magma temperatures in the range of ~750 °C.

In contrast, the closure temperature of the U–Pb system in zircon is higher than the magma temperature, so as soon as the zircon first crystallized, its U–Pb clock started to record time. Temperature-induced diffusional resetting of radiometric systems may make interpretations of ages a bit more difficult, but it creates a new field – thermochronometry – that constructs time-temperature histories for various geologic processes (Reiners and Ehlers 2005). Though many radiometric systems can be used for thermochronometry, the oldest radiometric dating technique, the decay of U and Th to He, has been reenergized (Farley 2002) because the low closure temperature of the U,Th – He system can be used to address geological processes such as the rate of mountain uplift or surface erosion by telling the time when a rock arrived close enough to Earth’s surface to cool below the closure temperature of the radiometric clock (Reiners and Brandon 2006).

While the appreciation and use of thermochronometry has been the primary growth field in radiometric geochronology in the last decade, other efforts are directed at finding ways to avoid reset ages. The traditional way to do this is to move to sample sizes much larger than diffusion lengths. This is the basis for “whole rock isochrons” where kilogram rock samples are homogenized and measured as an individual point on an isochron. Although this approach has produced some accurate ages, it suffers two main weaknesses. First, deciding whether all the rocks included on an isochron are the same age is not always easy, particularly for ancient, highly deformed, gneisses where this approach often is applied. Second, in order to obtain a sufficient range in parent/daughter ratio to precisely determine an isochron, compositionally distinct rocks are usually



Absolute Age Determinations: Radiometric, Fig. 2 Closure temperatures for various radiometric systems in different minerals. The fission track technique counts the density of damage tracks made in crystals due to the fission of ^{238}U . Coupled with the measurement of U concentration in the crystal, the density of fission tracks can be converted into a radiometric age. Figure modified from (Pollard, 2002 – <http://pangea.stanford.edu/~dpollard/NSF/main.html>) which was adapted from a figure produced by P. Fitzgerald, S. Baldwin, G. Gehrels, P. Reiners, and M. Ducea

included on the isochron. Judging from modern igneous rocks, large compositional distinctions often are associated with variable initial isotope composition, which would violate the requirement that all the samples used to construct an isochron have the same initial isotope composition. Where whole rock isochrons can be useful is in detecting when much older components are involved in the genesis of younger rocks. A good example of this are the Sm–Nd data for lunar basalts where mineral isochrons of individual samples provide eruption ages generally less than 4 billion years, but the whole rock data scatter about a line corresponding to an age near 4.45 Ga (Carlson and Boyet 2009). The precision of this older age is debatable, as is its interpretation, but it could imply that the younger basalts are made by melting source materials in the lunar interior that formed from some global differentiation event on the Moon that occurred some time near 4.45 Ga. Memory of this event was erased from the minerals in the

basalts, that know only of their crystallization from the magma, but the magmas themselves may “remember” the events that led to the formation of their source materials in the lunar interior.

The other approach to avoiding thermally reset ages is to use minerals whose crystal structure allows only very slow diffusion. One curious example is diamond. Diamond cannot be dated directly because it contains insufficient amounts of the radioactive elements, but it often contains as inclusions minerals that can be dated. Even though these minerals were formed, and have been stored, at temperatures above 1,000 °C and depths greater than 150 km in the mantle (see ► “[Lithosphere, Continental: Thermal Structure](#)”), the encapsulation by the diamond does not allow them to communicate through diffusion with their surroundings. As a result, dating of inclusions in diamond show that diamond formation in the mantle has occurred over a good part of Earth history, with the oldest diamonds being approximately 3.5 Ga (Pearson and Shirey 1999) (see ► “[Lithosphere, Continental](#)”).

By far, the most commonly used and most versatile way to avoid reset ages is with the application of the U–Pb system in zircon (Hanchar and Hoskins 2003). The strengths of U–Pb zircon dating include: (a) very high U–Pb closure temperature, (b) resistance to chemical modification through alteration and metamorphism, (c) high U concentrations, but initially vanishingly small Pb contents, (d) the U–Pb system includes two independent decay schemes that allow tests to determine whether or not the age has been disturbed (Wetherill 1956), (e) among the long-lived radioisotopes, the relatively short half-life of ^{235}U allows highly precise ages to be obtained for ancient rocks, (f) both the trace element concentrations and Hf isotope composition in zircon provide information on the nature of the source rock from which the zircon was derived, and (g) U–Pb ages can be obtained in samples as small as fragments of individual grains or even spots, tens of microns in diameter, ablated with either lasers or ion beams. Perhaps the only weakness of zircon is that it is not found in all rock types, occurring primarily in felsic igneous rocks. These characteristics explain the many varied applications of zircon U–Pb dating that include the discovery of the oldest dated material on Earth, 4.36 Ga zircons from quartzites in western Australia (Wilde et al. 2001), to high-precision calibration of the geologic time-scale (Bowring and Schmitz 2003), to the previously described use of zircons to determine the duration of crystallization in the magma chambers of young volcanic rocks.

Summary and Conclusions

Continually improving analytical capabilities coupled with expanding appreciation of the physical controls on what sets, and resets, radiometric clocks are allowing ever-increasing expansion of this now century-old approach. The

variety of radioactive species now in use provide the ability to determine absolute ages for geologic processes that occur over timescales that range from days to the age of the Earth and solar system.

Cross-References

- [Continental Drift](#)
- [Continental Rifts](#)
- [Earthquake, Archaeoseismology](#)
- [Energy Budget of the Earth](#)
- [Geodynamics](#)
- [Lithosphere, Continental](#)
- [Lithosphere, Continental: Thermal Structure](#)
- [Paleomagnetic Field Intensity](#)
- [Paleomagnetism, Magnetostratigraphy](#)
- [Paleoseismology](#)
- [Plate Tectonics, Precambrian](#)
- [Radiogenic Heat Production of Rocks](#)
- [Subduction Zones](#)

Bibliography

- Bard E, Frank M (2006) Climate change or solar variability: what's new under the sun? *Earth Planet Sci Lett* 248:1–14
- Bethke CM, Johnson TM (2008) Groundwater age and groundwater age dating. *Annu Rev Earth Planet Sci* 36:121–152
- Bowring SA, Schmitz MD (2003) High-precision U–Pb zircon geochronology and the stratigraphic record. *Rev Mineral Geochem* 53:305–326
- Boyett M, Carlson RW, Horan M (2010) Old Sm–Nd ages for cumulate eucrites and redetermination of the solar system initial $^{146}\text{Sm}/^{144}\text{Nd}$ ratio. *Earth Planet Sci Lett* 291:172–181
- Broecker WS (2005) Radiocarbon. In: Keeling RK (ed) *Treatise on geochemistry*. Elsevier, Amsterdam, pp 1–18
- Carlson RW, Boyett M (2009) Short-lived radionuclides as monitors of early crust–mantle differentiation on the terrestrial planets. *Earth Planet Sci Lett* 279:147–156
- Cooper KM, Reid MR (2008) Uranium-series crystal ages. *Rev Mineral Geochem* 69:479–544
- Dickin AP (2005) *Radiogenic isotope geology*, 2nd edn. Cambridge University Press, Cambridge
- Fairbanks RG, Mortlock RA, Chiu T-Z, Cao L, Kaplan A, Guilderson TP, Fairbanks TW, Bloom AL, Grootes PM, Nadeau M-J (2005) Radiocarbon calibration curve spanning 0–50,000 years BP based on paired $^{230}\text{Th}/^{234}\text{U}/^{238}\text{U}$ and ^{14}C dates on pristine corals. *Quat Sci Rev* 24:1781–1796
- Farley KA (2002) (U–Th)/He dating: techniques, calibrations, and applications. In: Porcelli D, Ballentine CJ, Wieler R (eds) *Reviews in mineralogy and geochemistry: noble gases in geochemistry and cosmochemistry*. Mineralogical Society of America, Washington, DC, pp 819–844
- Faure G, Mensing TM (2005) *Isotopes: principles and applications*, 3rd edn. Wiley, Hoboken
- Friedrich M, Kromer B, Spurk M, Hofmann J, Kaiser KF (1999) Paleoenvironment and radiocarbon calibration as derived from late glacial/early Holocene tree-ring chronologies. *Quat Int* 61:27–39
- Hanchar JM, Hoskins PWO (2003) Zircon. *Rev Mineral Geochem* 53:500

- Holmes A (1913) The age of the Earth. Harper and Brothers, London
- Jacobsen B, Yin Q, Moynier F, Amelin Y, Krot AN, Nagashima K, Hutcheon ID, Palme H (2008) ^{26}Al - ^{26}Mg and ^{207}Pb - ^{206}Pb systematics of Allende CAIs: canonical solar initial $^{26}\text{Al}/^{27}\text{Al}$ reinstated. *Earth Planet Sci Lett* 272:353–364
- Miller JS, Wooden JL (2004) Residence, resorption and recycling of zircons in Devils Kitchen rhyolite, Coso Volcanic field, California. *J Petrol* 45:2155–2170
- Morris JD (1991) Applications of cosmogenic ^{10}Be to problems in the Earth sciences. *Annu Rev Earth Planet Sci* 19:313–350
- Pearson DG, Shirey SB (1999) Isotopic dating of diamonds. In: Lambert DD, Ruiz J (eds) Application of radiogenic isotopes to ore deposit research and exploration. Boulder, Society of Economic Geologists, pp 143–172
- Reiners PW, Brandon MT (2006) Using thermochronology to understand orogenic erosion. *Annu Rev Earth Planet Sci* 34:419–466
- Reiners PW, Ehlers TA (2005) Low-temperature thermochronology: techniques, interpretation, and applications. *Rev Mineral Geochem* 58:622
- Rutherford E (1906) Radioactive transformations. Scribner's, New York
- Smyth HR, Hamilton PJ, Hall R, Kinny PD (2007) The deep crust beneath island arcs: inherited zircons reveal a Gondwana continental fragment beneath East Java, Indonesia. *Earth Planet Sci Lett* 258:269–282
- Truran JWJ, Heger A (2005) Origin of the elements. In: Davis AM (ed) Treatise on geochemistry. Elsevier, Amsterdam, pp 1–16
- von Blanckenburg F (2005) The control mechanisms of erosion and weathering at basin scale from cosmogenic nuclides in river sediments. *Earth Planet Sci Lett* 237:462–479
- Wetherill GW (1956) Discordant U-Pb ages. 1. *Trans Am Geophys Union* 37:320–326
- Wilde SA, Valley JW, Peck WH, Graham CM (2001) Evidence from detrital zircons for the existence of continental crust and oceans on Earth 4.4 Gyr ago. *Nature* 409:175–178

Anthropogenic Seismicity Related to Exploitation of Georesources

Stanislaw Lasocki and Beata Orlecka-Sikora
Department of Seismology, Institute of Geophysics Polish Academy of Sciences, Warsaw, Poland

Definition

Anthropogenic seismicity: Earthquakes induced or triggered by technological, usually georesources exploitation-related, activity.

Overview

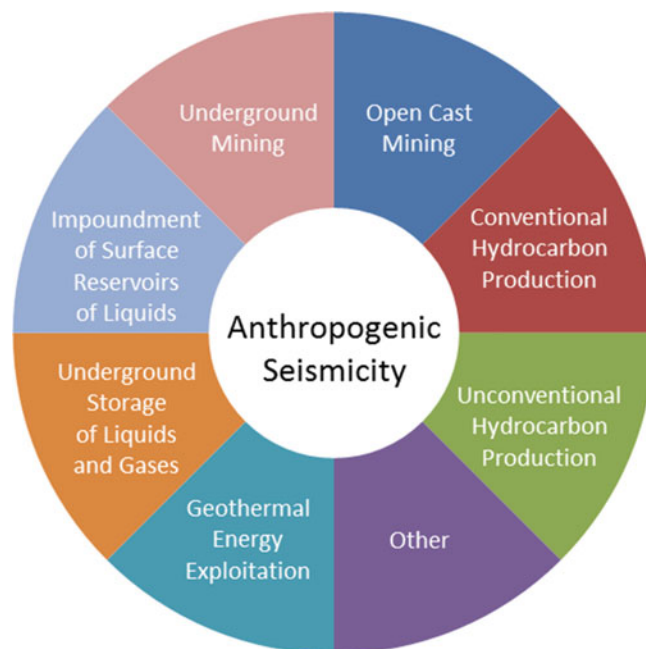
Exploitation of georesources alters the stress and strength conditions in the engaged rock mass. When such exploitation takes place in brittle rocks and yields stresses exceeding the rock strength at any point, the rock can fracture, violently emitting seismic waves. This anthropogenic seismicity was first observed as related to underground mining. The first

description of a damaging seismic event, called a rockburst, refers to the lead mine in Derbyshire in England in 1738. With rising demand for energy and minerals, anthropogenic seismicity has been appearing in association with diverse technological processes, and also in areas previously known as aseismic. Seismic events can be induced by underground and open-pit mining, both conventional and unconventional hydrocarbon exploitation, and geothermal energy production, as well as many other technological processes that perturb the boundary conditions in the affected rock mass (Fig. 1).

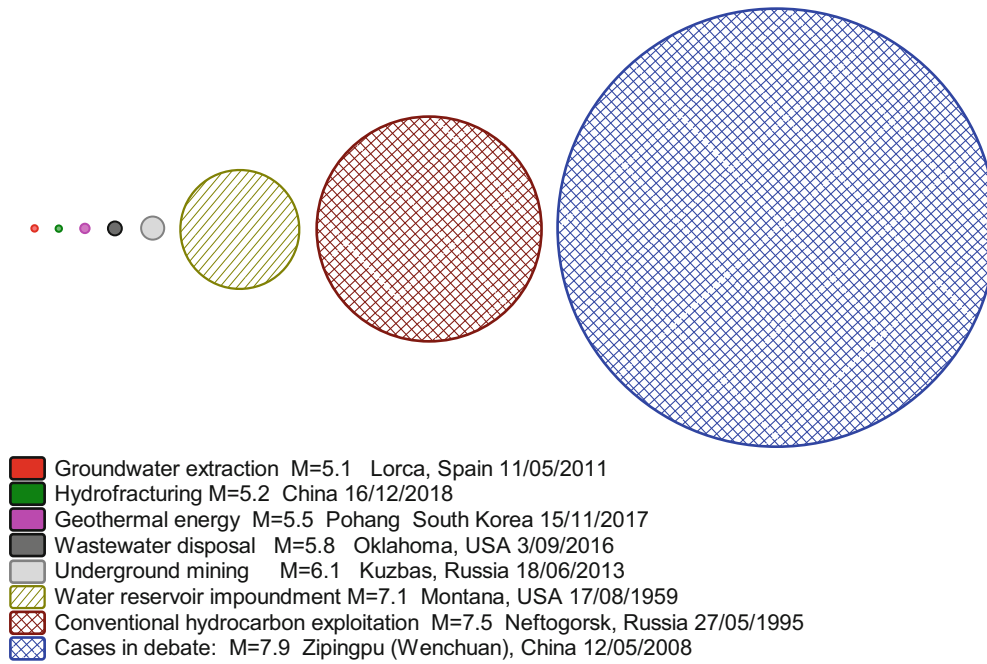
Although most of the induced seismic events are weak, hazards posed by anthropogenic seismicity can be significant. The severity of anthropogenic seismicity depends on the inducing technological activity. Figure 2 compares the largest seismic events induced by different georesources exploitation activities, respectively (source: The Human-Induced Earthquake Database (*HiQuake*), www.inducedearthquakes.org). The circled areas represent mutual proportions in the seismic energy of these events. All events presented here had significant damaging impacts.

Mechanisms of Interaction Between Technological Activity and the Rock Mass

A simplified mechanism of earthquake generation can be quantified by the Coulomb failure criterion (Fig. 3). Critical conditions leading to the seismic slip on a fault are met when the shear stress in the direction of slip, τ , equalizes the frictional strength of the fault, $\mu(\sigma_n - p) + c$, where σ_n is the compressive



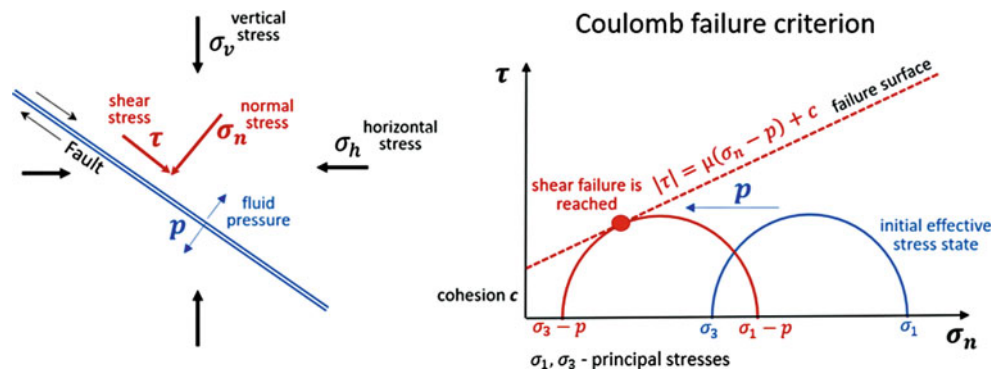
Anthropogenic Seismicity Related to Exploitation of Georesources, Fig. 1 Georesources exploitation activities that induce seismicity



Anthropogenic Seismicity Related to Exploitation of Georesources, Fig. 2 The strongest seismic events induced by georesources exploitation. The circled areas represent mutual

proportions in the seismic energy of these events. (Source: www.inducedearthquakes.org)

Anthropogenic Seismicity Related to Exploitation of Georesources, Fig. 3 Scheme of conditions leading to seismic slip on a fault



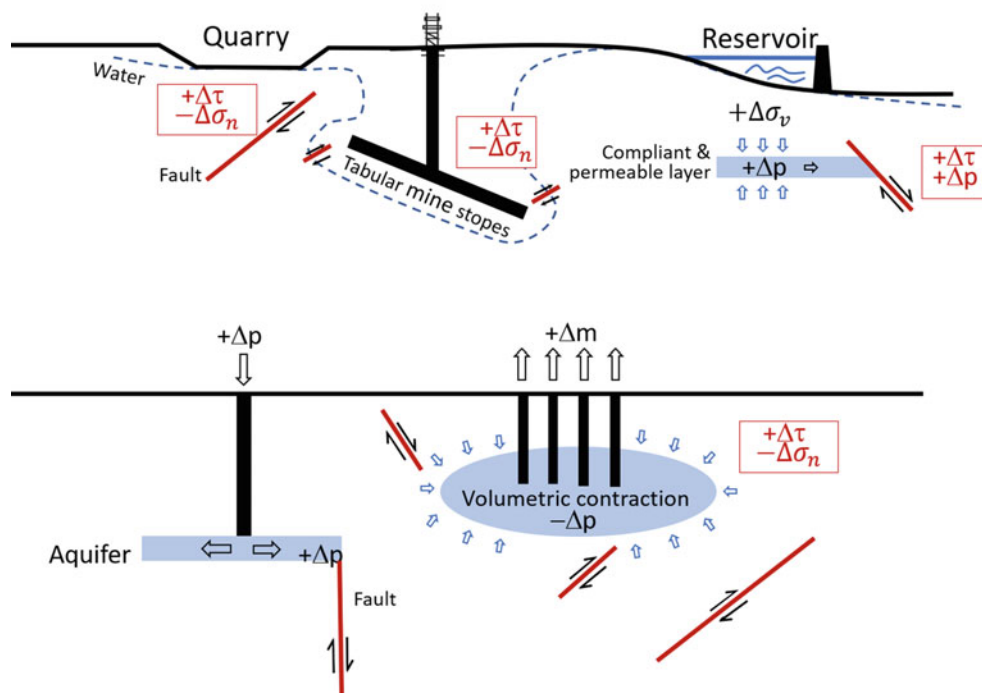
stress normal to the fault plane, p is the pore pressure, c is the cohesion, and μ is the coefficient of friction. The term $(\sigma_n - p)$ is called the effective normal stress. The Coulomb criterion indicates that a fault can be brought to failure by reducing the normal stress, increasing the pore pressure, increasing the shear stress, decreasing frictional strength of the fault, or combinations of these factors.

Different georesources exploitation activities interact in different ways with a rock mass, inducing seismicity. The schematic mechanisms of these interactions are shown in Fig. 4 (McGarr et al. 2002). The mass transfer leading to empty spaces formation results in the possible simultaneous reduction of the normal stress and increase of shear stress during mining. The

volume contraction caused by the reservoir depletion in conventional extraction of hydrocarbons decreases the pore pressure. However, for certain orientations of nearby faults the contraction can also lead to a decrease of normal stress and an increase of shear stress on the fault, as is demonstrated in the lower part of Fig. 4. The pore-pressure increase is a primary seismicity-inducing factor for all technology using pressurized fluid injection, including secondary hydrocarbon recovery from conventional deposits, hydraulic fracturing for unconventional hydrocarbon production, underground disposal of liquid or gaseous wastes including carbon capture and storage (CCS) projects, enhanced geothermal systems (EGS), and underground storage of fluids and gases.

Anthropogenic Seismicity Related to Exploitation of Georesources,

Fig. 4 Schematic mechanisms of inducing seismicity by georesources exploitation. (After McGarr et al. 2002)



Induced, Triggered, or Natural

Anthropogenic seismicity, for which technological activity provides most of the stress/rock strength perturbations needed for its generation, is called induced seismicity. For seismicity known as triggered, anthropogenic stress/rock strength disturbances meet preexisting, preloaded faults and push these faults to failures. In such cases human activity accounts for only a small fraction of the energy associated with the earthquakes, whereas tectonic loading has the primary role.

The stress perturbations introduced by georesources exploitation are rarely capable of producing major earthquakes. Most of the severely damaging or devastating seismic events linked to technological activities are triggered. Because tectonic loading of preexisting faults increases with time, the fault pushed to rupture by a technological activity would have ruptured over time solely from increased tectonic loading, which would have occurred when the increasing shear stress on the fault plane had equalized the frictional stress. Moreover, if earthquake triggering is the result of an increase of pore pressure that weakens the fault, the triggered earthquake is weaker than the earthquake which would have occurred without the triggering technological activity. However, this realization is by no means comforting for people affected by the triggered earthquake.

When an earthquake has occurred within the range of technological operations related to the exploitation of georesources, the essential question is posed: did the technological activity

triggered this earthquake, or was it a natural event caused solely by tectonic forcing? An accurate answer to this question would determine the extent of responsibility of the company whose activities may have led to the earthquake. However, because the reaction of a rock mass to georesources exploitation is complex and diverse, in spite of numerous attempts (e.g., Davis and Frohlich 1993; Dahm et al. 2013; Zhang et al. 2016) there are no general recommendations that, when followed, would result in unequivocally answering this question. Moreover, even for individual earthquake cases studied in detail, debates on whether these are natural or human triggered may continue for years. Such are the sequence of earthquakes in Gazli, Uzbekistan (1976: M7, M7; 1978: M5.7; 1984: M7) possibly linked to conventional gas production; the case of the Coalinga, USA M6.5 earthquake in 1983, which might have been connected with oil production; the M5.1 Lorca, Spain earthquake in 2011, suspected to have been triggered by long periods of water pumping; and many others. Presently, it is often only possible to state whether human influence in the occurrence of a specific earthquake can or cannot be ruled out. This question was answered in the case of the sequence of devastating earthquakes that occurred in the Emilia-Romagna region, Italy, in May 2012. The sequence, with two main shocks of magnitudes approaching 6.0, killed 27 people, left more than 20,000 people homeless, and damaged or destroyed hundreds of structures of historical significance. The total loss has been estimated as some 5.4 billion euros. The general region as the Apennine foreland is seismically active, and this region has also had a long

history of georesources exploitation. In proximity to the earthquake sequence location, there were three active licenses for hydrocarbon production, one active gas storage concession, and one geothermal license. In the aftermath of the devastating earthquakes of May 2012, the Department of Civil Protection of Italy appointed an International Commission, ICHESE, to evaluate the possibility of an impact of these technological activities on the occurrence of the earthquakes (Styles et al. 2014). Analysis of the operational data from the hydrocarbon production license, the operation closest to the location of the earthquake sequence, showed that all parameters fluctuated, that is, monthly production of oil, monthly production of gas, average monthly reinjected volume of water, and average injection wellhead pressure. In particular, a rapid simultaneous change in trend from decreasing to increasing of monthly extracted and injected volumes of fluid and of the wellhead pressures occurred 1 year before the seismic crisis, in April–May 2011 (Fig. 5). This increasing trend lasted until the crisis in May 2012. Statistical tests indicated that the seismicity rate became greater in the period after April–May 2011, when all production parameters were increasing, than in the period before April–May 2011, when those parameters were decreasing. This result pointed to a link between technological operations in the area and the seismicity before the seismic crisis. At the same time the time–space clustering analysis indicated a strong connection between the seismic events that preceded the earthquake of May 20 and the subsequent major shock sequence. Furthermore, it was ascertained that the directions of the P and T axes within the group of events that occurred in the periods of increasing trends of operational parameters differed from the directions of these axes within the group of events that occurred in the periods of decreasing trends. The statistical analyses pointed also to the correlation between the P and T axes directions of events from the ‘increasing trends’

group with these axes directions of major shocks from May–June 2012. These observations indicated that the seismic activity within the time–space cluster of events which concluded with the May 20th main event was correlated with an increase of extraction and injection activity at the studied hydrocarbon production license.

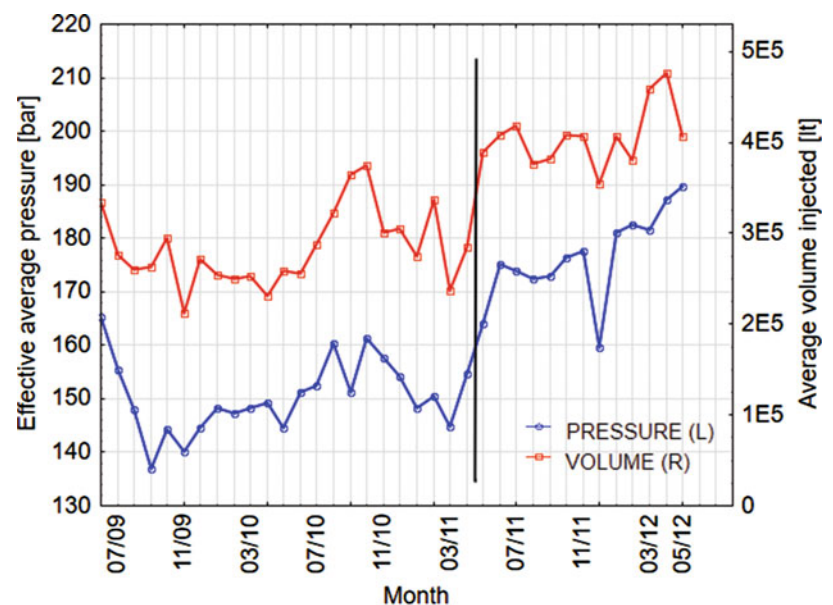
The ICHESE commission also carried out geomechanical modeling of potential triggers for earthquake initiation. It has been recognized that the stress changes generated by the activities of reservoir depletion were not sufficient in themselves to initiate movement on the fault resulting in the earthquake of May 2012. Similar results from modeling were obtained later by various groups investigating the Emilia–Romagna seismic crisis. These results might be seen as arguing in favor of a tectonic origin of the whole earthquake sequence. However, the significant statistical correlations between the changes in seismic activity characteristics before the May 20, 2012 event and the increase in production parameters since April–May 2011 could not be ignored. Bearing in mind limitations of modeling, and possible discrepancies between the actual behavior of the Earth and the modeled behavior, the ICHESE commission has concluded that the combined anthropogenic actions of extraction and injection of fluids in a tectonically active region, already close to criticality, may have contributed to producing the conditions required to produce a significant earthquake (Styles et al. 2014). This example illustrates the possible difficulties in discrimination between natural and human-triggered seismicity.

Specific Features

Anthropogenic seismicity is by definition related to georesource exploitation. These relationships, the dependence

Anthropogenic Seismicity Related to Exploitation of Georesources,

Fig. 5 Fluctuations of average monthly reinjected volume of water and average injection wellhead pressure in the hydrocarbon production license, the closest to the location of the Emilia-Romagna earthquake sequence of May 2012. In April–May 2011 (marked by the vertical bar), the trend of these parameters changed from decreasing to increasing. This change yielded an increase of seismicity rate. (After Styles et al. 2014)



of seismic processes on technological processes, mean that some vital properties of anthropogenic seismicity differ from the properties of natural seismicity. Locations of anthropogenic seismic sources are in some ways linked to locations of the inducing factors. Therefore, these sources are mostly, although not exclusively, shallower than the sources of tectonic earthquakes. This shallowness results in strong shaking in epicentral areas caused even by events that are relatively weaker, and in fast decay of amplitudes of shaking with increasing epicentral distance.

Mining seismicity forms well-defined time–space clusters that are correlated in time and space with the time-varying location of mining work. Peak concentrations of seismicity locate at close distances ahead of mining work at and above unmined areas. Seismicity that is located at greater distances occurs mostly above the goaf and is usually linked to tectonic or mining discontinuities or weak zones. However, the detailed distribution of seismic events is mine specific. Fluid-injection- and fluid-extraction- induced seismicity are primarily related to the distribution of stress field changes in the subsurface surrounding the open hole of a well, the changes being caused by a combined effect of changes in pore pressure and changes in elasticity of the bedrock. Nevertheless, the distance at which earthquakes are triggered by fluid injection and the decay of seismic activity depend on the type of rock (Goebel and Brodsky 2018). In sedimentary rocks the pore pressure is transmitted to the surroundings more readily than in crystalline rocks. Consequently, induced seismic events extend further in sedimentary than in crystalline rocks. Moreover, in sedimentary rocks the seismic activity decays steadily with the distance from the injection point, whereas in crystalline rock the activity rapidly disappears beyond a certain distance.

Resulting from the complicated seismogenesis of anthropogenic seismicity, the source mechanisms of induced earthquakes, contrary to those of tectonic earthquakes, often have prevailing nonshearing components (e.g., Vavryčuk 2011). These components are on the one hand indicators of local modifications of the stress field induced by exploitation, but they depend also on local tectonics and the location of a source with respect to the geometry of technological operations. In mines the pure shearing mechanism is associated with either the slip across a preexisting fault or the slip associated with fracturing of intact rock. Events with such focal mechanisms result from stress changes induced by mining, but they are usually remote from excavations, and their focal mechanisms are controlled mostly by ambient stresses (McGarr 2005). The mechanisms composed of shear and implosive components result from strong interaction between the seismic events and the adjacent mine stope and involve combinations of fault slip and co-seismic stope closure. The mechanisms composed of deviatoric and explosive components are associated with time-dependent development of a fracture zone in the immediate environment of mining excavations (McGarr 2005). The most hazardous seismic phenomena in underground mines are rockbursts and tunnel collapses. Mining collapses are

very shallow, located on excavating levels not deeper than about 3 km. Being this shallow, they generate significant surface vibrations and rapid ground deformation. Typically, their mechanisms are mostly isotropic with a high (more than 50%) implosive element. It has been shown by Rudziński et al. (2016) that a collapse source can be modeled with a simple source model called tabular cavity collapse using diagonal moment tensor density with the nonzero elements $m_{11} = m_{22} = -1$ and $m_{33} = -3$ (Fig. 6).

Massive fluid injections and extractions cause spatiotemporal variations of crustal stresses on the reservoir scale, resulting also in diverse source mechanisms of injection-induced seismicity (Segall and Fitzgerald 1998). Tensile components of source mechanisms observed in injection-induced seismicity result from high pore pressure. Within hydrocarbon reservoirs, the least horizontal stress decreases with declining reservoir pressure caused by extraction, leading to a relative horizontal tension within the reservoir. In extensional environments, dilatant fracturing and normal faulting are always promoted near the edges of the reservoir or in regions of a high pore-pressure gradient. In regional compressional environments, production may trigger reverse faulting above and below the reservoir. In geothermal reservoirs, thermoelastic effects are also present and they even may dominate the poroelastic stresses, especially near injection wells and steam-producing fractures (Segall and Fitzgerald 1998).

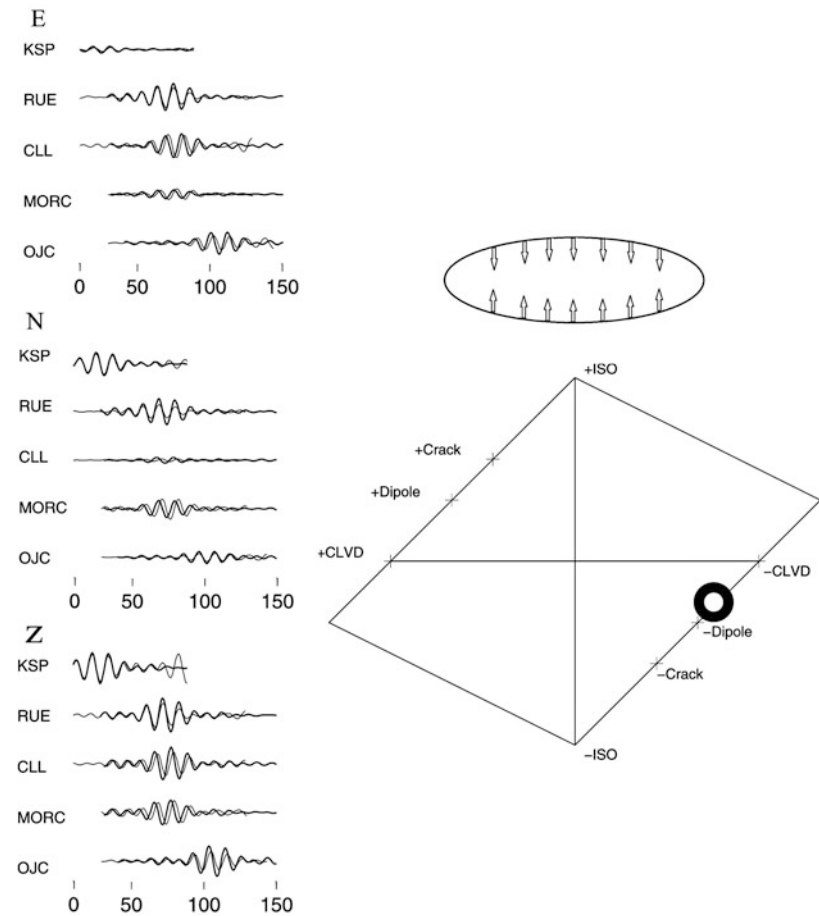
Inducing technological operations continue only in a finite (from the human point of view) time period, and the operational parameters change with time. These time constraints cause time limitations and changes in time of the induced seismic processes. The time–space location of anthropogenic seismicity is transient. In contrast with tectonic earthquakes, by detailed insight the anthropogenic seismic processes can rarely be considered as stationary. Moreover, interactions among induced seismic events caused by co-seismic stress changes have been also identified in mines and in a geothermal reservoir (Orlecka-Sikora et al. 2012; Catalli et al. 2013). These stress changes influence the space–time patterns of seismicity.

The simplest, although quite efficient, approach to infer the time changes of seismic activity assumes that the event occurrences follow a nonhomogeneous Poisson process with a slowly changing event rate so that interval stationarity is an acceptable approximation. The event rate is then modeled by the Poisson distribution in which the probability of n events occurring in the time period $[t - \Delta, t]$ is $Pr[N = n] = \frac{[\lambda(t)\Delta]^n}{n!} e^{-\lambda(t)\Delta}$, where Δ is sufficiently small to allow for assuming the event rate, $\lambda(t)$, to be constant in this interval (Lasocki 2017 and references therein).

There are also attempts to include operational parameters of the inducing technology into models of induced events occurrence (e.g., Shapiro et al. 2010; Mena et al. 2013; Garcia-Aristizabal 2018). However, such models are so far very case-specific.

Anthropogenic Seismicity Related to Exploitation of Georesources, Fig. 6

Left: observed seismograms recorded on broadband stations after a strong mining collapse that occurred in Rudna Mine, Poland, in March 2013 (gray line) and synthetic (black line) traces generated for a tabular cavity collapse. Right: sketch of the contraction of a tabular cavity collapse and its source on the Hudson source-type plot. (By courtesy of Lukasz Rudziński)



With a less granular view of the problem, when seismicity is induced by a number of activities distributed in time and space (e.g., seismicity in a broader area with many operating injection wells), the impact of time changes of these individual factors may average itself, and the seismicity can present itself as stationary in longer time periods. However, even in such cases the anthropogenic event occurrence processes often deviate significantly from Poissonianity. For a Poisson process the interevent time (the time laps between two consecutive events) distribution is exponential, but observed interevent time distributions conform to Weibull or to even more complex distribution forms (e.g., SHEER Project, D4.2 2017).

The magnitude distribution models most commonly used in tectonic seismology stem from a statistical relationship known as the Gutenberg–Richter relationship. This relationship links linearly the common logarithm of the number of events observed in consecutive, same size magnitude bins, n , to magnitude values of the bin centers, m_m , $\log n = a - b \cdot \log m_m$ or the common logarithm of the observed number, N , of events whose magnitudes are greater than or equal to any m to m , $\log N = a - b \cdot \log m$. The relationship is obeyed above a certain completeness magnitude, m_c , which is the lowest magnitude of events, which are statistically all recorded

in given monitoring conditions. The Gutenberg–Richter relation leads to exponential models of magnitude distribution, either unlimited or with upper limits.

For the same reason as does the event rate, the distribution of magnitudes of anthropogenic seismic events also changes in time. Moreover, the variety of combinations of technological and geological factors inducing anthropogenic earthquakes may result in significant deviations of the observed magnitude distribution from the models derived from the Gutenberg–Richter relation. Such deviations have been statistically ascertained (e.g., Lasocki and Orlecka-Sikora 2008; Urban et al. 2016; SHEER Project D4.2 2017). The multimodality of magnitude distribution was indicated in some cases, which suggested the observed magnitudes to be mixtures of outcomes from different fracturing processes. Usually the deviations of observed magnitude distributions from the Gutenberg–Richter relation-based models are not dramatic and do not preclude the use of these models as handy approximations. Then, the changes in time of the slope parameter, b , must be taken into account. However, these deviations have significant consequences for seismic hazard assessment (Lasocki and Orlecka-Sikora 2008; Lasocki 2017 and references therein).

Collaborative Undertakings for Research into Anthropogenic Seismicity

Research into anthropogenic seismicity and related hazards requires an insight into relationships between the cause—the inducing technological activity, and the effect—the seismicity. Because of the complexity of induced seismic processes and the diversity and time-variability of the technological conditions responsible for generating the seismic events, these relationships are complex and difficult to quantify and generalize. Much progress in the understanding of these processes requires both deep integration of research groups active in the field and an access to high-quality data sets comprehensively describing both the inducing technological activity and the seismicity. In response to the increasing need for integration among scientists and engineers interested in induced seismicity, in 2007 the Triggered and Induced Seismicity Working Group (TAIS WG) was established in the International Association of Seismology and Physics of the Earth's Interior (IASPEI). The goal of TAIS WG has been to summarize the present state of knowledge on induced and triggered seismic processes and to discuss future trends in the field.

The discussions conducted within the TAIS WG inspired the creation of open-access services for studying anthropogenic hazards associated with exploitation of georesources. This goal has been realized by international teams in the form of the Thematic Core Service Anthropogenic Hazards (TCS AH) in the framework of the European Plate Observing System Program (EPOS-ERIC) (<https://tcs.ah-epos.eu>, Lasocki and Orlecka-Sikora 2016). The TCS AH consortium runs an e-research infrastructure, an IS-EPOS platform connected with the international data nodes that provides access to large sets of relevant data and allows free experimentation in a virtual laboratory. As well as its value for scientific work, the free access for all world researchers to this international facility enables proliferating discoveries and societal applications. The TCS AH consortium promotes interdisciplinary collaborations among all stakeholders: the scientific community, industrial partners, and society. The roadmap for TCS AH includes integrating worldwide research to develop capacities for addressing anthropogenic hazards challenges; further integrating and developing research infrastructures; increasing public understanding of anthropogenic hazards; driving innovation by transferring ready-to-use solutions to the industry; and transferring the expert knowledge and experience from industry to science.

Summary

In response to rising demands for minerals, exploitation of georesources is being carried out in more and more difficult stress conditions. As a result, the appearance of anthropogenic seismicity in the affected rock mass has been, and will be, increasing.

Most induced seismic events are weak; however, the hazard posed by anthropogenic seismicity can be significant. In particular, when perturbations in the rock mass caused by exploitation of georesources meet tectonically preloaded faults, the triggered seismic events can cause material loss, injuries, and even fatalities.

Exploitation of georesources often takes place in densely populated areas. Therefore, the accompanying seismicity is of socioeconomic significance. Although the hazard posed by anthropogenic seismicity is frequently overrated, public perception of these hazards and risks cannot be neglected. Vital georesources exploitation activities can lose public confidence unless the accompanying seismic risks are accurately assessed and properly presented to the public.

Anthropogenic seismicity results from and is in the first place related to the evoking georesource exploitation processes. In particular, because the exploitation activities change in time, anthropogenic seismicity processes and the related hazards are intrinsically time dependent.

The link of anthropogenic seismicity to the respective georesource exploitation activities suggests that potentially it is possible to maintain the anthropogenic seismic risk at an acceptable level by modifying operational practices. However, research into these cause–effect relationships requires multidisciplinary data collection with mandatory inclusion of relevant technological data. Further advances in the recognition of anthropogenic seismicity problems can be achieved only through problem-oriented international integration of the research groups active in the field and deep science–industry synergy.

Bibliography

- Catalli F, Meier M-A, Wiemer S (2013) The role of Coulomb stress changes for injection-induced seismicity: the Basel enhanced geothermal system. *Geophys Res Lett* 40:72–77. <https://doi.org/10.1029/2012GL054147>
- Dahm T, Becker D, Bischoff M, Cesca S, Dost B, Fritschen R, Hainzl S, Klöse CD, Kühn D, Lasocki S, Meier T, Ohrnberger M, Rivalta E, Wegler U, Husen S (2013) Recommendation for the discrimination of human-related and natural seismicity. *J Seismol* 17:197–202
- Davis SD, Frohlich C (1993) Did (or will) fluid injection cause earthquakes? Criteria for a rational assessment. *Seismol Res Lett* 64:207–224
- Garcia-Aristizabal A (2018) Modelling fluid-induced seismicity rates associated with fluid injections: examples related to fracture stimulations in geothermal areas. *Geophys J Int* 186:793–807
- Goebel THW, Brodsky EE (2018) The spatial footprint of injection wells in a global compilation of induced earthquake. *Science* 361(6405):899–904
- Lasocki S (2017) Probabilistic assessment of mining-induced time-dependent seismic hazards, Chapter 11.3. In: Feng X-T (ed) *Rockburst mechanisms, monitoring, warning, and mitigation*. Butterworth-Heinemann (Elsevier), Oxford, UK, pp 366–380
- Lasocki S, Orlecka-Sikora B (2008) Seismic hazard assessment under complex source size distribution of mining-induced seismicity. *Tectonophysics* 456:28–37
- Lasocki S., Orlecka-Sikora B. (2016) Integrated approach to geophysical hazards induced by exploration and exploitation of georesources – to

- facilitate the way of attaining excellence. The EPOS Newsletter issue 01, July 2016, Article 03. <https://www.epos-ip.org/news-press/epos-ip-newsletter>, <http://www.rich2010.eu/rich-success-stories>
- McGarr A (2005) Observations concerning diverse mechanisms for mining-induced earthquakes. In: Potvin Y, Hudyma M (eds) Rockbursts and seismicity in mines: controlling seismic risk. Australian Centre for Geomechanics, Nedlands, pp 107–111
- McGarr A, Simpson D, Seeber L (2002) Case histories of induced and triggered seismicity. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology, part A. Academic Press, London, pp 647–661
- Mena B, Wiemer S, Bachmann C (2013) Building robust models to forecast the induced seismicity related to geothermal reservoir enhancement. *Bull Seismol Soc Am* 103:383–393
- Orlecka-Sikora B, Lasocki S, Lizurek G, Rudziński Ł (2012) Response of seismic activity in mines to the stress changes due to mining induced strong seismic events. *Int J Rock Mech Min Sci* 53:151–158
- Rudziński Ł, Cesca S, Lizurek G (2016) Complex rupture process of the March 19, 2013, Rudna mine (Poland) induced seismic event and collapse in the light of local and regional moment tensor inversion. *Seismol Res Lett* 87:274–284
- Segall P, Fitzgerald SD (1998) A note on induced stress changes in hydrocarbon and geothermal reservoirs. *Tectonophysics* 289:117–128
- Shapiro SA, Dinske C, Langenbruch C, Wenzel F (2010) Seismogenic index and magnitude probability of earthquakes induced during reservoir fluid stimulations. *Leading Edge* 29:304–309
- SHEER Project, D4.2 (2017) Report on modelling the probability distributions used in seismic hazard analysis. www.sheerproject.eu/images/deliverables/SHEER-Deliverable-4.2.pdf
- Styles P, Gasparini P, Huenges E, Scandone P, Lasocki S, Terlizze F, International Commission on Hydrocarbon Exploration and Seismicity in the Emilia Region (ICHESE) (2014) Report on the hydrocarbon exploration and seismicity in Emilia Region. <http://unmig.sviluppoeconomico.gov.it/unmig/agenda/dettagliatitizia.asp?id=175>
- Urban P, Lasocki S, Blascheck P, do Nascimento AF, Van Giang N, Kwiatek G (2016) Violations of Gutenberg–Richter relation in anthropogenic seismicity. *Pure Appl Geophys* 173:1517–1537
- Vavryčuk V (2011) Tensile earthquakes: theory, modeling, and inversion. *J Geophys Res Solid Earth* 116:B12320
- Zhang H, Eaton DW, Li G, Liu Y, Harrington RM (2016) Discriminating induced seismicity from natural earthquakes using moment tensors and source spectra. *J Geophys Res Solid Earth* 121:972–993

Archaeomagnetism

Donald H. Tarling
School of Earth, Ocean and Environmental Sciences,
University of Plymouth, Plymouth, UK

Synonyms

Magnetic dating, archaeological; Remanence dating

Definition

Archaeomagnetism is the study of all remanent magnetization associated with materials found in, or associated with, an

archaeological context. In practice, it is the application of paleomagnetic techniques to archaeological materials, predominantly in terms of their uses in dating or “sourcing” such materials. This generally means that magnetic surveys are not usually considered “archaeomagnetic,” other than when used to provide information on the actual directions and intensities of the remanent magnetizations of the materials causing the anomalies.

As most applications of archaeomagnetic studies relate to magnetic dating, these aspects are described first.

Basic Features

Most materials within archaeomagnetic contexts contain iron oxide minerals that are capable of carrying a magnetic remanence acquired at some past time. More rarely, iron sulphides and hydroxides can carry such remanences. Even more rarely, a remanence can be carried by pure iron, nickel and cobalt and their alloys, and in very specific conditions, iron sulphates. The commonness of magnetite (Fe_3O_4) and hematite (Fe_2O_3) makes these almost always the dominant minerals involved. Such ferromagnetic (*sensu lato*) minerals can acquire a *thermal remanence* in the Earth’s magnetic field (or of nearby magnetic objects) while cooling from higher than ambient temperatures, *chemical remanences* as a result of crystalline or compositional changes. Some archaeological materials can comprise iron objects, but these are generally poor recorders of early remanences, being generally magnetically “soft.” All archaeological materials can become comminuted and eventually deposited as sediments (*v.i.* Archaeological sediments) in which the individual magnetic grains can be magnetically aligned during their deposition in aqueous or aeolian conditions, resulting in a *detrital magnetization* (see ► “*Paleomagnetism, Principles*” for details).

In an archaeological environment, it is unlikely that only one form of remanence is present. It can also be presumed that some of the originally acquired remanence will have decayed with time, and other time-dependent magnetizations, *viscous magnetizations*, will have been acquired as the materials lay within a geomagnetic field that gradually changes in both direction and intensity. In order to determine the direction and intensity of the geomagnetic field for some specific time in the past (such as when last heated or deposited), it is necessary to remove such viscous magnetization. Where such viscous magnetizations are not associated with chemical changes, they can be removed by partial demagnetization in alternating magnetic fields or heating in zero-magnetic field in nonmagnetic furnaces. Alternating field demagnetization randomizes the low coercivity components of remanence, that is, the viscous remanences, leaving the high coercivity components unchanged. Partial thermal demagnetization reduces the relaxation time of all contained minerals, that is, the time

taken for the direction of individual grain magnetization to relax into the ambient field direction. As the ambient field is zero, the grains with low relaxation times (those that carry the viscous magnetizations) thereby become randomized, leaving the high relaxation grains still in original directions. (For further detail, see ► [“Paleomagnetism, Principles”](#).) Such partial demagnetization procedures are necessary for magnetic dating using directions or intensity. Statistical analyses, initially Principle Component Analyses, are then made to determine the number, direction, intensity, and precision of the magnetic vectors present in any single sample (see ► [“Paleomagnetism, Measurement Techniques and Instrumentation”](#) for details). The isolated individual sample vectors defined within some 5° that are considered to have been acquired originally are then combined, usually using Fisherian Statistics and giving equal weight to each sample vector, to provide a mean site direction. (Where individual oriented hand samples have been subsampled, these subsamples are combined to obtain the sample values that then yield the site mean value.)

Such treatments normally require different instrumentation for dating based on directions than those for dating based on intensity and are described separately. The theoretical bases are the same.

Archaeological Site Sampling

Directional Analyses

Generally between 7 and 20 individually oriented samples are taken. This enables orientation errors to be averaged, as well as measurement errors. Ideally such samples are distributed evenly throughout the structure being sampled, although concentration is made on materials that are most likely to yield meaningful results and where the physical stability (since acquiring remanence) is better established. Sample orientation in situ is preferably undertaken nonmagnetically (sun compass, gyro-theodolite, or by sightings) as the objects themselves commonly distort magnetic compass readings. It is also desirable that the nature and conditions at the site, with the in situ samples, is recorded so that, when completing the analyses, possible causes from anomalous determinations can be better evaluated (see ► [“Paleomagnetism, Measurement Techniques and Instrumentation”](#)).

Paleo-intensity Analyses

Samples for such studies do not need to be oriented while in situ nor need the samples be still in their original positions in which their remanence was acquired. However, such orientations can be useful in assisting assessments of samples that show more than one remanent magnetic vector. Generally, some three to five samples are selected for study from any one archaeological site (see ► [“Paleomagnetic Field Intensity”](#)).

Magnetic Dating Based on the Direction of Remanence

This dating method depends on the longtime spatial variations (>~10 years) of the geomagnetic field (see ► [“Geomagnetic Field, Secular Variation”](#)). Analyses of the present-day geomagnetic variations suggest that, at a medium latitude location, the observed geomagnetic field in a circular area of some 550 km radius (~1 Mkm) can be modeled as that of a geocentric inclined dipole field within an error of ~0.5° (solid angle) in direction. At greater distances, the error increases to >1° (solid angle) rendering the model increasingly unreliable for comparing the directions of remanence at more distant sites. Such modeling enables “Master Curves” (Paleo-Secular Variation Curves) of directional change to be constructed using previously studied sites for specific regions, usually for a specific country or groups of small countries. In the UK, most English and Welsh observations were corrected to a central location, Meriden, while French observations were generally corrected to Paris.

Master Curves are now more commonly calculated within a far more sophisticated statistical modeling, notably hierarchical Bayesian modeling (Lanos et al. 2005). In such analyses, the Gaussian errors for the date assignments as well as for directional properties are incorporated in the construction of the Master Curves. Such Bayesian statistics similarly incorporate harmonic analyses. Spherical Cap analyses have the objective to determine the geomagnetic field variations attributable to fields at the Earth’s core (Pavón-Carrasco et al. 2008). The latter method is therefore more appropriate for geomagnetic field modeling (see ► [“Geomagnetic Field, Theory”](#)).

Relative Dating

Nearby sites (within some 50–80 km) that acquired their remanence at the same time will exhibit identical (within 1°) directions of remanence. Consequently, nearby sites with differing directions of remanence can be considered to have acquired their remanence at different times. Conversely, similar directions can indicate similar ages for their remanence acquisition. However, identical geomagnetic directions can occur in the same location at different times. When such repeat directions occur at time intervals of a few 100 years, the archaeological context is often adequate to distinguish between alternative age assignments.

Chronological Dating

This dating method is dependent on the secular variation of the geomagnetic field being known for the region being investigated (see ► [“Geomagnetic Field, Secular Variation”](#)). As the current geomagnetic theories for the origin of the geomagnetic field are still evolving (see ► [“Geomagnetic Field, Theory”](#)), this usually requires previously well-dated

(by non-archaeomagnetic means) site values to be available. In newly studied areas, the method is clearly inhibited by the lack of such data, but as more and more observations are obtained, the past directions of the geomagnetic field become increasingly well defined. Archaeomagnetic dating is thus an unusual scientific dating technique as its precision increases as more observations accumulate.

Magnetic Polarity and Polarity Excursion Dating

This dating application is based on unusual aspects of the long-term geomagnetic behavior. On scales of a few thousand years, the geomagnetic field can undertake major excursions away from its normal direction, geomagnetic events, or excursions. On timescales of a few 100,000 years, the Earth's magnetic field changes polarity – the North magnetic pole becoming the South magnetic pole, and vice versa for the South magnetic pole. Consequently, some geomagnetic excursions occur on archaeological timescales, and the younger polarity changes occur on archaeological-anthropological timescales (see ► [“Geomagnetic Excursions”](#) and ► [“Geomagnetic Field, Polarity Reversals”](#) for more detailed considerations).

Magnetic Dating Based on the Intensity of Remanence

Relative dating can be based on the determination of paleointensities of the ancient field in a similar way as for direction (qv). Most current studies are directed toward defining the changes in geomagnetic intensity during archaeological time although sufficient quantities of high quality paleo-intensity determinations are now becoming available for specific regions, enabling chronological dating to be undertaken. As for directional studies, when past records of the Earth's magnetic field intensity are available, a Master Curve can be constructed for a similar circular area of $\sim 1 \text{ Mkm}^2$, radius of $\sim 550 \text{ km}$, as for directional studies.

Errors

General Errors

These apply to both directional and intensity dating techniques.

1. It is usually assumed that any effects of weathering and other environmental changes are minimal and that the magnetic grains carrying the remanence of interest are essentially unaffected. However, some weathering products, such as goethite and hematite, can carry high-stability remanence. It is desirable that, as far as practicable, unweathered materials are collected.

2. During partial demagnetization, it is assumed that no chemical or structural alterations occur to the existing magnetic minerals and that no new magnetic minerals are formed. Such changes are far more likely during thermal demagnetization than by alternating magnetic field techniques and are usually monitored by repeated measurements of the initial susceptibility. This is, however, a coarse measure and magnetic domain changes (see ► [“Magnetic Domains”](#)) which can occur with no clear change in susceptibility.
3. Both directions and intensities can be affected by magnetic anisotropy (shape and crystalline) that may be associated with the sample shape, the shape of contained minerals and the orientations of their crystallographic axes (see ► [“Magnetic Anisotropy”](#)).
4. Lightning strikes at times subsequent to the original acquisition of remanence, and have electromagnetic effects that can drastically alter the magnetic remanence. Although mostly associated with electrical charges that usually travel through surface waters, these effects can penetrate. Such effects can usually be recognized by their high intensities and tendency to be in random directions.
5. The local geomagnetic field direction and intensity may well be distorted if there were nearby magnetized objects, natural or manufactured, at the time of remanence acquisition.
6. The behavior of the geomagnetic field is still poorly defined and, particularly at high magnetic latitudes, can show major diurnal variations. It also appears that secular variations are not necessarily smooth, but may occur in “jerks” (see ► [“Geomagnetic Field, Secular Variation”](#)) although these appear to be mostly within the current experimental error of the technique.
7. The geomagnetic field over time can return to previous direction and intensity values. When such “crossover” points occur in the Master Curves at sufficiently different times, then the archaeological age constraints on the site may enable the most likely chronological age to be applied. However, most extant Master Curves are likely to still be distorted by some studies for which the archaeological or chronological age has subsequently been modified. However, such “anomalous” sites becoming increasingly recognized as databases increase.

Errors Specific to Directional Dating

In addition to the general errors, directional studies are sensitive to the oriented samples having remained in situ since they acquired their remanences. In a solid structure, such loose materials can be readily identified, but individual parts of some structures can have undergone motions relative to the rest of the structure, for example, kiln-wall “fall-out” or “fall-in.” On hillside, it is also possible that the whole structure may have undergone motion away from where the remanence was

acquired. Thus, a knowledge of the archaeological controls on possible motions of the site is essential.

Errors Specific to Intensity Dating

In addition to the general errors, paleo-intensity studies are sensitive to the “cooling rate” of the acquired magnetic properties. The naturally acquired remanence will commonly have been acquired over times that can be of the order of several days, while the laboratory studies are necessarily on far shorter times scales. The laboratory redox conditions during cooling are normally very different.

Overall Error Assessment

While there are many potential sources of error in archaeomagnetic dating, in most archaeological sites where the conditions are clear and the materials largely pristine, the errors arising appear to be only rarely of sufficient magnitude to warrant rejection.

Specialized Applications

Murals and Plaster

Remarkably, it has been shown that pigments can carry a remanence relating to the time when they were applied. Where such pigments have been applied to walls, their remanence has been shown to relate to the time that the pigments were applied, particularly in tempera paintings – either originally or during subsequent retouching. It seems most likely that this remanence is associated with hematite and was acquired “detritally,” that is, by rotation of the grains while still fluidized (Lanza et al. 2009). Similar observations have also been shown for plaster work (Soler-Arechalde et al. 2006).

Object Reconstruction

Specific artifacts, such as pottery, when fired, obtain a uniform magnetization in the ambient geomagnetic field as they cool within the furnace. If broken, individual shards retain this original remanence and can be reassembled to assist in defining the original shape of the pot (Burnham and Tarling 1975).

Archaeological Sediments

As sediments are deposited within an archaeological site, any contained already magnetized grains are influenced by the ambient geomagnetic field. As these alignment forces are weak (compared with the effects of gravity and flow motions) only a percentage of the magnetic grains are aligned and such alignments can be gravitational flattening during the deposition process. However, the more magnetically stable grains (usually with single-domain dimensions of $\sim 1 \mu\text{m}$) can rotate into full alignment while they are still fluidized immediately after deposition. However, such alignments can be modified by meniscus

forces as the water table subsequently rises or falls. Sedimentary minerals also undergo a series of chemical reactions as they adjust to their new environment. Consequently, it is often difficult to establish the true direction of the geomagnetic field at the time of deposition with sufficient certainty to assign ages based on the magnetic determinations. However, there are archaeological sites, such as basal sediments, deposited immediately after a ditch had been cut, that appear to be recording the geomagnetic field direction at that time.

Magnetic “Sourcing”

Magnetic sourcing attributes an object to a particular location on the basis of its magnetic properties. For example, an obsidian arrowhead can sometimes be sourced to the obsidian outcrop from which it was originally fashioned.

The physical basis for this is that the compositions and grains sizes of the magnetic minerals within a rock can be specific to a particular outcrop. Each outcrop also tends to have a slightly different cooling history and became magnetized by a specific geomagnetic field strength. Several magnetic properties can be rapidly and cheaply measured, such as its initial (low field) susceptibility, its saturation remanence, and its intensity of natural remanence, are sufficient to identify a specific outcrop. Generally, the range of magnetic properties in most sedimentary rocks is too great to allow such assignments, but obsidians, rapidly cooled volcanic glasses, commonly have a range of magnetic properties that are either unique to that outcrop or to only two or three outcrops.

Summary

The study of the magnetic properties of archaeological materials, archaeomagnetism, is particularly significant in geophysics as it enables secular variations in the direction and strength of the geomagnetic field over timescale that are far longer than those for direct measurement (~ 400 years). In comparable magnetic studies of rocks, paleomagnetism, these secular variations are commonly averaged out during the processes by which rocks acquire their remanent magnetization. Conversely, archaeomagnetic records provide an additional dating tool for archaeologists. As a comparative dating tool, archaeomagnetism enables relative dating between nearby samples (100–200 km) within ~ 10 –20 years. As a dating tool, it is mainly dependent on the validity of the ages used to compile the regional Master Curve. It is thus unique, as a scientific dating method, in that as more data are acquired, age “outliers” can be identified and their age reassessed. Consequently, the Master Curves are continually improving and, ultimately, could have the same precision as for relative dating. There are also interesting applications in using such observations in conservation studies and establishing past environmental conditions.

Cross-References

- [Geomagnetic Excursions](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Geomagnetic Field, Secular Variation](#)
- [Geomagnetic Field, Theory](#)
- [Magnetic Anisotropy](#)
- [Magnetic Domains](#)
- [Paleomagnetic Field Intensity](#)
- [Paleomagnetism, Measurement Techniques and Instrumentation](#)
- [Paleomagnetism, Principles](#)

Bibliography

- Batt CM, Zanarini I (eds) (2008) Archaeomagnetic applications for the rescue of cultural heritage (AARCH). *Phys Chem Earth* 33(6–7):403–608
- Burnham RJP, Tarling DH (1975) Magnetisation of shards as an assistance to the reconstruction of pottery vessels. *Stud Conserv* 20:152–158
- Lanos P, Le Goff M, Kovacheva M, Schnepf E (2005) Hierarchical modeling of archaeomagnetic data and curve estimation by moving average technique. *Geophys J Int* 160:440–476
- Lanza R, Zanella E, Sandino S (2009) Magnetic remanence of haematite-bearing minerals. *Geophys Res Lett* 36:L24302
- Pavón-Carrasco FJ, Osete ML, Torta JM, Gaya-Piqué LR, Lanos P (2008) Initial SCHADL00 regional archaeomagnetic model for Europe for the last 2000 years. *Phys Chem Earth* 33(6–7):597–608
- Soler-Arechalde AM, Sánchez F, Rodríguez M, Caballero-Miranda A, Goguitchaichvil A, Urrutia-Fufugauchi J, Manzanilla L, Tarling DH (2006) Archaeomagnetic investigation of oriented pre-Columbian lime-plasters at Teotihuacan, Mesoamerica. *Earth Planets Space* 58:1433–1439

Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India

Harsh K. Gupta

National Geophysical Research Institute, Council of Scientific and Industrial Research, Hyderabad, India

Definition and Introduction

Under certain suitable geological conditions, anthropogenic activity can trigger or induce earthquakes. The triggered/induced earthquakes are known to have occurred due to gold and coal mining, petroleum production, filling of artificial water reservoirs, high-pressure liquid injection into ground, geothermal energy, and natural gas production. The largest scientifically accepted triggered earthquake of M 6.3 occurred on 10 December 1967 in the vicinity of Koyna Dam near the west coast of India. It is debated

whether the M 7 Gazli earthquakes of May 1976 and 19 March 1984 were induced due to the production of large quantities of gas at the Gazli oil field in Uzbekistan. There is an argument that the Sichuan, China, M 7.9 earthquake of 12 May 2008, that claimed over 80,000 human lives, was triggered due to filling of the nearby Zipingpu reservoir. It has been also proposed that flooding of a river near San Andreas Fault in California caused at least two M~6 earthquakes. A good account of triggered/induced seismicity can be found in a review by (McGarr et al. 2002). Hudyma and Potvin (2005) and Drzewiecki and Piemikarczyk (2017) have extensively dealt with mining-induced seismicity. Gupta (2002) has reviewed artificial water reservoir-triggered seismicity.

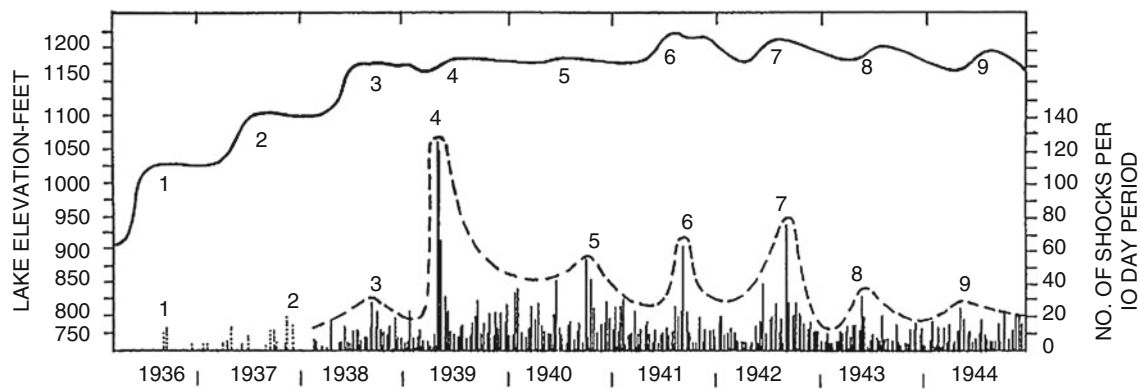
Here we present cases of artificial water reservoir-triggered earthquakes that occurred all over the world, with a special emphasis on earthquakes in the Koyna region near the west coast of India. This is a classical case where triggered earthquakes have been occurring since the impoundment of reservoir in 1962 till now in 2019.

Triggered vis-à-vis Induced Earthquakes

For a long time, the adjectives “induced” and “triggered” were used interchangeably whenever one talked of artificially simulated earthquakes. McGarr and Simpson (1997) have addressed this question and suggested that it would be important to draw a distinction between the two. They proposed that the adjective “triggered seismicity” should be used only when a small fraction of stress change or energy associated with earthquakes is accounted for by the causative activity. The term “induced seismicity” should be used where the causative activity is responsible for a substantial part of the stress change. In the case of triggered seismicity, tectonic loading plays an important role. The stress-level changes associated with filling of some of the deepest artificial water reservoirs are only of the order of 1 MPa or so, whereas the stress drop associated with the earthquakes is much larger. Therefore, all cases of earthquakes occurring subsequent to filling of the artificial water reservoirs fall in the category of “triggered earthquakes,” and hence it is appropriate to call it “reservoir-triggered seismicity” (RTS).

Artificial Water Reservoir-Triggered Earthquakes

The generation of hydroelectric power, flood control, and irrigation has led to the creation of huge artificial water reservoirs globally. For the first time, triggering of earthquakes was pointed out by Carder (1945) at Lake Mead in the USA. Figure 1 depicts Lake Mead water levels and local seismicity for the period 1936 through 1944. The rise in water level and corresponding bursts of seismic activity are numbered. The correspondence is indeed remarkable. Damaging triggered earthquakes exceeding M 6 occurred at Hsingfengkiang, China (1962); Kariba, Zambia-Zimbabwe border (1963);



Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Fig. 1 Lake Mead water levels and the local seismicity. For 1936 and 1937, only the felt shocks are plotted. The

risers in water levels and the corresponding bursts of seismic activity are numbered. General trend of tremor-frequency variation is shown by dotted lines (After Carder 1945)

Kremasta, Greece (1966); and Koyna, India (1967). Koyna earthquake of M 6.3 that occurred on 10 December 1967 is so far the largest scientifically accepted triggered earthquake. It claimed over 200 human lives, injured about 1500, and rendered thousands homeless. The occurrence and potential of triggered earthquakes has caused major modification of civil works and engineering projects. Anticipating a large triggered earthquake, the Hsingfengkiang Dam was strengthened twice before the occurrence of M 6.1 earthquake on 20 March 1962 (Shen et al. 1974). The disposal of waste fluid through injection into the ground at the Rocky Mountain Arsenal had to be discontinued due to triggered earthquakes (Evans 1966). The possibility of high-magnitude-triggered seismicity was responsible for terminating the Auburn Dam project in California (Allen 1978). There is a general reluctance on the part of Engineering Community, globally, to accept the significance or even existence of the phenomenon of triggered seismicity (Allen 1982; Simpson and Leith 1988). What Allen (1982) said some four decades ago, “From a purely economic point of view, not to speak of public safety, the problem of reservoir induced earthquakes deserves far more attention than it currently is receiving in most parts of the world,” is still true.

Reservoir-Triggered Seismicity (RTS): Global Distribution

Globally there are hundreds of sites where RTS is reported to have occurred. Gupta (2011), on the basis of magnitude of the largest triggered earthquake, grouped these sites in the following categories (Table 1):

1. Sites where largest earthquake exceeded M 6.0 (4 sites)
2. Sites where the largest earthquake M was 5.0 to 5.9 (10 sites)
3. Sites where the largest earthquake M was 4.0 to 4.9 (29 sites)

Over the years, RTS has been reported from many more sites. Among these, a mention must be made of Polyphyto Dam, Greece, where the largest triggered earthquake of M 6.5 occurred on 13 May 1995 (Pavlou 2019), and Tranh River No.2, Vietnam, where the largest triggered earthquake was of M 4.6 on 22 October 2012 (Trieu et al. 2014).

Important Factors for RTS

Several studies examined the correspondence among possible correlates like rate of loading, highest water level reached and the duration of retention of high water levels, and the occurrence of RTS. The most important correlate is the depth of water column in the reservoir (Baecher and Keeney 1982). Figure 2 demonstrates that when the water column depth exceeds 150 m, about one in five reservoirs has experienced RTS. A review of recent global examples gives support to this observation. It must also be noted that a reservoir with a water volume exceeding 1 km³ and/or water depth exceeding 100 m is called a large reservoir. Globally there are more than one thousand such reservoirs, and only a small percentage has evidenced RTS.

Common Characteristics of RTS Sequences

By the early 1970s, over a dozen cases of RTS were known. In a couple of detailed studies, Gupta et al. (1972a, b) discovered several common characteristics of RTS sequences, which discriminated them from the normal earthquake sequences occurring in close by regions but not associated with reservoirs. These characteristics are:

1. In the earthquake frequency-magnitude relation ($\log N = A - bM$, where N is the number of earthquakes with magnitude $\geq M$ and A and b are constants), the foreshock

Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Table 1 Reported cases of reservoir-triggered seismicity (RTS) where $M \geq 4$ earthquake occurred (Gupta 2011)

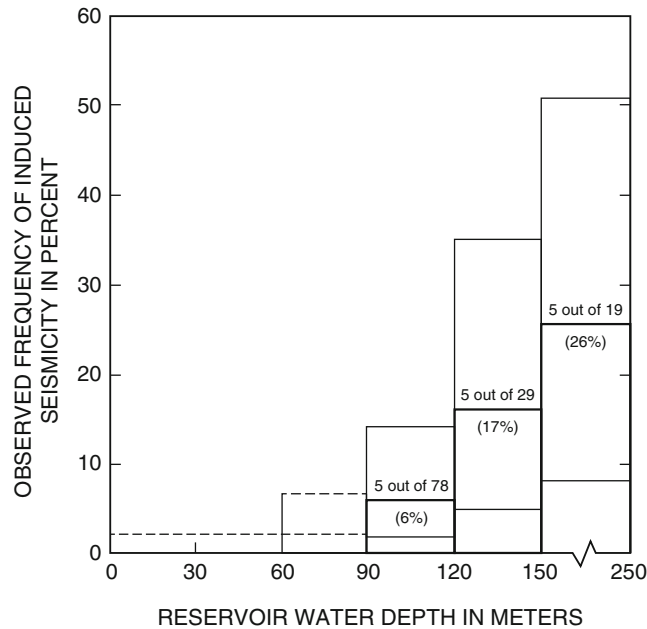
Name of the dam/reservoir	Country	Height of dam (m)	Reservoir volume (10^6 m^3)	Year of impounding	Year of the largest earthquake	Magnitude/intensity	References
Sites where earthquakes having magnitude ≥ 6.0 were triggered							
Hsingfengkiang	China (PRC)	105	13,896	1959	1962	6.1	1, 2, 3
Kariba, Zambia	Zimbabwe	128	175,000	1958	1963	6.2	1, 2, 6
Koyna	India	103	2,780	1962	1967	6.3	1, 2, 4, 5
Kremasta	Greece	160	4,750	1965	1966	6.2	1, 2, 4, 5
Sites where earthquakes having magnitude between 5.0 and 5.9 were triggered							
Aswan	Egypt	111	1,64,000	1964	1981	5.6	2, 7
Benmore	New Zealand	110	2,040	1964	1966	5.0	1, 2, 8
Charvak	Uzbekistan	148	2,000	1971	1977	5.3	25
Eucumbene	Australia	116	4,761	1957	1959	5.0	2
Geheyan	China	151	3,400	1993	1997	VI	22a
Hoover	USA	221	36,703	1935	1939	5.0	1, 2, 10
Marathon	Greece	67	41	1929	1938	5.7	1, 2, 4, 5
Oroville	USA	236	4400	1967	1975	5.7	2, 11
Srinagarind	Thailand	140	11,750	1977	1983	5.9	28
Warna	India	80	1,260	1985	1993	5.0	24
Sites where earthquakes having magnitude between 4.0 and 4.9 were triggered							
Akosombo Main	Ghana	134	148,000	1964	1964	V	2, 9
Bajina Basta	Yugoslavia	90	340	1966	1967	4.5–5.0	2, 5
Bhatsa	India	88	947	1981	1983	4.9	20
Bratsk	Russia	100	169		1996	4.2	22b
Camarillas	Spain	49	37	1960	1964	4.1	2, 4, 5
Canelles	Spain	150	678	1960	1962	4.7	2, 4, 5
Capivari-Cachoeira	Brazil	58	180	1970	1971	VI	17
Clark hill	USA	60	3517	1952	1974	4.3	2, 12
Dahua	China (PRC)	74.5	420	1982	1993	4.5	26
Danjiangkou	China (PRC)	97	16,000	1967	1973	4.7	16
Foziling	China (PRC)	74	470	1954	1973	4.5	16
Grandwal	France	88	292	1959	1963	V	1, 2, 4, 5
Hoa Binh	Vietnam	125		1988	1989	4.9	22c
Kastraki	Greece	96	1000	1968	1969	4.6	2
Kerr	USA	60	1505	1958	1971	4.9	1, 2, 9
Komani	Albania	130	1600	1985	1986	4.2	27
Kurobe	Japan	186	149	1960	1961	4.9	2, 13
Lake Baikal	Russia					4–4.8a	23
Lake Pukaki	New Zealand	106	9000	1976	1978	4.6	21
Manicouagan 3	Canada	108	10,423	1975	1975	4.1	2
Marimbondo	Brazil	94	6150	1975	1975	IV	18
Monteynard	France	155	275	1962	1963	4.9	1, 2, 4, 5
Nurek	Tajikistan	317	1000	1972	1972	4.6	1, 2, 14
P. Colombia/V. Grande	Brazil	40/56	1500/2300	1973–1974	1974	4.2	19
Piastra	Italy	93	13	1965	1966	4.4	2, 4, 5
Pieve di Cadore	Italy	116	69	1949	1950	V	2, 15
Shenwo	China (PRC)	50	540	1972	1974	4.8	16

(continued)

Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Table 1 (continued)

Name of the dam/reservoir	Country	Height of dam (m)	Reservoir volume (10^6 m^3)	Year of impounding	Year of the largest earthquake	Magnitude/intensity	References
Voughlans	France	130	605	1968	1971	4.4	2, 4, 5
Karun-III	Iran	185	2970	2005	2005	4.3	29

References: 1 = (Gupta and Rastogi 1976); 2 = (Packer et al. 1979); 3 = (Shen et al. 1974); 4 = (Rothe 1970, 1973); 5 = (Bozovic 1974); 6 = (Gough and Gough 1970b); 7 = (Toppozada 1982); 8 = (Adams 1974); 9 = (Simpson 1976); 10 = (Carder 1945); 11 = (Bufe et al. 1976); 12 = (Talwani 1976); 13 = (Hagiwara and Ohtake 1972); 14 = (Soboleva and Mamadaliev 1976); 15 = (Caloi 1970); 16 = (Oike and Ishikawa 1983); 17 = (Berrocal (personal communication), 1990); 18 = (Veloso et al. 1987); 19 = (Berrocal et al. 1984); 20 = (Rastogi et al. 1986); 21 = (Reyners 1988); 22a = (Chen et al. 1996), 22b = (Pavlenov and Sherman 1996), 22c = (Tung 1996), 23 = (Djadkov 1997), 24 = (Rastogi et al. 1997); 25 = (Plotnikova et al. 1992); 26 = (Guang 1995); 27 = (Muco 1991); 28 = (Chung and Liu 1992); 29 = (Kangi and Heidari 2008)



Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Fig. 2 Height of water column is the most important correlate (After Stuart-Alexander and Mark 1976)

and aftershock b values of the RTS sequences are higher than the b values for natural earthquake sequences in the regions concerned and the regional b values.

2. In addition to high b values, the magnitude ratio of the largest aftershock to the mainshock is also high.
3. Aftershock activity decays slowly compared to normal earthquake sequences.
4. The foreshock-aftershock sequence pattern belongs to Type II of Mogi's model (Mogi 1963), whereas the natural earthquake sequence pattern belongs to Type I of Mogi's model (Fig. 3 I).

The above mentioned observations are governed by the mechanical properties of the media, and their deviation from the normal implies changes in these properties consequent to impoundment of the artificial water reservoir. The sketch shown in Fig. 3 illustrates the changes. "A" in this figure is a homogeneous media rock volume. When the stress exceeds the strength

of the rock, there would be a major event releasing most of the strain, followed by peripheral adjustment aftershocks. In such an earthquake sequence, there would not be any foreshocks before the major event. The aftershock activity would be over in a short time, the ratio of the largest aftershock to the main event would be low, and the b value would also be low. This is typically the situation with the earthquake sequences in stable continental regions not associated with the reservoir loading. Due to filling of the water reservoir, the heterogeneity of the media increases ("B" in Fig. 3 II), and the rock volume gets fragmented. This results in the release of accumulated stresses by smaller rock volumes. In such a situation, the events would start occurring, as and when the strength of an individual rock volume is exceeded. The main event would correspond to the largest rock volume, and there would be events before it and after it, changing the pattern from Type I of Mogi's model to Type II. The ratio of the magnitude of the largest aftershock to the mainshock would also be high, and the b value of the foreshock sequence as well as the aftershock sequence would be high. This is what is observed with RTS sequences. These common characteristics were identified in the 1970s, based on around two dozen cases of RTS. Over the years, these are found to hold good for all cases of RTS investigated subsequently.

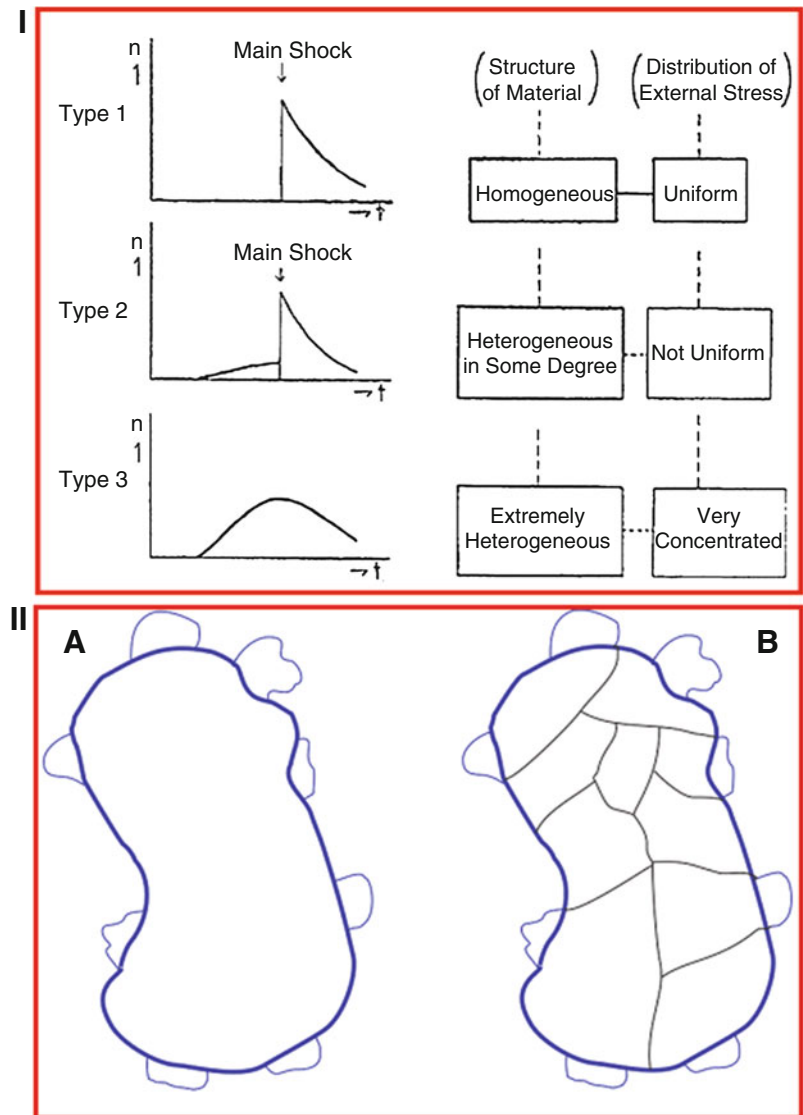
Mechanism of Triggered Earthquakes

In the following we give a gist of major milestones in comprehending the phenomenon of triggered earthquakes.

The foundation of understanding the phenomenon of triggered earthquakes was laid by the study of the waste fluid injection-induced earthquakes in the vicinity of the Rocky Mountain Arsenal well near Denver, Colorado, USA, in the early 1960s (Evans 1966). There are three main effects of reservoir loading relevant to triggering of earthquakes as pointed out by Gough and Gough (1970b) and several others:

1. The elastic stress increases following filling of the reservoir.
2. The increase in pore fluid pressure in saturated rocks, basically due to decrease in pore volume due to compaction, in response to increase in elastic stress.
3. Pore pressure changes related to fluid migration.

Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Fig. 3 (I) Depicts (Mogi 1963) classification of the earthquake sequences into three broad categories. (II) Cartoon for (A) homogenous rock mass and (B) fragmented rock mass. For details see the text



Gough and Gough (1970a, b) provided the first definitive quantitative analysis of the role of the load of the water reservoir in triggering earthquakes at Lake Kariba. The role of reservoir load in triggering earthquakes was also considered by Bell and Nur (1978) and Roeloffs (1988). They pointed out that reservoir-load-induced stresses at seismogenic depths are very small and can only perturb the ambient stress field. Gupta et al. (1972a, b) identified the rate of increase of reservoir water levels, maximum water levels reached, and the duration of retention of high water levels as factors affecting the frequency and magnitude of triggered earthquakes. The influence of pore fluid pressure in inducing earthquakes in very simple 1D reservoir models was presented by (Snow 1972). More sophisticated models were dealt by Withers and Nyland (1976), Bell and Nur (1978), and Roeloffs (1988) based on Biot's (1941) consolidation theory, which later on is generalized by Rice and Cleary (1976) by recognizing that the fluids too may be compressible. Very interesting results related to modeling of pore

pressure diffusion have been reported for Acu Reservoir in Brazil (Do Nascimento et al. 2004).

Pore pressure diffusion plays a very important role in the triggering of the earthquakes. However, there are a very few direct measurements of diffusivity, and it is mostly inferred from the temporal migration of seismicity. Talwani et al. (1999) have reported in situ measurements of hydrological diffusivity and Skempton's coefficient at the Bad Creek Reservoir in South Carolina, USA. At an observation well located 250 m away from the reservoir, a change in the water level of the well had a direct correspondence with the changes in the water level of the reservoir, initially with a delay of 98 h, which later stabilized at 72 h. This led to a frequency-independent estimate of diffusivity of $\sim 0.076 \text{ m}^2 \text{ s}^{-1}$ and Skempton's coefficient of 0.66 for an undrained response of the reservoir. Later, Talwani et al. (2007) analyzed more than 90 case histories of triggered seismicity and found diffusivity to vary between 0.1 and $10 \text{ m}^2 \text{ s}^{-1}$. This range of diffusivity values corresponds to a

range of intrinsic permeability between 5×10^{-16} and $5 \times 10^{-14} \text{ m}^2$. Talwani et al. (2007) referred to this range of the permeability of fractures as seismogenic permeability. Seismogenic permeability is an intrinsic property of fractures where pore pressure diffusion is associated with seismicity.

The in situ measurements of physical properties and examination of physical mechanism controlling triggered seismicity at Monticello Reservoir, South Carolina, USA, provided the much-needed field verification of the theoretical and model developments of the concept of triggered seismicity (Zoback and Hickman 1982). The effect of changes in lake levels and other derived parameters on triggered earthquakes have been dealt by Simpson and Negmatullaev (1981) for the Nurek Dam and by Gupta (1983) for the Koyna Dam. The part played by pore pressure diffusion in triggering earthquakes has been dealt by (Talwani and Acree 1984/1985). The effect of inhomogeneities in rock properties on triggering earthquakes was addressed by (Simpson and Narasimhan 1990).

Most of the theoretical models discussed the effect of pore fluid pressure in isotropic rocks. Chen and Nur (1992) point out that deviatoric effect of pore fluid pressure in anisotropic rocks has wide applications in comprehending triggered seismicity, earthquake precursors, and aftershocks. This approach needs to be applied to a few cases of RTS.

Other interesting studies include the notion of fault stability as a measure of the interplay of frictional stresses mobilized and the resolved shear stresses acting on a fault (Chander and Kalpna 1997). Kalpna and Chander (2000) have developed an algorithm for the simulation of stresses and pore pressure for more realistic laterally finite 3D models of reservoirs.

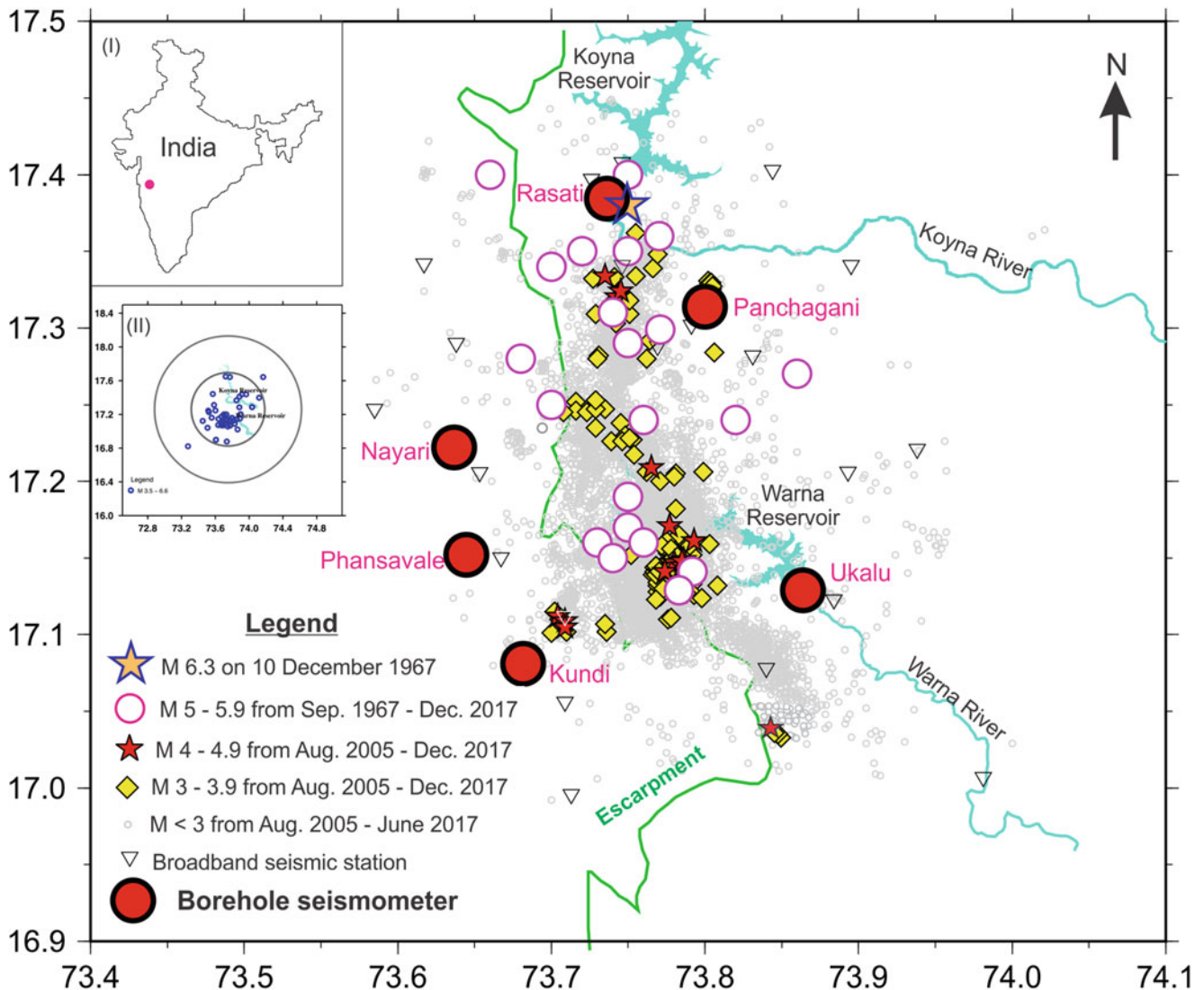
Koyna, India

Koyna reservoir located close to the west coast of India continues to be the most significant site of triggered earthquakes. Earthquakes began to occur soon after the impoundment of the Shivaji Sagar Lake created by the Koyna Dam in 1962. So far, globally the largest triggered earthquake of M 6.3 on 10 December 1967, 22 earthquakes of $M \geq 5$, about 200 earthquakes of $M \geq 4$, and several thousand smaller earthquakes have occurred in the region. Talking to the residents in the region in the 1960s revealed that they had not experienced any earthquake in the region in their living memory. There is a seismic station operating at Pune, about 120 km away from Koyna. A scrutiny of the seismograms revealed no earthquakes that could be assigned to the Koyna region. Another water reservoir, Warna, was impounded in 1985. This reservoir is located 35 km SSE of Koyna (Fig. 4). After the impoundment of Warna reservoir, the triggered activity got enhanced. For convenience, we shall call the Koyna and the Warna reservoir

region as the Koyna region. Major bursts of seismic activity associated with the Koyna reservoir occurred in 1967, 1993, 1980, 1993–1994, 2005, and 2009.

How Long Triggered Earthquakes Will Continue at Koyna?

Most RTS sites are located in stable continental regions (SCRs), where strain is of the order of 10^{-9} to 10^{-10} as compared to 10^{-5} at plate boundary and 10^{-7} to 10^{-9} for intraplate regions (Gupta and Johnston 1998). RTS typically gets initiated soon after the filling of the reservoir and may continue for a year to several years, with the largest event occurring within a few years of impoundment. The Hsingfengkiang Reservoir in China was impounded in 1959, and the largest RTS event of M 6.1 occurred on 19 March 1961. The frequency of RTS events dropped from thousands every month in the 1960s to a few every month by 1978 (Ishikawa and Oike 1982; Gupta 1992). Lake Kariba, located at Zambia-Zimbabwe border, was filled in 1958. The frequency of earthquakes increased several folds with the largest RTS event of M 6.2 occurring on 13 April 1963 (Gough and Gough 1970a, b). By 1974, RTS events dropped to less than 10% of 1963 level. At Kremasta, Greece, impoundment took place in 1965, and there was a very rapid filling of the reservoir in January 1966, and the largest RTS event of M 6.2 occurred on 5 February 1966 (Galanopoulos 1967; Stein et al. 1982). The RTS at Kremasta was short lived. However, at Koyna RTS was initiated in 1962 and has continued till now with the latest M 4 earthquake occurring on 20 June 2019. Gupta et al. (2002) have examined in detail the question as to how long the triggered earthquakes would continue at Koyna. The maximum credible earthquake for the Indian shield region has been estimated to be M 6.8. It is hypothesized that the region between Koyna and Warna was stressed close to critical before the impoundment of the Koyna reservoir and was capable of generating an M 6.8 earthquake. As demonstrated through the study of b values in earthquake magnitude-frequency relation, foreshock-aftershock patterns, the ratio of the magnitude of the largest aftershock to the mainshock, and the decay of aftershock activity in earthquake sequences at Koyna, the heterogeneity of the media has increased. None of the M~5 earthquake has occurred at the same location. M~5 earthquakes occur in Koyna region when the previous water maximum in the reservoir has been exceeded. Long time ago, Kaiser (1953) had reported that acoustic emission, under monotonically increasing stress, shows an appreciable increase after the applied stress exceeds the previously applied stress maxima. This approach has been successfully used by Yoshikawa and Mogi (1981) to estimate crustal stresses from cored samples. For the Nurek Dam also, it was reported that major triggered



Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Fig. 4 Updated from Gupta et al. (2017). Koyna-Warna region near west coast of India. Location of the Koyna main earthquake of 10 December 1967; earthquakes of $M \sim 5$ during August 1967 through June 2017; smaller earthquakes from August 2005 through June 2017; surface and borehole seismic

stations; green curve indicates the WGE (Western Ghat Escarpment). (I) Koyna location in India; (II) Distribution of $M \geq 3.7$ earthquakes for 1967–2015 (USGS) in the vicinity of Koyna and an outer circle of 100 km radius indicating that there is almost no seismic activity outside the Koyna region

events occurred when the water level in the reservoir reached the previous maximum and exceeded it (Simpson and Negmatullaev 1981). Gupta et al. (2002) had concluded that as of 2002, about one-half of an $M 6.8$ earthquake energy has been released in the Koyna region, and the activity should continue for another three to four decades. Due to the increase in heterogeneity, no fault segment, long enough to generate a $M > 6$ earthquake, is left intact, so an earthquake like the 10 December 1967 may not occur. The occurrence of $M \sim 5$ earthquakes would be governed by the fact whether the previous water maximum has been exceeded, and other factors such as the rate of loading, the highest water levels reached, and the duration of the retention of high water levels.

Short-Term Earthquake Forecast at Koyna

A case was made that short-term earthquake forecast may be feasible at Koyna (Gupta 2001). This was based on the fact that the shallow (depth ≤ 10 km) seismic activity in the vicinity of the Koyna is confined to an area of 20×30 km², and there is no other seismic source within 50 km radius. Every year, following the rainy season, the reservoirs get filled, and there is an enhancement of seismic activity. From among earthquake precursors, foreshocks are important and have a potential in forecasting earthquakes (Dodge et al. 1996; Helmstetter et al. 2003). Ohnaka (1992) has noted that immediate foreshock activity is a part of the nucleation process leading to mainshock

Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Table 2 Events preceding $M \geq 4.0$ earthquakes in Koyna region

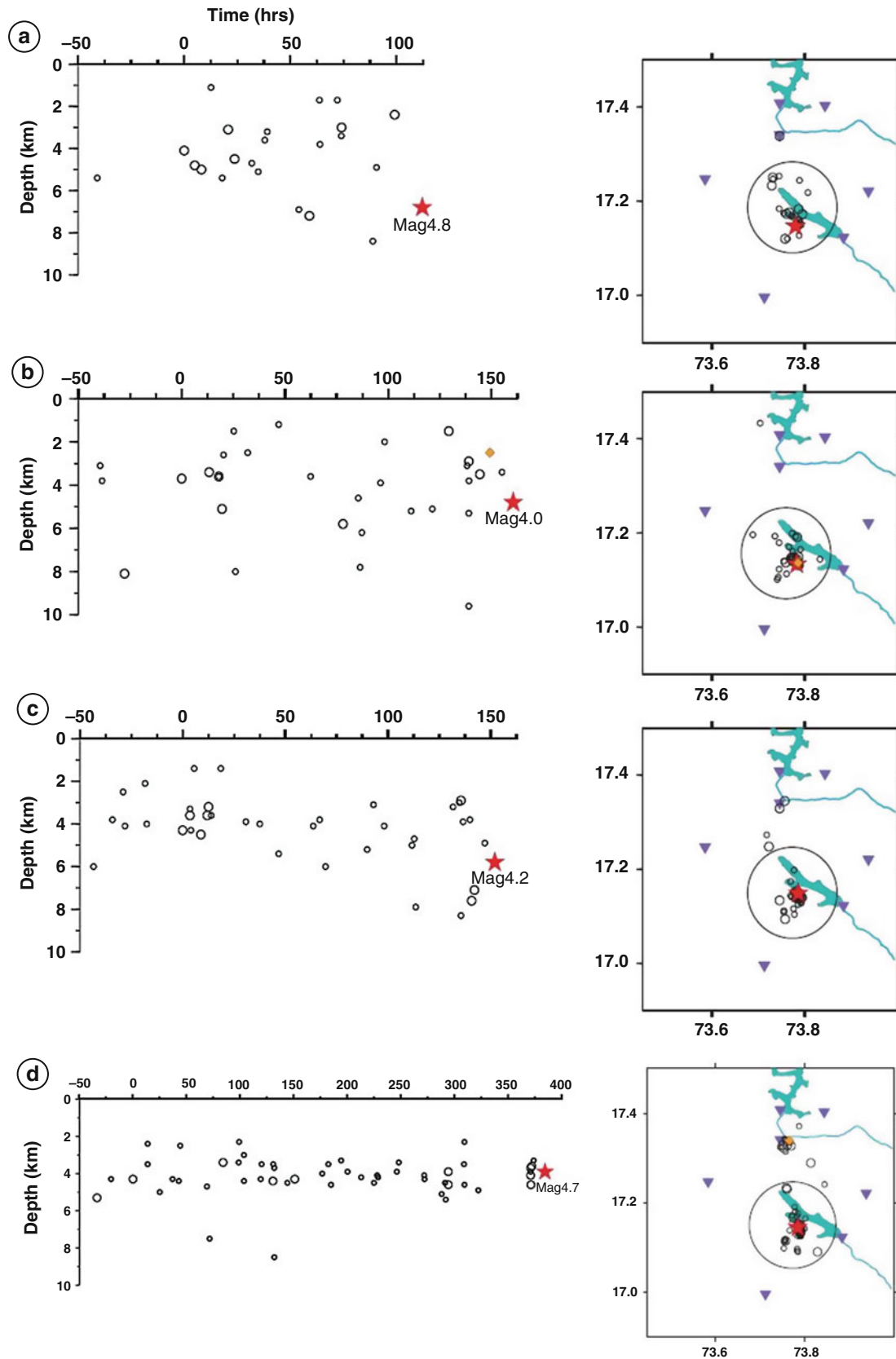
Main earthquake	Duration of nucleation period (hours)	No. of events before the main earthquake				
		<i>M</i> 1.0–1.9	<i>M</i> 2.0–2.9	<i>M</i> 3.0–3.9	Largest earthquake	50 h prior to the beginning of nucleation period
30 August 2005 M 4.8	110	13	8	0	2.6	2
13 November 2005 M 4.0	160	18	8	1	3.0	3
26 December 2005 M 4.2	150	22	8	0	2.9	6
17 April 2006 M 4.7	380	40	10	0	2.7	2

dynamic rupture. Gupta et al. (2007) have reported successful earthquake forecast in Koyna based on identification of nucleation in real time. The seismic activity in the Koyna region is monitored by a closely spaced network of seven modern seismic stations. Table 2 from Gupta et al. (2007) is the learning phase where four earthquakes of $M 4.0$ to $M 4.8$ were found to be preceded by well-defined nucleation where several events occurred in a tight cluster of less than 10 km radius (Fig. 5a–d). It was realized that if the formation of nucleation could be identified in real time before the occurrence of the mainshock, it may be possible to make a short-time forecast.

During the middle of May 2006, an interesting situation was developed. The events in the Koyna region for the period 11–16 May 2006 are depicted in Fig. 6a (Gupta et al. 2007). By the afternoon of 16 May 2006, some 50 events of $M \geq 1.0$, the largest being $M 2.7$, had occurred in the preceding 107 h in a small area, the focal depth being between 2 and 8 km. It was inferred that the region is going through a nucleation phase. Based on the experience of previous nucleation, the following forecast was made at 19:05 IST and communicated to the Editor of *Current Science*, Secretary of the Ministry of Earth Sciences of the Government of India, and the President of the Geological Society of India: “On the basis of the data available from seven seismic stations operating in the Koyna region, we have identified a nucleation which started on 12 May 2006. This may lead to the occurrence of an $M \sim 4$ earthquake in the next 15 days. This shallow earthquake (focal depth less than 8 km) will occur within a radius of 10 km centered at 17.1°N , 73.8°E . On the basis of our previous experience of studying nucleation-preceding earthquakes in the Koyna region, we expect this earthquake to occur over the next 15 days time (till 31 May 2006), with a 50% probability.” An earthquake of $M 4.2$ occurred on 21 May. Table 3 gives the comparison of forecasted parameters and that of 21 May 2006 earthquake. Since 2006, eight similar short-term forecasts have been made in the Koyna region, and all have come true.

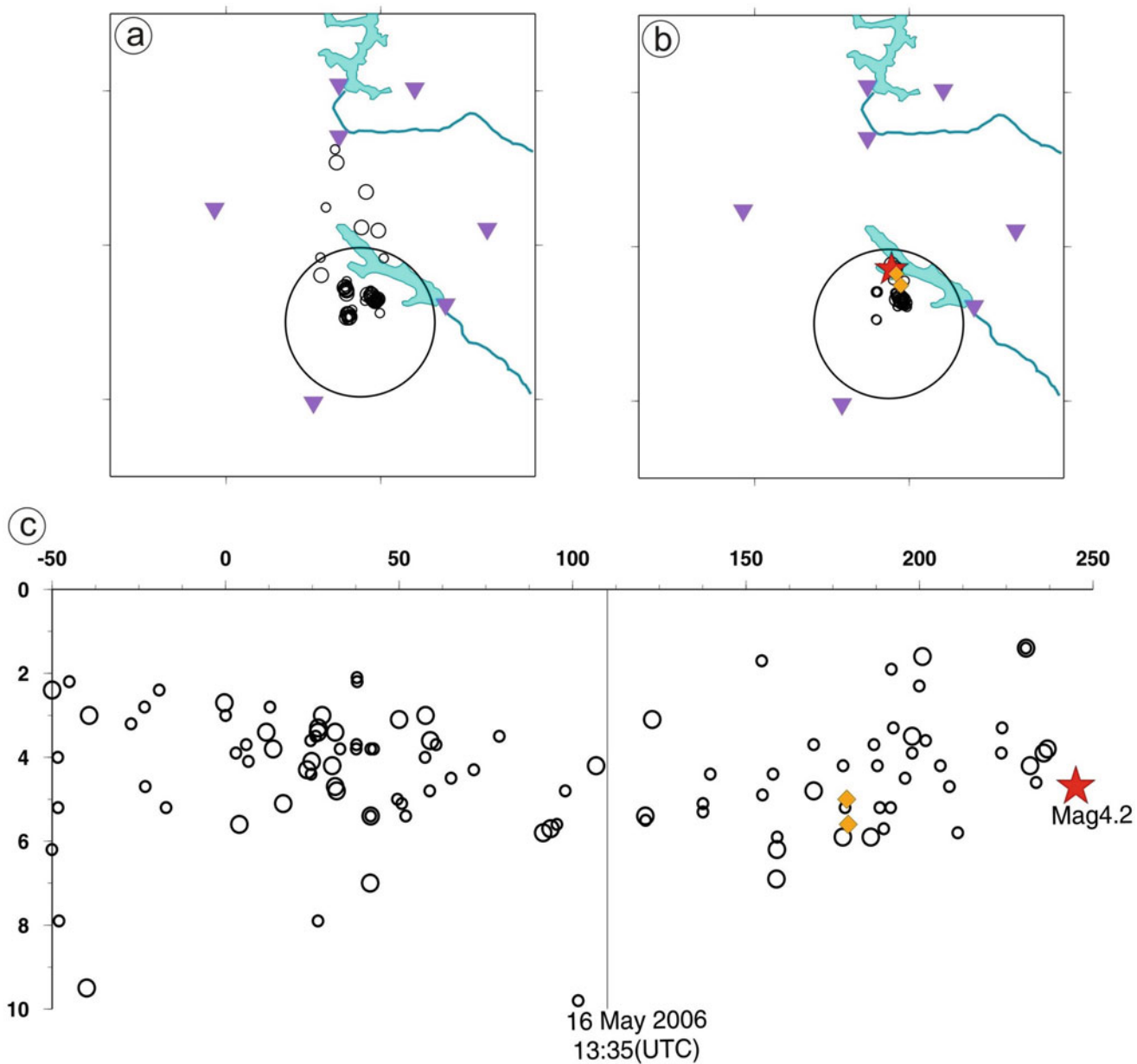
Near-Source Studies at Koyna

The correspondence between the water level variations in Koyna and Warna reservoirs and observed RTS is well established. However, in the absence of the knowledge of physical properties of rocks and fluids in the fault zone, the triggering mechanism is not comprehended. This necessitates near-source studies of earthquakes in the Koyna region. The fact that the RTS is confined to a small area of 30×20 km, most of the earthquakes occur within a depth of 2–9 km; there is no other source of earthquakes within 100 km of Koyna Dam, and the region is easily accessible for all kind of field observations, and temperatures are estimated to be $\sim 130^\circ\text{C}$ at a depth of 6 km, making Koyna an ideal site for setting up of a deep borehole laboratory. The project of setting up of a borehole laboratory for near-field study of earthquakes was discussed in an International Continental Drilling Program (ICDP) sponsored discussion meeting during 21–26 March 2011, where global experts participated and supported the proposed program and suggested certain field work and observations to be carried out before putting the borehole (Gupta and Nayak 2011). In the following 3 years, the recommended investigations were carried out. These included (1) drilling of nine boreholes into the basalt column and penetrating the granitic basement to a depth of 300–500 m, (2) airborne gravity gradient and magnetic surveys covering the entire RTS area, (3) magnetotelluric surveys to map the subsurface and estimate the thickness of basalt cover, (4) LiDAR surveys for getting a bare-earth model and getting a high-resolution topography, (5) instrumenting of six boreholes with three component seismometers for better estimate of the focal parameters of the earthquakes, and (6) heat flow measurements to estimate temperatures at depths of 6 km. The work carried out during 2011–2014 was reported in the 2nd ICDP workshop held during 16–18 May 2014 at Koyna and Karad (Gupta et al. 2014). The results were also presented elsewhere (Gupta et al.



Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Fig. 5 Seismic activity for the period 50 h before the start of nucleation till the occurrence of $M \geq 4$ earthquakes in Koyna region during August 2005–April 2006. Foreshocks clustering before the mainshocks in a region of 10 km radius (right). Temporal distribution of events within the circle with the depth. $\circ$, $\circ$, $\diamond$, $\star$

and $\star$ are events in M 1.0–1.9, 2.0–2.9, and 3.0–3.9 ranges and main earthquake, respectively (left). The onset of identified nucleation period is given "0" time mark. Nucleation period lasted for 110, 160, 150, and 380 h for (a) 30 August, (b) 13 November, and (c) 26 December 2005 and (d) 17 April 2006 earthquakes, respectively (Table 2)



Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Fig. 6 Symbols as in Fig. 5. (a) Events that occurred during 107 h (11–16 May) preceding the identification of nucleation. (c) Temporal-depth plot of the events within the 10-km-

radius circle. Nucleation was inferred to have started 107 h before, where 0 h is put on the time axis. (b) Seismic activity from 13:35 UTC on 16 May till the occurrence of the M 4.2 earthquake on 21 May 2006, on the right side of (c)

2016; Gupta 2017, 2018). Many of the 2nd ICDP workshop participants had participated in the 1st ICDP workshop as well. Satisfied with the progress made, a go ahead was given to put a 3-km-deep Pilot Borehole.

Pilot Borehole Location

Five possible locations were examined for suitability of placing the Pilot Borehole (Gupta et al. 2017). Finally, based on

Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Table 3 Forecast of 21 May 2006 M 4.2 earthquake in Koyna on 16 May 2006

	Forecast parameters	Occurrence
Epicenter	Within 10 km radius of 17.1° N, 73.8° E	17.171° N, 73.777° E
Magnitude	~ 4	4.2
Time	Within 15 days of 16 May 2006, i.e., until 31 May 2006	20:29:01 UTC On 21 May 2006
Focal depth	Less than 8 km	4.7 km

vicinity of site D (Fig. 7) to the Donachiwada fault, which hosted the 10 December 1967 M 6.3 earthquakes and many other events of $M > 4$, logistic convenience, and the trend of earthquakes for the period 2005–2015, it was chosen to place the Pilot Borehole. Drilling of the Pilot Borehole started on 20 December 2016, and the borehole was completed on 11 June 2017. Granitic basement was reached at a depth of 1247 m. No sediments were found below the basalt. Zones of huge fluid losses were encountered at several depths. An online gas analysis facility has been set up. Measurements of *in situ* stresses were made at depths of 1600 m and deeper. It is found that 10^{-16} to 10^{-14} permeability of the granites would not produce significant pore fluid pressure changes to trigger earthquakes. However, core samples of granites at 1522 m depth provided evidence of fractured zones that would contribute to water channelization and triggering of earthquakes (Hazarika et al. 2017). Measurement of elastic properties of granitoids from Koyna region showed low and variable rock strength compared with granitic rocks of aseismic areas (Goswami et al. 2017). The 3rd ICDP post-operation workshop was held during 14 to 16 October 2017 to

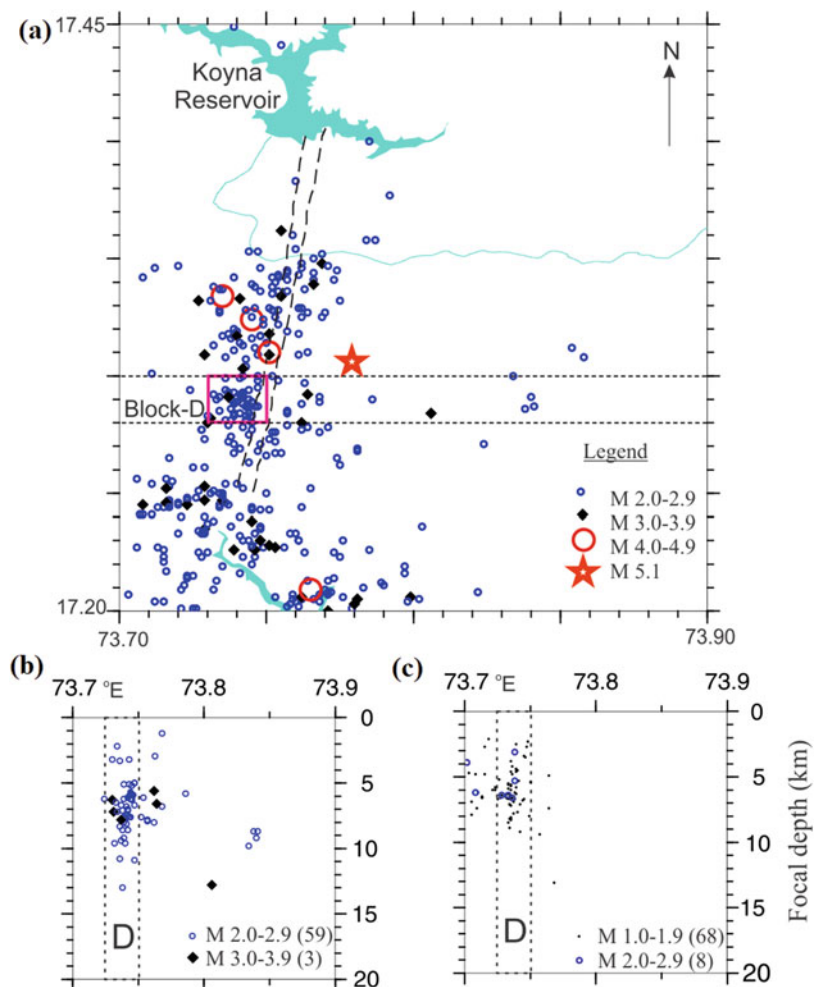
review the work carried out to develop plans for the Main Borehole of ~ 6 km depth, its instrumentation, and future course of action (Podugu et al. 2018). In a recent study, detailed geophysical logging revealed seven anomalous zones below 2100 m depth (Goswami et al. 2019), characterized by shear wave velocity anomaly of up to 25%. All the data acquired are being analyzed for finalizing the specifications of the proposed ~ 6 km deep Main Borehole.

Summary

In this article we provide a thumbnail review of studies of global occurrence of artificial water reservoir-triggered earthquakes, with a special emphasis on Koyna, India. Considering the small changes in the stress regime caused by filling of the deepest reservoirs compared to the stress drop of the associated earthquakes, it is appropriate to replace “reservoir-induced seismicity” by “reservoir-triggered seismicity.”

Koyna continues to be the most significant sight of RTS globally. The latest M 4 earthquake occurred on 20 June 2019.

Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India, Fig. 7 (a) Earthquakes of $M \geq 2.0$ in the vicinity of block D during August 2005–December 2015. A 2-km swath is indicated by dotted lines. The surface expression of the Donachiwada fault is indicated by dashed lines. (b) Depth section of hypocenters for the period 2005–2015 for the 2-km-wide swath. (c) Earthquakes in the 2-km swath for the period 2016–2017



At Koyna, earthquakes of M 4–5 are often preceded by well-defined clusters of foreshocks of $M \leq 3$, referred as nucleation that is found to last typically 100–400 h. Identification of nucleation in real time has led to successful short-term forecasts of $M \sim 4$ earthquakes.

The study of RTS provides exceptionally good opportunity to comprehend physics of earthquakes, finding safer sites of creating artificial water reservoirs and bringing us closer to accurate short-term earthquake forecasts.

Setting up of a deep borehole laboratory for near-field study of earthquakes is underway at Koyna. The first phase of setting up of a 3-km-deep Pilot Borehole has been successfully completed. One of the major discoveries has been that it is not the fluid flow through permeability, but through deep fractures discovered at depths of 3 km in the seismic zone that trigger earthquakes.

Cross-References

- [Earthquake Precursors and Prediction](#)
- [Earthquake Prediction, M8 Algorithm](#)
- [Earthquake, Aftershocks](#)
- [Earthquake, Foreshocks](#)
- [Earthquakes, Energy](#)
- [Seismicity, Intraplate](#)
- [Seismological Networks](#)

Bibliography

- Adams RD (1974) The effect of Lake Benmore on local earthquakes. *Eng Geol* 8:155–169
- Allen CR (1978) Evaluation of seismic hazard at the Auburn dam site. California, U.S. Bur. Reclam. Rep., Denver, CO, p 10
- Allen CR (1982) Reservoir-induced earthquakes and engineering policy. *Calif Geol* 35:248–250
- Baecher BG, Keeney RL (1982) Statistical examination of reservoir induced seismicity. *Bull Seismol Soc Am* 72:553–569
- Bell ML, Nur A (1978) Strength changes due to reservoir-induced pore pressure and stresses and application to Lake Oro-Berocal. *J Geophys Res* 83:4469–4483
- Berocal J, Assumpcao M, Antezana R, Dias Neto CM, Ortega R, Franca H, Velose J (1984) Sismicidade do Brasil Instituto Astronomico e Geofisico. Universidade de Sao Paulo, Sao Paulo, p 320
- Berocal J, Fernandes C, Antezana R, Shukowsky R, Barbosa JR, Shayani S, Pereira ES, (1989) Induced seismicity by the Sobradinho reservoir. Bahia (personal communication)
- Biot MA (1941) General theory of three-dimensional consolidation. *J Appl Phys* 12:155–164
- Bozovic A (1974) Review and appraisal of case histories related to seismic effects of reservoir impounding. *Eng Geol* 8:9–27
- Bufe CG, Lester FW, Lahr KM, Lahr JC, Seekins LC, Hanks TC (1976) Oroville earthquakes: normal faulting in Sierra Nevada foothills. *Science* 192:72–74
- Caloi P (1970) How nature reacts on human intervention—responsibilities of those who cause and who interpret such reaction. *Ann Geofis (Rome)* 23:283–305
- Carder DS (1945) Seismic investigations in the Boulder Dam area, 1940 – 1944, and the influence of reservoir loading on earthquake activity. *Bull Seismol Soc Am* 35:175–192
- Chen B, Li S, Yin Z (1996) On the characteristics and prediction of induced earthquakes of the Geheyan reservoir. Abstract Volume IASPEI Reg. Assembly in Asia, Tangshan, Aug 1–3, 1996
- Chen Q, Nur A (1992) Pore fluid pressure effects in an isotropic rocks: mechanisms of induced seismicity and weak faults. *Pure Appl Geophys* 139:463–480
- Chander R, Kalpna R (1997) On categorizing induced and natural tectonic earthquakes near new reservoirs. *Eng Geol* 46(2):81–92
- Chung WY, Liu C (1992) The reservoir-associated earthquakes of April 1983 in Western Thailand: source modelling and implications for induced seismicity. *PAGEOPH* 138(1):17–41
- Do Nascimento AF, Cowie PA, Lunn RJ, Pearce RG (2004) Spatio-temporal evolution of induced seismicity at Acu reservoir, NE Brazil. *Geophys J Int* 158:1041–1052
- Dodge DA, Beroza GC, Ellsworth WL (1996) Detailed observations of California foreshock sequences: implications for the earthquake initiation process. *J Geophys Res* 101:22371–22392
- Djadjkov PG (1997) Induced seismicity at the Lake Baikal; principal role of load rate, Abstract Volume IASPEI General Assembly, Thessaloniki, Aug 18–28, 1997
- Drzewiecki J, Piernikarczyk A (2017) The forecast of mining-induced seismicity and the consequent risk of damage to the excavation in the area of seismic event. *J Sustain Min* 16:1–7
- Evans MD (1966) Man made earthquakes in Denver. *Geotimes* 10:11–17
- Galanopoulos AG (1967) The large conjugate fault system and the associated earthquake activity in Greece. *Ann Geol Pays Hellen* 8:119–134
- Goswami D, Akkiraju VV, Singh SK, Shalivahan, Roy S (2017) Rock strength and elastic properties of basement granitoids from Koyna region, Deccan Volcanic Province, India. *J Geol Soc India* 90:783–787
- Goswami D, Roy S, Akkiraju VV (2019) Delineation of damage zones from 3 km downhole geophysical logs in the Koyna Seismogenic Zone, western India. *J Geophys Res Solid Earth*. <https://doi.org/10.1029/2018JB017257>
- Gough DI, Gough WI (1970a) Stress and deflection in the lithosphere near Lake Kariba, 1. *Geophys J* 21:65–78
- Gough DI, Gough WI (1970b) Load induced earthquakes at Kariba. 2. *Geophys J R Astron Soc* 21:79–101
- Guang YH (1995) Seismicity induced by cascade reservoirs in Dahan, Yantan Hydroelectric Power Stations. In: Proceedings international symposium on reservoir-induced seismicity. State Seismological Bureau, Beijing, pp 157–163
- Gupta HK, Rastogi BK, Narain H (1972a) Common features of the reservoir associated seismic activities. *Bull Seismol Soc Am* 62:481–492
- Gupta HK, Rastogi BK, Narain H (1972b) Some discriminatory characteristics of earthquakes near the Kariba, Kremasta and Koyna artificial lakes. *Bull Seismol Soc Am* 62:493–507
- Gupta HK, Rastogi BK (1976) Dams and earthquakes. Elsevier, Amsterdam, p 229
- Gupta HK (1983) Induced seismicity hazard mitigation through water level manipulation at Koyna, India: a suggestion. *Bull Seismol Soc Am* 73:679–682
- Gupta HK (1992) Are RIS events of $M > 5$ preceded by a couple of foreshocks of $M > 4$? *Bull Seismol Soc Am* 82(1):517–520
- Gupta HK, Johnston AC (1998) NOTE on “Chapman conference on stable continental region (SCR) earthquakes”. *J Geol Soc India* 52:115–117
- Gupta HK (2001) Short-term earthquake forecasting may be feasible at Koyna, India. *Tectonophysics* 338:353–357
- Gupta HK, Mandal P, Rastogi BK (2002) How long will triggered earthquakes at Koyna, India continue? *Curr Sci* 82:202–210

- Gupta HK (2002) A review of recent studies of triggered earthquakes by artificial water reservoirs with special emphasis on earthquakes in Koyna, India. *Earth Sci Rev* 58:279–310
- Gupta HK, Shashidhar D, Periera M, Mandal P, Rao NP, Kousalya M, Satyanarayana HVS, Dimri VP (2007) Earthquake forecast appears feasible at Koyna, India. *Curr Sci* 93:843–848
- Gupta HK (2011) Artificial water reservoir triggered earthquakes. *Encyclopedia of solid earth geophysics*. Springer+ Business media 1:15–24
- Gupta HK, Nayak S (2011) Deep scientific drilling to study reservoir triggered earthquakes in Koyna, western India. *Sci Drill* 12:53–54
- Gupta HK, Rao NP, Roy S, Arora K, Tiwari VM, Patro BPK, Satyanarayana HVS, Shashidhar D, Mallika K, Akkiraju VV, Goswami D, Vyas D, Ravi G, Srinivas KNSSS, Srihari M, Mishra S, Dubey CP, Raju CV, Borah U, Reddy KC, Babu N, Rohilla S, Dhar U, Sen M, Rao YJB (2014) Investigations related to scientific deep drilling to study reservoir-triggered earthquakes at Koyna, India. *Int J Earth Sci* 104:1511–1522
- Gupta HK, Arora K, Rao NP, Roy S, Tiwari VM, Patro PK, Satyanarayana HVS, Shashidhar D, Mahato CR, Srinivas KNSSS, Srihari M, Satyavani N, Srinu Y, Gopinadh D, Raza H, Jana M, Akkiraju VV, Goswami D, Vyas D, Dubey CP, Raju DCV, Borah U, Raju K, Reddy KC, Babu N, Bansal BK, Nayak S (2016) Investigations of continued reservoir triggered seismicity at Koyna, India. *J Geol Soc London, Special Publications* 445:151–188
- Gupta HK, Shashidhar D, Mahato CR, Satyanarayana HVS, Mallika K, Rao NP, Maity BS, Navitha K (2017) Location of the pilot borehole for investigations of reservoir triggered seismicity at Koyna, India. *Gondwana Res* 42:133–139
- Gupta HK (2017) Koyna, India, an ideal site for near field earthquake observations. *Geol Soc India* 90(6):645–652
- Gupta HK (2018) Review: reservoir triggered seismicity (RTS) at Koyna, India over the past 50 years. *Bull Seismol Soc Am* 108(5B):2907–2918
- Hagiwara T, Ohtake M (1972) Seismic activity associated with the failing of the reservoir behind Kurobe Dam, Japan, 1963–1970. *Tectonophysics* 15:241–254
- Hazarika P, Yadav A, Roy S (2017) Influence of permeability in modeling of reservoir triggered seismicity in Koyna region, western India. *J Geol Soc India* 90:728–732
- Helmstetter A, Sornette D, Grasso JR (2003) Mainshocks are aftershocks of conditional foreshocks: how do foreshock statistical properties emerge from aftershock laws? *J Geophys Res* 108B:2046
- Hudyma MR, Potvin Y (2005) Mining-induced seismicity in underground mechanised hard rock mines-results of a world-wide survey. In: Seminar on advanced geomechanics in mines, 3rd August 2005. Australian Centre for Geomechanics, Perth, p 47
- Ishikawa M, Oike K (1982) On reservoir-induced earthquakes in China. *Zisin* 35(2):171–181
- Kaiser J (1953) Erkenntnisse und Folgerungen aus der Messung von Gerauschen bei Zugbeanspruchung von metallischen Werkstoffen. *Archiv für das Eisenhüttenwesen* 24:43–45. (in German)
- Kalpna, Chander R (2000) Green's function based stress diffusion solutions in the porous elastic half space for time varying finite reservoir loads. *Phys Earth Planet Inter* 120:93–101
- Kangi A, Heidari N (2008) Reservoir-induced seismicity in Karun III dam (Southwestern Iran). *J Seismol* 12:519–527
- McGarr A, Simpson D (1997) Keynote lecture: a broad look at induced and triggered seismicity "Rockbursts and seismicity in mines.". In: Gibowicz SJ, Lasocki S (eds) Proceedings of 4th international symposium on rockbursts and seismicity in mines Poland, 11–14 Aug 1997. A. A. Balkema, Rotterdam, pp 385–396
- McGarr A, Simpson D, Seeber L (2002) Case histories of induced and triggered seismicity. *International handbook of earthquake and engineering seismology*. International Geophysics 81A:647–661
- Mogi K (1963) Some discussions on aftershocks, foreshocks and earthquake swarms – the fracture of a semi-infinite body caused by an inner stress origin and its relation to earthquake phenomena. *Bull Earthq Res Inst* 41:615–658
- Muco B (1991) The swarm of Nikaj-Merturi, Albania. *Bull Seismol Soc Am* 81:1015–1021
- Ohnaka M (1992) Earthquake source nucleation: a physical model for short-term precursors. *Tectonophysics* 211:249–278
- Oike K, Ishikawa Y (1983) Induced earthquakes associated with large reservoirs in China. *Chin Geophys II*(2):383–403
- Packer DR, Cluff LS, Knuepfer PL, Withers RJ (1979) A study of reservoir induced seismicity. Woodward-Clyde Consultants, U.S.A. U.S. Geol. Surv. Contract 14-08-0001-16809 (unpublished report)
- Pavlenov VA, Sherman SI (1996) Premises of induced seismicity on the reservoirs of the Angare River, Abstract Volume IASPEI Reg. Assembly in Asia, Tangshan, China, August 1–3, 1996
- Pavlou K (2019) Relationship between observer seismicity and water level fluctuations in Polyphyto dam area (North Greece). *J Geogr Environ Earth Sci Int* 21(2):1–10
- Plotnikova LM, Makhmudova VI, Sigalova OB (1992) Seismicity associated with the Charvak reservoir, Uzbekistan. *Pure Appl Geophys* 139:607–608
- Podugu N, Yadav A, Mallika K, Goswami D, Anuradha MU (2018) Report on ICDP post-operations on international workshop on "Scientific Deep Drilling at Koyna, India". *J Geol Soc India* 91(1):120–124
- Rastogi BK, Chadha RK, Raju IP (1986) Seismicity near Bhatsa reservoir, Maharashtra, India. *J Phys Earth Planet Inter* 44:179–199
- Rastogi BK, Chadha RK, Sarma CSP, Mandal P, Satyanarayana HVS, Raju IP, Kumar N, Satyamurthy C, Nageswara Rao A (1997) Seismicity at Warna reservoir (near Koyna) through 1995. *BSSA* 87(6):1484–1494
- Reyners M (1988) Reservoir induced seismicity at Lake Pukaki, New Zealand. *Geophys J* 93:127–135
- Rice JR, Cleary MP (1976) Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. *Rev Geophys Space Phys* 14(2):227–242
- Roeloffs EA (1988) Fault stability changes induced beneath a reservoir with cyclic variations in water level. *J Geophys Res* 93(B3):2107–2124
- Rothe JP (1970) The seismic artificials (man-made earthquakes). *Tectonophysics* 9:215–238
- Rothe JP (1973) A geophysics report. In: Ackermann WC, White GF, Worthington EB (eds) Man-made lakes: their problems and environmental effects. Geophysical Monograph. American Geophysical Union, Washington, D.C., pp 441–454
- Shen C, Chang C, Chen H, Li T, Hueng L, Wang T, Yang C, Lo H (1974) Earthquakes induced by reservoir impounding and their effect on the Hsingfengkiang Dam. *Sci Sinica* 17(2):239–272
- Simpson DW (1976) Seismicity changes associated with reservoir loading. *Eng Geol* 10:123–150
- Simpson DW, Leith WS (1988) Induced seismicity at Toktogul reservoir, Soviet Central Asia. U.S. geological survey no. 14-08- 0001-G1168, 32
- Simpson DW, Narasimhan TN (1990) Inhomogeneities in rock properties and their influence on reservoir induced seismicity. *Gerlands Beitrage Zur Geophysik* 99:205–219
- Simpson DW, Negmatullaev SK (1981) Induced seismicity at Nurek reservoir, Tadzhikistan, USSR. *Bull Seismol Soc Am* 71(5):1561–1586
- Snow DT (1972) Geodynamics of seismic reservoirs. In: Proceedings of symposium on percolation through fissured rocks, Deutsche Gessellschaft für Erd-und Grundbau, Stuttgart, T2-J:1–19
- Soboleva OV, Mamadaliev UA (1976) The influence of the Nurek reservoir on local earthquake activity. *Eng Geol* 10:293–305
- Stein S, Wiens DA, Fujita K (1982) The 1966 Kremasta reservoir earthquake sequence. *Earth Planet Sci Lett* 59:49–60
- Stuart-Alexander DE, Mark RK (1976) Impoundment induced seismicity associated with large reservoirs. U.S. geological survey open-file report 76–770

- Talwani P (1976) Earthquakes associated with Clark Hill reservoir, South Carolina – a case of induced seismicity paper presented at the 1st international symposium on induced seismicity. *Eng Geol* 10:239–253
- Talwani P, Acree S (1984/1985) Pore pressure diffusion and the mechanism of reservoir-induced seismicity. *PAGEOPH* 122:947–965
- Talwani P, Cobb JS, Schaeffer MF (1999) In situ measurements of hydraulic properties of a shear zone in northwestern South Carolina. *J Geophys Res* 104(B7):14993–15003
- Talwani P, Chen L, Gahalaut K (2007) Seismogenic permeability, k_s . *J Geophys Res* 112:B07309. <https://doi.org/10.1029/2006JB004665>, 1–18
- Topozada TR (1982) UNDP/Tokten report on Aswan earthquakes
- Trieu CD, Trong CD, Dung LV, Tuan TA, Van DQ, Long HV (2014) Triggered earthquake study in Tranh River no.2 (Vietnam) hydro-power reservoir. *J Geol Soc India* 84:319–325
- Tung NT (1996) The induced seismicity at Hoa Binh Reservoir region, Abstract Volume. IASPEI Reg. Assembly in Asia, Tangshan, Aug 1–3, 1996
- Veloso JAV, Assumpcao M, Concalves ES, Reis JC, Duarte VM, da Motta CBG (1987) Registro de sismicidade induzida em reservatorios da CEMIG e FURNAS. *An 50 Congr Bras Geol Eng* 1:135–146
- Withers RJ, Nyland E (1976) Theory for the rapid solution of ground subsidence near reservoirs on layered and porous media. *Eng Geol* 10:169–185
- Yoshikawa S, Mogi K (1981) A new method for estimating the crustal stress from cored rock samples; laboratory study in the case of uniaxial compression. *Tectonophysics* 74:323–339
- Zoback MD, Hickman S (1982) Physical mechanisms controlling induced seismicity at Monticello reservoir, South Carolina. *J Geophys Res* 87:6959–6974

Biogeophysics

Lee Slater¹ and Estella Atekwana²

¹Department of Earth and Environmental Sciences, Rutgers University Newark, Newark, NJ, USA

²College of Earth, Ocean, and Environment, University of Delaware, Newark, DE, USA

Definition

Biogeophysics Subdiscipline of exploration geophysics focusing on the geophysical signatures resulting from microbial communities and their interactions with geologic media.

Introduction

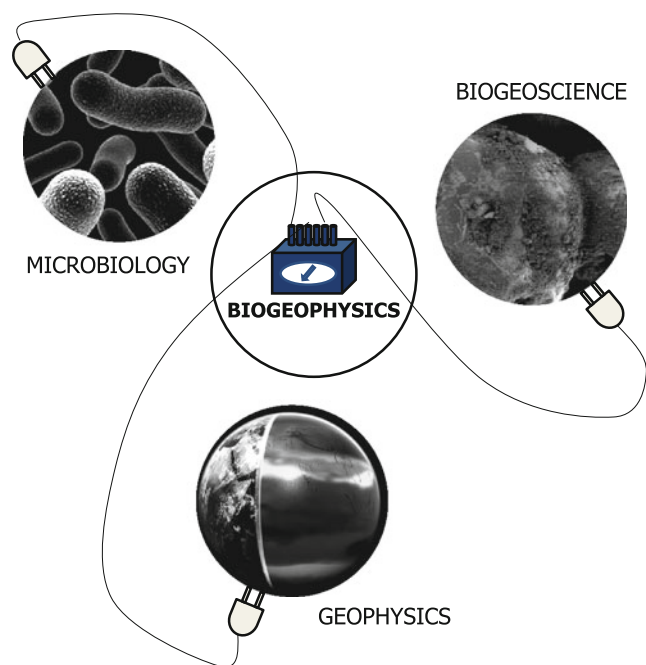
Geophysical imaging techniques have the potential to measure not just the subsurface physical and chemical properties, as geophysics is conventionally used, but also microbes, microbial processes, and microbe-mineral interactions. “Biogeophysics” is defined here as a rapidly evolving discipline of exploration geophysics concerned with the geophysical signatures of microbial communities and their interactions with geologic media that combines the fields of microbiology, biogeoscience, and geophysics (Atekwana and Slater 2009) (Fig. 1). Within this context, biogeophysics examines the links between dynamic subsurface microbial processes, microbial-induced alterations to geologic materials, and geophysical signatures. We note that the term biogeophysics is also used in other disciplines (a) to describe research into the origins of life and (b) to describe fluxes of energy, water, and momentum between earth surface and the atmosphere. Our use of the term describes how it has been adopted by the exploration geophysics community in recognition of the geophysical signatures of microbial activity.

These geophysical signatures may arise from (1) microbial cells and extracellular structures themselves, (2) growth of microorganisms (production of biomass) and biofilm formation, (3) generation of metabolic by-products and the interactions of these metabolites with the host material, and (4) microbially mediated processes. The geophysical signatures arising from each of these four source mechanisms are described below.

Geophysical Detection of Cells and Biofilms

Microbial cells and biofilms exhibit distinct electrical properties and certain (magnetotactic) bacteria also display unique magnetic characteristics. The membrane potential (the voltage drop across the membrane due to the negatively charged molecules inside cells) of live cells results in an accumulation of mobile electric charge carriers at membrane surfaces. When live cells are placed in time-oscillating electric fields, these charges move on the surface of the membrane, giving rise to high polarizations for cellular suspensions that are readily measured with dielectric spectroscopy. As the mobility of these surface charges is relatively small, this effect is manifest at low frequencies, such that the relative dielectric permittivity of live cell suspensions can be as high as 10^6 (Prodan et al. 2004; Stoy et al. 1982). The outer and inner cell radii, diffusion constants, and membrane potential all have a very distinct effect on broadband dielectric spectroscopy data (Prodan et al. 2008). Polarization enhancement is also observed when microbial cells are present in high concentrations in porous media (Ntarlagiannis et al. 2005). This polarization signature of porous media has been modeled in terms of the growth curve for microbial populations (Revil et al. 2012; Mellage et al. 2018). The mechanism may be enhanced when cells are more preferentially adsorbed onto mineral surfaces (Abdel Aal et al. 2009).

The magnetic properties of soils and rock are altered by a diverse group of prokaryotes that exert a significant control over magnetite formation in two ways that differ mechanistically: biologically induced mineralization (BIM) and



Biogeophysics, Fig. 1 Primary disciplinary themes encompassing biogeophysics

biologically controlled mineralization (BCM) (Jimenez-Lopez et al. 2010). Biologically induced magnetite formation is indistinguishable from magnetite formed by inorganic processes; however, biologically controlled magnetite formation is distinct and produced by magnetotactic bacteria (Jimenez-Lopez et al. 2010). These magnetotactic bacteria biomineralize intracellular, membrane-bound, single-magnetic-domain crystals of the magnetic minerals magnetite and/or greigite called magnetosomes (Bazylnski and Frankel 2004). Magnetotactic bacteria are typically found in a wide variety of aquatic environments (e.g., fresh water lakes) with the highest numbers occurring mostly at oxic-anoxic interfaces. The presence of magnetotactic bacteria in sediments is often used as a proxy for paleoenvironmental/climatic conditions, and the presence of magnetosomes in sediments and other secondary magnetic minerals produced by microbial activity likely impacts the magnetic properties of the subsurface recorded with magnetic geophysical techniques as has been demonstrated at hydrocarbon contaminated sites (Rijal et al. 2010).

Biofilms (an attached state of cell growth, whereby cells are closely packed and firmly attached to each other) further alter the geophysical properties of soils. Low frequency electrical measurements respond to self-limiting microbial growth/cell attachment and biofilm formation in porous media, as well as death and lyses of cells (Davis et al. 2006). Electrical polarization may result from the electrical properties of the biofilm itself or from the modification of grain surfaces due to cell attachment. Biofilms also alter the

acoustic properties of porous media, with biofilm growth in soil columns resulting in spatial complexity of acoustic amplitudes (Davis et al. 2009). Such variations likely arise from a nonuniform distribution of microbial activity or heterogeneity in the biomass distribution and biofilm morphology (e.g., variations in biofilm thickness, roughness, hydration, etc.).

Geophysical Detection of Metabolic By-Products and Mineral Weathering

Microbial metabolism enhances the weathering of minerals through the attachment, growth, and colonization of mineral surfaces by microorganisms (Bennett et al. 1996). Metabolic by-products, including biogenic gases (e.g., CO_2 , H_2S , CH_4 , etc.), organic acids, and biosurfactants all affect electrical properties of porous media. Microbial production of organic acids and biosurfactants adds ions to solution, increasing electrolyte concentration of pore fluids (Cassidy et al. 2001). Organic acids enhance mobility of sparingly soluble metals and also increase the number of reaction sites, thus accelerating mineral weathering (Hiebert and Bennett 1992). Enhanced mineral dissolution catalyzed by increased organic acid concentration can lead to physical changes in grain surface morphology, surface area, surface roughness, and the generation of secondary porosity and increased permeability (McMahon et al. 1992). These changes in porosity affect acoustic wave propagation by altering grain contact coupling and changes in surface area/roughness drive changes in electrical polarization. Biogenic gases reduce bulk electrical conductivity and enhance attenuation of seismic signal amplitudes.

The effects of metabolic by-products on geophysical properties have been recorded in hydrocarbon-contaminated environments (Abdel Aal et al. 2004; Werkema et al. 2003). Such sites are natural bioreactors where excess organic substrates stimulate microbial activity. Enhanced mineral weathering in hydrocarbon-contaminated aquifers (Hiebert and Bennett 1992; McMahon et al. 1995) increases pore fluid conductivity and thereby bulk conductivity sensed with a range of electromagnetic methods (Abdel Aal et al. 2004; Atekwana et al. 2004; Sauck et al. 1998). Elevated bulk conductivity is therefore found where intrinsic bioremediation and enhanced mineral weathering are occurring. These zones of highest bulk electrical conductivity may coincide with the highest percentages of oil-degrading microbial populations, with spatial heterogeneity in the microbial community structure and shifts in the microbial community concomitant with vertical changes in bulk electrical conductivity (Allen et al. 2007). Biogenic methane production by archaea in anaerobic soils can result in extensive free phase gas production that reduces dielectric permittivity and electrical conductivity (Comas and Slater 2007; Parsekian et al. 2010).

Geophysical Detection of Microbially Mediated Redox Processes

Microbial metabolic activity is a critical driver of redox chemistry because microbes derive energy from oxidation-reduction reactions (the transfer of electrons from one reactant to the other). Terminal electron acceptors (TEAs) govern nutrient utilization by microbes during the breakdown of organic carbon (Cozzarelli et al. 1999). Microbial respiration consequently results in reduced conditions and strong redox gradients in the Earth typically develop in the presence of heterotrophic bacteria. Significant changes in Eh and pH result in new mineral stability fields in which some minerals become unstable and are dissolved and mobilized, whereas others may precipitate from solution. As TEAs are consumed, changes in pore fluid chemistry may drive changes in pore fluid conductivity, thereby affecting bulk electrical conductivity.

Strong electrical potentials associated with current sources in the Earth are correlated with redox gradients recorded at sites where microbial degradation of hydrocarbons is occurring (Minsley et al. 2007; Naudet et al. 2003). Such large potentials require geobatteries traditionally invoked to explain very large (~1.0 V) potential gradients due to internal current sources recorded over electronically conductive ore bodies straddling the large redox gradient provided by the water table (Sato and Mooney 1960). Biogeobatteries may occur in conjunction with a strong redox gradient between highly reducing conditions below the water table within a contaminant plume and an oxidized zone above the water table if microbial activity can generate the required electron bridge (Revil et al. 2010; Heenan et al. 2017). Possible mechanisms facilitating electron migration include iron oxides, clays, and conductive biological materials (Revil et al. 2010). Metal-reducing organisms, such as *Shewanella* and *Geobacter*, produce electrically conductive appendages called bacterial nanowires that may facilitate electron transfer to solid phase electron acceptors (Reguera et al. 2005). However, the ability of biofilms to facilitate electron transport over the scale of the groundwater interface is unknown although new evidence suggests that such electron transfer can take place at mm scales (Nielsen et al. 2010). Nonetheless, more studies are needed to confirm this finding.

Geophysical Detection of Microbe-Mineral Transformations

Biogeochemical processes result in mineral precipitation that alters the physicochemical properties of the grain surfaces. When microbial-induced precipitation is extensive,

the volumetric physical properties of porous media are modified. In anaerobic environments, iron-reducing bacteria and sulfate-reducing bacteria can use Fe(III) and sulfate, respectively, as terminal electron acceptors. Ferrous iron produced by iron-reducing bacteria promotes the precipitation of electrically conductive secondary minerals such as siderite (FeCO_3), magnetite (Fe_3O_4), and goethite (FeOOH) (Fredrickson et al. 1998). Magnetic susceptibility is a simple and powerful geophysical parameter for monitoring iron mineral transformations coupled to microbial induced oxidation of hydrocarbons (Atekwana et al. 2014; Lund et al. 2017). Sulfide produced during microbial sulfate reduction can react with iron (II) produced by iron-reducing bacteria to precipitate iron sulfide minerals. Strong seismic and electrical signatures are generated as a result of microbe-induced ZnS and FeS precipitation (Williams et al. 2005). Decreases in seismic wave amplitude result from the development of differential elastic moduli associated with accumulation of metal sulfide-encrusted microbes within pores (Williams et al. 2005). Electrical signals result from the formation, movement, and dissolution of electronically conductive biominerals that profoundly enhance the polarization of a porous medium. These electrical signals may be diagnostic of both the concentration and the distribution of the biominerals throughout a porous medium (Slater et al. 2007).

Geophysical signals also result when microbial processes involve the precipitation of semimetallic or nonmetallic minerals, e.g., metabolically induced calcite precipitation by bacterial hydrolysis of urea (Ferris et al. 1995). Relative to metallic minerals, smaller electrical signals result from changes in pore volume/pore tortuosity and/or surface area/surface roughness driven by precipitation of nonmetallic minerals (Wu et al. 2010). Calcite precipitation induced by bacteria has been shown to form cements in porous media and affect subsurface fluid flow (Ferris et al. 1995). Such cements profoundly change the elastic properties of soils and rocks, particularly when ureolysis is stimulated to form calcite cement that acts to stiffen the soil matrix (DeJong et al. 2006). Shear waves are well suited to monitoring changes in the particle soil matrix due to precipitation as shear velocity (V_s) is largely unaffected by pore fluid composition and directly dependent on void ratio, coordination number (average number of surrounding particles a given particle is in contact with), and confining stress (DeJong et al. 2010). Large changes in shear wave velocity accompany stiffening result from initial binding of microbes to the soil matrix, suggesting that relatively small volumes of precipitates can generate large geophysical signals. Electrical signals can also be used to monitor soil strengthening by microbial precipitation of calcite (Saneiyan et al. 2019).

Summary

Twenty years ago, it seemed inconceivable to suggest that microbial processes could potentially impact geophysical signatures. However, today it is clear that geophysical signatures result from a range of microbial processes covering the scale of an individual cell to the scale of contaminant plumes in the Earth. A pressing question in biogeophysics remains how to better quantify the geophysical signatures of microbial processes through the development of appropriate modeling frameworks. Success in this venture will require new multidisciplinary research between geophysicists, biogeochemists, and microbiologists.

Cross-References

- [Electrical Properties of Rocks](#)
- [Electrical Resistivity Surveys and Data Interpretation](#)
- [Geomagnetic Field, Measurement Techniques](#)
- [Geomagnetic Field, Theory](#)
- [Geophysical Well Logging](#)
- [Magnetic Anomalies: Interpretation](#)
- [Magnetic Methods, Principles](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Remanent Magnetism](#)
- [Seismic Properties of Rocks](#)
- [Seismic Viscoelastic Attenuation](#)

Bibliography

- Abdel Aal GZA, Atekwana EA, Slater LD (2004) Effects of microbial processes on electrolytic and interfacial electrical properties of unconsolidated sediments. *Geophys Res Lett* 31(12):L12505. <https://doi.org/10.1029/2004gl020030>
- Abdel Aal G, Atekwana E, Radzikowski S, Rossbach S (2009) Effect of bacterial adsorption on low frequency electrical properties of clean quartz sands and iron-oxide coated sands. *Geophys Res Lett* 36, L04403 (pages 1–5). <https://doi.org/10.1029/2008GL036196>
- Allen JP, Atekwana EA, Duris JW, Werkema DD, Rossbach S (2007) The microbial community structure in petroleum-contaminated sediments corresponds to geophysical signatures. *Appl Environ Microbiol* 73(9):2860–2870. <https://doi.org/10.1128/aem.01752-06>
- Atekwana EA, Slater L (2009) Biogeophysics: a new frontier in Earth science research. *Rev Geophys*. <https://doi.org/10.1029/2009RG000285>
- Atekwana EA, Atekwana EA, Werkema DD Jr, Allen J, Smart L, Duris J, Cassidy DP, Sauck WA, Rossbach S (2004) Evidence for microbial enhanced electrical conductivity in hydrocarbon-contaminated sediments. *Geophys Res Lett* 31:L23501. <https://doi.org/10.1029/2004GL021359>
- Atekwana EA, Mewafy FM, Abdel Aal G, Werkema DD, Revil A, Slater LD (2014) High-resolution magnetic susceptibility measurements for investigating magnetic mineral formation during microbial mediated iron reduction. *J Geophys Res Biogeosci* 119:80–94. <https://doi.org/10.1002/2013JG002414>
- Bazylnski DA, Frankel RB (2004) Magnetosome formation in prokaryotes. *Nat Rev Microbiol* 2(3):217–230
- Bennett PC, Hiebert FK, Choi WJ (1996) Microbial colonization and weathering of silicates in a petroleum-contaminated groundwater. *Chem Geol* 132(1–4):45–53
- Cassidy DP, Werkema DD, Sauck WA, Atekwana E, Rossbach S, Duris J (2001) The effects of LNAPL biodegradation products on electrical conductivity measurements. *J Environ Eng Geophys* 6(1):47–52
- Comas X, Slater L (2007) Evolution of biogenic gases in peat blocks inferred from noninvasive dielectric permittivity measurements. *Water Resour Res* 43(5):W05424. <https://doi.org/10.1029/2006wr005562>
- Cozzarelli IM, Herman JS, Baedeker MJ, Fischer JM (1999) Geochemical heterogeneity of a gasoline-contaminated aquifer. *J Contam Hydrol* 40(3):261–284
- Davis CA, Atekwana E, Slater LD, Rossbach S, Mormile MR (2006) Microbial growth and biofilm formation in geologic media is detected with complex conductivity measurements. *Geophys Res Lett* 33(18):L18403. <https://doi.org/10.1029/2006gl027312>
- Davis CA, Pyrak-Nolte LJ, Atekwana EA, Werkema DD, Haugen ME (2009) Microbial-induced heterogeneity in the acoustic properties of porous media. *Geophys Res Lett* 36:L21405. <https://doi.org/10.1029/2009gl039569>
- DeJong, JT, Mortensen BM, Martinez BC, Nelson DC (2010) Bio-mediated soil improvement. *Ecol Eng* 36(2):197–210
- DeJong JT, Fritzges MB, Nusslein K (2006) Microbially induced cementation to control sand response to undrained shear. *J Geotech Geoenviron* 132(11):1381–1392. [https://doi.org/10.1061/\(asce\)1090-0241\(2006\)132:11\(1381\)](https://doi.org/10.1061/(asce)1090-0241(2006)132:11(1381))
- Ferris FG, Frattton CM, Gertis JP, Schultzelam S, Lollar BS (1995) Microbial precipitation of a strontium calcite phase at a groundwater discharge zone near rock-creek, British-Columbia, Canada. *Geomicrobiol J* 13(1):57–67
- Fredrickson JK, Zachara JM, Kennedy DW, Dong HL, Onstott TC, Hinman NW, Li SM (1998) Biogenic iron mineralization accompanying the dissimilatory reduction of hydrous ferric oxide by a groundwater bacterium. *Geochim Cosmochim Acta* 62(19–20):3239–3257
- Heenan JW, Ntarlagiannis D, Slater LD, Beaver CL, Rossbach S, Revil A, Atekwana EA, Bekins B (2017) Field-scale observations of a transient geobattery resulting from natural attenuation of a crude oil spill. *J Geophys Res Biogeosci* 122:918–929. <https://doi.org/10.1002/2016JG003596>
- Hiebert FK, Bennett PC (1992) Microbial control of silicate weathering in organic-rich ground water. *Science* 258(5080):278–281
- Jimenez-Lopez C, Romanek CS, Bazylnski DA (2010) Magnetite as a prokaryotic biomarker: a review. *J Geophys Res* 115:G00G03. <https://doi.org/10.1029/2009JG001152>
- Lund AL, Slater LD, Atekwana EA, Ntarlagiannis D, Cozzarelli I, Bekins BA (2017) Evidence of coupled carbon and iron cycling at a hydrocarbon-contaminated site from time lapse magnetic susceptibility. *Environ Sci Technol* 51(19):11244–11249
- McMahon PB, Chapelle FH, Falls WF, Bradley PM (1992) Role of microbial processes in linking sandstone diagenesis with organic-rich clays. *J Sediment Petrol* 62(1):1–10
- McMahon PB, Vroblecky DA, Bradley PM, Chapelle FH, Gullett CD (1995) Evidence for enhanced mineral dissolution in organic acid-rich shallow ground-water. *Ground Water* 33(2):207–216
- Mellage A, Smeaton CM, Furman A, Atekwana EA, Rezanezhad F, Van Cappellen P (2018) Linking spectral induced polarization (SIP) and subsurface microbial processes: Results from sand column incubation experiments. *Environ Sci Technol* 52(4):2081–2090
- Minsley B, Sogade J, Morgan FD (2007) Three-dimensional self-potential inversion for subsurface DNAPL contaminant detection at

- the Savannah river site, South Carolina. *Water Resour Res* 43: W04429. <https://doi.org/10.1029/2005WR003996>.
- Naudet V, Revil A, Bottero J-Y, Bégassat P (2003) Relationship between self-potential (SP) signals and redox conditions in contaminated groundwater. *Geophys Res Lett* 30(21):2091
- Nielsen LP, Risgaard-Petersen N, Fossing H, Christensen PB, Sayama M (2010) Electric currents couple spatially separated biogeochemical processes in marine sediment. *Nature* 463(7284): 1071–1074
- Ntarlagiannis D, Williams KH, Slater L, Hubbard S (2005) The low frequency electrical response to microbially induced sulfide precipitation. *J Geophys Res* 110:G02009
- Parsekian AD, Slater L, Comas X, Glaser PH (2010) Variations in free-phase gases in peat landforms determined by ground-penetrating radar. *J Geophys Res* 115:G02002. <https://doi.org/10.1029/2009JG001086>.
- Prodan C, Mayo F, Claycomb JR, Miller JHJ (2004) Low-frequency, low-field dielectric spectroscopy of living cell suspensions. *J Appl Phys* 95(7):3754–3756
- Prodan E, Prodan C, Miller JH (2008) The dielectric response of spherical live cells in suspension: an analytic solution. *Biophys J* 95(9):4174–4182. <https://doi.org/10.1529/biophysj.108.137042>
- Reguera G, McCarthy KD, Mehta T, Nicoll JS, Tuominen MT, Lovley DR (2005) Extracellular electron transfer via microbial nanowires. *Nature* 435:1098–1101
- Revil A, Mendonça CA, Atekwana EA, Kulesa B, Hubbard SS, Bohlén KJ (2010) Understanding biogeobatteries: Where geophysics meets microbiology. *J Geophys Res* 115. <https://doi.org/10.1029/2009jg001065>
- Revil A, Atekwana E, Zhang C, Jardani A, Smith S (2012) A new model for the spectral induced polarization signature of bacterial growth in porous media. *Water Resour Res* 48:W09545. <https://doi.org/10.1029/2012WR011965>
- Rijal ML, Appel E, Petrovský E, Blaha U (2010) Change of magnetic properties due to fluctuations of hydrocarbon contaminated groundwater in unconsolidated sediments. *Environ Pollut* 158(5): 1756–1762
- Saneiyani S, Ntarlagiannis D, Ohan J, Lee J, Colwell F, Burns S (2019) Induced polarization as a monitoring tool for in-situ microbial induced carbonate precipitation (MICP) processes. *Ecol Eng* 127:36–47
- Sato M, Mooney HM (1960) The electrochemical mechanism of sulfide self-potentials. *Geophysics* 25(1):226–249
- Sauk WA, Atekwana E, Nash MS (1998) High conductivities associated with an LNAPL plume imaged by integrated geophysical techniques. *J Environ Eng Geophys* 2(3):203–212
- Slater L, Ntarlagiannis D, Personna YR, Hubbard S (2007) Pore-scale spectral induced polarization signatures associated with FeS biomineral transformations. *Geophys Res Lett* 34(21):L21404. <https://doi.org/10.1029/2007gl031840>
- Stoy RD, Foster KR, Schwan HP (1982) Dielectric properties of mammalian tissues from 0.1 to 100 MHz: a summary of recent data. *Phys Med Biol* 27:501–513
- Werkema DD Jr, Atekwana EA, Endres AL, Sauk WA, Cassidy DP (2003) Investigating the geoelectrical response of hydrocarbon contamination undergoing biodegradation. *Geophys Res Lett* 30(12):1647–1651
- Williams KH, Ntarlagiannis D, Slater LD, Dohnalkova A, Hubbard SS, Banfield JF (2005) Geophysical Imaging of Stimulated Microbial Biomineralization. *Environ Sci Technol* 39(19): 7592–7600
- Wu Y, Hubbard S, Williams KH, Ajo-Franklin J (2010) On the complex conductivity signatures of calcite precipitation. *J Geophys Res*. <https://doi.org/10.1029/2009JG001129>

Body Waves

Mahmoud Mohamed Selim Saleh

Department of Nature and Applied Sciences, Al-Aflaj Community College, AL-Kharj University, Al-Aflaj, Riyadh, Saudi Arabia

Synonyms

Energy waves generated by an earthquake or an artificial explosion

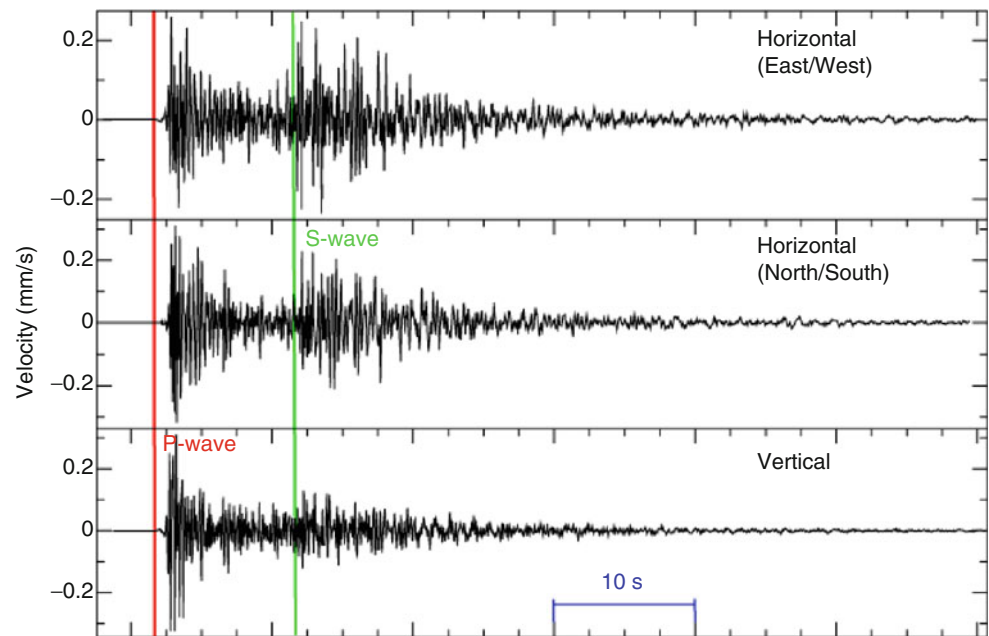
Definition

Waves consist of a disturbance in materials (media) that carries energy and propagates. However, the material that the wave propagates in generally does not move with the wave. The movement of the material is generally confined to small motions, called particle motion, of the material as the wave passes. After the wave has passed, the material usually looks just like it did before the wave, and, is in the same location as before the wave. When stress builds up in the Earth to such a level that rapid slip occurs on a fracture (i.e., an earthquake takes place) or when an explosion or mechanical device is used to initiate a seismic disturbance artificially, a complex field of seismic waves is generated. This wave field propagates much like the waves that travel away from the point at which a stone is thrown into a pond. Waves that travel through the interior of the Earth are called body waves. They follow ray paths bent by the varying density and modulus (stiffness) of the Earth's interior. The density and modulus, in turn, vary according to temperature, composition, and phase. These waves are usually observed at higher frequency.

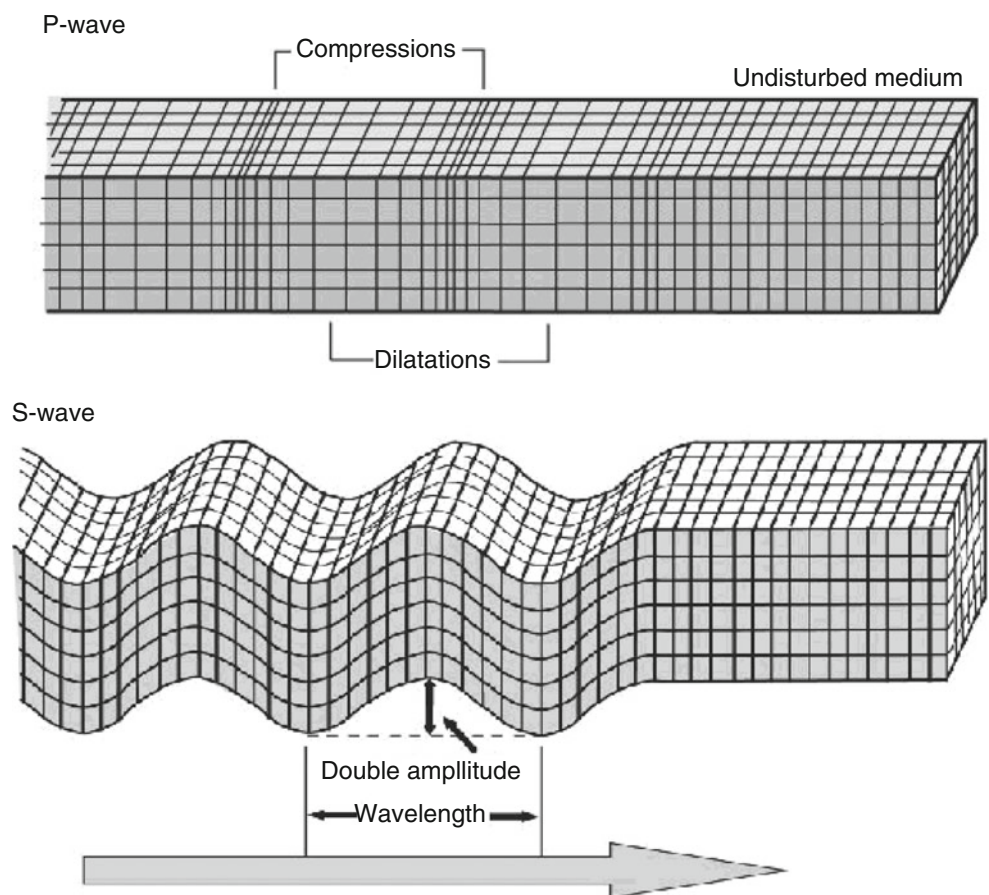
Types of Body Waves

There are two main kinds of body waves: P-waves and S-waves, so-named because they are the primary and secondary waves detected by a seismograph (Fig. 1). The P-wave is always the first wave to arrive at a seismometer, closely followed by its reflection from the surface and S-waves arrive next. P-waves, or compressional waves, are longitudinal waves (wave motion in the direction the wave is traveling). S-waves are transverse waves or shear waves, involving a back-and-forth shearing motion at right angles to the direction the wave is traveling (Fig. 2). Body waves are reflected and transmitted at interfaces where the seismic velocity and/or density changes, obeying Snell's Law. At such an interface,

Body Waves, Fig. 1 P- and S-wave recorded in a three-component seismograph.



Body Waves, Fig. 2 P- and S-waves motion. The arrow shows the direction that the wave is moving



or discontinuity, some of the energy of an incident body wave is reflected as a P-wave, some as an S-wave, some is transmitted as a P-wave and some as an S-wave. The notation for the various seismic ray paths within the Earth are as follows:

When an earthquake occurs, seismic waves are emitted from the focus (hypocenter); there are several paths that it can take through the Earth before emerging again at the surface. These paths (refer to Fig. 3) are symbolized by the letters: p = P-wave arrival from a path that traveled upward from the focus (hypocenter); pP = P-wave arrival from a path that traveled upward from the focus, reflected off the surface of the Earth, then arrived back at the surface; P = P-wave arrival from a path that traveled downward from the focus (hypocenter); PP = P-wave reflected off the surface once; PPP = P-wave reflected off the surface twice; c = a reflection off the outside of the outer surface of the outer core – note that this is the principle cause of multiple arrivals of P- and S-waves right at the epicenter; K = a travel path (refraction) through the outer core; KK = one reflection on the inside outer surface of the outer core; KKK = two reflections off the inside outer surface of the outer core; i = a reflection off the outside of the outer surface of the inner core; I = a travel path (refraction) through the inner core. These letters can be used to indicate the path of a seismic wave through the Earth (Fig. 3). For example PKiKP indicates that the wave traveled downward from the focus, refracted through the outer core, reflected off the surface of the inner core, traveled through the outer core, and then traveled through the mantle to arrive at the surface. SKiKS is the same path, but an S-wave. Because liquids have no resistance to shear and cannot sustain a shear wave, S-waves cannot travel through liquid material. The Earth's outer core is

believed to be liquid because S-waves disappear at the mantle–core boundary, while P-waves do not.

Wave Propagation

Seismic body waves propagate through the Earth's interior. Because of the elastic properties of the Earth's materials (rocks) and the presence of the Earth's surface, four main types of seismic waves propagate within the Earth. Compressional (P) and Shear (S) waves propagate through the Earth's interior.

Waves on a Seismogram

If we look at a seismogram, we expect to see the first wave to arrive to be a P-wave (the fastest), then the S-wave, and finally, the Love and Rayleigh (the slowest) waves. As you might expect, the difference in wave speed has a profound influence on the nature of seismograms. Since the travel time of a wave is equal to the distance the wave has traveled divided by the average speed the wave moved during the transit, we expect that the fastest waves arrive at a seismometer first. The fact that the waves travel at speeds that depend on the material properties (elastic moduli and density) allows us to use seismic-wave observations to investigate the interior structure of the planet. We can look at the travel times, or the travel times and the amplitudes of waves to infer the existence of features within the planet, and this is an active area of seismological research. To understand how we “see” into Earth using vibrations, we must study how waves interact with the rocks that make up Earth. Several types of interaction between waves and the subsurface geology (i.e., the rocks) are commonly observable on seismograms. We will examine the two simplest types of interaction, refraction and reflection.

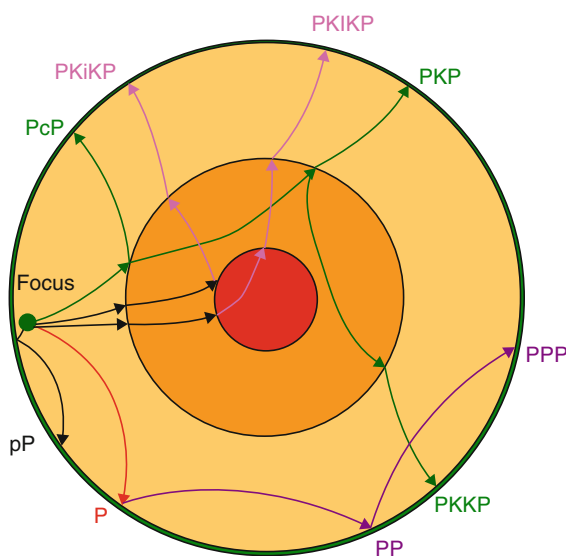
Refraction

As a wave travels through Earth, the path it takes depends on the velocity. Perhaps you recall from high school a principle called Snell's law, which is the mathematical expression that allows us to determine the path a wave takes as it is transmitted from one rock layer into another. Snell's law describes the relationship between the angles and the velocities of the waves. Snell's law equates the ratio of material velocities V_1 and V_2 to the ratio of the sines of incident (θ_1) and refracted (θ_2) angles, as shown in the following equation:

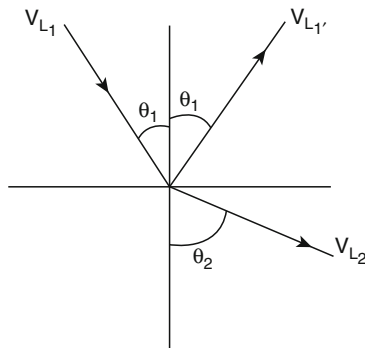
$$\frac{\sin \theta_1}{V_{L_1}} = \frac{\sin \theta_2}{V_{L_2}} \quad (1)$$

where V_{L_1} is the longitudinal wave velocity in material 1 and V_{L_2} is the longitudinal wave velocity in material 2.

Note that in Fig. 4, there is a reflected longitudinal wave ($V_{L'}$) shown. This wave is reflected at the same angle as the incident wave because the two waves are traveling in the same



Body Waves, Fig. 3 P- and S-wave paths in the Earth's interior

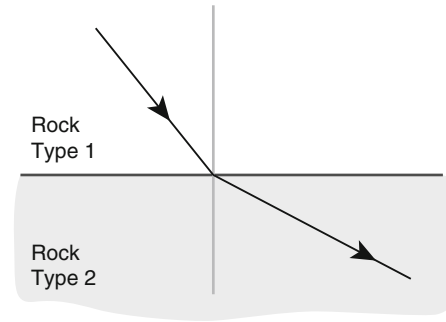


Body Waves, Fig. 4 Reflected ($V_{L1'}$) and refracted (V_{L2}) longitudinal waves

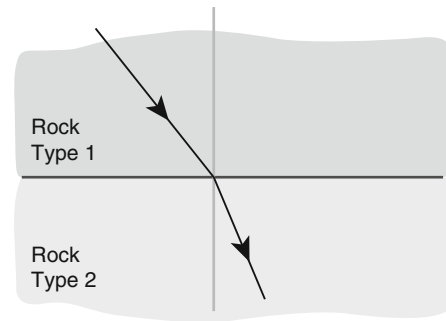
material, and hence have the same velocities. This reflected wave is unimportant in our explanation of Snell's law, but it should be remembered that some of the wave energy is reflected at the interface. When a longitudinal wave moves from a slower to a faster material, there is an incident angle that makes the angle of refraction for the wave 90° . This is known as the first critical angle. The change in direction depends on the ratio of the wave velocities of the two different rocks. When waves reach a boundary between different rock types, part of the energy is transmitted across the boundary. The actual interaction between a seismic wave and a contrast in rock properties is more complicated because an incident P-wave generates transmitted and reflected P- and S-waves and also an incident S-wave generates transmitted and reflected P- and S-waves, so five waves are involved. The transmitted wave travels in a different direction, which depends on the ratio of velocities of the two rock types (Figs. 5 and 6). Part of the energy is also reflected backwards into the region with rock type 1, but we have not shown that on these figures. Refraction has an important effect on waves that travel through Earth. In general, the seismic velocity in Earth increases with depth (there are some important exceptions to this trend) and refraction of waves causes the path followed by body waves to curve upward (Fig. 7).

Reflection

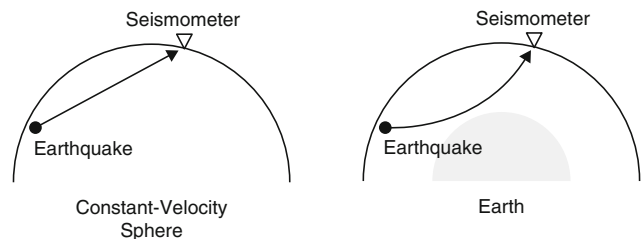
The second wave interaction with variations in rock type is reflection. I am sure you are familiar with reflected sound waves: we call them echoes. And your reflection in a mirror or pool of water is composed of reflected light waves. In seismology, reflections are used to prospect for petroleum and investigate Earth's internal structure. In some instances, reflections from the boundary between the mantle and crust may induce strong shaking that causes damage about 100 km from an earthquake (we call that boundary the "Moho" in honor of Mohorovicic, the scientist who discovered it). A seismic reflection occurs when a wave impinges on a change in rock type (which usually is accompanied by a change in seismic-wave speed). A part of the energy carried



Body Waves, Fig. 5 Velocity in rock type 2 is greater than velocity in rock type 1

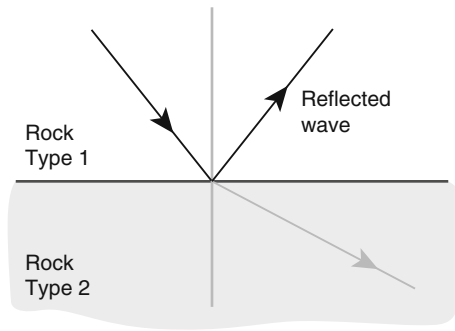


Body Waves, Fig. 6 Velocity in rock type 2 is less than velocity in rock type 1



Body Waves, Fig. 7 The overall increase in seismic-wave speed with depth in the Earth produces an upward curvature to rays that pass through the mantle. A notable exception is caused by the decrease in velocity from the mantle to the core. This speed decrease bends waves backwards and creates a "P-wave shadow zone" between about 100° and 140° distance ($1^\circ = 111.19$ km)

by the incident wave is transmitted through the material (that is the refracted wave described above), and a part is reflected back into the medium that contained the incident wave (Fig. 8). The amplitude of the reflection depends strongly on the angle that the incident wave makes with the boundary and the contrast in material properties across the boundary. For some angles, all the energy can be returned into the medium containing the incident wave. An incident P-wave generates transmitted and reflected P- and S-waves and so five waves are involved. Likewise, when an S-wave interacts with a boundary in rock properties, it too generates reflected and refracted P- and S-waves. Because major boundaries between different rock types within the Earth are normally



Body Waves, Fig. 8 When a wave encounters a change in material properties (seismic velocities and or density), its energy is split into reflected and refracted waves

approximately parallel to the Earth's surface, S-wave particle motion is commonly in the SV (perpendicular to the ray path and vertical) and SH (perpendicular to the ray path and horizontal) directions. P-waves travel faster than S-waves, so there will be separate wave front representations for the P- and S-waves (Fig. 9). If the physical properties of the material through which the waves are propagating are constant, the wave fronts will be circular (or spherical in three dimensions). If the physical properties vary in the model, the wave fronts will be of more complex shapes. In the transverse or shear wave, the particles oscillate at a right angle or transverse to the direction of propagation. Shear waves require an acoustically solid material for effective propagation, and therefore, are not effectively propagated in materials such as liquids or gases. Shear waves are relatively weak when compared to longitudinal waves. In fact, shear waves are usually generated in materials using some of the energy from longitudinal waves.

Velocity

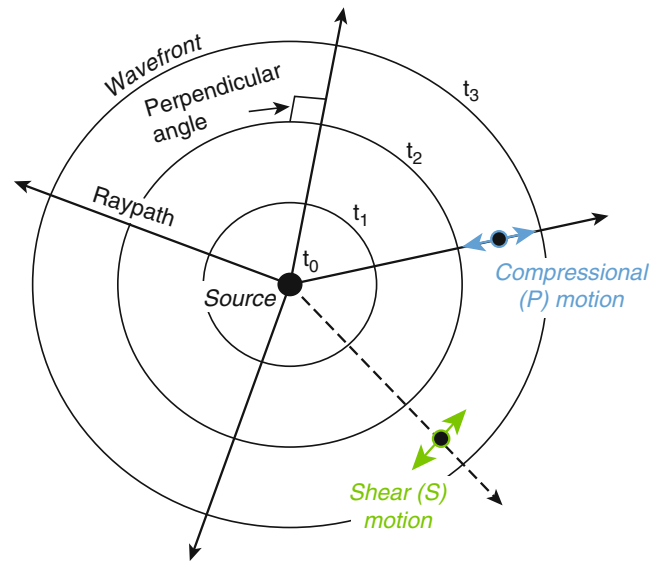
Among the properties of waves propagating in isotropic solid materials are wavelength, frequency, and velocity. The wavelength (λ) is directly proportional to the velocity of the wave (C) and inversely proportional to the frequency (f) of the wave. The phase velocity of a seismic wave can be written as:

$$C = \omega/k, \quad (2)$$

where $\omega = 2\pi f$ is the angular frequency and $k = \lambda/2\pi$ is wave number.

Velocity of P- and S-waves

Seismic waves travel fast, of the order of kilometers per second (km/s). The precise speed that a seismic wave travels depends on several factors, the most important of which is the composition of the rock. We are fortunate that



Body Waves, Fig. 9 Wave fronts and ray paths in seismic-wave propagation

the speed depends on the rock type because it allows us to use observations recorded on seismograms to infer the composition or a range of compositions of the planet. But the process is not always simple, because sometimes different rock types have the same seismic-wave velocity, and other factors also affect the speed, particularly temperature and pressure. Temperature tends to lower the speed of seismic waves and pressure tends to increase the speed. Pressure increases with depth in Earth because the weight of the rocks above increases with increasing depth. Usually, the effect of pressure is larger and in regions of uniform composition, the velocity generally increases with depth, despite the fact that the increase of temperature with depth works to lower the wave velocity. When the different seismic-wave types are described below, I will quote ranges of speed to indicate the range of values we observe in common terrestrial rocks. But you should keep in mind that the specific speed throughout Earth will depend on composition, temperature, and pressure.

P-waves

P-waves are the first waves to arrive on a complete record of ground shaking because they travel the fastest (their name derives from this fact—P is an abbreviation for primary, first wave to arrive). They typically travel at speeds between ~ 1 and ~ 14 km/s. The slower values correspond to a P-wave traveling in water, the higher number represents the P-wave speed near the base of Earth's mantle. The velocity of a wave depends on the elastic properties and density of a material. If we let κ represent the bulk modulus of a material, μ the shear modulus, and ρ the density, then the P-wave velocity, which we represent by C_P , is defined by:

$$C_P = \sqrt{\frac{\lambda + 2\mu}{\rho}} \quad (3)$$

A modulus is a measure of how easy or difficult it is to deform a material. For example, the bulk modulus is a measure of how a material changes volume when pressure is applied and is a characteristic of a material. For example, foam rubber has a lower bulk modulus than steel. P-waves are sound waves, it is just that in seismology we are interested in frequencies that are lower than humans' range of hearing (the speed of sound in air is about 0.3 km/s). The vibration caused by P-waves is a volume change, alternating from compression to expansion in the direction that the wave is traveling. P-waves travel through all types of media – solid, liquid, or gas.

S-waves

Secondary, or S-waves, travel slower than P-waves and are also called “shear” waves because they do not change the volume of the material through which they propagate – they shear it. S-waves are transverse waves because they vibrate the ground in a direction “transverse,” or perpendicular, to the direction that the wave is traveling.

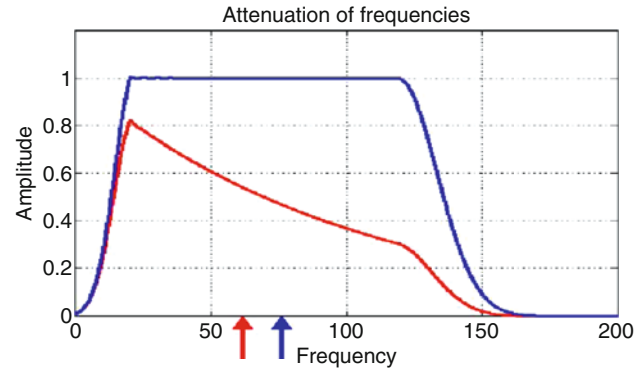
The S-wave speed, call it C_S , depends on the shear modulus and the density

$$C_S = \sqrt{\frac{\mu}{\rho}} \quad (4)$$

Even though they are slower than P-waves, the S-waves move quickly. Typical S-wave propagation speeds are of the order of 1–8 km/s. The lower value corresponds to the wave speed in loose, unconsolidated sediment, the higher value is near the base of Earth's mantle. An important distinguishing characteristic of an S-wave is its inability to propagate through a fluid or a gas because fluids and gases cannot transmit shear stress and S-waves are waves that shear the material. The velocity of seismic waves depends on the elastic properties and density of a material in which the wave is traveling. P-waves are sound waves, it is just that in seismology we are interested in frequencies that are lower than humans' range of hearing (the speed of sound in air is about 0.3 km/s). The vibration caused by P-waves is a volume change, alternating from compression to expansion in the direction that the wave is traveling. In general, earthquakes generate larger shear waves than compressional waves and much of the damage close to an earthquake is the result of strong shaking caused by shear waves.

Attenuation

Near the source of a strong disturbance, such as a large explosion or earthquake, the wave-generated deformation can be large enough to cause permanent deformation, which



Body Waves, Fig. 10 Amplitude spectrum in red is for signal after attenuation

will be visible as cracks, fault offsets, and displacements of the ground after the disturbance has passed. A source of energy creates the initial disturbance and the resulting waves propagate (travel) outward from the disturbance. Because there is finite energy in a confined or short-duration disturbance, the waves generated by such a source will spread out during propagation and become smaller (attenuate) with distance away from the source or with time after the initial source, and thus, will eventually die out (Fig. 10). Knowledge of attenuation can be very useful in seismic data processing, as its removal increases resolution.

Attenuation often serves as a measurement tool that leads to the formation of theories to explain physical or chemical phenomenon that decreases the body wave's intensity. The amplitude change of a decaying plane wave can be expressed as:

$$A = A_0 e^{-\alpha z} \quad (5)$$

In this expression, A_0 is the unattenuated amplitude of the propagating wave at some location. The amplitude A is the reduced amplitude after the wave has traveled a distance z from the source. The quantity α is the attenuation coefficient of the wave traveling in the z -direction. The more common unit of the attenuation value is Decibels. Attenuation is generally proportional to the square of the frequency. Quoted values of attenuation are often given for a single frequency, or an attenuation value averaged over many frequencies may be given. The quoted values of attenuation only give a rough indication of the attenuation and should not be automatically trusted. Generally, a reliable value of attenuation can only be obtained by determining the attenuation experimentally for the particular material being used.

Summary: Results of Regional and Global Studies

The study of seismic body waves permits us to derive important knowledge about the internal constitution of the Earth.

The problems related to propagation of elastic seismic body waves has been discussed in a number of books and in numerous papers (e.g., Bromwich 1898; Sezawa 1927; Haskell 1953; Ewing et al. 1957; Brekhovskikh 1960; Bullen 1963; Bath 1968; Achenbach 1975 and others). Attenuations of seismic waves have been investigated by many authors (e.g., McDonald et al. 1958; Knopoff and Macdonald 1958; Kolsky 1963; White 1965; Kuster and Toksoz 1974; Shoenberger and Levin 1974; Kennett 1983; Hong and Kennett 2003 and others). The asymptotic theory of body-wave propagation in anisotropic media is also well developed (e.g., Crampin 1984; Thomsen 1986; Tsvankin 1997; Cerveny 2001; Chapman 2004).

The effect of initial stresses present in the medium is not considered in the above studies. Recently, Dey and Dutta (1998) have studied the problem of propagation and attenuation of seismic waves, taking the effect of initial stresses present in the medium into account. Biot incremental deformation theory has been used (Biot 1965) in their study. Selim and Ahmed (2006) have studied the velocities of propagation, damping, and attenuations of seismic body waves in compressible and dissipative medium under the effect of both initial and couple stresses.

Body-Wave Modeling

Seismological methods for determining Earth structure are often classified as being either active or passive in nature. Passive methods involve waiting for an earthquake to occur and provide the seismic source for recording. In contrast, controlled-source seismology refers to active methods where the experimenter provides the source by an explosion or a mechanical device such as a hammer, a weight that is dropped, or a vibrator. Active methods can, in turn, be divided into two basic classifications. The first is seismic reflection profiling, which is a clearly defined approach in which the goal is to produce an image of the subsurface in which structures can be seen directly, in the way that an X-ray image reveals features inside an object. Other active seismic methods infer seismic velocities and the presence of discontinuities in velocity and structure (such as faults) using a variety of approaches that analyze the arrival times and sometimes the shape of seismic waves traveling along different paths through the Earth. As a consequence of advances in seismic instrumentation and national programs increasing the number of instruments available, there have been many recent developments in these techniques that do not directly image the Earth. A convenient aspect of the theoretical basis for virtually all active-source techniques is that they are largely independent of scale. Thus, detailed studies to address environmental problems and regional studies to determine deep Earth structure employ the same basic techniques of analysis (Hancock and Skinner 2000). Recently, several efforts have been made to complete our knowledge of the structure of the

Earth's upper mantle by detailed observation of seismic body waves, especially by their amplitude–distance curves.

Cross-References

- ▶ Deep Seismic Reflection and Refraction Profiling
- ▶ Earth's Structure, Global
- ▶ Earthquake Source Theory
- ▶ Earthquakes and Crustal Deformation
- ▶ Earthquakes, Energy
- ▶ Propagation of Elastic Waves: Fundamentals
- ▶ Seismic Phase Nomenclature: The IASPEI Standard
- ▶ Seismic Ray Theory
- ▶ Seismic Signals in Well Water Observations
- ▶ Seismic Tomography
- ▶ Seismic Viscoelastic Attenuation
- ▶ Seismic Wave Propagation in Real Media: Numerical Modeling Approaches
- ▶ Seismic Waves, Scattering
- ▶ Seismic, Velocity, and Density Relationships
- ▶ Seismic, Waveform Modeling and Tomography
- ▶ Seismology, Global Earthquake Model
- ▶ Traveltime Tomography Using Controlled-Source Seismic Data

Bibliography

- Achenbach JD (1975) Wave propagation in elastic solids. North-Holland, Amsterdam, p 425
- Bath MA (1968) Mathematical aspects of seismology. Elsevier, Amsterdam, p 414
- Biot MA (1965) Mechanics of incremental deformation. Wiley, New York, p 181
- Brekhovskikh LM (1960) Waves in layered media. Academic, New York
- Bromwich TJLA (1898) On the influence of gravity on the elastic waves and in particular on the vibration of an elastic globe. *Proc Lond Math Soc* 30:98–120
- Bullen KE (1963) An introduction to the theory of seismology. Cambridge University Press, Cambridge
- Cerveny V (2001) Seismic ray theory. Cambridge University Press, Cambridge
- Chapman C (2004) Fundamentals of seismic wave propagation. Cambridge University Press, Cambridge
- Crampin S (1984) An introduction to wave propagation in anisotropic media. *Geophys J R Astron Soc* 76:17–28
- Dey S, Dutta D (1998) Propagation and attenuation of seismic body waves in initially stressed dissipative medium. *Acta Geophysica Polonica XLVI(3)*:351–365
- Ewing WM, Jardetzky WS, Press F (1957) Elastic waves in layered media. McGraw-Hill, New York, p 380
- Hancock P, Skinner BJ (2000) Controlled-source seismology. In: The Oxford companion to the Earth. Encyclopedia.com
- Haskell NA (1953) The dispersion of surface waves in multilayered media. *Bull Seismol Soc Am* 43:17–34
- Hong TK, Kennett BLN (2003) Scattering attenuation of 2D elastic waves: theory and numerical modeling using a wavelet-based method. *Bull Seismol Soc Am* 93(2):922–938

- Kennett BLN (1983) Seismic wave propagation in stratified media. Cambridge University Press, Cambridge, p 342
- Knopoff L, Macdonald JF (1958) Attenuation of small amplitude stress waves in solid. *Rev Mod Phys* 30:1178
- Kolsky H (1963) Stress waves in solid. Dover, New York
- Kuster GT, Toksoz MN (1974) Velocity and attenuation of seismic waves in two-phase media, part I: theoretical formulations. *Geophysics* 39(5):587–618
- McDonald FJ, Angona FA, Mills RL, Sengbush RL, Van Nostrand RG, White JE (1958) Attenuation of shear and compressional waves in Pierre shale. *Geophysics* 23(3):421–439
- Selim MM, Ahmed MK (2006) Propagation and attenuation of seismic body waves in dissipative medium under initial and couple stresses. *Appl Math Comput* 182:1064–1074
- Sezawa K (1927) Dispersion of elastic waves propagated on the surface of stratified bodies and on curves surfaces. *Bull Earthquake Res Inst (Tokyo)* 3:1–18
- Shoenberger M, Levin FK (1974) Apparent attenuation due to intrabed multiples. *Geophysics* 39(3):278–291
- Thomsen L (1986) Weak elastic anisotropy. *Geophysics* 51:1954–1966
- Tsvankin I (1997) Anisotropic parameters and p-wave velocity for orthorhombic media. *Geophysics* 62:1292–1309
- White JE (1965) Seismic waves, radiation, transformation and attenuation. McGraw-Hill, New York, p 302

Borehole Seismic Networks and Arrays

Marco Bohnhoff and Peter Malin
GFZ German Research Center for Geosciences,
Geomechanics and Scientific Drilling, Potsdam, Germany

Synonyms

Borehole strings and chains; Buried grids; Downhole networks and arrays; Vertical arrays; VSP

Definition

In the broadest sense, *borehole networks and arrays* are assemblies of *geophysical* instruments placed in the subsurface via boreholes. This article focuses on their seismological applications. While common usage often blurs the distinction between a *network* and an *array*, this article explicitly distinguishes between their operational theories and engineering practices, and in the case histories and examples provided. To strictly qualify as an *array*, the assembly of instruments must include matching or similar sensors located at different depths in one well or lateral separated in different wells. The combined natural-frequency responses and spacing of the sensors must also include a range over which they can be used to improve signal detection and analysis via time series processing. Outside of this range a *borehole* assembly acts as a *network* of independent detectors.

Introduction

The most common borehole geophysical instruments measure elastic and permanent displacements, electromagnetic fields, and fluid pressures. These sensors are used to variously record underground accelerations, velocities, tilts, tensor and volumetric strains, electrical conductivities, and pore pressures. While accelerometers, tiltmeters, and stainmeters are common borehole sensors, most borehole installations are currently done with seismometers.

The primary driver behind this application is the significant improvement in seismic signal quality with installation depth (e.g., Prevedel et al. 2015). Boreholes not only increase the distance of seismometers from near-surface environmental and cultural noise, they are also placed below the signal-trapping-and-attenuating properties of near-surface materials. As a result, borehole seismometers can dramatically improve the detection and location of natural and induced microearthquakes and enhance the quality of controlled-source data (e.g., Kwiatak et al. 2019). It has allowed the observation of phenomena such as fault zone guided waves (Ellsworth and Malin 2011). Recently, it has revealed the existence of resonances in fluid-filled fractures (Sicking and Malin 2019).

This entry describes some of the early developments in borehole seismology, its current instrumentation and installation methods, its noise-suppression and signal-detection advantages, and the differences between networks and arrays (e.g., Bohnhoff et al. 2018). It also presents representative examples and case histories that illustrate the key aspects of this method. The article concludes with a brief discussion of the new borehole fiber-optic sensor-cables method for borehole seismology. Such cables are showing promise in the measurement of seismic waves, rock strains, and subsurface temperatures and pore pressures (e.g., Lellouch et al. 2019). As their installation and analysis methods progress, it is likely that these types of sensors will become standard choices for borehole networks and arrays.

Early Developments

The general use of boreholes for making subsurface seismic measurements appears to have begun nearly 100 years ago. McCollum and LaRue (1931) report that “during the past year or more, use has been made of existent wildcat wells for the purpose of making a seismic exploration [of salt dome structures].” By the 1940s, use of a single borehole seismograph to profile the flanks of a salt dome was an established practice (Gardner 1949; Musgrave et al. 1960).

These developments parallel the first test of the borehole electric log by Conrad and Marcel Schlumberger in 1927 (Johnson 1962). The diversity of what became known as well logging tools grew rapidly, as did their use in make

point measurements of near-well rock properties and structures. The development of borehole seismic networks and arrays can be viewed as efforts to penetrate further into the rock mass for the same purposes. From this effort came the finding that placing seismometers even a few tens of meters underground significantly increased the quality of seismic data. These include increases in signal-to-noise ratios, recorded signal frequency bandwidths, and the detection and resolution of remote lateral velocity variations and structures.

Current Instrumentation and Installation

Examples of the type of seismic instruments used to achieve these advantages are shown in Fig. 1. In the case of traditional three-component, mass and spring sensors, these range from greater than ~ 10 – 12 Hz omni directional seismometers in fixed orthogonal mounts, to lower frequency ~ 1 – 10 Hz, gimble mounted, vertical and horizontal components. In both cases, the installed orientation of the components needs to be determined, either by special downhole tools or by calibrations shots or distant earthquakes. For a fixed depth installation, the latter two methods are most cost-effective.

In addition to the borehole sensors and their communication cables, a critical component of their deployment is the method used to couple them to the surrounding rock. This can be by cementing of the entire well, including instruments and cables. The disadvantage is that nothing is retrievable, and the well is permanently blocked. With this type of installation, the use of mechanically robust and electrically isolated sensors is essential: no active components requiring electrical power or corrodible materials, and triple barriers from sensor-compromising borehole fluids and gases.

For some applications an adequate degree of coupling can also be achieved by lowering the sensors on a heavy but flexible backbone – e.g., oil production tube – the spring action of which, when standing on the borehole bottom, pushes the sensors against the wall. This system is retrievable and works well for signals with accelerations significantly less than 1 g. Its disadvantage is the need for heavy surface equipment to handle the weight of the backbone plus sensors and cable. A more robust coupling alternative to a flexible backbone is the use of collapsible bladders or the equivalent of a firehose, the filing of which pushes the sensor against the borehole wall. Neither of these methods are suitable for moving the array along the well for profiling purposes.

For relocatable networks or arrays, passive coupling can be accomplished with bow-spring or magnetic clamps, the continuously coupled sensors pulled into the borehole by a heavy weight or mechanical tractor. The most sophisticated moveable couplers include electrically or hydraulically driven arms, the controlled forces of which guarantee the best connection to the borehole wall.

Noise-Suppression and Signal-Detection Advantages

Whether in the form of a borehole network or array, placing seismic receivers underground significantly improves the quality of seismic signals and the amount of information they contain about seismic events, velocities, and geological structures (Figs. 2 and 3).

Perhaps the most significant improvements are the increases in the signal-to-noise ratio and bandwidth of upward propagating seismic waves. As shown in Fig. 2a, several factors signal-to-noise improvement can be achieved by installing sensors just a few tens of meters of downhole. These improvements are most dramatic for signal frequencies above a few Hz, as these are most affected by attenuation and scattering near the earth's surface. Since these effects apply equally to downward propagating noises generated at the surface, their damping-out with depth adds to the improvement as well.

The combined effects on signal-to-noise and bandwidth can be seen in Fig. 2 by comparing the frequency-dependent improvements at depth with surface values: while it is not common to record 100 Hz seismic signals at ground levels, they are usually well recorded at a depth of 250–500 m.

These signal improvements have significant effects on the detection and location of small seismic events – both natural and fabricated. As shown in Fig. 2b, by lowering the sensors of an established seismic monitoring network tens to hundreds of meters below their current positions, the network's event-detection threshold can be lower by one to three magnitude units. Further improvements relate to more accurate determination of lateral velocity changes, observation of near-vertical structures, and the location of other subsurface sources – e.g., well drilling bits (Fig. 3).

Networks Versus Arrays

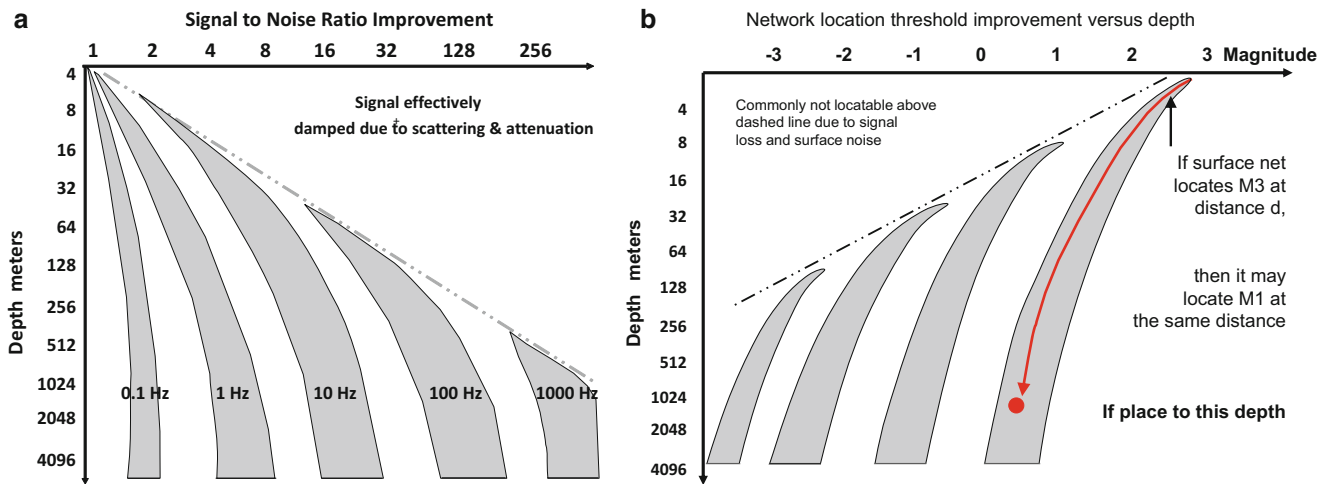
Data from multiple borehole sensors can be processed in two modes: as travel-time/ray path signals or as fully sampled waveforms. This distinction depends on the separation of the sensors, laterally or vertically, and the apparent wavelength of the signals to be processed. A simple rule that separates these two domains is given by Rayleigh's criterion – the rule of thumb being that sensors separated by less than a quarter of the apparent wavelength observed can be used in practical signal enhancement procedures. Otherwise the data must be handled as independent arrivals.

In the ray domain, borehole seismometer data have been used in two fashions: back-projecting rays to their sources based on P-wave polarization and S-P times and Kirchhoff migration studies of event amplitudes. In the waveform domain, signal-to-noise improvement have been accomplished



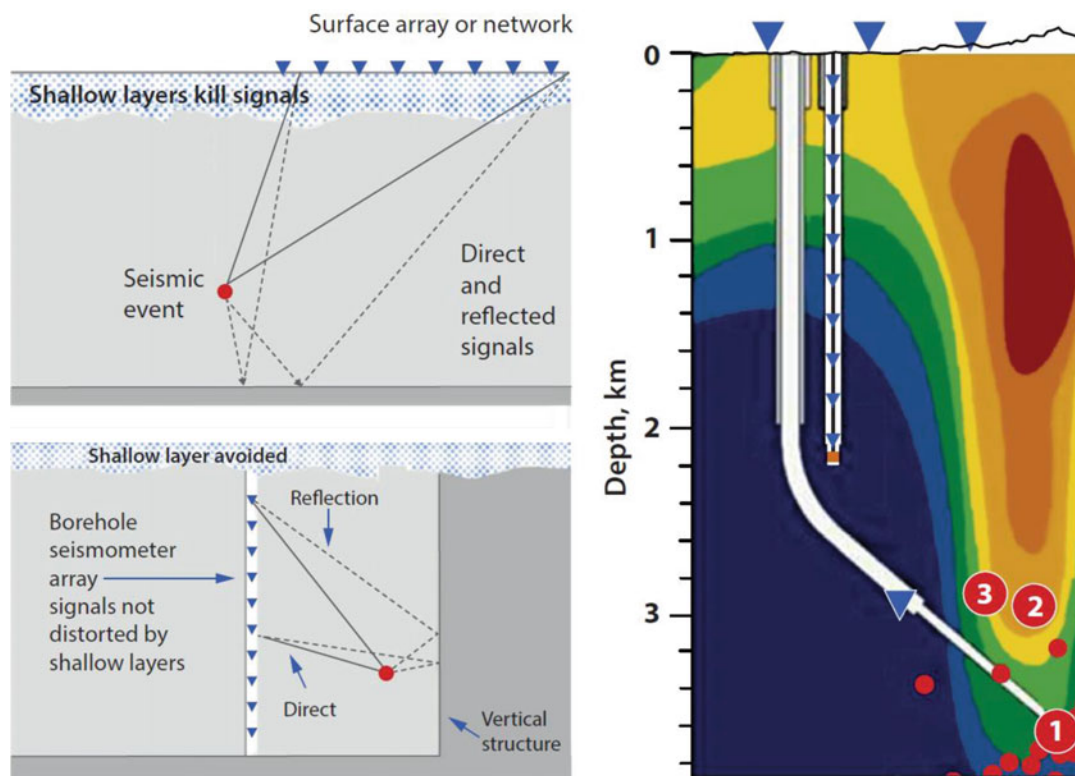
Borehole Seismic Networks and Arrays, Fig. 1 (a) An example of analog borehole seismographs for deployment in a hybrid, four-level vertical borehole array and seven-site lateral network configuration. These sensor arrays of 1, 2, and 15 Hz seismographs were deployed at 75 m spacing in each of the seven matching, 300 m deep, GONAF stations surrounding the eastern Marmara Sea (Prevedel et al. 2015; Bohnhoff et al. 2017.) While the seven stations are too far apart to be used as a lateral array for signal improvement methods such as beam forming, the arrays in each of the seven can be used to separate up-coming signals from down-going surface noise. These deployments can also be used to determine near-surface site effects as they impact earthquake locations (Raub et al. 2016). (b) An example of a multilevel Vertical Seismic Profiling array, with clamping arms extended. The

availability of electrical power for the clamps allows for the inclusion of digitizing and telemetry electronics in each of the 15 Hz-seismometer-based sounds. The spacing between levels can be adjusted from just a few meters to tens by changing the interconnects between levels. (c) An example of a marine hydrophone array – known as a streamer – rolled up on a reel. When deployed as a vertical array, the end of the streamer is anchored on the ocean floor and held vertical with a surface or near-surface buoy. (d) An example of a Fiber-Optic Distributed Acoustic Sensing (DAS) system. ((a) Modified from Prevedel et al. 2015; (b) courtesy of Section 4.2 ‘Geomechanics and Scientific Drilling’ at the GFZ German Geosciences Research Center, Potsdam, Germany; (c) https://commons.wikimedia.org/wiki/File:Streamer-detail_hg.jpg; (d) courtesy of J.-A. Chavarria, Horizon 2020)



Borehole Seismic Networks and Arrays, Fig. 2 (a) A typical range of signal-to-noise and bandwidth improvement for different frequency bands as a function of seismometer depth. This diagram, obtained from a variety of geological sites, shows the limited improvements resulting for burial of low frequency signals versus the significant one for high frequency ones. It also shows the increase in bandwidth versus depth. The width of the frequency band improvements is intended to cover the effects of different rock types and surface weathering and fracturing.

(b) Improvement in network detection and location capacity as a function of placing sensors below existing stations. The sequence of curved shapes is intended to show how elimination of near surface attenuation and noise by sensor burial lowers the network's threshold. In the example indicated on the right side of the diagram, a surface-based network that detects and locates ~M3 events at a given distance d can do the same for events one to two magnitude units lower by placing the networks sensors downhole modified from Bohnhoff et al. 2018



Borehole Seismic Networks and Arrays, Fig. 3 (Left) A schematic comparison between a surface array or network recording a subsurface seismic event (red dot) (upper left) and a borehole array or network (lower left). Blue triangles are sensors. The main advantages of the borehole installation are avoiding upward propagation through the attenuation and scattering of the near surface layers (e.g., Blakeslee and Malin 1990), attenuation and trapping of surface noise by these same layers, the appropriate geometry to record signals from vertical and subvertical structures,

and sensitivity to lateral velocity gradients. (Right) Improvements in event location at the SAFOD drill site: (1) are earthquake locations derived from (near) surface sensors (Zoback et al. 2011; E. Shalev, Personal Communication, 2003), (2) shows the refinement by the addition of the 32-level Pilot Hole vertical array (PHA, blue triangles) (Thurber et al. 2003), and then (3) by the subsequent installation of the Main Hole seismometer (MHS, large blue triangle). (Background color shows electrical resistivity structure modified from Unsworth et al. (1997).)

with various signal stacking and correlation procedures. It is important to note that in both domains it is essential that the orientation of the three component sensors be accurately known.

Even when processed as a network of sensors along a borehole versus as a signal-enhancement array, vertically separated seismometers have advantages such as identifying down-going surface noise from upcoming earthquake signals. For impulsive signals, this can be done by noting arrival orders. For long noise trains, it can be done by noting the lack of noise reaching the deepest sensors.

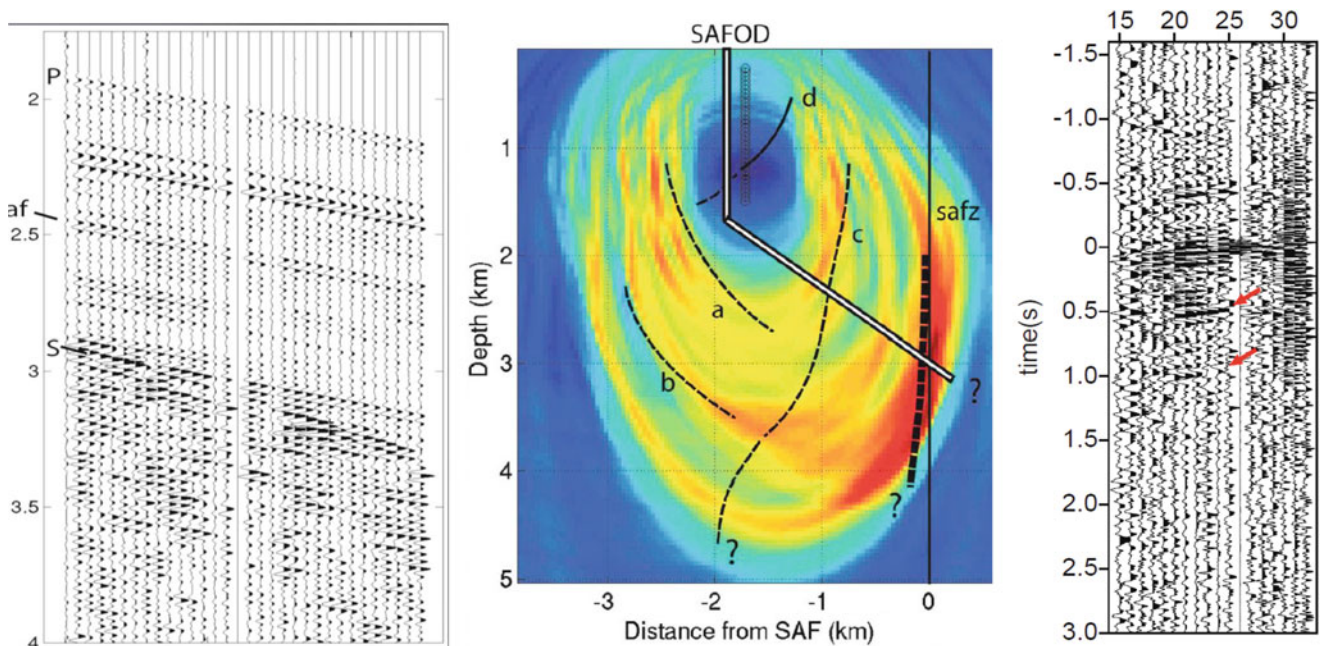
Illustrative Examples

Prime examples of borehole seismic networks made up of single-seismometer stations are government run local, national, and global monitoring systems. These include the high-resolution networks at Parkfield in California (Parkfield HRSN 2014) and Taipei in Taiwan (Huang et al. 2010). The most extensive national network is the constantly growing Japanese Hi-Net, with many hundreds of stations placed more than 100 m underground (Takahashi 1982; Hamada et al. 1985; Obara 2003). Global borehole network stations have served both research and international security needs. The borehole-based Seismic Research Observatory made

significant contributions to the study of the Earth's deep structure (<https://www.fdsn.org/networks/detail/SR/>). Many of the stations of the Global Seismic Network and Comprehensive Test Ban Treaty nuclear verification network are borehole-based (e.g., Bent 2013; <https://www.iris.edu/hq/programs/gsn>; Richards 2016; <https://www.ctbto.org/>).

In the ray domain, borehole seismometer strings like the one in the San Andreas Fault Observatory at Depth (SAFOD; Zoback et al. 2011) have been used in two forms: (1) back-projecting rays to their earthquake sources based on P-wave polarization and S-P times (Oye et al. 2004) and (2) Kirchhoff Migration studies of secondary arrivals originating from subvertical faults and other structures (Fig. 4a, b; Chavarria et al. 2003).

The ability of deep seismometer installations to accurately locate earthquake hypocenters is illustrated by the target event for the San Andreas Fault Observatory at Depth project (Fig. 3, right; Zoback et al. 2011; Thurber et al. 2004; Zhang et al. 2009; Unsworth et al. 1997). Initially, a large network of several dozen surface stations was used to establish a preliminary target event location using travel time tomography (Thurber et al. 2003). Shortly thereafter, arrival times from both earthquakes and explosions recorded on a 32-level string of borehole seismometers installed in the 2 km deep SAFOD Pilot Hole were added to the inversion (Thurber et al. 2004). The result was that the location of the target event moved some 700–800 m, mostly upwards, and less so to the



Borehole Seismic Networks and Arrays, Fig. 4 (a) The vertical component seismograms from the 32-level SAFOD 2 km deep Pilot Hole seismometer string. The signals are upcoming waves from an earthquake located on the San Andreas Fault. (b) A data set of both earthquakes and explosions were Kirchhoff migrated by Chavarria et al. (2003) to find potential reflection structures along the path of the SAFOD Main Hole. (c) Interferometric processing of SAFOD Main Hole Seismic

While Drilling data recorded on the Pilot Hole string to yield a clear reflection from fault structures within the San Andreas Fault Zone (Vasconcelos et al. 2008). The signals indicated by the red arrows are from structures inside the SAFZ system. As part of their analysis, Vasconcelos et al. (2008) also used Pilot Hole data to complete a wave field migrated image of the SAFZ structure. ((a and b) Modified from Chavarria et al. 2003; (c) modified from Vasconcelos et al. 2008)

southwest (Zoback et al. 2011; E. Shalev, Personal Communication, 2003). Afterwards, drilling of the SAFOD Main Hole (MH) was stopped short of this target to allow further location improvement by placing a seismometer at its bottom. The results were a further 200 m move mostly to the southwest.

An unexpected advance in fault zone velocity structure imaging at Parkfield came from borehole seismometer placed within tens of meters of the SAF (Ellsworth and Malin 2011; Malin: Fault Zone Guided Waves, this volume). These revealed head waves refracted along the faster side of the SAFZ (Ben-Zion and Malin 1991). In this same data set, there were clear, normally dispersed, fault zone guided waves (F_g). These fit Love-type trapped waves (F_L) propagating in a several km deep, ~100–120 m thick, low velocity channel (Li and Malin 2008). This picture was further refined by data from the SAFOD MH sensor, which contained Rayleigh-type trapped waves (F_R) and a new, leaky mode, fault guided wave F_ϕ . The latter phase fits partially trapped PSV waves travel in a several km deep, ~30 m wide fault core of very low P velocity.

A unique example of a local borehole-based network is the Geophysical Observatory on the North Anatolian Fault (GONAF) near Istanbul, Turkey (Bohnhoff et al. 2017). This seismic observatory is a hybrid network of borehole array stations, each station consisting of a four-level vertical array of seismometers spaced ~75 m apart, going up from ~300 m (Fig. 1a). The different levels has been used to follow up-coming and surface-reflected microearthquake signals for site effect studies (Raub et al. 2016) and to discriminate between down going surface noise and actual microearthquake signals (Malin et al. 2018). The former study uncovered a reversal in seismic velocities that grossly affected the motion of upcoming signals. In the latter, the noise suppression was used to lower the magnitude threshold by ~2–3 units, allowing detection of many microearthquakes preceding a 4.2 mainshock.

Some of the networks mentioned above can also be operated as arrays in a narrow range of signal frequencies, as in the case of surface wave analysis (Luo et al. 2008). However, lateral arrays and multilevel borehole sensor strings have been developed with the specific purpose of signal enhancement in mind (Fig. 1b). Early work in this type of Vertical Seismic Profiling in boreholes was begun in the 1950s by Y. I. Galperin (Galperin 1961; Galperin et al. 1986) and extensively expanded upon in the following decades (Balch and Lee 1984; Balch et al. 1986). The array aspects of lateral borehole networks have been extensively developed by the nuclear verification community (Richards 2016), primarily in the area of beam forming. These methods are now being adapted by energy industry seismic contractors (McClellan et al. 2018).

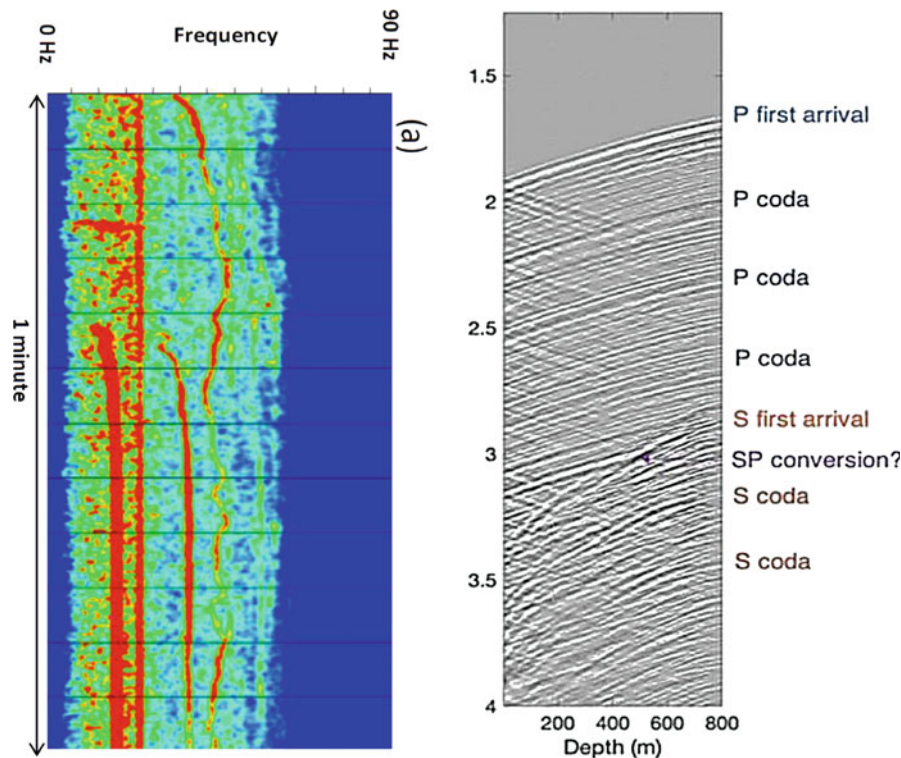
Full waveform signal-enhancement processing has also been applied to Seismic-While-Drilling data that was

collected with the Pilot Hole string during the drilling of the SAFOD MH. Using vibroseis-type deconvolution and seismic reflection signal processing methods, Coates et al. (2006) were able to obtain an image showing steeply dipping Tertiary Great Valley strata immediately south of the SAFOD site – an unexpected juxtaposition based on surface and area borehole geology. Because of their subvertical attitude and signal loss to near-surface attenuation and noise, this structure would have been difficult to image with standard surface reflection profiling. Further signal processing of the Seismic While Drilling data was done by Vasconcelos et al. (2008) (Fig. 4c). They used an interferometric approach to signal stacking to obtain an image of the internal structure within the fault zone.

In addition to downhole seismometers and accelerometers, multisensor hydrophone (streamer) arrays have been adapted for vertical deployments in boreholes on land and in the open ocean (Fig. 1c; Marzetta et al. 1988; Toksöz et al. 1992; Ikelle and Amundsen 2018). On land, early tests revealed that incoherent tube waves in long boreholes could completely cover secondary arrivals, initially limiting their usefulness. This problem was by and large solved by various signal processing and mechanical tube-wave-damping methods. Moreover, such arrays allowed detection of new seismic source processes related to migrating fluids and gases due to wellbore leakage (Bohnhoff and Zoback 2010; Bohnhoff et al. 2010; Martinez-Garzon et al. 2013). In the ocean, while near-bottom sensors are significantly quieter than near-surface ones, ocean currents can cause vertical arrays to tilt in various directions over minutes, hours, and days, making signal processing a challenge.

On land, the elimination of near-surface attenuation and suppression of surface noise has also allowed observation of what appear to be Krauklis wave resonances in subsurface fractures (Sicking and Vermilye 2019; Liang et al. 2017; Tary et al. 2014a, b). Figure 5a shows a spectrogram obtained by the stacking of frequency-time-amplitude data from a 12-level 2.5 km deep borehole seismometer string at a geothermal development site in Finland.

All these previous studies collected point-observation data using conventional borehole seismographic instruments: passive spring-and-mass and hydrophone sensors, force feedback, and similar nodal devises. Now on the near-horizon for broad application are borehole fiber-optic Distributed Acoustic Sensing (DAS) systems (Fig. 1d). A high-level account of this new technology was recently published by Lellouch et al. (2019). DAS systems have now effectively achieved continuous sampling of broadband seismic wavefields. Lellouch et al. (2019) have compared borehole DAS observations of earthquake and explosion data from multichannel VSP instruments at the SAFOD site. Their analysis included travel time picking, slant stack beam forming, and interferometry. Due to its redundancy, it was



Borehole Seismic Networks and Arrays, Fig. 5 (a) Spectrogram of resonances from major subsurface fractures recorded by a 12-level borehole seismometer string during hydraulic stimulation of a geothermal well in Finland. The string was placed between 1970 and 2420 m below ground and ~3600 m above the stimulation zone. While the resonance could be seen on individual-level spectrograms, the current plot is the stack of spectrograms from 11 levels. Based on the work of Tary et al. (2014a, b) and Liang et al. (2017), and correlations with stimulation pumping and induced earthquakes, Sicking and Malin (2019) have interpreted these signals as coming from interfering

Krauklis wave on major subsurface fractures (Figure courtesy of P. E. Malin 2019). (b) Broadband Fiber-Optic DAS recordings of M1.33 and M2.4 earthquakes on the San Andreas Fault at Parkfield. These data were collected with a DAS cable along the first 800 m of the SAFOD Main Hole. The data have been bandpass filtered between 0 and 120 Hz. Compared to the point observation data in Fig. 4a, b, these data meet the 1/4 wavelength criterion for high-frequency signal enhancement. The result is a nearly continuous wavefield image that provides a much richer look at the velocity structure surrounding the drill site and SAF. (Figure from Lellouch et al. 2019)

possible to detect both S-waves and clear P-S reflections in the DAS data (Fig. 5b). In addition to the technical description of DAS operation, their paper includes important velocity and structural results from SAFOD.

Conclusions

Progressing from a single spring-mass seismometer in shallow water well to continuous Fiber-Optical DAS cables over the past 100 years, borehole seismic networks and arrays have opened up a previously unobtainable view of the earth's interior. It facilitates the detection and monitoring of deep and remote, and otherwise inaccessible, seismic events and tectonic processes. This includes the signals of very small earthquakes, very large explosions, and, now, evidently, resonances in fractures stimulated by hydraulic pumping. Used to image geological resources, tectonic structures and potentially hazardous faults, borehole seismology allows for less

intrusive remote exploration of the crust and refinement of seismic hazards and risks.

With further technological and methodological improvements, downhole seismic networks and arrays may become the norm of high-quality seismic observations. This evolution is currently taking place with the advent of Fiber-Optic DAS systems. Combined with lowered costs in drilling and the needs for a more fundamental understanding of the earth and its resources and hazards, borehole seismic networks and arrays are likely to be the next step in the history of seismology.

Cross-References

- [Body Waves](#)
- [Continental Crustal Structure](#)
- [Earthquake, Focal Mechanism](#)
- [Earthquake, Magnitude](#)

- Earthquakes and Crustal Deformation
- Earthquakes, Strong-Ground Motion
- Energy Partitioning of Seismic Waves
- Fault Zone Guided Waves
- Propagation of Elastic Waves: Fundamentals
- Seismic Diffraction
- Seismic Imaging, Overview
- Seismic Instrumentation
- Seismic Phase Nomenclature: The IASPEI Standard
- Seismic Signals in Well Water Observations
- Seismic Wave Propagation in Real Media: Numerical Modeling Approaches
- Seismic Waves, Scattering
- Seismic, Reflectivity Method
- Seismogram Interpretation
- Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting
- Single and Multichannel Seismics
- Slow Earthquake
- Surface Waves

Bibliography

- Balch AH, Lee MW (1984) Vertical seismic profiling: technique, applications, and case histories. International Human Resources Development Corporation/Reidel, Boston/Dordrecht/Lancaster
- Balch AH, Lee MW, Paillet FL (1986) Vertical seismic profiling – technique, application, and case histories. *J Acoust Soc Am* 80(3). <https://doi.org/10.1121/1.393897>
- Bent A (2013) Global seismograph network (GSN). In: Bobrowsky PT (ed) *Encyclopedia of natural hazards. Encyclopedia of earth sciences series*. Springer, Dordrecht
- Ben-Zion Y, Malin PE (1991) San Andreas Fault zone head waves near Parkfield, California. *Science* 251:1592–1594
- Blakeslee SN, Malin PE (1990) A comparison of earthquake coda waves at surface versus subsurface seismometers. *J Geophys Res Solid Earth* 95:309–326. <https://doi.org/10.1029/JB095iB01p00309>
- Bohnhoff M, Zoback MD (2010) Oscillation of fluid-filled cracks triggered by degassing of CO₂ due to leakage along wellbores. *J Geophys Res* 115:B11305. <https://doi.org/10.1029/2010JB000848>
- Bohnhoff M, Zoback MD, Chiaramonte L, Gerst JL, Gupta N (2010) Seismic detection of CO₂ leakage along monitoring wellbores. *Int J Greenhouse Gas Control* 4(4):687–697. <https://doi.org/10.1016/j.ijggc.2010.01.009>
- Bohnhoff M, Dresen G, Ceken U, Kadiroglu FT, Karal RF, Kilic T, Nurlu M, Yanik K, Acarel D, Bulut F, Ito H, Johnson W, Malin PE, Mencin D (2017) GONAF – a borehole geophysical observatory around the north Anatolian fault in the eastern Sea of Marmara. *Sci Drill* 22:19–28
- Bohnhoff M, Malin P, ter Heege J, Deflandre J-P, Sicking C (2018) Suggested best practice for seismic monitoring and characterization of non-conventional reservoirs. *First Break* 36:59–64
- Chavarría JA, Malin PE, Shalev E, Catchings RD (2003) A look inside the San Andreas Fault at Parkfield through vertical seismic profiling. *Science* 302:1746–1748
- Coates R, Haldorsen JBU, Miller D, Malin P, Shalev E, Taylor ST, Stolte C, Verliac M (2006) Oilfield technologies for earthquake science. *Oilfield Rev* 18(2):24–33
- Eisner L, Hulseley BJ, Duncan P, Jurick D, Werner H, Keller W (2010) Comparison of surface and borehole locations of induced seismicity. *Geophys Prospect* 58(5):809–820. <https://doi.org/10.1111/j.1365-2478.2010.00867.x>
- Ellsworth WL, Malin PE (2011) Deep rock damage in the San Andreas Fault revealed by P- and S-type fault-zone-guided waves. *Geol Soc London Spec Publ* 359:39–53. <https://doi.org/10.1144/SP359.3>
- Galperin YI (1961) Three component downhole seismic acquisition. Moscow (in Russian)
- Galperin EI, Nersesov IL, Galperina RM (1986) Seismological observations in boreholes. In: *Borehole seismology and the study of the seismic regime of large industrial centres. Seismology and exploration geophysics, vol 2*. Springer, Dordrecht
- Gardner LW (1949) Seismograph determination of a salt dome boundary using a well detector deep on the dome flank. *Geophysics* 14:29–38
- Hamada K, Ohtake M, Okada Y, Matsumura S, Sato H (1985) A high-quality digital network for microearthquake and ground tilt observations in the Kanto-Tokai area, Japan. *Earthquake Pred Res* 3:447–469
- Haring MO, Schanz U, Ladner F, Dyer BC (2008) Characterization of the Basel 1 enhanced geothermal system. *Geothermics* 37:469–495
- High-Resolution Seismic Network (HRSN) (2014) High resolution seismic network. UC Berkeley Seismological Laboratory, Berkeley. <https://doi.org/10.7932/HRSN>
- Huang W-G, Huang B-S, Wang I-H, Chen K-C, Wen K-L, Tsao S, Hsieh Y-C, Chen C-H (2010) Seismic observations in the Taipei Metropolitan Area using the downhole network. *Terr Atmos Ocean Sci* 21(3). [https://doi.org/10.3319/TAO.2009.12.11.03\(TH\)](https://doi.org/10.3319/TAO.2009.12.11.03(TH))
- Ikelle LT, Amundsen L (2018) Introduction to petroleum seismology. Society of Exploration Geophysicists, Tulsa. <https://doi.org/10.1190/1.9781560801702.ch5>
- Johnson HM (1962) A history of well logging. *Geophysics* 27(4):427–541. ISSN: 0016-8033; 1942-2156. <https://doi.org/10.1190/1.1439054>
- Kwiatak G, Saarno T, Ader T, Bluemle F, Bohnhoff M, Chendorain M, Dresen G, Heikkinen P, Kukkonen I, Leary P, Leonhardt M, Malin P, Martínez-Garzón P, Passmore K, Passmore P, Valenzuela S, Wollin C (2019) Controlling fluid-induced seismicity during a 6.1-km-deep geothermal stimulation in Finland. *Sci Adv* 5(5):eaav7224. <https://doi.org/10.1126/sciadv.aav7224>
- Lellouch A, Yuan S, Spica Z, Biondi B, Ellsworth WL (2019) Seismic velocity estimation using passive downhole distributed acoustic sensing records – examples from the San Andreas Fault Observatory at Depth. *J Geophys Res Solid Earth*. <https://doi.org/10.1029/2019JB017533>
- Li Y-G, Malin PE (2008) San Andreas Fault damage at SAFOD viewed with fault-guided waves. *Geophys Res Lett* 35:L08304. <https://doi.org/10.1029/2007GL032924>
- Liang C, O'Reilly O, Dunham EM, Moos D (2017) Hydraulic fracture diagnostics from Krauklis-wave resonance and tube-wave reflections. *Geophysics* 82(3):D171–D186
- Luo Y, Xia J, Miller RD et al (2008) Rayleigh-wave dispersive energy imaging by high resolution linear radon transform. *Pure Appl Geophys* 165(5):903–922
- Malin PE, Bohnhoff M, Blumle F, Dresen G, Martínez-Garzon P, Nurlu M, Ceken U, Kadiroglu FT, Kartal RF, Kilic T, Yanik K (2018) Microearthquakes preceding a M4.2 Earthquake Offshore Istanbul. *Sci Rep* 8:Article number: 16176. <https://doi.org/10.1038/s41598-018-34563-9>
- Martínez-Garzon P, Bohnhoff M, Zambrano-Narvaez G, Chalaturnyk R (2013) Microseismic monitoring of CO₂ injection at the Penn West Pembina Cardium EOR Project, Canada. *Sensors* 13:11522–11538. <https://doi.org/10.3390/s130911522>
- Marzetta T, Orton M, Krampe A, Johnston L, Wuenschel P (1988) A hydrophone vertical seismic profiling experiment. *Geophysics* 53(11):1437–1444. <https://doi.org/10.1190/1.1442423>

- McClellan JH, Eisner L, Liu E, Iqbal N, Al-Shuhail AA, Kaka SI (2018) Array processing in microseismic monitoring detection, enhancement, and localization of induced seismicity. *IEEE Signal Process Mag.* <https://doi.org/10.1109/MSP.2017.2776798>
- McCollum B, LaRue WW (1931) Utilization of existing wells in seismograph work. *Bull Am Assoc Pet Geol* 15:1409–1417
- Musgrave AW, Woolley WC, Gray H (1960) Outlining of salt masses by refraction methods. *Geophysics* 25:141–167
- Obara K (2003) Hi-Net: high sensitivity seismograph network, Japan. In: *Methods and applications of signal processing in seismic network operations. Lecture notes in earth sciences*, vol 98. Springer, Berlin
- Oye V, Chavarria J-A, Malin P (2004) Determining SAFOD area micro-earthquake locations solely with the Pilot Hole seismic array data. *Geophys Res Lett* 31:L12S10. <https://doi.org/10.1029/2003GL019403>
- Prevedel B, Bulut F, Bohnhoff M, Raub C, Kartal R, Fatih Alver F, Malin PE (2015) Downhole geophysical observatories: best installation practices and a case history from Turkey. *Int J Earth Sci* 104:1537. <https://doi.org/10.1007/s00531-015-1147-5>
- Raub C, Bohnhoff M, Petrovic B, Parolai S, Malin PE, Yanik K, Kartal RF, Kiliç T (2016) Seismic wave propagation in shallow layers at the GONAF-Tuzla site, Istanbul, Turkey. *Bull Seismol Soc Am* 106:912–927. <https://doi.org/10.1785/0120150216>
- Richards PG (2016) The history and outlook for seismic monitoring of nuclear explosions in the context of the Comprehensive Nuclear-Test-Ban Treaty. *Nonprolif Rev.* <https://doi.org/10.1080/10736700.2016.1272207>
- Sicking CJ, Malin PE (2019) Fracture seismic: mapping subsurface connectivity. *Geosciences* 2019(9):508
- Sicking CJ, Vermilye J (2019) Resonance frequencies in passive recordings map fracture systems: Eagle Ford and New Albany Shale examples. Extended abstract, unconventional resources technology conference (URTEC). <https://doi.org/10.15530/urtec-347>
- Takahashi H (1982) The deep borehole observatories and their contribution for revealing the characteristics of microearthquake activity in the Kanto district. *Rep Natl Res Cent Disaster Prev* 28:1–104. (in Japanese with English abstract)
- Tary JB, Van der Baan M, Eaton DW (2014a) Interpretation of resonance frequencies recorded during hydraulic fracturing treatments. *J Geophys Res Solid Earth* 119(2):1295–1315
- Tary JB, Van der Baan M, Sutherland B, Eaton DW (2014b) Characteristics of fluid induced resonances observed during microseismic monitoring. *J Geophys Res* 119:8207–8222
- Thurber C, Roecker S, Roberts K, Gold M, Powell L, Rittger K (2003) Earthquake locations and three-dimensional fault zone structure along the creeping section of the San Andreas Fault near Parkfield, CA: preparing for SAFOD. *Geophys Res Lett* 30(3):1112. <https://doi.org/10.1029/2002GL016004>
- Thurber C, Roecker S, Zhang H, Baher S, Ellsworth W (2004) Fine-scale structure of the San Andreas Fault zone and location of the SAFOD target earthquakes. *Geophys Res Lett* 31(12):L12S02. <https://doi.org/10.1029/2003GL019398>
- Toksöz MN, Cheng CH, Cicerone RD (1992) Fracture detection and characterization from hydrophone vertical seismic profiling data. *Int Geophys Ser* 51:389–414. [https://doi.org/10.1016/S0074-6142\(08\)62831-4](https://doi.org/10.1016/S0074-6142(08)62831-4)
- Unsworth MJ, Malin PE, Egbert GD, Booker JR (1997) Internal structure of the San Andreas Fault at Parkfield, California. *Geology* 25:359–362
- Vasconcelos I, Sneider R, Sava P, Taylor T, Malin P, Chavarria A (2008) Drill bit noise illuminates the San Andreas Fault. *EOS Trans Am Geophys Union* 89(38):349–360
- Zhang H, Thurber C, Bedrosian P (2009) Joint inversion for Vp, Vs, and Vp/Vs at SAFOD, Parkfield, California. *Geochem Geophys Geosyst* 10:Q11002. <https://doi.org/10.1029/2009GC002709>
- Zoback M, Hickman S, Ellsworth W, the SAFOD Science Team (2011) Scientific drilling into the San Andreas Fault zone – an overview of SAFOD's first five years. *Sci Drill* 11:14–28. <https://doi.org/10.2204/iodp.sd.11.02.2011>

Characteristic Earthquakes and Seismic Gaps

David D. Jackson and Yan Y. Kagan
Department of Earth and Space Sciences, University of
California Los Angeles, Los Angeles, CA, USA

Definition

Fault slip	Relative motion, either steady or sudden as in earthquakes, between rock units on either side of a fault.
Rupture	Sudden fault slip in an earthquake.
Elastic rebound	Sudden release, as by an earthquake, of slowly accumulated strain energy.
Segment	A section of a fault or plate interface bounded by features thought to serve as strong barriers to earthquake rupture. Features postulated to form such barriers include changes in fault orientation or in rock type across parts of the fault, and intersections with other faults.
Characteristic earthquake	An earthquake rupturing an entire fault segment. Alternately, one of a sequence of earthquakes rupturing the same area of fault.
Recurrence interval	The time between characteristic earthquakes on a given segment or fault area. <i>Quasiperiodic</i> Occurring at approximately equal recurrence intervals.
Seismic cycle	A sequence of events on a segment starting with a large earthquake, followed by aftershocks, then by steady stress accumulation, and culminating with another large earthquake. The term “cycle” is sometimes but not always meant to imply quasiperiodic recurrence.

Seismic gap A segment for which the time since the previous characteristic earthquake approaches or exceeds the average recurrence interval.

Introduction

The seismic gap hypothesis holds that most long-term geologic slip on faults or plate boundaries is accomplished by characteristic earthquakes on segments. Such quakes are presumed to reduce the stress substantially, necessitating a substantial recurrence time for elastic stress to recover before the next characteristic earthquake. The dates and rupture extent of past earthquakes may be determined by modern seismic networks; by historic reports of faulting, damage, or strong shaking; or by paleoseismic investigation of trenches across faults. The average recurrence time may be determined either from a sequence of dates of past characteristic earthquakes, or by the time required for steady slip on a fault to accumulate the slip experienced in a characteristic earthquake (Working Group on California Earthquake Probabilities 1988, 1990).

History

Well before plate tectonics became accepted in the 1960s, Gilbert (1884) argued that large earthquakes should be separated by substantial time intervals. Reid (1910, 1911) proposed that faults slowly accumulate energy later released suddenly by earthquakes (“elastic rebound”), and that the time to a future earthquake could be estimated by surveying the strain released by a previous one. Fedotov (1965) noticed that most of the Kamchataka–Kurile trench in the northwest Pacific had been ruptured by earthquakes in the early twentieth century, with the exception of one zone southeast of Hokkaido. He concluded that a future earthquake was likely in that area, a forecast realized in 1973 when the Nemuro-oki earthquake struck with magnitude about 7.5 (Kasahara 1981, 182).

Plate tectonics now provides a compelling, steady source of the strain energy driving elastic rebound, and that fact led many to conclude that occurrence of large earthquakes must be separated by enough time to allow recharge. Scholz (1990, 260) remarked that “A tenet of plate tectonics is that the rates of plate motion must be steady over geologic time periods and must be continuous along plate boundaries. If it is assumed further that a significant portion of this motion must be released seismically, then it follows that segments of plate boundaries that have not ruptured for the longest time are those most likely to rupture in the near future. These places are called seismic gaps.” The seismic gap hypothesis was celebrated when Nishenko (1989) published a list of 13 earthquakes that fit, at least approximately, the descriptions of previously expected earthquakes.

The seismic gap model has been applied to long-term forecasting of earthquakes in many regions (Sykes 1971; Kelleher 1972; Kelleher et al. 1973; McCann et al. 1979; Working Group on California Earthquake Probabilities 1988; Nishenko 1991). The model and its definition have evolved along with the quality and quantity of data that go into it. Fedotov (1965) defined the Nemuro-oki gap as the last remaining unbroken segment. McCann et al. (1979) used elapsed time and a color code to label segments around the Pacific Rim: more than 100 years had elapsed in red gaps, between 30 and 100 years in orange gaps, and less than 30 years in apparently safe green zones. In the most comprehensive forecast ever using the seismic gap model, Nishenko (1991) actually made testable probabilistic forecasts for about 125 plate boundary segments around the Pacific Rim. In that study he estimated the mean recurrence time and the elapsed time for each segment, and assumed that their ratio obeys a log-normal probability density function. With that information, he calculated the probability that each zone would be ruptured by a characteristic earthquake whose magnitude he also listed for each zone within 5-, 10-, and 30-year periods. A very similar model was used by the Working Group on California Earthquake Probabilities (1988) in its official earthquake probability estimates.

Assumptions

All the applications mentioned above share several important assumptions. First, their authors assume that faults and plate boundaries are segmented and that rupture does not cross segment boundaries. Second, they assume that each characteristic earthquake ruptures to both ends of its segment, reducing the stress to a uniform base level and beginning the process of stress recharge. Third, they assume that the time to the next characteristic earthquake depends almost entirely on the time of the previous one: not on other

earthquakes, nonelastic stress redistribution, or other causes. To make useful forecasts, scientists must obviously be able to distinguish characteristic earthquakes from all others in order to know the elapsed time since the last one.

Small Characteristic Earthquakes

Recent studies in California (e.g., Nadeau and McEvilly 1999), Japan (e.g., Igarashi et al. 2003; Okada et al. 2003), and elsewhere have identified sequences of small earthquakes fitting the alternative definition of “characteristic earthquake” above. In each sequence, the events are approximately the same size, and rupture approximately the same fault area. They recur at nearly equal time intervals or in some cases variable intervals consistent with variations in their size or in fault slip rate. In most cases, the slipped areas appear to be surrounded by areas where displacement occurs by steady slip rather than earthquakes. Because of that special circumstance, these small repeating earthquakes are not relevant to the discussion of seismic gaps.

Modified Seismic Gap Hypothesis

As time, earthquake experience, and theoretical sophistication have accumulated, earth scientists have modified the seismic gap theory to rely less on the assumptions above. The Working Group on California Earthquake Probabilities (1990) allowed adjustments to account for stresses from earthquakes not on the relevant segment. The 1992 Landers, California earthquake (magnitude about 7) presented a particularly important observation. In that event, rupture jumped segment boundaries and even faults, making use of up to five faults mapped as separate before 1992. In a seismic hazard model produced by the Southern California Earthquake Center (Working Group on California Earthquake Probabilities 1995), the seismic gap model was modified to allow rupture to jump segment boundaries with a modest probability. Later uses of the seismic gap model in California for official hazard estimates employ increasingly complex assumptions, especially about conditions under which rupture is likely to involve more than one segment (Working Group on California Earthquake Probability 2002, 2008).

The more complex versions of the model raise interesting questions. What truly constitutes a segment? How effective are weak barriers in stopping rupture, and what controls their strength? Are the boundaries fixed in space, or can they move as stress conditions change? When rupture crosses a boundary, does it consistently continue to the next? If not, does it reset the stress and the clock on the partially ruptured segment? Do the elapsed times on adjacent segments control the

probability that rupture will jump the barrier between them? If so, which segment is most important? Modelers must answer these questions, implicitly or explicitly, to forecast using the modified gap models. So far, there is no clear consensus on the answers.

Challenges to the Seismic Gap Model

Despite some reported successes, the seismic gap hypothesis has often been questioned. Critics point to the difficulty of verifying the rather strong assumptions behind the hypothesis, and to its limited success in forecasting earthquakes.

The basic assumption that faults and plate boundaries are segmented has provoked significant debate. Even the few apparent successes (e.g., Loma Prieta, CA, 1989; Parkfield, CA 2004; Chile, 2010), are equivocal at best. The rupture of the Loma Prieta earthquake was about the length of the nearest segment mapped by the Working Group on California Earthquake Probabilities (1988), but it shifted south and spilled over the southern segment boundary. Moreover, the event occurred near, but not actually on the San Andreas Fault for which the segment was defined. Before 2004, the Parkfield segment was defined in several different ways, so its location at Parkfield does not confirm the segmentation hypothesis (Jackson and Kagan 2006). The 2010 Chile earthquake went well beyond the segment boundaries specified by Nishenko (1991). In addition, several events have clearly violated pre-assigned boundaries. The 1992 Landers quake was mentioned above, and the great Sumatra tsunami earthquake of 2004 breached several boundaries along its 1300-km rupture zone (Nalbant et al. 2005).

The assumption that earthquakes rupture to both ends of segment boundaries also lacks verification. A basic problem is that the locations of segment boundaries are usually estimated inaccurately from the extent of past earthquake ruptures. For earthquakes identified by paleoseismic investigations, rupture can generally be pinpointed at only a few widely spaced locations. For historical earthquakes, rupture extent is typically estimated with great uncertainty from the extent of damage or reported shaking. Even for modern instrumentally recorded earthquakes, the extent of the after-shock zone or fault scarp may not accurately represent the rupture at depth where the elastic rebound occurs. In many cases, the extent of rupture for older earthquakes is assumed to be similar to that of the most recent event, a clear case of circular reasoning.

The connection between plate tectonics and earthquake recurrence referred to in the words of Scholz above depends on the assumption that characteristic earthquakes release most of the slowly accumulated fault slip. However, that assumption fails in many examples. For instance, the cumulative slip of the

Parkfield, CA earthquakes since 1857, often regarded as an archetypal characteristic sequence, accounts for only a small fraction of expected slip at the long-term geological rate (Jackson and Kagan 2006). In such cases, the alternative definition of characteristic earthquakes listed above, and the times of past events, may provide valuable information on the causes of some earthquakes but the direct link to plate tectonics is lost.

Proponents of the seismic gap theory cite many examples in which identified gaps have been filled by earthquakes. The positive examples are appealing but insufficient for two reasons. First, the definitions of gaps and forecasted earthquakes were quite general, making the target easier to hit at random. Second, they included only successes; a fair evaluation needs to consider failures as well.

Kagan and Jackson (1991, 1995), and Rong et al. (2003) applied statistical tests to the seismic gap theory as articulated by McCann et al. (1979) and Nishenko (1991). Earthquakes were actually more frequent in McCann's green zones than in the red ones, opposite to what the gap theory assumes. The 1991 gap model implied far more earthquakes, in different places, than actually occurred. Kagan and Jackson (1995) and Rong et al. (2003) also tested a simple alternative to the gap model, assuming that earthquakes occur randomly in time and near past earthquakes. The alternative model fits the total number and the locations of future earthquakes much better than the 1991 gap model.

These statistical tests necessarily applied to the earlier versions of the gap hypothesis, in which segments were assumed to be independent of one another. Since then, more complex versions of the model have been applied. Most of these applications involve one or a few purported gaps with estimated recurrence times of decades or centuries. Published models generally do not provide probabilities for multi-segment ruptures, and they cannot be effectively evaluated until several seismic cycles have elapsed. To be rigorously testable, such a model must forecast a few tens of well-defined earthquakes. Unfortunately, no systematic, well-specified version of the more complex seismic gap model has been applied broadly enough to be tested.

Conclusions

The intuitively appealing seismic gap model encompasses the virtually unassailable principle that earthquakes release strain energy accumulated over a long time. Although many large events have occurred in previously identified gaps, the same is true of locations outside them. Simple versions of the gap theory, in which characteristic earthquakes occur on independent segments, are no longer tenable. Modified versions of the gap theory have not yet been formulated in a rigorously testable way.

Cross-References

- [Earthquake Precursors and Prediction](#)
- [Earthquake Rupture: The Inverse Problem](#)
- [Earthquakes and Crustal Deformation](#)
- [Earthquakes, Energy](#)
- [Earthquakes, Location Techniques](#)
- [GPS, Tectonic Geodesy](#)
- [Great Earthquakes](#)
- [Paleoseismology](#)
- [Seismic Hazard](#)
- [Seismicity, Subduction Zone](#)
- [Seismology, Global Earthquake Model](#)
- [Statistical Seismology](#)

Bibliography

- Fedotov SA (1965) Regularities of the distribution of strong earthquakes in Kamchatka, the Kurile Islands and northeastern Japan. *Tr Inst Fiz Zemli Akad Nauk SSSR* 36(203):66–93 (in Russian)
- Gilbert GK (1884) A theory of the earthquakes of the Great Basin, with a practical application. *Am J Sci* 27(157):49–54. Ser. 3
- Igarashi T, Matsuzawa T, Hasegawa A (2003) Repeating earthquakes and interplate aseismic slip in the northeastern Japan subduction zone. *J Geophys Res* 108(B5):2249. <https://doi.org/10.1029/2002JB001920>
- Jackson DD, Kagan YY (2006) The 2004 parkfield earthquake, the 1985 prediction, and characteristic earthquakes: lessons for the future. *Bull Seismol Soc Am* 96:S397–S409
- Kagan YY, Jackson DD (1991) Seismic gap hypothesis: ten years after. *J Geophys Res* 96(21):21419–21431
- Kagan YY, Jackson DD (1995) New seismic gap hypothesis: five years after. *J Geophys Res* 100(B3):3943–3959
- Kasahara K (1981) *Earthquake mechanics*. Cambridge University Press, Cambridge
- Kelleher JA (1972) Rupture zones of large South American earthquakes and some predictions. *J Geophys Res* 77:2087–2103
- Kelleher JA, Sykes LR, Oliver J (1973) Possible criteria for predicting earthquake locations and their applications to major plate boundaries of the Pacific and Caribbean. *J Geophys Res* 78:2547–2585
- McCann WR, Nishenko SP, Sykes LR, Krause J (1979) Seismic gaps and plate tectonics: seismic potential for major boundaries. *Pure Appl Geophys* 117:1082–1147
- Nadeau RM, McEvilly TV (1999) Fault slip rates at depth from recurrence intervals of repeating microearthquakes. *Science* 285:718–721
- Nalbant S, Steacy S, Sieh K, Natawidjaja D, McCloskey J (2005) Seismology: earthquake risk on the Sunda trench. *Nature* 435(7043):756–757
- Nishenko SP (1989) Earthquakes: hazards and predictions. In: James DE (ed) *The encyclopedia of solid earth geophysics*. Van Nostrand Reinhold, New York, pp 260–268
- Nishenko SP (1991) Circum-Pacific seismic potential – 1989–1999. *Pure Appl Geophys* 135:169–259
- Okada T, Matsuzawa T, Hasegawa A (2003) Comparison of source areas of $M_{4.8} \pm 0.1$ repeating earthquakes off Kamaishi, NE Japan: are asperities persistent features? *Earth Planet Sci Lett* 213:361–374
- Reid HF (1910) *The California earthquake of April 18, 1906. The mechanics of the earthquake*, vol 2. Carnegie Institution of Washington, Washington, DC
- Reid HF (1911) The elastic-rebound theory of earthquakes. *Univ Calif Dep Geol Bull* 6:413–444

- Rong Y-F, Jackson DD, Kagan YY (2003) Seismic gaps and earthquakes. *J Geophys Res* 108(B10):2471, ESE-6, 1–14
- Scholz C (1990) *The mechanics of earthquakes and faulting*. Cambridge University Press, Cambridge
- Sykes LR (1971) Aftershock zones of great earthquakes, seismicity gaps, and earthquake prediction for Alaska and the Aleutians. *J Geophys Res* 76:8021–8041
- Working Group on California Earthquake Probabilities (1988) Probabilities of large earthquakes occurring in California, on the San Andreas Fault. U.S. Geological Survey Open-File Report, 88–398, 62 ~ pp
- Working Group on California Earthquake Probabilities (1990) Probabilities of large earthquakes in the San Francisco Bay Region, California, USGS Circular 1053
- Working Group on California Earthquake Probabilities (1995) Seismic hazards in southern California: probable earthquakes, 1994–2024. *Bull Seismol Soc Am* 85:379–439
- Working Group on California Earthquake Probabilities (2002) Earthquake probabilities in the San Francisco Bay region: 2002 to 2031, USGS Circular 1189
- Working Group on California Earthquake Probabilities (2008) The Uniform California Earthquake Rupture Forecast, Version 2 (UCERF 2): U.S. Geological Survey Open-File Report 2007–1437 and California Geological Survey Special Report 203. <http://pubs.usgs.gov/of/2007/1437/>

Continental Crustal Structure

Rolf Meissner (Deceased) and Hartmut Kern
Kiel, Germany

Definition

Seismic reflection and refraction profiling supplying velocity-depth functions and relations to petrology is the key to probe the continental crust.

Introduction

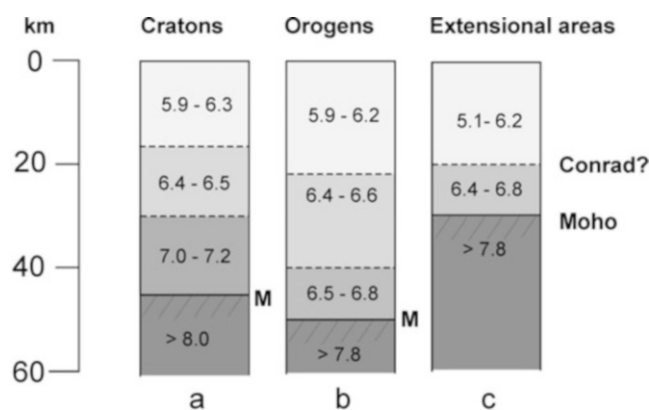
In the last 20 years, much progress has been made in the development of geophysical techniques that increased our knowledge about the variability of crustal structure substantially (see ► [“Deep Seismic Reflection and Refraction Profiling”](#)). A big misunderstanding, still often presented today, is the use of the phrase “the” crust. This is a marked oversimplification. Continental crust has been found to be extremely different (Christensen and Mooney 1995; Mooney 2015; Rudnick and Gao 2014; Hacker et al. 2015). There are at least three main types to be distinguished (see ► [“Lithosphere, Continental”](#)).

1. Thick, old cratons, shields, and platforms
2. Orogens, often related to continental subduction
3. Extensional areas, including shelves and rifts

Thick Precambrian cratons, disrupted by early plate tectonics, are presently found in all continents; they contribute more than 70% to the continental lithosphere. They are characterized by a thick crust (45–50 km) with high velocities (>7.2 km/s) in the lower part. The two other crustal types were generated by modern plate tectonics: the *orogens* by convergence and plate collisions, most of them showing deep crustal roots, and the *extensional areas* by divergence and wrench tectonics, often accompanied by thermal and magmatic processes. They generally show a shallow crustal thickness without a high-velocity lower crust. Special forms are rifts, basins, and igneous provinces.

Type sections of the three main crustal structures are presented in Fig. 1.

In general, the continental crust has lower seismic velocities and consists of lighter material than the underlying mantle; V_p is smaller than 7.8 km/s, V_s smaller than 4.3 km/s, and ρ below 3.1 g/cm^3 . Crustal thickness is between 20 km (shelves) and more than 70 km (some orogens) (Mooney et al. 2005). Mohorovičić was the first to determine the base of the crust, today called the “Moho” for short. It is observed worldwide and is clearly recognized by seismic methods. It is worthy to note that the position of the seismic Moho is not always identical with the petrological Moho (defined as the boundary between non-peridotitic crustal rocks and olivine-dominated rocks). In case that mafic rocks are transformed into eclogites at the base of orogenically thickened crust, P-wave velocities of eclogites largely overlap with those of peridotites so that seismic field studies may not detect eclogites as crustal rocks (e.g., Mengel and Kern 1992; O’Reilly and Griffin 2013). An intracrustal boundary – the “Conrad” – seems to separate upper sialic from more mafic lower crust in some areas. Transport of crustal material into the mantle and delamination of the lithosphere and the lowermost part of a thickened crust may give rise to a reduction of crustal thicknesses.



Continental Crustal Structure, Fig. 1 Type sections of the three main crustal structures (a–c) and ranges of P-wave velocities in the upper, middle, and lower crust. Numbers are velocities in km/s

Methods

Reliable studies of the continental crust were started in the last 40 years by seismic methods, applying first refraction studies with man-made explosions and lines of simple geophones. Wide-angle reflection seismics followed, and steep-angle reflection seismics became more and more important (see ► “[Deep Seismic Reflection and Refraction Profiling](#)”). In the last 30 years, digital technology, computer programs for ray tracing and tomography, and large, national, and international research programs with an increased number of recording devices and energy sources were initiated. Large, national reflection programs, often with a substantial support by the industry or large funding agencies, were developed in nearly all industrialized countries, spearheaded by COCORP in the USA. Both methods – refraction-wide-angle and steep-angle reflection – complement each other, the former supplying velocity-depth functions and relations to petrology and the latter providing reflectivity images of the subsurface that reveal past tectonic processes. Both methods form the backbone of crustal interpretation. They confirmed identical Mohos, either by a jump of velocities or by termination of (crustal) reflectivity because the mantle is less reflective (Meissner and Brown 1991). Vibroseis and dynamite shooting are used for reflection programs, the latter one generally providing a higher energy output, which is sometimes used for piggy-back wide-angle work. In the last 20 years, a relatively cheap seismological method, the “receiver function method” (see ► “[Seismic, Receiver Function Technique](#)”), supported interpretation of seismic data, making use of global seismicity and a wave change at seismic boundaries. This “receiver function method” is based on seismic phase conversions of teleseismic waves at the Moho and at boundaries in the crust and mantle.

Non-seismic geophysical techniques are complementary to seismic methods: gravity, for instance, for investigating isostatic problems or selecting between competing seismic interpretations or – together with magnetics – looking for inhomogeneities or controlling structure. Electromagnetic and magnetotelluric methods help to find and to interpret conductivity anomalies. Measurements of heat flow help to observe heat sources and decipher the thermal or magmatic history of a specific area. Depending on the heat flow, temperature-depth functions and viscosities that control rheology can be estimated. The brittle upper crust extends to a depth of 12–20 km and is a zone where generally earthquakes occur. It is separated from the more ductile, aseismic part of the middle and lower crust (some exceptions). Various experimental studies of creep, stress, and strength support seismic interpretation.

Cratons

Cratons are relatively stable Precambrian continental units with a rather deep lithosphere (Fig. 1a). They are called

shields when a sediment cover is lacking and platforms when covered by sediments. Today, cratons are found in all continents, possibly created from a few early supercontinents that were later disrupted by plate tectonics. No consensus exists about the number of cratons. They comprise up to ten in Africa and even more in North America. The latter were finally combined into the big “North American Craton” forming the whole interior of North America (without its margins in the west and east). Even “stable” cratons are influenced by Proterozoic and Phanerozoic processes like rifts or plumes. In North America there is a 2500-km-long Proterozoic midcontinental rift with a thick cover of sediments and volcanic intrusions and a reduced crustal thickness. There are also Paleozoic rifts in the southern part of the East European Craton, and there is an unusual 60-km-thick crust in the cold Baltic Shield, well within the eclogite stability field in the lowermost part. However, high velocities (> 7.8 km/s) corresponding to eclogitic rocks are not observed in the lowermost crust of the Baltic Shield (Korja et al. 2009).

The crystalline crust of cratons is similar worldwide. Cratons are characterized by low heat flow values. Their thickness is in the range 35–55 km with a “classical” three-velocity structure (Pawlenkova 1996), as shown by “deep seismic sounding” (wide-angle plus refraction studies). In general, cratons have an upper crust with V_p around 6 km/s, a middle crust around 6.5 km/s, and a roughly 10-km-thick lower crust with velocities around 7.0–7.2 km/s, sometimes 6.7–7.5 km/s (Fig. 1a). A typical example is the EUROBRIDGE profile (Guterch et al. 1997) crossing the crust of the East European platform (Fig. 2).

Reflection seismic experiments generally show no special enhancement of reflectivity in the (mafic) lower crust, as shown by the large transects of LITHOPROBE in Canada,

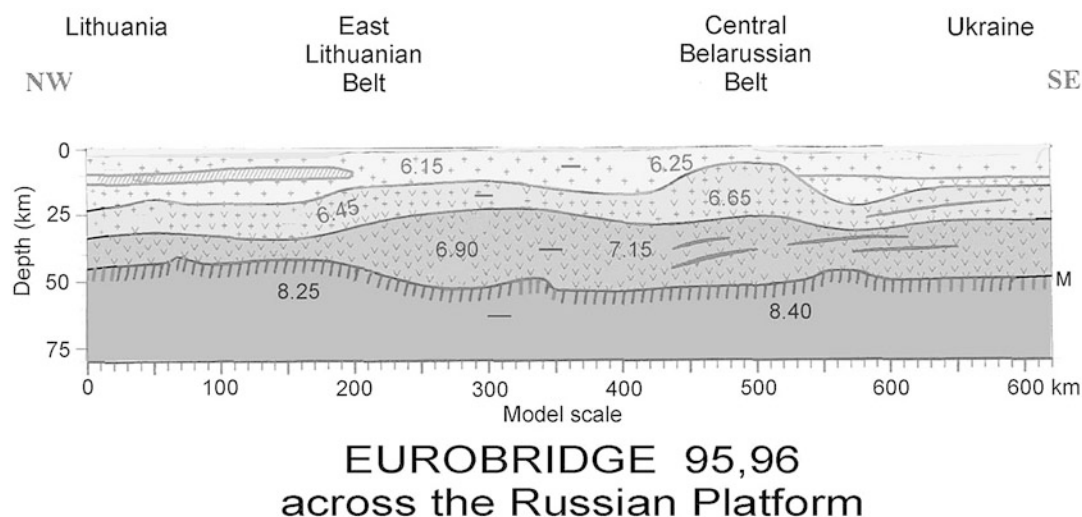
by the COCORP data in the USA, or by the BABEL surveys in the northern Baltic Sea (some exceptions).

Orogens

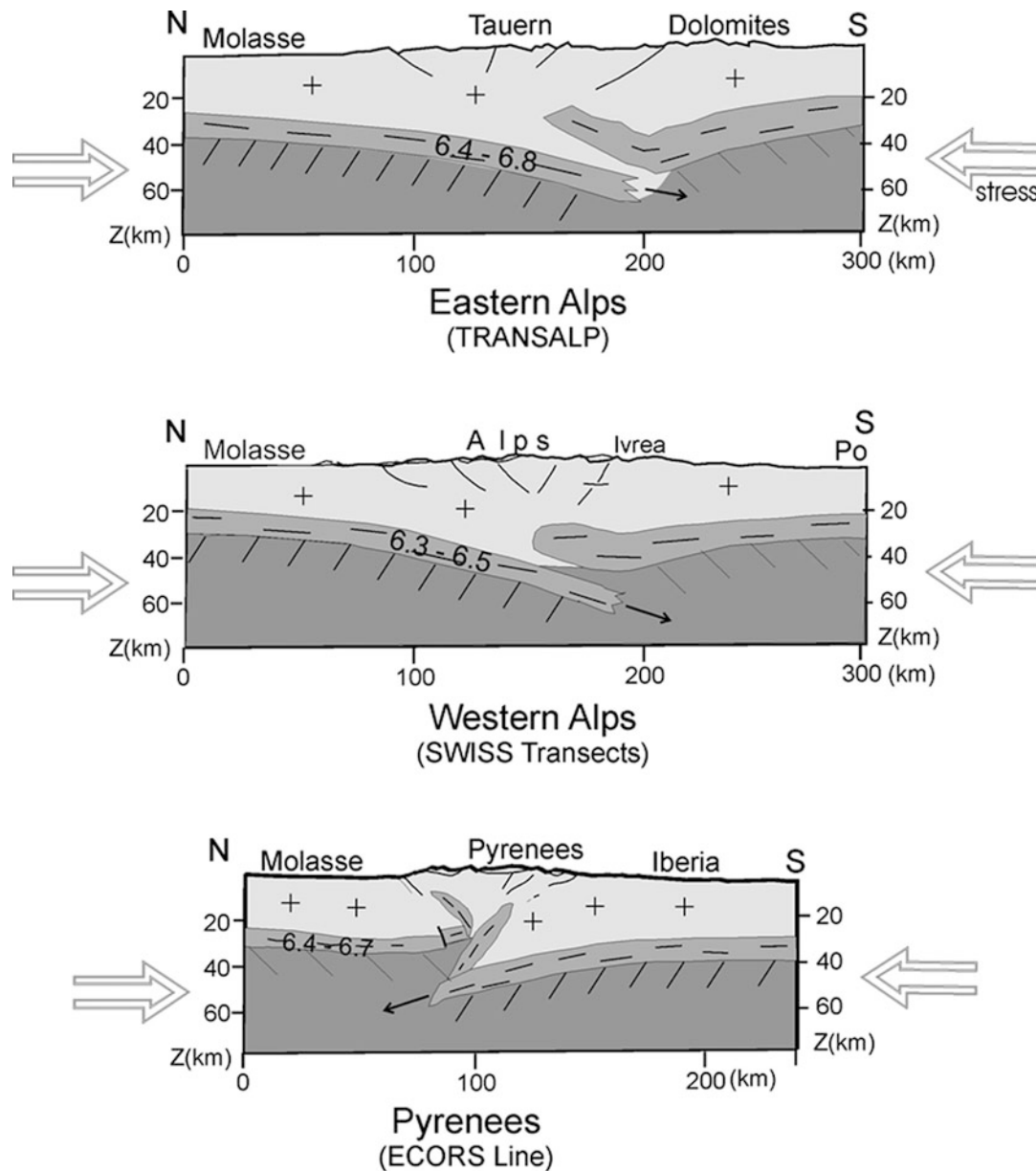
Collisions of plates or terranes create orogens by compressing material upward (mountain ranges) and downward (roots), and dynamic processes initiate oceanic or continental subduction. While the elevation of an orogen apparently depends on the time and speed of the colliding plates (e.g., Himalaya, Tibet), processes of crust/mantle interaction are quite complex. Magmatic processes are found behind subducting plates in many orogens (Alps, Sierra Nevada, Andes). In continental collision zones, rock types of the lower crust resemble those of the neighboring plates, guided into the orogens. There are high-velocity (cratonic) crustal rocks with a mafic lower crust in the Uralides (from the East European platform) and sialic, extensional crusts with a reflecting lower crust in the Alps (mostly from the Variscides). Continental subduction often follows oceanic subduction. Remnants of these processes may be reflected by the crustal “roots” (see ► “[Seismic, Reflectivity Method](#)”). Figure 3 shows simplified cross sections through the Alps and Pyrenees. Orogens are more or less isostatically compensated, and its Bouguer anomaly is strongly negative (Ebbing 2004).

Extensional Areas

Extensional processes cover many dimensions: from rifts, basins, (the basin and range of North America), and continental shelves up to large Phanerozoic areas like



Continental Crustal Structure, Fig. 2 Typical cratonic crust from the East European platform (after Guterch et al. 1997)



Continental Crustal Structure, Fig. 3 Typical sections of orogenic crust (Alps and Pyrenees)

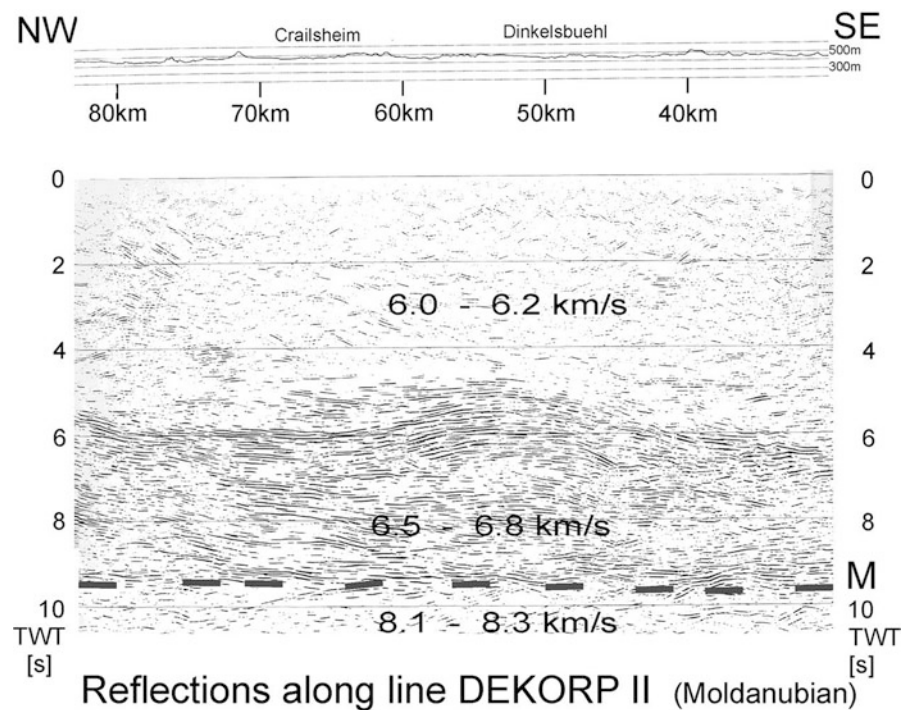
Europe and west of the Teisseyre-Tornquist line (see ► [“Trans-European Suture Zone”](#)). This large extensional area was created in post-Avalonian and post-Variscan time at the end of the Paleozoic. A surprisingly uniform crust has been formed, showing three prominent differences compared to the cratons: their thickness is only 30 km (Fig. 1c), their velocities are lower than 6.8 km/s, and in many places, they show prominent reflectivity (lamellae) in the lower crust.

Apparently, all three characteristic crustal patterns are genetically interrelated. While a former lower crust (in the basin and range province and in the Variscides) got lost in the previous orogeny (by delamination or subduction),

subsequent moderate, but widespread, extension dominated. Mafic-ultramafic dykes intruded into the lower crust and formed sill-like intrusions, giving rise to high impedance contrasts after cooling. Importantly, sills are mostly rather thin and do not generally increase the lower-crustal overall (sialic) seismic velocities. Figure 4 gives an example of strong lower crust reflectivity and low average velocities.

In contrast, the formation of rifts and basins is a result of local (extensional) stress and/or strong local heat anomalies. As a consequence, large volumes of mantle-derived basaltic magmas intrude into the crust and after cooling the mafic rocks (gabbros or metamorphic equivalents) cause the

Continental Crustal Structure,
Fig. 4 Multiple reflections
 (laminations) in the Moldanubian
 lower crust along the DEKORP II
 line, along with ranges of V_p
 velocities at different depths



observed high velocities at the base of most rifts and basins. Examples are the Oslo Graben, the Danish Basin and other adjacent rifts and basins north of the Variscan front, the great East African Rift, the Donetz Basin, and the midcontinental North American rift. In general, the massive intrusions created a transparent lower crust, but some basins with high lower crust velocities (above 7.2 km/s) show some thin lamellae (apparently ultramafic) on top of the Moho. Also in some areas of continental/ocean transition, high-velocity lamellae are observed in the lower crust (Roberts et al. 2009).

In general, lower crustal reflectivity seems to depend on the volume and thickness of the mafic intrusions compared to the host rock. Whereas sill-like intrusions are suggested to generate the observed lower crustal reflectivity, massive intrusions in the lower crust reduce reflectivity and enhance the average velocities to values of 6.8–7.2 km/s. The volcanic Messum ring complex in Namibia shows no lamellae in the high-velocity interior but a good reflectivity in the surrounding (Bauer et al. 2003). In contrast, the Viking Graben in the North Sea shows lower crust reflectivity and no reflection around it (McBride et al. 2004).

Most continental shelves show extension and a stretched continental crust with a thickness of only 20–30 km, often containing aborted rifts (Korja et al. 2009). Reflectivity in the lower crust is found in the West Atlantic margin (Faroe or Hatton) and along the Greenland coast. Often, seaward dipping reflections, caused by intrusive and extrusive magmatism, mark the onset of oceanic crust with high velocities. BIRPS WAM – profile from the British Isles toward the

Atlantic Ocean – shows a preferred lower crust reflectivity below the (continental) Mesozoic basins, while the oceanic crust is free from reflections.

Constraints on Crustal Petrology

The measured seismic velocity structures are very helpful in providing a rough characterization of the Earth's crust in different tectonic environments, but they are nonunique so that inferences about composition cannot be drawn from wave velocities (at least P-wave velocities). The seismic properties at depth are determined by a number of lithologic and physical factors that control the in situ rock properties in a very complex manner. Velocities are controlled by the intrinsic properties of the rocks (mineralogical composition, chemical composition, metamorphic grade, crystallographic preferred orientation of constituent minerals, etc.) and by the physical environment of the crust (temperature, pressure, porosity, fluid content, etc.) (see ► “Seismic Properties of Rocks”).

Direct information about the composition and structure of the deep crust can be obtained either from crustal terrains exposed at the surface (e.g., Fountain and Salisbury 1981; Kern and Schenk 1988) or from xenoliths brought to the surface by magmas (Downes et al. 1990). Deeply eroded Precambrian terrains (e.g., Sri Lanka) and upthrust tectonic slices in orogenic belts (e.g., Ivrea Zone, N. Italy; Serre Mountains, S. Calabria; Kapuskasing Zone, Canada)

provide perhaps the best geologic guides to structural style and composition at depth. Such rocks are important in providing direct data although they contain a mixed message as they are no longer the in situ deep crust.

A unique ability to correlate the seismic data with the structure and composition of the in situ deep crust can be provided by coupling experimentally determined or calculated P- and S-wave velocities for relevant crustal rocks collected from surface outcrops or from xenoliths, simulating in situ conditions: (1) by laboratory seismic measurements at actual PT conditions (e.g., Christensen and Wepfer 1989; Kern et al. 1999; Ji et al. 2015) and (2) by calculations from modal analyses and elastic properties of the rock-forming minerals and their pressure and temperature derivatives (e.g., Jackson et al. 1990; Barruol and Kern 1996; Almquist and Mainprice 2017).

High P-wave velocities (>6.9 km/s) generally defining the lowermost crust (below about 20–25 km depth) are typical for metamorphic rocks in the granulite facies. These are mafic granulites, mafic amphibolite facies rocks, anorthosites, and high-grade metapelites.

Intermediate P-wave velocities (6.5–6.9 km/s) represent the upper-lower to mid-crust and are characteristic for intermediate granulites and metapelites.

Low P-wave velocities (6.2–6.5 km/s) occurring at mid-crustal levels in Paleozoic and more recent fold belts and in the uppermost crust in shields and platforms are likely to be composed of felsic rocks in the amphibolite and granulite facies. It should be noted, however, that not all felsic (granitic) upper crustal rocks are intermediate- to high-metamorphic grade. The Superior Province (Canada) is an example of low metamorphic shield rocks.

Mineralogically, feldspar (K-feldspar, plagioclase) is the most abundant mineral, followed by quartz and hydrous minerals (such as mica and amphiboles).

The bulk chemical composition of the upper crust is felsic, and the middle crust is intermediate in composition. The average composition of the lower crust is in general mafic in cratons, but it tends to be intermediate or even felsic in some regions. The continental crust as the whole (average of all crustal types) is suggested to be intermediate in average bulk composition (Rudnick and Fountain 1995; Mooney 2015).

Importantly, most of the rocks constituting the Earth's crust exhibit marked anisotropy of elastic properties (e.g., Kern et al. 2001; Ji et al. 2002; Almquist and Mainprice 2017). Typical values of intrinsic P-wave anisotropy (quartz-mica-schists, felsic gneisses, granulite-facies metapelites, amphibolites) range from 5.4% to 10.7%. Anisotropy may be caused by crystallographic (CPO) and morphological (or shape) preferred orientation (SPO), by oriented microcracks or by thin layers of otherwise isotropic materials with different properties (see ► “Magnetic Anisotropy”).

Whereas oriented cracks may contribute a large fraction of observed seismic anisotropy in the uppermost crust (Crampin 1987; Crampin and Gao 2018), CPO- and SPO-related seismic anisotropy is a major contribution in ductily deformed and foliated crustal structures (shear zones, gneiss sections) of the deeper crust. It is mainly due to the alignment of mica and hornblende minerals and their strong single crystal anisotropies (Barruol and Kern 1996; Meissner and Kern 2008). In addition to mineral and chemical composition and rock fabric, pore fluids may have an important bearing on in situ seismic properties. Saturation of pore space ($P_{pore} \approx 0$) increases V_p , whereas V_s remains unaffected (Nur and Simmons 1969). In case of pore pressure approaching lithostatic pressure ($P_{pore} \approx P_{lith}$), microcracks and grain boundaries are more or less kept open. As a consequence, both P- and S-wave velocities are smaller than in dry rocks, due to a reduction of the effective pressure ($P_{eff} = P_{lith} - n \times P_{pore}$; $n \approx 1$). Importantly, V_p anisotropy is significantly higher in dry rocks than in water-saturated rocks at atmospheric pressure conditions (Popp and Kern 1994). As pressure is increased, differences progressively decrease. The corresponding S-wave velocities and S-wave splitting data are only weakly affected by intergranular fluids.

Seismic anisotropy is becoming increasingly important also in understanding the evolution of the Earth's crust (and upper mantle). Since the measured seismic anisotropy has a structural (macro- and/or micro-) origin, seismic azimuthal anisotropy manifested by shear wave splitting, in particular, can provide important geophysical evidence of deformation because the orientation and magnitude of anisotropy are, in general, strongly related to the internal strain-induced rock fabric. Although variations of lithology are suggested to be most important for the generation of seismic reflections in the lower crust (laminated lower crust) (e.g., Meissner et al. 2006), seismic reflectivity (Warner 1960) may be enhanced by the fabric-related anisotropy of ductily deformed structures (Kern and Wenk 1990; Vasin et al. 2017). It should be noted, however, that in crustal rocks, the variability due to variation in composition is generally higher than the variation caused by velocity anisotropy (except for schists). The reverse is true in the olivine-rich upper mantle rocks. This probably implies that on a regional scale, fabric-related anisotropy is more important in the oceanic and continental upper mantle than in the continental crust.

Summary

Seismic refraction and reflection surveys have revealed three main types of continental crust: (1) the thick old cratons, constituting large parts of all continents; (2) the orogens, generated by plate collisions; and (3) the extensional areas with a relatively thin crust, exhibiting ample reflectivity

(lamellae) in the lower part. Crustal thickness is between 20 km (shelves) and more than 70 km (some orogens). The Mohorovičić discontinuity (Moho), defined by a jump in velocity and density, separates the lowermost crust from the upper mantle. High P-wave velocities (>6.9 km/s) are typical for the lowermost crust of platforms and orogens; they stand for high-grade mafic rocks in the granulite facies. In the thin lower crust of extensional areas, high-velocity mafic rocks are almost lacking. The upper-lower to mid-crust is characterized by intermediate P-wave velocities (6.5–6.9 km/s) and is likely to be composed of intermediate granulites and metapelites. Low P-wave velocities (6.2–6.5 km/s) of mid-crustal levels in Paleozoic and more recent fold belts as well as in the uppermost crust in shields and platform point to felsic rocks in the amphibolite and granulite facies. Seismic anisotropy is an important property of most rocks constituting the Earth's crust. It is basically caused by crystallographic and morphological (or shape) preferred orientation, by oriented microcracks, or by thin layers of otherwise isotropic material with different properties.

Cross-References

- [Continental Drift](#)
- [Geodynamics](#)
- [Lithosphere, Continental](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Imaging, Overview](#)
- [Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting](#)
- [Seismic Tomography](#)
- [Seismicity, Subduction Zone](#)
- [Subduction Zones](#)

Acknowledgments We greatly appreciate the assistance of Gudrun Reim (Kiel) in preparing the figures, and we thank Walter Mooney (Menlo Park) for his review, leading to many improvements of the manuscript.

Bibliography

- Almqvist BSG, Mainprice D (2017) Seismic properties and anisotropy of the continental crust: predictions based on mineral texture and rock microstructure. *Rev Geophys* 55. <https://doi.org/10.1002/2016RG000552>. 524
- Barruol G, Kern H (1996) Seismic anisotropy and shear-wave splitting in lower-crustal and upper-mantle rocks from the Ivrea Zone – experimental and calculated data. *Phys Earth Planet Inter* 95:175–194
- Bauer K, Trumbull RB, Victor T (2003) Geophysical images and a crustal model of the intrusive structures beneath the Messum ring complex, Namibia. *Earth Planet Sci Lett* 216:65–80
- Christensen NI, Mooney WD (1995) Seismic velocity structure and composition of the continental crust: a global view. *J Geophys Res* 100:9761–9788
- Christensen NI, Wepfer WW (1989) Laboratory techniques for determining seismic velocities and attenuations, with applications to the continental lithosphere. In: Pakiser LC, Mooney WD (eds) *Geophysical framework of the continental United States*. Geophysical Society of America Memoir, 172. Geological Society of America, Boulder
- Crampin S (1987) The geological and industrial implications of extensive-dilatancy anisotropy. *Nature* 328:491–496
- Crampin S, Gao Y (2018) Evidence supporting new geophysics. *Earth Planet Phys* 2:173–188. <https://doi.org/10.26464/epp2018018>
- Downes H, Dupuy C, Leyreloup A (1990) Crustal evolution of the Hercynian belt of Western Europe: evidence from lower crustal granulitic xenoliths. *Chem Geol* 68:291–303
- Ebbing J (2004) The crustal structure of the Eastern Alps from a combination of 3D gravity modeling and isostatic investigations. *Tectonophysics* 350:89–104
- Fountain D, Salisbury MH (1981) Exposed cross sections through the continental crust; Implications for the crustal structure, petrology and evolution. *Earth Planet Sci Lett* 56:263–277
- Guterch A, Grad M, Thybo H, Keller GR, POLONAISE Working Group (1997) POLONAISE '97 – an international seismic experiment between Precambrian and Variscan Europe in Poland. *Tectonophysics* 314:101–121
- Hacker BR, Keleman PB, Behn MD (2015) Continental lower crust. *Annual Review of Earth and Planetary Sciences* 43:167–205
- Jackson I, Rudnick RL, O'Reilly SY, Bezant C (1990) Measured and calculated elastic wave velocities for xenoliths from the lower crust and upper mantle. *Tectonophysics* 173:207–210
- Ji SC, Wang Q, Xia B (2002) *Handbook of seismic properties of minerals, rocks, and ores*. Polytechnic International Press, Montreal, 630 pp
- Ji S, Shao T, Michibayashi K, Oya S, Satsukawa T, Wang Q, Zhao W, Salisbury M (2015) Magnitude and symmetry of seismic anisotropy in mica- and amphibole-bearing metamorphic rocks and implications for tectonic interpretation of seismic data from southeast Tibetan Plateau. *J Geophys Res* 120. <https://doi.org/10.1002/2015JB012209>
- Kern H, Schenk V (1988) A model of velocity structure beneath Calabria, South Italy, on laboratory data. *Earth Planet Sci Lett* 87:325–337
- Kern H, Wenk H-R (1990) Fabric related velocity anisotropy and shear wave splitting in rocks from the Santa Rosa mylonite Zone, California. *J Geophys Res* 95:11213–11223
- Kern H, Gao S, Jin Z, Popp T, Jin S (1999) Petrophysical studies on rocks from the Dabie ultrahigh-pressure (UHP) metamorphic belt, Central China: implications for the composition and delamination of the lower crust. *Tectonophysics* 301:191–215
- Kern H, Popp T, Gorbatshevich F, Zharikov A, Lobanov KV, Smirnov YP (2001) Pressure and temperature dependence of V_p and V_s in rocks from the superdeep well and from surface analogues at Kola and the nature of velocity anisotropy. *Tectonophysics* 338:113–134
- Korja A, Hyönen T, Tira T, Heikkinen P (2009) Examining three-dimensional crustal heterogeneity in Finland. *Eos Trans AGU* 90(15):129–130
- McBride JH, White RS, Smallwood JR, England RW (2004) Must magmatic intrusion in the lower crust produce reflectivity? *Tectonophysics* 388:271–297
- Meissner R, Brown L (1991) Seismic reflections from the Earth's crust: comparative studies of tectonic patterns. *Geophys J Int* 105:1–2
- Meissner R, Kern H (2008) Earthquakes and strength in the laminated lower crust – can they be explained by the corset model? *Tectonophysics* 448:49–59

- Meissner R, Rabbel W, Kern H (2006) Seismic lamination and anisotropy of the lower continental crust. *Tectonophysics* 416:81–99
- Mengel K, Kern H (1992) Evolution of the petrological and seismic Moho – implications for the continental crust-mantle boundary. *Terra Nova* 4:109–116
- Mooney W (2015) Crust and lithospheric structure – global crustal structure. In: *Treatise of geophysics*, vol I, 2nd edn, pp 339–390
- Mooney WD, Rao VV, Chulik GS, Detweiler ST (2005) Comparison of the deep crustal structure and seismicity of North America with the Indian Subcontinent. *Curr Sci* 88:1639–1651
- Nur A, Simmons C (1969) The effect of saturation on velocity in low-porosity rocks. *Earth Planet Sci Lett* 7:183
- O'Reilly S, Griffin W (2013) Moho vs crust-mantle boundary: evolution of an idea. *Tectonophysics* 609:535–546
- Pawlenkova N (1996) Crust and mantle structure in Northern Eurasia from seismic data. In: Dmowska R, Saltzman B (eds) *Advances in geophysics*, vol 37. Academic, San Diego
- Popp T, Kern H (1994) The influence of dry and water-saturated cracks on seismic velocities of crustal rocks – a comparison of experimental data with theoretical model. *Surv Geophys* 15:443–465
- Roberts AW, White RS, Christie PAF (2009) Imaging igneous rocks on the North Atlantic rifted continental margin. *Geophys J Int* 179:1029–1038
- Rudnick RL, Fountain DM (1995) Nature and composition of the continental crust: a lower crustal perspective. *Rev Geophys* 33:267–309
- Rudnick R, Gao S (2014) Composition of the continental crust. In: *Treatise on geochemistry*, vol 4, 2nd edn. Elsevier, Amsterdam, pp 1–51
- Vasin RN, Kern H, Lokajicek T, Svitek T, Lehmann E, Mannes DC, Chaousche M, Wenk H-R (2017) Elastic anisotropy of Tambo gneiss from Promontogno, Switzerland: a comparison of crystal orientation and microstructure-based modelling and experimental measurements. *Geophys J Int* 209:1–20
- Warner M (1960) Absolute reflection coefficients from deep seismic reflections. *Tectonophysics* 173:15–23

Continental Drift

Alan G. Smith
(Deceased)

Definition

Continental drift The name given to the relative movement between continents.

Introduction

Apparently, continental drift was first postulated by Abraham Ortelius, who made one of the first atlases, in the third edition of his *Thesaurus Geographicus* (Ortelius 1596) to account for the similarity between the coastlines of western Africa, western Europe, and the eastern Americas. During the next three

centuries or so, several other thinkers used this morphological similarity to come to the same conclusion.

Alfred Wegener

Although Ortelius speculated that earthquakes and floods had torn the continents apart, Alfred Wegener, an Austrian meteorologist, was the first to systematically gather the geological evidence for continental drift in his synthesis (Wegener 1929a), also translated into English (Wegener 1929b), parts of which had been published as early as 1912 (Wegener 1912).

Wegener was puzzled in particular by the present-day distribution of former ice-age deposits, or tillites, of Permo-Carboniferous age, now known to be about 300 million years old that are found today in South America, Africa, Antarctica, India, and Australia. Wegener's meteorological background led him to assume that the present-day climatic zones, with cold polar regions and a hot tropical belt, was a fundamental property of the Earth's atmosphere that had been established before these glacial deposits had formed and had persisted to the present-day. He realized that if all the southern continents had been joined together to form a supercontinent lying near the south geographic pole, then the *present-day* distribution of all the Permo-Carboniferous tillites would have a logical explanation. This supercontinent is known as Gondwana. But Wegener went further and postulated that the northern continents had also been joined to form a northern supercontinent known as Laurasia, which, with Gondwana, formed a huge continental area incorporating all the major continents, known as Pangea.

Wegener's solution to the tillite distribution had a compelling elegance about it, partly because it also placed the Carboniferous forests of Europe and North America, (whose compressed remains gave the name to the Carboniferous period) in what would have been the tropical region of that time. However, elegance is not a scientific proof, and Wegener's ideas were rejected by most geophysicists and many geologists, principally because there were no physical measurements that demonstrated drift and there was also no known *physical* mechanism that could move continents independently of one another (e.g., Jeffreys 1929).

Wegener (1929a) made the reassembly by positioning the continents at where he presumed the edges of the continents lay, that is, somewhat beyond the present-day coastlines, but his maps were rather schematic. A more convincing Gondwana map, supported by much geological evidence, was made by Du Toit (1937), but the best precomputer map showing the excellence of the fit between the continental edges of South America and Africa was made by Carey (1958). Jeffreys (1964) denied that there was a fit, and it was his denial that in part led Bullard and others to examine how well the continental edges fitted

together (Bullard et al. 1965; Everett and Smith 2008). These ideas were supported by the changes in the Earth's magnetic field preserved in rocks (see ► [“Paleomagnetism, Principles”](#)) determined by workers such as Irving (1964), but the results were *not fully accepted by the geoscience community* until after the plate tectonic revolution.

Conclusions

It is now established that the Earth's lithosphere is divided into a number of rigid *Plate Tectonics, Precambrian* in relative motion. Tracing these motions back in time shows that Gondwana, Laurasia, and Pangea did exist in the past, *vindicating Wegener's idea completely*.

Cross-References

- [Lithosphere, Continental](#)
- [Lithosphere, Continental: Thermal Structure](#)
- [Lithosphere, Mechanical Properties](#)
- [Lithosphere, Oceanic](#)
- [Paleomagnetism, Principles](#)
- [Plate-Driving Forces](#)
- [Plate Motions in Time: Inferences on Driving and Resisting Forces](#)
- [Plate Tectonics, Precambrian](#)

Bibliography

- Bullard E, Everett JE, Smith AG (1965) The fit of the continents around the Atlantic. *Philos Trans R Soc Lond A* 258:41–51
- Carey SW (1955) The orocline concept in geotectonics – part I, Publication no. 28. The papers and proceedings of the Royal Society of Tasmania, vol 89, pp 255–288
- Carey SW (1958) A tectonic approach to continental drift. In: Carey SW (ed) *Continental drift: a symposium*. University of Tasmania, Hobart, pp 177–355
- Du Toit AL (1937) *Our wandering continents*. Oliver & Boyd, Edinburgh
- Everett JE (1965) The fit of the continents around the Atlantic. PhD, thesis, Cambridge
- Everett JE, Smith AG (2008) Genesis of a geophysical icon: the Bullard, Everett and Smith reconstruction of the circum-Atlantic continents. *Earth Sci Hist* 27(1):1–11. <http://pubs.usgs.gov/gip/dynamic/historical.html>
- Irving E (1964) *Palaeomagnetism and its application to geological and geophysical problems*. Wiley, New York, p 399
- Jeffreys H (1929) *The Earth*. Cambridge University Press, Cambridge
- Jeffreys H (1964) How soft is the Earth? *Q J R Astron Soc* 5:10–22
- Ortelius A (1596) *Thesaurus geographicus, recognitus et auctus*. Ex Officina Plantiniana, Antwerp
- Ortelius A (1603) *Theatrum orbis terrarum*. Abridged edition printed by James Shawe, London
- Smith AG, Hallam A (1970) The fit of the southern continents. *Nature* 225:139–144
- Wegener AL (1912) Die Entstehung der kontinente (the origin of continents). *Geol Rundsch* 3:276–292
- Wegener AL (1929a) Die Entstehung der kontinente und Ozeane. (The origin of continents and oceans). Brunswick, Vieweg
- Wegener AL (1929b) The origin of continents and oceans (English translation). Methuen, London

Continental Rifts

A. M. Celâl Şengör

Faculty of Mines, Department of Geology and Eurasia
Institute of Earth Sciences, Istanbul Technical University,
Istanbul, Turkey

Synonyms

Graben; Taphrogen

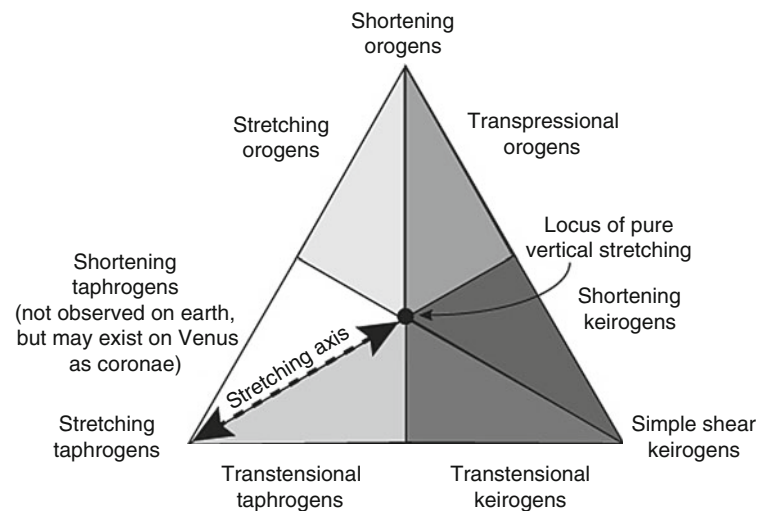
Definition

A continental rift (Gregory 1894; Quennell 1982, 1985) is a fault-bounded elongate trough under or near which the entire thickness of the lithosphere (► [“Lithosphere, Continental”](#) and ► [“Lithosphere, Mechanical Properties”](#)) has been reduced in extension during its formation. Just as old mountain ranges may no longer have any topographic expression because of tectonic and/or erosional events, some, especially old, rifts may no longer appear as troughs for the same reasons, but their original trough shape is recognized by their stratigraphically younger fills, or metamorphically lower grade of their down-dropped central blocks (► [“Sedimentary Basins”](#)).

Introduction

Rifts form one of the three main categories of lithospheric-scale structures resulting from differential motion of parts of the lithosphere. Lithospheric shortening creates orogens, simple shear creates keirogens, and stretching creates taphrogens that are collections of rifts (Fig. 1). However, at present only 20.5% of the active plate boundaries show normal convergence and 21% normal divergence. Some 8% show pure dextral strike-slip and some 6% pure sinistral strike-slip. The remaining 58.5% display some deviation from the three end-members, with relative motion vector angles to boundary strikes varying between 0° and 67° (Woodcock 1986). Plate boundaries must have shown a similar behavior in the past, so only about half of all the rifts a geologist may encounter is likely to show normal extension (► [“Plate Motions in Time: Inferences on Driving and Resisting Forces”](#) and ► [“Plates and Paleoreconstructions”](#)).

Continental Rifts, Fig. 1 A ternary diagram of strain showing the three lithospheric megastructure categories resulting from various strains



Rifts are important structures from the viewpoints of our understanding of the behavior of our planet and our exploitation of its resources. They record evidence of continental fragmentation in diverse tectonic settings including all three types of plate boundaries and in plate interiors (► [“Plate Motions in Time: Inferences on Driving and Resisting Forces”](#) and ► [“Plates and Paleoreconstructions”](#)). Also, at different stages of their evolution they present opportunities of studying the continental crust (and in extreme cases even the upper mantle) from its surficial sedimentary rocks down to the crust-mantle interface (► [“Lithosphere, Continental”](#)). Especially the lacustrine sedimentary sections of rifts are useful for studies on the past climates (e.g., Olsen and Kent 1999; Kravchinsky et al. 2003; Felton et al. 2007), and they have enabled geologists to refine stratigraphic correlations down to thousands of years as far back as in the early Mesozoic (e.g., Olsen and Kent 1999).

Rifts house important economic reserves such as hydrocarbons (the Sirt rift in Libya being the most productive with about 45 billion barrels of oil and 33 trillion cubic feet of gas; for the hydrocarbon potential of rifts, see: Harding 1983; MacKenzie and McKenzie 1983: a classic on hydrocarbon genesis in rifts; Ziegler 1994a; Lambiase 1995), hydrothermal and stratiform copper, copper-nickel, molybdenum, giant lead-zinc, and even uranium deposits (for metallogenesis in rifts, see Sawkins 1990). Most of the geothermal areas in the world are located in active rifts (e.g., Jaffé 1971, Fig. 1) (► [“Geothermal Heat Pumps”](#)). At present 1/5 of the entire fresh-water reserve of the Earth resides in a rift basin, namely, Lake Baykal. Rifts are thus significant not only for increasing our geological knowledge, but also for contributing to our well-being. It is perhaps of some interest in this context to remember that our genus and species were born in a rift, namely, in the East African.

For good overviews of rifts and rifting, see Burke and Dewey (1973), Burke and Wilson (1976), Coward et al.

(1987), McKenzie (1978 a “must” reading for understanding the evolution of rifts), Manspeizer (1988), Ziegler (1992, 1994b), Landon (1994), Şengör (1995), Şengör and Natal’in (2001: a detailed catalogue of the world’s rifts with tabulated summaries of origin and history and references), and Ring and Wernicke (2009). For the geology of some of the classic rift regions of the world, with increasing age, see: Africa as a whole: Kampunzu and Lubala (1991), Burke (1996), and Burke and Gunnell (2008). East Africa: Frostick et al. (1986), Frostick (1997: a good introduction for beginner), Schlüter (1997; this book has a bibliography of some 1300 items), Morley (1999; has an excellent collection of seismic sections). Basin-and-Range: Wernicke (1990), Beratan (1996), Faulds and Stewart (1998), Snow and Wernicke (2000), Dickinson (2002: best introduction for an outsider), and McQuarrie and Wernicke (2005; best kinematic reconstruction of any rift I know). Rio Grande rift, USA: Ingersoll (2001). The Upper Rhine rift: see the special issue on it in the *International Journal of Earth Sciences (Geologische Rundschau)*, v. 94 (2005); also: Hüttner (1991), Schumacher (2002), Rotstein et al. (2006). The Lake Baykal rift: Crane et al. (1991), Kuz’min et al. (2001), ten Brink and Taylor (2002), Kravchinsky et al. (2003). The West Siberian Basin: Surkov et al. (1994). The North Sea Basin with the North German (Lower Saxony) Basin: Blundell and Gibbs (1990), Evans et al. (2003; this lavishly illustrated and richly documented mammoth book makes all earlier publications on the subject either obsolete or redundant; also available as a two CD-ROM set), Wilson et al. (2004), Littke et al. (2008). The Mediterranean region: Durand et al. (1999), Corti et al. (2006; an excellent paper about a little-known rift cluster).

Willis (1936, 1–30) presents a good historical review of ideas pertaining to rifting. For some of the important early ideas on rifting, see Şengör (2018). For an excellent recent review of ideas and studies on the East African rifts, see Mohr (2009). For a history of uplift-related rift ideas, see Şengör

(2003). For a historical review of recent developments in our understanding of extensional tectonics, see Wernicke (2009). Quennell (1982, 1985) are anthologies of some historical papers on rift studies. They are not satisfactory, but are the only ones of their kind.

Rift, Graben, and Taphrogen

“Rift” and “graben” are two words used interchangeably in the international literature designating elongate troughs bounded by normal faults. However, not all such troughs are parts of structures that rupture, or even thin by rock creep, the entire lithosphere. Some normal-fault-bounded troughs result from sliding and stretching of surficial rocks over lubricating sedimentary horizons such as evaporites or shales. The normal-fault bounded basins that formed during a landslide associated with the 27 March 1964 Prince William Sound earthquake in Alaska or the similar troughs in the Canyonlands National Park in Utah formed by sliding and stretching the sedimentary section above the Upper Carboniferous evaporites are examples. Such structures have nothing to do with lithospheric stretching and thus they do not fit the definition of rift given above. Şengör (1995) therefore suggested that the word rift be confined to those structures that actually disrupt or otherwise (by creep) thin the lithosphere and graben be reserved for those that do not. His usage is what is recommended and followed in this entry. In rifts when the extension factor β , the ratio of extended width to unextended width (McKenzie 1978), goes beyond 3 the continental lithosphere generally ruptures completely and ocean generation begins by sea-floor spreading (Le Pichon and

Sibuet 1981) (► “Seafloor Spreading”) either by dyke injection or by cold extrusion of subcontinental mantle in the form of ultramafic “core complexes” as seen in such places as along the boundary between the Upper Penninic and the Lower Austroalpine units of the Alps (Bernoulli et al. 2003 and the references therein) or along the South Galicia Bank (Boillot et al. 1987). For obvious reasons, this cannot happen to grabens; however, much they may be extended.

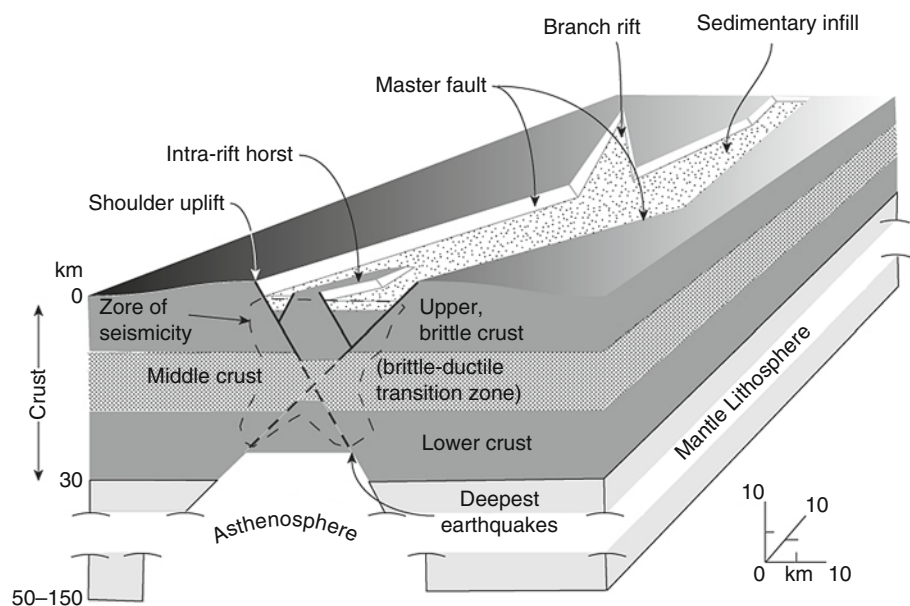
Large regions of localized extension exist on our planet where many rifts and grabens occur together. It is inappropriate to call such regions simply rifts. It would be like calling whole orogens simply “thrusts.” These large regions of extension are the extensional counterparts of orogens (Figs. 1 and 7) and therefore it is appropriate to use for them the term “taphrogen” derived from Krenkel’s term “taphrogeny.”

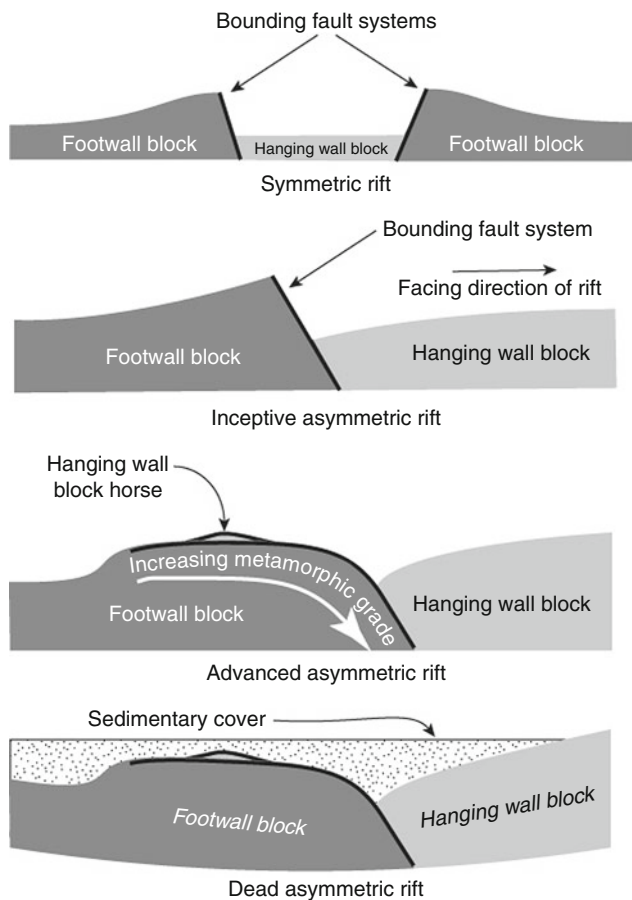
General Properties of Rifts

Structure of Rifts

Figure 2 is an idealized and highly simplified block diagram of a rift. Rifts have two fundamental structural entities: a main hanging wall block and one or two footwall blocks depending on whether the rift is symmetric (as in Fig. 2) or asymmetric (see Fig. 3). In symmetric rifts, the extension represented by the rift is almost entirely localized between two master bounding normal (or oblique extensional) fault systems and it is these two systems that take up much of the extension, although the hanging wall block delimited by them is also intensely broken up by smaller normal faults with strikes largely parallel with those of the major bounding fault families (Fig. 4). The hanging wall is commonly dropped down

Continental Rifts, Fig. 2 A highly schematic block diagram of an “ideal” rift drawn to emphasize the main architectural elements of rifts. The straight marginal faults, for example, do not mean that all rifts must have straight faults or that they are single faults instead of fault families



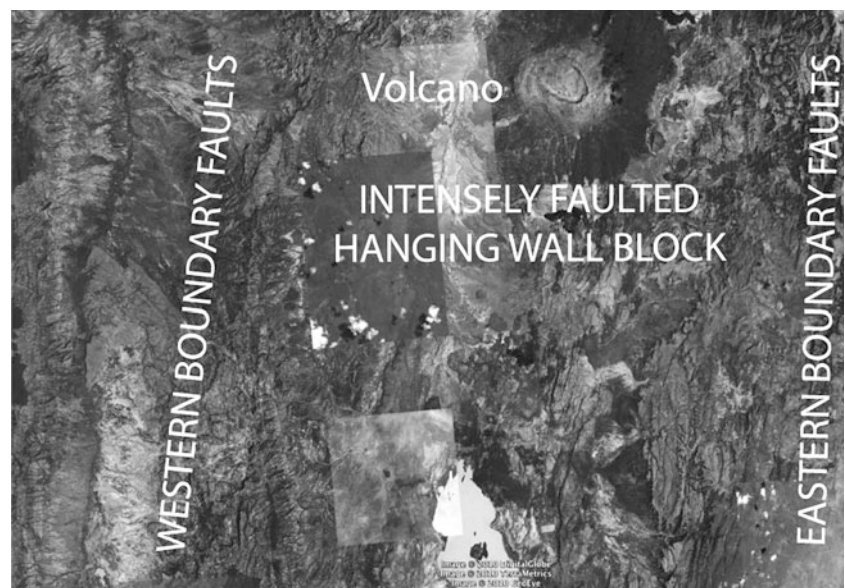


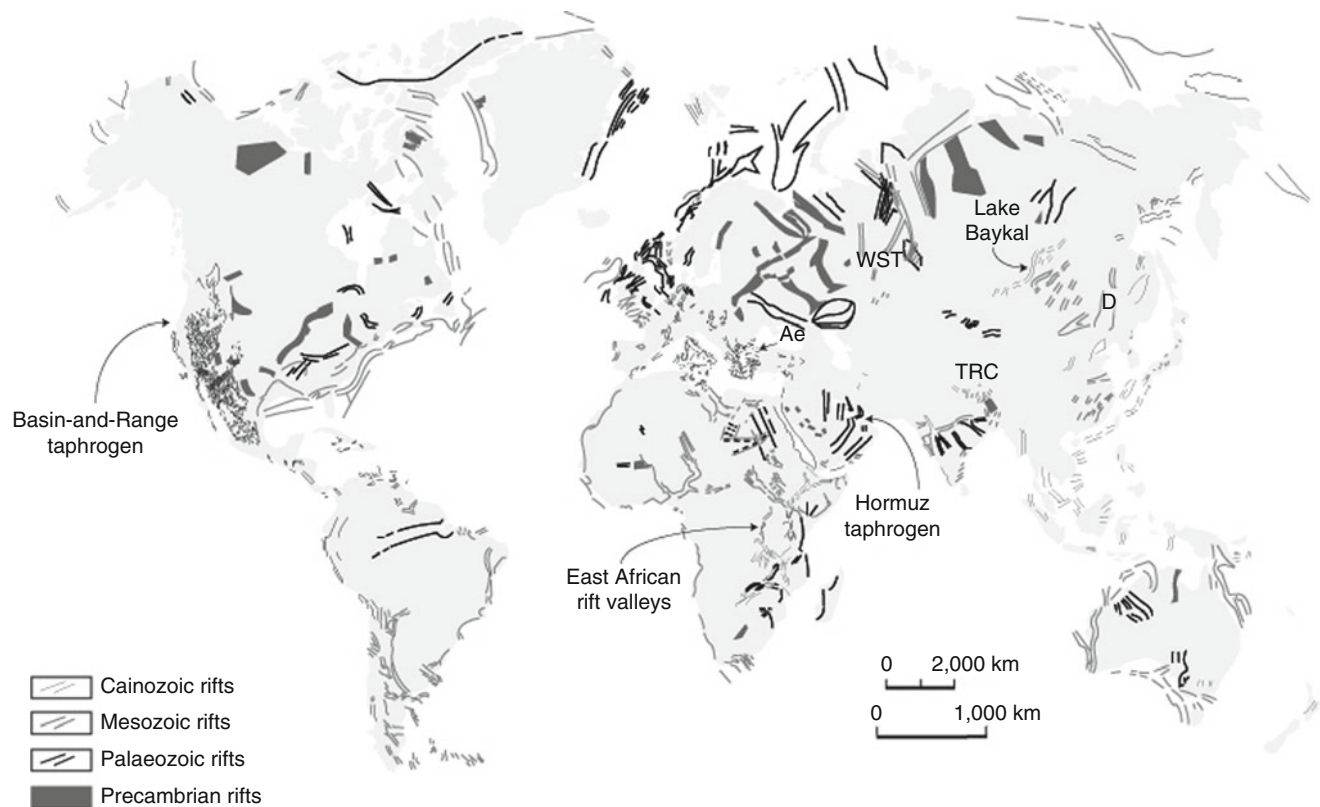
Continental Rifts, Fig. 3 Some properties of symmetric and asymmetric rifts

because of crustal thinning affecting it and in the basin thus formed sediments accumulate whose thicknesses can be in excess of 15 km in large rifts such as the Dneper-Donetz rift in the Ukraine (Ulmishek et al. 1994), which is very close to the maximum possible thickness (17 km) of sediment accumulation on continental crust (Dewey 1982). Most rifts of average size, however, have sediment thicknesses ranging from 2 to 5 km. Individual rift lengths vary from about a 1000 km to 100 km and widths from about 150 to 10 km. Rift size is a function not only of the amount of extension, but also the thickness of the lithosphere it disrupts. The thicker the lithosphere, the wider the rift. In Fig. 5, the distribution of continental rifts on Earth is shown. In areas of thick lithosphere (e.g., in the North American craton, eastern Europe and Siberian craton), the rifts are very wide, whereas in areas of thin lithosphere such as the western USA or in the Aegean they are very narrow, in places less than 10 km wide.

The bounding faults of the rifts have long been thought to be listric (shovel-shaped), that is, flattening with depth, although studies on earthquakes have not supported this. Wherever precise hypocenter location is possible and where the earthquake has caused a surface rupture (► “[Earthquake, Focal Mechanism](#)”, ► “[Earthquakes and Crustal Deformation](#)”, and ► “[Earthquakes, Location Techniques](#)”), it is seen that the associated fault is planar down to depths of 8 to 10 km. Eyidogan and Jackson (1985) documented that the 28 March 1969 Demirci earthquake in western Turkey consists of two discrete subevents: one along a northerly dipping planar normal fault and another one on a horizontal surface north of, but at the same depth as, the hypocenter of the first one. The horizontal rupture is believed to be located in a

Continental Rifts, Fig. 4 A part of the eastern (“Gregory”) rift in Kenya showing the intense faulting of the down-dropped hanging wall block between the master fault families. Google Earth image





Continental Rifts, Fig. 5 Distribution of continental rifts in the world. (Updated from Şengör and Natal'in 2001)

region that deforms ductilely under long-term strain rates, but ruptures when that rate is suddenly increased as in an earthquake (Fig. 2). This suggests a kinked cross-section for the master fault. Such “kinked” faulting events may be the explanation of why hanging wall blocks in some rifts are so much more disrupted than the rift shoulders. In areas of thicker lithosphere, however, hypocenters along the master faults of rifts may be as deep as 30 km as exemplified by the 10 March 1989 Malawi earthquake in east Africa (Fig. 2; Jackson and Belknap 1993) (► “Seismicity, Intraplate”).

The straight fault hypothesis is difficult to generalize especially where the master fault families bounding rifts are strongly curved in map view such as those that created the Lake Tanganyika rift system (Burgess et al. 1988). It is possible that individual segments of large faults rupture as straight fractures, but eventually the entire fault acquires a listric geometry. This seems supported by the fact that in Lake Tanganyika, where the rift floor has deeply subsided, faulting within the hanging wall block is less intense (Burgess et al. 1988). This is incompatible with a major kink of the master fault at depth. Wernicke and Burchfiel (1982) pointed out that where numerous, parallel-striking planar normal faults have rotated in unison around horizontal axes, one listric fault at the head of the series is a kinematic necessity to allow the others to rotate.

Asymmetric rifts are said to “face” in the direction of the dip of their master faults (Fig. 3). Many major rifts are made up of half rifts that change their facing along the strike. In such cases, “transfer” or “accommodation zones” form along their boundaries that take the form of areas of complex oblique faulting and crustal block rotations (see especially the excellent discussions on transfer and accommodation zones in Faulds and Stewart 1998; also Derer et al. 2005). Smaller and shallower transfer faults, essentially tear faults disrupting the hanging wall block with complicated displacement histories, form where the hanging wall is differentially stretched along the strike of the rift (cf. Şengör 1987).

Rift shoulders are commonly uplifted. In symmetric rifts, both shoulders go up because of isostatic adjustment of the crustal wedges that are formed by the master faults dipping towards the rift (► “Isostasy”). Such uplift not only rotates the land surface on the shoulders away from the rift thus guiding much of the regional drainage in the same direction, but it rotates also the bounding normal faults (Buck 1988). Each increment of faulting further rotates the shoulder and further rotates the previously formed fault segment. In Fig. 3, in the case of a young asymmetric rift the fault is shown to be straight. As the rift ages and extension advances, the fault surface curves away from the rift as a result of incremental rotations and finally its oldest parts acquire a flat position,

possibly with “horses” of the hanging wall block stuck on it as “extensional allochthons” (Fig. 3). This seems to be the origin of many near-horizontal normal faults in areas of thin lithosphere and large amounts of extension such as the extensional metamorphic core complexes of the North American Cordillera (Armstrong 1982), or those of the Tethysides (Verdel et al. 2007), although the existence of “primary” low-angle normal faults is also known both from detailed field studies (e.g., Miller and John 1988) and from the seismicity (Eyidogan and Jackson 1985; Abers et al. 1997). The amount of extension in many rifts, especially those that are symmetric, does not go that far, however, and their bounding faults rarely dip less than 30°.

In places where intracontinental rifting along a narrow rift valley has advanced far, as, for example, along the Wonji Fault Belt in Ethiopia, it has been noted that faulting is driven by dyke injection below the brittle upper carapace of about 10 km thickness and that segments with active deformation at the surface correlate with those of active magmatism (Kendall et al. 2005; Casey et al. 2006).

Cloos (1939) believed that the shoulder uplift in symmetric rifts resulted from the original doming that had led to key-stone drop of the hanging wall block, but it was quickly realized that the observed angles of slope on crustal domes rarely exceed 1°, whereas to give rise to substantial rifts one needs about 10° dome slopes, which are excessive. Table 1 lists values of extension and shoulder uplift in some selected rifts. In none, except possibly in the Española Basin, is the shoulder uplift sufficient to cause the estimated amount of stretching.

Sedimentation in Rifts

Because rifts form elongate basins, they are ideal sediment receptacles. The type of sediment to be found in a rift varies according to the size and geometry of the rift, the climatic zone(s) in which it is located and its evolutionary history. Although most rifts are filled slowly by sedimentation in the sea, lakes, rivers and, in a subordinate degree, subaerially by wind, some are catastrophically invaded by the sea by breaching the barriers of sub-sea-level rift complexes (e.g., the Permian North Sea or the medial Cretaceous South Atlantic). In such cases, either vast evaporite deposits (as in the Aptian of the South Atlantic: Burke and Şengör 1988) or thick shales (as in the late Permian of the North sea: Evans et al. 2003) are laid down in rift troughs.

Subaerial rifts are normally fed by rivers that locally create lakes within them because of endorheic conditions. Near the fault-bounded shoulders, stream valleys are commonly short (as the crest of the shoulder often is also the water-divide) but the streams in them are energetic because of the steep slopes. They thus carry clastic sediments of a variety of sizes that are horizontally sorted in large alluvial fans at the feet of the fault-escarpments. Along these escarpments, many alluvial fans

Continental Rifts, Table 1 Amounts of extension and shoulder uplift at selected continental rifts

Name of rift	Amount extended (km)	Uplift of shoulder area (km)
N. Kenya	35–40	>1.4–1.7
Gregory	10	2
Malawi	7.5 (from transfer fault offset)	2.8 (strike-slip affected?)
Suez N	16.3	1.5
Central	17.7–32.6	1
S	29.3 (all from fault geometry)	~500 m
Upper Rhine	17 (from crustal configuration), 5–7 (from fault geometry)	2.2
Oslo	28–36 (from crustal configuration), 11–13 (from fault geometry)	1
Viking	100–130 (from crustal configuration), 30 (from fault geometry)	3
Central (North Sea)	100–105 (from crustal configuration), 15 from fault geometry)	1.5
Benue	100 ^a	?
Baykal	15–20 (from crustal configuration), 10 (from fault geometry)	2–3
Española basin	5.5	1.5
Albuquerque Basin N	10 (from fault geometry)	~1
Albuquerque Basin S	16 (from fault geometry)	0.9

^aBenkhelil et al. (1988) established at least 10 km of shortening during the following compression

form and coalesce and ongoing subsidence leads to the accumulation of large thicknesses of conglomerates, sandstones, and shales near the faulted rift margin(s). Because of rotation towards the fault plane of the basin floors, the sedimentary thicknesses systematically decrease away from rift shoulders towards rift median lines. Steep slopes and catastrophic rain in both tropical and desert climates frequently lead to slope failures resulting in landslides complicating the sedimentary record. Permanent or ephemeral lakes may occupy the middle parts of rift floors. Ephemeral lakes commonly occur in dry climates in the form of playas and deposit evaporites and clays that may interfinger with marginal alluvial fan deposits. Permanent and deep lakes are important sediment repositories in rifts. They may accumulate very thick sedimentary sections in them exhibiting complex facies dictated by the evolution of the geomorphology of the rift and its surroundings. Such lakes can be very deep (Lake Baykal in Siberia: with 1637 m the world’s deepest lake in addition to an 8 km rift sedimentary fill; Lake Tanganyika in East Africa: 1457 m, but with half the sediment thickness of Lake Baykal). They are

well-aerated in cold climates because the cold upper layers of the lake waters make convection and overturning possible (as in Lake Baykal), but they are stratified with anoxic deeps in hot regions because hot surface waters cannot sink and thus inhibit overturning (as in Tanganyika). As the anoxic conditions in Lake Tanganyika exemplify, deep rift lakes in hot climates may generate good hydrocarbon source rocks in reducing environments. This is corroborated by ancient rift fills that formed under similar conditions such as the Permian oil-bearing sequences in the Junggar and Turfan rift basins in Central Asia (e.g., Graham et al. 1990; Miao et al. 2006) or the organic-rich (up to 8% organic C) Triassic-Jurassic lacustrine sediments of the Newark-type rifts of the eastern United States (Manspeizer and Olsen 1981).

Rift valley floors are seldom level along the strike of the master bounding faults and this commonly induces longitudinal drainage in rifts. In fact most of world's major rivers flow in two kinds of valleys: those that are located in rift troughs or in depressions inherited from them and those that are placed in foreland basin troughs (cf. Audley-Charles et al. 1977). The Rhine in Europe, the Irrawady in Asia, flow along their almost entire lengths in active rift troughs. Along the western arm of the East African rift valleys, Lake Tanganyika receives (surface elevation at 773 m a.s.l.) the excess waters of Lake Kivu (at 1460 m) by means of the south flowing river Ruzizi that came into being only 10,600 years ago and may have stopped flowing between 8 and 6000 years ago (Felton et al. 2007) showing the sensitive dependence of rift sedimentation on the global climate. Farther north, just beyond Kivu, Lake Edward (at 912 m) is connected to Lake Albert (at 619 m) through the northerly flowing river Semliki. Lake Albert in turns feeds the Nile northwards via the river Albert Nile (White Nile or "Bahr al Jebel," that is, "the river of the mountain").

In the sedimentary record, the rift facies is commonly characterized by thick (2–15 km) clastic sequences with rapid thickness and facies changes both laterally and vertically. If the rifts are underwater and far from clastic sources, they may house carbonate and even siliceous sedimentary sections as illustrated by the Alpine Triassic-early Jurassic facies. There, a continuous passage from late Triassic shallow water carbonates via a horizon of extensional-faulting-related neptunian dykes filled with progressively deepening facies and finally covered entirely by radiolarian cherts indicates a stretching basin floor as it deepened. That no shoulder uplift is indicated by any unconformities suggests that the Alpine rift was asymmetric with a major south-dipping detachment and that the rocks on the hanging wall block never experienced normal-fault-related upheaval. If catastrophic marine invasion occurs, then sediments at once blanket the preexisting topography.

For good reviews of sedimentation in rifts, see Frostick et al. (1986), Lorenz (1988), Olsen and Schlische (1990), Leeder (1995), Şengör (1995), Beratan (1996), also the section entitled "Rift Basins" in Einsele (2000, 543–560).

Magmatism in Rifts

Rift magmatism is extremely variable, because the tectonic environments in which rifts form are so diverse (see below). Rifts that form above mantle plumes (► ["Mantle Plumes"](#)) in plate interiors away from plate boundaries may begin their activity with restricted volumes of alkalic basalts indicating limited partial melting of mantle material at depth and may evolve towards abundant tholeiitic effusions as the temperature at depth increases and more and more mantle rock is melted, as, for example, illustrated, by the evolution of the Afar plume in east Africa. Most geologists take alkalic to peralkalic magmatism as indication of extensional tectonics. The frequent association of nepheline syenites, kindred alkaline igneous rocks, and carbonatites ("alkalic rocks and carbonatites": ARC) with intracontinental rifts has been taken by Burke et al. (2003, 2008) to argue that deformed rocks of the same compositions ("deformed alkalic rocks and carbonatites": DARCs) that have been shown to characterize suture zones in regions in Asia, Europe, Africa and North America are nothing but ARC rocks that erupted into intracontinental rifts that became involved in a Wilson cycle of ocean opening and closing. This has been shown to form a powerful tool to map suture zones where other suture indicators such as ophiolites or subduction-related rocks may be missing. However, to think that rifts generate only alkalic magmatic rocks is clearly too simplistic. Those rifts that form along magmatic arc systems, for instance, inevitably have calc-alkalic magmas and those that form above recently thickened continental crust, may generate vast quantities of calc-alkalic andesitic/rhyolitic melts. Rifts forming in back-arc regions, such as those in the Aegean, contain volcanic rocks showing evidence for a metasomatized mantle beneath them (e.g., the so-called Kulaites, that is, hornblende basalts, in western Turkey: Richardson-Bunbury 1996). Large amounts of silica-rich magmatic rocks may be generated if mantle plumes manage to melt large quantities of continental crust while ascending or if fractional crystallization is allowed to proceed undisturbed (e.g., the Cretaceous tin granites in Nigeria or the Cainozoic granites in Scotland; see Wiart and Oppenheimer 2004 for an active example in Afar). Some rifts such as Tanganyika or Baykal are almost entirely devoid of magmatism, while others, such as the late Palaeozoic Oslo Rift, have floors made up entirely of magmatic rocks. Some Atlantic-type continental margins, which grew out of disrupted rifts, are called "volcanic," whereas others "non-volcanic," because they descended from rifts with similar magmatic characteristics.

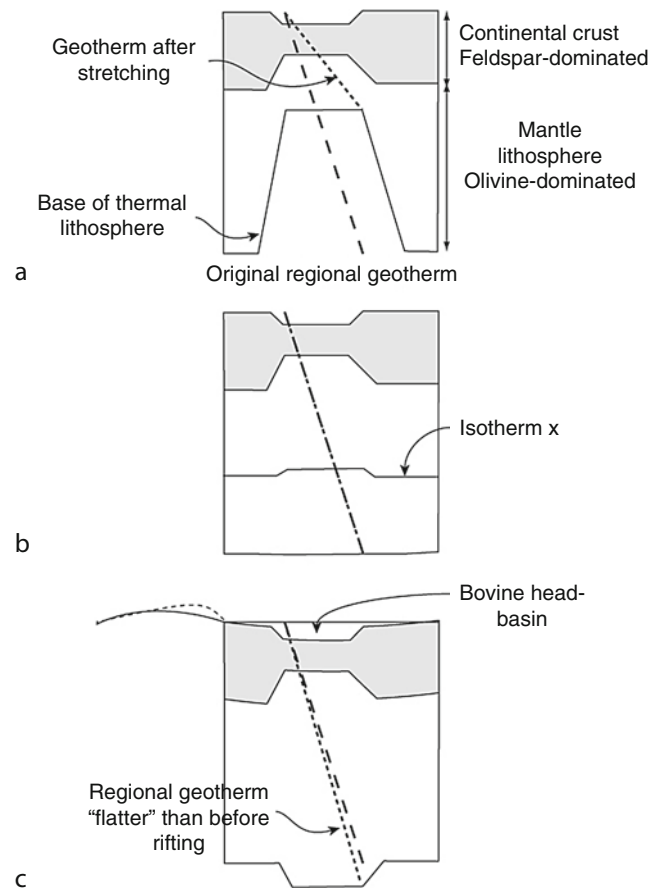
It is naive to think that lithospheric extension generates only certain kinds of magmatic rocks without considering the tectonic environments that gave rise to the extension and in which the extension proceeds (► ["Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity"](#)). It is therefore imperative to look at the diversity of the environments which generate rifts.

For good reviews of magmatism in rifts, see Kampunzu and Lubala (1991), Kinnaird (1998), Ulrych et al. (1999), Burke et al. (2003, 2008), Wilson et al. (2004), Yirgou et al. (2006); also the special section on magmatism and extension in *Journal of Geophysical Research* 100 (1995).

Metamorphism in Rifts

Continental rifts are not commonly thought of as prime loci of metamorphic events, but five different kinds of metamorphic rocks do occur in them (► “[Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity](#)”). In increasing importance: (1) contact metamorphic rocks around rift-related magmatic rocks, (2) hydrothermal metamorphism around vents, (3) Buchan-type metamorphism due to rise of the asthenosphere below highly stretched rift floors and associated abundant intrusions, (4) burial metamorphism due to large sediment thicknesses, and (5) metamorphic rocks that form within the continental crust and are brought to surface by rifting.

There is little to say about the first kind of metamorphic rocks in rifts as they are no different from any other contact metamorphic rock. Hydrothermal metamorphism is important in many rifts, because of the high heat flow and the continuous opening of extensional fractures allowing hot waters to circulate (e.g., Crane et al. 1991). Buchan-type metamorphism in rifts recently attracted much attention as a result of the early metamorphic history of the Pyrenees (Wickham and Oxburgh 1985), Mouthumet and the Montagne Noir massifs in southern France during the late Palaeozoic (Şengör 2013). Rift-related metamorphism creating high-grade gneisses has been reported on the basis of condensed metamorphic zones and absence of associated plutonism also elsewhere (e.g., St-Onge and King 1987). Burial metamorphism is not common in rifts unless they accumulate very large thicknesses of sediment quickly, sufficient to raise the temperature and pressure conditions to initiate metamorphism. If we assume a normal geothermal gradient ($\sim 3^\circ\text{C}/100\text{ m}$; see ► “[Heat Flow, Continental](#)”), an accumulation of about 4 km of sediment would be sufficient to initiate metamorphism by raising the temperature beyond 120°C . Jackson (1987) argued, however, that because earthquake hypocenters in rifts are seldom deeper than 10 km, at least beyond that depth rocks must be metamorphic (the 600°C is a thermal limit above which earthquakes do not occur). In a few places, there are earthquakes down to 30 km, as in the east African Rift System. It is true however, that rift sediments are seldom metamorphosed beyond greenschist grade, because they are seldom stretched enough and seldom accumulate sufficient thicknesses of sediment to heat the rocks beyond 320°C . Smith (1976) pointed out, for example, that at the bottom of the about 16 km-thick Gulf Coast sediment pile, the temperature may be just about 320°C under 4 kb pressure. But this ignores both the radioactive heating of the pile (► “[Radiogenic Heat Production of Rocks](#)”) and raising the geotherm by stretching. However,



Continental Rifts, Fig. 6 Thermal regime of rifts with a simple geotherm. (a) Immediately after the stretching both the crust and the lithospheric mantle thin and isotherms become crowded under the rift increasing the heat flow in the rift. Although no particle in the rift basement will become any hotter. (b) After stretching ceases, isotherms begin to subside. (c) When the lithosphere “recovers” the rift will end up with more lithospheric mantle beneath it and will have a heavier substratum that will subside and pull its surroundings with it creating an “intracratonic basin” overlying the rift. This is called a “bovine head” cross-section because of its resemblance to a long horn ox. While the basin subsides, its margins will be elastically uplifted (solid line above c), but may relax visco-elastically in time (dotted line above c). Here the margins are shown only on one side for graphic convenience, but naturally the events will take place all around the basin

even when the crust is stretched, every particle in the stretched crust must cool, unless the rift trough is rapidly sedimented (Fig. 6; McKenzie 1978; Jackson 1987). I am unaware of any well-documented metamorphic terrane that formed as a consequence solely of rift burial.

By contrast, large normal faults of the kind shown for the advanced asymmetric rifts in Fig. 3 may pull up rocks from the lower continental crust that may have been metamorphosed even up to granulite facies as shown by many extensional metamorphic core complexes. But in this case the metamorphism is not *caused* by rifting, but it is *exposed* by rifting. In fact, as the rocks ascend in the footwall, they get retrograded, which is directly caused by rifting.

For metamorphism in rifts, see: Armstrong (1982), Wickham and Oxburgh (1985), Jackson (1987), Lister and Davis (1989), Block and Royden (1990), Verdel et al. (2007), and Şengör (2013).

Kinds of Rifts

A classification of rifts must serve to answer the questions: how many different sorts of environments lead to lithospheric extension and what kinds of rifts form in these environments? In essence one must specify where and how taphrogens commonly form. If taphrogeny stops before producing ocean, it causes subsidence and creates large basins (called “koilogens” by Spizaharsky and Borovikov 1966, pp. 113ff.) overlying taphrogens (Fig. 6; cf. McKenzie 1978; Şengör 1995). A comprehensive synthesis of various taphrogen types does not yet exist. Below is a restatement of Şengör’s (Şengör 1995; Şengör and Natal’in 2001) classification of types of rifts that make up the main elements of taphrogens. That classification is hierarchical and goes from pure geometry to dynamics (Fig. 8). Its hierarchical nature allows the environment and the path of formation of a given rift to be determined. For numerous examples of each of Şengör’s categories, see Şengör and Natal’in (2001). In this section, references already given in Şengör (1995) and in Şengör and Natal’in (2001) are not repeated.

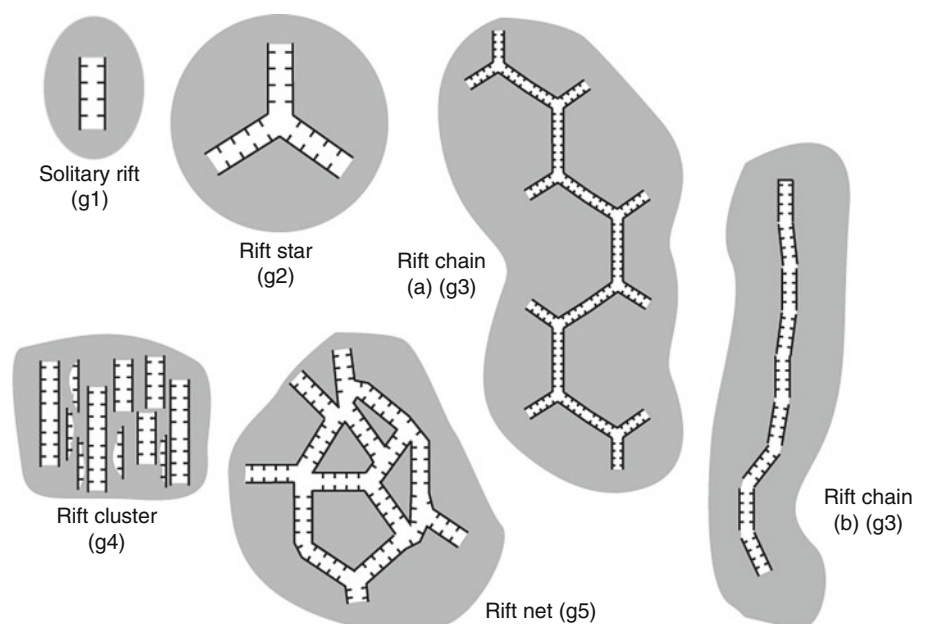
Şengör’s classification has three different categories that do not completely overlap, namely *geometric*, *kinematic*, and *dynamic*. In the following, the three different categories are identified with their initials, that is, *g*, *k*, and *d*, respectively.

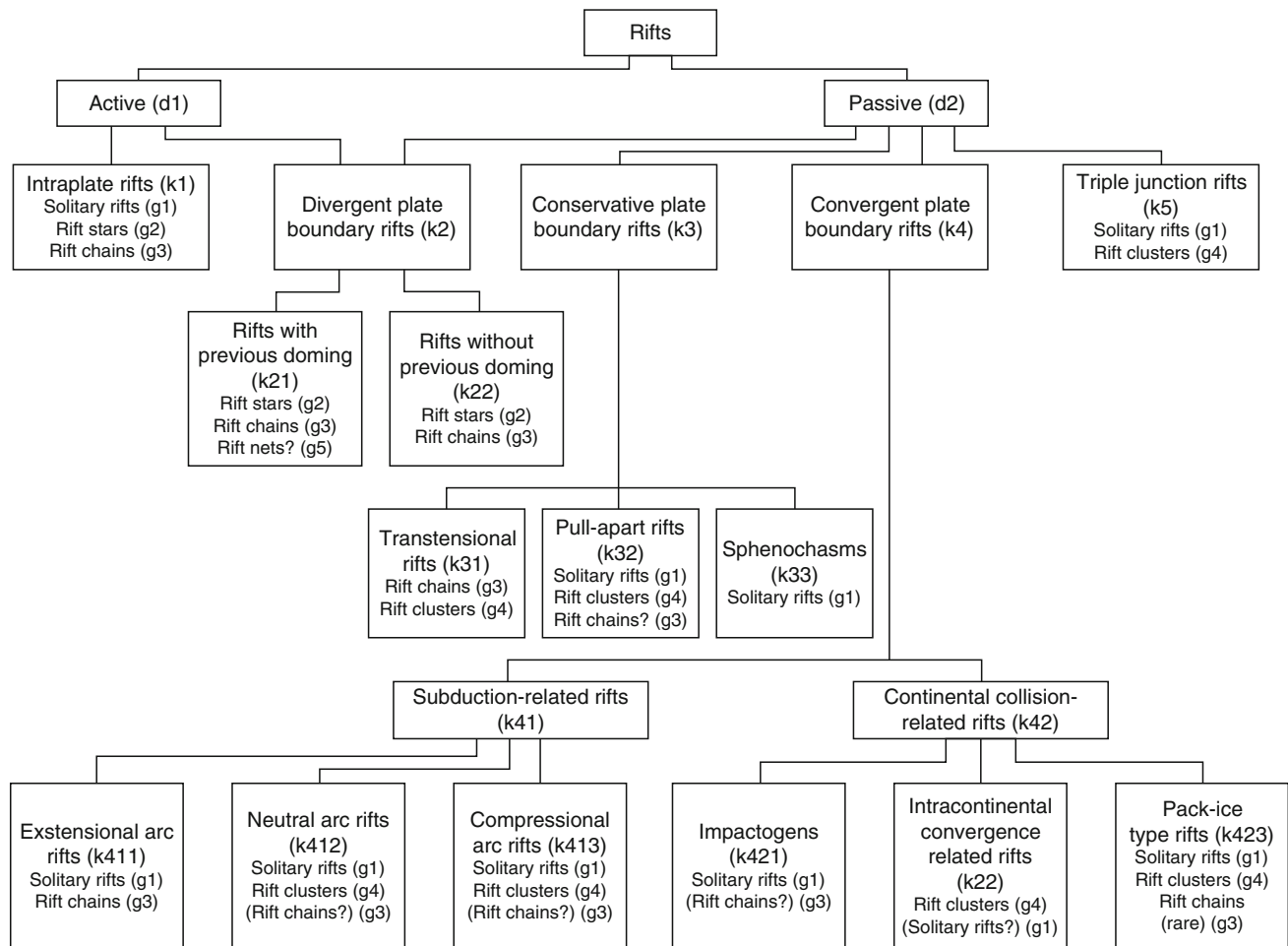
Geometric Classification of Rifts (See Figs. 7 and 8)

Rifts or groups of rifts forming individual taphrogenic units display five kinds of fundamental map patterns. From simplest to more complex, these are:

- g1) Solitary rifts:* Solitary rifts form small, fairly insignificant, and very rare taphrogens and are extremely difficult to ascertain in the geological record, because it is commonly hard to tell whether a given rift fragment is isolated or part of a larger taphrogen (for a comparable difficulty in shortening structures, imagine a big solitary fold or a large thrust fault!).
- g2) Rift stars:* Rift stars form when more than two rifts radiate away from a common center, building a roundish taphrogen. Rift stars are very common features of the structural repertoire of our planet today.
- g3) Rift chains:* When several rifts are aligned end-to-end along linear/arcuate belts of rifting, they form *rift chains*. The East African Rift System constitutes the best known active rift chain in the world. Solitary rifts or rift stars or combinations of these may be connected to form different kinds of rift chains (e.g., Fig. 5 rift chains a or b).
- g4) Rift clusters:* When several subparallel rifts occur in roughly equant areas, they are said to form a *rift cluster*. The two best-known active rift clusters in the world are the Basin-and-Range extensional area in North America and the Aegean Sea and the surrounding regions (see Fig. 5).
- g5) Rift nets:* Rift nets constitute a rare pattern, which comes about when rifts form a roughly checkered pattern as in the Proterozoic basement of the East European platform or in the late Mesozoic in central North Africa (cf. Şengör and

Continental Rifts, Fig. 7 The geometric classification of taphrogens. (After Şengör 1995)





Continental Rifts, Fig. 8 The classification of rifts and taphrogens. (After Şengör 1995)

Natal'in 2001). They resemble chocolate-bar boudinage, as seen in the Proterozoic basement of Eastern Europe (Fig. 5), and may have a similar origin, but more commonly rift nets form in complex and rapidly shifting stress environments in which dominant extension directions change fast. Many rift nets in fact may represent two superimposed rift clusters.

Kinematic Classification of Rifts (See Fig. 8)

Because rifts are ubiquitous in all stages of the Wilson Cycle of ocean opening and closing, the kinematic characteristics of the plate boundaries have been taken as a basis for classifying them according to the environment of the overall displacement and strain in which they form. There are three types of plate boundaries plus the plate interiors, to which four types of rift families correspond. In addition, incompatibilities arise around some unstable intracontinental triple junctions leading to complex rifting that should be treated separately from the other four classes, thus creating a fifth kinematic class, called "triple-junction rifts."

k1) Intraplate rifts: Rifts surrounded entirely by underformed lithosphere occupy this category. Such rifts are usually solitary, small, and rare (the difficulty in forming them is analogous to that forming a solitary fold or a nappe surrounded by entirely underformed terrain) and are not easy to identify in the geological history. The Lake George and Lake Champlain rifts in the northeastern United States are active examples (► "Seismicity, Intraplate").

k2) Rifts associated with divergent plate boundaries: These rifts result from plate separation along nascent extensional boundaries. All the Cainozoic rifts in east Africa belong here. This category of rifts may be further subdivided into two classes as follows:

k21) Rifts that form following an episode of doming: The divergent boundary along which rifts form is in this case preceded by an episode of lithospheric doming. The East African Rift Valleys are the best-known examples of such rifting (► "Isostasy, Thermal").

k22) Rifts that form with no pre-rift doming: In this case rifts form without a prelude of uplift, as is the case in the

Salton Trough and the Gulf of California. A good fossil example is the rifting of the Alpine Neo-Tethys in the earlier Mesozoic.

k3) *Rifts that form in association with conservative plate boundaries*: Conservative, that is, transform fault, boundaries are those along which neither extension nor shortening is required by the regional slip vectors. However, various reasons conspire to induce both extension and shortening to occur along considerable stretches of these boundaries. Rifts along conservative plate boundaries form in three different settings:

k31) *Transtensional conservative boundaries*: If a conservative boundary is opening up all along its length because of a component of extension, it is called *transtensional*. Many active rifts have a transtensional component and fossil examples of such rifts may be recognized largely through the structures they contain as shown by Olsen and Schlische (1990). Dewey (2002) gave an exhaustive analysis of the strain that develops in such rifts.

k32) *Pull-apart basins along conservative boundaries*: Major strike-slip faults commonly have bends along them that either facilitate (“releasing bends”) or obstruct (“restraining bends”) slip along them. Extensional basins form along the releasing bends, in which the magnitude of extension equals the magnitude of cumulative strike-slip offset along the strike-slip fault since the formation of the releasing bend. Such basins are called “pull-apart basins.” Crowell’s (1974) fault-wedge basins are nothing more than special cases of pull-apart basins.

k33) *Sphenochasms*: Not all basins created by secondary extension associated with strike-slip faults are pull-apart basins. Some represent tears caused by either an asperity or differential drag along the strike-slip fault in one of the fault walls, in which the amount of extension changes from a maximum along the fault to zero at the pole of opening of the tear-basin. S. W. Carey called such wedge-shaped rifts that open towards a major strike-slip fault *sphenochasms*.

k4) *Rifts that form in association with convergent plate boundaries*. A large family of rifts forms in association with convergent plate boundaries. In this group, a first-order subdivision is between rifts associated with subduction zones and rifts associated with continental collision, although this may artificially split some genetic groups, such as those rifts that presumably form because of the tension generated by excessive crustal thickening. The usefulness of the present grouping is that it enables a rapid overview of the presently active rift environments and comparison with ancient ones.

k41) *Rifts associated with subduction zones*: Environments of rifting associated with subduction zones

correspond to three different types of arc behavior, namely, extensional, neutral, and compressional arcs.

k411) *Rifts associated with extensional arcs*: An extensional arc generally splits along the magmatic axis (if such an axis is already in existence) forming a small rift chain. Such a situation is today known from both the Okinawa rift and the Izu-Bonin arc system. Such rifts commonly do not get preserved intact, both because of the complications of the tectonic evolution of arcs involving common changes of behavior and because of later collisions with other arcs or continents.

In extensional arcs, rifts also develop orthogonal to arc trend owing to the extension of the arc as it bows out in front of an expanding marginal basin (as, for instance, in Crete).

k412) *Rifts associated with neutral arcs*: Neutral arcs have neither shortening nor extension across them. Therefore, the only rifts that may form in neutral arcs are those associated with arc-parallel strike-slip faults, which may be classified in the same way as the rifts that form along conservative plate boundaries. More complex rift basins may originate along such arc-parallel strike-slip faults, if the sliver plate in the forearc area and its various pieces rotate about vertical axes.

Pull-apart basins in arcs are difficult to recognize, but the Sumatra Fault has several well-developed examples along it.

Sphenochasms along strike-slip faults in arcs are rarer still. Davis et al. (1978) have discussed two possible examples, the more recent of which may have created the “Columbia Embayment” by motion along the Straight Creek fault in the latest Cretaceous and the earliest Cainozoic.

k413) *Rifts associated with compressional arcs*: In compressional arcs crust commonly thickens and lithosphere thins, both by heating and by eventual delamination. The arc becomes shortened across, and elongated along, its trend. This elongation commonly generates rifts at high angles to the trend of the arc. Rifts on the Altiplano in the Andes are examples of such high-angle rifts.

k42) *Rifts associated with zones of continental collision*: Three different environments of rifting form associated with the collision of continents: (1) *Lines of extension* that radiate from points at which collision commences, (2) *Regions of extension* abutting against sutures, and (3) *Nodes of extension* in areas of complex deformation in fore- and hinterlands shattered by collisions. Impactogens (k421), rifts forming in intracontinental convergence belts (k422), and pack-ice-type rifts (k423) correspond with these three environments, respectively.

k5) *Triple Junction Rifts*: Triple junction rifts form at or near unstable intracontinental triple junctions, at which plate evolution dictates the generation of “holes” because continental lithosphere cannot be continuously subducted.

Dynamic (Genetic) Classification of Rifts (See Fig. 6)

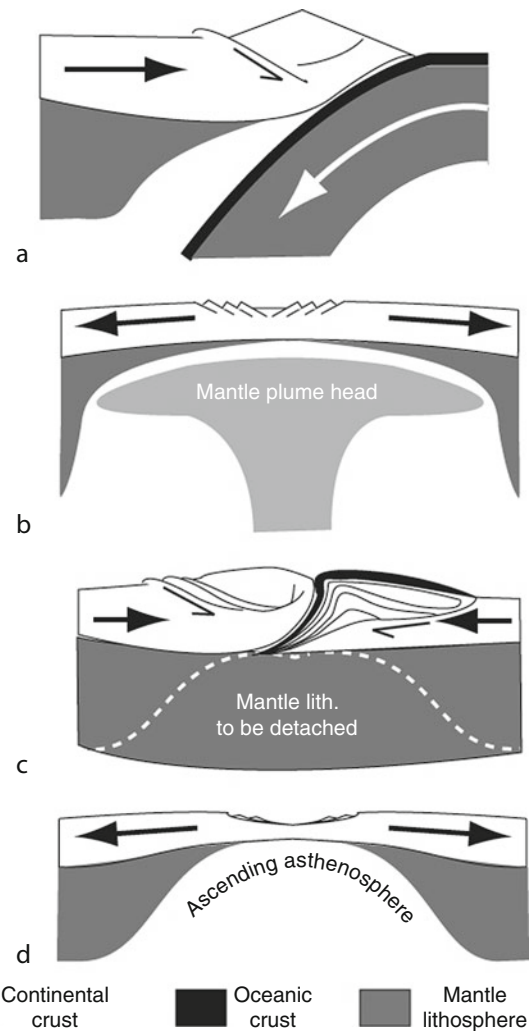
Rifts are also classified according to the origin of forces (► “*Geodynamics*”) that lead to rifting. Şengör and Burke (1978) proposed that stresses which cause rifting may be imposed on the lithosphere directly by the mantle beneath it (cf. Şengör 2001) or they may result from two-dimensional plate evolution. Accordingly, they termed these two modes of rifting “active” and “passive.” Gans (1987) tried to replace these terms with “open-system” and “closed-system” rifting, respectively, which, however, did not find general acceptance.

d1) *Active rifting*: “Active rifting” is rifting caused by mantle upwelling (► “*Earth, Density Distribution*” and ► “*Mantle Convection*”) associated with plumes in the mantle (► “*Mantle Plumes*”). Two views have been advanced to explain the origin of the extension in domes rising above hot-spot jets: one ascribes the rifting to basal shear stresses induced by a spreading plume head beneath a dome. The other holds the potential energy of the rising dome responsible for driving the rifting. All of these factors probably do contribute to maintaining the active rifting process at its habitually slow pace of considerably less than 1 cm/a (► “*Geoid*”).

Plume-caused taprogens may be also termed deeply-rooted rifts and would be the extensional counterparts of subduction-related orogens (Fig. 9a, b). These can be contrasted to collisional orogens that have lost their subductive “anchors” and the passive rifts with no deep mantle connexions (Fig. 9c, d).

d2) *Passive rifting*: In the passive rifting mode, extension is caused by the two dimensional motions of the lithospheric plates and not by an autonomous interference from the mantle (Fig. 11d). In this mode of rifting, there is no prerifting doming associated with a hot-spot (Şengör and Burke 1978). Kinematic mechanisms reviewed above under the headings k22, k31, k32, k33, k411, k412, k413, k421, k422, k423, and k5 all may form rifts in a “passive rifting mode.”

There is only one kind of rift this classification does not consider: rifts that form by propagating from an already existing rift. Since propagation may take many forms, it might be sufficient to indicate such rifts with the notation d2 to indicate their passive mode of opening. The regional geologist may use Fig. 8 as a “flow chart” to follow the evolutionary histories of the various kinds of rift basins he encounters.



Continental Rifts, Fig. 9 Deeply rooted mega-structures ((a) subduction-controlled orogens, (b) mantle-plume controlled, i.e., active rifts) and rootless mega-structures ((c) collisional orogens, (d) plate motion-controlled, i.e., passive rifts)

Concluding Note

Although they constitute one of the three great continental structure types and have serious relevance to human well-being, rifts remain poorly known. This is largely because there are still important technological difficulties in studying them directly because of the large sediment thicknesses they contain and because their diversity is little appreciated. They are so ubiquitous in all tectonic environments that they tend to be overlooked in favor of other, rarer, structures that may appear more prominent. It is commonly not recognized that all kinds of structures such as magmatic arcs, collisional orogens, keirogens, intracratonic basins all have associated rifts and, consequently, people wishing to know about rifts tend not to concern themselves with them. A stereotypic rift that commonly bedevils textbooks is one that disrupts a craton

and has alkalic or peralkalic vulcanism with a thick, mainly nonmarine sedimentary fill, although there are many more rifts in the world than those with such characteristics. What is now needed in rift studies is a comprehensive synthesis that takes into account this diversity.

Bibliography

- Abers GA, Mutter CZ, Fang J (1997) Shallow dips of normal faults during rapid extension: earthquakes in the Woodlark-d'Entrecasteaux rift system, Papua New Guinea. *J Geophys Res* 102:15,301–15,307
- Armstrong RL (1982) Cordilleran metamorphic core complexes – from Arizona to Southern Canada. *Ann Rev Earth Planet Sci Lett* 10: 129–154
- Audley-Charles MG, Curran JR, Evans G (1977) Location of major deltas. *Geology* 5:341–344
- Beratan KK (ed) (1996) Reconstructing the history of basin and range extension using sedimentology and stratigraphy. Geological Society of America Special Paper 303, vi+212 pp+ 1 folded plate in pocket
- Bernoulli D, Manatschal G, Desmurs L, Müntener O (2003) Where did Gustav Steinmann see the trinity? Back to the roots of an Alpine ophiolite concept. In: Dilek Y, Newcomb S (eds) Ophiolite concept and the evolution of geological thought. Geological Society of America Special Paper 373, pp 93–110
- Block L, Royden LH (1990) Core complex geometries and regional scale flow in the lower crust. *Tectonics* 9:557–567
- Blundell DJ, Gibbs AD (eds) (1990) Tectonic evolution of the North Sea rifts. Clarendon, Oxford
- Boillot G, Recq M, Winterer EL, Meyer AW, Applegate J, Baltuck M, Bergen JA, Comas MC, Davies TA, Dunham K, Evans CA, Girardeau J, Goldberg G, Haggerty J, Jansa LF, Johnson JA, Kasahara J, Loreau JP, Luna-Sierra E, Moullade M, Ogg J, Sarti M, Thurow J, Williamson M (1987) Tectonic denudation of the upper mantle along passive margins: a model based on drilling results (ODP leg 103, western Galicia margin, Spain). *Tectonophysics* 132: 335–342
- Buck R (1988) Flexural rotation of normal faults. *Tectonics* 7:959–973
- Burgess CF, Rosendahl BR, Sander S, Burgess CA, Lambiase J, Derksen S, Meader N (1988) The structural and stratigraphic evolution of Lake Tanganyika: a case study of continental rifting. In: Manspeizer W (ed) Triassic-Jurassic rifting – continental breakup and the origin of the Atlantic Ocean and passive margins B. Elsevier, Amsterdam, pp 859–881
- Burke K (1996) The African plate. *S Afr J Geol* 99:339–409
- Burke K, Dewey J (1973) Plume-generated triple junctions: key indicators in applying plate tectonics to old rocks. *J Geol* 81:406–433
- Burke K, Gunnell Y (2008) The African Erosion surface: a continental-scale synthesis of geomorphology, tectonics, and environmental change over the past 180 million years. Geological Society of America Memoir 201, iv+66 pp
- Burke K, Şengör AMC (1988) Ten metre global sea-level change associated with South Atlantic Aptian salt deposition. *Mar Geol* 83:309–312
- Burke K, Wilson JT (1976) Hot spots on the earth's surface. *Sci Am* 235:46–57
- Burke K, Ashwal LD, Webb S (2003) New way to map old sutures using deformed alkaline rocks and carbonatites. *Geology* 31:391–394
- Burke K, Khan SD, Mart RW (2008) Grenville Province and Montereian carbonatite and nepheline syenite distribution related to rifting, collision, and plume passage. *Geology* 36:983–986
- Casey M, Ebinger C, Keir D, Gloaguen R, Mohamed F (2006) Strain accommodation in transitional rifts: extension by magma intrusion and faulting in Ethiopian rift magmatic segments. In: Yirgou G, Ebinger CJ, Maguire PKH (eds) The Afar volcanic province within the East African rift system. Geological Society Special Publication 259, pp 143–163
- Cloos H (1939) Hebung-Spaltung-Vulkanismus – Elemente einer Geometrischen Analyse Irdischer Großformen. *Geol Rundsch* 30(Zwischenheft 4A):405–527
- Corti G, Cuffaro M, Doglioni C, Innocenti F, Manetti P (2006) Coexisting geodynamic processes in the Sicily Channel. In: Dilek Y, Pavlides S (eds) Postcollisional tectonics and magmatism in the Mediterranean region and Asia. Geological Society of America Special Paper 409, pp 83–96
- Coward MP, Dewey JF, Hancock PL (eds) (1987) Continental extensional tectonics. Geological Society Special Publication 321 (Albert M. Quennell volume), 637 pp
- Crane K, Hecker B, Golubev V (1991) Heat flow and hydrothermal vents in Lake Baikal, U.S.S.R. *EOS Trans Am Geophys Union* 72:585 and 588
- Crowell JW (1974) Sedimentation along the San Andreas Fault, California. In: Dickinson WR (ed) Plate tectonics and sedimentation. Society of Economic Paleontologists and Mineralogists Special Publication 19, pp 292–303
- Davis GA, Monger JWH, Burchfiel BC (1978) Mesozoic construction of the Cordilleran “collage,” central British Columbia to central California. In: Howell DG, McDougall KA (eds) Mesozoic paleogeography of the Western United States, Pacific Coast Paleogeography Symposium 2, Pacific section. Society of Economic Paleontologists and Mineralogists, Los Angeles, pp 1–32
- Derer CE, Schumacher ME, Schäfer A (2005) The Upper Rhine Graben: basin geometry and early syn-rift tectono-sedimentary evolution. *Int J Earth Sci (Geologische Rundschau)* 94:640–656
- Dewey JF (1982) Plate tectonics and the evolution of the British Isles. *J Geol Soc Lond* 139:371–412
- Dewey JF (2002) Transtension in arcs and orogens. *Int Geol Rev* (George A. Thompson Symposium volume) 44:402–439
- Dickinson WR (2002) The Basin and Range province as a composite extensional domain. *Int Geol Rev* (George A Thompson Symposium volume) 44:1–38
- Durand L, Jolivet, F., Horváth, F. and Séranne M (eds) (1999) The Mediterranean basins: tertiary extension within the Alpine Orogen. Geological Society Special Publication 156, 570 pp
- Einsele G (2000) Sedimentary basins – evolution, facies and sediment budget, 2nd, completely revised edn. Springer, Berlin
- Evans D, Graham C, Armour A, Bathurst P (2003) The millenium atlas: petroleum geology of the central and northern North Sea. The Geological Society of London, London
- Eyidogan H, Jackson J (1985) Seismological study of normal faulting in the Demirci, Alasehir and Gediz earthquakes of 1969–70 in western Turkey: implications for the nature and geometry of deformation in the continental crust. *Geophys J R Astron Soc* 81:569–607
- Faulds JE, Stewart JH (eds) (1998) Accommodation zones and transfer zones: the regional segmentation of the Basin and Range province. Geological Society of America Special Paper 323, vii+257 pp. + 1 folded plate in pocket
- Felton AA, Russell JM, Cohen AS, Baker ME, Chesley JT, Lezzar KE, McGlue MM, Pigati JS, Quade J, Curt Stager J, Tiercelin JJ (2007) Paleolimnological evidence for the onset and termination of glacial aridity from Lake Tanganyika, Tropical East Africa. *Palaeogeogr Palaeoclimatol Palaeoecol* 252:405–423. <https://doi.org/10.1016/j.palaeo.2007.04.003>
- Frostick LE (1997) The east African rift basins. In: Selley RC (ed) African basins. Elsevier, Amsterdam, pp 187–209
- Frostick LE, Renaut RW, Reid I, Tiercelin JJ (eds) (1986) Sedimentation in the African rifts. Geological Society Special Publication 25, 382 pp
- Gans PB (1987) An open-system, two-layer crustal stretching model for the eastern Great Basin. *Tectonics* 6:1–12

- Graham SA, Brassell S, Carroll AR, Xiao X, Demaison G, McKnight CL, Liang Y, Chu J, Hendrix MS (1990) Characteristics of selected petroleum source rocks, Xianjiang Uygur autonomous region, North-west China. *Am Assoc Pet Geol Bull* 74:493–512
- Gregory JW (1894) Contributions to the physical geography of British East Africa. *Geogr J* 4:289–315
- Harding TP (1983) Graben hydrocarbon plays and structural styles. *Geol Mjnb* 62:3–23
- Hüttner R (1991) Bau und Entwicklung des Oberrheingrabens – Ein Überblick mit historischer Rückschau. *Geol Jahrb* E48:17–42
- Ingersoll RV (2001) Structural and stratigraphic evolution of the Rio Grande rift, northern New Mexico and southern Colorado. *Int Geol Rev* 43:867–891
- Jackson JA (1987) Active continental deformation and metamorphism. *Philos Trans R Soc Lond A321*:47–66
- Jackson J, Belknap T (1993) The Malawi earthquake of 10 March 1989: deep faulting within the east African rift system. *Tectonics* 12:1131–1139
- Jaffé FC (1971) Geothermal energy: a review. *Bulletin der Vereinigung Schweizerischer Petroleum-Geologen und -Ingenieure* 38:17–40
- Kampunzu AB, Lubala RT (eds) (1991) Magmatism in extensional structural setting – the Phanerozoic African plate. Springer, Berlin
- Kendall J-M, Stuart GW, Ebinger CJ, Bastow ID, Keir D (2005) Magma-assisted rifting in Ethiopia. *Nature* 433:146–148
- Kinnaid JA (1998) Aspects of tensional magmatism. *J Afr Earth Sci* 26:1–150
- Kravchinsky VA, Krainov MA, Evans ME, Peck JA, King JW, Kuzmin MI, Sakai H, Kawai T, Williams DF (2003) Magnetic record of Lake Baikal sediments: chronological and paleoclimatic implication for the last 6.7 Myr. *Palaeogeogr Palaeoclimatol Palaeoecol* 195:281–293
- Kuz'min MI, Karabanov EB, Kawai T, Williams D, Bychinsky VA, Kerber EV, Kravchinsky VA, Bezrukova EV, Prokopenko AA, Geletii VF, Kalmychikov GV, Goregryad AV, Antipin VS, Khomutova MY, Soshina NM, Ivanov EV, Khrusevich GK, Tkachenko LL, Solotchina EP, Ioshida N, Gvozdkov AN (2001) Deep drilling on Lake Baikal: main results. *Russ Geol Geophys* 42:3–28
- Lambiase JJ (ed) (1995) Hydrocarbon habitat in rift basins. Geological Society Special Publication 80, 381 pp
- Landon SM (ed) (1994) Interior rift basins. American Association of Petroleum Geologists Memoir 59, 276 pp
- Le Pichon X, Sibuet JC (1981) Passive margins: a model of formation. *J Geophys Res* 86:3708–3720
- Leeder MR (1995) Continental rifts and proto-oceanic rift troughs. In: Busby CJ, Ingersoll RV (eds) *Tectonics of sedimentary basins*. Blackwell, Oxford, pp 119–148
- Lister GS, Davis GA (1989) The origin of metamorphic core complexes and detachment faults formed during tertiary continental extension in the northern Colorado River region, U.S.A. *J Struct Geol* 11:65–94
- Littke R, Bayer U, Gajewski D, Nelskamp S (2008) Dynamics of complex intracontinental basins – the central European basin system. Springer, Berlin
- Lorenz JC (1988) Triassic-Jurassic Rift-Basin sedimentology – history and methods. Van Nostrand Reinhold, New York
- MacKenzie AS, McKenzie D (1983) Isomerization and aromatization of hydrocarbons in sedimentary basins formed by extension. *Geol Mag* 120:417–470
- Manspeizer W (1988) Triassic-Jurassic rifting – continental breakup and the origin of the Atlantic Ocean and passive margins, vol 2. Elsevier, Amsterdam
- Manspeizer W, Olsen PE (1981) Rift basins of the passive margin: tectonics, organic-rich lacustrine sediments, basin analysis. In: Hobbs GW (ed) *Field guide to the geology of the Palaeozoic, Mesozoic, and tertiary rocks of New Jersey and the Central Hudson Valley*. Petroleum Exploration Society of New York, New York, pp 25–103
- McKenzie D (1978) Some remarks on the development of sedimentary basins. *Earth Planet Sci Lett* 40:25–32
- McQuarrie N, Wernicke BP (2005) An animated tectonic reconstruction of southwestern North America since 36 MA. *Geosphere* 1:147–172. <https://doi.org/10.1130/GES00016.1>
- Miao JY, Kou HS, Zhou LF, Han ZY (2006) Sedimentary environments of organic matter from Middle Permian source rocks in northern Xinjiang, China. *Chin J Geochem* 25:258–265
- Miller JMG, John BE (1988) Detached strata in a tertiary low-angle normal fault terrane, southeastern California: a sedimentary record of unroofing, breaching, and continued slip. *Geology* 16:645–648
- Mohr P (2009) Africa beckoning. Millbrook Nova Press, Galway
- Morley CK (1999) Geoscience of rift systems – evolution of East Africa. AAPG studies in geology 44. The American Association of Petroleum Geologists, Tulsa, 242 pp. + 5 appendices+ 16 pp. Index
- Olsen PE, Kent DV (1999) Long-period Milankovich cycles from late Triassic and early Jurassic of eastern North America and their implications for the calibration of the early Mesozoic time-scale and the long-term behaviour of the planets. *Philos Trans R Soc Lond A357*:1761–1786
- Olsen PE, Schlische RW (1990) Transtensional arm of the early Mesozoic Fundy rift basin: penecontemporaneous faulting and sedimentation. *Geology* 18:695–698
- Quennell AM (1982) Rift valleys Afro-Arabian. Benchmark papers in geology, vol 60. Hutchison Ross, Stroudsburg
- Quennell AM (ed) (1985) Continental rifts. Benchmark papers in geology series. Van Nostrand Reinhold, New York
- Richardson-Bunbury JM (1996) The Kula volcanic field, western Turkey: the development of a Holocene alkali basalt province and the adjacent normal-faulting graben. *Geol Mag* 133:275–283
- Ring U, Wernicke B (eds) (2009) Extending a continent: architecture, rheology and heat budget. Geological Society Special Publication 321, 272 pp
- Rotstein Y, Edel J-B, Gabriel G, Boulanger D, Schaming M, Munsch M (2006) Insight into the structure of the Upper Rhine Graben and its basement from a new compilation of Bouguer Gravity. *Tectonophysics* 425:55–70
- Sawkins FJ (1990) Metal deposits in relation to plate tectonics. Springer, Berlin
- Schlüter T (1997) *Geology of East Africa* with contributions by Craig Hampton: Beiträge zur Regionalen Geologie der Erde, 27. Gebrüder Borntraeger, Berlin
- Schumacher ME (2002) Upper Rhine Graben: role of preexisting structures during rift evolution. *Tectonics* 21:1006–1022. <https://doi.org/10.1029/2001TC900022>
- Şengör AMC (1987) Cross-faults and differential stretching of hangingwalls in regions of low-angle normal faulting: examples from western Turkey. In: Coward MP, Dewey JF, Hancock PL (eds) *Continental extensional tectonics*. Geological Society Special Publication 321 (Albert M. Quennell volume), pp 575–589
- Şengör AMC (1995) Sedimentation and tectonics of fossil rifts. In: Busby CJ, Ingersoll RV (eds) *Tectonics of sedimentary basins*. Blackwell, Oxford, pp 53–117
- Şengör AMC (2001) Elevation as indicator of mantle plume activity. In: Ernst R, Buchan K (eds). Geological Society of America Special Paper 352, pp 183–225
- Şengör AMC (2003) The large wavelength deformations of the lithosphere: materials for a history of the evolution of thought from the earliest times to plate tectonics. Geological Society of America Memoir 196, xvii+347 pp. + 3 folded plates in pocket
- Şengör AMC (2013) The Pyrenean Hercynian Keirogen and the Cantabrian Orocline as genetically coupled structures. *J Geodyn* 65:3–21. (Ali Koçyiğit Festschrift)
- Şengör AMC (2018) Some notes towards a history of ideas on rifts and extensional tectonics in general. *Earth Sci Hist* 37:144–156
- Şengör AMC, Burke K (1978) Relative timing of rifting and volcanism on Earth and its tectonic implications. *Geophys Res Lett* 5:419–421

- Şengör AMC, Natal'in BA (2001) Mantle Plumes: Their Identification Through Time. Rifts of the world. In: Ernst R, Buchan K (eds). Geological Society of America, Boulder, Colorado 352, pp 389–482
- Smith AG (1976) Orogeny: a review. *Tectonophysics* 33:215–285
- Snow JK, Wernicke B (2000) Cenozoic tectonism in the Central Basin and range: magnitude, rate and distribution of upper crustal strain. *Am J Sci* 300:659–719
- Spizaharsky TN, Borovikov LI (1966) Tectonic map of the Soviet Union on a scale of 1: 2 500 000. In: Scientific communications read to the Commission for the Geological Map of the World, Delhi, 22nd International Geological Congress, pp 111–120
- St-Onge MR, King JE (1987) Evolution of regional metamorphism during back-arc stretching and subsequent crustal shortening in the 1.9. Ga Wopmay Orogen, Canada. *Philos Trans R Soc Lond A* 321:199–218
- Surkov VS, Smimov LV, Zhero OG (1994) Early Mesozoic rifting and evolution of the West Siberian Basin. In: Roure F, Elluz N, Shein VS, Skvortsov I (eds) *Geodynamic evolution of sedimentary basins, international symposium, Moscow*. Éditions Technip, Paris, pp 135–143
- ten Brink US, Taylor MH (2002) Crustal structure of central Lake Baikal: insights into intracontinental rifting. *J Geophys Res* 107:ETG-2-1–ETG-2-15. <https://doi.org/10.1029/2001JB000300>
- Ulmishek GF, Bogino VA, Keller MB, Poznyakevich ZL (1994) Structure, stratigraphy, and petroleum geology of the Pripyat and Dnieper-Donets Basins, Byelorussia and Ukraine. In: Landon SM (ed) *Interior rift basins*. American Association of Petroleum Geologists Memoir 59, pp 125–156
- Ulrych J, Cajz V, Adamovič J (eds) (1999) Magmatism and rift basin evolution. *GeoLines* 9:1–135
- Verdel C, Wernicke BP, Ramezani J, Hassanzadeh J, Renne PR, Spell TL (2007) Geology and thermochronology of tertiary Cordilleran-style metamorphic core complexes in the Saghand region of central Iran. *Geol Soc Am Bull* 119:961–977
- Wernicke B (ed) (1990) Basin and range extensional tectonics near the latitude of Las Vegas, Nevada. Geological Society of America Memoir 176, xii+511 pp. + numerous separate plates
- Wernicke B (2009) The detachment era (1977–1982) and its role in revolutionizing continental tectonics. In: Ring U, Wernicke B (eds). Geological Society Special Publication 321, pp 1–8
- Wernicke B, Burchfiel BC (1982) Modes of extensional tectonics. *J Struct Geol* 4:105–111
- Wiat P, Oppenheimer C (2004) Large magnitude silicic volcanism in north Afar: the Nabro Volcanic Range and Ma'alalta volcano. *Bull Volcanol* 67:99–115
- Wickham S, Oxburgh ER (1985) Continental rifts as a setting for regional metamorphism. *Nature* 318:330–333
- Willis B (1936) East African plateaus and rift valleys. Studies in comparative seismology. Carnegie Institution of Washington, Washington, DC
- Wilson M, Neumann E-R, Davies GR, Timmermann MJ, Heremans M, Larsen BT (eds) (2004) Permo-carboniferous magmatism and rifting in Europe. Geological Society Special Publication 223, 498 pp
- Woodcock NH (1986) The role of strike-slip fault systems at plate boundaries. *Philos Trans R Soc Lond A* 317:13–29
- Yirgou G, Ebinger CJ, Maguire PKH (eds) (2006) The Afar volcanic province within the East African rift system. Geological Society Special Publication 259, 331 pp
- Ziegler PA (ed) (1992) *Geodynamics of rifting*. Elsevier, Amsterdam
- Ziegler PA (1994a) Hydrocarbon habitat in rift basins. In: Roure F, Elluz N, Shein VS, Skvortsov I (eds) *Geodynamic evolution of sedimentary basins, international symposium, Moscow*. Éditions Technip, Paris, pp 85–94
- Ziegler PA (1994b) Geodynamic processes governing development of rift basins. In: Roure F, Elluz N, Shein VS, Skvortsov I (eds) *Geodynamic evolution of sedimentary basins, international symposium, Moscow*. Éditions Technip, Paris, pp 19–67

Core Dynamo

Ulrich R. Christensen

Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany

Synonyms

Geodynamo

Definition

Dynamo: Process for generating electrical current and magnetic field by electromagnetic induction in a moving conducting medium

Geodynamo: Dynamo process in the fluid outer core that generates Earth's main magnetic field

Introduction

It has been firmly established that the geomagnetic field must be generated by a magnetohydrodynamic dynamo operating in the Earth's outer core. Thermal and compositional buoyancy forces drive a convective circulation of the liquid iron alloy. Assuming that a magnetic field is present, electrical currents are induced in the moving fluid. When the magnetic field associated with these currents has the strength and geometry that is suitable for the induction process and no external source for the field is required, this is called a self-sustained dynamo. To first approximation the Earth's core is a sphere with uniform electrical conductivity, in contrast to technical dynamos, where the currents are guided by a highly inhomogeneous distribution of electrical conductors. In the former case we speak of a homogeneous dynamo. From a theoretical point of view, homogeneous dynamos are more difficult to understand. Simple motions, such as differential rotation (as in an electrical generator), are unable to drive a homogeneous dynamo and flow patterns of a certain complexity are required. Until late into the twentieth century dynamo theory has been concerned mainly with the conceptual understanding of how, in principle, a magnetic field can be generated in such an environment. Starting with models by Glatzmaier and Roberts (1995), Kageyama and Sato (1995), and Kuang and Bloxham (1997), realistic self-consistent numerical simulations of the geodynamo became available and have been successful in reproducing many of the observed properties of the geomagnetic field. The more fundamental aspects of the geodynamo are discussed elsewhere (► “Geomagnetic Field, Theory”). Here, the progress in

understanding the geodynamo based on numerical modeling and comparing its results with specific properties of the geomagnetic field is addressed.

Dynamo Model Concept and Equations

Model Setup

In contrast to earlier kinematic dynamo models, where the flow is prescribed, modern geodynamo models aim at a fully self-consistent treatment of fluid flow and magnetic field generation in the Earth's outer core. There are some basic requirements for a realistic model: (i) The mechanism for driving flow by thermal or compositional buoyancy must be part of the model. (ii) Because an axisymmetric field cannot be generated by a dynamo (Cowling's theorem), the model must be fully three-dimensional. (iii) The model must be rotating, because Coriolis forces are important to generate a flow pattern that is conducive for the dynamo process. The basic setup of geodynamo models is that of a rotating spherical shell of outer radius r_o and inner radius r_i , filled with a fluid of uniform conductivity, which represents Earth's outer core (Fig. 1). For this system the coupled equations for convection-driven flow and electromagnetic induction are solved. A detailed account on fundamental aspects of convection in rotating spheres is given in Jones (2015) and modeling aspects and the commonly employed numerical schemes are discussed in Christensen and Wicht (2015).

Dynamo Equations

The relevant magnetohydrodynamic equations are usually written in nondimensional form. A possible scheme for

scaling the equations is to use the shell thickness $D = r_o - r_i$ as length scale, the viscous diffusion time D^2/ν as time scale (ν is kinematic viscosity), $(\rho\Omega/\sigma)^{1/2}$ for the scale of the magnetic field $\mathbf{B}$, and the imposed temperature contrast ΔT between inner and outer boundary for temperature T (ρ is density, σ electrical conductivity, and Ω rotation rate). The Navier-Stokes equation for the velocity $\mathbf{u}$, augmented by rotational and electromagnetic forces, is

$$E \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) + 2\hat{\mathbf{z}} \times \mathbf{u} + \nabla \Pi = E \nabla^2 \mathbf{u} + \frac{RaE}{Pr} \frac{\mathbf{r}}{r_o} T + \frac{1}{P_m} (\nabla \times \mathbf{B}) \times \mathbf{B}. \quad (1)$$

The terms on the left hand side describe in order the inertial force, the Coriolis force (with $\hat{\mathbf{z}}$ the unit vector parallel to the rotation axis), and the gradient of the non-hydrostatic pressure Π . The terms on the right hand side stand for viscous friction, thermal buoyancy forces, and the Lorentz force.

The magnetic induction equation, obtained from Maxwell's equations and Ohm's law for a moving incompressible conductor, is

$$\frac{\partial \mathbf{B}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{B} = (\mathbf{B} \cdot \nabla) \mathbf{u} + \frac{1}{P_m} \nabla^2 \mathbf{B}, \quad (2)$$

where the second term on the LHS describes magnetic field advection and the terms on the RHS magnetic field generation and diffusion, respectively. Magnetic diffusion is a consequence of the ohmic resistance that damps the electrical currents associated with the magnetic field.

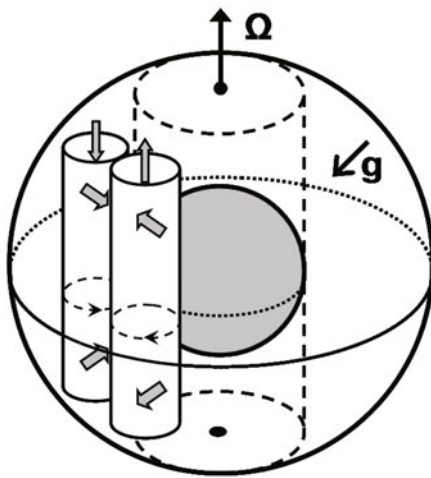
The advection-diffusion equation for temperature is

$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = \frac{1}{Pr} \nabla^2 T + \epsilon, \quad (3)$$

with a heat source term ϵ on the RHS. For compositional convection an equivalent equation holds where the concentration of light components replaces temperature. The set of equations is completed by the condition of incompressibility, which seems to be justified for Earth's core where density differences are moderate, and the condition that $\mathbf{B}$ is solenoidal:

$$\nabla \cdot \mathbf{u} = 0, \quad \nabla \cdot \mathbf{B} = 0. \quad (4)$$

The equations are complemented by boundary conditions for the velocity, usually $\mathbf{u} = 0$, fixed temperature or fixed heat flux at r_o and r_i , and a continuity condition for the magnetic field at r_o that links the field in the dynamo with an external potential field that decays with radius. Simple models assume the inner core to be insulating, but in many cases it is taken as an electrical conductor with the same conductivity as in the



Core Dynamo, Fig. 1 Columnar convection in a rotating spherical shell. The inner core tangent cylinder is shown by broken lines. Under Earth's core conditions the columns would be thinner and more numerous. (From Christensen et al. (2010). Copyright: Cambridge University Press)

outer core. It has been suggested that the finite conductivity of the inner core plays an important role for preventing frequent dipole reversals of the geodynamo, because a dipole reversal can only be completed successfully if the new polarity persists for the time scale of diffusive dipole decay in the inner core, $r_i^2/(\pi^2\lambda)$, which is several thousand years. However, this is contested by dynamo models that show little differences between simulations with a conducting or an insulating inner core.

Parameters

The four nondimensional control parameters in these equations are defined in Table 1. In dynamo models the values of most of these parameters fall short of their extreme values in the Earth's core. For practical reasons it is not possible to resolve the very small spatial structures that occur at realistic parameter values. The *Rayleigh number* Ra describes the ratio between buoyancy driving convection and retarding effects. In the core it is much larger than the critical value for the onset of convection, Ra_c , while models are more moderately supercritical (Table 1). The discrepancy is by at least eight orders of magnitude for the *Ekman number* E , the ratio of viscous forces to Coriolis forces. The very small core value of E indicates that viscous forces are negligible, except in very thin Ekman layers at the boundaries to the solid mantle and inner core. The *magnetic Prandtl number* Pm , the ratio of viscosity to magnetic diffusivity, is very small for liquid metals, but must be set to a value of order one to obtain a self-sustained dynamo in present models. Only the (hydrodynamic) *Prandtl number* Pr , the ratio of viscosity to thermal diffusivity, is of order one in the core and in dynamo models. In terms of physical parameters, the viscosity and the thermal diffusivity are far too large in the models, and in most cases, the rotation rate is too small.

Several dimensionless diagnostic numbers can be formed with the characteristic flow velocity U and magnetic field strength B in the dynamo (Table 1). The velocity in Earth's core can be estimated from the geomagnetic secular variation under the assumption that the magnetic flux at the core surface

is approximately frozen into the moving fluid (Alfvén's theorem), which is the case when the diffusion term in Eq. 2 can be neglected. In this case the variation of the radial magnetic field component B_r at the core-mantle boundary can be written

$$\partial B_r / \partial t + \nabla_h \cdot (\mathbf{u} B_r) = 0. \quad (5)$$

where the suffix h indicates the horizontal part of the divergence operator. Additional assumptions are needed to infer the velocity at the top of the core (below the Ekman layer) from the observed B_r and its time derivative, but the flow pattern obtained from these inversions are broadly similar and the characteristic velocity is approximately 0.5 mm/s (Holme 2015). The characteristic field strength inside Earth's core is probably in the range of a few mT (milliTesla), as will be discussed further below.

The most important diagnostic number is the *magnetic Reynolds number* Rm , which describes the ratio of magnetic advection and induction to magnetic diffusion. In order for induction effects in a dynamo to overcome the diffusive dissipation of magnetic field, it is expected that Rm must be larger than one. Practically, it turns out that self-sustained magnetic field generation occurs in geodynamo models when Rm exceeds approximately 40. Using $U \approx 0.5$ mm/s the magnetic Reynolds number is of order 1000, safely above the critical limit. Also, $Rm \gg 1$ is a condition for Eq. 5 to apply. Still, the value of Rm in the Earth's core is moderate and can be handled in direct numerical simulations. This contrasts with the much larger values of Rm in other cosmic dynamos, which requires that magnetic induction and diffusion effects at small unresolved scales are parameterized by applying concepts of the mean-field dynamo theory. The ability to solve for magnetic induction and diffusion in Earth's core directly without parameterizations is probably the main reason for the success of geodynamo models.

The (hydrodynamic) *Reynolds number* Re is much smaller in the models than it is in the core where it is of the order 10^9 , indicating a highly turbulent regime. Consequently, dynamo

Core Dynamo, Table 1 Dynamo parameters. α thermal expansivity, g_o gravity at core surface, ΔT superadiabatic temperature contrast across core, κ thermal diffusivity, ν kinematic viscosity, Ω rotation rate, D outer

core thickness, $\lambda = 1/(\mu_o\sigma)$ magnetic diffusivity, U characteristic flow velocity, B characteristic magnetic field strength, μ_o magnetic permeability, ρ density

	Control parameters			
	Rayleigh no.	Ekman no.	Magn. Prandtl no.	Prandtl no.
Definition	$Ra = \alpha g_o \Delta T D^3 / (\kappa \nu)$	$E = \nu / (\Omega D^2)$	$Pm = \nu / \lambda$	$Pr = \nu / \kappa$
Core Models	$10^4 Ra_c$ (1–100) Ra_c	10^{-15} – 10^{-14} 10^{-7} – 10^{-2}	10^{-6} – 10^{-5} 0.1–10	0.1–1 0.1–1
	Diagnostic numbers			
	Magn. Reynolds no.	Reynolds no.	Rossby no.	Elsasser no.
Definition	$Rm = UD/\lambda$	$Re = UD/\nu$	$Ro = U/(\Omega D)$	$\Lambda = B^2/(2\mu_o\lambda\rho\Omega)$
Core Models	10^3 40–2000	10^9 <10,000	10^{-7} 10^{-2} – 10^{-4}	0.1–10 1–100

models miss the small eddies that are expected to exist in the core. Viscosity kicks in only at length scales of some meters to suppress even smaller structures in the flow. However, the small eddies are unimportant for the magnetic induction process, because the magnetic diffusivity, which is $Pm^{-1} = 10^6$ times larger than the kinematic viscosity, homogenizes the magnetic field at scales of less than several tens of kilometers.

The *Elsasser number* Λ describes the ratio of Lorentz forces to Coriolis forces. In Earth's core both forces are assumed to be of leading order and to balance approximately in the so-called magnetostrophic state, which is often associated with an Elsasser number of order one. Many present dynamo models match this value of Λ .

The *Rossby number* Ro describes the ratio of inertial forces to the Coriolis force and is very small in the core. In models it is less than one, although larger than the core value. Instead of forming the Rossby number with the global length scale D , a more appropriate measure for the ratio of the two forces is obtained by using the characteristic length scale ℓ of the flow (local Rossby number Ro_ℓ). The value of Ro_ℓ is difficult to estimate for the core, but model values and the Earth value may be in better agreement than they are for Ro .

Flow in the Earth's core is thought to be in a MAC-balance, from **M**agnetic, **A**rchimedian, **C**oriolis, that is, these three forces balance to first order in Eq. 1 when putting pressure forces aside. Geodynamo models have been suspected to miss the correct force balance and to be controlled by viscous forces. However, a direct evaluation of force contributions shows that at least in simulations with an Ekman number smaller than 10^{-5} a MAC-balance is approached (Yadav et al. 2016). Simulations that have further pushed the viscosity (or E and Pm) to smaller values (Schaeffer et al. 2017) basically confirm the validity of many results that have been obtained with more viscous geodynamo models.

Energetics of the Geodynamo

Although the decay of radioactive isotopes in the core may contribute to some degree to the core heat flux, most of the buoyancy that drives convection in the outer core is associated with secular cooling. The sluggish convection in the mantle controls the heat flux through the core-mantle boundary in terms of its magnitude and its spatial distribution, which is expected to be rather heterogeneous.

Estimates put the heat flow at the core-mantle boundary into the range of 7–17 TW (Nimmo 2015). For thermal convection to occur, the temperature gradient must reach or exceed the adiabatic gradient. Because iron is a comparatively good conductor 10–15 TW can be transported by thermal conduction along an adiabatic gradient. This fraction of the heat flux is lost for driving the dynamo. It is possible that the

top of the core is thermally stably stratified. A consequence of core cooling is the slow growth of the solid inner core by freezing iron on its outer boundary. The associated latent heat release balances roughly one half of the heat loss from the core and renders the deep fluid core thermally unstable even if the top layer is not. The relative depletion of light alloying elements in the solid core is balanced by an enrichment in the residual fluid above the growing inner core, which drives compositional convection. Estimates for the work done by thermal and compositional buoyancy forces are in the range 0.8–5 TW (Nimmo 2015). Compositional buoyancy contributes roughly two-thirds to it. This energy flux is available to generate magnetic field and balance the associated ohmic dissipation. Viscous dissipation is usually assumed to play a minor role in the core.

Earth has a global magnetic field since at least 3.5 Gyr. At times before the inner core nucleated, estimated to have occurred 0.5–1.5 Gyr ago, only thermal convection could have driven the geodynamo. If the present core heat flux is less than the flux that can be conducted along an adiabat, it must have been significantly higher prior to inner core nucleation when compositional driving had been unavailable. A possible alternative is the exsolution of mantle components that had been dissolved in iron at very high temperatures during the Earth's formation. This would drive a form of compositional convection that is independent of the existence of an inner core.

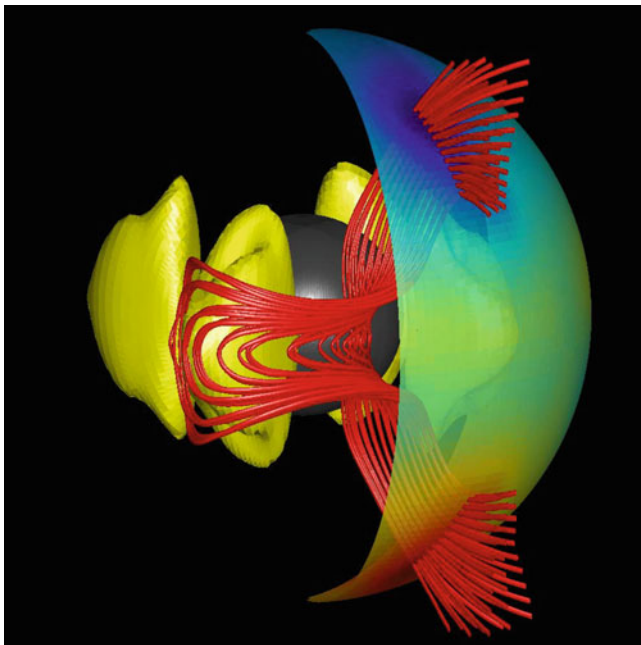
Flow Pattern and Dynamo Mechanism

The pattern of flow is strongly controlled by the Coriolis force. It is different for the regions inside and outside the *inner core tangent cylinder*. This is an imaginary cylinder aligned with the rotation axis and touching the inner core at the equator, which is shown by broken lines in Fig. 1.

Flow Outside Tangent Cylinder

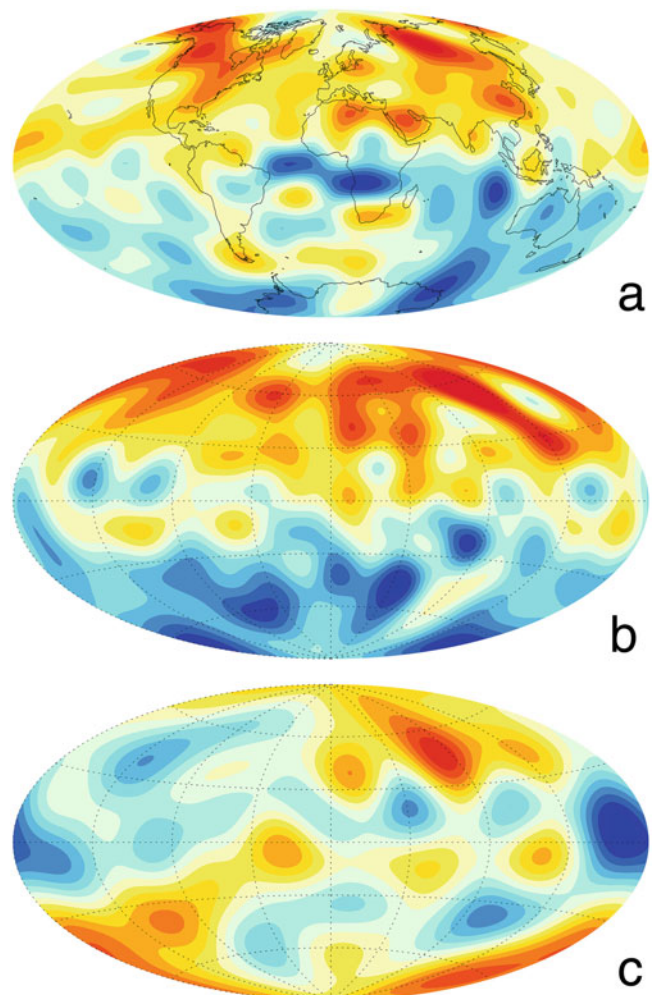
Outside the tangent cylinder, the flow is organized in convection columns aligned with the vector of rotation, with the flow encircling the column axis (Fig. 1). Superimposed is a circulation along the column axis that is directed toward the equatorial plane in columns with a cyclonic sense of rotation and away from the equator in anticyclonic columns. This results in helical flow, in which particles move on corkscrew-like trajectories. The helicity, $H = \mathbf{u} \cdot (\nabla \times \mathbf{u})$, is coherently negative in the northern hemisphere and positive in the southern hemisphere. Such helical flow is the basis for the α -effect, which can convert an axisymmetric toroidal field into an axisymmetric poloidal field, or vice versa, through intermediate steps involving non-axisymmetric small-scale fields (Moffatt 1978). Because purely axisymmetric fields cannot be maintained by a dynamo, this mechanism plays an important

role in the understanding of the dynamo process. The α -effect has been introduced in the context of mean-field dynamo theory, where it describes the induction effects of small unresolved flow scales. In geodynamo models and probably to some extent in the Earth's core it is due to fairly large-scaled convection columns (Busse 1975). The models suggest that the axisymmetric toroidal field in the Earth's core (which is invisible from the outside) consists of a torus of westward directed field north of the equatorial plane and eastward field south of the equatorial plane. The helical columnar flow converts this to a dipolar poloidal field whose field lines can leave the core (Fig. 2). In some dynamo models the toroidal magnetic field is regenerated by the same “macroscopic” α -effect from the poloidal magnetic field (α^2 dynamo, Olson et al. 1999). In other geodynamo models the shearing of poloidal field lines by differential rotation, the Ω -effect, has been found to play an important role for toroidal field generation ($\alpha\Omega$ -dynamo), but essentially all models confirm that the helical columnar convection produces the dipolar poloidal field component. Although at low Ekman number and high Rayleigh number the simple columns give way to a more complex and turbulent pattern, the vortices are still elongated in the direction of the rotation axis and retain statistically a coherent helicity in each hemisphere.



Core Dynamo, Fig. 2 Bundle of magnetic field lines in a simple geodynamo model ($E = 10^{-3}$, $Ra/Ra_c = 1.8$, $Pm = 5$, $Pr = 1$). The yellow blobs are anticyclonic vortices. Strong inward magnetic flux is shown in blue on the outer core boundary and outward flux in red. The NS-running part of the field lines represents poloidal magnetic field and the EW-directed segments form the axisymmetric toroidal magnetic field, which consists of two bundles of opposite directions in the two hemispheres. The same field structure repeats periodically in longitude. (Courtesy of Julien Aubert)

Strong concentrations of magnetic flux in the present geomagnetic field centered below North America and Siberia, which have counterparts at the same longitudes in the southern hemisphere, are interpreted as evidence for the presence of convection columns in the fluid core (Gubbins and Bloxham 1987) (Fig. 3a). The convergence at the core-mantle boundary associated with the equatorward flow in the axis of cyclonic columns (Fig. 1) concentrates the poloidal magnetic field lines into these flux lobes, assuming that the frozen flux assumption is approximately valid. In simple models, such as that in Fig. 2, the flux lobes are associated with a single columnar vortex. Dynamo models at lower Ekman number and strongly supercritical Rayleigh number show similar high-latitude flux lobes in their magnetic fields when they are low-pass filtered to the resolution of the geomagnetic core



Core Dynamo, Fig. 3 Radial magnetic field at the core-mantle boundary, red (full lines) inward and blue (broken lines) outward. (a) Geomagnetic field in 2010. (b) Dynamo model with dominant dipole contribution ($E = 10^{-5}$, $Pm = 0.8$, $Pr = 1$, $Ra/Ra_c = 114$). (c) Multipolar dynamo model ($E = 10^{-5}$, $Pm = 0.5$, $Pr = 1$, $Ra/Ra_c = 161$). The model magnetic fields are low-pass filtered to the resolution of the geomagnetic field

field, roughly up to spherical harmonic degree 13 (e.g., Fig. 3b). In this case, and probably in the Earth's core, the large-scale magnetic flux concentrations are due to the superposition of strong but complex and smaller-scaled magnetic field structures, each associated with a separate convection column.

Flow Inside Tangent Cylinder

In the polar regions inside the tangent cylinder, many dynamo models show an upwelling plume at or near the rotation axis. Because of the action of Coriolis forces, the plume is associated with an eastward vortex motion near the inner core and by westward vortex motion near the core-mantle boundary, which is called a thermal wind circulation. A westward vortex motion in the polar cap region at the top of the core has also been inferred from the observed geomagnetic secular variation (Hulot et al. 2002). A polar plume is associated with flow that diverges away from the rotation axis below the core surface, which disperses poloidal magnetic field lines. This is seen in many geodynamo models (e.g., Fig. 3b) and can explain the weak magnetic flux in the north polar region of the core (Fig. 3a). It has been suggested that the solid inner core couples electromagnetically to the eastward vortex motion of the overlying fluid, which leads to a superrotation of the inner core with respect to the mantle and crust. While perfect coupling to the outer core flow would lead to superrotation rates in excess of a degree per year (Glatzmaier and Roberts 1996), gravitational coupling between the inner core and the mantle can slow down the rate. Seismological evidence indicates a superrotation at a fraction of a degree per year.

Field Properties

Field Strength Inside the Core

How strong is the magnetic field inside Earth's core? Observations allow to constrain the strength of the poloidal field at the core-mantle boundary, which is presently 0.26 mT for the dipole field alone and 0.39 mT in spherical harmonic degrees from 1 to 13. It is not known how much the small unresolved length scale add to the field at the top of the core, but possibly they double this value. Inside Earth's core the poloidal field is expected to be stronger than at the outer boundary and the toroidal component also contributes. Suggestions that the toroidal field, if generated by an intense Ω -effect, could be much stronger than the poloidal field are not supported by geodynamo simulations. In most of them the magnetic energy is equipartitioned between the two field components. The dynamo models suggest that the characteristic field strength inside the dynamo is around 5–15 times the dipole field strength on the outer boundary, which would put it into the range of 1–4 mT.

The field strength inside Earth's core can be probed directly by the observation of torsional waves and oscillations. To the extent that viscous and inertial forces are negligible, the azimuthal Lorentz force integrated over the surface of cylinders that are aligned with the rotation axis (geostrophic cylinders) must vanish in the so-called Taylor state. Perturbations of the Taylor state will result in torsional oscillations or torsional waves, consisting of an oscillatory differential rotation between the geostrophic cylinders. Magnetic field lines that permeate adjacent cylinders act like elastic strings and provide a restoring force when the cylinders are displaced relative to each other. The frequency of torsional eigenmodes and the wavespeed of traveling waves are directly related to the *rms* strength of the field component B_s in the cylindrically radial direction. Some component of the large-scale flow in the core, which has been inferred from secular variation using Eq. 5, can be interpreted as a torsional wave that recurs every 6 years. This is in good agreement with a signal in the length-of-day variation at this period, that is, a change of the rotation rate of the solid part of the Earth caused by angular momentum exchange between core and mantle. From the time needed by the torsional wave to traverse the fluid core, approximately 4 years, a mean value of B_s in the range 2–3 mT has been deduced (Gillet et al. 2010).

Scaling Laws

An important question is what controls the magnetic field strength of the geodynamo and of planetary dynamos in general. It has been suggested that because of the requirement for a magnetostrophic state, the magnetic field inside Earth's core is driven toward a value that makes the Elsasser number (Table 1) to be of order one, that is,

$$B \propto (2\rho\Omega/\sigma)^{1/2}. \quad (6)$$

This results in $B \approx 2$ mT, in good agreement with the above estimates. A different line of argument is based on the power available to balance ohmic dissipation. With some assumptions on how the length scale of the magnetic field in the core depends on the vigor of convection, this leads to

$$B \propto \mu_o^{1/2} \rho^{1/6} P^{1/3}, \quad (7)$$

where $P = W/(4\pi r_o^2)$ stands for the power per unit area of the core surface. With the current estimates for W Eq. 7 results in a similar value of the field strength as the Elsasser number rule. However, the two scaling rules are radically different; in Eq. 6 the field strength is independent of the energy flux and in Eq. 7 it does not depend on rotation rate or electrical conductivity. Sets of dynamo simulations that cover a substantial range in these properties, or their counterparts in terms of the nondimensional control parameters, support the power-

based scaling law (Eq. 7). Furthermore, they show no significant dependence of the field strength on the viscosity. This may indicate that although current dynamo simulations are far too viscous (too high Ekman number), the viscosity is sufficiently low to not play a dominant role. The validity of the power-based scaling rule is also supported by the finding that it gives the right order of magnitude of the field strength for various convection-driven dynamos in rapidly rotating cosmic objects including Earth, Jupiter, and low-mass stars. A detailed discussion of dynamo scaling laws is given in Christensen (2010).

Magnetic Field Geometry

While the rotation rate seems to have little influence on the field strength inside dynamo models, it probably controls the magnetic field morphology. In some geodynamo models the magnetic field on the outer boundary resembles Earth's field at the core-mantle boundary (Fig. 3a, b) in terms of a dominant axial dipole and also of structures associated with multipole contributions. In other dynamo models the dipole contribution is weak. In some cases the field is dominated by the axial quadrupole or is restricted to one hemisphere, but in most cases it is irregular with contributions from a broad spectrum of multipole components without a pronounced dominance of a particular mode (Fig. 3c). The selection of the field geometry was found to depend on the (local) Rossby number. At low values (rapid rotation and moderate flow velocities) dynamos are dipolar, whereas for high values (slower rotation and fast flow) they tend to be multipolar. In addition boundary conditions and the mode of driving convection, that is, either by secular cooling of the fluid core or by the latent heat and light element flux from the growing solid inner core, can play a role. The driving mode has changed over geological time. Before the Earth's inner core nucleated 0.5–1.5 Gyr ago the dynamo has probably been driven more weakly than today by secular cooling alone, whereas thereafter the buoyancy source is mainly located at the inner core boundary. However, dynamo simulations for the different scenarios (Aubert et al. 2009) suggest that neither the geometry of the magnetic field nor its strength has changed very strongly. This is in accord with the lack of paleomagnetic evidence for drastic changes in the geomagnetic field since the Archean.

Secular Variation

Time Scales

The characteristic time scale of secular variation can be defined by the square root of the power of the magnetic field over the power in the time derivative of the field. This can be done separately for each harmonic degree n . The time scales decrease with increasing n . Aside from this scale dependence,

the secular variation time has been found in dynamo models to vary with the inverse of the magnetic Reynolds number (Christensen and Tilgner 2004), as would be expected if most of the secular variation is caused by the frozen-flux advection of magnetic field. Using the observed secular variation time scales to estimate the magnetic Reynolds number (or flow velocity) in the core gives essentially the same result, $Rm \approx 1000$, as the more complex inversion of the secular variation for the core flow.

Advection vs. Diffusion of Field

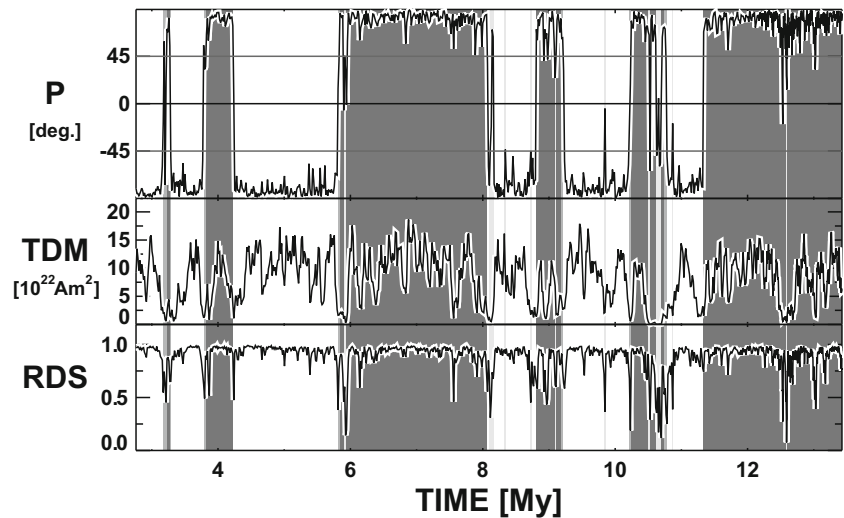
The westward drift of magnetic field structures is a long-standing issue. Some dynamo models show a persistent and global westward drift, whereas others do not. Over the past 400 years magnetic field structure have systematically migrated westward flow in the Atlantic hemisphere, but not in the Pacific hemisphere, where the amplitude of secular variation is generally weaker. Dynamo models support the notion that such hemispherical dichotomy can be caused by coupling between the mantle and the core, in particular by the heterogeneity of the core heat flow caused by lower mantle thermal structure.

While much of the geomagnetic secular variation can be attributed to frozen-flux advection in a large-scale circulation, magnetic diffusion must play some role, in particular diffusion in the radial direction. It has been suggested that intense flux spots of both polarities, which are found in the geomagnetic field at low latitudes under Africa and the Atlantic (Fig. 3a), have been formed by the expulsion of toroidal magnetic flux bundles through the core surface Bloxham (1986), somewhat analogous to the formation of sunspots. Because the toroidal field is confined to the core, the emergence of such spots necessarily involves diffusion, whereas their migration can be caused by frozen flux advection. Some geodynamo simulations support such origin of equatorial flux spots. In these simulations the columnar convection acts on a pair of toroidal flux bundles with opposite polarities and produces (anti)symmetric pairs of spots north and south of the equator. However, the geomagnetic low-latitude field pattern conforms at best marginally with this picture. An alternative explanation to the concept of field expulsion by convective columns and migration due to a large-scale westward flow is that propagating hydromagnetic waves in the core create a periodic pattern of upwelling and downwelling that produces the belt of wave-train-like magnetic field structures in the low-latitude Atlantic hemisphere.

Magnetic Field Reversals

Several dynamo simulations show stochastic dipole reversals. Often the details resemble remarkably well what is known from the paleomagnetic record about geomagnetic reversals

Core Dynamo, Fig. 4 Time series of dipole tilt, true dipole moment (TRM) and dipole strength relative to total field strength (RDS) from a numerical dynamo simulation (Wicht 2005). (Copyright Blackwell)



(e.g., Wicht et al. 2009). They are rather short events that interrupt long periods of stable polarity. Aside from complete reversals also excursions (aborted reversals) occur. During a reversal the magnetic field is weak and not dipole-dominated (Fig. 4). Most model studies of reversal behavior have been performed with simple dynamo models at rather high values of the Ekman number, because long simulation times are needed to capture a fair number of these events. Reversing dipolar dynamos are typically found in a parameter regime between that of stable dipolar dynamos and of multipolar dynamos. Driving convection more strongly for otherwise identical conditions leads to very frequent and erratic dipole reversals that have little significance because the dipole is hardly ever dominating the magnetic field. Lowering the Prandtl number has the same effect, suggesting that strengthening the nonlinear inertial forces relative to the Coriolis force favors non-dipolar dynamo. Reducing the convective driving or increasing the Prandtl number results in strong dipolar fields that rarely or never reverse. In addition, boundary conditions affect the reversal frequency. In particular, it has been shown that the pattern of heat flux at the core-mantle boundary, which is controlled by the thermal state of the lowermost mantle and varies on the long time scale of mantle convection, has a strong influence on the occurrence of reversals. This supports the assumption that the drastic secular change in the reversal rate over the past 200 Myr, which is observed in the paleomagnetic record, is caused by changes in the amplitude and pattern of core heat flow associated with a reorganization of mantle convection, such as formation and fading of superplumes. Although model studies have led to a number of suggestions concerning the cause for individual reversals, for example by the formation of strong new plumes in the core, so far no generally accepted concept for the reversal mechanism has emerged.

Summary

Helical convection columns outside the inner core tangent cylinder region play an essential role for our present understanding of the geodynamo. Theory and numerical simulations suggest that they are a fundamental element of the circulation in the outer core and that they play a key role for the generation of the Earth's dipole-dominated magnetic field. Certain details of the magnetic field structure at the core-mantle boundary are explained by such columns. Somewhat less certain, but supported by several lines of evidence, is the notion of rising plumes in the tangent cylinder near the rotation axis that are associated with weak magnetic field and westward vortex flow in the polar regions at the core-mantle boundary. Much of our current understanding is based on numerical dynamo models that are far too viscous compared to the core. Even if models that use actual core values of the viscosity are out of reach, recent simulations have lowered the viscosity to an extent that was previously numerically unfeasible. They suggest that many of the basic findings are not compromised by unrealistic viscosity values. Modeling supports the notion that the magnetic field strength is controlled by the available power that drives the dynamo. Several dynamo models show dipole reversals that resemble in many respects geomagnetic reversals, but a full understanding of what causes reversals or controls the frequency of geomagnetic reversals is missing.

Cross-References

- [Core-Mantle Coupling](#)
- [Differential Rotation of the Earth's Inner Core](#)
- [Earth's Structure, Core](#)

- Energy Budget of the Earth
- Geomagnetic Excursions
- Geomagnetic Field, Global Pattern
- Geomagnetic Field, Polarity Reversals
- Geomagnetic Field, Secular Variation
- Geomagnetic Field, Theory
- Paleomagnetic Field Intensity
- Radioactivity in Earth's Core

Bibliography

- Aubert J, Labrosse S, Poitou C (2009) Modelling the paleo-evolution of the geodynamo. *Geophys J Int* 179:1414–1429
- Bloxham J (1986) The expulsion of magnetic flux from the Earth's core. *Philos Trans R Soc Lond* 87:669–678
- Busse FH (1975) Thermal instabilities in rapidly rotating systems. *Geophys J R Astron Soc* 42:437–459
- Christensen UR (2010) Dynamo scaling laws: applications to the planets. *Space Sci Rev* 152:565–590
- Christensen UR, Tilgner A (2004) Power requirement of the geodynamo from ohmic losses in numerical and laboratory dynamos. *Nature* 429:169–171
- Christensen UR, Wicht J (2015) Numerical dynamo simulations. In: Schubert G (ed) *Treatise on geophysics*, 2nd edn. Core dynamics, vol 8. Elsevier, Amsterdam, pp 245–277
- Christensen UR, Aubert J, Hulot G (2010) Conditions for earth-like geodynamo models. *Earth Planet Sci Lett* 296:487–496
- Gillet N, Jault D, Canet E, Fournier A (2010) Fast torsional waves and strong magnetic fields within the Earth's core. *Nature* 465:74–77
- Glatzmaier GA, Roberts PH (1995) A three-dimensional convective dynamo solution with rotating and finitely conducting inner core and mantle. *Phys Earth Planet Inter* 91:63–75
- Glatzmaier GA, Roberts PH (1996) Rotation and magnetism of Earth's inner core. *Science* 274:1887–1891
- Gubbins D, Bloxham J (1987) Morphology of the geomagnetic field and implications for the geodynamo. *Nature* 325:509–511
- Holme R (2015) Large-scale flow in the core. In: Schubert G (ed) *Treatise on geophysics*. Core dynamics, vol 8. Elsevier, Amsterdam, pp 91–113
- Hulot G, Eymin C, Langlais B, Mandea M, Olson N (2002) Small-scale structure of the geodynamo inferred from Ørsted and Magsat satellite data. *Nature* 416:620–623
- Jones CA (2015) Thermal and compositional convection in the outer core. In: Schubert G (ed) *Treatise on geophysics*, 2nd edn, vol 8. Elsevier, Amsterdam, pp 115–159
- Kageyama A, Sato T (1995) Computer simulation of a magnetohydrodynamic dynamo. II. *Phys Plasmas* 2:1421–1431
- Kuang W, Bloxham J (1997) An earth-like numerical dynamo model. *Nature* 389:371–374
- Moffatt HK (1978) *Magnetic field generation in electrically conducting fluids*. Cambridge University Press, Cambridge
- Nimmo F (2015) Energetics of the core. In: Schubert G (ed) *Treatise on geophysics*, 2nd edn. Core dynamics, vol 8. Elsevier, Amsterdam, pp 27–55
- Olson P, Christensen UR, Glatzmaier GA (1999) Numerical modeling of the geodynamo: mechanism of field generation and equilibration. *J Geophys Res* 104:10,383–10,404
- Schaeffer N, Jault D, Nataf H-C, Fournier A (2017) Turbulent geodynamo simulations: a leap towards Earth's core. *Geophys J Int* 211:1–29
- Wicht J (2005) Palaeomagnetic interpretation of dynamo simulations. *Geophys J Int* 162:371–380
- Wicht J, Stellmach S, Harder H (2009) Numerical models of the geodynamo: from fundamental Cartesian models to 3D simulations of field reversals. In: Glassmeier KH, Soffel H, Negendank JFW (eds) *Geomagnetic field variations*. Springer, Berlin, pp 107–158
- Yadav RK, Gastine T, Christensen UR, Wolk SJ, Poppenhaeger K (2016) Approaching a realistic force balance in geodynamo simulations, *proceed. Nat Acad Sci* 113:12065–12070

Core-Mantle Coupling

Paul H. Roberts¹ and Jonathan M. Aurnou²

¹Institute of Geophysics and Planetary Physics, University of California, Los Angeles, CA, USA

²Department of Earth and Space Sciences, University of California, Los Angeles, CA, USA

Definition, Scope, and Aims

The Earth is not a perfect timekeeper, and the spectrum of the variations in the mantle's angular velocity $\hat{\Omega}$ spans a wide range of frequencies. Of particular interest here are the comparatively large amplitude decadal and semidecadal variations in which changes in length of day, P , of up to 2 ms occur. These would not be explained even if the global circulations of the atmosphere and oceans could be reversed. This is confirmed by a more detailed argument given in our recent review (Roberts and Aurnou 2012), which will be referred to here as “RA12.”

The origin of these length of day (LOD) variations must be sought in the Earth's core, and Fig. 1 suggests that the task is not an easy one. Fig. 1b shows dP/dt , derived by differentiating smoothed LOD data from the last half century, with atmospheric, oceanic, and tidal signals removed; semi-decadal time variations are clearly seen with a period τ_{LOD} of about 6 years (e.g., Abarca del Rio et al. 2000). Figure 1c shows the implied $\hat{\Gamma}_z$ as a function of time t , where $\hat{\Gamma}_z$ is the component parallel to the polar axis O_z of the torque $\hat{\Gamma}$ exerted by the core on the mantle. This shows that $\hat{\Gamma}_z$ of nearly 10^{18} Nm is generated. We shall call this the “target torque” and seek its origin.

In addition, to the semi-decadal oscillations, larger variations seem to exist having longer periods, one of which is estimated to have a roughly 60 year period (e.g., Roberts et al. 2007). This time scale is reminiscent of the geomagnetic secular variation, and it is natural to seek a connection between them (e.g., Braginsky 1970; Gillet et al. 2010). Figure 1c supports this quest. It shows the LOD data (Gross 2001; Holme and de Viron 2005) plotted against the estimated LOD variations inferred from the core flow models of Jackson (1997) that are based on inversion of geomagnetic secular variation data. The qualitative agreement implies that the

variations in LOD are due to core-mantle angular momentum exchange associated with magnetohydrodynamic (MHD) processes in the core.

This review will focus on variations in LOD, i.e., changes in $\Omega_z (\approx \Omega)$. Precession and nutation of the Earth's axis, which describe variations in Ω_x and Ω_y , are also phenomena that cannot be satisfactorily explained without invoking core-mantle torques, but they are beyond the scope of this review. It will become clear that torques on the Solid Inner Core (SIC) are also of interest. Variables in the SIC will be distinguished by a tilde ($\sim$) and those in the mantle by a hat ($\hat{\cdot}$). Except when making general statements, unadorned letters will refer to variables in the Fluid Outer Core (FOC).

The Four Coupling Processes

The torque on the mantle about the geocenter O is

$$\hat{\Gamma} = \oint_S \mathbf{r} \times \hat{\mathbf{T}} dA, \quad \text{where } \hat{T}_i = -\hat{S}_{ij} n_j, \quad (1a, b)$$

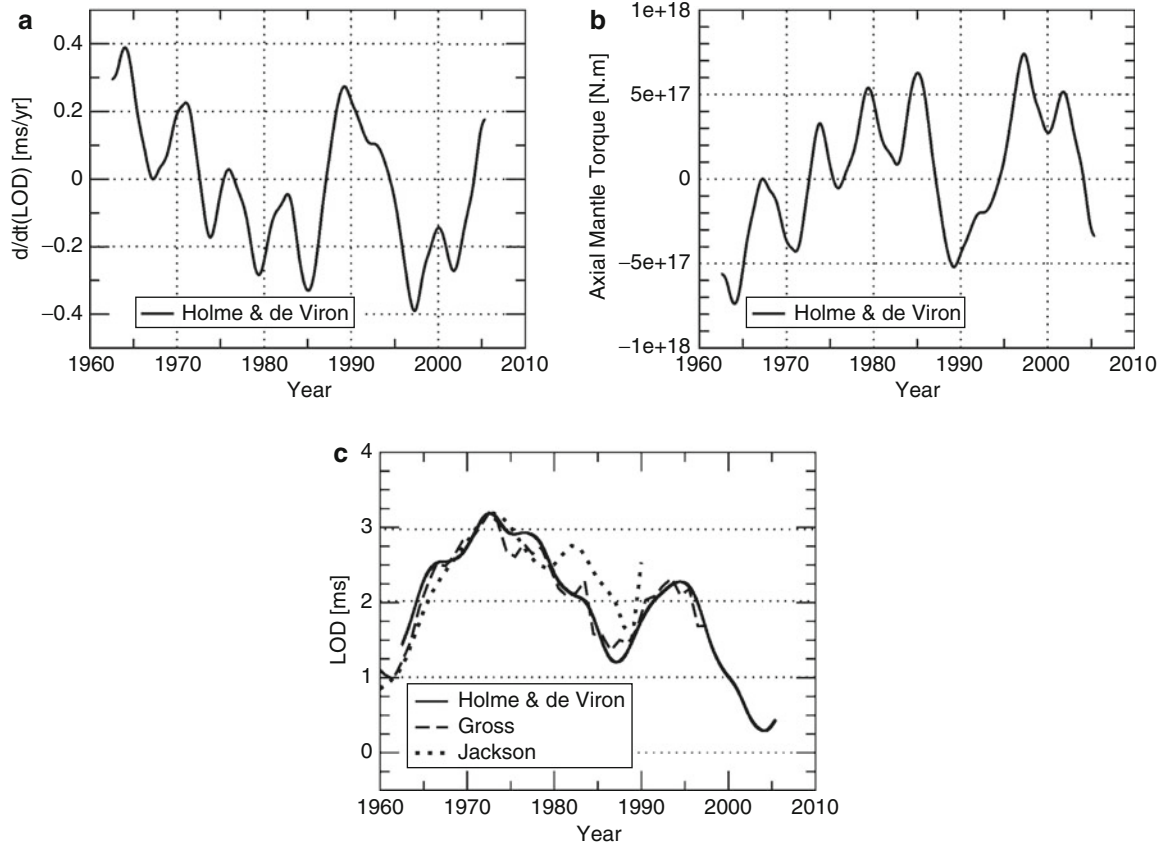
sometimes called the “surface traction,” is the stress associated with the normal and $\hat{S}_{ij}$ is the total stress tensor; $\mathbf{r} = r\mathbf{1}_r$ is the radius vector from O, $r = |\mathbf{r}|$. The minus sign in Eq. 1b arises because our unit normal, $\mathbf{n}$, to $\hat{\mathbf{S}}$, the Core Mantle Boundary (CMB), points into the mantle. The axial torque, i.e., the component of $\hat{\Gamma}$ along the rotation axis Oz, is

$$\hat{\Gamma}_z = - \oint_S s \hat{S}_{\phi n} dA, \quad (1c)$$

where s is distance from the z -axis. For the torque on the SIC, we have similarly

$$\begin{aligned} \tilde{\Gamma} &= \oint_S \mathbf{r} \times \tilde{\mathbf{T}} dA, & \tilde{T}_i &= \tilde{S}_{ij} n_j, \\ \tilde{\Gamma}_z &= \oint_S s \tilde{S}_{\phi n} dA; \end{aligned} \quad (1d, e, f)$$

there is no minus sign in Eq. 1e, f because our unit normal, $\mathbf{n}$, to the Inner Core Boundary (ICB) points out of the SIC.



Core-Mantle Coupling, Fig. 1 (a) LOD time series data from Holme and de Viron (2005) and Gross (2001) compared with LOD model from the “smooth” core flow inversion of Jackson (1997). The time series include the variation due to lunar tidal drag. (b) Temporal derivative, dP/dt , of the smoothed time series data of Holme and de Viron (2005), where P is the

length of day. (c) The implied axial torque, $\hat{\Gamma}_z = -(2\pi\hat{C}/P^2) dP/dt$, on the mantle, where $\hat{C}$ is the axial moment of inertia of the mantle. The LOD in panel (a) is arbitrary to within an additive constant, chosen here to obtain agreement at 1972.5 with Gross (2001)

Equation 1a tacitly assumes that the core alone exerts a torque on the mantle; sources of torque from outside the Earth are ignored, and therefore

$$\begin{aligned}\mathbf{\Gamma} + \hat{\mathbf{\Gamma}} + \tilde{\mathbf{\Gamma}} &= \mathbf{0}, \\ \mathbf{M} + \hat{\mathbf{M}} + \tilde{\mathbf{M}} &= \text{constant},\end{aligned}\quad (1g, h)$$

where $\mathbf{\Gamma}$ is the torque on the FOC; $\mathbf{M}$, $\hat{\mathbf{M}}$ and $\tilde{\mathbf{M}}$ are the corresponding angular momenta. Eq. 1g implies that, if one of the three torques changes, so do one or two of the others, in the opposite sense. The system is, in this respect, self-regulating. We shall treat the mantle and SIC as rigid bodies below.

Stress is exerted on the CMB in four ways: through viscosity, topography, gravity, and magnetic field, and there are correspondingly four parts to each of S_{ij} , $\mathbf{T}$ and $\mathbf{\Gamma}$, e.g., $\hat{\mathbf{\Gamma}} = \hat{\mathbf{\Gamma}}^V + \hat{\mathbf{\Gamma}}^T + \hat{\mathbf{\Gamma}}^G + \hat{\mathbf{\Gamma}}^M$. The topographic and gravitational torques depend on the non-sphericity of $\hat{\mathbf{S}}$ and $\tilde{\mathbf{S}}$. For the others, negligible error is made by replacing the CMB and ICB by spheres, $\hat{\mathbf{S}}$, and $\tilde{\mathbf{S}}$, of radii r_o and r_i . The torques will be estimated below.

The Viscous Torque

Assuming uniform fluid density, ρ , and kinematic viscosity, ν , the part of the viscous stress tensor, $\hat{S}_{ij}^V$, responsible for $\hat{T}_i^V$ is $\rho\nu\nabla_j V_i$, i.e., $\hat{\mathbf{T}}^V = -\rho\nu(\mathbf{n} \cdot \nabla)\mathbf{V} \approx -\rho\nu\partial_r\mathbf{V}$ so that, by Eq. 1a, c,

$$\begin{aligned}\hat{\mathbf{\Gamma}}^V &= -\rho\nu \oint_{\hat{\mathbf{S}}} \mathbf{r} \times \partial_r \mathbf{V} \, dA, \\ \hat{\Gamma}_z^V &= -\rho\nu \oint_{\hat{\mathbf{S}}} s \partial_r V_\phi \, dA.\end{aligned}\quad (2a, b)$$

These torques involve only the radial gradient of the fluid velocity, $\mathbf{V}$. This tends to drag the mantle in the direction of the subsurface flow.

The viscosity of core fluid is hard to estimate. First principles calculations (de Wijs et al. 1998; Dobson et al. 2000; Vočadlo et al. 2000) suggest that ν at the CMB is within a factor of 3 of $10^{-6} \text{ m}^2 \text{ s}^{-1}$. Suppose the viscous stress transfers momentum between the core and mantle over a length scale $d_v = E^{1/2} r_o \sim 0.2 \text{ m}$, which is the thickness of the laminar Ekman boundary layer. Here

$$E = \nu / \Omega \mathcal{L}^2 \quad (2c)$$

is the Ekman number ($\approx 3 \times 10^{-15}$), which quantifies the ratio of the viscous and Coriolis forces; $\mathcal{L}$ is a characteristic scale of motion outside the boundary layer, for which we take $\mathcal{L} = r_o$. Assuming velocities of order $\mathcal{V} \approx 10^{-4} \text{ m s}^{-1}$ (e.g., Jackson

1997), the surface fraction, $\hat{T}^V \approx \rho\nu\mathcal{V}/d_v$, is about 10^{-5} N m^{-2} , implying that $\hat{\mathbf{\Gamma}}^V \approx 5 \times 10^{14} \text{ Nm}$.

In a highly turbulent medium such as the FOC, small scale motions transport macroscale quantities, such as angular momentum, far more effectively than molecular diffusion. When these scales are too small to be resolved by numerical computations, they are termed “sub-grid scales.” Their effect must then be included in some other way. Appeal is often made to an analogy with molecular diffusion. Molecular transport depends on the molecular mean-free-path, ℓ , and the rms molecular speed, u . Simple dynamical arguments show that the molecular diffusivity, ν , for momentum is of order $u\ell$. The analogy pictures small scale eddies replacing molecules as the transporters of macroscale momentum. The correlation length, l_{cor} , of the turbulence replaces ℓ and the rms turbulent velocity, v , replaces u . The transport of macroscale momentum is then governed by a kinematic “turbulent viscosity” $\nu_T \approx vl_{\text{cor}}$, which greatly exceeds ν . We use $v = 10^{-4} \text{ m s}^{-1}$ for the velocity scale. The length scale is harder to estimate. Here we take $l_{\text{cor}} \sim E^{1/3} r_o \approx 100 \text{ m}$, which is based upon the characteristic length scale for finite amplitude rapidly rotating convection (e.g., Stellmach and Hansen 2004; Sprague et al. 2006). This gives $\nu_T \approx 10^{-2} \text{ m}^2 \text{ s}^{-1}$. Though it is strictly inconsistent to use v and l_{cor} when estimating the shear on the resolved scales, an upper bound on the turbulent traction follows from doing so: $\hat{T}^V \approx \rho\nu_T v / l_{\text{cor}} \approx \rho(vl_{\text{cor}})v / l_{\text{cor}} \approx 10^{-4} \text{ N m}^{-2}$, implying $\hat{\Gamma}_z^V \approx 5 \times 10^{15} \text{ Nm}$.

This may overestimate $\hat{\Gamma}_z^V$. A stable layer may exist at the top of the FOC; see, e.g., Loper (2007), Buffett (2010). Braginsky (1999) pointed out that the light material released during the freezing of the SIC may preferentially congregate near the ICB, and that this may answer unresolved questions about the geomagnetic secular variation (Braginsky 1984). Turbulent motions in a buoyantly stable layer tend to be damped preferentially in the direction of stratification, and this reduces the macroscale momentum transport across the layer (e.g., Gargett 1984; Davidson 2004). Waves in such a layer may increase what ν alone can do in transporting macroscale momentum (e.g., Rogers and Glatzmaier 2006). But it is doubtful if they can transport it as effectively as ν_T for fully convective turbulence far from boundaries.

Though these and similar arguments lack rigor, the above estimates of $\hat{\Gamma}^V$ are less than the target torque, but not vastly so (cf. Kuang and Bloxham 1997; Brito et al. 2004; Deleplace and Cardin 2006; Buffett and Christensen 2007).

The Topographic Torque

The likelihood that there are inverted mountains and valleys on the CMB, and that these might create topographic torques large enough to explain the observed changes in LOD, was first suggested by Hide (1969). These irregularities are often

collectively called “bumps,” and their study was jokingly christened “geophrenology” by the late Keith Runcorn. Hide’s idea generated much interest and literature, e.g., Anufriev and Braginsky (1975, 1977a, b), Jault and Le Mouél (1989), Kuang and Bloxham (1993, 1997), Buffett (1998) and Kuang and Chao (2001). Seismic studies infer the bump height, $\mathcal{H}$, is of order 1 km (e.g., Tanaka 2010). See *Earth’s Structure, Lower Mantle*.

Topographic coupling between the FOC and its boundaries depends on deviations from sphericity in the shapes of the CMB and ICB. The fluid pressure, p , creates surface fractions, $\hat{\mathbf{T}}^T = p\mathbf{n}$ and $\tilde{\mathbf{T}}^T = -p\mathbf{n}$, that are not purely radial. The resulting topographic torques on the CMB and ICB are

$$\begin{aligned}\hat{\mathbf{T}}^T &= \oint_{\hat{S}} p \mathbf{r} \times \mathbf{n} \, dA, \\ \tilde{\mathbf{T}}^T &= -\oint_{\tilde{S}} p \mathbf{r} \times \mathbf{n} \, dA.\end{aligned}\quad (3a, b)$$

Equal but opposite torques act on the FOC, so that (cf. Eq. 1g)

$$\mathbf{T}^T = -\hat{\mathbf{T}}^T - \tilde{\mathbf{T}}^T = -\int_V \mathbf{r} \times \nabla p \, dV. \quad (3c)$$

We define the CMB and ICB by

$$r = r_o + \hat{h}(\theta, \phi), \quad r = r_i + \tilde{h}(\theta, \phi), \quad (4a, b)$$

where (r, θ, ϕ) are spherical coordinates. Eq. 3a, b can be simplified if it is assumed that $|\hat{h}|/r_o \ll 1$, $|\tilde{h}|/r_i \ll 1$, $|\nabla \hat{h}| \ll 1$ and $|\nabla \tilde{h}| \ll 1$; see RA12. The projection of Eq. 3a, b onto the spheres $\hat{S}$ and $\tilde{S}$ are then

$$\begin{aligned}\hat{\mathbf{T}}^T &= \oint_{\hat{S}} \hat{h} \mathbf{r} \times \nabla p \, dA., \\ \tilde{\mathbf{T}}^T &= -\oint_{\tilde{S}} \tilde{h} \mathbf{r} \times \nabla p \, dA.. \end{aligned}\quad (4c, d)$$

The torques $\hat{\mathbf{T}}^T$ and $\tilde{\mathbf{T}}^T$ obviously depend on variations in p in the FOC, and we consider next the causes and magnitudes of these. This necessitates a considerable digression. The first step is to develop a reference state.

Convection mixes the FOC so well that, except in thin boundary layers at the CMB and ICB, it is chemically and thermodynamically homogeneous. It is therefore isentropic, i.e., its specific entropy, S , is uniform. The core is an unknown mixture of all elements, but the basics can be understood by assuming that it is a binary alloy of Fe and a lighter element, X , whose mass fraction is X . It is usually supposed

that X is mostly Si or S, but it is unnecessary to be specific here. Except in boundary layers,

$$\begin{aligned}S &= S_a = \text{Constant}, \\ X &= X_a = \text{Constant, in the FOC},\end{aligned}\quad (5a, b)$$

where the suffix a stands for “adiabatic”.

Although fast enough to mix the core thoroughly, core flows are slow compared with the speed of sound, u_s ($\approx 10^4 \text{ m s}^{-1}$). The primary dynamical balance is therefore hydrostatic and, allowing for centrifugal forces, it is governed by

$$\begin{aligned}\nabla p_a &= \rho_a (\mathbf{g}_a - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r})) \\ &= \rho_a \left(\mathbf{g}_a + \frac{1}{2} \nabla (\boldsymbol{\Omega} \times \mathbf{r})^2 \right),\end{aligned}\quad (5c)$$

where $\boldsymbol{\Omega} = \Omega \mathbf{1}_z$ is the angular velocity of the Earth and $\mathbf{g}$ is the gravitational acceleration. Newtonian gravitation theory requires that

$$\nabla \times \mathbf{g} = \mathbf{0}, \quad \nabla \cdot \mathbf{g} = -4\pi G \rho, \quad (5d, e)$$

where G is the constant of gravitation. It follows that

$$\nabla p_a = \rho_a \mathbf{g}_e = -\rho_a \nabla \Phi_e, \quad \text{where } \mathbf{g}_e = -\nabla \Phi_e \quad (5f, g)$$

is the effective gravitational field and $\Phi_e = \Phi_a - \frac{1}{2}(\boldsymbol{\Omega} \times \mathbf{r})^2$ is the “effective gravitational potential.” Eq. 5d shows that p_a , ρ_a , ... are constant on equipotential surfaces:

$$\begin{aligned}p_a &= p(\Phi_e, S_a, X_a), \\ \rho_a &= \rho(\Phi_e, S_a, X_a) = -\partial p(\Phi_e, S_a, X_a) / \partial \Phi_e, \dots\end{aligned}\quad (5h, i)$$

The effect of the centrifugal forces is quantified by $\varepsilon_\Omega = \Omega^2 r/g$. This is small, varying from 1.7×10^{-3} at the CMB to 1.5×10^{-3} at the ICB. The deviation of the equipotential surfaces from spheres is therefore small, as are deviations caused by the gravitational field produced by density anomalies in the mantle and SIC, quantified below. It is therefore helpful to refer to departures from a “reference state” that is spherically symmetric: $\Phi = \Phi^s(r)$, $p = p^s(r)$, $\rho = \rho^s(r)$, $T = T^s(r)$, etc., where T is temperature. There are large departures from adiabaticity in boundary layers and smaller departures throughout the FOC that are of two main types:

- (a) *Deviations from Φ^s due to rotation and density anomalies in the mantle and SIC.* These will be denoted by a prime, e.g., $p'_a = p_a - p^s$. Rotation produces the larger deviation but, as will be seen below, this is less relevant to LOD variations than the anomaly created by the mantle, which in the FOC is dominantly

$$\Phi'_a = A'(r/r_o)^2 \sin^2 \theta \cos 2\phi, \quad (A' > 0) \quad (6)$$

Wahr and de Vries (1989), Forte et al. (1994). The value of the constant A' is about $1,300 \text{ m}^2 \text{ s}^{-2}$, according to Defraigne et al. (1996) and about $2,300 \text{ m}^2 \text{ s}^{-2}$ according to Forte et al. (1994); our compromise will be $A' = 1,800 \text{ m}^2 \text{ s}^{-2}$. By Eq. 5i, $p'_a \approx -\rho_a^s \Phi'_a$. Therefore $\varepsilon_a = p'_a/p_a^s$ is 10^{-4} on the CMB equator and 10^{-5} at the ICB equator.

(b) *Deviations created by convection.* We use the notation

$$\rho = \rho_a + \rho_c, \quad (7a)$$

and similarly for other variables. The Earth radiates energy into space at a rate estimated to exceed 43 TW (e.g., §4.1.5 of Schubert et al. 2001). We take an extreme position by supposing the entire 43 TW emerges from the core. The outward heat flux in the FOC is the sum of the convective heat flux q_c and the adiabatic heat flux $q_a = -KdT^s/dr$, where K is thermal conductivity. Though the latter may be as much as 5 TW, we ignore it. Then $q_c = 0.28 \text{ W m}^{-2}$ at the CMB. We take

$$q_c = \rho c_p \overline{T_c V_r}, \quad (7b)$$

where c_p ($\approx 800 \text{ J kg}^{-1} \text{ K}^{-1}$) is the specific heat at constant pressure, p . The overline denotes a horizontal average over the flow, which is undoubtedly highly turbulent. For $V_r = 10^{-4} \text{ m s}^{-1}$, it follows that $T_c = 3.5 \times 10^{-4} \text{ K}$. As $\hat{T}_a \approx 4,000 \text{ K}$ (e.g., Kawai and Tsuchiya 2009), $\varepsilon_c = T_c/T_a$ is less than 10^{-7} , which is 3–4 orders of magnitude smaller than ε_a . Even this is an overestimate; it has never been suggested that Q_c is larger than 15 TW. The smallness of ε_Ω , ε_a , and ε_c means that all these effects can be treated as perturbations of the reference state.

As $p'_a(r, \theta, \phi) \gg p_c(r, \theta, \phi)$, it is sensible to consider first the effect of p'_a by introducing the “adiabatic topographic torques” on the mantle, ICB, and FOC:

$$\begin{aligned} \hat{\Gamma}_a^T &= \oint_S p_a \mathbf{r} \times \mathbf{n} \, dA, \\ \tilde{\Gamma}_a^T &= -\oint_S p_a \mathbf{r} \times \mathbf{n} \, dA, \\ \Gamma_a^T &= -\int_V \mathbf{r} \times \nabla p_a \, dV. \end{aligned} \quad (8a, b, c)$$

If the CMB (ICB) were an equipotential or if it were spherical, $\hat{\Gamma}_a^T$ ($\tilde{\Gamma}_a^T$) would vanish according to Eq. 8a, b, but generally these torques are nonzero. They can be evaluated only if $\hat{p}(\tilde{p})$ is known on the CMB (ICB). It will be

shown in the next section that $\hat{\Gamma}_a^T$ and $\tilde{\Gamma}_a^T$ are intimately related to what we shall call “adiabatic gravitational torques.” We therefore postpone further discussion and estimation of adiabatic topographic torques.

Consider the torque $\hat{\Gamma}_c^T$ associated with the convective motions in the FOC:

$$\begin{aligned} \hat{\Gamma}_c^T &= \oint_S \hat{h} \mathbf{r} \times \nabla p_c \, dA, \\ \hat{\Gamma}_{c,z}^T &= \oint_S \hat{h} \partial_\phi p_c \, dA. \end{aligned} \quad (9a, b)$$

Reasons will be given in the section on “The Magnetic Torque” why core flow may be well described by the magnetostrophic approximation Eq. 20a, and why, deep in the core, Coriolis and Lorentz forces are comparable, implying a magnetic field strength $\mathcal{B}$ there of about 2 mT, or about four times greater than the typical field strength $\mathcal{B}_o$ on the CMB. The Lorentz force is therefore 16 times less on the CMB than in the bulk of the core. Also, as g_ϕ is small, $\partial_\phi p_c \approx -2\Omega\rho_0 r_o V_\theta \cos \theta \sin \theta$ should be a good approximation to the ϕ -component of Eq. 20a on the CMB. Therefore (Hide et al. 1993)

$$\hat{\Gamma}_{c,z}^T = -2\Omega\rho_0 r_o \oint_S \hat{h}(\theta, \phi) V_\theta(r_o, \theta, \phi) \cos \theta \sin \theta \, dA. \quad (9c)$$

In principle, $\hat{\Gamma}_{c,z}^T$ can be estimated by extracting $\hat{h}$ from seismological analysis, and by using the $V_\theta(r_o, \theta, \phi)$ inferred from the core surface motion. In practice, this is difficult and has generated controversy. Eq. 9c suggests that

$$\hat{\Gamma}_{c,z}^T = O(2\Omega\rho\mathcal{V}\mathcal{H}r_o^3), \quad (9d)$$

which is 10^{18} Nm for a bump height of $\mathcal{H} = 100 \text{ m}$. Such a bump height is well within the bounds set by recent seismic investigations (e.g., Tanaka 2010). Equation 9b indicates, however, that Eq. 9d may be a serious overestimate because p_c is a single-valued function and $\partial_\phi p_c$ is as often positive as negative in the integrand of Eq. 9b. Though $\pm 2\Omega\rho\mathcal{V}r_o$ is a reasonable estimate of $\partial_\phi p_c$ at most points on the CMB, considerable cancellation is likely when evaluating the integral in Eq. 9b. There is even an outside chance that the cancellation might be complete; see Anufriev and Braginsky (1978b).

Reliable estimation of the convective topographic torque must probably await careful experiments and allied theory. No argument has so far convincingly demonstrated that topography can create torques of the target magnitude of 10^{18} Nm but, equally, none have shown that it cannot.

The Gravitational Torque

The gravitational torque on a body V of density $\rho(\mathbf{x})$ in a gravitational field $\mathbf{g}(\mathbf{x})$ is

$$\mathbf{\Gamma}^G = \int_V \rho \mathbf{r} \times \mathbf{g} dV. \quad (10a)$$

This volume integral can be usefully transformed into a surface integral by drawing on the analogy between the theories governing Newtonian gravitation and electrostatics, the only difference between these theories being one of sign: like charges repel but all bodies attract gravitationally. It can be shown that

$$\begin{aligned} \rho g_i &= \nabla_j S_{ij}^G, \quad \text{where} \quad S_{ij}^G \\ &= -\frac{1}{4\pi G} \left(g_i g_j - \frac{1}{2} g^2 \delta_{ij} \right) \end{aligned} \quad (10b, c)$$

is the gravitational stress tensor, the gravitational pressure $-g^2/8\pi G$ being also the gravitational energy density. Equations 10b, c enable 10a to be written as a surface integral:

$$\mathbf{\Gamma}^G = -\frac{1}{4\pi G} \oint_S \mathbf{r} \times \left[(\mathbf{n} \cdot \mathbf{g}) \mathbf{g} - \frac{1}{2} g^2 \mathbf{n} \right] dA, \quad (10d)$$

where $\mathbf{n}$ points out of V . See RA12 and Appendix B of Braginsky and Roberts (1995) for derivations of these results. When they are applied below, $\mathbf{n}$ on the CMB and SIC will, as previously defined, be oriented approximately parallel to $\mathbf{r}$.

By Eq. 10a, the gravitational torque on the FOC due to the mantle and SIC is

$$\mathbf{\Gamma}^G = \int_V (\rho_a + \rho_c) \mathbf{r} \times (\mathbf{g}_a + \mathbf{g}_c) dV = \mathbf{\Gamma}_a^G + \mathbf{\Gamma}_c^G, \quad (11a)$$

where $\mathbf{\Gamma}_a^G$ and $\mathbf{\Gamma}_c^G$ are the adiabatic and convective parts of $\mathbf{\Gamma}^G$:

$$\begin{aligned} \mathbf{\Gamma}_a^G &= \int_V \rho_a \mathbf{r} \times \mathbf{g}_a dV, \\ \mathbf{\Gamma}_c^G &= \int_V \mathbf{r} \times (\rho_c \mathbf{g}_a + \rho_a \mathbf{g}_c + \rho_c \mathbf{g}_c) dV. \end{aligned} \quad (11b, c)$$

It was pointed out earlier that $\rho_c = O(10^{-4} \rho'_a)$, $g_c = O(10^{-4} g'_a)$, etc. Apparently therefore $\mathbf{\Gamma}_c^G = O(10^{-4} \mathbf{\Gamma}_a^G)$, making it sensible to focus first on the adiabatic torque.

Consistent with hydrostatic balance in the FOC (see Eq. 5f), Eqs. 8c and 11b, give

$$\mathbf{\Gamma}_a^{G+T} = \mathbf{\Gamma}_a^G + \mathbf{\Gamma}_a^T = \int_V \mathbf{r} \times (\rho_a g_a - \nabla p_a) dV = \mathbf{0}. \quad (12a)$$

By Eqs. 8a, b and 10d and the continuity of $\mathbf{g}$ and p , we also have

$$\begin{aligned} \hat{\mathbf{\Gamma}}_a^{G+T} &= \oint_S \mathbf{r} \\ &\times \left\{ p_a \mathbf{n} + (4\pi G)^{-1} \left[(\mathbf{n} \cdot \mathbf{g}_a) g_a - \frac{1}{2} g_a^2 \mathbf{n} \right] \right\} dA, \end{aligned} \quad (12b)$$

$$\begin{aligned} \tilde{\mathbf{\Gamma}}_a^{G+T} &= -\oint_S \mathbf{r} \\ &\times \left\{ p_a \mathbf{n} + (4\pi G)^{-1} \left[(\mathbf{n} \cdot \mathbf{g}_a) g_a - \frac{1}{2} g_a^2 \mathbf{n} \right] \right\} dA, \end{aligned} \quad (12c)$$

from which, in agreement with Eq. 12a,

$$\hat{\mathbf{\Gamma}}_a^{G+T} + \tilde{\mathbf{\Gamma}}_a^{G+T} = \mathbf{0}. \quad (12d)$$

If the torques $\hat{\mathbf{\Gamma}}_a^{G+T}$ and $\tilde{\mathbf{\Gamma}}_a^{G+T}$ are nonzero, $\hat{\mathbf{\Omega}}$ and $\tilde{\mathbf{\Omega}}$ evolve. The consequent change in the relative orientation of the mantle and SIC modifies $\mathbf{g}_a$ and p_a in Eq. 12b, c so that the system evolves toward a configuration of minimum energy E . In this configuration

$$\hat{\mathbf{\Gamma}}_a^{G+T} = \mathbf{0}, \quad \tilde{\mathbf{\Gamma}}_a^{G+T} = \mathbf{0}. \quad (13a, b)$$

If this minimum energy, torque-free state is perturbed, the restoring GT-torques (as we shall call them) set up a “gravitational oscillation” of the SIC relative to the mantle.

To give a simple example, suppose that the gravitational anomaly defined by Eq. 6 is created by sources entirely within the mantle, the CMB having no bumps, so that $\hat{\mathbf{\Gamma}}_{a,z}^{G+T} = \hat{\mathbf{\Gamma}}_{a,z}^G$. Similarly, suppose the SIC is spherical but has internal sources that above the ICB produce the gravitational anomaly

$$\Phi_a'' = A''(r_i/r)^3 \sin^2 \theta \cos 2(\phi - \varphi), \quad (A'' > 0), \quad (14a)$$

where φ is the angular displacement of the system from the stable state, $\varphi = 0$, in which Eq. 13a, b hold. Eq. 12b, c show that

$$\begin{aligned} \hat{\mathbf{\Gamma}}_{a,z}^G &= -\tilde{\mathbf{\Gamma}}_{a,z}^G = \Gamma_0^G \sin 2\varphi, \quad \text{where} \\ \Gamma_0^G &= 8A'A''r_i^3/3Gr_o^2 (> 0). \end{aligned} \quad (14b, c)$$

These torques vanish for the stable minimum energy states $\varphi = 0, \pi$ (and also for the unstable $\varphi = \pm \frac{1}{2}\pi$). Small departures from a stable state satisfy

$$\hat{C}d_t^2\tilde{\varphi} = 2\Gamma_0^G(\tilde{\varphi} - \hat{\varphi}), \quad \tilde{C}d_t^2\tilde{\varphi} = 2\Gamma_0^G(\tilde{\varphi} - \hat{\varphi}), \quad (14d, e)$$

where $\hat{C}$ ($= 7.12 \times 10^{37} \text{ kg m}^2$) and $\tilde{C}$ ($= 5.86 \times 10^{34} \text{ kg m}^2$) are the polar moments of inertia of mantle and SIC, respectively. The frequency, ω^G , of the oscillation is

$$\omega^G = \left[2(\hat{C} + \tilde{C})\Gamma_0^G/\tilde{C}\tilde{C} \right]^{1/2} \approx \left[2\Gamma_0^G/\tilde{C} \right]^{1/2}. \quad (14f)$$

Although an anomaly of the form of Eq. 14a could be produced by density variations within the SIC, it is more plausibly created by SIC topography for the following reasons. It is generally believed that the SIC is the result of freezing of the FOC, an ongoing process even today (Jacobs 1953). An alloy generally changes its composition when it changes phase. The rather large density jump at the ICB, $\tilde{\Delta} = \tilde{\rho}(r_i) - \rho(r_i) \approx 600 \text{ kg m}^{-3}$, is hard to attribute to contraction on freezing but can be readily explained as a small reduction in X on freezing. Phase equilibrium at the ICB implies $T_a^s(\Phi_a, S_a, X_a) = T_m(p_a, X_a)$, where T_m is the melting temperature. This implies that the ICB is an equipotential surface. Since $\Phi = \Phi_a^s(r) + \Phi_a'(r, \theta, \phi)$ where $|\Phi_a'/\Phi_a^s| \ll 1$, Taylor expansion, using Eq. 4b, shows that $\Phi_a^s(r_i) + \tilde{g}\tilde{h}(\theta, \phi) + \Phi_a'(r_i, \theta, \phi)$ is approximately constant, where $\tilde{g} = -g_{a,r}^s(r_i) = \partial_r \Phi_a^s(r_i) (> 0)$ is gravity at the ICB. It follows that

$$\begin{aligned} \tilde{h} &= -\Phi_a'(r_i, \theta, \phi)/\tilde{g} = \tilde{\varepsilon} \sin^2 \theta \cos 2\phi, \\ \tilde{\varepsilon} &= -A'(r_i/r_0)^2/\tilde{g}. \end{aligned} \quad (14g, h)$$

This shows how the gravitational anomaly in the mantle imposes its $n = m = 2$ preference on the SIC. It makes the otherwise ad hoc assumption of Eq. 14a seem perfectly reasonable. The condition that $\tilde{h}$ creates Eq. 14a, for $\varphi = 0$ and $r - r_i \gg \tilde{h}$ is

$$A''/A' = 4\pi G r_i^3 \tilde{\Delta}/5r_0^2 \tilde{g} \approx 0.0034. \quad (14i)$$

The maximum bump height on the ICB is $|\tilde{\varepsilon}| \approx 50 \text{ m}$. Nonzero φ corresponds to a rotated SIC. Such a rotation is to be expected; core turbulence continually subjects the SIC to (topographic) torques that continuously change its orientation. Two relaxation processes act to restore the ICB to its equipotential: (i) flow within the SIC, (ii) new freezing/melting on the ICB. If either were instantaneous, there would be no torque between the mantle and SIC, but both appear to act on much longer time scales than core turbulence, so that SIC topography is almost “frozen” to the SIC as it turns.

Concerning (i), the viscosity of the SIC is plausibly much less than the viscosity of the mantle. According to Schubert et al. (2001) $\hat{\nu} \approx 5 \times 10^{19} \text{ m}^2 \text{ s}^{-1}$ in the deep mantle but, according to Mound and Buffett (2006), $\hat{\nu} \gtrsim 10^{13} \text{ m}^2 \text{ s}^{-1}$. Whereas mantle anomalies are essentially “frozen in,” slow motions within the SIC created by stresses exerted by the

FOC on the ICB can gradually restore equilibrium (Yoshida et al. 1996; Buffett 1997). Concerning (ii), the thermodynamic disequilibrium created by the misalignment of the ICB from its equipotential surface is slowly removed by new freezing of the FOC or new melting of the SIC. This processes has not been fully explored (but see Fearn et al. 1981; Morse 1986). The possible significance of melting/freezing processes on SIC structure has been recently investigated by Alboussière et al. (2010) and Monnereau et al. (2010).

Interest in gravitational torques and oscillations was sparked by Buffett (1996). We follow him but by a different method, making use of Eq. 14c, i:

$$\Gamma_0^G = 32\pi r_i^6 \tilde{\Delta} A'^2 / 15r_0^4 \tilde{g} \approx 6.7 \times 10^{19} \text{ Nm} \quad (15a)$$

By Eq. 14f, the frequency of the oscillation is $\omega^G \approx 4.8 \times 10^{-8} \text{ s}^{-1}$, with a period of

$$\tau^G = 2\pi/\omega^G \approx 4.1 \text{ years}. \quad (15b)$$

According to Buffett et al. (2009), gravitational oscillations are mainly responsible for the LOD variations shown in Fig. 1. Mound and Buffett (2006) obtain $\Gamma_0^G \approx 1.5 \times 10^{20} \text{ Nm}$.

Equations 14d, e imply, for some t_0 ,

$$\begin{aligned} \hat{\Omega} &= \hat{\Omega}_0 + d_t \hat{\phi} \\ &= \hat{\Omega}_0 + \left(2\Gamma_0^G/\tilde{C}\omega^G \right) \left(\tilde{\phi} - \hat{\phi} \right)_{\max} \sin \omega^G(t - t_0). \end{aligned} \quad (15c)$$

The amplitude of the gravitational oscillation is therefore related to that of the variation, ΔP , in LOD by

$$\left(\tilde{\phi} - \hat{\phi} \right)_{\max} = \pi \hat{C} \omega^G \Delta P / \Gamma_0^G P_0^2 \approx 1.2, \quad (15d)$$

for $\Delta P = 1 \text{ ms}$. This gives a maximum angular velocity difference of $\omega^G \left(\tilde{\phi} - \hat{\phi} \right)_{\max} \approx 2 \text{ year}^{-1}$. Furthermore the peak-to-peak variation in the radial gravitational acceleration at the Earth's surface, $r = r_E$, is

$$\Delta g_r''(r_E) = \frac{12A''}{r_i} \left(\frac{r_i}{r_E} \right)^4 \left(\hat{\phi} - \tilde{\phi} \right)_{\max}^2 \simeq 4 \text{ nGal}. \quad (15e)$$

This value is too small by roughly a factor of 5 to be detectable by the GRACE satellite system (e.g., Wahr et al. 2006).

One limitation of this analysis is the neglect of electromagnetic stresses at the boundaries when they are in motion relative to the core fluid. It has been implicitly assumed that the SIC is completely decoupled from the fluid within the tangent cylinder (TC), the imaginary cylinder that touches the

ICB at its equator. This is particularly significant because the fluid dynamics inside and outside the TC are quite dissimilar. See Hide and James (1983), Heimpel and Aurnou (2007), and the next section. Because the SIC is as good an electrical conductor as the FOC (or better), it may be tightly coupled magnetically to C^N and C^S , as suggested by Braginsky (1970); see the next section. To examine the effect of this coupling, we make the extreme assumption that all the fluid in the TC is completely locked to the SIC. Because the mantle is a poor electrical conductor, this fluid is not well coupled to the mantle, so that the entire fluid column within the TC can co-rotate about O_z almost freely with the SIC. This suggests that, instead of Eq. 1g, h, a more useful division of the total angular momentum of the Earth might be based on

$$\begin{aligned}\hat{\Gamma} + \Gamma^{\text{TC}} + \Gamma^{\text{XTC}} &= 0, \\ \hat{\mathbf{M}} + \mathbf{M}^{\text{TC}} + \mathbf{M}^{\text{XTC}} &= \text{constant},\end{aligned}\quad (16a, b)$$

where $^{\text{TC}}$ refers to the TC and SIC locked together, and $^{\text{XTC}}$ refers to the part of the FOC exterior to the TC. The moment of inertia of the fluid within the TC is $2.12 \times 10^{35} \text{ kg m}^2$ which, when added to $\tilde{C}$, gives $C^{\text{TC}} = 2.71 \times 10^{35} \text{ kg m}^2$. Using this instead of $\tilde{C}$ in Eq. 14f, τ^G is lengthened from 4.1 years to

$$\begin{aligned}\tau_{\text{TC}}^G &= 2\pi/\omega_{\text{TC}}^G \approx 8.9 \text{ years}, \\ \text{where } \omega_{\text{TC}}^G &= \left[2(\hat{C} + C^{\text{TC}})\Gamma_0^G/\hat{C}C^{\text{TC}}\right]^{1/2}.\end{aligned}\quad (16c, d)$$

See also Mound and Buffett (2006).

Even though ρ_c/ρ_a , p_c/p_a , etc., are of order 10^{-4} , this does not mean that $\Gamma_a^{G+T}/\Gamma_c^{G+T}$ is as small as that. In fact, Eq. 13a shows that $\Gamma_a^{G+T} = 0$ in the minimum energy state. The adiabatic GT-torques dominate the convective torques only if φ is sufficiently large. Stated another way, a convective torque can be nullified by a small departure from the minimum energy state. Earlier, the torque on the SIC created by core turbulence was held responsible for causing φ to deviate from zero. This torque is essentially (magneto-)convective, and is nullified by the GT-torque for a tiny change in φ . Another way of estimating how tiny this φ is equates the magnitudes of the GT-torque, taken as $1.3 \times 10^{20} \varphi \text{ Nm}$ (see Eqs. 14b, c and 15d), and the convective torque, taken to have its target value of 10^{18} Nm . This gives $\varphi \approx 0.5^\circ$. Within this angle, the mantle and SIC are gravitationally locked together, over short time scales compared with those of the relaxation processes in the SIC described above. See Buffett and Glatzmaier (2000).

The Magnetic Torque

This section assumes that readers are familiar with pre-Maxwell EM theory and the fundamentals of MHD, including

the frozen flux theorem and Alfvén waves. Davidson (2001) contains the necessary background.

It may be useful to remind readers that the magnetic torque about O on a body V carrying a current of density $\mathbf{J}$ is the integrated moment of the Lorentz force, $\mathbf{J} \times \mathbf{B}$:

$$\Gamma^M = \int_V \mathbf{r} \times (\mathbf{J} \times \mathbf{B}) dV = \int_V r[B_r \mathbf{J} - J_r \mathbf{B}] dV. \quad (17a)$$

The Lorentz force can be expressed as a divergence:

$$\begin{aligned}(\mathbf{J} \times \mathbf{B})_i &= \nabla_j S_{ij}^M, \\ \text{where } S_{ij}^M &= \mu_0^{-1} \left(B_i B_j - \frac{1}{2} B^2 \delta_{ij} \right)\end{aligned}\quad (17b, c)$$

is the magnetic stress tensor. An alternative form of Eq. 17a is therefore

$$\Gamma^M = \mu_0^{-1} \oint_S \mathbf{r} \times \left[(\mathbf{n} \cdot \mathbf{B}) \mathbf{B} - \frac{1}{2} B^2 \mathbf{n} \right] dA, \quad (17d)$$

where the unit normal $\mathbf{n}$ to S points out of V . Therefore

$$\begin{aligned}\hat{\Gamma}_z^M &= -\mu_0^{-1} \oint_S \hat{\phi}_r \hat{B}_r \hat{B}_\phi dA, \\ \tilde{\Gamma}_z^M &= \mu_0^{-1} \oint_S \tilde{s} \tilde{B}_r \tilde{B}_\phi dA.\end{aligned}\quad (17e, f)$$

These results can be used as they stand to assess the magnetic coupling between the inferred core surface flow and the mantle. See Stix and Roberts (1984), Love and Bloxham (1994), Holme (1998). Here, however, we are more interested in forging a link between the observed changes in LOD and torsional waves. To explain what the latter are, it is necessary to consider some dynamical issues.

Most studies of core MHD are based on the Boussinesq approximation; see, e.g., Braginsky and Roberts (2007). This assumes constant density, $\rho_0 (\approx 10^4 \text{ kg m}^{-3})$, and expresses conservation of mass and momentum as

$$\begin{aligned}\nabla \cdot \mathbf{V} &= 0, & \partial_t \mathbf{V} + \mathbf{V} \cdot \nabla \mathbf{V} + 2\boldsymbol{\Omega} \times \mathbf{V} \\ &= -\nabla(p_c/\rho_0) + C\mathbf{g}_e + \mathbf{J} \times \mathbf{B}/\rho_0 + \nu_T \nabla^2 \mathbf{V}.\end{aligned}\quad (18a, b)$$

The accelerations in Eq. 18b are from inertia ($\partial_t \mathbf{V}$ and $\mathbf{V} \cdot \nabla \mathbf{V}$), rotation ($2\boldsymbol{\Omega} \times \mathbf{V}$), pressure (p_c), buoyancy ($C\mathbf{g}_e$), magnetic field ($\mathbf{J} \times \mathbf{B}/\rho_0$), and viscosity ($\nu_T \nabla^2 \mathbf{V}$). Thermal and compositional buoyancy, combined in the codensity C (Braginsky and Roberts 1995), maintains the flow and the magnetic field; see *Core Dynamo*.

The Coriolis force is generally more significant than the inertial and viscous forces. This is indicated by the smallness of the Ekman and Rossby numbers:

$$Ro = \mathcal{V}/\Omega\mathcal{L}. \quad (19)$$

See Eq. 2c for the definition of E . From $\mathcal{V} = 10^{-4} \text{ m s}^{-1}$, $\mathcal{L} = r_o$ and $v_T = 10^{-2} \text{ m}^2 \text{ s}^{-1}$ follow $Ro \approx 10^{-6}$ and $E \approx 10^{-11}$. This suggests that the inertial and viscous terms can be safely omitted from Eq. 18b, except on small length scales.

If the inertial and viscous forces are ejected from Eq. 18b, it becomes

$$2\rho_0\mathbf{\Omega} \times \mathbf{V}^N = -\nabla p_c + \rho_0 C\mathbf{g}_e + \mathbf{J} \times \mathbf{B}, \quad (20a)$$

where N stands for “non-geostrophic,” and “geostrophic” is defined below. Equations 18a and 20a define the “magnetostrophic approximation,” often used to describe the quasi-steady macroscales of core MHD. As the viscous term has been ejected, the only boundary condition that $\mathbf{V}^N$ must obey is

$$\mathbf{n} \cdot \mathbf{V}^N = 0, \quad \text{on the CMB and ICB.} \quad (20b)$$

The full boundary conditions of continuous $\mathbf{B}$ and $\mathbf{n} \times \mathbf{E}$ still apply. Because $\partial_t \mathbf{V}$ has been ejected from Eq. 20a, there are no Alfvén waves. Instead, the system evolves on the much longer ageostrophic time scale,

$$\begin{aligned} \tau^N &= 2\Omega\mathcal{L}^2/V_A^2 = \tau_\eta/\Lambda, \\ \text{where } \tau_\eta &= \mathcal{L}^2/\eta \approx 10^5 \text{ years} \end{aligned} \quad (20c, d)$$

is the free decay time for magnetic fields of scale $\mathcal{L}$, and Λ is the “Elsasser number”:

$$\begin{aligned} \Lambda &= V_A^2/2\Omega\eta, \quad \text{where} \quad V_A \\ &= \mathcal{B}/\sqrt{(\mu_0\rho_0)} \end{aligned} \quad (20e, f)$$

is the Alfvén velocity. Elsasser (1946) suggested that $\mathcal{B}$ is determined by a balance of Lorentz and Coriolis forces. Taking $j \approx \sigma\mathcal{V}\mathcal{B}$ this implies $\Lambda = 1$, $\mathcal{B} \approx 2 \text{ mT}$, and $V_A \approx 2 \text{ cm s}^{-1}$. It also gives $\tau^N \approx \tau_\eta \approx 2 \times 10^5 \text{ years}$ for $\mathcal{L} = r_o$.

In cylindrical coordinates (s, ϕ, z) , the ϕ -component of Eq. 20a is

$$2\Omega\rho_0 V_s^N = -\partial_\phi p_c + (\mathbf{J} \times \mathbf{B})_\phi. \quad (21a)$$

Integrate this over the surface, $\mathcal{C}(s)$, of the circular cylinder of radius s ($> r_i$) coaxial with Oz. The left-hand-side vanishes by mass conservation, as can be verified by integrating Eq. 18a over the interior, $\mathcal{L}(s)$, of $\mathcal{C}(s)$ and applying Eq. 20b to $\hat{\mathcal{N}}(s)$ and $\hat{\mathcal{S}}(s)$, the spherical caps of $\mathcal{C}(s)$ on the CMB that complete the boundary of $\mathcal{I}(s)$. It follows that

$$\int_{\mathcal{C}(s)} (\mathbf{J} \times \mathbf{B})_\phi \, dA = 0 \quad (21b)$$

If $s < r_i$, there are two cylinders, $\mathcal{C}^N(s)$ and $\mathcal{C}^S(s)$, of radius s to the north and south of the SIC for which

$$\begin{aligned} \int_{\mathcal{C}^N(s)} (\mathbf{J} \times \mathbf{B})_\phi \, dA &= 0, & \int_{\mathcal{C}^S(s)} (\mathbf{J} \times \mathbf{B})_\phi \, dA \\ &= 0. \end{aligned} \quad (21c, d)$$

Equations 20b–d are examples of “Taylor’s constraint” (Taylor 1963). The cylinders $\mathcal{C}(s)$ are termed “Taylor cylinders.” Of these, $\mathcal{C}(r_i)$ is the tangent cylinder (TC).

It is obviously possible to assign a $\mathbf{J}$ which creates a $\mathbf{B}$ that contradicts Eqs. 21b–d, at least initially. This shows that Eq. 20a is an oversimplification. That approximation rested on discarding the inertial force in comparison with the Coriolis force. Consider however the class of “geostrophic flows”:

$$\mathbf{v} = v(s, t) \mathbf{1}_\phi. \quad (22a)$$

The corresponding Coriolis acceleration is

$$\begin{aligned} 2\rho_0\mathbf{\Omega} \times \mathbf{v} &= -\nabla\chi, \\ \text{where } \chi &= 2\Omega\rho_0 \int v(s, t) \, ds \end{aligned} \quad (22b, c)$$

can be absorbed into p_c . Coriolis forces are therefore totally ineffective when they act on geostrophic flows. Other forces previously abandoned in comparison with Coriolis forces become influential, especially the inertial forces, which must be restored when analyzing the geostrophic part of core flow. This recovers the Alfvén wave, or something very like it, called the “torsional wave.” They share a common time scale:

$$\tau_A = r_o/V_A \approx 6 \text{ years.} \quad (22d)$$

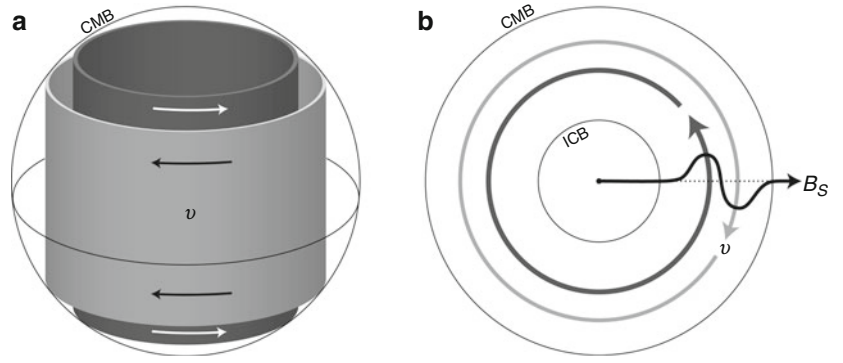
That this is also the time scale t_{LOD} of the semi-decadal variations of in Fig. 1a, b may not be a coincidence, as argued by Gillet et al. (2010).

The geostrophic part $\mathbf{v}$ of $\mathbf{V}$ can be extracted from $\mathbf{V}$ by taking the “geostrophic average,” $\langle V_\phi \rangle$, of V_ϕ : for $s > r_i$, this average is defined by

$$\begin{aligned} v &= \langle V_\phi \rangle \equiv \frac{1}{\hat{A}}(s) \int_{\mathcal{C}(s)} V_\phi \, dA, & \text{so that} \\ \mathbf{V}^N &= \mathbf{V} - v\mathbf{1}_\phi, \end{aligned} \quad (22e)$$

Core-Mantle Coupling,

Fig. 2 Schematics showing (a) geostrophic flows in the core, V^g , and (b) plan view of an initially cylindrical magnetic field (dashed line) distorted by v . The restoring Lorentz torques on the distorted magnetic field, B_s (solid line), lead to the cylindrical propagation of torsional waves. (Adapted from Dumberry 2007)



where $\hat{A}(s) = 4\pi s z_1$ is the area of $C(s)$, and $z_1(s) = \sqrt{(r_o^2 - s^2)}$ is the semi-length of its sides. The axial angular momentum of the FOC is carried by v . Therefore, insofar as the rotation of the SIC is locked to that of the fluid in the TC, the angular momentum, $M_z + \tilde{M}_z$, of the entire core can be derived from the zonal part of the inferred core surface flow. The LOD record provides $\tilde{M}_z$. Therefore Eq. 1h can be tested: $\hat{M}_z = -(M_z + \tilde{M}_z)$. Results have been gratifying; see Jault et al. (1988), Jackson (1997) and Fig. 1c above. The previous section indicates however that the mantle and SIC are not locked together but take part in a gravitational oscillation having a period (4–9 years) similar to the torsional wave period $\tau_A \approx 6$ years. The implied convolvment of gravitational and magnetic coupling complicates the task of extracting information about either; see Buffett et al. (2009).

In a torsional wave, the geostrophic cylinders are in relative angular motion about their common (polar) axis; see Fig. 2a. The response **b** of **B** to the motion **v** can, as for an Alfvén wave, be visualized using the frozen flux theorem, the field lines behaving like elastic strings threading the cylinders together and opposing their relative motion; see Fig. 2b. The resulting torque on a cylinder supplies the restoring force for a wave, the mass of the cylinders providing the inertia. Whenever **J** and **B** contradict Eqs. 21b–d, a torsional wave is launched that travels in the $\pm s$ -directions.

The canonical torsional wave equation is

$$\frac{\partial^2 \zeta}{\partial t^2} = \frac{1}{s^2 \hat{A}(s)} \frac{\partial}{\partial s} \left[s^2 \hat{A}(s) V_A^2(s) \frac{\partial \zeta}{\partial s} \right], \quad (23)$$

where $\zeta(s, t) = v/s$ is the angular velocity of $C(s)$ and $V_A(s) = \mathcal{B}_s(s)/(\mu_0 \rho_0)^{1/2}$ is the Alfvén velocity based on the mean $(B_s^N)^2$ over $C(s)$: $\mathcal{B}_s^2(s) = \langle (B_s^N)^2 \rangle$. Equation 23 is called canonical because it displays the essential nature of torsional waves clearly. It is however not only incomplete but also ignores magnetic coupling to the mantle and SIC. Equation 23 presupposes that the field, $\mathbf{B}^N$, on which the waves ride is axisymmetric. In this case, Eq. 23 has a severe

singularity at $s = 0$ which excludes physically acceptable solutions. This difficulty can be evaded by supposing that the TC rotates as a solid body, as suggested by Braginsky (1970), and by applying Eq. 23 only in XTC. For general, non-axisymmetric $\mathbf{B}^N$, the regular singularity at $s = 0$ is harmless, but unfortunately Eq. 23 is then incomplete. The terms missing from Eq. 23 represent the transmission of torque from one Taylor cylinder to another by the potential field outside the core. As the terms are troublesome if retained, they are usually abandoned, with an unsubstantiated claim that they are too small to be worth keeping. Whether they are retained or abandoned, the wave equation admits normal mode solutions, i.e., solutions in which ζ is proportional to $\exp(-i\omega t)$, where every ω is real, as shown in RA12. When magnetic coupling to the mantle is included, ω acquires a negative imaginary part, representing the ohmic losses in the mantle. The inclusion of magnetic coupling is highly relevant to our focus here, but it clearly adds another layer of complexity.

In view of these theoretical obstacles, the reader may wonder whether computer simulation would not provide a simpler approach. It is however difficult to extract evidence of torsional waves from geodynamo simulations. This is because it is not yet possible to attain geophysically realistic magnetic Prandtl numbers, Pm , in numerical integrations. The importance of viscous damping of torsional waves can be assessed by the ratio of the torsional wave time scale, $\tau_A = r_o/V_A$, to the spin-up time scale, $\tau_{SU} = \sqrt{(r_o^2/2\Omega\nu)}$, which is their viscous decay time; see Roberts and Soward (1972). This ratio,

$$\tau_A/\tau_{SU} = 2(\Omega\nu)^{1/2}/V_A = \sqrt{(Pm/\Lambda)}, \quad (24)$$

is small for the Earth ($\sim 10^{-3}$) but inconveniently large in simulations (≥ 0.1). See Wicht and Christensen (2010) for recent simulations of torsional waves.

Space limitations do not permit the mathematical theory of torsional waves to be presented here; see Braginsky (1970), Roberts and Soward (1972), Jault (2003), Dumberry (2007) and RA12. The principal aim in what follows is to outline the

underlying physics, that has been employed in theoretical studies, and that will have to be incorporated in numerical models in the future, when computer technology has advanced far enough to permit core MHD to be modeled more faithfully. Our discussion here sidesteps interesting and as yet incompletely answered questions, the first of which is whether torsional waves are involved in any essential way with the semi-decadal LOD periodicity; might not the gravitational oscillation described in the last section be mainly responsible? The answer is unclear but core field strengths of 2 mT suggest that torsional waves very effectively link together geostrophic motions throughout the core, and that torsional waves therefore necessarily accompany gravitational oscillations. This obviously does not imply that torsional waves couple well to the mantle; if the mantle were an electrical insulator, there would be no magnetic coupling of core and mantle at all. The waves would then be detectable only through the inferred core surface motion. Strong magnetic coupling of the waves to the mantle raises difficult questions about both the excitation and dissipation of the waves. As shown below, the waves are mainly and heavily damped by ohmic dissipation in the mantle. What maintains the waves against this energy loss? No definitive answer has been given to this very significant question. Core turbulence may be directly responsible, or indirectly through the Reynolds magneto-stresses it exerts on the SIC or on the TC. Buffett et al. (2009) find evidence for wave generation at the TC. Core turbulence would presumably excite many damped normal modes. This might help to explain some of the recently discovered short time scales of core surface motions (Olsen and Manda 2008; Finlay et al. 2010). It has not been established unequivocally that the 6 year period is the fundamental mode. Though often questioned, evidence of a 60 periodicity exists. This longer period signal might be representative of the fundamental mode but would require current estimates of $\mathcal{B}_s(s)$ to be revised downward.

These matters are beyond the scope of this review. We shall merely sketch, mostly in a qualitative way, the approximate theory that is currently the most complete. We also aim to expose its strengths and weaknesses.

Underlying the entire theory is the idea that core flow can be neatly separated into geostrophic motions of short time scale, τ_A , and non-geostrophic motions of long time scale τ^N obeying Eq. 20a. In other words, the theory focusses on the large length scales of core flow, in the belief that this includes the fundamental torsional wave mode of greatest interest. The torsional wave therefore rides on a flow satisfying Taylor's constraints Eqs. 21a–c. Because the Lehnert number, $\lambda = \tau_A / \tau^N = V_A / 2\Omega r_o = \omega_A / 2\Omega \approx 3 \times 10^{-5}$, is small, the time dependence of $\mathbf{B}^N$ can be ignored in torsional wave theory.

It was pointed out earlier that torsional waves are geostrophic motions in which the inertial force is crucial. The first

step in deriving the torsional wave equation is therefore to restore the time-dependent inertial force to Eq. 21a, obtaining

$$\rho_0 s \partial_t V_\phi + 2\Omega \rho_0 s V_s = -\partial_\phi p_c + s(\mathbf{J} \times \mathbf{B})_\phi, \quad (25a)$$

from which $\partial_t \zeta$ is extracted, by taking the geostrophic average; see Eq. 22e.

The evaluation of $(\mathbf{J} \times \mathbf{B})_\phi$ is simplified because E is small and the Lundquist number, $Lu = \tau_\eta / \tau_A = V_A r_o / \eta \approx 3 \times 10^4$, of the waves is large. In a first approximation, $E = Lu^{-1} = 0$. Then viscous and ohmic diffusion and the associated boundary conditions are discarded. In particular, the electric field created in the FOC by the waves simplifies to

$$\mathbf{e} = -\mathbf{v} \times \mathbf{B}^N - \mathbf{V}^N \times \mathbf{b} - \mathbf{v} \times \mathbf{b}, \quad (25b)$$

where $\mathbf{b}$ is the magnetic field of the waves. The term $-\mathbf{V}^N \times \mathbf{B}^N$ does not appear because it already acts on the Taylor state. The last term in Eq. 25b is also ignored because the wave amplitude is assumed to be small. The ratio of $-\mathbf{V}^N \times \mathbf{b}$ to $-\mathbf{v} \times \mathbf{B}^N$ is of order $A^N [b/v \sqrt{(\mu_0 \rho_0)}]^{-1}$, where $A = V/V_A$ is the Alfvén number, which for the non-geostrophic flow is about 0.1 (see above). As in an Alfvén wave, $v = O[b/\sqrt{(\mu_0 \rho_0)}]$, so that $|\mathbf{V}^N \times \mathbf{b}| \approx 0.1 |\mathbf{v} \times \mathbf{B}^N|$. This, combined with the fact that the inclusion of $-\mathbf{V}^N \times \mathbf{b}$ in $\mathbf{e}$ adds severe complications, encourages the neglect of $-\mathbf{V}^N \times \mathbf{b}$ in Eq. 25b, leaving

$$\mathbf{e} = -\mathbf{v} \times \mathbf{B}^N. \quad (25c)$$

The mantle and core are linked across a boundary layer of Ekman-Hartmann type; e.g., see Dormy et al. (2007). Because $Pm \ll 1$, this has a double structure. Viscosity acts only in an Ekman layer of thickness $d_v = \sqrt{(\nu/2\Omega)} \approx 0.1$ m; magnetic diffusion acts in a layer whose thickness is comparable with the EM skin depth, $d_\eta = (\frac{1}{2}|\omega|\mu_0\sigma)^{-1/2} \approx 10$ km. We therefore call this a “magnetic diffusion layer” (MDL), even though Coriolis and Lorentz forces affect its structure too. In the MDL, Eq. 25c is replaced by

$$\mathbf{e} = -\mathbf{v} \times \mathbf{B}^N - \eta \nabla \times \mathbf{b} \approx -\mathbf{v} \times \mathbf{B}^N - \eta \mathbf{1}_r \times \partial_r \mathbf{b}. \quad (26)$$

As $d_\eta/d_v \gg 1$, the Ekman layer is only a tiny part of the MDL. Ekman-Hartmann theory simplifies by discarding the Ekman part, setting $\nu = 0$ and abandoning the no-slip boundary conditions. The structure of the MDL still depends on rotation, and on the Elsasser number, Λ , defined in Eq. 20e.

The boundary layers play a vital role in linking ζ in the main bulk of the FOC to $\hat{\mathbf{b}}$ on the CMB and $\tilde{\mathbf{b}}$ on the ICB. They are therefore essential in determining $\hat{\mathbf{I}}^M$ and $\tilde{\mathbf{I}}^M$. At the CMB, $\mathcal{B} \approx 0.5$ mT and $\Lambda \approx 0.07$. The prevailing magnetic field, $\mathbf{B}^N$, therefore has very little effect on the boundary layer, which is controlled almost entirely by Coriolis forces and

magnetic diffusion. At the ICB, where $\mathcal{B}$ may be even an order of magnitude greater than at the CMB, $\Lambda > 1$, and Lorentz forces are too significant to ignore in the boundary layer. Further details are given in RA12.

The torque exerted by the waves on the mantle is proportional to the electrical conductivity, $\hat{\sigma}(v, \theta, \phi)$, of the mantle, which we assume is nonzero only in the layer $r_o < r < r_1 = r_o + d$ at the base of the mantle. The conductance of this layer is

$$\hat{\Sigma}(\theta, \phi) = \int_{r_o}^{r_1} \hat{\sigma}(r, \theta, \phi) dr. \quad (27a)$$

It is commonly assumed that $10^7 \text{ S} < \hat{\Sigma} < 10^9 \text{ S}$; we take $\hat{\Sigma} = 10^8 \text{ S}$ below.

Electric currents, $\hat{\mathbf{j}}$, flow in the mantle, $\hat{\mathbf{V}}$, either by leaking from the core or by electromagnetic induction, through the time dependence of the EM field on the CMB. We shall be interested in the penetration of the fields $\mathbf{b}$ and $\mathbf{e}$ of the waves, at frequencies ω of order $5 \times 10^{-8} \text{ s}^{-1}$. The resulting magnetic and electric fields, $\hat{\mathbf{b}}$ and $\hat{\mathbf{e}}$, in the mantle depend on ω and on L , the horizontal length scale imposed by $\mathbf{b}$ and $\mathbf{e}$ on the CMB. Associated with ω is the skin depth of the mantle:

$$\hat{d}_\eta = \left(\frac{1}{2} |\omega| \mu_0 \hat{\sigma}_o \right)^{-1/2}. \quad (27b)$$

Starting with Glatzmaier and Roberts (1995), theoreticians have usually simplified EM theory in the mantle by adopting the “thin-layer approximation” (TLA). This originated from the modeling of laboratory MHD experiments (see, e.g., Müller and Bühler 2001). It is easily applied and therefore popular, although the conditions for its validity are seldom mentioned or questioned. The TLA demands that $d \ll \hat{d}_\eta (\ll L)$. The horizontal part, $\hat{\mathbf{e}}_H$, of the electric field $\hat{\mathbf{e}}$ is then independent of r , so that

$$\hat{\mathbf{j}}_H = \hat{\sigma} \hat{\mathbf{e}}_H, \quad \text{and} \quad \hat{\mathcal{J}} = \hat{\Sigma} \hat{\mathbf{e}}_H, \quad (27c, d)$$

where $\hat{\mathcal{J}}$ is the total horizontal current carried by the layer.

It may be helpful to visualize the TLA as a mathematical limit, $d \rightarrow 0$, $\hat{\sigma} \rightarrow \infty$ with $\hat{\Sigma}$ held fixed. Then $\hat{\mathcal{J}}$ is a surface current responsible for a discontinuity in the magnetic field. If $\hat{\mathbf{b}} \left(= -\nabla \hat{\mathbf{w}} \right)$ is the potential field above the conducting layer,

$$\begin{aligned} \hat{\mathbf{b}}(r_o, \theta, \phi) - \hat{\mathbf{b}}(r_o, \theta, \phi) &= \mu_0 \hat{\mathcal{J}}(\theta, \phi) \times \mathbf{1}_r \\ &= \mu_0 \hat{\Sigma}(\theta, \phi) \hat{\mathbf{e}}_H(r_o, \theta, \phi) \times \mathbf{1}_r. \end{aligned} \quad (27e)$$

The potential field does not affect the torque, Γ^M , on the mantle, although it does contribute to the torque that each Taylor cylinder exerts on the others.

We contrast two proposed conductivity distributions. Buffett et al. (2002) inferred from their studies of nutational resonances that d is only 200 m and that $\hat{\sigma} = 3 \times 10^5 \text{ S m}^{-1}$, which is comparable to the core conductivity σ , and gives $\hat{d}_\eta = 10 \text{ km}$. The TLA should therefore be excellent in most applications. In contrast, the laboratory experiments of Ohta et al. (2008) suggest $\hat{\sigma} = 100 \text{ S m}^{-1}$ and $d = 3 \times 10^5 \text{ m}$; see also Yoshino (2010). This gives $\hat{d}_\eta = 2 \times 10^6 \text{ m}$, so it is doubtful if the TLA can be validly applied. The similar conductances of the models ($\hat{\Sigma} = 3/6 \times 10^7 \text{ S}$) are insufficient to justify the use of the TLA.

This completes our critique of the basics of torsional wave coupling to the mantle. Some of its consequences are unexpected; but most are not. Even dimensional reasoning leads to

$$\hat{\Gamma}_z^M = \hat{\Gamma}_0^M (\bar{\zeta} - \hat{\zeta}), \quad (28a, b)$$

$$\text{where } \hat{\Gamma}_0^M = O\left(r_o^4 \hat{\Sigma} \mathcal{B}_o^2\right) \approx 4 \times 10^{27} \text{ Nms},$$

for $\mathcal{B} = 0.5 \text{ mT}$. Here $\mathcal{B}_o^2$ is an average of $(B_r^N)^2$ over the CMB, and $\bar{\zeta}$ is defined by:

$$\bar{\zeta}(t) = \frac{1}{s^2 \hat{A}(s)} \int_0^{r_o} s^2 \hat{A}(s) \zeta(s, t) ds \quad (28c)$$

Perhaps unexpectedly, the boundary layer on the CMB described above reduces $\hat{\Gamma}_0^M$ by a factor of $\Sigma / (\Sigma + \hat{\Sigma})$, where Σ is the conductance of the boundary layer, defined by

$$\Sigma = \frac{1}{2} (1 + \varsigma) \sigma d_\eta = (1 + \varsigma) (\sigma / 2 \mu_0 |\omega|)^{1/2}, \quad (28d)$$

and $\varsigma = \text{sgn}(\omega)$. This gives $|\Sigma| \approx 3 \times 10^9 \text{ S}$, which is comparable with $\hat{\Sigma}$. Ignoring this factor, the magnetic interaction of mantle and FOC is governed by

$$\hat{C} d_t \hat{\zeta} = \hat{\Gamma}_0^M (\bar{\zeta} - \hat{\zeta}), \quad C d_t \bar{\zeta} = -\hat{\Gamma}_0^M (\bar{\zeta} - \hat{\zeta}), \quad (28e, f)$$

where $C = 9.08 \times 10^{36} \text{ kg m}^2$ is the polar moment of inertia of the solidly rotating FOC. These equations provide an estimate of the e -folding time, $\hat{\tau}_\eta$, taken by mantle conduction to kill $\hat{\zeta}$ and $\bar{\zeta} : \hat{\tau}_\eta = C \hat{C} / (C + \hat{C}) \hat{\Gamma}_0^M \approx 64 \text{ years}$. This is greater than the time taken by the waves to cross the core, which is $\tau_A = r_o / V_A \approx 5.6 \text{ years}$, for $\mathcal{B} = 2 \text{ mT}$.

If we take $\delta \hat{\zeta} \approx 3 \times 10^{-12} \text{ s}^{-1}$, as indicated by the LOD data of the first section above, Eq. 28e, f suggest that $\bar{\zeta} - \hat{\zeta} \approx (\hat{C} + C) \hat{\zeta} / C$ is about $2 \times 10^{-11} \text{ s}^{-1}$, so that $r_o (\bar{\zeta} - \hat{\zeta}) \approx 7 \times 10^{-5} \text{ m s}^{-1}$, which is less than, but comparable with, the inferred core surface flow. It also gives $\hat{\Gamma}_z^M \approx$

$8 \times 10^{16} \text{ Nm}$. This increases to the target torque if we take $\hat{\Sigma} = 1.2 \times 10^9 \text{ S}$, but that reduces $\hat{\tau}_\eta$ to 5.3 years, which is less than τ_A .

This highlights a difficulty that might be called the “magnetic coupling paradox,” and quantified by a quality factor:

$$\Pi = \hat{\tau}_\eta / \tau_A = (\mathcal{B} / \mathcal{B}_o)^2 (\mu_0 \hat{\Sigma} V_A)^{-1}. \quad (29)$$

There is a window for $\hat{\Sigma}$, that may be narrow or nonexistent, in which $\hat{\Gamma}_z^M$ is large enough to explain variations in LOD by torsional waves, but simultaneously is small enough to ensure that $\Pi > 1$ so that the waves are not over damped by mantle conduction. According to the admittedly imprecise, order of magnitude estimates made here, the window is nonexistent. For the target torque to be reached or exceeded, $\hat{\Gamma} \gtrsim 1.2 \times 10^9 \text{ S}$, but $\Pi > 1$ requires $\hat{\Gamma} \lesssim 1.1 \times 10^9 \text{ S}$. See also Dumberry and Mound (2008).

So far the existence of the SIC has been ignored, almost totally. We have however recognized that, for $s < r_i$, two Taylor cylinders exist, $\mathcal{C}^N(s)$ and $\mathcal{C}^S(s)$, in which $\zeta^N(s)$ and $\zeta^S(s)$ may be expected to differ. For simplicity, we assume here that they are equal. To evaluate $\tilde{\Gamma}_z$, it is necessary to link $\mathbf{b}$ to $\tilde{\mathbf{b}}$ across a boundary layer strongly influenced by Λ . An analysis of the boundary layer leads to (see RA12)

$$\tilde{b}_\phi = \mu_0 \tilde{\sigma} \tilde{d}_\eta |s B_r^N(\tilde{\zeta} - \tilde{\zeta})| \quad \text{on } r = r_i, \quad (30a)$$

where $\tilde{\zeta}$ is the angular velocity of the SIC, and $\tilde{d}_\eta = (\frac{1}{2} |\omega| \mu_0 \tilde{\sigma} i)^{-1/2} \approx d_\eta \approx 10 \text{ km}$. This leads to an expression for $\tilde{\Gamma}_z^M$, in which the main part that couples to the TC is

$$\begin{aligned} \tilde{\Gamma}_z^M &= \tilde{\Gamma}_0^M (\tilde{\zeta} - \tilde{\zeta}), \quad \text{where} \quad \tilde{\Gamma}_0^M \\ &= O(r_i^4 \tilde{\sigma} \tilde{d}_\eta \mathcal{B}_i^2); \end{aligned} \quad (30b, c)$$

cf. Eq. 28a, b. For $\mathcal{B}_i = 5 \text{ mT}$, $\tilde{\Gamma}_0^M \approx 2 \times 10^{29} \text{ Nms}$. This large torque acts on the SIC whose moment of inertia is less than $10^{-3} \hat{C}$. The coupling time, $\tilde{\tau}_\eta$, is therefore very much less than $\hat{\tau}_\eta$. Equations analogous to Eq. 28a, b give (for $\tilde{\sigma} = \sigma$)

$$\tilde{\tau}_\eta \approx \tilde{C} / |\tilde{\Gamma}_0^M| \approx 4 \text{ days}. \quad (30d)$$

This is the time taken for a mismatch between ζ and $\tilde{\zeta}$ to be wiped out by magnetic torques. Clearly the coupling between TC and SIC is substantial at frequencies of order ω_A . This supports the opinion, advanced several times in this review, that on semi-decadal time scales, the TC is effectively locked to the SIC in its rotation about Oz.

Synthesis

In this review we have analyzed the various ways in which Earth’s core is coupled to the mantle and have presented estimates of the amplitudes of these couplings in order to show which may plausibly explain the available LOD data. In our first section, we provide observational evidence for core-mantle coupling. We show that Earth’s rotation rate has a roughly semi-decadal time variability, such that the LOD fluctuates at the ms level. To explain these LOD fluctuations, an internal coupling must exist between the mantle and the core that provides torques of order 10^{18} Nm , which we named “the target torque.”

In the later text, we develop estimates of the strength of the viscous, topographic, gravitational, and electromagnetic torques. Only the viscous torque, Γ_z^V , appears to be too weak to explain the LOD signal. This is true even when we allow for the enhanced coupling that turbulence can provide.

The topographic torque on the mantle, $\hat{\Gamma}_z^T$, is created by core flow interacting with bumps on the CMB. In analyzing $\hat{\Gamma}_z^T$, it became clear that exchange of angular momentum with the SIC is significant. Order of magnitude arguments showed that potentially the largest part of $\hat{\Gamma}_z^T$ is $\hat{\Gamma}_{a,z}^T$, which is produced by the gravitational field of density anomalies in the mantle and possibly SIC, including bumps on their surfaces. This part of $\hat{\Gamma}_z^T$ is therefore intimately related to the gravitational torque $\hat{\Gamma}_z^G$. When the two are treated together as $\hat{\Gamma}_{a,z}^{G+T}$, there is an equal but opposite torque, $\hat{\Gamma}_{a,z}^{G+T}$ on the SIC. Gravitational oscillations (Buffett 1996) occur when the system is perturbed from a state of minimum gravitational energy in which $\hat{\Gamma}_{a,z}^{G+T} = \tilde{\Gamma}_{a,z}^{G+T} = 0$. An oscillation period of $\tau^G = 4.1 \text{ years}$ was derived. If, as seems likely, strong magnetic coupling exists between the tangent cylinder (TC) and the SIC, the gravitational oscillation period increases to $\tau_{TC}^G = 8.9 \text{ years}$.

The remaining part of $\hat{\Gamma}_z^T$ is $\hat{\Gamma}_{c,z}^T$, and is produced by core convection. Its importance is uncertain. From what is known today, $\Gamma_{c,z}^T$ may be 0 Nm or may exceed the target torque (cf. Kuang and Bloxham 1997; Hide 1998; Jault and Le Mouél 1999; Kuang and Chao 2001).

A simple model of torsional waves traversing FOC can explain oscillations of period, $\tau \approx 6 \text{ years}$, but there is a paradox: The target torque cannot be attained by magnetic coupling between the waves and the mantle unless a dimensionless “paradox parameter,” Π , defined in Eq. 29, is large enough. If this parameter is too large, however, the waves are damped out before they can cross the core. Whether the core can evade the paradox seems uncertain.

The topographic, gravitational and magnetic torques all have significant uncertainties in their amplitudes, but the target torque falls within these uncertainties, i.e., conceivable any of them could explain the semi-decadal LOD signals. The coupling processes may be convolved. The recent model of

Buffett et al. (2009) allows for this, but argues that the gravitational torque dominates. In contrast, Gillet et al. (2010) infer that torsional oscillations in the FOC can explain the LOD observations without strong gravitational coupling. Improvements in data and modeling of Earth's rotation (e.g., Gross 2009), the geomagnetic field (e.g., Hulot et al. 2002; Jackson 2003), core seismology (e.g., Dai and Song 2008), and the time-variations in the gravity field (e.g., Velicogna and Wahr 2006; Dumberry 2010) will all prove important in testing these core-mantle coupling arguments.

This review has focussed on explaining variations in LOD owing to internal coupling between the mantle and core. This coupling produces changes primarily in the axial angular rotation rate, Ω_z , on semi-decadal time scales. Detailed measurements now exist of variations in Earth's full rotation vector on many time scales (e.g., Gross 2007), with the different directional components providing information on different geophysical phenomena (e.g., Mathews et al. 2002). Furthermore, rotation vector and magnetic field measurements now exist for other planets (e.g., Margot et al. 2007; Uno et al. 2009), and will improve in quality in the coming decades. Such measurements will allow the development of models of deep interior structure and dynamics in planetary bodies throughout the solar system (e.g., Tyler 2008; Noir et al. 2009; Goldreich and Mitchell 2010).

Cross-References

- Core Dynamo
- Earth's Structure, Lower Mantle
- Energy Budget of the Earth
- Geomagnetic Field, Theory

Acknowledgments We thank Richard Gross, Richard Holme, and Andrew Jackson for sharing their insights and their data. We are also grateful to Bruce Buffett and the referee (Mathieu Dumberry) for giving helpful advice.

Bibliography

- Abarca del Rio R, Gambis R, Salstein DA (2000) Interannual signals in length of day and atmospheric angular momentum. *Ann Geophys* 18:347–364
- Alboussière T, Deguen R, Melzani M (2010) Melting induced stratification above the Earth's inner core due to convective translation. *Nature* 466:744–747
- Anufriev AP, Braginsky SI (1975) Influence of irregularities of the boundary of the Earth's core on the velocity of the liquid and on the magnetic field. *Geomagn Aeron* 15:754–757
- Anufriev AP, Braginsky SI (1977a) Influence of irregularities of the boundary of the Earth's core on the fluid velocity and the magnetic field, II. *Geomagn Aeron* 17:78–82
- Anufriev AP, Braginsky SI (1977b) Influence of irregularities of the boundary of the Earth's core on the fluid velocity and the magnetic field, III. *Geomagn Aeron* 17:742–750
- Braginsky SI (1970) Torsional magnetohydrodynamic vibrations in the Earth's core and variations in day length. *Geomagn Aeron* 10:1–8
- Braginsky SI (1984) Short-period geomagnetic secular variation. *Geophys Astrophys Fluid Dyn* 30:1–78
- Braginsky SI (1999) Dynamics of the stably stratified ocean at the top of the core. *Phys Earth Planet Inter* 111:21–34
- Braginsky SI, Roberts PH (1995) Equations governing convection in Earth's core and the geodynamo. *Geophys Astrophys Fluid Dyn* 79:1–97
- Braginsky SI, Roberts PH (2007) Anelastic and Boussinesq approximations. In: Gubbins D, Herrero-Bervera E (eds) *Encyclopedia of geomagnetism and paleomagnetism*. Springer, Heidelberg, pp 11–19
- Brito D, Aurnou JM, Cardin P (2004) Turbulent viscosity measurements relevant to planetary core-mantle dynamics. *Phys Earth Planet Inter* 141:3–8
- Buffett BA (1996) Gravitational oscillations in the length of day. *Geophys Res Lett* 23:2279–2282
- Buffett BA (1997) Geodynamic estimates of the viscosity of the Earth's inner core. *Nature* 388:571–573
- Buffett BA (1998) Free oscillations in the length of day: inferences on physical properties near the core-mantle boundary. *Geodynamics* 28:153–165
- Buffett BA (2010) Chemical stratification at the top of Earth's core: constraints from nutation observations. *Earth Planet Sci Lett* 296:367–372
- Buffett BA, Christensen UR (2007) Magnetic and viscous coupling at the core-mantle boundary; inferences from observations of the Earth's nutations. *Geophys J Int* 171:145–152
- Buffett BA, Glatzmaier GA (2000) Gravitational braking of inner-core rotation in geo-dynamo simulations. *Geophys Res Lett* 27:3125–3128
- Buffett BA, Mathews PM, Herring TA (2002) Modeling of nutation and precession: effects of electromagnetic coupling. *J Geophys Res* 107:2070. <https://doi.org/10.1029/2000JB000056>
- Buffett BA, Mound J, Jackson A (2009) Inversion of torsional oscillations for the structure and dynamics of Earth's core. *Geophys J Int* 177:878–890
- Dai W, Song X (2008) Detection of motion and heterogeneity in Earth's liquid outer core. *Geophys Res Lett* 35:L16311
- Davidson PA (2001) *An introduction to magnetohydrodynamics*. Cambridge University Press, Cambridge, UK
- Davidson PA (2004) *Turbulence*. Oxford University Press, Oxford, UK
- de Wijs GA, Kresse G, Vocadlo I, Dobson DP, Alfè D, Gillan MJ, Price GD (1998) The viscosity of liquid iron at the physical conditions of Earth's core. *Nature* 392:805–807
- Defraigne P, Dehant V, Wahr J (1996) Internal loading of an inhomogeneous compressible mantle with phase boundaries. *Geophys J Int* 125:173–192
- Deleplace B, Cardin P (2006) Viscomagnetic torque at the core-mantle boundary. *Geophys J Int* 167:557–566
- Dobson DP, Crichton WA, Vocadlo I, Jones AP, Wang Y, Uchida T, Rivers M, Sutton S, Brodhardt JP (2000) In situ measurements of viscosity of liquids in the Fe-FeS system at high pressures and temperatures. *Am Mineral* 85:1838–1842
- Dormy E, Roberts PH, Soward AM (2007) Core, boundary layers. In: Gubbins D, Herrero Bervera E (eds) *Encyclopedia of geomagnetism and paleomagnetism*. Springer, Heidelberg, pp 111–116
- Dumberry M (2007) Taylor's constraint and torsional oscillations. In: Cardin P, Cugliandolo LF (eds) *Dynamos*. Elsevier, Amsterdam, pp 383–401
- Dumberry M (2010) Gravity variations induced by core flows. *Geophys J Int* 180:635–650
- Dumberry M, Mound J (2008) Constraints on core-mantle electromagnetic coupling from torsional oscillation normal modes. *J Geophys Res* 113:B03102. <https://doi.org/10.1029/2007JB005135>
- Elsasser WM (1946) Induction effects in terrestrial magnetism, II. The secular variation. *Phys Rev* 70:202–212

- Fearn DR, Loper DE, Roberts PH (1981) Structure of the Earth's inner core. *Nature* 292:232–233
- Finlay CC, Dumberry M, Chulliat A, Pais MA (2010) Short timescale core dynamics: theory and observations. *Space Sci Rev* 155:177–218. <https://doi.org/10.1007/s11214-010-9691-6>
- Forte AM, Woodward RJ, Dziewonski AM (1994) Joint inversion of seismic and geo-dynamic data for models of three dimensional mantle heterogeneity. *J Geophys Res* 99:21857–21877
- Gargett AE (1984) Vertical eddy diffusivity in the ocean interior. *J Mar Res* 42:359–393
- Gillet N, Jault D, Canet E, Fournier A (2010) Fast torsional waves and strong magnetic field within the Earth's core. *Nature* 465(7294):74–77. <https://doi.org/10.1038/nature09010>
- Glatzmaier GA, Roberts PH (1995) A three-dimensional convective dynamo solution with rotating and finitely conducting inner core and mantle. *Phys Earth Planet Inter* 91:63–75
- Goldreich PM, Mitchell JL (2010) Elastic ice shells and synchronous moons: implications for cracks on Europa and non-synchronous rotation on Titan. *Icarus*. <https://doi.org/10.1016/j.icarus.2010.04.013>
- Gross RS (2001) A combined length-of-day series spanning 1832–1997: LUNAR97. *Phys Earth Planet Inter* 123:65–76
- Gross RS (2007) Earth rotation variations – long period. In: Herring TA (ed) *Physical geodesy*. Treatise on geophysics, vol 3. Elsevier, Oxford, pp 239–294
- Gross RS (2009) Ocean tidal effects on Earth rotation. *J Geodyn* 48:219–225
- Heimpel MH, Aurnou JM (2007) Turbulent convection in rapidly rotating spherical shells: a model for equatorial and high latitude jets on Jupiter and Saturn. *Icarus* 187:540–557
- Hide R (1969) Interaction between the Earth's liquid core and solid mantle. *Nature* 222:1055–1956
- Hide R (1998) A note on topographic core-mantle coupling. *Phys Earth Planet Inter* 109:91–92
- Hide R, James IN (1983) Differential rotation produced by potential vorticity mixing in a rapidly rotating fluid. *Geophys J R Astron Soc* 74:301–312
- Hide R, Clayton RW, Hager BH, Speith MA, Voorhies CV (1993) Topographic core-mantle coupling and fluctuations in Earth's rotation. In: Aki K, Dmowska R (eds) *Relating geophysical structures and processes: the Jeffreys volume*. Geophysical monograph series, vol 76. AGU, Washington, DC, pp 107–120
- Holme R (1998) Electromagnetic core-mantle coupling-I. Explaining decadal changes in the length of day. *Geophys J Int* 132:167–180
- Holme R, de Viron O (2005) Geomagnetic jerks and a high-resolution length-of-day profile for core studies. *Geophys J Int* 160:435–439
- Hulot G, Eymin C, Langlais B, Manda M, Olsen N (2002) Small-scale structure of the geodynamo inferred from Oersed and Magsat satellite data. *Nature* 416:620–623
- Jackson A (1997) Time-dependency of tangentially geostrophic core surface motions. *Phys Earth Planet Inter* 103:293–311
- Jackson A (2003) Intense equatorial flux spots on the surface of Earth's core. *Nature* 424:760–763
- Jacobs JA (1953) The Earth's inner core. *Nature* 172:297–298
- Jault D (2003) Electromagnetic and topographic coupling, and LOD variations. In: Jones CA, Soward AM, Zhang K (eds) *Earth's core and lower mantle*. Taylor and Francis, London, pp 46–76
- Jault D, Le Mouél JL (1989) The topographic torque associated with a tangentially geostrophic motion at the core surface and inferences on the flow inside the core. *Geophys Astrophys Fluid Dyn* 48:273–296
- Jault D, Le Mouél JL (1999) Comment on 'On the dynamics of topographic core-mantle coupling' by Weijia Kuang and Jeremy Bloxham. *Phys Earth Planet Inter* 114:211–215
- Jault D, Gire C, Le Mouél J-L (1988) Westward drift, core motions and exchanges of angular momentum between core and mantle. *Nature* 333:353–356
- Kawai K, Tsuchiya T (2009) Temperature profile in the lowermost mantle from seismological and mineral physics joint modeling. *Proc Natl Acad Sci U S A*. <https://doi.org/10.1073/pnas.0905920106>
- Kuang W-J, Bloxham J (1993) The effect of boundary topography on motions in the Earth's core. *Geophys Astrophys Fluid Dyn* 72:161–195
- Kuang W-J, Bloxham J (1997) On the dynamics of topographic core-mantle coupling. *Phys Earth Planet Inter* 99:289–294
- Kuang W-J, Chao BF (2001) Topographic core-mantle coupling in geodynamo modeling. *Geophys Res Lett* 28:1871–1874
- Loper DE (2007) Turbulence and small-scale dynamics in the core. In: Olson PL (ed) *Core dynamics*. Treatise on geophysics, vol 8. Elsevier, Amsterdam, pp 187–206
- Love JJ, Bloxham J (1994) Electromagnetic coupling and the toroidal magnetic field at the core-mantle boundary. *Geophys J Int* 117:235–256
- Margot JL, Peale SJ, Jurgens RF, Slade MA, Holin IV (2007) Large longitude libration of Mercury reveals a molten core. *Science* 316:710–714
- Mathews PM, Herring TA, Buffett BA (2002) Modeling of nutation and precession: new nutation series for nonrigid Earth and insights into the Earth's interior. *J Geophys Res* 107:2068. <https://doi.org/10.1029/2001JB000390>
- Monnereau M, Calvet M, Margerin L, Souriau A (2010) Lopsided growth of Earth's inner core. *Science* 328:1014–1017
- Morse SA (1986) Adcumulus growth of the inner core. *Geophys Res Lett* 13:1466–1469
- Mound JE, Buffett BA (2003) Interannual oscillations in length of day: implications for the structure of the mantle and core. *J Geophys Res* 108:2334. <https://doi.org/10.1029/2002JB002054>
- Mound JE, Buffett BA (2005) Mechanisms of core-mantle angular momentum exchange and the observed spectral properties of torsional oscillations. *J Geophys Res* 110:B08103. <https://doi.org/10.1029/2004JB003555>
- Mound J, Buffett B (2006) Detection of a gravitational oscillation in length-of-day. *Earth Planet Sci Lett* 243:383–389
- Müller U, Bühler L (2001) *Magnetofluidynamics in channels and containers*. Springer, Heidelberg
- Noir J, Hemmerlin F, Wicht J, Baca SM, Aurnou JM (2009) An experimental and numerical study of librational driven flow in planetary cores and subsurface oceans. *Phys Earth Planet Inter* 173:141–152
- Ohta K, Onada S, Hirose K, Sinmyo R, Shimizu K, Saya N, Ohishi Y, Yasuhara A (2008) The electrical conductivity of post-perovskite in Earth's D'' layer. *Science* 320:89–91
- Olsen N, Manda M (2008) Rapidly changing flows in the Earth's core. *Nat Geosci* 1:390–394
- Roberts PH, Aurnou JM (2012) On the theory of core-mantle coupling. *Geophys Astrophys Fluid Dyn* (to appear)
- Roberts PH, Soward AM (1972) *Magnetohydrodynamics of the Earth's core*. *Annu Rev Fluid Mech* 4:117–154
- Roberts PH, Yu ZJ, Russell CT (2007) On the 60-year signal from the core. *Geophys Astrophys Fluid Dyn* 43:321–330
- Rogers TM, Glatzmaier GA (2006) Angular momentum transport by gravity waves in the solar interior. *Geophys Astrophys Fluid Dyn* 653:756–764
- Schubert G, Turcotte DL, Olson P (2001) *Mantle convection in the earth and planets*. Cambridge University Press, Cambridge, UK
- Sprague M, Julien K, Knobloch E, Werne J (2006) Numerical simulation of an asymptotically reduced system for rotationally constrained convection. *J Fluid Mech* 551:141–174
- Stellmach S, Hansen U (2004) Cartesian convection driven dynamos at low Ekman number. *Phys Rev E* 70:056312
- Stix M, Roberts PH (1984) Time-dependent electromagnetic core-mantle coupling. *Phys Earth Planet Inter* 36:49–60

- Tanaka S (2010) Constraints on the core-mantle boundary topography from *P4KP-PcP* differential travel times. *J Geophys Res* 115: B04310. <https://doi.org/10.1029/2009JB006563>
- Taylor JB (1963) The magnetohydrodynamics of a rotating fluid and the Earth's dynamo problem. *Proc R Soc Lond A* 274:274–283
- Tyler RH (2008) Strong ocean tidal flow and heating on moons of the outer planets. *Nature* 456:770–773
- Uno H, Johnson CL, Anderson BJ, Korth H, Solomon SC (2009) Modeling Mercury's internal magnetic field with smooth inversions. *Earth Planet Sci Lett* 285:328–339
- Velicogna I, Wahr J (2006) Acceleration of Greenland ice mass loss in spring 2004. *Nature* 443:329–331
- Vočadlo I, Alfè D, Price GD, Gillan MJ (2000) First principles calculation of the diffusivity of FeS at experimentally accessible conditions. *Phys Earth Planet Inter* 120:145–152
- Wahr J, de Vries D (1989) The possibility of lateral structure inside the core and its implications for nutation and Earth tide observations. *Geophys J Int* 99:511–519
- Wahr J, Swenson S, Velicogna I (2006) Accuracy of GRACE mass estimates. *Geophys Res Lett* 33:L06401. <https://doi.org/10.1029/2005GL025305>
- Wicht J, Christensen UR (2010) Torsional oscillations in dynamo simulations. *Geophys J Int* 181:1367–1380
- Yoshida S, Sumita I, Kumazawa M (1996) Growth model of the inner core coupled with outer core dynamics and the resulting elastic anisotropy. *J Geophys Res* 101:28085–28103
- Yoshino T (2010) Laboratory electrical conductivity measurement of mantle minerals. *Surv Geophys* 31:163–206. <https://doi.org/10.1007/s10712-009-9084-0>

Crustal Reflectivity (Oceanic) and Magma Chamber

Satish C. Singh

Laboratoire de Géoscience Marines, Institut de Physique du Globe de Paris, Paris, France

Synonyms

Melt lens; Spreading center and ridge

Definition

Axial magma chamber (Melt lens) is a thin melt lens observed at ocean spreading centers.

Layer 2A or *Lava* is the top layer of oceanic igneous crust.

Layer 2B is a dike sequence and lies above the axial melt lens.

Layer 3 (gabbro) forms the lower oceanic crust.

Moho is a boundary between the crust and mantle.

Pg is a seismic ray that travels through the crust.

PmP is a reflection from the crust–mantle boundary.

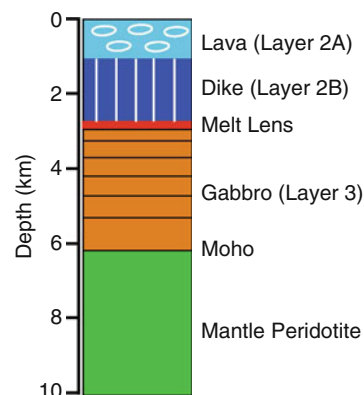
Pn is a ray that travels in the upper mantle.

Tomography is a technique to image the velocity structure of the earth.

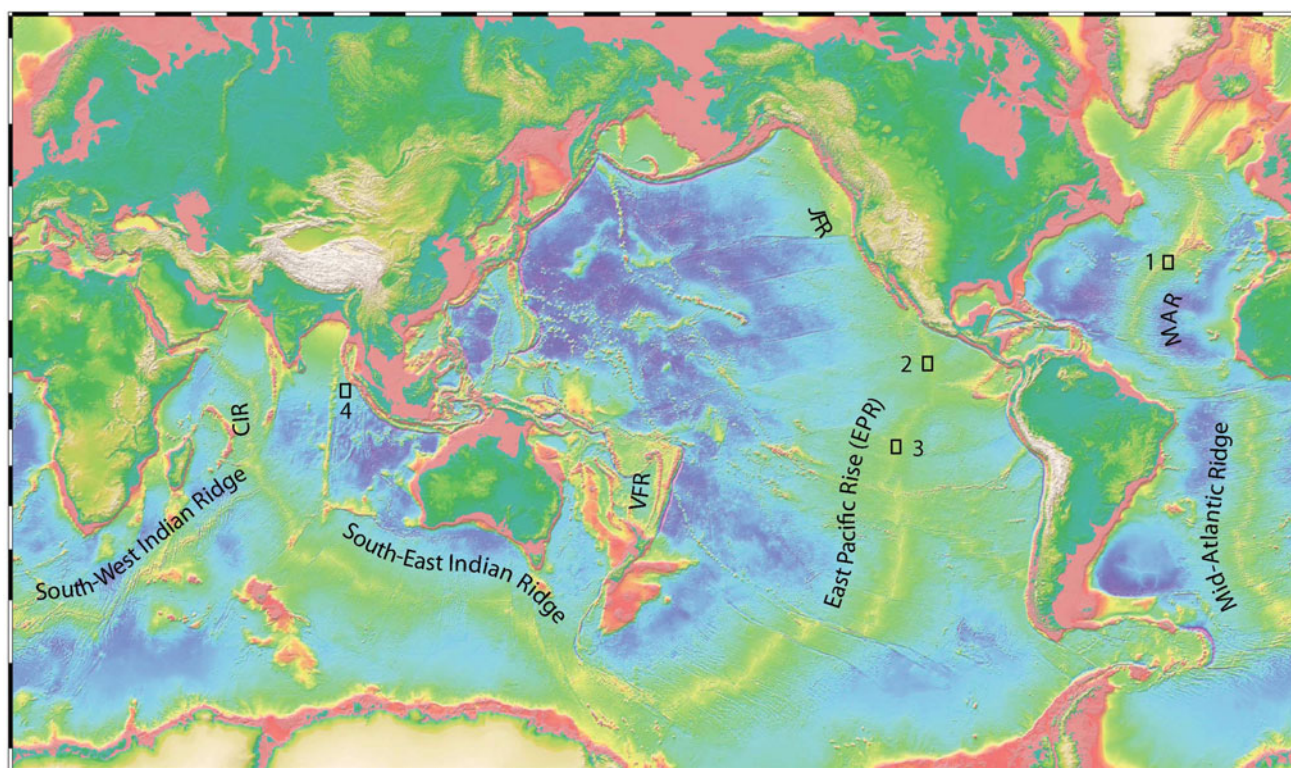
Introduction

Over 70% of the earth's crust is formed by the cooling and crystallization of melt at ocean spreading centers, which represent over 55,000 km of chains of volcanoes in the middle of the oceans, called mid-ocean ridges. At ocean spreading centers, the oceanic plate separates causing the mantle to move upward, reducing the pressure and causing the melting of the mantle. Since the newly formed melt is lighter than the surrounding mantle material, it moves upward toward the surface of the earth. Part of the melt is erupted on the seafloor as lava, which cools very rapidly forming a cap of solid extrusive layer, also known as Layer 2A (Fig. 1). As there is a significant amount of water present at mid-ocean ridges, the water circulates deep in the crust. Therefore, the melt stays mainly in the middle of the crust and erupts along thin dikes. When these dikes are cooled and crystallized, they form a thick intrusive layer or Layer 2B. Below the dikes, the melt could reside for a long period, forming a steady state melt lens, called axial melt lens or axial magma chamber (AMC). The magma cools and crystallizes in this melt lens, forming a crystalline lower crust. The melt lens forms the lower limit for further penetration of water, and therefore, partial melt is generally present beneath the melt lens down to the crust–mantle boundary. Sometimes hot melt ascending from the mantle may get injected in this partially molten region. Based on this basic process, the oceanic crust is divided into three layers, lava (extrusive), dikes (intrusive), and gabbroic crust. The relative thicknesses of these layers depend on the spreading rate, which can vary from a few millimeters up to 200 mm per year.

Based on the spreading rate, mid-ocean ridges are divided into four groups: fast, intermediate, slow, and ultraslow. East Pacific Rise (EPR) is a fast spreading center, where the spreading rate varies from 70 to 180 mm per year (Fig. 2). Juan de Fuca Ridge and Valu Fa Ridge are intermediate



Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 1 Classical model of the oceanic crust. Layers 2A and 2B form the upper crust whereas the gabbro layer corresponds to the lower crust



Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 2 Major spreading centers on the earth. JFR, Juan de Fuca Ridge; VFR, Valu Fa Ridge; CIR, Central Indian Ridge. The rectangles indicate the positions of data/results shown in this entry: 1, Lucky Strike Segment at Mid-

Atlantic Ridge (Figs. 5, 7, 9, and 12); 2, 9° N East Pacific Rise (Figs. 8, 11, and 14); 3, 14° S East Pacific Rise (Figs. 6, 10, and 13); and 4, Wharton basin (Figs. 15 and 16)

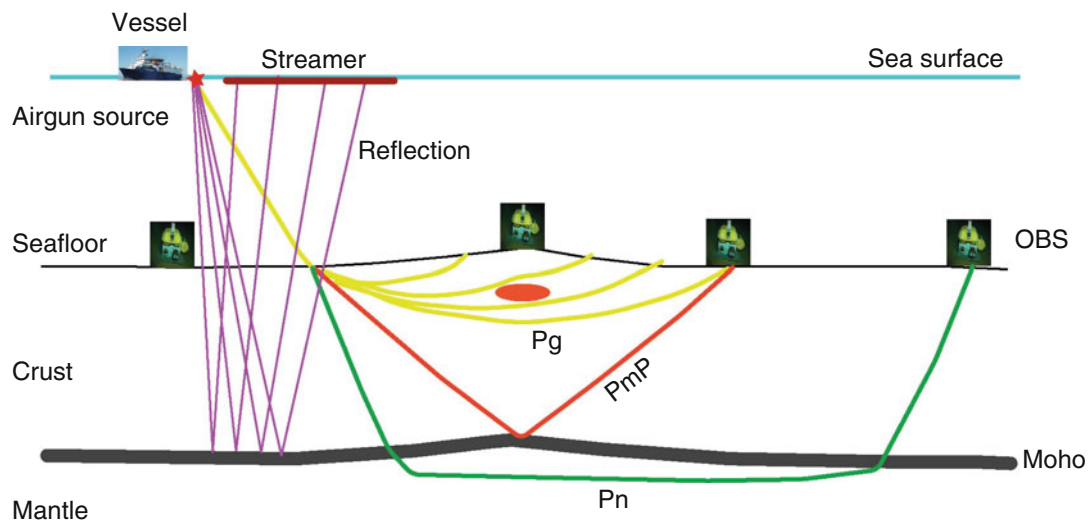
spreading centers, where the spreading rate varies from 55 to 70 mm per year. Mid-Atlantic Ridge and Central Indian Ridge are slow spreading centers with spreading rates of 20–55 mm per year. South-West Indian Ridge and Gakkel Ridge are ultraslow spreading centers with spreading rates of less than 20 mm per year. The melt production depends on the spreading rate and, therefore, the crustal thickness also varies from 3.5 to 8 km, with an average thickness of 6 km.

Methods

There are two main methods to study the oceanic crustal structures: seismic refraction or tomography and reflection methods. In seismic refraction method, a set of ocean bottom seismometers (OBS) are placed on the seafloor and an array of air guns is used to generate acoustic energy that travels through water column into the crust and mantle and is recorded on OBS (Fig. 3). There are three types of waves that arrive on an OBS: waves that travel in the crust (Pg) and the mantle (Pn) and waves reflected from the crust–mantle boundary, Moho (PmP). The travel times of arrivals of energy on these OBS are used to determine the velocity structure of the crust and upper mantle. Since OBS are small in numbers,

they are placed 5–20 km apart on the seafloor, and hence provide a very large-scale (5–10 km) velocity structure of the crust. Since the compressional (P) waves travel fast, they arrive first on OBS, and hence mainly P-wave velocities are estimated using OBS data. However, sometimes secondary (S) arrivals, converted phases Ps, are recorded on these OBS and provide information on S-wave velocity.

Seismic reflection method is most commonly used to obtain crustal reflectivity. A streamer containing pressure sensors (hydrophones) is towed behind a vessel recording seismic energy generated by an array of air guns as the vessel moves at 4.5–5 knots (Fig. 3). The length of the streamer varies from a few hundred meters up to 15 km (Singh et al. 2009), and hydrophones are generally spaced at 12.5 m. For crustal imaging the air gun shot spacing is ~50 m. Depending upon the target depth, the recording length varies from 10 to 20 s (20–60 km depth). This technique is routinely used for oil and gas exploration, and has become a standard technique for crustal studies. There are standard tools to analyze these data. Since the source receiver distance is small as compared to the depth of the target, the data are initially presented as a function of time, and then are depth-converted using velocities obtained from the OBS study or from other sources.



Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 3 Schematic diagram showing the seismic reflection and refraction survey to study the oceanic crust. Red star indicates the air gun source.

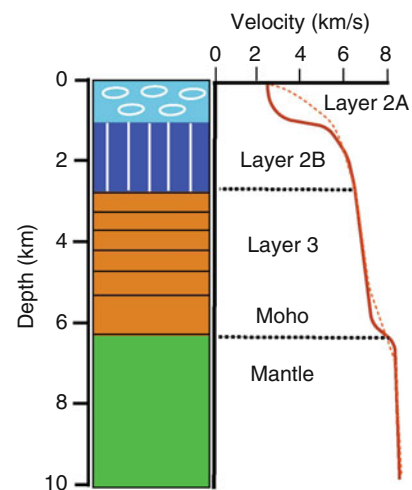
OBS, ocean bottom seismometers; Pg, crustal arrival; Pn, mantle arrival; and PmP, reflection from the crust–mantle boundary. Red ellipse indicates melt lens

Crustal Structure at Fast Spreading Centers

Based on seismic refraction and reflection studies at fast spreading centers and on ophiolites from Oman, the crust at fast spreading centers is considered as a layer cake and is divided into three distinct layers: Layer 2A (pillow lava), Layer 2B (dike sequence), and Layer 3 (gabbro) (Fig. 4). The velocity in Layer 2A varies from 2.5 to 4.5 km/s, that in Layer 2B from 4.5 to 6 km/s, and in gabbro from 6 to 7 km/s. On the ridge axis, a magma chamber is generally present between the dike and the gabbro layer.

The structure of Layer 2A is inferred either using the velocity from OBS study or from the reflection method. The boundary between Layer 2A and 2B is not sharp but there is a high velocity gradient where the velocity increases from 2.5 to 4–4.5 km/s in a thin transition zone, which leads to three arrivals (triplication) from Layer 2A (Fig. 5). However, since the first arrival tomography is performed on OBS data, spaced at 5–10 km distance, one can only get a smooth velocity of the crust. Therefore, a velocity contour of 4.5 km/s is generally used at the base of Layer 2A (Fig. 4, *dashed curve*).

In seismic reflection data, the triplication from the Layer 2A/2B boundary lies in a narrow offset range at intermediate offsets, depending on the water depth and the thickness of Layer 2A (Fig. 5) (Harding et al. 1993; Seher et al. 2010a). Since the cusp of triplication has strong amplitude, it can be used to image the base of Layer 2A, and leads to a very nice image of Layer 2A. Since the energy lies in a limited offset range, the two-way travel time for these images varies depending on the velocity used for stacking. This becomes particularly important when the thickness map needs to be prepared from 3D seismic reflection data or compared along long profiles. In order to objectively compare Layer 2A

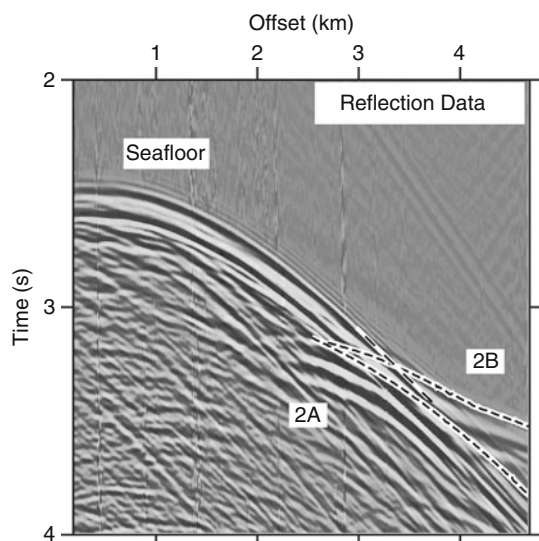


Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 4 Average one-dimensional velocity of oceanic crustal layers and mantle (solid brown line). Thin dashed line indicates velocity determined using tomographic type techniques

thickness, a constant velocity that provides the best image of Layer 2A is used to stack Layer 2A arrival and the two-way travel time is converted into depth using the tomographic velocity obtained using either OBS data or streamer tomography (Seher et al. 2010a).

Plot Triplication

Figure 6 shows the image of Layer 2A obtained along the southern East Pacific Rise (Kent et al. 1994). The thickness of the layer varies from ~200 m on the ridge axis and increases to 600 m off-axis (Kent et al. 1994; Harding et al. 1993). The thickening of Layer 2A off-axis could be associated with the



Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 5 Seismic reflection from the Lucky Strike Segment of Mid-Atlantic Ridge showing triplication (dashed curve) from Layer 2A/2B boundary. (Modified from Seher et al. 2010a)

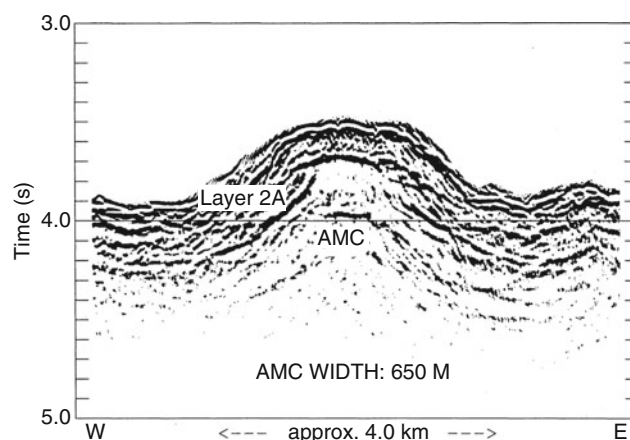
accumulation of lavas away from the ridge axis. On slow spreading ridges, Layer 2A could be up to 1 km thick (Singh et al. 2006a; Seher et al. 2010a). Instead of thickening of Layer 2A away from the spreading center, its thickness decreases, which could be due to thinning by stretching (faulting) of the crust (Fig. 7). In fact, the thinnest (300 m) Layer 2A at the Lucky Strike segment is observed near the Median bounding faults (Seher et al. 2010a).

There is still debate about the causes of Layer 2A reflection: it could be due to pillow lava and dike boundary as shown in Fig. 4 or due to alteration front associated with hydrothermal circulation or pressure collapse boundary (Christeson et al. 2007).

Below Layer 2A, the velocity increases from 4.5 up to 6 km/s, corresponding to dike sequence, below which a magma chamber reflection might be present. The thickness of the dike sequence is ~1.5 km on a fast spreading center (Vera et al. 1990) and up to 2 km on a slow spreading center (Seher et al. 2010b).

Axial Magma Chamber (Melt Lens)

On a fast spreading center, the axial magma chamber (AMC) reflector marks the boundary between the upper crust (lava and dikes) and the lower crust (gabbro). It is observed along a significant part of the fast spreading axis at 1.5–2 km below the seafloor. The width of the AMC varies from a few hundred meters to 4.5 km (Kent et al. 1993). The largest melt lens is observed beneath 9° N overlapping spreading center at the



Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 6 Seismic reflection image of Layer 2A and axial magma chamber at 14° S East Pacific Rise (Position 3 in Fig. 2). (From Kent et al. 1994). The Layer 2A thickness varies from 200 m on the ridge axis to 400 m off-axis

East Pacific Rise (Kent et al. 2000; Singh et al. 2006a) (Fig. 8). Recently, Canales et al. (2009) have imaged an off-axis melt lens. A 3D seismic reflection study of the 9° N EPR suggests that there might be an extensive presence of melt sills off-axis (Carton et al. 2009).

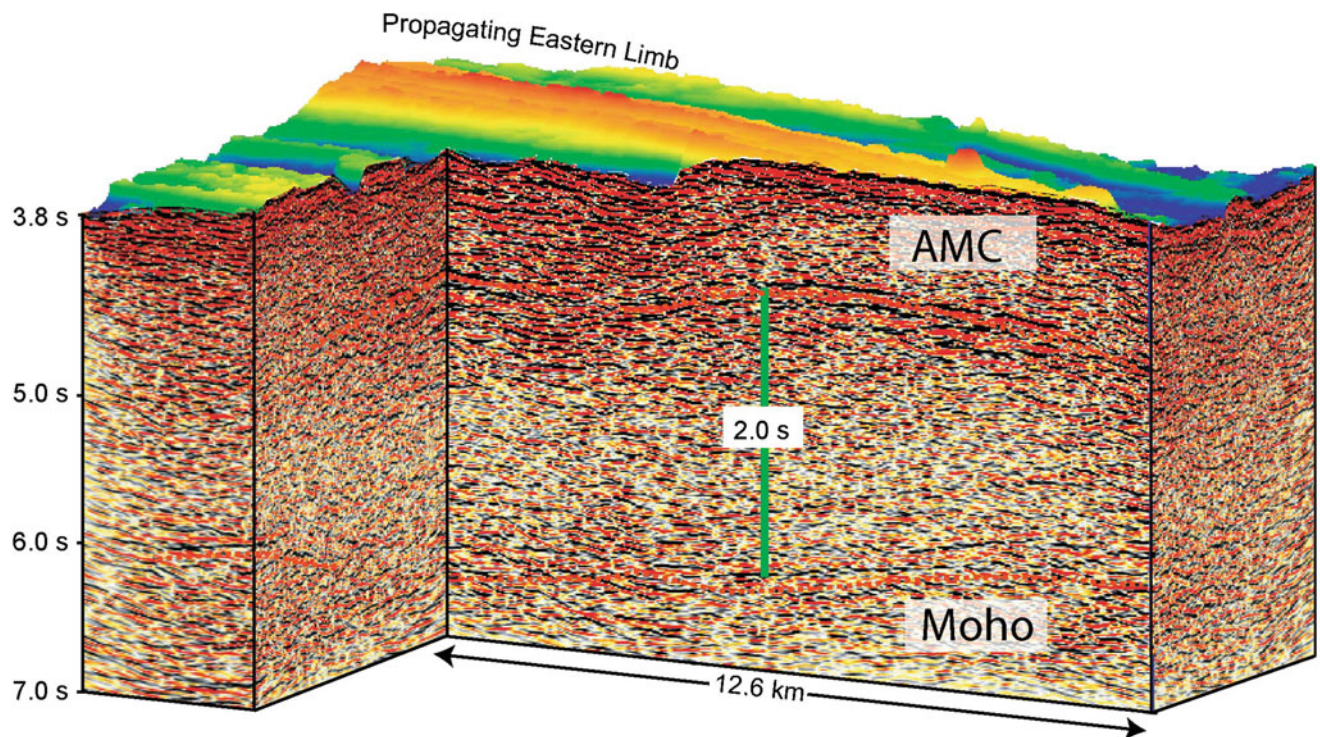
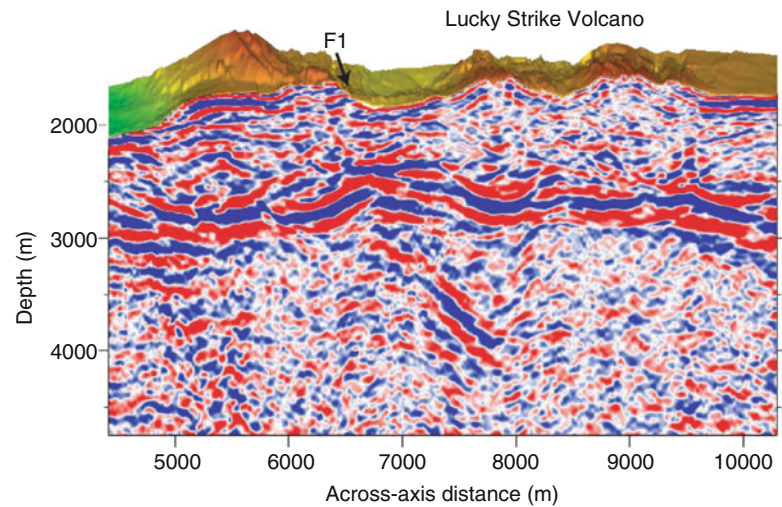
The AMC have also been observed beneath intermediate spreading centers such as Juan de Fuca Ridge (Canales et al. 2006) and Valu Fa Ridge (Collier and Sinha 1990). They are generally observed about 3 km below the seafloor. It has been difficult to image AMC reflection on slow spreading ridges, which has been due to strong scattering on the seafloor and complex 3D bathymetry. Using 3D seismic reflection technique, Singh et al. (2006b) discovered AMC beneath the Lucky Strike segment of the Mid-Atlantic Ridge (Fig. 9). The AMC is about 3 km wide and 7 km long at ~3 km below the seafloor.

The thickness of the melt lens is difficult to determine. Forward modeling of seismic reflection data from 9° N at EPR suggests that it could be 50–100 m thick (Kent et al. 1993). Seismic full waveform inversion of data from two locations at the EPR suggests that it should be 50–60 m thick (Collier and Singh 1997; Singh et al. 1998, 1999). Using full waveform of seismic reflection data from 14° S at EPR, Singh et al. (1999) found that a 60 m thick melt lens is overlain by a 60 m thick solid roof and underlain by a 200 m thick solid floor. Above the roof layer, they find a 200 km thick low velocity layer, which they associate with the presence of hydrothermal circulation (Singh et al. 1999) (Fig. 10). The roof layer could be the transition zone between the hot melt (1,200 °C) below and the hydrothermal layer (400 °C) above.

The presence of a solid floor suggests that magma cools and crystallizes in the melt lens and forms the solid floor.

Crustal Reflectivity (Oceanic) and Magma Chamber,

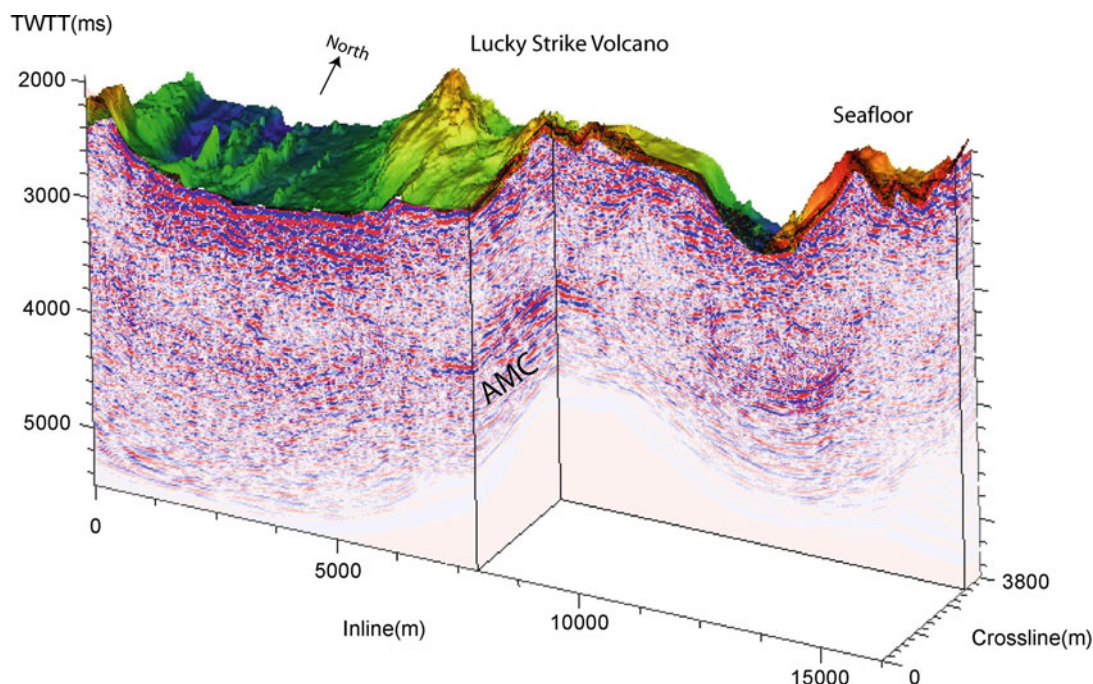
Fig. 7 Seismic reflection of Layer 2A beneath the Lucky Strike volcano at slow spreading Mid-Atlantic Ridge (Position 1 in Fig. 2). F1 indicates an active fault. Layer 2A is thick beneath the ridges and thin beneath the fault



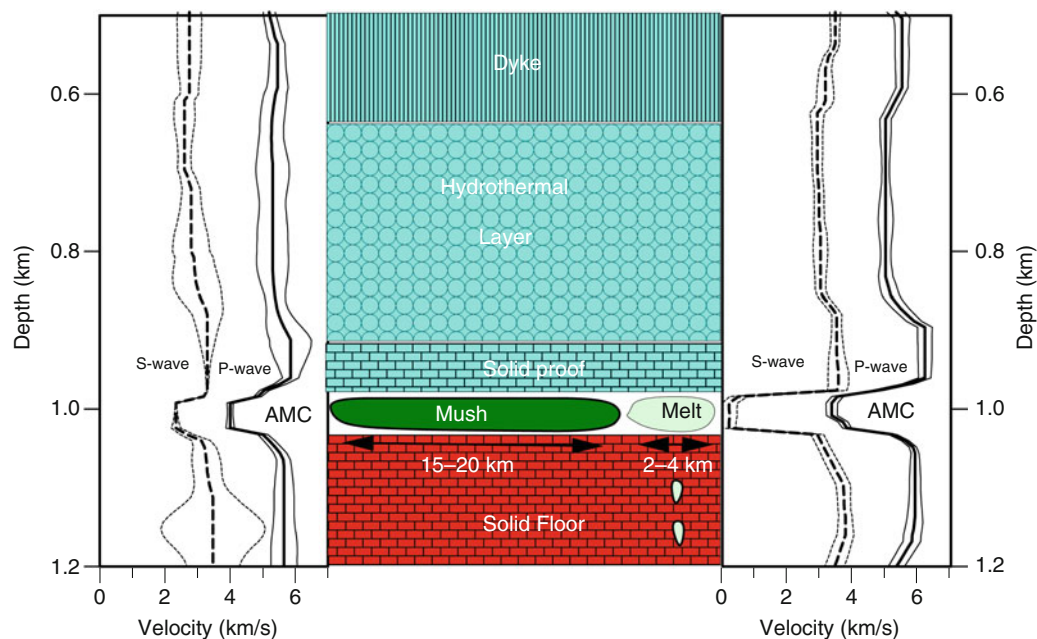
Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 8 Seismic reflection image of axial magma chamber (AMC) and Moho reflection at 9° N overlapping spreading center at the East Pacific Rise (Position 2 in Fig. 2). (Modified from Singh et al. 2006a)

Using 3D seismic reflection data, we can image the base of the melt lens. Here the melt lens is about 4.5 km wide and could be up to 250 m thick (as compared to 50–100 m observed elsewhere). This could be due to its very large size (4.5 km) and the presence of low velocity underneath it (Bazin et al. 2003). Singh et al. (2006a) suggest that up to 40% melt could be present between 1.5 and 6 km depth, making it the largest observed melt lens in the crust observed on the earth so far (Fig. 11).

Using full waveform and partial stacking technique of seismic reflection data, Singh et al. (1998) show that the melt along 14° S EPR consists of 2–4 km long pure melt zones at 15–20 km interval along the ridge axis (Fig. 10). They associate pure melt region where the shear wave velocity is nearly zero and partial melt zone where the shear wave velocity is nonzero (Taylor and Singh 2002). They suggest that a pure melt region corresponds to fresh supply of magma from the mantle and a partial melt region corresponds to



Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 9 Axial magma chamber reflection image from the Lucky Strike Segment of the Mid-Atlantic Ridge (Position 1 in Fig. 2)

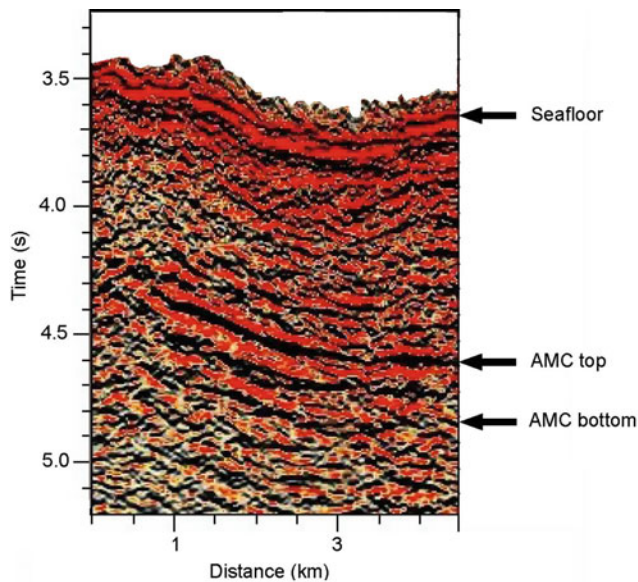


Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 10 Full waveform inversion results at two different locations along the 14° S EPR (Position 3 in Fig. 2). Inversion results show the presence of 60 m AMC, 60 m thick roof, and 200 m thick floor of the AMC. It also shows a hydrothermal layer above the roof of the AMC. The result on the

left panel corresponds to a partial melt region (less than 50% of liquid melt) whereas that on the right corresponds to a pure melt (more than 80% of liquid) region. The depth is from the seafloor. (Modified from Singh et al. 1999)

cooled and crystallized state of melt lens. They also find that a pure melt region is associated with hydrothermal activities on the seafloor, linking the supply of magma from the mantle with hydrothermal circulation on the seafloor. In order for the

melt lens to be in steady state, it would require a supply of magma from the mantle every 30 years as it will take 50 years to completely solidify a 50 m thick melt lens (Singh et al. 1999).

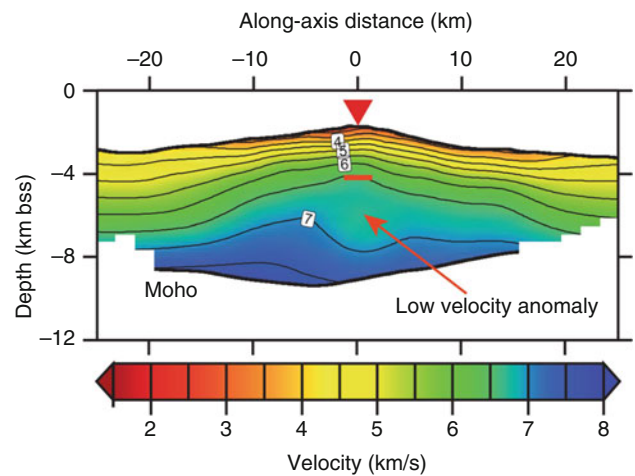


Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 11 Seismic reflection image showing the top and bottom of the axial magma chamber at 9° N EPR (Position 2 in Fig. 2)

Lower Crust

So far, no other crustal melt lenses have been observed beneath the axial melt lens. Therefore, only tomographic methods using OBS data are used to determine the velocity structure of the lower crust. Using two-ship expanding spread profile, Vera et al. (1990) and Harding et al. (1989) found that the velocity in the lower crust (5.5–6 km/s) on-axis was lower than that off-axis (6.8–7 km/s), which suggested the presence of partial melt in the lower crust. Using 3D tomography technique, Toomey et al. (1990) and Dunn et al. (2000) showed the presence of low velocity in the lower crust, suggesting the presence of partial melt. Singh et al. (2006b) found a large anomaly in two-way travel time between the Moho reflection and AMC and suggest that up to 40% of melt might be present in the lower crust (Fig. 8). A low velocity anomaly has also been observed beneath the Lucky Strike segment of the Mid-Atlantic Ridge beneath the melt lens (Seher et al. 2010c; Singh et al. 2006b). These results suggest that partial melt is present in the lower crust beneath the melt lens (Fig. 12).

Based on the study of Oman ophiolite, Kelemen et al. (1997) and Boudier et al. (1996) suggested that the lower crust is formed by injection of melt sills in the lower crust instead of cooling and crystallization of magma in the upper crustal melt sill. However, no melt lenses have been imaged in the lower crust so far, even using 3D seismic reflection technique (Singh et al. 2006a), which suggest that the melt in the lower crust must be in small pockets, not in large melt sills.



Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 12 Tomographic results from the Lucky Strike Segment of the Mid-Atlantic Ridge (Position 1 in Fig. 2) showing melt lens (red line), low velocity in the lower crust, and Moho across the ridge axis. The inverse triangle is the center of the ridge axis. (Modified from Seher et al. 2010c)

Oceanic Moho

The Moho is a boundary between the crust and mantle, where the velocity changes from 6.8–7.0 km/s to 7.8–8.1 km/s. For a sharp boundary, a clear Moho reflection is observed. However, the Moho is not just a simple boundary. Kent et al. (1994) identified three types of Moho reflection: (1) Impulsive Moho where a clear single reflection is observed (Fig. 13), (2) Diffuse Moho where reflection from Moho is patchy, and (3) Shingled Moho where reflection is shingled.

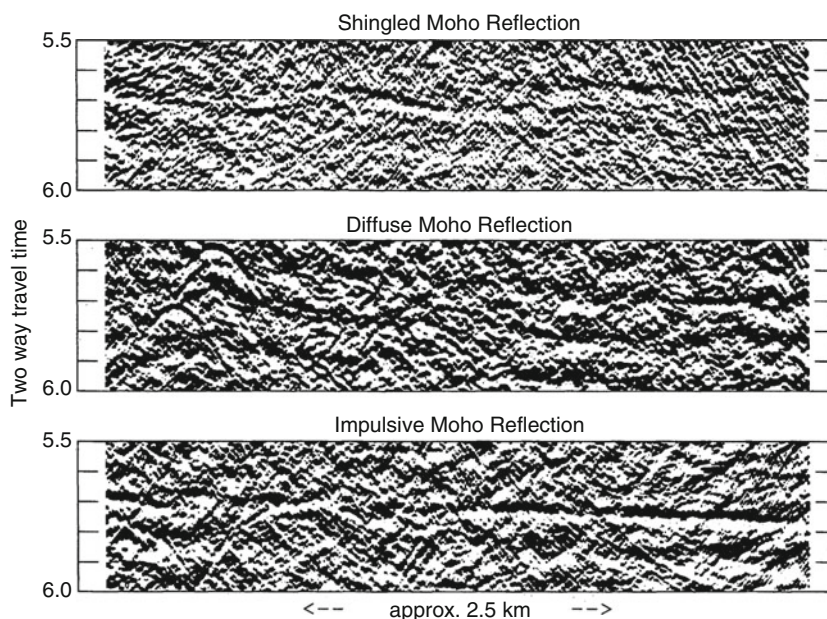
Moho reflections are generally observed away from the spreading center, not beneath the melt lens. This is because the upper crustal melt lens and associated lower crustal melt would attenuate seismic energy and hence, it would be difficult to image Moho using conventional seismic reflection techniques. Secondly, it was accepted that Moho is formed away from the ridge axis. However, using 3D seismic reflection technique, Singh et al. (2006a) have imaged Moho beneath the wide axial melt lens (Fig. 14), suggesting the Moho is formed at zero age.

Moho reflections are observed ~2 s two-way travel time below the seafloor. The average crustal velocity is about 6 km/s, and hence the average crustal thickness is about 6 km. However, Singh et al. (2011) have observed Moho reflection 1.3 s below the seafloor offshore NW Sumatra, where the crust was formed at the fast spreading Wharton ridge about 55 Ma ago. If we assume an average velocity of 6.0 km/s, which means at the crust there is only 3.5 km thick, it would be the thinnest crust ever observed in a fast spreading environment (Fig. 15).

Since there is a positive velocity contrast at the crust–mantle boundary, large amplitude critical angle reflections

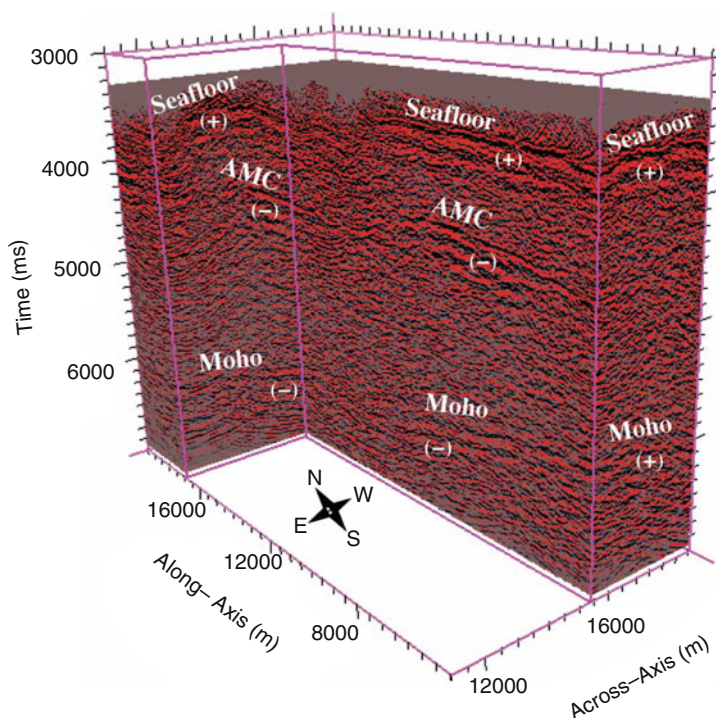
Crustal Reflectivity (Oceanic) and Magma Chamber,

Fig. 13 Different types of reflection for Moho boundary observed at Southern East Pacific Rise (Position 3 in Fig. 2) (Kent et al. 1994). Top: Shingled Moho, Middle: Diffuse Moho, Bottom: Impulsive Moho



Crustal Reflectivity (Oceanic) and Magma Chamber,

Fig. 14 Seismic reflection image of Moho beneath the melt lens at the 9° N East Pacific Rise (Position 2 in Fig. 2)



(PmP) are generated that arrive at large offsets (15–30 km), and are recorded on OBS. These data are then used to map the Moho structure. However, there are two problems in using arrival times of PmP arrivals. The velocity gradient in the lower crust is very small or close to zero, and therefore rays that sample the lower crust do not turn in the lower crust making it difficult to estimate velocity in this zone. Since the arrival time of PmP arrival depends on the velocity in

the crust, particularly the lower 3 km of the crust, it is difficult to estimate both the velocity in the lower crust and the Moho structure. Since the velocity in the lower crust can be approximated to a large extent ~ 6.8 km/s, these data provide a reasonable constrain on the Moho structure, and are routinely used. Secondly, the OBS spacing is generally 5–20 km and the lateral resolution of the Moho structure is very poor.

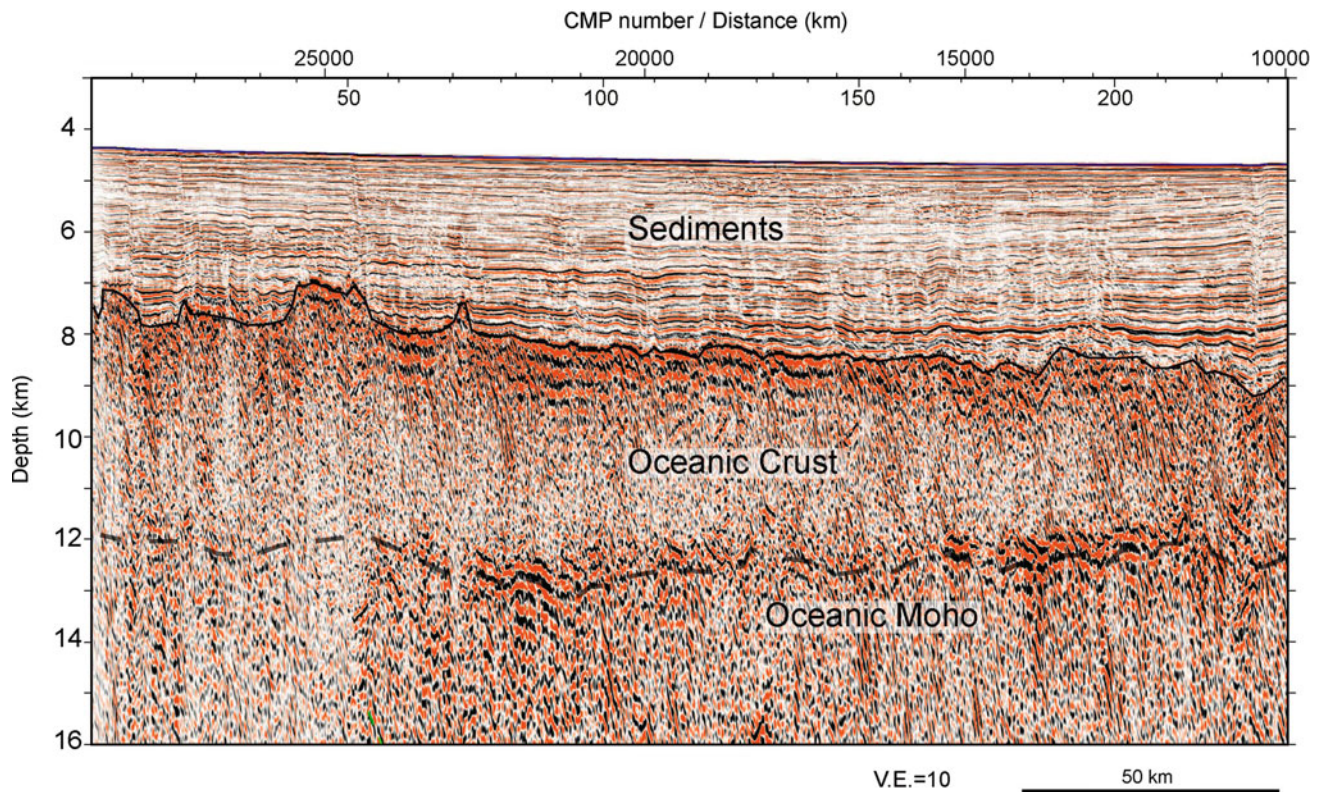
Crustal Thickness

Once we determine the velocity in the crust and Moho structure, we can estimate the crustal thickness. So far, most of the crustal thickness estimations are based on wide-angle seismic studies, and there are significant uncertainties in crustal thickness estimations for the reason explained above. However, combined analyses of seismic reflection, refraction, and gravity data suggest that wide-angle estimation of the crustal thickness can be very reliable.

The crust formed at fast spreading centers is generally uniform and 6 km thick (White et al. 1992; Eittreim et al. 1994) but some variations (5–8 km) have been observed recently (Canales et al. 2003; Singh et al. 2006a; Barth and Mutter 1996). There is no significant crustal thickness variation across the fracture zone in a fast spreading environment, where it is also 5.5–6 km thick (Van Avendonk et al. 2001). The crust formed at slow spreading centers is generally thick at the center of the segment (6–7 km) and thin at segment ends (4–5 km) (Barclay et al. 1998; Seher et al. 2010b). The crust beneath fracture zones in a slow spreading environment is generally thin (3–4.5 km) (Detrick et al. 1993). The thinnest crust (3.5–4.5 km) is formed at ultraslow spreading centers (Jokat and Schmidt-Aursch 2007). These observations

suggest that the thickness of the crust depends on the spreading rate and melt distribution in the crust. On the other hand, thicker crusts are found beneath large igneous provinces, across hotspot tracks or at the interaction of plume and ridges, which are believed to be formed by higher mantle temperatures due to the presence of a plume (e.g., Parkin and White 2008; Watts et al. 1985; Charvis et al. 1999; Grevenmeyer et al. 2001). For example, the crust beneath Iceland could be up to 30 km thick, and beneath La Reunion Island it is 13 km (Charvis et al. 1999).

However, anomalously thin crust has been reported. A 5 km thick crust is observed near the South American trench (Grevenmeyer et al. 2007) and in the area of IODP Hole 1256 (Hallenborg et al. 2003) for crusts formed at the EPR about 20 Ma. Rodger et al. (2006) found an ultrathin crust (4 km) that was formed near the central MAR about 85 Ma ago, which they associate to be due to a reduction in the spreading rate from slow to ultraslow. The crustal study at ODP holes 504B also found thin crust (5 km), which Collins et al. (1989) associate to be due to depleted mantle. Singh et al. (2011) have observed ultrathin crust (3.5–4.5 km) in the Indian Ocean (Fig. 15) that was formed at fast spreading Wharton center 55 Ma ago, which they associate to be due to the interaction of the Kerguelen plume with the Wharton



Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 15 Seismic reflection image of extremely thin crust in the Indian Ocean (Position 4 in Fig. 2). (Modified from Singh et al. 2011)

Spreading center due to channeling of the cold lithosphere around a plume and its interaction with the spreading center.

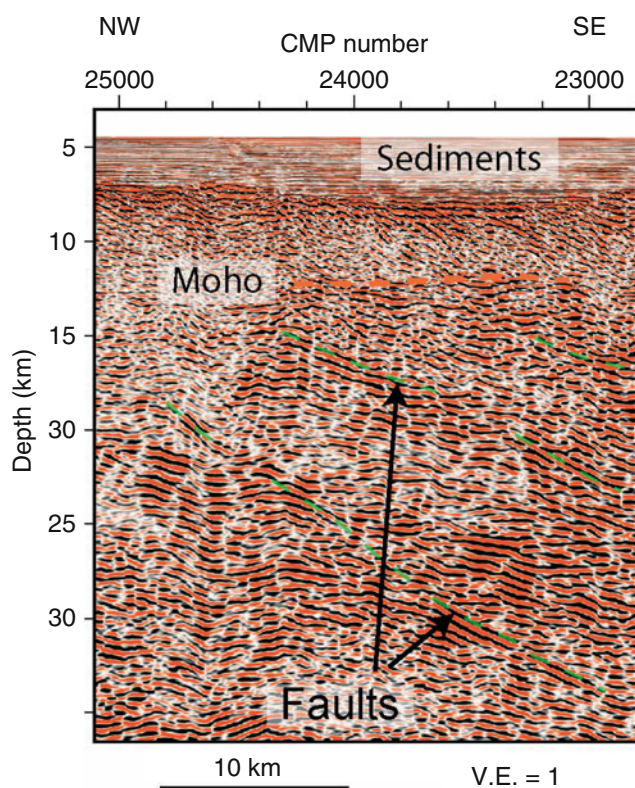
Upper Mantle

The velocity in the upper mantle is generally determined using rays that turn below the Moho, which are called Pn arrivals. The velocity in the mantle beneath ridge axis is ~ 7.6 – 7.8 km/s and is 8–8.1 km/s away from the ridge axis (Vera et al. 1990; Harding et al. 1989). An extensive 3D tomography and undershoot experiment were carried out at 9° N EPR. Using these data, Dunn et al. (2000, 2001) obtained a symmetric low velocity anomaly below the Moho in the mantle. Using the same data and some new data Toomey et al. (2007) found asymmetric low velocity anomaly at 10 km away from the ridge axis. There are serious problems with these results. First, the crustal thickness in this area varies from 6 to 8 km. Secondly, the velocity anomaly depends on the background velocity used during the inversion. Dunn et al. (2000) use a starting mantle velocity of 8.2 km/s whereas one year later Dunn et al. (2001) use 7.6 km/s. Although the crustal thickness is up to 8 km (Canales et al. 2003), Dunn et al. (2001) interpret velocity at 7 km below the seafloor; the part of the anomaly could be due to crustal variations. Toomey et al. (2007) show that the ray only penetrates down to 8 km below the seafloor but interpret the velocity at 9 km below the seafloor. These conflicting results and the inaccuracy in inversion led Singh and Macdonald (2009) to suggest that these results are not reliable. The only robust solution we have is that the velocity in the mantle is ~ 7.6 – 7.8 km/s below the ridge axis and 8–8.1 km/s away from the ridge axis (Vera et al. 1990; Harding et al. 1989). However, it is possible that melt sills get trapped in the mantle as shown by Nedimovic et al. (2005) and may lead to a low velocity anomaly in the mantle.

Faults in the Oceanic Crust and Upper Mantle

Along with magmatic process, tectonic processes also play an important role in shaping the oceanic crust. Singh et al. (2006b) have shown that median valley bounding faults could be imaged down to 3 km, close to the melt lens. They also found extensive faults above the melt lens. It is possible that some of these faults penetrate down to the crust–mantle boundary (Dusunur et al. 2009).

In a subduction zone environment, there are two forces that act on the incoming oceanic plate: plate bending leading to normal faulting and compressive forces leading to thrust faulting. Water can enter into the mantle leading to serpentinization of the oceanic mantle. Normal faulting down to



Crustal Reflectivity (Oceanic) and Magma Chamber, Fig. 16 Seismic reflection of deep penetrating faults in the mantle in the Indian Ocean (Position 4 in Fig. 2)

5–8 km below the Moho, related to bending of the subducting plate and the associated serpentinization, has been reported for the Central American trench (Ranero et al. 2003). Faults have also been imaged in the mantle offshore Sumatra down to 35 km depths (Fig. 16) suggesting that a significant part of the oceanic lithosphere is involved during oceanic earthquakes.

Summary

The reflectivity of the oceanic crust and axial magma chambers at ocean spreading centers are described in detail using real seismic images. The oceanic crust consists of three distinct layers: Layer A is 200–1,000 m thick and consists mainly of pillow lavas. Layer B is 1–2 km thick and consists mainly of cooled basalts in dikes. Layer 3 is 2–3 km thick and is formed by the cooling and crystallization of magma in the melt lens. Axial melt lens has been imaged on fast and intermediate spreading centers and recently on a slow spreading center (Singh et al. 2006b). Melt lenses are 50–60 m and have a roof and a floor. Between the melt and Moho, lower crust, partial melt is generally present. The oceanic crust is about 6 km thick, which could vary from 3.5 to 8 km. The

P-wave velocity in the mantle lie between 7.6 and 8.1 km/s, but the detailed nature of the oceanic mantle is poorly constrained and requires further investigations. Faults due to earthquakes have been observed down to 35 km depth.

Cross-References

- [Continental Crustal Structure](#)
- [Deep Seismic Reflection and Refraction Profiling](#)
- [Lithosphere, Oceanic](#)
- [Ocean Bottom Seismics](#)
- [Ocean, Spreading Center](#)
- [Seismic Data Acquisition and Processing](#)
- [Seismic Structure at Mid-Ocean Ridges](#)
- [Single and Multichannel Seismics](#)

Bibliography

- Barclay AH, Toomey DR, Solomon SC (1998) Seismic structure and crustal magmatism at the Mid-Atlantic Ridge, 35° N. *J Geophys Res* 103:17827–17844
- Barth GA, Mutter JC (1996) Variability in oceanic crustal thickness and structure: multichannel seismic reflection results from the northern East Pacific Rise. *J Geophys Res* 101:17951–17975
- Bazin S, Harding AJ, Kent GM, Orcutt JA, Singh SC, Tong CH, Pye JW, Barton PJ, Sinha MC, White RS, Hobbs RW, van Avendonck HJA (2003) A three-dimensional study of axial low velocity beneath the 9° 03' N overlapping spreading center. *Geophys Res Lett* 30:1–4
- Boudier F, Nicolas A, Ildefonse B (1996) Magma chambers in the Oman ophiolite: fed from the top and the bottom. *Earth Planet Sci Lett* 144:238–250
- Canales J-P, Detrick RS, Toomey DR, Wilcock WSD (2003) Segment-scale variations in the crustal structure of 150–300 kyr old fast spreading oceanic crust (East Pacific Rise, 8° 15' N–10° 50' N) from wide-angle seismic refraction profiles. *Geophys J Int* 152:766–794
- Canales J-P, Singh SC, Detrick RS, Carbotte SM, Harding AJ, Kent GM, Diebold JB, Babcock J, Nedimovic MR (2006) Seismic evidence for variations in axial magma chamber properties along the southern Juan de Fuca Ridge. *Earth Planet Sci Lett* 246:353–366
- Canales JP, Nedimovic MR, Kent GM, Carbotte SM, Detrick RS (2009) Seismic reflection images of a near-axis melt sill within the lower crust at the Juan de Fuca Ridge. *Nature* 460:89–94
- Carton HD, Carbotte SM, Mutter JC, Canales J, Nedimovic MR, Marjanovic M, Aghaei O, Xu M, Han S, Stowe L (2009) Characteristics of the crustal magma body in the 2005–2006 eruption area at 9° 50' N on the East Pacific Rise from 3D multi-channel seismic data. AGU, Fall meeting, Abstract OS11B-02
- Charvis P, Laesanpura A, Gallart J, Hirn A, Lepine J, de Voogd B, Minshull TA, Hello Y, Pontoise B (1999) Spatial distribution of hotspot material added to the lithosphere under La Reunion, from wide-angle seismic data. *J Geophys Res* 104:2875–2893
- Christeson G, McIntosh K, Karson J (2007) Inconsistency in correlation of layer 2A and lava layer thickness in oceanic crust. *Nature* 445:418–421
- Collier JS, Singh SC (1997) Detailed structure of the top of the melt body beneath the East Pacific Rise at 9°40' N from waveform inversion of seismic reflection data. *J Geophys Res* 102:20287–20304
- Collier JS, Sinha M (1990) Seismic images of a magma chamber beneath the Lau Basin back-arc spreading centre. *Nature* 346:646–648
- Collins JA, Purdy GM, Brocher TM (1989) Seismic velocity structure at Deep Sea Drilling Site 504B, Panama Basin: evidence for thin oceanic crust. *J Geophys Res* 94:9283–9302
- Detrick RS, Buhl P, Vera E, Mutter J, Orcutt J, Madsen J, Brocher T (1987) Multi-channel seismic imaging of a crustal magma chamber along the East Pacific Rise. *Nature* 326:35–41
- Detrick RS, White RS, Purdy GM (1993) Crustal structure of North Atlantic Fracture Zone. *Rev Geophys* 31:439–458
- Dunn RA, Toomey DR, Solomon SC (2000) Three-dimensional seismic structure and physical properties of the crust and shallow mantle beneath the East Pacific Rise 9° 30' N. *J Geophys Res* 105:23537–23555
- Dunn RA, Toomey DR, Detrick RS, Wilcock WSD (2001) Continuous mantle melt supply beneath an overlapping spreading centre on the East Pacific Rise. *Science* 291:1955–1958
- Dusunur D, Seher T, Combier V, Escartin J, Crawford W, Cannat M, Singh SC, Matias LM, Miranda JM (2009) Seismological constraints on the thermal structure and interaction of tectonic and magmatic processes around the Lucky Strike magma chamber, slow spreading Mid-Atlantic Ridge. *Mar Geophys Res* 30:105–120
- Eittreim SL, Gribidenko H, Helsley CE, Sliter R, Mann D, Ragozin N (1994) Oceanic crustal thickness and seismic character along a central Pacific transect. *J Geophys Res* 99:3139–3145
- Grevemeyer I, Flueh ER, Reichert C, Bialas J, Klaschen D, Kopp C (2001) Crustal architecture and deep structure of the Ninetyeast Ridge hotspot trail from active-source ocean bottom seismology. *Geophys J Int* 144:414–431
- Grevemeyer I, Ranero CR, Flueh ER, Klaschen D, Bialas J (2007) Passive and active seismological study of bending-related faulting and mantle serpentinisation at the Middle American trench. *Earth Planet Sci Lett* 258:528–542
- Hallenborg E, Harding AJ, Kent GM (2003) Seismic structure of 15 Ma oceanic crust formed at an ultra-fast spreading East Pacific Rise: evidence for kilometer-scale fracturing from dipping reflectors. *J Geophys Res* 108:2532. <https://doi.org/10.1029/2003JB002400>
- Harding AJ, Kappus ME, Orcutt JA, Vera EE, Buhl P, Mutter JC, Detrick R, Brocher T (1989) The structure of young oceanic crust 13° N on the East Pacific Rise from expanding spread profiles. *J Geophys Res* 94:12163–12196
- Harding AJ, Kent GM, Orcutt AJ (1993) A multichannel seismic investigation of upper crustal structure at 9° N on the East Pacific Rise: implications for crustal accretion. *J Geophys Res* 98:13925–13944
- Jokat W, Schmidt-Aursch MC (2007) Geophysical characteristics of the ultra-slow spreading Gakkel Ridge, Arctic Ocean. *Geophys J Int* 168:983–998
- Kelemen PB, Koga K, Shimizu N (1997) Geochemistry of gabbro sills in the crust–mantle transition zone of the Oman ophiolite: implications for the origin of the oceanic lower crust. *Earth Planet Sci Lett* 146:475–488
- Kent GM, Harding AJ, Orcutt JA (1993) Distribution of magma beneath the East Pacific Rise between the Clipperton Transform and the 9° 17' N Deval from forward modelling of common depth point data. *J Geophys Res* 98:13945–13969
- Kent GM, Harding AJ, Orcutt JA, Detrick RS, Mutter JC, Buhl P (1994) Uniform accretion of oceanic crust south of the Garrett transform at 14° 15' S on the East Pacific Rise. *J Geophys Res* 99:9097–9116
- Kent GM, Singh SC, Harding AJ, Sinha MC, Tong V, Barton PJ, Hobbs R, White R, Bazin S, Pye J (2000) Evidence from three-dimensional reflectivity images for enhanced melt supply beneath mid-ocean-ridge discontinuities. *Nature* 406:614–618
- Nedimovic MR, Carbotte SM, Harding AJ, Detrick RS, Canales JP, Diebold JB, Kent GM, Tischer M, Babcock JM (2005) Frozen magma lenses below the oceanic crust. *Nature* 436:1149–1152

- Parkin CJ, White RS (2008) Influence of the Iceland mantle plume on the oceanic crustal generation in the North Atlantic. *Geophys J Int* 173:168–188
- Ranero CR, Reston TJ, Belykh I, Gnibidenko H (1997) Reflective oceanic crust formed at a fast spreading centre in the Pacific. *Geology* 25:499–502
- Ranero CR, Phipps Morgan J, McIntosh K, Reichert C (2003) Bending-related faulting and mantle serpentinization at the Middle America trench. *Nature* 425:367–373
- Rodger M, Watts AB, Greenroyd CJ, Peirce C, Hobbs RW (2006) Evidence for unusually thin oceanic crust and strong mantle beneath the Amazon Fan. *Geology* 34:1081–1084
- Seher T, Crawford W, Singh SC, Cannat M (2010a) Seismic layer 2A variations in the Lucky Strike segment at the Mid-Atlantic Ridge from reflection measurements. *J Geophys Res* 115:B07107. <https://doi.org/10.1029/2009JB006783>
- Seher T, Singh SC, Crawford W, Escartin J (2010b) Upper crustal velocity structure beneath the central Lucky Strike Segment from seismic refraction measurements. *Geochem Geophys Geosyst* 11:5. <https://doi.org/10.1029/2009GC002894>
- Seher T, Crawford W, Singh SC, Cannat M, Combier V, Dusunur D, Canales J-P (2010c) Crustal velocity structure of the Lucky Strike segment of the Mid-Atlantic Ridge (37°N) from seismic refraction measurements. *J Geophys Res* 115:B03103. <https://doi.org/10.1029/2009JB006650>
- Singh SC, Macdonald K (2009) Mantle skewness and ridge segmentation. *Nature* 458:E11–E12
- Singh SC, Kent GM, Collier JS, Harding AJ, Orcutt JA (1998) Melt to mush variations in crustal magma properties along the ridge crest at the southern East Pacific Rise. *Nature* 394:874–878
- Singh SC, Collier JS, Kent GMJS, Harding AJ, Orcutt JA (1999) Seismic evidence for a hydrothermal layer above the solid roof of axial magma chamber at the southern East Pacific Rise. *Geology* 27:219–222
- Singh SC, Harding A, Kent G, Sinha MC, Combier V, Hobbs R, Barton P, White R, Tong V, Pye J, Orcutt JA (2006a) Seismic reflection images of Moho underlying melt sills at the East Pacific Rise. *Nature* 442:287–290
- Singh SC, Crawford W, Carton H, Seher T, Combier V, Cannat M, Canales J, Dusunur D, Escartin J, Miranda M (2006b) Discovery of a magma chamber and faults beneath a hydrothermal field at the Mid-Atlantic Ridge. *Nature* 442:1029–1033
- Singh SC, Midenet S, Djajadihardja Y (2009) Seismic evidence of the locked and unlocked Sumatra subduction zone. *Eos* 90:471–472
- Singh SC, Carton H, Chauhan A et al (2011) Extremely thin crust in the Indian Ocean possibly resulting from Plume-Ridge interaction. *Geophys J Int*. <https://doi.org/10.1111/j.1365-246X.2010.04823.x>
- Taylor M, Singh SC (2002) Composition and microstructure of magma bodies from effective medium theory. *Geophys J Int* 149:15–21
- Toomey DR, Purdy GM, Solomon SC, Wilcock WSD (1990) The three-dimensional seismic velocity structure of the East Pacific Rise near latitude 9°30' N. *Nature* 347:639–645
- Toomey DR, Joussetin D, Dunn RA, Wilcock WSD, Detrick RS (2007) Skew of mantle upwelling beneath the East Pacific Rise governs segmentation. *Nature* 446:409–414
- Van Avendonk HJA, Harding AJ, Orcutt JA, McClain JS (2001) Contrast in crustal structure across the Clipperton transform fault from travel time tomography. *J Geophys Res* 106:10961–10981
- Vera EE, Mutter JC, Buhl P, Orcutt JA, Harding AJ, Kappus ME, Detrick RS, Brocher TM (1990) The structure of 0 to 0.2 m.y. old oceanic crust at 9° N on the East Pacific Rise from expanding spread profiles. *J Geophys Res* 95:15529–15556
- Watts AB, ten Brink US, Buhl P, Brocher TM (1985) A multi-channel seismic study of lithospheric flexure across the Hawaiian Emperor seamount chain. *Nature* 315:105–111
- White RS, McKenzie D, O'Nions RK (1992) Oceanic crustal thickness from seismic measurements and rare earth element inversions. *J Geophys Res* 97:19683–19715

Curie Temperature

Vincenzo Pasquale

Department of Earth, Environment and Life Sciences,
University of Genoa, Genoa, Italy

Synonyms

Curie point

Definition

A rock containing magnetic minerals loses its permanent magnetism when heated up to a critical temperature, referred to as Curie temperature or Curie point, T_c . On the atomic level, below T_c , the magnetic moments are aligned in their respective domains, and even a weak external magnetic field results in a magnetization. As the temperature increases to T_c and above, fluctuations due to the increase in thermal energy destroy that alignment, and the rock becomes paramagnetic.

Magnetic Minerals and Curie Temperature

The most important magnetic rock-forming minerals are oxides of iron and titanium, and their compositions can be represented in the FeO-TiO₂-Fe₂O₃ ternary diagram. Among the three major solid-solution series identified in this system, the titanomagnetite series, Fe_{3-x}Ti_xO₄ (with $0 \leq x \leq 1$), plays a major role in controlling rock magnetism (Stacey 1992). Above 600 °C there is a continuous solid solution in the two-component series magnetite (Fe₃O₄) and ulvöspinel (Fe₂TiO₄), which upon cooling is restricted toward the end-members. Magnetite, the most abundant and strongly magnetic mineral, is of ferrimagnetic type, i.e., neighboring magnetic moments are aligned antiparallel, as in antiferromagnetism, but unequal numbers or strengths, thus giving a net magnetization.

There are many other magnetic minerals, but they are rare (iron sulfide), unstable (maghemite), or having a weak spontaneous magnetization due to the canting of its equal and nearly opposite atomic moments (canted antiferromagnetism), as hematite occurring in sedimentary rocks often in solid solution with ilmenite.

Variation of Curie temperature with titanium substitution in the titanomagnetite series is approximately given by $T_c = 580 (1 - 1.26 x)$. In pure magnetite, T_c is 575–585 °C, but titaniferous inclusions can reduce T_c , which approximates room temperature for $x = 0.77$ (Hunt et al. 1995). Saturation magnetization is also a function of temperature, and it disappears above T_c ; at room temperature it decreases from 90 to 92 A m² kg⁻¹ for $x = 0$ to about zero for $x = 0.8$. Because Ti⁴⁺

has no unpaired spins, the saturation magnetization decreases with increasing x (Alva-Valdivia and López-Loera 2011). The cell dimensions increase with increasing x . As a result of the increased cell dimension, there is a decrease in Curie temperature.

T_c undergoes a small increase with pressure. At the boundary between the crust and the mantle (or Moho), the T_c increase should be no more than a few degrees of the experimental values at normal pressure.

Geomagnetic Implications

There is a connection between the anomalies of geomagnetic field and the temperature-depth distribution, since rock magnetization is strongly affected by temperature variations. The rocks lose their magnetization at a depth where the temperature is greater than T_c and, consequently, their ability to generate detectable magnetic anomalies disappears. By transforming magnetic anomaly data into the Fourier domain and analyzing their spectra, it is possible to infer the depth of the magnetic layer bottom (MLBD), i.e., where magnetic rocks are replaced with nonmagnetic material. Its comparison with information about thermal state of the Earth's crust can provide interpretation for both the magnetic source distribution and the Curie temperature.

Chiozzi et al. (2005) gave an example of MLBD determination and its relation with Moho depth, thermal structure, and geothermal flow in Central and Southern Europe (Table 1). MLBD corresponds to the Moho in the Variscan units, and the expected temperature is close to T_c of magnetite. Beneath the Alpine units, the magnetic layer bottom is shallower than the Moho and corresponds to temperature, as inferred from the geothermal flow, of 550 °C, indicating the presence of Ti. In the Ligurian basin, the temperature at the MLBD minimum depth is of 570 °C, hence compatible with the Curie point of magnetite. However, this depth is slightly larger than the Moho. Like observed in some oceanic regions (Counil et al. 1989), this might mean that the uppermost part of the mantle has magnetization.

Curie Temperature, Table 1 Moho depth, depth and temperature of magnetic layer bottom, and geothermal flow of main tectonic units in Central and Southern Europe

Tectonic unit	Moho depth (km)	Magnetic layer bottom		Geothermal flow (mW m ⁻²)
		Depth (km)	Temperature (°C)	
Variscan units (Central Europe, Corsica-Sardinia block)	29–33	29–33	540–580	60–70
Alpine units (Alps, Apennines, Molasse and Po basins)	25–50	22–28	500–600	50–80
Ligurian basin	17–24	20–24	570–650	80–100

Cross-References

- ▶ [Continental Crustal Structure](#)
- ▶ [Geomagnetic Field, Global Pattern](#)
- ▶ [Geomagnetic Field, Theory](#)
- ▶ [Heat Flow, Continental](#)
- ▶ [Magnetic Domains](#)
- ▶ [Remanent Magnetism](#)

Bibliography

- Alva-Valdivia LM, López-Loera H (2011) A review of iron oxide transformations, rock magnetism and interpretation of magnetic anomalies: El Morro mine (Brazil), a case study. *Geofis Int* 50:341–362
- Chiozzi P, Matsushima J, Okubo Y, Pasquale V, Verdoya M (2005) Curie-point depth from spectral analysis of magnetic data in Central-Southern Europe. *Phys Earth Planet Inter* 152:267–276
- Counil JL, Achache J, Galdeano A (1989) Long-wavelength magnetic anomalies in the Caribbean: Plate boundaries and allochthonous continental block. *J Geophys Res* 94:7419–7431
- Hunt CP, Moskowitz BM, Banerjee SK (1995) Magnetic properties of rocks and minerals. In: Ahrens TJ (ed) *Rock physics and phase relations: a handbook of physical constants*. American Geophysical Union, Washington, pp 189–204
- Stacey FD (1992) *Physics of the earth*, 3rd edn. Brookfield Press, Brisbane

Deep Scientific Drilling

Ulrich Harms¹ and Harold Tobin²

¹Scientific Drilling, GFZ German Research Centre for Geosciences, Potsdam, Germany

²Department of Earth and Space Sciences, University of Washington, Seattle, WA, USA

Definition

Scientific Drilling	Applying drilling methods developed in the hydrocarbon and mineral exploration industry to dig holes and retrieve samples of rock and fluid for scientific, not-for-profit purposes.
Deep Scientific Drilling	Drilling for scientific research goals beyond near-surface geological and shallow groundwater exploration to depths greater than several hundred meters subsurface, through land, sea ice, or below the sea/lake floor.

Introduction

Geoscientific insight into the deep Earth is mostly based on observations from the surface, indirect evidence through deep geophysical investigations or, increasingly, modeling. However, several critical Earth system processes and records of environmental and paleoclimate change are difficult to observe from ground level only. Hence, despite the great progress achieved in solid Earth science of the past decades, truly ground-truthed knowledge of the dynamics of Earth's crust is limited.

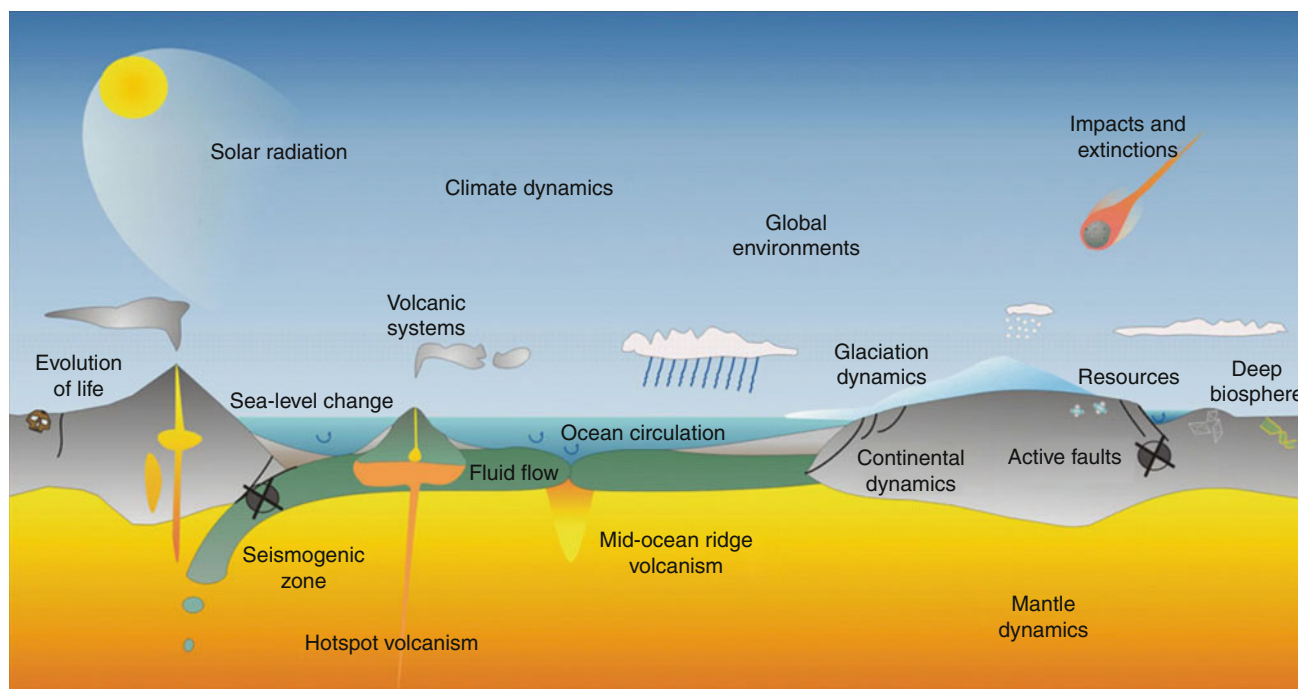
Today, direct access to the interior of the Earth's crust is still confined almost exclusively to sedimentary basins where hydrocarbon resources are recovered through deep industry-financed wells, if typically shallow exploration and mining for groundwater and minerals are neglected. Deep

drilling for mainly scientific reasons has so far remained a relatively rare exception because the very high cost of drilling is a challenging barrier in the design of research projects. Nevertheless, this threshold can be overcome if societal interests are pressing, if political and financial support is gained, and typically if international long-term cooperation is achieved. A small number of international Earth science programs devoted to scientific drilling have been established with widely agreed-upon strategic scientific goals, international co-funding, and a necessary minimum of infrastructure and operational capabilities. While the overarching research goals are almost identical, the drilling programs are separated, based on largely technological grounds, into oceanic (Integrated Ocean Discovery Program 2011) and continental (International Continental Scientific Drilling Program, Horsfield et al. 2014).

Goals of Scientific Drilling

The international programs for scientific drilling address topics of broad international interest which can contribute significantly to solving important Earth science themes of high societal relevance (Fig. 1). The research themes of these programs can be summarized in three overarching branches:

- The evolution and dynamics of Earth's crust and lithosphere, especially in subduction zones and orogenic belts, with special emphasis on the physicochemical background of geological hazards such as earthquakes, volcanic eruptions, landslides, and meteorite impacts.
- Environmental, climate, and paleoceanographic changes as recorded in geological archives of sediments in the oceans, in lakes, in ice shields and in other depositional environments, including the resulting biological consequences, as well as the exploration of the widely unknown microbial biosphere living at depth.



Deep Scientific Drilling, Fig. 1 Sketch of research themes addressed by major international scientific drilling programs

- Basic science of unconventional resources such as methane hydrates or geothermal energy, and of novel utilization of the deep underground environment such as CO₂-sequestration.

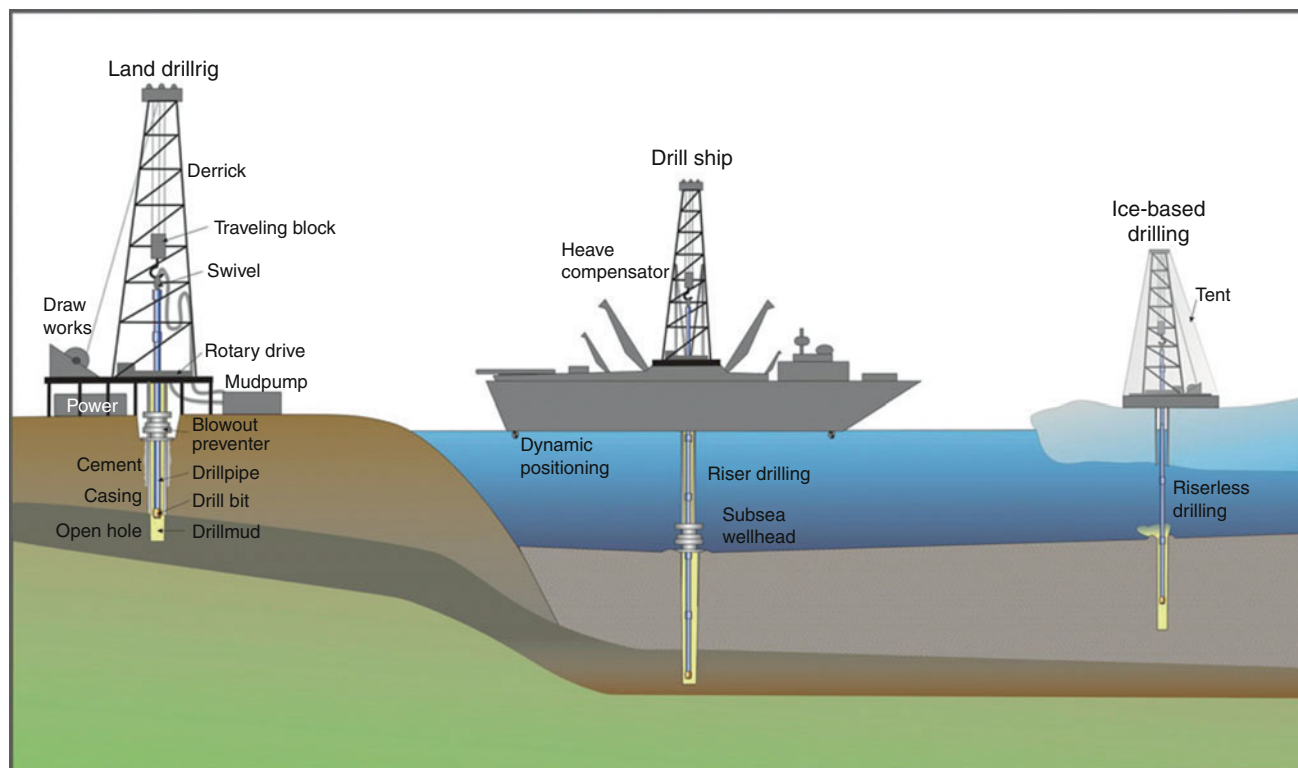
This chapter addresses mainly the first topic on lithospheric dynamics, hazard, and evolution of the Earth because most deep scientific drilling projects are dealing to a large extent with questions about internal forcing, whereas climate, environmental, and subsurface biosphere research primarily employs shallow drilling techniques.

Drilling, Sampling, and Monitoring Techniques

A standard in most deep drilling operations (e.g., hydrocarbon exploration) is the rotary drilling technique. A cutting bit is lowered into the ground on a steel tube rotated through a rotary drive in a drilling derrick. The propelled drill string consists of connected pipe elements through which a drilling fluid is pumped down the well. The drill mud, usually water with clay minerals and other added solids to provide viscosity, cools the bit and flushes out chips of the destroyed volume of rock (drill cuttings) to the surface through the annulus between the borehole wall and the drill string. The weight-on-bit is controlled through adjusted variable lifting of the string which is pulled down and kept in tension by heavy

weight sections of drill pipe installed right above the drill bit. The drilling progress is supported by extending the string in the derrick with additional sections, or stands, of pipe. When drilling from a barge, ship or floating platform, the annulus extends only to the sea/lake floor, so mud and cuttings do not return to the drill rig. In this setting, drilling can be performed with water in place of drilling mud, and cutting spilling out on sea or lake bottom around the well. However, if pressure control and mud return is required, an outer second pipe, a so-called “riser,” is put in place, so the mud and cuttings can be pumped back to the deck. Furthermore, in waves or swell, a heave compensation device is required to ensure constant weight-on-bit. Drilling through tidally lifted ice necessitates defrosting of the pipe in the ice section. Figure 2 summarizes basic drilling methods and important elements used in scientific drilling.

In normal rotary drilling technique, coring is performed with a hollow core bit that leaves a central column of rock that slides into the pipe while drilling progresses. After a few meters of coring, the whole assembly has to be pulled back out of the hole (called “pipe tripping”) to get the core to surface. In scientific drilling projects, by contrast, continuous coring is often desirable. To avoid time-consuming pipe tripping runs, wireline coring techniques are applied. An inner core barrel is lowered through the drill string and latches in above the core bit, collecting and holding the drilled-out rock column, and is finally retrieved by a winch-drawn steel cable, obviating the need for a complete pipe round trip with each



Deep Scientific Drilling, Fig. 2 Principle technical components and styles of drilling operations in different environments

core. The actual formation-cutting method varies depending on the type of rock or sediment present. Typically, thin-kerfed diamond core bits with high-rotation speed are used for hard rock drilling, roller cone abrasion bits are used for softer sedimentary rock, and nonrotating sharp edged hollow metal pistons of several meters length are hydraulically shot (forced) into soft sediments to collect cores and advance the borehole (Fig. 3).

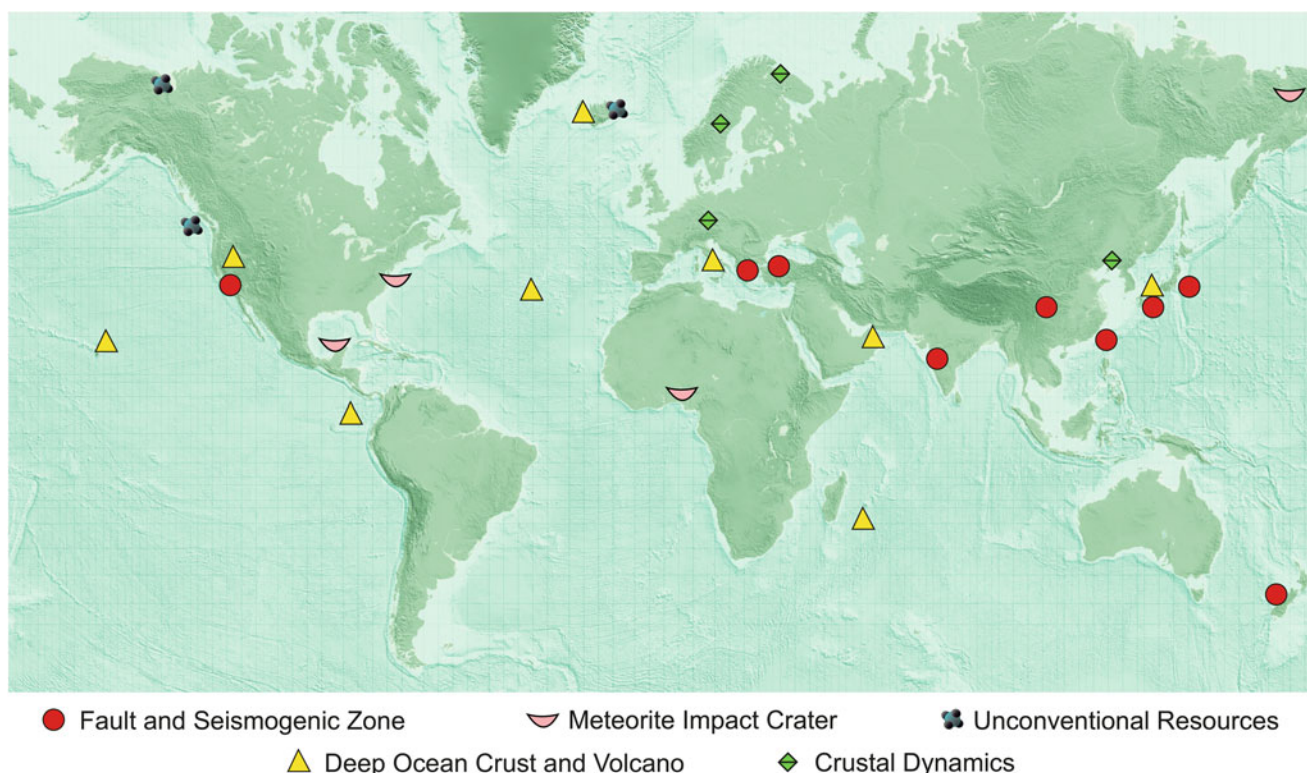
After research wells are drilled, in situ measurements and monitoring of active processes in rocks and fluids allow for unprecedented data acquisition undisturbed by surface processes. A classic example is seismometer recordings in boreholes whose resolution and noise floor is improved by orders of magnitudes in comparison to surface measurements (Abercrombie 1997). Sealed long-term borehole observatories have been installed in many places on land and at sea that are designed to record changes in temperature, pressure, strain, tilt, acceleration, and hydrologic parameters, and to sample fluids or microbiology (e.g., Fischer et al. 2008; Kastner et al. 2006; Prevedel and Kück 2006; Araki et al. 2017).

Dynamics of Continental Crust

An unprecedented and still standing depth record in ultradeep scientific drilling was achieved in 1989 in the former Soviet Union when the Kola Superdeep Borehole reached 12,261 m

after 25 years of drilling (Kozlovsky 1987; Fuchs et al. 1990). Temperatures of less than 200 °C, low tectonic stress, and special lightweight aluminum tools were key to this success. The well intersected Paleoproterozoic to Archean meta-volcano-sedimentary sequences and plutonic rocks in the Kola Peninsula of the northeastern Baltic Shield. The main scientific targets were the origin and extent of ore deposits as well as the principal structure and composition of the deeper crust. Important findings include – beside the falsification of a hypothesized basaltic layer in an onion-skin-like deep crust – the discovery of high permeability and circulating fluids throughout the crust.

The German Continental Deep Drilling Program (KTB) aimed at principal physical and chemical processes and the geological evolution of the mid-European continental crust in an amalgamated collision zone (Emmermann and Lauterjung 1997). After 4 years of drilling of high-grade metamorphic rocks, a final depth of 9101 m was reached in 1994 at formation temperatures of ~270 °C. A Paleozoic suture zone cut by the borehole was found to be strongly deformed by post-orogenic Mesozoic fault zones in an antiformal stack. Down to 9 km, faults contain saline, gas-rich free Ca-Na-Cl-fluids, the geothermal gradient increases from about 21 to 28 °C/km, and fluids percolate freely along extensive fracture pathways. A detailed correlation of geophysical data with petrophysical properties at depth allowed for the calibration of surface geophysics to in situ conditions. Furthermore, for the first



Deep Scientific Drilling, Fig. 3 World map with important scientific deep drilling sites

time, the state of stress and the rheological behavior of an ultra deep crustal profile were analyzed in-situ. Determinations of the minimal and maximum horizontal principal stress between 3 and 9 km depth yielded a uniform stress direction except for a major fault at 7 km depth (Brudy et al. 1997). In conjunction with the deep main well, a 4000-m deep pilot hole just 200 m away allowed for the installation of a depth laboratory, for example, fluid level (Schulze et al. 2000) and induced seismicity experiments (Baisch et al. 2002).

Ultrahigh pressure metamorphic (UHPM) rocks form the 1000-km-long Qinling-Dabie-Sulu Belt in China, which was formed in a Triassic continental plate collision zone. From 2002 to 2005, a 5158-m deep core hole of the Chinese Continental Scientific Drilling project served to retrieve a vertical sample profile of the protracted multiphase deep (>150 km) subduction and rapid exhumation history of UHPM rocks (e.g., Liu et al. 2007; Zhang et al. 2006). The complex metamorphic path and its timing could be determined specifically in high-pressure minerals which survived late stage fluid-enhanced retrogression and overprint encased in zircons and other refractory accessories (Fig. 3).

Volcanic Systems

Volcanism is one of the most spectacular expressions of the dynamics of tectonic processes and the gigantic scale of

recycling working in the Earth. The scientific investigation of volcanic processes is also crucial for societal security and thus an important mission for science. Since most of the history of volcanism is encapsulated in rocks that are overlain by sediments, younger lava flows or beneath the sea, drilling is in many cases a critical tool for the investigation of the different types of magmatism reaching surface.

Mid-ocean Ridge Magmatism

As defined by Conference Participants (1972), ophiolites represent a suite of oceanic crust plus upper mantle. Beneath a sediment cover lava flows and pillow basalts overlay sheeted dike complexes, gabbroic rocks and, as part of the upper mantle, serpentinized peridotites and similar ultramafic rocks. Several holes drilled by the Deep Sea Drilling Program 1969–1985 and the Ocean Drilling Program 1985–1996 confirmed this structure but at the same time provided a much more complex picture. After several meters of lava flows were encountered in the early years of these programs, deeper parts of the oceanic plutonic rock succession were drilled in composite sections in tectonic windows and contributed much to the understanding of the ocean crust and mantle. Several drillholes in the Mid-Atlantic Ridge provided insight into slow-spreading modes with emplacement of ultramafic rocks. A 1508-m-deep hole (735B) penetrated a long section of gabbroic rocks of the slow-spreading Southwest Indian Ridge off Madagascar (Dick et al. 2000).

The first lava-dike transition was drilled in 6 Ma age crust of the Nazca Plate off Ecuador (Hole 504B) down to 2111 m below seafloor (Alt et al. 1996). And the first complete section of the upper oceanic crust was recovered by the Integrated Ocean Drilling Program at Site 1256 in 2005 and 2006 when the lava-sheeted dyke (1061 m) and the sheeted dyke-gabbro (1407 m depth) transitions of the Cocos Plate off Costa Rica were penetrated. The hole was deepened to 1521.6 m in 2011. The 15 Ma crust was formed in the East Pacific Rise by superfast spreading producing a thick succession (750 m) of lava flows, a relatively thin sheeted dyke complex and an upper gabbro section of still unknown extent (Alt et al. 2007). As the drilled fractionated gabbros (1407–1521.6 m) are not the residues of the overlying volcanics, the full sequence of plutonic ocean crust and the Moho transition to the upper mantle remains to be detected in future ocean drilling (Ildefonse et al. 2007; Teagle et al. 2011).

Active Volcano Drilling and Subduction-Induced Volcanism

The most common and typically most dangerous types of volcanoes on Earth are aligned along the subduction-related arc systems. They are typically fed by highly viscous andesitic to rhyolitic magmas whose degassing behavior is critical for their eruption mode, either effusive or explosive. While erosion and tectonics provide detailed insight into solidified and cooled volcanic structures, little is known about the physical and chemical conditions or permeability close to and in feeder fissures and magma conduits of active volcanoes. Despite challenging engineering and safety issues, some attempts have been made to drill into active volcanism. Still notable because of temperatures of above 1400 °C is the shallow drilling of the Kilauea Iki lava lake on Hawaii in the 1970s (Hardee et al. 1981). Furthermore, in Eastern California, the recent, rhyolitic Inyo Lava Domes were sampled to understand degassing mechanisms (Eichelberger et al. 1986).

A series of devastating eruptions with frequently occurring pyroclastic flows and vulcanian explosions in the early 1990s at Mount Unzen Volcano in southeastern Japan was accompanied by volcano-tectonic earthquakes and isolated tremor allowing for a geophysical pinpointing of the ascent path and surface appearance of magma. The identified conduit locale was within depth reachable by drilling, and justified drilling to sample in situ conditions after eruptions ended. Within the framework of the International Continental Scientific Drilling Program, the Unzen Volcano Drilling Project hit 2004 the conduit through an almost 2000 m long but 1500 m deep well (Sakuma et al. 2008). Deviated drilling steered by down-hole motors from the northern flank of Unzen was utilized to meet environmental and safety restrictions. Large-diameter casings were cemented in place to allow mud circulation for cooling of high-temperatures which were modeled to reach up to 600 °C. However, maximum temperatures were as low as

160 °C, and no gas kicks appeared due to enhanced permeability and 9 years of convection and groundwater circulation after the last eruptions. At 1996 m depth, rocks were cored that could be identified as feeder dyke material (Nakada et al. 2005).

In the next stage of scientific drilling into volcanic features, calderas will be explored because they form through collapse during super-eruptions – the most explosive volcanism on Earth. An outstanding example of an actively inflating caldera is the Campi Flegrei volcanic field in Naples, Italy (Troise et al. 2007). A shallow crustal magma chamber, strong geothermal activity with enormous release of volcanic gases, and recent episodes of unrest call for drilling as a tool to investigate the driving mechanisms. Similarly to the Unzen drilling, a deviated well is planned to be drilled from land under the Bay of Naples towards the center of seismic unrest in the heart of the structure at 3.5 km depth.

Mantle Plume Volcanism

Hot spots cut across lithospheric oceanic and continental plates and build up giant volcanic edifices in Large Igneous Provinces (LIP) and Ocean Island Volcanoes. Their build up requires vast melt extraction from the deep mantle and provides insight into mantle compositions, transport mechanisms, and the evolution of volcanic structures over space and time. Several boreholes of the Ocean Drilling Program opened windows into the Ontong-Java Plateau of the Western Pacific (Fitton et al. 2004) and the Kerguelen Plateau in the Southern Indian Ocean (Frey et al. 2000) to confirm their LIP origin.

In the Hawaii Scientific Drilling Project, the Mauna Kea volcano was cored with 95% recovery to a total depth of 3520 m (3508 m below sea level) to sample the evolution of plume melts over about 700,000 years as the volcanoes have drifted with the Pacific Plate over the Hawaii Plume (Stolper et al. 2009). The formation temperatures increase only from about 10 to 50 °C in the borehole (19 °C/km gradient below 2000 m) due to deep circulation of seawater in a complex hydrological regime including freshwater inflow as deep as 3000 m. The circa 850 m subaerial and 2440 m submarine layered lava and hyaloclastite accumulations represent almost 400 lithological units such as individual lava flows. They show geochemical and isotopic heterogeneity with tholeiitic versus alkaline magma transitions. This bimodal distribution of source components indicates a radial zoning within the melting region with a hot inner zone which does not contain entrained mantle material and must be based in the lowermost mantle region (Bryce et al. 2005).

A further drilling target are supercritical fluids at 4–5 km depth in the Iceland Deep Drilling Project (Fridleifsson and Elders 2007). Through coupled geothermal industry research and scientific investigation of in situ fluid-magma exchange, the first well at Krafla, Northern Iceland, hit rhyolitic magma

at about 2000 m depth before the well was cased for testing (Elders et al. 2011). The Yellowstone-Snake River HOTSPOT drilling project documented the volcanic history of this mantle plume through three boreholes (1821, 1912, and 1958 m deep) that cut through basaltic and rhyolitic series (Shervais et al. 2013).

Scientific drilling into volcanic rocks of the ocean crust provided the most important test and confirmation of the seafloor-spreading hypothesis. Samples retrieved from thickened ocean crust confirmed another type of seafloor volcanism forming Large Igneous Provinces which contribute large magma volumes to the continental crust as well. These LIPs seem to represent transient plume heads while the plume tails persist in Ocean Island Volcanoes. Hawaii served as pivotal example to drill deep into a single volcano to sample different parts of the plume and the volcanic history. Furthermore, recent penetrations into active volcanoes demonstrated the technical feasibility of such operations at very high temperatures. This access to deep volcanic processes is opening novel possibilities for monitoring.

Impact Craters

Almost 180 craters on Earth are currently known that have been formed by astrophysical chance when celestial bodies collided with Earth. The effects on life and the environmental consequences of the destructive forces as well as the input of extraterrestrial material, including organic compounds, is an overarching question in Earth systems history. Together with geophysical illumination, deep drilling into impact structures provides crucial data to study cratering, informing models of the impactors' size, impact angle, the resulting energy release through melting, evaporation, ejection, and, most importantly, provides data for charting the aftermaths of such dramatic events. The best studied example to date is the Chicxulub crater (Dressler et al. 2003) in which the approximately 10-km-large impactor penetrated the whole crust and caused upper mantle rebound and an exemplary drastic biotic turnover 65.5 Ma ago (Schulte et al. 2010). The impact ejected a spherule-rich ash layer all over the whole globe and released vast amounts of carbonate and sulfate gases previously bound in sedimentary rocks at the site of impact. The resulting darkness, cooling, geochemical cycle perturbations and sudden change of environment interrupted the food chain and led to the major extinction event at that time.

Since the year 2002, four main impact craters have been studied by scientific drilling: the 200-km-wide Cretaceous-Paleogene Chicxulub Crater in Mexico was drilled on land and offshore (Hecht et al. 2004; Gulick et al. 2019), the 1 Ma old and 12 km-wide Bosumtwi Crater in Ghana (Ferrière et al. 2008), the late Eocene, 60 km Chesapeake Bay Crater in the Eastern United States (Gohn et al. 2008), and the 14 km Lake

Elgygytgyn Crater in Chukotka, Siberia (Brigham-Grette and Melles 2009). One of the important results is an improved understanding of the cratering process. Giant km-sized megablocks of target rocks were drilled in the lower sections of the huge Chesapeake and Chicxulub craters at 800–1700 m depth. In Chicxulub, uplifted basement collapsed outward and formed a peak ring that was first covered in brecciated impact melt and coarse-grained suevite rocks followed by well-sorted suevites. Resurge deposits and reflected rim-wave tsunami layers sedimented atop within about a day. This impact rock layer is about 130 m thick and shows the extremely turbulent event sequence with, e.g., charcoal from impact induced wildfires in the tsunami layer (Gulick et al. 2019).

Active Fault Zones

Since the late 1990s, researchers have undertaken a wide range of ambitious projects aimed at drilling into active faults in a range of tectonic settings, both on land and at sea. These projects have been motivated by the recognition that fault zone process are poorly understood from surface data, and that a combination of (a) samples of fault materials from present-day active systems in the subsurface and (b) access through boreholes to *in situ* measurements and monitoring of ambient conditions in active faults, can provide key new observations for learning about these processes. Our understanding of the mechanics and dynamics of plate boundary faulting is severely limited by a lack of information on ambient conditions and mechanical properties of active faults at depth. A major goal in earthquake mechanics research is thus direct *in situ* sampling and instrumentation by drilling into the seismogenic zone of active faults, including both inter- and intraplate fault systems in many tectonic settings.

Understanding of the complex physics of tectonic faulting, earthquakes, and the generation of tsunami in the Earth's crust is one of the grand challenges of the geosciences in general and seismology in particular (Lay 2009). Because earthquakes take place deep below the Earth's surface, direct observation of processes acting in the faults themselves has generally not been possible; what we know about earthquake processes is derived from seismologic observations, geodetic deformation, and studies of exhumed rocks containing products of slip at depth long ago. Researchers have learned and inferred a great deal from these studies, but they are inherently limited. Much of fault zone drilling is aimed at understanding earthquake slip, but in fact fault physics encompasses a spectrum of slip and slip rates ranging from purely aseismic behavior, such as steady creep, through events such as slow earthquakes and episodic creep, to true seismicity. It is not yet clear how similar or distinct the physics of these modes of fault slip may be (Ide et al. 2007).

Hence, some of the key questions addressed through fault zone drilling projects include the following: What are the deformation mechanisms of fault slip in both aseismic and seismic conditions? Why is slip sometimes seismic, sometimes aseismic, and sometimes in an intermediate state, and what governs the transitions among these processes? How do earthquakes nucleate, and how does rupture grow and stop? Are there precursory phenomena in the fault or near field that indicate a preparatory phase to earthquake or slip nucleation? If so, can a prediction or early warning strategy based on borehole observation be developed? What are the processes governing tsunamigenic slip? What is the role of fluid content and fluid pressure in modulating faulting processes, especially during rapid seismic slip? How does permeability of the fault zone evolve interseismically and coseismically? What is the stress tensor in and around a fault zone throughout the earthquake cycle or other slip? What controls localization of slip vs. distributed strain in faulting?

Pursuing these questions, a number of major national and international efforts have been carried out or planned in recent years. These prominently include the San Andreas Fault Observatory at Depth (SAFOD), the Taiwan Chelungpu Fault Drilling Project (TCDP), and the Gulf of Corinth drilling on land, the Nankai Trough Seismogenic Zone Experiment (NanTroSEIZE), the Japan Trench Fast Drilling Program (JFAST), and Hikurangi subduction drilling at sea.

Fault zone drilling projects are different in many respects from most other scientific drilling efforts. Rather than the complete stratigraphic intervals or sampling of broader rock volumes targeted by most other deep drilling, fault zone targets include *discontinuities* and small anomalous intervals. The targets of interest are often those locations of poorest drilling conditions, such as a highly fractured damage and gouge zone with potential for excess pore pressure above hydrostatic condition, anomalous stress state, weak rock, and other drilling hazards. The scientific approach commonly focuses on obtaining very extensive downhole measurements, logs, cores, and seismic structure in and around these fault zone environments. Measurements at the time of drilling as well as long-term observations are emphasized.

Nojima Fault Projects, Japan

Soon after the Kobe earthquake in 1995 (M6.9), several research teams drilled multiple boreholes into and across the shallow portion of the Nojima fault (the locus of that event) from ~750 m to ~1800 m depth (Ito et al. 1999). Detailed compositional and structural analyses identified a narrow fault core of gouge and surrounding damage zone, with asymmetric structure in the footwall and hanging wall (Boullier et al. 2001; Fujimoto et al. 2001; Ohtani et al. 2001; Tanaka et al. 2001). Shear wave splitting studies of aftershocks showed a rapid evolution of fast direction after the

earthquake, interpreted as evidence of fracture healing during a 12 monthly post-seismic period (Tadokoro and Ando 2002).

SAFOD, California

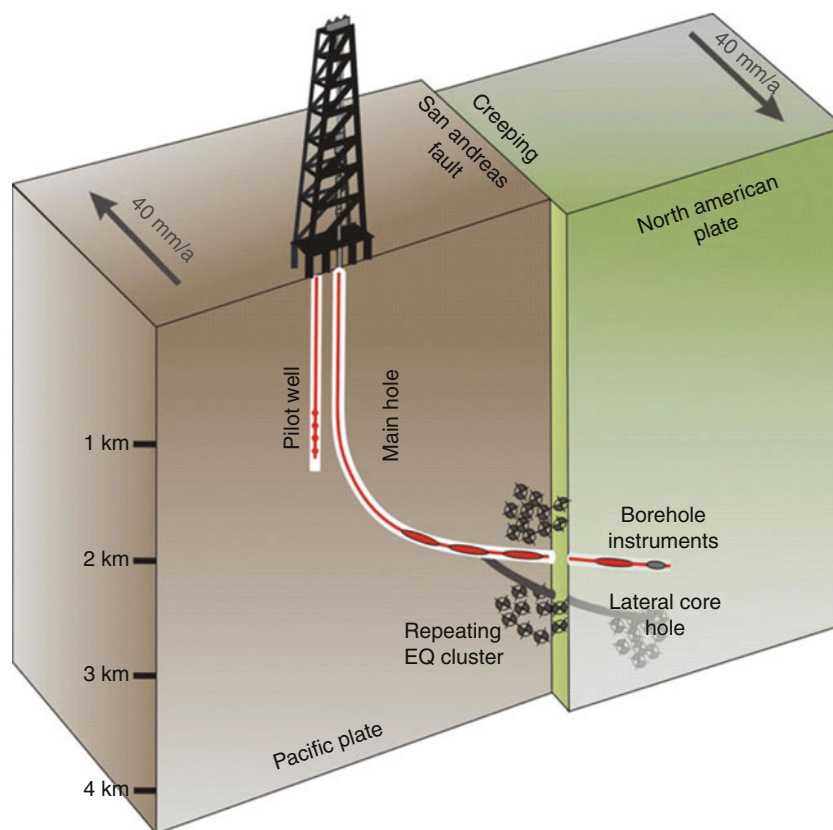
The San Andreas Fault Observatory at Depth borehole was drilled from 2004 through 2007 into the San Andreas fault in central California at the northern end of the locked Parkfield segment and the creeping section of the fault zone (Williams et al. 2010). A vertical pilot hole was drilled in 2002 to circa 2 km depth (Hickman et al. 2004), and the main hole was later drilled in several phases in a “dogleg” deviated 55° across the steeply dipping San Andreas fault zone to a vertical depth of 2.6 km (Fig. 4; Hickman et al. 2007). The initial main borehole was logged and cased; repeated casing diameter logs over time showed that two intervals of casing several meters wide were actively deforming, interpreted as a manifestation of creep in two discrete fault zones (Hickman et al. 2007). In 2007, the hole was reentered and sidetracks were drilled to obtain core samples (Williams et al. 2010). A core interval of approximately 40 m was obtained from one of the two identified major fault zones and additional cores were collected from adjacent wall rock intervals. Borehole seismometers installed at SAFOD detected earthquakes as small as M2 within a few hundred meters of the source, allowing unprecedented high-frequency detection of source processes (Ellsworth et al. 2007). A notable result from SAFOD has been the absence of evidence for either a thermal anomaly from frictional heating around the fault core or an excess pore fluid pressure substantially above hydrostatic anywhere in or around the faults, as well as geochemical evidence that the fault zone is a permeability barrier to cross-fault flow (Hickman et al. 2007; Wiersberg and Erzinger 2007), simultaneously adding evidence for absolute fault weakness in the San Andreas and opening new questions about the source of that weakness.

Taiwan Chelungpu Drilling Project, Taiwan

The devastating 1999 M7.6 Chi-Chi earthquake in Taiwan was a thrust event in a collisional tectonic setting that exhibited very large slip (>10 m), including large surface displacements, on the Chelungpu fault. The principal objective of drilling was to sample the main slip surface or zone, if one could be identified, at a relatively short post-seismic interval. The fault zone was drilled in 2004 and 2005 in an area of high-velocity large slip during the main shock, and crossed the fault zone and several hundred meters of the footwall in two adjacent holes (Ma et al. 2006). The main slip zone and several subsidiary (or older) fault cores were identified in the borehole at ~1100 m depth (Hung et al. 2007). Studies of the fault gouge particle size, thickness, and composition have yielded quantitative estimates for the work done in forming gouge and damage (fracture energy), contributing to the understanding of energy budgets during earthquakes (Ma et al. 2006).

Deep Scientific Drilling,

Fig. 4 Sketch of drilling design in the San Andreas zone observatory at depth



Wenchuan Earthquake Fault Scientific Drilling Project (WFSD), Sichuan, China

The devastating Wenchuan earthquake (M7.9) of May 12, 2008, caused at least 68,000 deaths. It occurred on a reverse fault in crystalline basement rocks, and slip propagated to the near surface. Only 178 days after the earthquake, the Wenchuan Fault drilling program was begun in order to access the co-seismic slip zone for sampling and in situ observations, in six boreholes along the Longmenshan Fault system. The boreholes elucidated a complex structure of the fault system, thermal pressurization as fault-weakening mechanism, an extremely low frictional coefficient for faulting coincident with graphite occurrence and a relationship between deep fluid appearances and seismicity (Xu and Li 2019).

Geophysical Observatory in the Eastern Sea of Marmara, GONAF

A prime example of earthquake-endangered megacities is Istanbul, Turkey, due to its close proximity to the extremely active North Anatolian Fault Zone that routes compartmentalized just a few kilometers south of the city center through the Marmara Sea. In contrast to the majority of other fault zone drilling projects the GONAF boreholes did not penetrate the fault itself but were utilized to deploy extremely sensitive

borehole seismometers and strainmeters down to 300 m together with surface stations as a permanent observatory of deformation around the Sea of Marmara (Bohnhoff et al. 2017). The acquired high-resolution data from depth have meanwhile been used to elucidate the coeval release of slow strain and enhanced local seismic moment (Martinez-Garzon et al. 2019). These accomplishments are leading to an ongoing expansion of the GONAF and related networks for monitoring of the whole Marmara region and beyond.

Deep Scientific Drilling to Study Reservoir-Triggered Earthquakes in Koyna, India

A key societal challenge in Earth science remains predicting earthquakes. There is an earthquake-prone zone in Western India where such forecasts are possible as artificial water reservoir-triggered earthquakes occur repeatedly after monsoon rains since 1962. While the region is seismically stable, rooted in Precambrian basement overlain by about 1200 m of Deccan Trap basalt flows, a secluded zone of recurrent seismicity stretches North-South of two man-made water reservoirs. Periodic earthquakes at depths between 2 and 7 km offer a unique opportunity to install an observatory to monitor mechanical and physicochemical properties of rocks and fluids in the near-field of these events (Gupta and Nayak 2011).

The 3-km-deep pilot well KFD1 drilled in 2017 transects 1247 m of basalt and 1767 m of granitic-gneissic basement rocks to shed light on deformation processes. Borehole geophysical logs analysis revealed anomalous zones with stress-induced shear wave anisotropy indicating the orientation of maximum horizontal compressive stress related to the active subsurface fault damage zone (Goswami et al. 2019, 2020). For the next phase, it is planned to drill a main well towards the proper fault zone down to about 6 km depth to install a borehole laboratory at depth for direct and permanent monitoring of an intraplate seismic zone.

Subduction Zone Fault Drilling at Sea

For well over 30 years, drilling efforts have targeted active large offset thrusts and decollement faults in subduction zone settings, but these have mostly been in areas shallower than the depth of seismogenic strain accumulation and release. Beginning in the 1980s with the Barbados Accretionary Wedge at the Atlantic Caribbean plate boundary, drilling was aimed at understanding the faulting conditions, ambient pore fluid pressure, and mechanics in a very low angle (~2 degree dip) décollement thrust fault at the base of an accretionary wedge (Moore 1989). Drilling, downhole logging, and long-term borehole observatory measurements there demonstrated that the fault zone was a mechanically very weak interval with high permeability, high excess pore fluid pressure, and patchily distributed areas of anomalously high porosity (Moore et al. 1998; Fisher 2005). Drilling at Costa Rica off the Nicoya peninsula also elucidated the structure of a shallow decollement zone. Structural studies and geochemical tracers of hydrological flow paths suggest that two strongly decoupled systems exist above and below the ~20-m-thick main fault zone, which acts simultaneously as a low-permeability barrier to cross fault flow and a high-permeability channel for along-fault flow (Tobin et al. 2001). Installation of a sealed borehole monitoring system there documented an apparently weakly overpressured décollement zone that exhibits temporal pressure excursions tied to apparent aseismic transients recorded geodetically on land (Davis and Villinger 2006). Similarly, instrumented boreholes at the Nankai Trough off Shikoku Island in Japan show pressure transients associated temporally and spatially with swarms of very low frequency earthquakes, suggesting a long-distance (tens of km) poroelastic connection between the earthquake events and the boreholes (Davis et al. 2006). The same data exhibit long-term pressure trends interpreted as a response to gradual interseismic strain accumulation.

Drilling to understand megathrust earthquake processes and fault zone properties has been a major focus of both the Integrated Ocean Drilling Program and the International Ocean Discovery Program (IODP), and was a primary justification for the construction of the riser drilling vessel *Chikyu*. Major fault drilling programs have been undertaken by IODP

at the Nankai Trough, Sumatra, Japan Trench, Costa Rica, and Hikurangi (New Zealand) subduction zones. The spate of extremely large earthquakes that began with the Sumatra 2004 M9.2 and continued with the 2010 Maule M8.8, the Tohoku 2011 M9.1, and others has led to renewed urgency to investigate these events. Scientific ocean drilling has contributed to a new view of how fault locking, slip, and friction work. IODP Expedition 362 drilling at Sumatra showed that diagenesis of well-cemented sediments led to unexpected conditions of frictional instability to the trench for the 2004 event (Hüpers et al. 2017), facilitating large slip. Nankai drilling on Expedition 316 in 2007 revealed evidence of frictional heating of fault zone rocks during past slip, indicative of rupture reaching shallower levels of the megathrust than was thought possible theretofore (Sakaguchi et al. 2011).

Beginning in 2007, IODP embarked on its most ambitious fault zone drilling project to date, to drill a transect of boreholes at the Nankai Trough, including one ultradeep site targeting the plate boundary fault at 5 km below the sea floor. The Nankai Trough Seismogenic Zone Experiment (NanTroSEIZE) project represents the first attempt to drill, sample, and instrument the seismogenic portion of a plate-boundary fault or megathrust within a subduction zone, where the 1944 M8.1 Tonankai earthquake slip zone is located, and where more recently discovered very low frequency earthquakes, seismic tremor, and transient slow slip have all been detected (e.g., Sugioka et al. 2012). The fundamental goal of NanTroSEIZE science plan (Tobin and Kinoshita 2006) has been to access, sample, and then instrument the faults in aseismic and seismogenic regions of the megathrust system. Over the period from 2007 through 2019, 12 *Chikyu* expeditions have targeted this zone, reaching a maximum depth of 3262 m below the sea floor, a record in scientific ocean drilling. Key results have included the delineation of the fault zones, and the establishment of the first three real-time cabled borehole observatories in a subduction zone system which have detected many transient slip events (Araki et al. 2017; Tobin et al. 2019). The long-range objective of drilling into and instrumenting the plate interface fault zone at 5–7 km depth remains as a future objective.

Japan Trench Fast Drilling Project (JFAST)

The unprecedented discovery of very large (40–80 m) fault slip to the trench region during the 2011 Tohoku M9.1 earthquake and tsunami led to “rapid-response” drilling being planned and executed within 13 months of the earthquake. JFAST was carried out as IODP Expeditions 343 and 343 T (Chester et al. 2013). The primary objective was to identify and sample the fault zone that slipped near its upper end at the trench, and measure the residual thermal anomaly from the frictional heating of fault slip. A challenging aspect of the drilling was the extreme water depth of nearly 7000 m at the sea floor. A borehole of 844 m was drilled, the slipped fault

material recovered, and its frictional properties documented (Ujiie et al. 2013). Subsequently, a thermistor string was sealed into the borehole for several months, and detected the frictional heat signature of the earthquake, measured as 0.3 °C above the background. From these data, Fulton et al. (2013) concluded that slip had occurred under extremely low effective friction, with a coefficient of only 0.08. These are the first data to quantify fault slip conditions for large earthquakes.

Unconventional Resources: Geoenergy

Among the unexplored geological resources requiring drilling for basic scientific research are gas hydrates, in which methane and other gaseous hydrocarbons combine with water to form crystalline, ice-like structures (clathrates). Two widespread environments of their occurrence have been known for a long time: in sediments of the continental slopes where they simulate bottom reflectors in seismic images, and below permafrost in polar regions (Kvenvolden 1988).

Little is known about the global abundance and distribution of gas hydrates, and even less about the physical and chemical parameters that affect the formation and stability of gas hydrates. They bear a high potential of generating continental slope instabilities causing giant submarine landslides and contributing significantly to the atmospheric greenhouse effect when released into atmosphere (Dickens 2003). At sea, the best studied, drilled areas are in the Eastern Pacific off Vancouver Island and at Blake Ridge in the western Atlantic Ocean off Georgia (Tréhu et al. 2006). Continental gas hydrate accumulations were drilled in the Mackenzie River delta in northern Canada (Dallimore and Collett 2005). Two observation wells and one central production well drilled to 1200 m depth were used for an extensive field program including coring, downhole logging, cross-hole experiments, and production testing. All three wells show massive thermogenic methane gas hydrate accumulations between 890 and 1100 m depth with the highest concentration at the base of the gas hydrate stability zone (Takahashi et al. 2005). Similarly, in marine sediments, the hydrates are concentrated in coarse-grained, high-permeability layers.

The Iceland Deep Drilling Project, IDDP

Global decarbonization plans require enhanced regenerative energy production. Accordingly one of the emerging research objectives in Earth science is geothermal energy. Much effort is invested in identifying (and tapping) hot water and steam and heat exchangers at depth or creating such systems artificially. Although such system can be tapped for up to some tens of years, the utilization of high enthalpy supercritical

magmatic fluids or the exploitation of magmatic heat and accompanying crystallization energy promises ten times higher energy output. In an attempt to test the theory of supercritical fluids (22 MPa and 374 °C for pure water) utilization, the IDDP project drilled in 2009 at Krafla in Iceland, a well in basaltic rocks that was intruded at 2100 m depth by rhyolitic magma (Elders et al. 2011). Drilling was stopped due to challenging 900 °C conditions but the well was tested for directly utilizing magmatic heat and inspired planning for a magma test site (Eichelberger 2019).

The second IDDP well drilled on the Reykjanes Peninsula in SW Iceland to about 4500 m depth reaching about 535 °C undisturbed formation temperature (436 °C measured borehole temperature) and 34 MPa fluid pressure. The drill cores retrieved from different depths confirm increasing hydrothermal alteration of basaltic dike rocks up to pyroxene hornfels facies (Fridleifsson et al. 2018). Although cold-water injection was seismically monitored and fluid inflow zones could be characterized, a test confirming energy gain from supercritical magmatic fluids remains pending.

Summary

This chapter summarizes the most significant, but not all, deep scientific drilling projects that have been carried out in recent years by international research teams. Key goals of these projects encompass topics of high societal relevance such as a better understanding of the physics driving geological disasters and a wise utilization of resources. Various several kilometer deep drilling ventures allowed for unprecedented sampling and in situ monitoring of dynamic processes governing formation and restructuring of the Earth's crust such as collision events, volcanism, faulting, or impact cratering. Despite grand engineering challenges, deep Earth access techniques are now available for novel burgeoning scientific exploration. Scientific drilling promises to be an integral tool of research generating new observations for the time to come.

Cross-References

- [Earthquakes and Crustal Deformation](#)
- [Geophysical Well Logging](#)
- [Impact Craters on Earth](#)
- [KTB Depth Laboratory: A Window into the Upper Crust](#)
- [Lithosphere, Continental](#)
- [Lithosphere, Oceanic](#)
- [Mantle Plumes](#)
- [Seismic Instrumentation](#)
- [Slow Earthquake](#)
- [Vertical Seismic Profiling](#)

Bibliography

- Abercrombie RE (1997) Near surface attenuation and site effects from comparison of surface and deep borehole recordings. *Bull Seismol Soc Am* 87:731–744
- Alt JC, Laverne C, Vanko DA, Tartarotti P, Teagle DAH, Bach W, Zuleger E, Erzinger J, Honnorez J, Pezard PA, Becker K, Salisbury MH, Wilkens RH (1996) Hydrothermal alteration of a section of upper oceanic crust in the eastern equatorial Pacific: a synthesis of results from Site 504 (DSDP Legs 69, 70, and 83, and ODP Legs 111, 137, 140, and 148). In: Alt JC, Kinoshita H, Stokking LB, Michael P (eds) *Proceedings of the ocean drilling program: scientific results*, vol 148, pp 417–434
- Alt JC, Teagle DAH, Umino S, Miyashita S, Banerjee NR, Wilson DS, IODP Expeditions 309 and 312 Scientists (2007) IODP Expeditions 309 and 312 drill an intact section of upper oceanic basement into gabbros. *Sci Drill* 6:4–10. Copernicus Publications, Göttingen, Germany. <https://doi.org/10.2204/iodp.sd.4.01.2007>
- Araki E, Saffer D, Kopf A, Wallace L, Kimura T, Machida Y, Ide S, Davis E, IODP Expedition 365 Shipboard Scientists (2017) Recurring and triggered slow-slip events near the trench at the Nankai Trough subduction megathrust. *Science* 356(6343):1,157–1,160. <https://doi.org/10.1126/science.aan3120>
- Baisch S, Bohnhoff M, Ceranna L, Tu Y, Harjes H-P (2002) Probing the crust to 9 km depth: fluid injection experiments and induced seismicity at the KTB superdeep drilling hole, Germany. *Bull Seismol Soc Am* 92:2369–2380
- Bohnhoff M, Dresen GH, Ceken U, Kadarioglu FT, Kartal RF, Kilic T, Nurlu M, Yanik K, Acarel D, Bulut F, Ito H, Johnson W, Malin PE, Mencin D (2017) GONAF – the borehole geophysical observatory at the North Anatolian Fault in the eastern sea of Marmara. *Sci Drill* 22:19–28. <https://doi.org/10.5194/sd-22-19-2017>
- Boullier A-M, Ohtani T, Fujimoto K, Ito H, Dubois M (2001) Fluid inclusions in pseudotachylytes from the Nojima fault, Japan. *J Geophys Res* 106:21965–21977
- Brigham-Grette J, Melles M (2009) Complex drilling logistics for lake El'gygytyn, NE Russia. *Sci Drill* 7:38–39. <https://doi.org/10.2204/iodp.sd.7.05.2009>
- Brudy M, Zoback MD, Fuchs K, Rummel F, Baumgärtner J (1997) Estimation of the complete stress tensor to 8 km depth in the KTB scientific drill holes: implications for crustal strength. *JGR* 102(B8):18,453–18,475
- Bryce JG, DePaolo DJ, Lassiter JC (2005) Geochemical structure of the Hawaiian plume: Sr, Nd, and Os isotopes in the 2.8 km HSDP-2 section of Mauna Kea volcano. *Geochim Geophys Geosyst* 6: Q09G18. <https://doi.org/10.1029/2004GC000809>
- Chester FM, Rowe C, Ujiie K, Kirkpatrick J, Regalla C, Remitti F, Moore JC, Toy V, Wolfson-Schwehr M, Bose S, others (2013) Structure and composition of the plate-boundary slip zone for the 2011 Tōhoku-oki earthquake. *Science* 342(6163):1,208–1,211. <https://doi.org/10.1126/science.1243719>
- Conference Participants (1972) Penrose field conference on ophiolites. *Geotimes* 17:24–25
- Dallimore SR, Collett TS (2005) Summary and implications of the Mallik 2002 gas hydrate production research well program. In: Dallimore SR, Collett TS (eds) *Scientific results from the mallik 2002 gas hydrate production research well program, mackenzie delta, Northwest Territories, Canada*. Geological Survey of Canada, Vancouver, Bulletin, 585
- Davis EE, Villinger HW (2006) Transient formation fluid pressures and temperatures in the Costa Rica forearc prism and subducting oceanic basement: CORK monitoring at ODP Sites 1253 and 1255. *Earth Planet Sci Lett* 245(1–2):232–244. <https://doi.org/10.1016/j.epsl.2006.02.042>
- Davis EE, Becker K, Wang K, Obara K, Ito Y, Kinoshita M (2006) A discrete episode of seismic and aseismic deformation of the Nankai trough subduction zone accretionary prism and incoming Philippine Sea plate. *Earth Planet Sci Lett* 242:73–84
- Dick HJB, Natland JH, Alt JC, Bach W, Bideau D, Gee JS, Haggas S, Hertogen JGH, Hirth G, Holm PM, Ildefonse B, Iturrino GJ, John BE, Kelley DS, Kikawa E, Kingdon A, LeRoux PJ, Maeda J, Meyer PS, Miller JD, Naslund HR, Niu Y, Robinson PT, Snow J, Stephen RA, Trimby PW, Worm HU, Yoshinobu A (2000) A long in-situ section of the lower ocean crust: results of ODP Leg 176 drilling at the Southwest Indian Ridge. *Earth Planet Sci Lett* 179:31–51
- Dickens GR (2003) A methane trigger from global warming? *Science* 299:1017–1020
- Dressler BO, Sharpton VL, Morgan J, Buffler R, Moran D, Smit J, Stöffler D, Urrutia J (2003) Investigating a 65-Ma-old smoking gun: deep drilling of the Chicxulub impact structure. *Eos* 84(14):125–130. <https://doi.org/10.1029/2003EO140001>
- Eichelberger J (2019) Planning an international magma observatory. *Eos* 100. <https://doi.org/10.1029/2019EO125255>. Published on 25 June 2019
- Eichelberger JC, Carrigan CR, Westrich HR, Price RH (1986) Non-explosive silicic volcanism. *Nature* 323:598–602
- Elders WA, Friðleifsson GÓ, Zierenberg RA, Pope EC, Mortensen AK, Guðmundsson Á, Lowenstern JB, Marks NE, Owens L, Bird DK, Reed M, Olsen NJ, Schiffman P (2011) Origin of a rhyolite that intruded a geothermal well while drilling in a basaltic volcano, at Krafla, Iceland. *Geology* 39:231–234
- Ellsworth W, Malin PE, Imanishi K, Roecker SW, Nadeau R, Oye V, Thurber CH, Waldhauser F, Boness NN, Hickman SH, Zoback MD (2007) Seismology inside the fault zone: applications to fault-zone properties and rupture dynamics. *Sci Drill* (Special Issue 1):85–88
- Emmermann R, Lauterjung J (1997) The German continental deep drilling program KTB: overview and major results. *J Geophys Res* 102:18179–18201. <https://doi.org/10.1029/96JB03945>
- Ferrière L, Koeberl C, Ivanov BA, Reimold WU (2008) Shock metamorphism of Bosumtwi impact crater rocks, shock attenuation, and uplift formation. *Science* 12(322):1678–1681. <https://doi.org/10.1126/science.1166283>
- Fischer AT, Davis EE, Becker K (2008) Borehole-to-borehole hydrologic response across 2.4 km in the upper oceanic crust: implications for crustal scale properties. *J Geophys Res* 113:B07106. <https://doi.org/10.1029/2007JB005447>
- Fisher A (2005) Marine hydrogeology: recent accomplishments and future opportunities. *Hydrogeol J* 13:69–97. <https://doi.org/10.1007/s10040-004-0400-y>
- Fitton JG, Mahoney JJ, Wallace PJ, Saunders AD (2004) Leg 192 synthesis: origin and evolution of the Ontong Java Plateau. In: Fitton JG, Mahoney JJ, Wallace PJ, Saunders AD (eds) *Proceeding of ODP, scientific results*, p 192. <https://doi.org/10.2973/odp.proc.sr.192.101.2004>
- Frey FA, Coffin MF, Wallace PJ, Weis D, Zhao X, Wise SE, Wähnert V, Teagle DAH, Saccoccia PJ, Reusch DN, Pringle MS, Nicolaysen KE, Neal CR, Müller RD, Moore CL, Mahoney JJ, Keszthelyi L, Inokuchi H, Duncan RA, Delius H, Damuth JE, Damasceno D, Coxall HK, Borre MK, Boehm F, Barling J, Arndt N, Antretter M (2000) Origin and evolution of a submarine large igneous province: the Kerguelen Plateau and Broken Ridge, Southern Indian Ocean. *Earth Planet Sci Lett* 176:73–89
- Fridleifsson GO, Elders WE (2007) Progress report on the Iceland Deep Drilling Project (IDDP). *Sci Drill* 4:26–29. <https://doi.org/10.2204/iodp.sd.4.04.2007>
- Fridleifsson GO, Elders WA, Zierenberg RA, Fowler AP, Weisenberger TB, Mesfin KG, Sigurðsson Ó, Nielsson S, Einarsson G, Óskarsson F, Guðnason EÁ, Tulinius H, Hokstad K, Benoit G, Nono F, Loggia D, Parat F, Cichy SB, Escobedo D, Mainprice D (2018 online) The Iceland Deep Drilling Project at Reykjanes: drilling into the root zone of a black smoker analog. *J Volcanol Geotherm Res*. <https://doi.org/10.1016/j.jvolgeores.2018.08.013>

- Fuchs K, Kozlovsky YA, Krivtsov AI, Zoback MD (1990) Super-deep continental drilling and deep geophysical sounding. Springer, Berlin
- Fujimoto K, Tanaka H, Higuchi T, Tomida N, Ohtani T, Ito H (2001) Alteration and mass transfer inferred from the Hirabayashi GSJ drill penetrating the Nojima fault, Japan. *Island Arc* 10:401–410
- Fulton PM, Brodsky EE, Kano Y, Mori J, Chester F, Ishikawa T, Harris RN, Lin W, Eguchi N, Toczko S, Expedition 343, 343T, and KR13-08Scientists (2013) Low coseismic friction on the Tōhoku-oki fault determined from temperature measurements. *Science* 342(6163):1,214–1,217. <https://doi.org/10.1126/science.1243641>
- Gohn GS, Koeberl C, Miller KG, Reimold WU, Browning JV, Cockell CS, Horton JW, Kenkmann T, Kulpecz AA, Powars DS, Sanford WE, Voytek MA (2008) Deep drilling into the Chesapeake Bay impact structure. *Science* 27(320):1740–1745. <https://doi.org/10.1126/science.1158708>
- Goswami D, Roy S, Akkiraju VV (2019) Delineation of damage zones from 3 km downhole geophysical logs in the Koyna Seismogenic Zone, western India. *J Geophys Res Solid Earth* 124:6101–6120. <https://doi.org/10.1029/2018JB017257>
- Goswami D, Hazarika P, Roy S (2020) In Situ Stress orientation from 3 km Borehole image logs in the Koyna Seismogenic Zone, Western India: implications for transitional faulting environment. *Tectonics* 39:e2019TC005647. <https://doi.org/10.1029/2019TC005647>
- Gulick SPS, Bralower TJ, Ormö J et al (2019) The first day of the Cenozoic. *Proc Natl Acad Sci USA* 116(39):19342–19351. <https://doi.org/10.1073/pnas.1909479116>
- Gupta H, Nayak S (2011) Deep scientific drilling to study reservoir-triggered earthquakes in Koyna, Western India. *Sci Drill* 12. <https://doi.org/10.2204/iodp.sd.12.07.2011>
- Hardee HC, Dunn JC, Hills R, Ward RW (1981) Probing the melt zone of Kilauea Iki lava lake, Kilauea volcano, Hawaii. *Geophys Res Lett* 8:1211–1214
- Harms U, Koeberl C, Zoback MD (2007) Continental scientific drilling. A decade of progress, and challenges for the future. Springer, Heidelberg, Germany
- Hecht L, Wittmann A, Schmitt RT, Stöffler D (2004) Composition of impact melt particles and the effects of post-impact alteration in suevitic rocks at the Yaxcopoil-1 drill core, Chicxulub crater, Mexico. *Meteorit Planet Sci* 39:1169–1186
- Hickman S, Zoback MD, Ellsworth WE (2004) Introduction to special section: preparing for the San Andreas fault observatory at depth. *Geophys Res Lett* 31:L12S01. <https://doi.org/10.1029/2004GL020688>
- Hickman S, Zoback M, Ellsworth W, Boness N, Malin P, Roecker S, Thurber C (2007) Structure and properties of the San Andreas Fault in central California: recent results from the SAFOD experiment. *Sci Drill* (Special Issue 1):29–32. <https://doi.org/10.2204/iodp.sd.s01.39.2007>
- Horsfield B, Knebel C, Ludden J, Hyndman R (eds) (2014) Unravelling the workings of planet earth. International Continental Scientific Drilling Program, Potsdam. <https://doi.org/10.2312/icdp.2014.001>
- Hung J, Ma K, Wang C, Song S, Ito H, Lin W, Yeh E (2007) Subsurface structure, fault zone characteristics and stress state in scientific Drill Holes of Taiwan Chelungpu Fault Drilling Project. *Sci Drill* (Special Issue 1):55–58
- Hüpers A, Torres ME, Owari S, McNeill LC, Dugan B, Henstock TJ, Milliken KL, Petronotis KE, Backman J, Bourlange S, others (2017) Release of mineral-bound water prior to subduction tied to shallow seismogenesis off Sumatra. *Science* 356(6340):841–844. <https://doi.org/10.1126/science.aal3429>
- Ide S, Beroza GC, Shelly DR, Uchide T (2007) A scaling law for slow earthquakes. *Nature* 447:76–79
- Ildefonse B, Christie DM, Mission Moho Workshop Steering Committee (2007) Mission Moho workshop: drilling through the oceanic crust to the mantle. *Sci Drill* 4:11–17. <https://doi.org/10.2204/iodp.sd.4.02.2007>
- International Ocean Discovery Program (2011) International ocean discovery program science plan 2013–2023, illuminating earth's past, present, and future, 92 pp. <http://www.iodp.org>
- Ito H, Fujimoto K, Tanaka H, Lockner D (1999) Proceedings of the international workshop on the Nojima fault core and borehole data analysis. Geological Survey of Japan Interim Report, No. EQ/00/1, USGS Open-File 895 Report 000-129
- Kastner M, Becker K, Davis EE, Fischer AT, Jannasch HW, Solomon EA, Wheat G (2006) New insights into the hydrogeology of the oceanic crust through long-term monitoring. *Oceanography* 19(4):46–57
- Kozlovsky YA (1987) The superdeep well of the Kola Peninsula. Springer, Heidelberg, Germany
- Kvenvolden KA (1988) Methane hydrate – a major reservoir of carbon in the shallow geosphere? *Chem Geol* 71:41–51
- Lay T (ed) (2009) Seismological grand challenges in understanding earth's dynamic systems. Report to the National Science Foundation, IRIS Consortium, 76 pp
- Liu FL, Xu ZQ, Liou JG, Dong HL, Xue HM (2007) Ultrahigh-pressure mineral assemblages in zircons from the surface to 5158 m depth in cores of the main drill hole, Chinese Continental Scientific Drilling project, SW Sulu belt, China. *Int Geol Rev* 49:454–478
- Ma K-F, Tanaka H, Song S-R, Wang C-Y, Hung J-H, Tsai Y-B, Mori J, Song Y-F, Yeh E-C, Soh W, Sone H, Kuo L-W, Wu H-Y (2006) Slip zone and energetics of a large earthquake from the Taiwan Chelungpu-fault Drilling Project. *Nature* 444:473–476. <https://doi.org/10.1038/nature05253>
- Martinez Garzon P, Bohnhoff M, Mencin D, Kwiatak G, Dresen G, Hodgkinson K, Nurlu M, Kadiroglu FT, Kartal RF (2019) Slow strain release along the eastern Marmara region offshore Istanbul in conjunction with enhanced local seismic moment release. *Earth Planet Sci Lett* 510:209–218. <https://doi.org/10.1016/j.epsl.2019.01.001>
- Moore JC (1989) Tectonics and hydrogeology of accretionary prisms: role of the decollement zone. *J Struct Geol* 11:95–106
- Moore JC, Klaus A, Bangs NL, Bekins B, Buecker CJ, Brueckmann W, Erickson SN, Hansen O, Horton T, Ireland P, Major CO, Moore GF, Peacock S, Saito S, Sreaton EJ, Shimeld JW, Stauffer PH, Taymaz T, Teas PA, Tokunaga T (1998) Consolidation patterns during initiation and evolution of a plate-boundary décollement zone: Northern Barbados accretionary prism. *Geology* 26:811–814. [https://doi.org/10.1130/0091-7613\(1998\)026<0811:CPDIAE>2.3.CO;2](https://doi.org/10.1130/0091-7613(1998)026<0811:CPDIAE>2.3.CO;2)
- Nakada S, Uto K, Sakuma S, Eichelberger JC, Shimizu H (2005) Scientific results of conduit drilling in the Unzen Scientific Drilling Project (USDP). *Sci Drill* 1:18–22
- Ohtani T, Tanaka H, Fujimoto K, Higuchi T, Tomida N, Ito H (2001) Internal structure of the Nojima fault zone from the Hirabayashi GSJ drill core. *Island Arc* 10:392–400
- Prevedel B, Kück J (2006) New developments in long-term downhole monitoring arrays. *Sci Drill* 3:34–37. <https://doi.org/10.2204/iodp.sd.2.08.2006>
- Sakaguchi A, Chester F, Fabbri O, Goldsby D, Li C, Kimura G, Tsutsumi A, Ujiie K, Yamaguchi A, Curewitz D (2011) Paleothermal condition of the shallow mega-splay fault based on vitrinite reflectance: core analysis of IODP NanTroSEIZE stage 1, Eos Trans. AGU 90(52), Fall Meet. Suppl, Abstract T12A-08
- Sakuma S, Kajiwara T, Nakada S, Uto K, Shimizu H (2008) Drilling and logging results of USDP-4 – Penetration into the volcanic conduit of Unzen Volcano, Japan. *J Volcanol Geotherm Res* 175:1–12. <https://doi.org/10.1016/j.jvolgeores.2008.03.039>
- Schulte P, Alegret AI, Arz JA, Barton PJ, Bown PR, Bralower TJ, Christeson GL, Claeys P, Cockell CS, Collins GS, Deutsch A, Goldin TJ, Goto K, Grajales-Nishimura JM, Grieve RAF, Gulick SPS, Johnson KR, Kiessling W, Koeberl C, Kring DA, MacLeod KG, Matsui T, Melosh J, Montanari A, Morgan JV, Neal CR, Nichols DJ, Norris RD, Pierazzo E, Ravizza G, Rebolledo-Vieyra M, Reimold WU, Robin E, Salge T, Speijer RP, Sweet AR, Urrutia-Fucugauchi J, Vajda V, Whalen MT, Willumsen PS (2010) The

- Chicxulub Asteroid impact and mass extinction at the cretaceous-paleogene boundary. *Science* 327:1214–1218. <https://doi.org/10.1126/science.1177265>
- Schulze K, Kumpel HJ, Huenges E (2000) In-situ petrohydraulic parameters from tidal and barometric analysis of fluid level variations in deep wells: some results from KTB. In: Stober I, Bucher K (eds) *Hydrogeology of crystalline rocks*. Kluwer, Dordrecht, The Netherlands, pp 79–104
- Shervais JW, Schmitt DR, Nielson D, Evans JP, Christiansen EH, Morgan L, Pat Shanks WC, Prokopenko AA, Lachmar T, Liberty LM, Blackwell DD, Glen JM, Champion LD, Potter KE, Kessler JA (2013) First results from HOTSPOT: The Snake River Plain Scientific Drilling Project, Idaho, U.S.A. *Sci Drill* 15:36–45. <https://doi.org/10.2204/iodp.sd.15.06.2013>
- Stolper EM, DePaolo DJ, Thomas DM (2009) Deep drilling into a mantle Plume Volcano: the Hawaii scientific drilling project. *Sci Drill* 7:4–14. <https://doi.org/10.2204/iodp.sd.7.02.2009>
- Sugioka H, Okamoto T, Nakamura T, Ishihara Y, Ito A, Obana K, Kinoshita M, Nakahigashi K, Shinohara M, Fukao Y (2012) Tsunamigenic potential of the shallow subduction plate boundary inferred from slow seismic slip. *Nat Geosci* 5:414–418. <https://doi.org/10.1038/NGEO1466>
- Tadokoro K, Ando M (2002) Evidence for rapid fault healing derived from temporal changes in S wave splitting. *Geophys Res Lett* 29(4):1047. <https://doi.org/10.1029/2001GL013644>
- Takahashi H, Fercho ED, Allimore SR (2005) Drilling and operations overview of the Mallik 2002 production research well program. In: Allimore SR, Collett TS (eds) *Scientific results from the Mallik 2002 gas hydrate production research well program*, Mackenzie Delta, Northwest Territories, Canada. Geological Survey of Canada, Vancouver, Bulletin
- Tanaka H, Fujimoto K, Ohtani T, Ito H (2001) Structural and chemical characterization of shear zones in the freshly activated Nojima fault, Awaji Island, southwest Japan. *J Geophys Res* 106:8789–8810
- Teagle DAH, Ildefonse B, Blum P, Expedition 335 Scientists (2011) Proceedings of the integrated ocean drilling program, p 335. <https://doi.org/10.2204/iodp.pr.335.2011>
- Tobin H, Kinoshita M (2006) NanTroSEIZE: the IODP Nankai Trough seismogenic zone experiment. *Sci Drill* 2:23–27. <https://doi.org/10.2204/iodp.sd.2.06.2006>
- Tobin H, Vannuchi P, Meschede M (2001) Structure, inferred mechanical properties, and implications for fluid transport in the décollement zone, Costa Rica convergent margin. *Geology* 29:907–910. [https://doi.org/10.1130/0091-7613\(2001\)029<0907:SIMPAT>2.0.CO;2](https://doi.org/10.1130/0091-7613(2001)029<0907:SIMPAT>2.0.CO;2)
- Tobin HJ, Kimura G, Kodaira S (2019) Processes governing giant subduction earthquakes: IODP drilling to sample and instrument subduction zone megathrusts. *Oceanography* 32(1):80–93. <https://doi.org/10.5670/oceanog.2019.125>
- Tréhu AM, Ruppel C, Holland M, Dickens GR, Torres ME, Collett TS, Goldberg D, Riedel M, Schultheiss P (2006) Gas Hydrates in Marine Sediments: lessons from scientific ocean drilling. *Oceanography* 19(4):124–142
- Troise C, De Natale G, Pingue F, Obrizzo F, De Martino P, Tammaro U, Boschi E (2007) Renewed ground uplift at Campi Flegrei caldera (Italy): new insight on magmatic processes and forecast. *Geophys Res Lett* 34:L03301. <https://doi.org/10.1029/2006GL028545>
- Ujiie K, Tanaka H, Saito T, Tsutsumi A, Mori JJ, Kameda J, Brodsky EE, Chester FM, Eguchi N, Toezko S, Expedition 343 and 343T Scientists (2013) Low coseismic shear stress on the Tōhoku megathrust determined from laboratory experiments. *Science* 342(6163):1,211–1,214. <https://doi.org/10.1126/science.1243485>
- Wiersberg T, Erzinger J (2007) Hydrogen anomalies at seismogenic depths of the San Andreas Fault, Abstracts of the 17th annual V. M. Goldschmidt conference: *Geochimica et Cosmochimica Acta*, 71,15S, A1110
- Williams ML, Fischer KM, Freymueller JT, Tikoff B, Tréhu AM, others (2010) Unlocking the secrets of the North American continent: an earthscope science plan for 2010–2020, EarthScope.org, 78 pp
- Xu Z, Li H (2019) The Wenchuan earthquake fault scientific drilling (WFSD) project. In: Li YG (ed) *Earthquake and disaster risk: decade retrospective of the Wenchuan earthquake*. Springer, Singapore
- Zhang ZM, Xiao YL, Xu ZQ, Hoefs J, Yang JS, Liu FL, Liou JG, Simons K (2006) UHP metamorphic rocks from the Chinese continental scientific drilling project: I. petrology and geochemistry of the main hole (0–2050 m). *Contrib Mineral Petrol* 152:421–441

Deep Seismic Reflection and Refraction Profiling

Kabir Roy Chowdhury

Department of Earth Sciences, Utrecht University,
Utrecht, The Netherlands

Synonyms

Active source seismology; Controlled source seismology; Deep seismic sounding; Explosion seismology; Wide-angle reflection/refraction profiling

Definition

Deep Seismic Reflection and Refraction Profiling

Classically, multichannel recording, along a measurement line, of (mostly) seismic P-waves, artificially generated using large energy sources, after these have traveled deep thru the earth's crust (and upper mantle). Later developments include multicomponent recording enabling analysis of shear waves. Deep reflection profiling is mostly done using vibrators (on land) or air guns (in water) at near-vertical distances (8–12 km) to image the structure of the crust and upper mantle. Wide-angle reflection/refraction profiling uses large explosions and recording distances (200–300 km), primarily to obtain velocity information down to the upper mantle.

Notational Notes

Below, all capitals (e.g., DSRRP) will be used for acronyms and italicized phrases within double quotes (e.g., “*Seismic Noise*”) will refer to articles elsewhere in *this* volume.

Introduction

Vibrations caused by earthquakes are known to travel great distances and provide valuable information about the internal structure of the earth. Explosions and other artificial sources could also be used to generate such waves. Man-made

earthquake signals have several advantages of their own, e.g., their location – both in space and time – is precisely known. Further, due to relatively shorter distances of observation, higher frequencies are better preserved and allow a better resolution of the subsurface structure. Development of *marine seismics* has extended the coverage of this technique to offshore regions too.

As a science, *deep seismics* is positioned between exploration seismics and seismological studies. With the shared theory of wave propagation in (visco)elastic media, similar instrumentation, and computer-intensive processing/interpretation techniques, there is increasing *overlap* between these two approaches. Improvements in data acquisition and processing have extended the depth range of exploration seismics from basin level downwards and now cover the entire continental crust down to *Moho* and the upper mantle region immediately below it (50–100 km). Such studies are called deep seismic sounding, long-range reflection-refraction, wide-angle crustal-seismics, etc. depending upon their focus. The acronym *DSRRP*, derived from the title, will denote studies using seismics to study the (continental) crust, including active and passive continental margins, and the upper mantle of the earth, using man-made and natural sources.

There are several good resources about the general theory of wave propagation in elastic media (e.g., Aki and Richards 2002) and exploration seismics (e.g., Sheriff and Geldart 1995 and ► “Seismic Data Acquisition and Processing”); special topics pertaining to the latter in the context of DSRRP will be elaborated whenever needed.

Basic Principles

Figure 1 illustrates the principle of the seismic experiment using a simplified earth model. Here, a horizontal boundary separates the top layer of velocity (actually wave speed) v_1 and thickness d from a half-space of velocity v_2 , with $v_1 < v_2$, mostly valid within the earth. Wave energy, generated at the source, travels through the media and is recorded by many (usually hundreds of) receivers on the surface as *seismic traces*, their collection being a *seismic record*. Plot of a record – amplitudes as a function of recording distance versus travel time – shows distinct spatiotemporal *alignments*, representing different wave types (*phases*), three of which are sketched in the figure. Depending upon their travel path, the phases show up in the record as *first* or later arrivals and carry information about some properties of the medium. The amplitudes of these phases – and their variation with the recording distance (*offset*) – yield additional information about material properties of the medium, e.g., *density*, *elasticity*, etc.

The subject matter of this article forms an extension of exploration seismics. Sheriff and Geldart (1995) and Liner

(2004) provide good coverage of its various aspects, e.g., theory and data acquisition, processing, and interpretation. More detailed treatment of these topics is provided by Aki and Richards (2002; theory of wave propagation), Vermeer (2002; data acquisition), and Yilmaz (2001; data processing). DSRRP-specific modifications/extensions to the standard field and processing procedures will be described later. Based on Fig. 1, a brief overview of some key concepts from *seismics* – relevant to DSRRP – follows.

Linear arrivals represent the travel times of both *direct* and *refracted* arrivals, if v_1 and v_2 are constant, the slopes (dt/dx) being inversely proportional to the corresponding velocities. Using these, and the *intercept time* (t_{int} , Fig. 1) of the refracted arrival, the layer thickness d can be calculated.

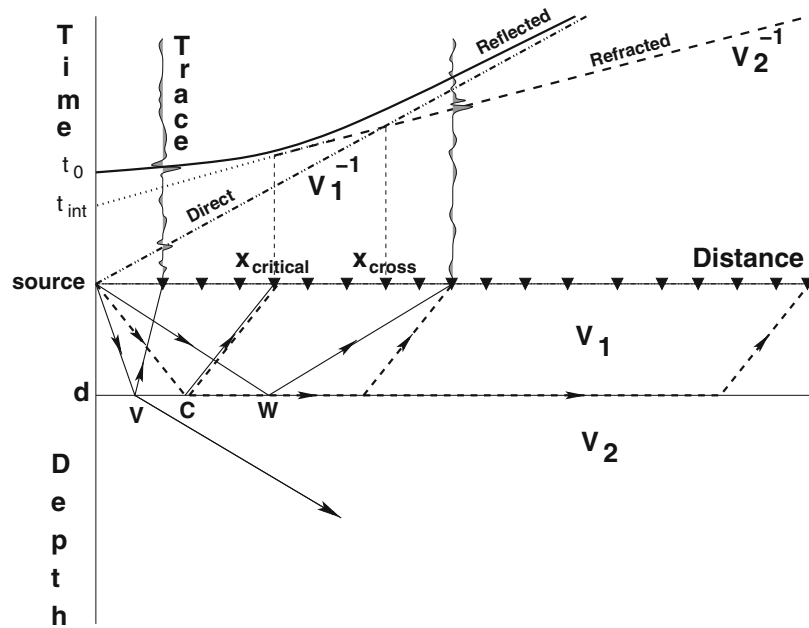
Critical distance is the minimum recording distance for observing the refracted wave. According to Snell’s law, only when the energy is incident on the boundary at the critical angle ($= \sin^{-1} \frac{v_1}{v_2}$) will it be refracted *along* the boundary, will propagate with the velocity v_2 of the lower medium, and will keep sending energy back into the upper medium according to Huygens’ principle.

Crossover distance denotes the distance, beyond which the faster refracted phase overtakes the direct wave ($v_2 > v_1$) and becomes the first arrival; the latter can also be used to determine the thickness d . In a layered earth with velocities increasing with depth, refraction arrivals from deeper boundaries become the first arrival after their respective crossover distances. *Historical note:* A. Mohorovicic used the overtaking of the direct P-phase from an earthquake by a faster (refracted) one, which he called P_n , to infer the thickness of the earth’s crust and the presence of a higher-velocity medium underneath (IGCP-559 2010a); the boundary defining the base of the crust has since been named *MOHO* in his honor.

Reflections represent that part of wave energy incident upon a layer boundary, say at the point “V” in Fig. 1, which will be transmitted back into the upper medium, according to Snell’s law. The *reflected* arrival is hyperbolic in shape, with its apex at time t_0 , corresponding to the vertical two-way time ($= 2d/v_1$) from the source down to the boundary and back. Reflection amplitude is sensitive to the (contrast of) material properties across the boundary and the incidence angle and provides valuable additional information regarding the media.

Supercritical incidence occurs, as the recording distance increases, with the incidence angle changing in Fig. 1 from (near) vertical (e.g., point “V”) through the critical incidence (e.g., point “C”) to supercritical (e.g., point “W”). Note that the reflection hyperbola becomes asymptotic to the linear arrival belonging to the layer above but never crosses it. This relationship also holds in a multilayered situation with increasing velocities.

Wide-angle incidence implying large recording distances (compared to the target depth) sees a sharp increase of



Deep Seismic Reflection and Refraction Profiling, Fig. 1 Schematics of a multichannel seismic experiment to image the earth (layer of wave speed V_1 overlying a half-space of wave speed V_2). Energy travels from the source through the medium and is recorded by the receivers (filled triangles), planted in the ground, as *seismic traces*. Dash-dots represent energy *directly* traveling from source to the

reflection coefficient and enables imaging of crustal boundaries with small contrasts. The quantity $t_x - t_0$ (Fig. 1), called normal moveout (NMO), can be measured well in this region and helps velocity estimation. The seismic signal window gets compressed though at large offsets due to the geometry (kinematics) of the travel time curves (Fig. 1, top).

Modern DSRRP represents an integration of the *near-vertical* and *wide-angle* modes of seismic investigation. While near-vertical multichannel reflection seismic (NMRS) provides powerful tools for structural imaging developed for the exploration industry, wide-angle reflection-refraction (WARR) is valuable for constraining the velocities of the crustal units and thus provides the geological framework.

Data Acquisition

Seismic field measurements need three components, viz., man-made *source* to produce seismic energy (usually at the earth surface or within water), which then propagates within the earth, *receivers* (usually also at the surface) to pick up a part of this energy returning back after reflection/refraction/diffraction (scattering in general) from within, and *recorder* to store the received signals (nowadays after digitization); see ► “[Seismic Data Acquisition and Processing](#)” for details.

Sources used in DSRRP have to be stronger than those in exploration, as the distances traveled by the seismic

receivers along the surface, continuous lines show *reflections* from the layer boundary, and dashes indicate energy *critically refracted* along the boundary. The upper half uses the same patterns to show the *travel times* of these three phases. Letters *V*, *C*, and *W* indicate points representing (near) vertical, critical, and wide-angle incidence of energy on the boundary

waves – both vertically and especially horizontally – are much larger. Such sources including quarry blasts using several kilotons of TNT – sometimes under water – were recorded over long distances, often across international boundaries (Steinhart and Meyer 1961; György 1972; Kaila et al. 1978). Calibrating and monitoring of nuclear explosions necessitated several long-distance (mostly) refraction profiles, with nuclear explosions as sources (see ► “[Traveltime Tomography Using Controlled-Source Seismic Data](#)”). Later, very long-range crustal seismic measurements were carried out in the Soviet Union using dedicated *peaceful nuclear explosions* (Pavlenkova and Pavlenkova 2006). Although chemical explosions continue to be used for WARR studies, powerful hydraulically driven mechanical vibrators on land and large compressed air sources (air guns) under water are preferred nowadays for NMRS; these have sufficient energy to return signals back to the surface from Moho and upper mantle.

Receivers used in DSRRP are similar to those used in seismology and in exploration seismics but need some special characteristics. Long measurement profiles imply deployments over a large distance/area, often necessitating stand-alone capabilities, especially for power supply and storage capacity; see Mereu and Kovach (1970), in which mention is also made of an early German effort (MARS-66) using a portable four-channel analog tape recorder. Modern DSRRP receivers digitize in real time and can often be

programmed to record at certain pre-arranged time windows (see [Texan](#) in the references). For near-vertical land/marine deployments, standard industry equipment is used; these are frequently cable-free and use radio communication for data transfer.

Time-of-shot recording was the weakest link in DSRRP – especially for WARR deployment, covering distances of several hundred kilometers. Early experiments used short-wave radio link or the parallel recording of a continually broadcasting radio transmitter. Availability of GPS-based time signals and high-quality internal clocks has mitigated this problem.

Field Layouts

Receiver arrays were utilized early on for long-range refraction-seismic studies for their directivity vis-a-vis possible nuclear test sites – some becoming permanent installations, e.g., NORSAR (Norway), GBA (India), and WKA (Australia). Recently, the Program for Array Seismic Studies of the Continental Lithosphere (see [PASSCAL](#) in the references) has made hundreds of identical instruments available, enabling temporary deployment in several countries. Using these, with fixed and portable deployments recording both active and passive sources, a 15-year observation has been completed in the USA. See Fig. 2 and [USARRAY](#) and [EarthScope](#) in the references.

Multichannel near-vertical land/marine studies use standard industry equipment and deployments (see also ► “[Seismic Data Acquisition and Processing](#)”), the main difference being a much longer recording time (15–30 s or longer). When using vibratory sources on land, long sweep and listening times are needed too. If the continental crust is covered by a shallow water layer (e.g., British Isles), marine data acquisition offers twin advantages of speed and better signal-to-noise ratio, enabling better imaging.

DSS-DSRRP onshore-offshore recorded earlier from near-vertical outward till at least the p_n -phase (refraction below Moho) was recorded as the first arrival (Gamburtsev 1952). Outside the Soviet Union, this technique was first extensively used in India starting 1972, to study the tectonic framework of the subcontinent (Reddy et al. 1999) and was later adapted for sub-basalt (*Deccan Trap*) exploration. Many DSRRP investigations nowadays combine the two modes – an industry standard *common midpoint* profile (NMRS) is frequently interspersed with a few widely spaced explosions, observed at large distances (WARR). Variations include two-ship marine recording and simultaneous onshore-offshore measurements, useful for investigating continental/accretionary margins.

Special Processing and Interpretation Tools

Processing of DSRRP data followed initially the standard industry scheme. Soon, however, the special needs of deep seismic data were realized. Consequently, schemes using modules newly developed by the industry/academia were established, and older data were frequently reprocessed, often resulting in *improved* images and interpretation (Cook and Vasudevan 2006).

Line drawings were used initially to prepare DSRRP data for migration and interpretation. NMRS signals are generally weak and laterally discontinuous and embedded in a noisy background. Lines were therefore drawn on the paper section to form more or less continuous alignments – taken to represent boundaries for interpretation. The process was strongly subjective. This approach paralleled interpreting WARR data with long refraction segments, which neglected signals with limited lateral continuity.

Coherency filter was proposed to mitigate this. Kong et al. (1985) formulated a procedure to *automatically* identify *signals* present in the seismic section. Using the concept of *phase coherency*, it uses a few user-defined parameters, to yield *repeatable* results, and can detect *weak but laterally continuous* signals in the presence of incoherent noise – although it does not preserve (relative) amplitudes. Some form of coherency filtering is nowadays commonly used for processing NMRS data.

Noise reduction versus amplitude preservation is, however, an important issue because modern processing techniques can use amplitude information to advantage. An example of amplitude-preserving noise reduction is provided by Kumar et al. (2011), wherein curvelets are used to suppress incoherent seismic noise.

Statistical processing of DSRRP data, for objective high-grading using a coherency criterion, was proposed by Hansen et al. (1988), based upon statistical hypothesis testing. It provided some estimate of the robustness of the results, albeit at the cost of additional computation time. Vasudevan and Cook (1998) introduced the concept of *skeletonization* to delineate regions of the deep crust based upon their seismic signature; van der Baan (2000) included local signal statistics for high grading the signals.

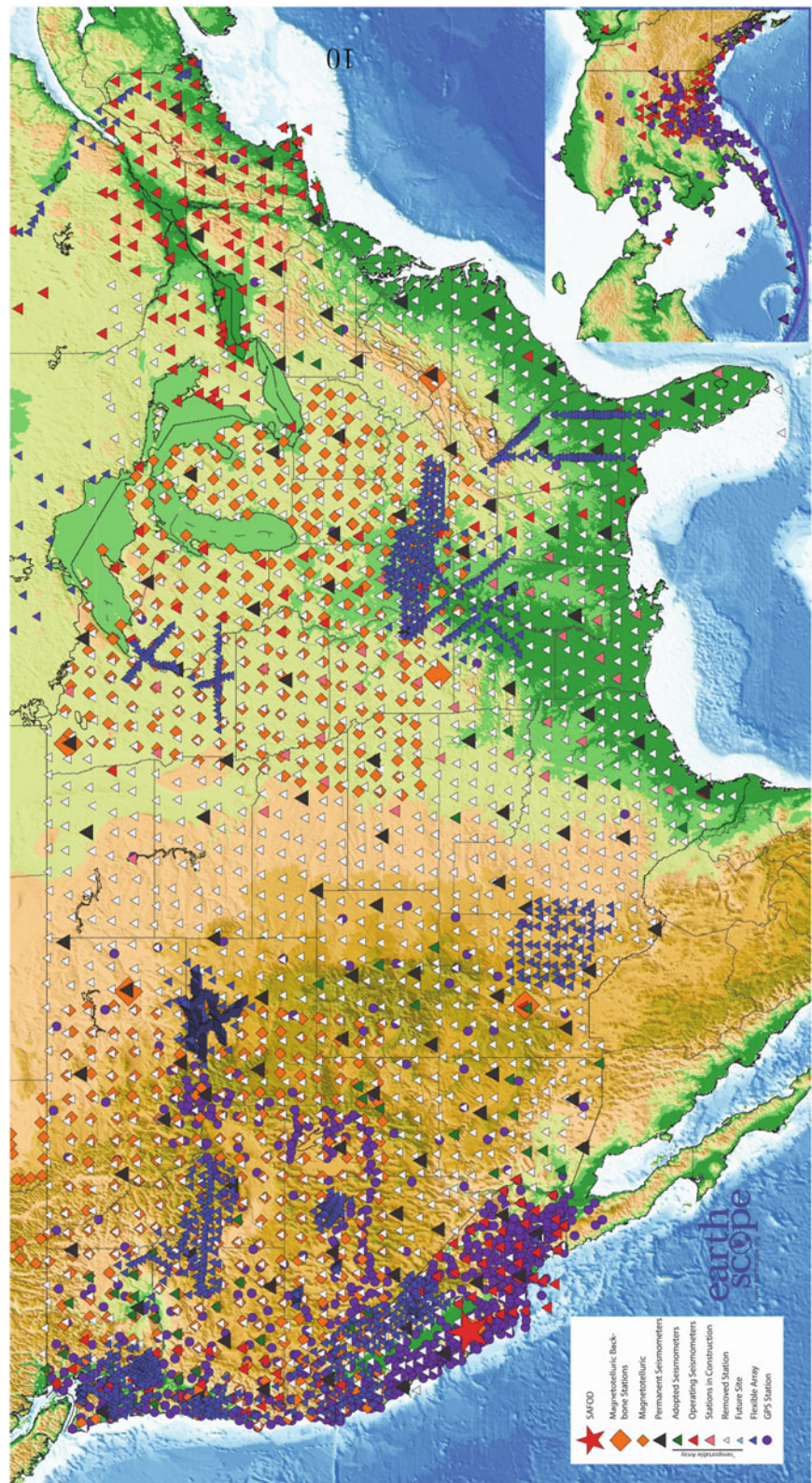
Attempts have also been made to treat the entire deep-reflection wave field as backscattering from a random medium (Hurich 1996; Hurich and Kocurko 2000; Pullammanappallil et al. 1997) and analyze it to extract parameters describing the medium (see below).

Vertical vis-a-vis horizontal tectonics used preferentially could lead to different interpretations of the same data. DSS profiles in György (1972) and later literature (e.g., Kaila et al. 1979) frequently contain intra-crustal normal faults. With

Deep Seismic Reflection and Refraction Profiling,

Fig. 2 Plot of EarthScope stations – permanent and temporary – including those of USARRAY. For details, see legend in the inset. Full resolution original at <http://www.earthscope.org> → Research → Maps → EarthScope Overall Maps (Archive) → June 2015. (Figure courtesy of www.earthscope.org)

EarthScope Stations Status as of June 2015



the paradigm shift associated with the formulation of plate tectonics, some of these may need to be revisited. Gibbs (1986) illustrated this by using a DSS section sans the earlier interpretive lines for an alternate interpretation.

Main Results

NMRS recordings from 12 to 15 s contain coherent reflected energy from down to 35–45 km, depending upon the presence/absence of sedimentary cover. Crystalline crust appears to be much more reflective than assumed earlier, although there are transparent zones too. In general, the mature continental middle crust is less reflective than the lower crust – probably indicating differences in their rheology (brittle vs. ductile). Intra-crustal dipping reflective zones – of both thrust and normal fault types – are encountered frequently. At times, these cut through the Moho into the upper mantle as *frozen* evidence of paleo-subduction. Moho appears frequently as the termination of *diffuse* reflectivity, the boundary itself occasionally showing sharp displacement (sign of strength).

WARR recordings can be modeled with longer boundaries separating tectonic/velocity blocks at a regional scale. These also provide occasional evidence of *sharp* offsets in crustal boundaries including Moho; the latter seems, in some cases, to be *doubled*.

The role of deeper structures – especially *deep faults* – in controlling the evolution of shallower geology, e.g., deposition, deformation, fluid movement, etc., is being increasingly appreciated. The knowledge gained is of economic significance – for understanding *systems* associated with economic accumulation of hydrocarbons and minerals and to help steer search for them.

Early DSS(RP) in Eastern Europe has been nicely summarized in György (1972), from which Fig. 3 is taken.

It shows a part of a long-offset (≈ 233 km) DSS profile recorded in Ukraine in the 1960s, as part of extensive DSS surveys in Eastern Europe, which had established the observability of deep refracted and reflected phases in the wide-angle range. High apparent velocity of the refracted P_n -phase (first arrival at large offsets) and its relationship with (later) reflected phase help identify the base of the crust as their origin. Some figures in the above reference also show near-vertical reflections from the Moho.

György (1972, pp. 44–68) includes on page 66 a contour map of the Moho depth below Ukraine, the detail of which, although debatable, is impressive, especially considering its vintage. It is based on an astounding 6000 line km of DSS profiling with a dense network, following the methodology described in Gamburtsev (1952). The contours indicate depths between 22.5 and 55 km, with rapid lateral variations at places. The latter were interpreted as *deep faults* displacing the Moho.

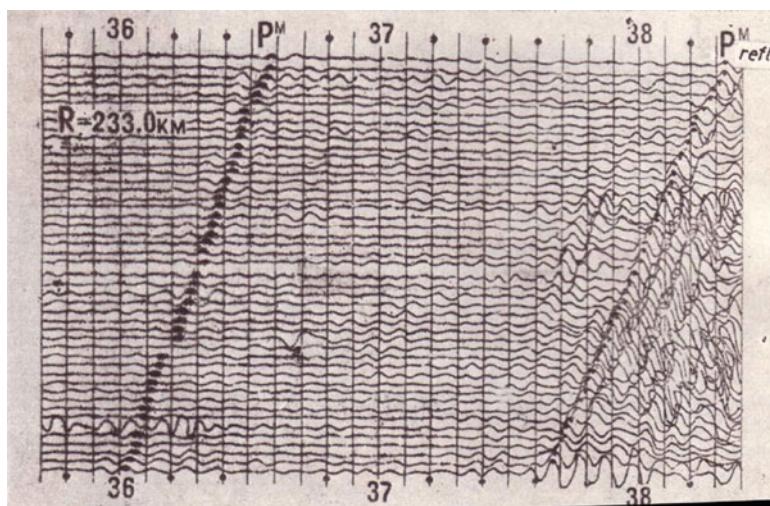
(Inter)national consortia in DSRRP started in 1975, when an industry crew measured a reflection profile in Texas, using a vibratory source, with a recording time of 15 s (Finlayson 2010b). A large amount of data has since then been collected/analyzed in several national academic industrial collaborations – Table 1 provides a partial overview; see also IGCP-559 (2010c) for more details.

Recently, some international DSRRP experiments have studied specific geological problems, e.g., Himalaya and Tibetan Plateau (Zhao et al. 1993), active continental margin in Central Andes (ANCORP Working Group 2003), and structures in East Antarctica from SEAL geotranssect (Kanao et al. 2011). Such transects are often multidisciplinary in character, e.g., Palomeras et al. (2011).

International biennial symposia have been organized (roughly) every 2 years since 1984 to showcase the data and discuss the results from DSRRP surveys. The publications resulting from these meetings provide a historical record of

Deep Seismic Reflection and Refraction Profiling,

Fig. 3 DSRRP data recorded in Ukraine in the 1960s. Both refracted and reflected energy is clearly visible at a distance of about 233 km and is correlatable trace to trace with a geophone spacing of 100 m. The first arrival, here marked P^M , is the head wave from Moho (P_n), and the strong later phase, marked P^M_{refl} , is the wide-angle reflection from the base of the crust. (Figure courtesy of *Geophysical Transactions* from György (1972, p. 50))



the progress of DSRRP – with respect to both technological advancement and scientific knowledge: Barazangi and Brown (1986a, b), Matthews and Smith (1987), Leven et al. (1990), Meissner et al. (1991), Clowes and Green (1994), White et al. (1996), Klemperer and Mooney (1998a, b), Carbonell et al. (2000), Thybo (2002), Davey and Jones (2004), Snyder et al. (2006), Heikkinen et al. (2011), Rawlinson and Goleby (2012), Santosh et al. (2014), Carbonell et al. (2016), Rawlinson et al. (2017), and Malinowski et al. (2019). Some additional information is also available in IGCP-559 (2010b).

(Re)processing, synthesis, and interpretation of the vast amount of DSRRP data – near-vertical and wide-angle – are not easy. The data quality depends upon the geological settings and the data acquisition technique and parameters used. Uniform processing of this dataset of mixed quality/vintage is necessary though for regional syntheses and interpretation, to understand the internal architecture of the continental crust (e.g., Phinney and Roy Chowdhury 1989; Cook and Vasudevan 2006).

Results from DSRRP contain several surprising features – some of which are still being interpreted – and have yielded new insights into the processes that shape the continental crust. Below are some of these highlights; the acronyms referring to the consortia/projects are explained in Table 1.

Imaging deeper with multichannel exploration seismics began in 1975. The COCORP consortium – the acronym is reported to have been coined past midnight at a bar in Mexico – pioneered the use of industry standard sources (vibrators), recording layout (NMRS), and processing for deep seismic profiling on land and obtained useful signals from depths of 40–50 km, by using 4–5 vibrators simultaneously and extending the recording time (Oliver et al. 1976). Later, similar studies confirmed that the crust underlying the basement possesses variable reflectivity, including some *transparent* regions. Moho, the base of the crust, often showed up on such images as the termination of a zone of diffuse reflectivity, and not as a long and sharp boundary inferred from earlier (refraction) studies.

One of the early surprises of the COCORP lines was the discovery of a mid-crustal zone of strong reflectivity below southern Georgia, USA; the *surrency bright spot* was reconfirmed during a later survey (Pratt et al. 1993). Such zones have since been reported in other surveys too, e.g., the *Quebrada Blanca Bright Spot* (QBBS) in Andean subduction zone (ANCORP Working Group 2003) and below the Dnieper-Donets paleorift (Pylypenko et al. 2011). The possible causes of this strong reflectivity remain controversial (see below).

Besides WARR recordings of quarry blasts, Germany had an early start in near-vertical recordings of deep reflections and statistical evaluation of their amplitudes (e.g., Dohr 1957; Dohr and Fuchs 1967). More recently, their DEKORP program has included investigations across active collisional

Deep Seismic Reflection and Refraction Profiling, Table 1 Some (inter)national DSRRP efforts

Acronym	Location	Period	Remarks
COCORP	USA	1975 onwards	Pioneered near-vertical imaging
BIRPS	UK	1981–1998	Marine seismic imaging
DEKORP	Germany	1983–1997	Deep drilling (KTB)
LITHOPROBE	Canada	1984–2003	Multidisciplinary
ECORS	France	1983–1991	IFREMER (marine)
MONA LISA	UK	1993	On/Off-shore
IBERSEIS	Spain	2001	Spanish universities and institutes
SEAL	<i>E. antarctica</i>	2002–2004	Japan
...			
INDEPTH	China	1992 onwards	USA and other countries
ANCORP	Chile	1996	Germany and other countries
KRISP	Kenya	1985–1994	European and US universities
BEST	Russia	2002	Denmark and Poland
BABEL	Baltic Sea	1989	European groups, On/Off-shore
EAGLE	Ethiopia	2003	European and US universities
SINOPROBE	China	2008–2012	Multidisciplinary
...			

zones, e.g., Alps (Gebrande et al. 2006) and the Andes (ANCORP Working Group 2003). For details, see DEKORP in the references.

Marine seismic imaging of the continental crust was seized upon to (partly) alleviate the unfavorable *signal-to-noise ratio* (SNR) for deep seismic data acquisition on land, e.g., noise from traffic, industry, etc. – although marine seismics has its noise sources too. Phinney (1986) used data from a 48-channel marine survey (recorded by USGS during 1973–1979) over the Long Island Platform. The original stack sections of lines 36 and 23, reprocessed to 12 s, showed clear evidence of a rich crustal architecture, with half grabens, wedges, and other tectonic features indicating both compressional and extensional phases of a Wilson cycle.

Existence of known – and expected – hydrocarbon-bearing structures had attracted marine seismic exploration activity in the 1970s and 1980s to the waters around the British Isles. The latter, surrounded by North Sea and the northeast margin of the Atlantic Ocean, are a part of the European continental shelf. Starting 1981, this situation was utilized to great advantage by the BIRPS consortium – essentially by extending the marine seismic exploration recording time to 15 s. The

very first profile, MOIST (recorded by the preceding BURPS group), contained strong reflections from the lower crust, Moho, and upper mantle (see Fig. 4), which could be connected to surface geology on land (Smythe et al. 1982), and even prompted correlation of tectonic evolution across the Atlantic (Brewer and Smythe 1984).

Later BIRPS profiles, e.g., WINCH (Brewer et al. 1983), provided evidence for shallower/younger tectonics being controlled by older/deeper crustal structures. The DRUM profile extended the recording to 30 s. The density of coverage and the quality of data allowed Flack and Warner (1990) to map deep reflections in 3D and enabled Chadwick and Pharaoh (1998, p. 268) to produce a contour map of Moho beneath the UK, which may be compared with a similar map below Ukraine (György 1972, p. 66) mentioned earlier.

Integrated transects, using additional geophysical tools, e.g., magnetotellurics, electromagnetics, and geochronology, characterized the LITHOPROBE program in Canada. It took advantage of the geology to investigate both ancient processes, e.g., assembly of continents and modern crustal dynamics of active subduction, detachment, and imbrication. For example, the Kapuskasing Uplift, one of the few lower crustal exposures on earth surface, was imaged in the KSZ transect (Percival et al. 1989), whereas the transects SNORCLE and SC examined younger mountain building processes (e.g., Clowes et al. 1983).

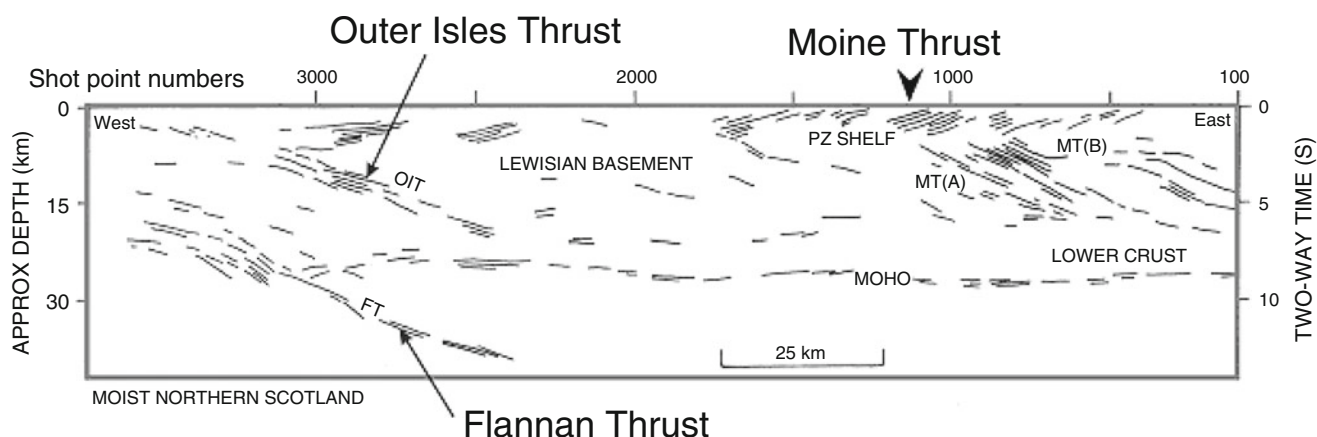
Relating to the ground truth, the strength of exploration seismics, is the Achilles' heel of DSRRP (and seismology). Superdeep drill holes provide the only opportunities of directly correlating observations to the rock properties; the German DEKORP program was able to utilize this in a symbiotic manner.

DEKORP played an important role in the site selection phase of the German superdeep drilling program, KTB, which was set up to drill $\approx 10,000$ m down through an ancient geodynamic suture – see Emmermann and Lauterjung (1997) for an overview of KTB. Later, besides providing a reference point for the seismics at the drill site, the KTB program has yielded direct evidence of shear zones, anisotropy, and substantial amounts of fluids deep in the crust; the analysis of the latter has provided new insights into their origin and role in controlling/influencing geodynamic processes.

The Russian super-deep drilling program in the Kola Peninsula (Smythe et al. 1994) reached the record depth of 12,200 m and also provided valuable *ground truth* regarding the macrostructure of the mature continental crust and the origin of crustal reflectivity.

Upper-mantle reflectivity-frozen subduction has also undergone a paradigm shift as a result of DSRRP investigations. Reflections from upper mantle were already reported by György (1972). DSRRP has provided vivid proof that reflective zones may extend into the upper mantle and has helped understand their structure. Dipping reflectivity in this region has provided information regarding paleotectonics and may – in some cases (see Fig. 4) – be evidence for paleo-subduction (see also BABEL Working Group 1990; Morgan et al. 1994).

Rheology of crust, Moho, and upper mantle depends upon their composition and in situ physical conditions. Intra-crustal faulting, affecting even the Moho, was already inferred by György (1972) and Kanasewich et al. (1969). DSS data from the Indian shield (Kaila et al. 1979) was used by Roy Chowdhury and Hargraves (1981) to infer constraints about the thermo-tectonic evolution of the early earth. With

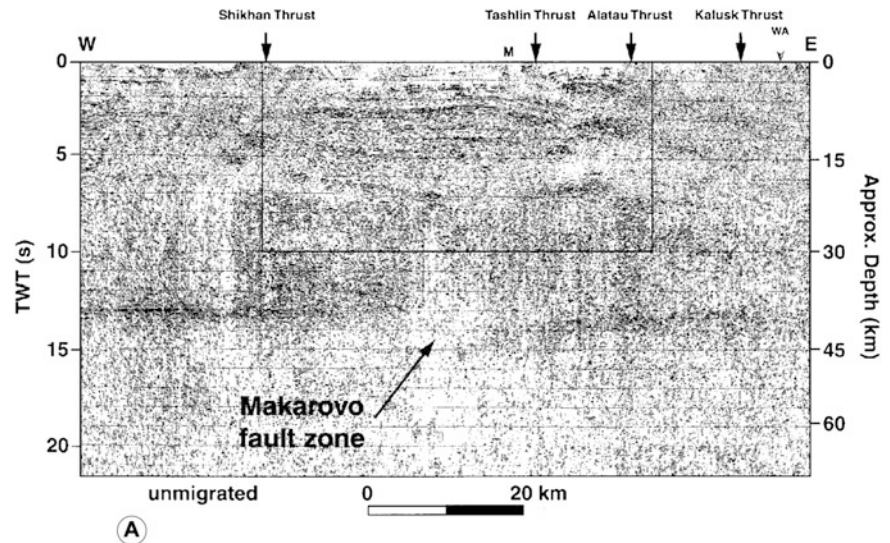


Deep Seismic Reflection and Refraction Profiling, Fig. 4 Annotated line drawing interpretation of the MOIST profile data, showing coherent reflections from throughout the crustal column. Moho shows up as a more or less continuous boundary. Several dipping thrust-like features (e.g., Outer Isles Thrust) can be seen at all depth

ranges, which also include transparent zones. One dipping reflective zone (Flannan Thrust) is seen to penetrate through the Moho into the upper mantle. (Figure from Finlayson (2010a); see Brewer and Smythe (1984) for details)

Deep Seismic Reflection and Refraction Profiling,

Fig. 5 Unmigrated seismic section from URSEIS profile across southern Urals. (Figure from Diaconescu et al. (1998), with permission from *Geology*). Moho (older than 1 billion years) is bright on the left and is offset sharply in the middle of the figure (Makarovo fault zone) by ≈ 5 km. Migrated image of the boxed upper part (not shown) contains laterally continuous stratification over this fault zone



improved acquisition/processing, such features are now regularly reported from different parts of the world.

(Near) vertical offset in the Moho, inferred at many places from teleseismic, gravimetric, and other geophysical observations, is now often imaged in NMRS. This has important bearings regarding the *nature* of Moho. This first-order seismic boundary is often thought to be a surface that re-equilibrates after tectonic events above (e.g., thrusting) or below (e.g., underplating). DSRRP images contain counter examples too, showing that Moho topography can survive later orogenic cycles, e.g., BABEL Working Group (1990) and Diaconescu et al. (1998). The reason for this behavior is not well understood.

The thermal- and the stress- regimes in the lower crust and upper mantle determine the interaction between the two during the formation of Moho as the boundary layer. Chadwick and Pharaoh (1998) interpret a DSRRP line by associating increased reflectivity of the Moho *there* to its being a detachment surface resulting from low-angle shear (Fig. 6). Local *doubling* of Moho is seen especially in some WARR data – its evolutionary mechanism remains unclear though.

WARR had taken a back seat with the increasing use of NMRS. With progress on some of the processing issues, it has made a comeback following the adage: *structures from reflection and velocities from refraction*. It is, sometimes, a part of onshore/offshore measurements and uses three-component receivers to study crustal anisotropy. Mooney and Brocher (1987) provide a review of earlier coincident WARR studies. Another example is BABEL Working Group (1993). Stand-alone WARR remains useful at places with difficult logistics; see the KRISP experiment in the Kenyan Rift Valley (Khan et al. 1999).

Deep seismic investigations using nuclear explosions: During 1965–1988, the then USSR carried out a series of

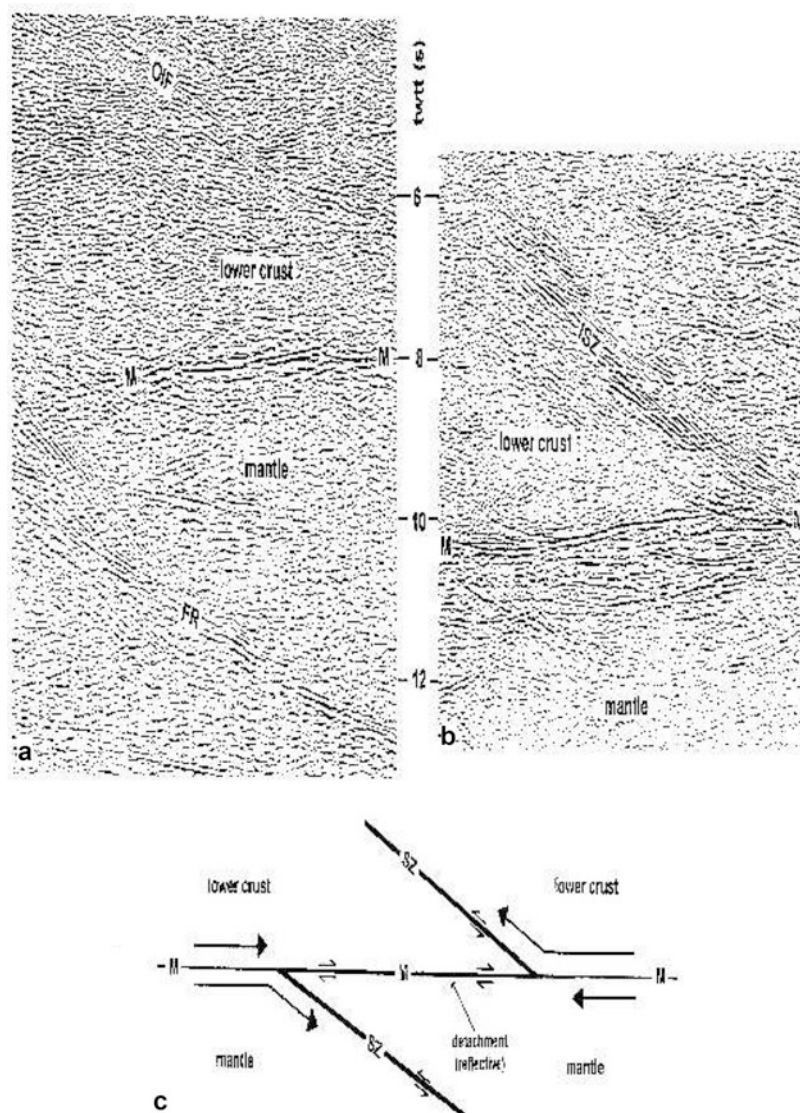
very long-range seismic experiments using *peaceful nuclear explosions* (PNEs), supplemented with chemical explosions placed every 100–150 km (Benz et al. 1992). Employing three-component receivers, typically 10 km apart, and recording out to 1550–3000 km (Pavlenkova and Pavlenkova 2006), these datasets provide an invaluable basis for current and future research; QUARTZ, the profile studied most, used 400 three-component receivers, 3 PNEs, and 51 chemical explosions (Morozov et al. [undated](#) IRIS communication; see references); analysis of its data has generated new ideas about wave propagation in the lower crust and upper mantle.

Analysis/modeling of WARR data typically starts by picking *travel times* of first and later arrivals that are interesting, the choice being decided by the data quality and the geological aim. Usually, some prominent (mid) crustal reflected/refracted arrivals are identified and *picked*, along with arrivals from Moho, and possibly deeper ones. Methods of deriving crustal models from this data include ray tracing (Zelt and Smith 1992), tomography (Hole 1992), computation of synthetics, etc. (see also ► [“Traveltime Tomography Using Controlled-Source Seismic Data”](#)). Frequently, a preliminary *velocity model* is iteratively fine-tuned to obtain a desired level of fit to the observed travel times. There are several schemes for *ray tracing*; see ► [“Seismic Ray Theory”](#) and Zelt (1995) for overviews. Of late, amplitude information has also been incorporated in such schemes.

Maguire et al. (2006) contains examples of the different aspects of the procedure in the context of the EAGLE experiment (Fig. 7). Different modeling/inversion schemes yield different results from same input. The DOBRefraction’99 Working Group (2003) provides an example of this. Pavlenkova (2011) has interpreted PNE arrival times to infer upper mantle rheology and suggests layering along the travel paths with boundaries at 100 and 200 km (fluids?).

Deep Seismic Reflection and Refraction Profiling,

Fig. 6 Schematic representation of low-angle shear near Moho (c) with parts of BIRPS seismic sections GRID (a) and NEC (b). (For details, see Chadwick and Pharaoh (1998); figure courtesy of *Tectonophysics*)

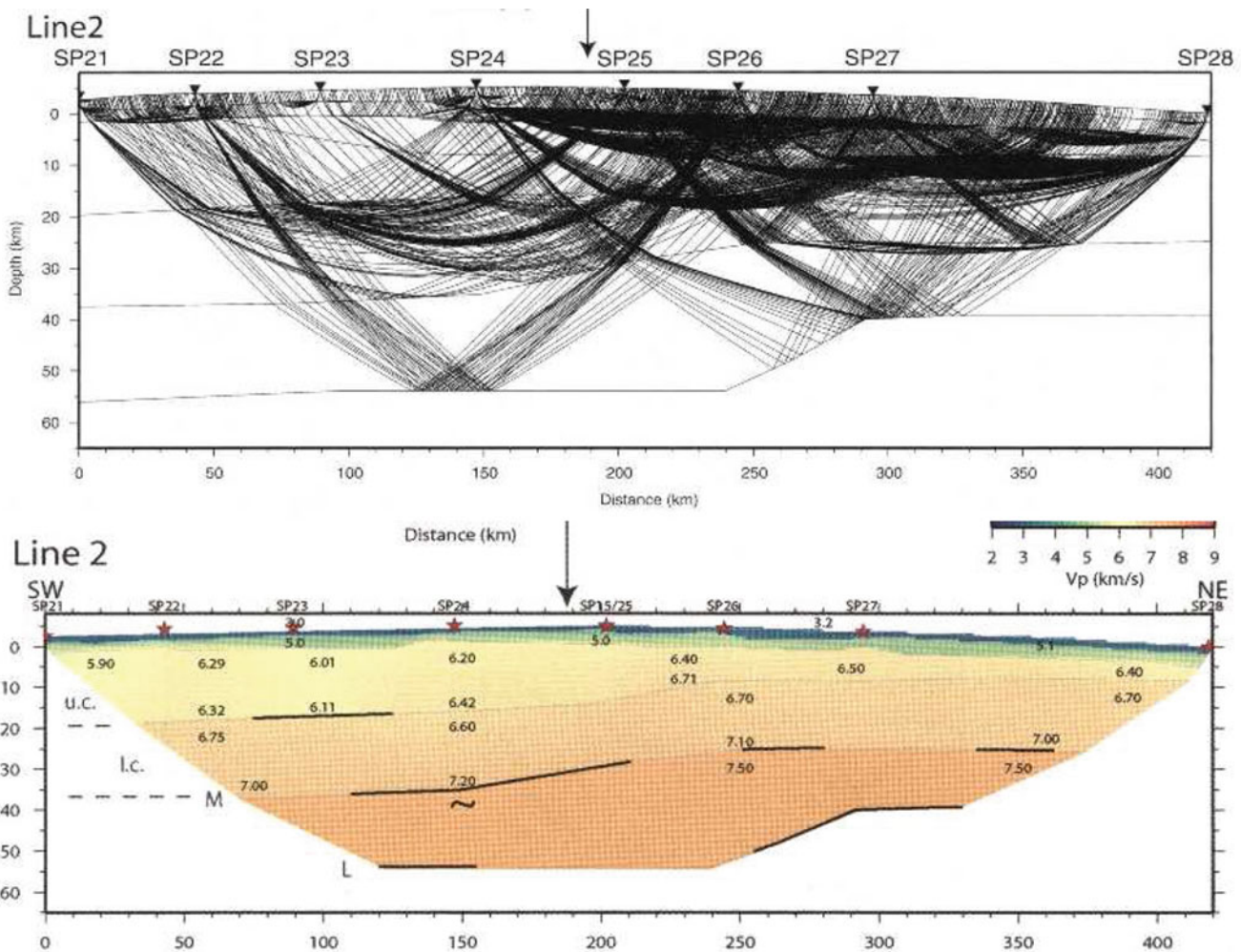


Analyzing amplitude of DSRRP signals is a crucial step, to differentiate between competing models of the crust. Improvements in the data quality, and careful processing, allow attempts to quantify the material properties that influence the strength of the recorded signals. Properties that are of primary interest are reflection coefficients (RC) across various boundaries, their variation with the angle of incidence (AVO), and the Q(uality) factor along the path (see ► [“Seismic Data Acquisition and Processing”](#) for definitions). These properties are especially important while considering the probable cause(s) of the bright spots, e.g., fluids, intrusions, layering, etc.

Warner (1990) reported RC values of about 0.1 for lower crustal layers and about 0.15 for Moho – both derived from the WAM profile (Peddy et al. 1989). The polarity (sign) of RC is important in differentiating between likely causes of bright spots sometimes seen in DSRRP images. However,

even using polarity-preserving processing, it is often only possible to determine relative values of RC. ANCORP Working Group (2003) discusses the issues involved for a couple of strong reflectors – *Nazca* and *QBBS* – while reporting $RC > 0.2$ for the latter. Combining NMRS and WARR data, Makovsky and Klemperer (1999) report RC values between -0.3 and -0.4 for the NBS bright spot in the Tibetan middle crust.

Obtaining estimates of Q along the travel path of NMRS signals (relatively high-frequency body waves) is also difficult – the effects due to scattering, conversion, etc. (apparent-Q) cannot be readily separated from the intrinsic attenuation (see also ► [“Seismic Viscoelastic Attenuation”](#)). Combining the two, Hobbs (1990) obtained a value of 500 ± 200 for effective-Q for the lower crust below the WAM profile. Morozov et al. (1998) used the data from the QUARTZ profile mentioned above to obtain a 2D



Deep Seismic Reflection and Refraction Profiling, Fig. 7 Velocity modeling by ray tracing for the EAGLE project (in the Ethiopian rift). Note the variable coverage of the ray tracing (above) and the *smooth* result, with a few long segments, in the final P-wave velocity model

(below) – both these characteristics are typical for WARR data analysis. (For details, see Maguire et al. (2006); figures courtesy of *Geol. Soc. London*, with permission from the author)

Q-model for the upper mantle below this PNE profile down to a depth of 500 km.

Some Recent Developments

DSSRP has traditionally positioned itself between basin-level seismics (using artificial sources) and deep-earth seismology (using natural sources). Recent methodological advances in the above fields have blurred this distinction. This is exhibited by the themes of the biannual meetings:

Seismix-2010: “Seismic imaging of continents and their margins: New research at the confluence of active and passive seismology”

Seismix-2012: “Advances in seismic imaging of crust and mantle: Preface”

Seismix-2016: “Seismic imaging at the cross-roads: Active, passive, exploration and solid Earth”

These may be contrasted with *seismic probing of continents and their margins*, which, with small variations, was used for almost the first three decades.

Note: most submissions at the latest biannual meeting (Seismix-2018) have not (yet) been published but are available at Seismix-2018-abs (2018). These abstracts (56 talks and 48 posters) further show the maturing of DSSRP – both regarding methodology and inclusion of other geophysical data for a better (geodynamic) interpretation.

Improvements in data acquisition are reported in Seismix-2018-abs (2018, pp. 40, 41) using the newly developed DAS (distributed acoustic sensing) technology. Both of these use the data for VSP (vertical seismic profiling) in a mineral exploration setting. Use of OBS (ocean bottom

seismometers) is reported on page 45 of the same abstract collection to improve lower crustal velocity models and also better calibrate pressure measurements from air gun sources.

Noise as a seismic source is now an exciting field of research and is referred to as *seismic interferometry* or *day-light imaging*. The basic idea is that *noise* contains useful information. Noise recorded at two locations can, for example, be used to obtain relevant properties of the intervening medium. See Curtis et al. (2006), Wapenaar and Snieder (2007), and Snieder and Wapenaar (2010) for further details and also ► “Seismic, Ambient Noise Correlation” and ► “Seismic Noise.”

Rawlinson and Goleby (2012) mentioned the use of noise for crustal imaging in their introductory article for Seismix-2010. Ito et al. (2012) use autocorrelation functions of ambient noise to obtain seismic images in a subduction zone. An interesting case of using autonomous recording *nodes* to record noise in between the active recording to obtain better V_s models is reported by Behm et al. (2019).

Inversion – Full/constrained/joint, i.e., the ability to reproduce each wiggle of the seismic observations is the ultimate goal for its interpretation, in terms of the viscoelastic properties of the causative medium (see ► “Seismic, Waveform Modeling and Tomography”); DSRRP benefits from research in this field. For a quick introduction to this topic, including underlying problems, see Virieux and Operto (2009), Virieux et al. (2017), and Brittan and Jones (2019).

True full waveform inversion (*FWI*) of seismic remains an unattained goal, but some progress is reported by, e.g., Górszczyk et al. (2019), which use both OBS and standard multichannel marine reflection data to optimize the crustal velocity model by a joint full waveform inversion. Rawlinson et al. (2016) use a detailed crustal starting model from ambient noise seismics for a *constrained inversion* in teleseismic tomography.

Multidisciplinary studies have been largely not discussed in this entry, which has focused on P-wave measurements. Availability of high-quality *multi-geophysical* data does enable extraction of additional information though. Shear waves, for example, can provide valuable constraints regarding lithology and composition – see Eccles et al. (2011) for an example.

Nowadays, data from other geophysical surveys, e.g., gravity, magnetic, magnetotelluric, geodetic, etc., are frequently available near the DSRRP transects; these can be used to improve both structural and petrophysical interpretation. See Dong et al. (2013) reporting results from the *SinoProbe* Programme (2008–2012) covering deep lithospheric exploration of China.

Integrating diverse geophysical datasets, however, needs special attention during interpretation, to obtain a model which satisfies all data in some optimal fashion. This approach may be further subdivided into *joint assessment* and *true joint/constrained inversion*.

Palomeras et al. (2011) falls in the first category, so does Yegorova and Pavlenkova (2014), who report inverting for density along the PNE profiles, with their starting model taken from seismic data.

Roberts et al. (2012) is an example of attempting to constrain the final model for a multidisciplinary dataset (seismic refraction, MT, and gravity) using computer emulation.

The road ahead will see DSRRP results being increasingly applied to geodynamic problems related to the continental crust/lithosphere (e.g., Seismix-2018-abs 2018, pp. 20, 75). Better acquisition/processing will result in crustal images/properties with a higher resolution, enabling more nuanced conclusions regarding their evolutionary history. Inroads are already being made into more fundamental questions like *earthquake processes* (Górszczyk et al. 2019).

True joint inversion of disparate datasets will remain a challenging problem – not in the least due to the relative scaling for the different datasets. Syracuse et al. (2017) report results from a joint inversion using P- and S-arrival times and dispersion data from Rayleigh waves from USARRAY deployment and Bouguer gravity anomalies. Several important issues related to FWI of different datasets are mentioned by Seismix-2018-abs (2018, p. 58), e.g., optimal acquisition design, large-scale inversions involving *hundreds of millions* of parameters, problem of nonlinearity, and optimizing high- and low-frequency contents of the result.

Research Problems

The massive amount of DSRRP data collected during the past half century has produced many new insights but has also brought up some (yet unsolved) problems.

Origin of crustal reflectivity, clearly visible in DSRRP images of mature continental crust, cannot be explained uniquely. Surface exposures of basement rocks mostly show steep dips due to earth’s surface being stress-free and explain transparent zones at shallow depths. Starting at intermediate depths, these structures are expected to become subhorizontal and seismically imageable and do reveal layering including strongly reflective zones. Smithson and Brown (1977, Fig. 5) had already proposed a *complex* crustal model with a three-fold subdivision based on geo-scientific data. Processes and materials to explain such low- and high-angle deformation, sometimes even affecting Moho and upper mantle, remain a challenge.

Shear zones, decollements, imbrications, laminae, metamorphism (facies changes, mylonitization), sill-like intrusions, and fluids (water, brine, melt, magma) have all been proposed as candidates for crustal (and Moho) reflectivity. Both Kola (Smythe et al. 1994) and KTB (Emmermann and Lauterjung 1997) super-deep drill holes have identified such conditions/structures at depth; extrapolations would point to

the presence of crustal fluids in quantities more than that earlier expected, both in bound and free form. See also Meissner et al. (2006), ► “Continental Crustal Structure” and ► “Crustal Reflectivity (Oceanic) and Magma Chamber”.

Bright spots – zones of very strong reflectivity – present an extra challenge. Pratt et al. (1993) investigated the currency bright spot (Georgia, USA), first assumed to be fluid related due to its flat nature and concluded an (ultra)mafic body instead. Makovsky and Klemperer (1999) inferred 10% (volume) of free aqueous fluids as the cause for the Tibetan bright spot. Figure 8 is an interesting example of two reflective zones, probably with different, but related, origins. Simancas et al. (2003) reported a mid-crustal highly reflective body below SW-Iberia from NMRS observations; later modeling of dense WARR data suggested the presence of high-velocity material. Palomeras et al. (2011) used additional data (heat flow, gravity, etc.) to further study this zone and inferred the presence of sill-like lenses of mantle material.

Origin and nature of (continental) Moho are also not well understood, probably because this boundary – geophysically defined as a first-order transition of P-wave velocity from ≈ 6.8 to ≈ 8.2 km/s – does not always have the same evolutionary history. However, DSRRP has replaced the earlier model for (seismological) Moho, consisting of long refraction segments – more or less flat – by a (seismic) boundary with a complex and variable structure. It seems, at places, to be the equilibrium surface established under the deviatoric stress regime after a tectono-thermal event involving the lower crust and upper mantle. At others, it

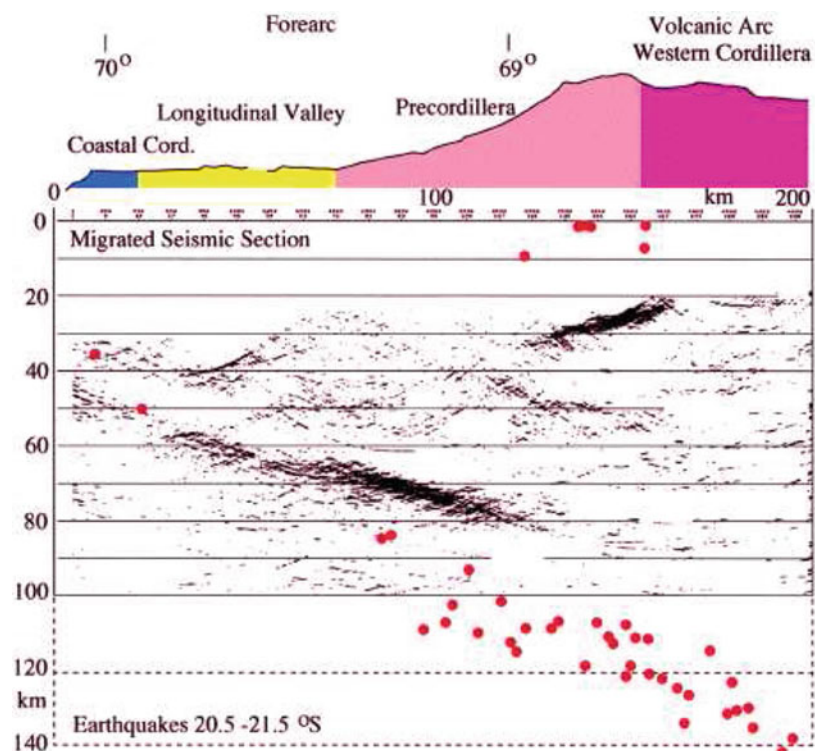
seems to be underplated or overlain by sill-like intrusions, from a later magmatic episode. Elsewhere, Moho is only identifiable as the terminal zone of a diffused lower-crustal reflectivity. At many places, it seems to exhibit enough strength, to retain its earlier structure through later tectonic events; examples include *frozen subduction* (Fig. 4) and *offset Moho* (Fig. 5).

Role of (multiple) scattering in lower crust has attracted attention lately – both to explain observations and to relate to geological evidence. Surface outcrops of Moho (read: lower crustal and upper mantle rocks) are extremely rare in the continental setting, e.g., *Kapusking Uplift* (Eastern Canada), *Musgrave Range* (Central Australia), and *Ivrea-Verbano Zone* (IVZ, Northern Italy). While the first two have been studied by DSRRP (LITHOPROBE transect KSZ and Central Australian Basin transect, respectively), IVZ – although sans seismics – has been extensively studied in the past by geologists, e.g., Zingg (1990), and geophysicists, e.g., Berckhemer (1969).

Recently, statistical analyses of detailed geological maps of several exposed (lower) crustal rocks have yielded the tantalizing possibility of a self-similar (fractal) description of their shapes in terms of typical *horizontal* and *vertical scale lengths*. For the IVZ, a 2D von Karman distribution of the structure with bimodal petrology has been derived (Holliger and Levander 1992). Holliger et al. (1993) computed synthetics for such a simulation of IVZ and were able to qualitatively reproduce lower-crustal reflectivity observed in NMRS. At large distances (WARR configuration), the

Deep Seismic Reflection and Refraction Profiling,

Fig. 8 Migrated seismic section from the ANCORP line 2 in the Central Andean subduction zone. (Reproduced from <http://www-app1.gfz-potsdam.de/www/pb3/dekorp/an-fig/Amline2.gif>, with permission from the author). The strongly reflective zone on upper right is the *Quebrada Blanca Bright Spot* (QBBS), and the reflective zone dipping to the right is the *Nazca reflector*. Superimposed red dots show seismicity in this area of active subduction. (For further details, see ANCORP Working Group (2003))



synthetics from the random medium contained laterally correlatable events, which could be erroneously used for velocity analysis, migration, etc., a possibility already suggested earlier (e.g., Levander and Gibson 1991; Emmerich et al. 1993).

Another explanation offered for increased reflectivity in the lower crust, including transition zones, is lamellar structures with associated wave propagation effects (amplitude, anisotropy); see Meissner et al. (2006) and ► “Magnetic Anisotropy.”

Both these possibilities, random heterogeneity and lamination, bring up the role of multiple scattering in the DSRRP wave field and question the propriety of using conventional tools from basin exploration for processing such data. Douma and Roy Chowdhury (2001) used synthetics to show that multiple scattering has a limited effect for a 1D bimodal model of IVZ but also mentioned the need for 2D full-wave numerical simulations.

In the upper mantle too, multiple scattering seems to play an important role. Menke and Chen (1984) had invoked this to explain long-range propagation of P_n -phase from earthquakes. More recently, DSRRP data from PNE profiles, e.g., QUARTZ, has shown surprisingly strong propagation of high-frequency (5 Hz) P_n -phase to distances of 3,000 km!

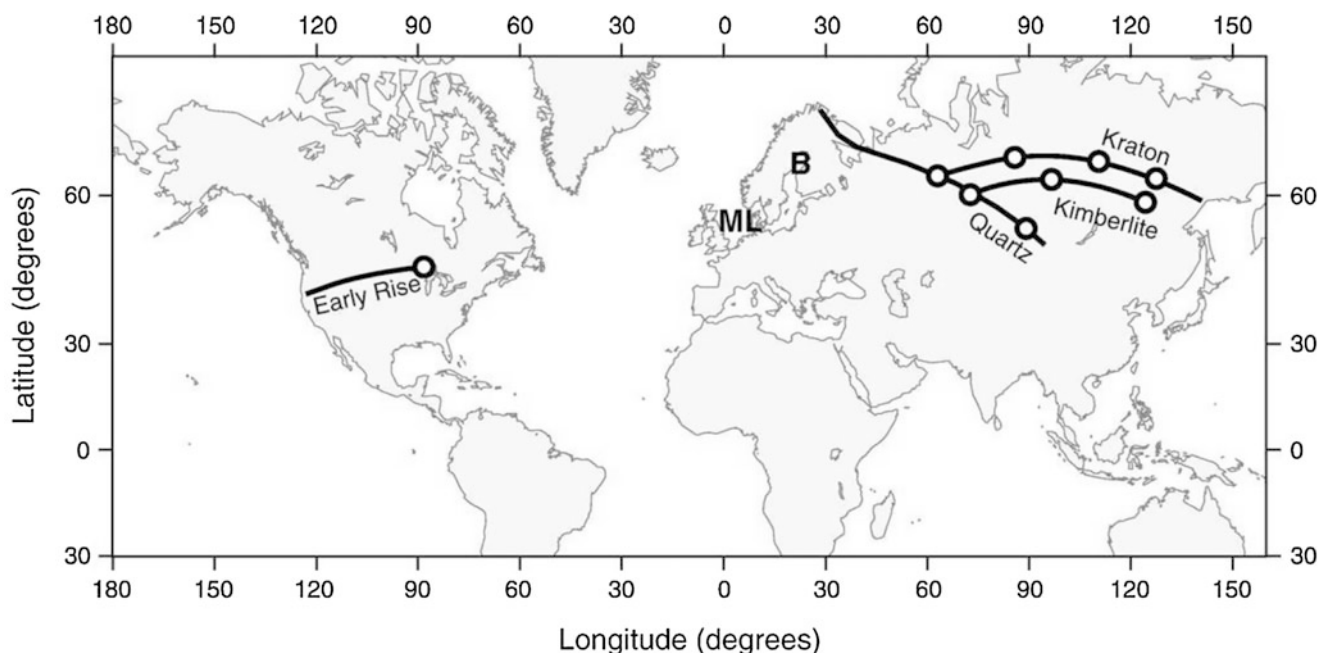
Estimating descriptive parameters of a possibly random medium in lower crust and the upper mantle can be done by modeling or direct estimation. To explain the long-distance P_n -phase in the QUARTZ data, Ryberg et al. (1995) modeled an upper mantle zone of *horizontally stretched, randomly*

distributed velocity anisotropy. Nielsen and Thybo (2006) modeled a larger dataset (Fig. 9) and inferred random heterogeneity for both lower crust and upper mantle.

Following some earlier work (Hurich 1996; Pullammanappallil et al. 1997; Hurich 2003; Carpentier and Roy Chowdhury 2007), Carpentier et al. (2011) have recently analyzed the data from Line 48 of the LITHOPROBE transect AG statistically. Assuming a 2D *von Karman* medium, they estimated the horizontal scale length of the medium directly from the seismic wave field (Fig. 10), which also shows their interpretation. The comparison with an earlier line drawing-based interpretation (above) illustrates the similarities and differences between the two approaches.

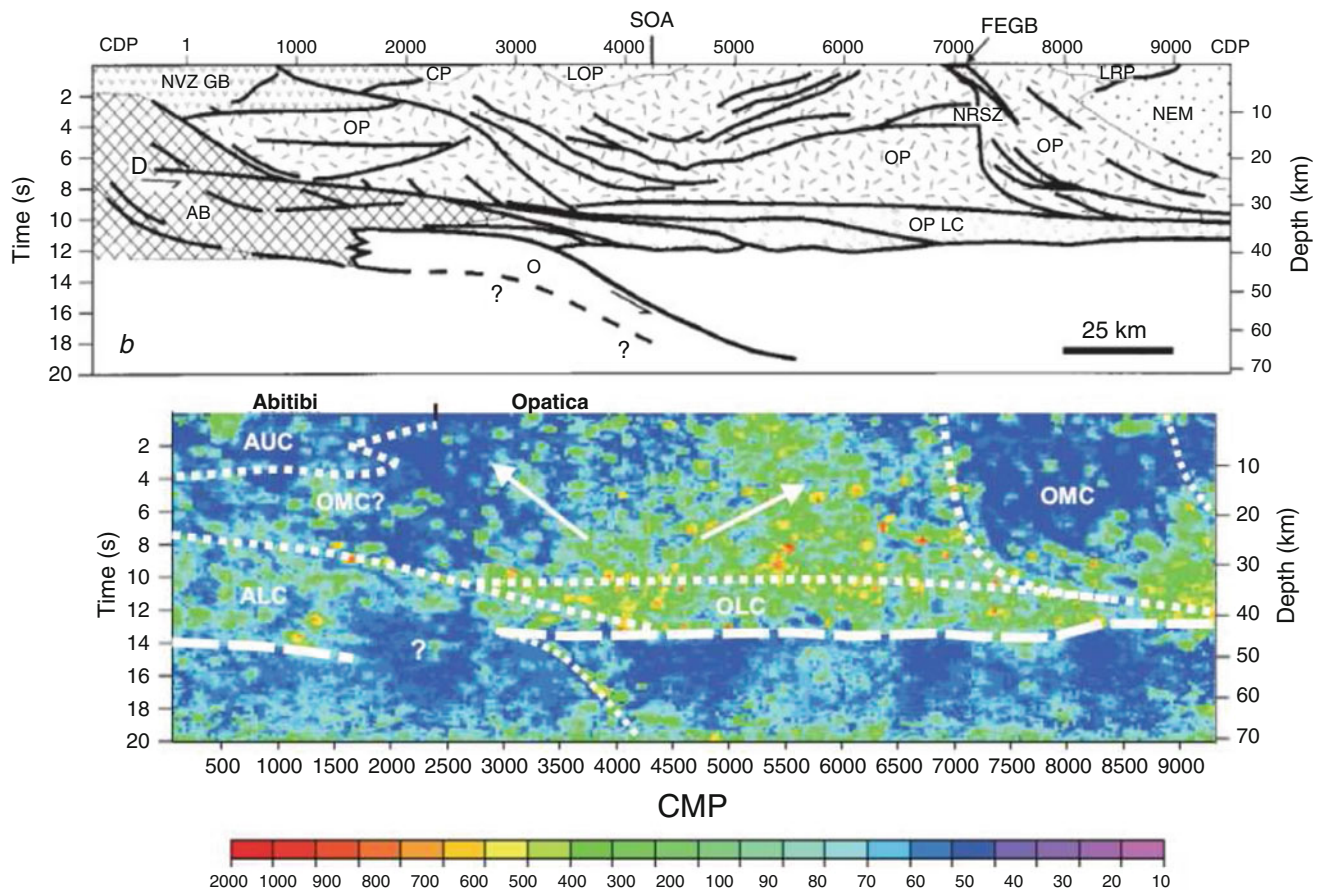
Summary

DSRRP has, over half a century, produced quality images of the continental crust and its margins, revealing their complex structure. These – including some unexpected results, e.g., *frozen subduction* – have contributed significantly to our ideas about the processes, current and ancient, involved in their evolution. As the deep structures, mostly inaccessible, play an important role in the development of the shallower geology, understanding these (*deep faults*) also helps in the optimal exploration of economic resources, e.g., hydrocarbons, ore deposits, etc., and in the study of natural hazards associated with volcanism and earthquakes.



Deep Seismic Reflection and Refraction Profiling, Fig. 9 Long-range seismic lines used for modeling by Nielsen and Thybo (2006). *Early Rise* used chemical explosives; the other lines used PNEs. B and

ML mark the locations of NMRS lines *BABEL* and *MONA LISA*, respectively. (Figure courtesy of *Tectonophysics*)



Deep Seismic Reflection and Refraction Profiling, Fig. 10 Above: Tectonic interpretation of LITHOPROBE line AG-48 (Abitibi-Grenville, Canada, taken from Calvert et al. (1995); figure courtesy of

Nature). Below: interpretation by Carpentier et al. (2011) overlain on their estimation of the horizontal scale length (a_x) of the medium from the seismic data. (Figure courtesy of *Tectonophysics*)

Cross-References

- Continental Crustal Structure
- Continental Rifts
- Crustal Reflectivity (Oceanic) and Magma Chamber
- Earth's Structure, Upper Mantle
- Lithosphere, Continental
- Magnetic Anisotropy
- Plate Tectonics, Precambrian
- Seismic Data Acquisition and Processing
- Seismic Imaging, Overview
- Seismic Instrumentation
- Seismic Monitoring of Nuclear Explosions
- Seismic Noise
- Seismic Ray Theory
- Seismic Tomography
- Seismic Viscoelastic Attenuation
- Seismic Waves, Scattering
- Seismic, Ambient Noise Correlation
- Seismic, Waveform Modeling and Tomography
- Traveltime Tomography Using Controlled-Source Seismic Data

Bibliography

- Aki K, Richards PG (2002) Quantitative seismology, 2nd edn. University Science Books, Mill Valley, USA
- ANCORP Working Group (2003) Seismic imaging of a convergent continental margin and plateau in the central Andes (Andean Continental Research Project 1996 (ANCORP'96)). *J Geophys Res* 108(B7):2328
- BABEL Working Group (1990) Evidence for early Proterozoic plate tectonics from seismic reflection profiles in the Baltic shield. *Nature* 348:34–38
- BABEL Working Group (1993) Deep seismic reflection/refraction interpretation of critical structure along BABEL profiles A and B in the southern Baltic Sea. *Geophys J Int* 112:325–343
- Barazangi M, Brown L (eds) (1986a) Reflection seismology: a global perspective. Geodynamic series, vol 13. American Geophysical Union, Washington, DC
- Barazangi M, Brown L (eds) (1986b) Reflection seismology: the continental crust. Geodynamic series, vol 14. American Geophysical Union, Washington, DC
- Behm M, Cheng F, Patterson A, Soreghan G (2019) Passive processing of active nodal seismic data: estimation of V_p/V_s ratios to characterize structure and hydrology of an alpine valley infill. *Solid Earth* 10:1337–1354
- Benz HM, Huger JD, Leith WS, Mooney WD, Solodilov LN, Egorkin AV, Ryaboy VS (1992) Deep seismic sounding in Northern Eurasia. *EOS Trans Am Geophys Union* 73:297–300

- Berckhemer H (1969) Direct evidence for the composition of the lower crust and the Moho. *Tectonophysics* 8:97–105
- Brewer JA, Matthews DH, Warner MR, Hall J, Smythe DK, Whittington RJ (1983) BIRPS deep seismic reflection studies of the British Caledonides. *Nature* 305:206–210
- Brewer JA, Smythe DK (1984) MOIST and the continuity of crustal reflector geometry along the Caledonian–Appalachian orogen. *J Geol Soc Lond* 141:105–120
- Brittan J, Jones I (2019) FWI evolution – from a monolith to a toolkit. *Lead Edge* 38:179–184
- Calvert AJ, Sawyer EW, Davis WJ, Ludden JN (1995) Archaean subduction inferred from seismic images of a mantle suture in the Superior Province. *Nature* 375:670–674
- Carbonell R, Gallart J, Tome M (eds) (2000) Deep seismic profiling of the continents and their margins, *Tectonophysics* 329
- Carbonell R, Sallares V, Ranero CR, Booth-Rea G (eds) (2016) Special issue on deep seismix, *Tectonophysics* 689
- Carpentier SFA, Roy Chowdhury K (2007) Underestimation of scale lengths in stochastic fields and their seismic response: a quantification exercise. *Geophys J Int* 169:547–562
- Carpentier SFA, Roy Chowdhury K, Hurich CA (2011) Mapping correlation lengths of lower crustal heterogeneities together with their maximum-likelihood uncertainties. *Tectonophysics* 508(1–4): 117–130. <https://doi.org/10.1016/j.tecto.2010.07.008>
- Chadwick RA, Pharaoh TC (1998) The seismic reflection Moho beneath the United Kingdom and adjacent areas. *Tectonophysics* 299:255–279
- Clowes RM, Green AG (eds) (1994) Seismic reflection probing of the continents and their margins, *Tectonophysics* 232
- Clowes RM, Ellis RM, Hajnal Z, Jones IF (1983) Seismic reflections from subducting lithosphere? *Nature* 03:668–670
- Cook FA, Vasudevan K (2006) Reprocessing and enhanced interpretation of the initial COCORP Southern Appalachian traverse. *Tectonophysics* 420:161–174
- Curtis A, Gerstoft P, Sato H, Snieder R, Wapenaar K (2006) Seismic interferometry – turning noise into signal. *The Leading Edge* 25(9): 1082–1092
- Davey FJ, Jones L (eds) (2004) Special issue – Continental lithosphere, *Tectonophysics* 388
- DEKORP. German continental seismic reflection program. <https://www.gfz-potsdam.de/en/section/near-surface-geophysics/projects/past-projects/dekorp-german-continental-reflection-seismic-program>
- Diaconescu CA, Knapp JH, Brown LD, Steer DN, Stiller M (1998) Precambrian Moho offset and tectonic stability of the East European platform from the URSEIS deep seismic profile. *Geology* 26: 211–214
- Dohr G (1957) Ein Beitrag der Reflexionsseismik zur Erforschung des tieferen Untergrundes. *Geol Rundsch* 46:17–26
- Dohr G, Fuchs K (1967) Statistical evaluation of deep crustal reflections in Germany. *Geophysics* 32(6):951–967
- Dong S, Li T, Lü Q, Gao R, Yang J, Chen X, Wei W (2013) Progress in deep lithospheric exploration of the continental China: a review of the SinoProbe. *Tectonophysics* 606:1–13
- Douma H, Roy Chowdhury K (2001) Amplitude effects due to multi-scale impedance contrasts and multiple scattering: implications for Ivrea-type continental lower crust. *Geophys J Int* 147(2):435–448
- EarthScope. Exploring the structure and evolution of the North American continent. <http://www.earthscope.org>
- Eccles JD, White RS, Christie PAF (2011) The composition and structure of volcanic rifted continental margins in the North Atlantic: further insight from shear-waves. *Tectonophysics* 508:22–33
- Emmerich H, Zwieli J, Muller G (1993) Migration of synthetic seismograms for crustal structures with random heterogeneities. *Geophys J Int* 113:225–238
- Emmermann R, Lauterjung J (1997) The German Continental Deep Drilling Program KTB: overview and major results. *J Geophys Res* 102(B8):18,179–18,201
- Finlayson DM (2010a) BIRPS startup: deep seismic profiling along the MOIST line on the continental shelf around the British Isles, 1981. <http://www.earthscrust.org.au/science/startups/birps-su.html>
- Finlayson DM (2010b) COCORP startup: the first deep seismic profiling of the continental crust in USA, Hardeman County, Texas, 1975. <http://www.earthscrust.org.au/science/startups/cocorp-su.html>
- Flack C, Warner M (1990) Three-dimensional mapping of seismic reflections from the crust and upper mantle, northwest of Scotland. *Tectonophysics* 173:469–481
- Gamburtsev GA (1952) Deep seismic soundings of the earth's crust. *Doklady Akad Nauk SSSR* 87:943–945
- Gebrande H, Castellarin A, Luschen E, Neubauer F, Nicolich R (eds) (2006) TRANSALP – a transect through a young collisional orogen, *Tectonophysics* 414
- Gibbs AK (1986) Seismic reflection profiles of precambrian crust: a qualitative assessment. In: Barazangi M, Brown L (eds) *Reflection seismology: the continental crust*. Geodynamic series, vol 14. American Geophysical Union, Washington, DC, pp 95–106
- Górszczyk A, Operto S, Schenini L, Yamada Y (2019) Crustal-scale depth imaging via joint full-waveform inversion of ocean-bottom seismometer data and pre-stack depth migration of multichannel seismic data: a case study from the eastern Nankai Trough. *Solid Earth* 10:765–784
- György S (ed) (1972) The crustal structure of central and southeastern Europe based on the results of explosion seismology. Hungarian Geophysical Institute Roland Eötvös, Budapest. *Geophysical Transactions*
- Hansen K, Roy Chowdhury K, Phinney RA (1988) The sign filter for seismic event detection. *Geophysics* 53(8):1024–1033
- Heikkinen P, Kukkonen I, Thybo H (eds) (2011) Special issue – Seismix 2008, active and passive plate margins and subduction zones, *Tectonophysics* 508
- Hobbs RW (1990) Effective Q determination using frequency methods on BIRPS data. *Tectonophysics* 173:25–30
- Hole JA (1992) Nonlinear high-resolution three-dimensional seismic travel time tomography. *J Geophys Res* 97(B5):6553–6562
- Holliger K, Levander AR (1992) A stochastic view of lower crustal fabric based on evidence from the Ivrea zone. *Geophys Res Lett* 19(11):1153–1156
- Holliger K, Levander AR, Goff JA (1993) Stochastic modeling of the reflective lower crust: petrophysical and geological evidence from the Ivrea zone (Northern Italy). *J Geophys Res* 98:11967–11980
- Hurich CA (1996) Statistical description of seismic reflection wave fields: a step towards quantitative interpretation of deep seismic reflection profiles. *Geophys J Int* 125:719–728
- Hurich CA (2003) The nature of crustal seismic heterogeneity: a case study from the Grenville Province. In: *Heterogeneity in the crust and upper mantle: nature, scaling, and seismic properties*. Kluwer Academic, New York, pp 299–320
- Hurich CA, Kocurko A (2000) Statistical approaches to interpretation of seismic reflection data. *Tectonophysics* 329:251–267
- IGCP-559 (2010a) Andrija Mohorovicic (1857–1936) – defining the earth's crust. <http://www.earthscrust.org.au/science/historic/andrija.html>
- IGCP-559 (2010b) International symposium on deep structure of the continents and their margins series. http://www.earthscrust.org.au/deep_structure_of_the_continents.htm
- IGCP-559 (2010c) Seismic imaging programs. <http://www.earthscrust.org.au/links.html>
- Ito Y, Shiomi K, Nakajima J, Hino R (2012) Autocorrelation analysis of ambient noise in northeastern Japan subduction zone. *Tectonophysics* 572–573:38–46
- Kaila KL, Krishna VG, Roy Chowdhury K, Narain H (1978) Structure of the Kashmir Himalaya from deep seismic soundings. *J Geol Soc India* 19:1–20
- Kaila KL, Roy Chowdhury K, Reddy PR, Krishna VG, Narain H, Subbotin SI, Sollogub VB, Chekunov AV, Kharechko GE,

- Lazarenko MA, Ilchenko TV (1979) Crustal structure along Kavali-Udipi profile in the Indian peninsular shield from deep seismic sounding. *J Geol Soc India* 20:307–333
- Kanao M, Fujiwara A, Miyamachi H, Toda S, Ito K, Tomura M, Ikawa T, The SEAL Geotransect Group (2011) Reflection imaging of the crust and the lithospheric mantle in the Lützow-Holm complex, Eastern Dronning Maud Land, Antarctica, derived from SEAL transects. *Tectonophysics* 508:73–84
- Kanasewich ER, Clowes RM, McLoughan CH (1969) A buried precambrian rift in western Canada. *Tectonophysics* 8:513–527
- Khan MA, Mechie J, Birt C, Byrne G, Gaciri S, Jacob B, Keller GR, Maguire PKH, Novak O, Nyambok IO, Pate JP, Prodehl C, Riaroh D, Simiyu S, Thybo H (1999) The lithospheric structure of the Kenya Rift as revealed by wide-angle seismic measurements. In: MacNiocaill C, Ryan P (eds) *Continental tectonics*. Special publication, vol 164. Geological Society, London, pp 257–269
- Klemperer SL, Mooney WD (eds) (1998a) Deep seismic profiling of the continents I: general results and new methods, *Tectonophysics* 286
- Klemperer SL, Mooney WD (eds) (1998b) Deep seismic profiling of the continents II: a global survey, *Tectonophysics* 288
- Kong SM, Phinney RA, Roy Chowdhury K (1985) A nonlinear signal detector for enhancement of noisy seismic record sections. *Geophysics* 50(4):539–550
- Kumar V, Oueity J, Clowes RM, Herrmann F (2011) Enhancing crustal reflection data through curvelet denoising. *Tectonophysics* 508:106–116
- Levander AR, Gibson BS (1991) Wide-angle seismic reflections from two-dimensional random target zones. *J Geophys Res* 96(B6): 10251–10260
- Leven JH, Finlayson DM, Wright C, Dooley JC, Kennet BLN (eds) (1990) *Seismic probing of continents and their margins*, *Tectonophysics* 173
- Liner CL (2004) *Elements of 3D seismology*. PennWell corporation, Tulsa
- Maguire PKH, Keller GR, Klemperer SL, Mackenzie GD, Keranen K, Harder S, O'Reilly B, Thybo H, Asfaw L, Khan MA, Amha M (2006) Crustal structure of the northern Main Ethiopian Rift from the EAGLE controlled-source survey; a snapshot of incipient lithospheric break-up; special publication. In: Yirgu G, Ebinger CJ, Maguire PKH (eds) *The Afar volcanic province within the East African rift system*, vol 259. Geological Society, London, pp 269–292
- Makovsky Y, Klemperer SL (1999) Measuring the seismic properties of Tibetan bright spots: evidence for free aqueous fluids in the Tibetan middle crust. *J Geophys Res* 104(B5):10795–10825
- Malinowski M, Czapczyk CM, Carbonell R, Rawlinson N (eds) (2019) Special issue – Advances in seismic imaging across the scales. European Geosciences Union. <https://www.solid-earth.net/special-issue986.html>
- Matthews D, Smith C (eds) (1987) *Deep seismic reflection profiling of the continental lithosphere*. Royal Astronomical Society, vol 89. Blackwell, Oxford, UK
- Meissner R, Brown L, Dürbaum H-J, Franke W, Fuchs K, Siefert F (eds) (1991) *Continental lithosphere: deep seismic reflections*. Geodynamic series, vol 22. American Geophysical Union, Washington, DC
- Meissner R, Rabbel W, Kern H (2006) Seismic lamination and anisotropy of the lower continental crust. *Tectonophysics* 416:81–99
- Menke W, Chen R (1984) Numerical studies of the coda falloff rate of multiply scattered waves in randomly layered media. *Bull Seismol Soc Am* 74(5):1605–1621
- Mereu RF, Kovach RJ (1970) A portable inexpensive seismic system for crustal studies. *Bull Seismol Soc Am* 60(5):1607–1613
- Mooney WD, Brocher TM (1987) Coincident seismic reflection/refraction studies of the continental lithosphere: a global review. *Rev Geophys* 25:723–742. American Geophysical Union
- Morgan JV, Hadwin M, Warner MR, Barton PJ, Morgan RPL (1994) The polarity of deep seismic reflections from the lithospheric mantle: evidence for a relict subduction zone. *Tectonophysics* 232:319–328
- Morozov IB, Morozova EA, Smithson SB, Solodilov LN (1998) 2-D Image of seismic attenuation beneath the deep seismic sounding profile “Quartz,” Russia
- Morozov IB, Morozova EA, Smithson SB, Solodilov LN (undated) Long range profile Quartz. www.adc1.iris.edu/data/reports/misc.ext/QUARTZ.pdf
- Nielsen L, Thybo H (2006) Identification of crustal and upper mantle heterogeneity by modelling of controlled-source seismic data. *Tectonophysics* 416:209–228
- Oliver J, Dobrin M, Kaufman S, Meyer R, Phinney R (1976) *Continuous seismic reflection profiling of the deep basement*, Hardeman County, Texas. *Geol Soc Am Bull* 87:1537–1546
- Palomeras I, Carbonell R, Ayarza P, Fernández M, Simancas JF, Poyatos DM, Lodeiro FG, Pérez-Estaún A (2011) Geophysical model of the lithosphere across the Variscan Belt of SW-Iberia: multidisciplinary assessment. *Tectonophysics* 508:42–51
- PASSCAL. Program of array seismic studies of continental lithosphere. <http://iris.edu/hq/files/programs/passcal>
- Pavlenkova NI (2011) Seismic structure of the upper mantle along long-range PNE profiles – rheological implication. *Tectonophysics* 508:85–95
- Pavlenkova GA, Pavlenkova NI (2006) Upper mantle structure of Northern Eurasia for peaceful nuclear explosion data. *Tectonophysics* 416:33–52
- Peddy C, Pinet B, Masson D, Scrutton R, Sibuet JC, Warner MR, Lefort JP, Shroeder IJ, BIRPS, ECORS (1989) Crustal structure of the Goban Spur continental margin, Northeast Atlantic, from deep seismic reflection profiling. *J Geol Soc* 146:427–437
- Percival JA, Green AG, Milkerei B, Cook FA, Geis W, West GF (1989) Seismic reflection profiles across deep continental crust exposed in the Kapuskasing uplift structure. *Nature* 342:416–420
- Phinney RA (1986) A seismic cross-section of the New England Appalachians: the orogen exposed. In: Barazangi M, Brown L (eds) *Reflection seismology: the continental crust*. Geodynamic series, vol 14. American Geophysical Union, Washington, DC, pp 151–172
- Phinney R, Roy Chowdhury K (1989) Reflection seismic studies of crustal structure in the Eastern United States. In: Pakiser L, Mooney W (eds) *Geophysical framework of the continental United States: GSA memoir*, vol 172. Geological Society of America, Boulder, pp 613–653
- Pratt TL, Mondary JF, Brown LD (1993) Crustal structure and deep reflector properties: wide angle shear and compressional wave studies of the midcrustal surrency bright spot beneath southeastern Georgia. *J Geophys Res* 98(B10):17,723–17,735
- Pullammanappallil S, Levander A, Larkin S (1997) Estimation of crustal stochastic parameters from seismic exploration data. *J Geophys Res* 102(B7):15,269–15,286
- Pylypenko VN, Verpakhovska OO, Starostenko VI, Pavlenkova NI (2011) Wave imags of the crustal structure from refraction and wide-angle reflection migrations along the DOBRE profile (Dnieper-Donets paleorift). *Tectonophysics* 508:96–105
- Rawlinson N, Goleby B (eds) (2012) Special issue – Seismic imaging of continents and their margins: new research at the confluence of active and passive seismology, *Tectonophysics* 572–573
- Rawlinson N, Pilia S, Young M, Salmon M, Yang Y (2016) Crust and upper mantle structure beneath southeast Australia from ambient noise and teleseismic tomography. *Tectonophysics* 689:143–156
- Rawlinson N, Stephenson R, Carbonell R (eds) (2017) Special issue – Seismic 2016: advances in active and passive seismic imaging of continents and their margins, *Tectonophysics* 718
- Reddy PR, Venkateswaralu N, Koteswar Rao P, Prasad ASSRS (1999) Crustal structure of peninsular shield, India from DSS studies. *Curr Sci* 77:1606–1611

- Roberts AW, Hobbs RW, Goldstein M, Moorkamp M, Jegen M, Heincke B (2012) Crustal constraint through complete model space screening for diverse geophysical datasets facilitated by emulation. *Tectonophysics* 572–573:47–63
- Roy Chowdhury K, Hargraves RB (1981) Deep seismic soundings in India and the origin of continental crust. *Nature* 291(5817):648–650
- Ryberg T, Fuchs K, Egorkin A, Solodilov L (1995) Observation of high-frequency teleseismic P_n on the long-range Quartz profile across northern Russia. *J Geophys Res* 100(B9):18151–18163
- Santosh M, Carbonell R, Artemieva I, Badal J (eds) (2014) Special issue – Advances in seismic imaging of crust and mantle, *Tectonophysics* 627
- Seismix-2018-abs (2018) Book of abstracts, 18th international SESIMIX symposium, seismology between the poles, Cracow. <http://seismix2018.pl/abstract>
- Sheriff RE, Geldart LP (1995) *Exploration seismology*, 2nd edn. Cambridge University Press, Cambridge, UK
- Simancas F, Carbonell R, González-Lodeiro F, Estaún AP, Juhlin C, Ayarza P, Kashubin A, Azor A, Poyatos DM, Almodóvar G, Pascual E, Sáez R, Expósito I (2003) Crustal structure of the transpressional Variscan of SW Iberia: SW Iberia deep seismic reflection profile (IBERSEIS). *Tectonics* 22(6):1–1–1–19
- Smithson SB, Brown SK (1977) A model for lower continental crust. *Earth Planet Sci Lett* 35:134–144
- Smythe DK, Dobinson A, McQuillin R, Brewer JA, Matthews DH, Blundell DJ, Kelk B (1982) Deep structure of the Scottish Caledonides revealed by the MOIST reflection profile. *Nature* 299:338–340
- Smythe DK, Smithson SB, Gillen C, Humphreys C, Kristoffersen Y, Karev NA, Garipov VZ, Pavlenkova NI, the Kola-92 Working Group (1994) Project images crust, collects seismic data in world's largest borehole. *EOS Trans Am Geophys Union* 75:473–476
- Snieder R, Wapenaar K (2010) Imaging with ambient noise. *Phys Today* 63(9):44–49
- Snyder DB, Eaton DW, Hurich CA (eds) (2006) *Seismic probing of continents and their margins*, *Tectonophysics* 420
- Steinhart JS, Meyer RP (1961) *Explosion studies of continental structure*. Carnegie Institution, Washington, DC. publ 622
- Syracuse EM, Zhang H, Maceira M (2017) Joint inversion of seismic and gravity data for imaging seismic velocity structure of the crust and upper mantle beneath Utah, United States. *Tectonophysics* 718:105–117
- Texan. Miniature seismic recorder. <http://iris.edu/hq/files/programs/passcal/docs/125A-01.pdf>
- The DOBREFraction'99 Working Group (2003) “DOBREFraction'99” – velocity model of the crust and upper mantle beneath the Donbas Foldbelt (East Ukraine). *Tectonophysics* 371:81–110
- Thybo H (ed) (2002) *Deep seismic profiling of the continents and their margins*, *Tectonophysics* 355
- USARRAY. A continental-scale seismic observatory. <http://www.usarray.org>
- van der Baan M (2000) Recognition and reconstruction of coherent energy with application to deep seismic reflection data. *Geophysics* 65(2):656–667
- Vasudevan K, Cook FA (1998) Skeletons and fractals – a statistical approach to deep crustal seismic data processing and interpretation. *Tectonophysics* 286:93–109
- Vermeeer GJO (2002) 3-D seismic survey design. Society of Exploration Geophysicists, Tulsa
- Virieux J, Operto S (2009) An overview of full-waveform inversion in exploration geophysics. *Geophysics* 74(6):wcc1–wcc26
- Virieux J, Asnaashari A, Brossier R, Métivier L, Ribodetti A, Zhou W (2017) An overview of full-waveform inversion in exploration geophysics. In: *Encyclopedia of exploration geophysics*. SEG, Tulsa, Oklahoma, pp R1-1–R1-40. <https://doi.org/10.1190/1.9789560803027.entry6>
- Wapenaar K, Snieder R (2007) Chaos tamed. *Nature* 447:643
- Warner M (1990) Absolute reflection coefficients from deep seismic reflections. *Tectonophysics* 173:15–23
- White DJ, Ansorge J, Bodoky TJ, Hajnal Z (eds) (1996) *Seismic reflection probing of the continents and their margins*, *Tectonophysics* 264
- Yegorova T, Pavlenkova G (2014) Structure of the upper mantle of Northern Eurasia from 2D density modeling on seismic profiles with peaceful nuclear explosions. *Tectonophysics* 627:57–71
- Yilmaz Ö (2001) *Seismic data analysis, processing, inversion and interpretation of seismic data*. Investigations in geophysics, vol I, 2nd edn. Society of Exploration Geophysicists, Tulsa
- Zelt CA (1995) Modelling strategies and model assessment for wide-angle seismic traveltimes data. *Geophys J Int* 139:183–204
- Zelt CA, Smith RB (1992) Seismic travel time inversion for 2-D crustal velocity structure. *Geophys J Int* 108:16–34
- Zhao W, Nelson KD, Project INDEPTH team (1993) Deep seismic reflection evidence for continental under thrusting beneath southern Tibet. *Nature* 366:557–559
- Zingg A (1990) The Ivrea crustal cross-section (Northern Italy and Southern Switzerland). In: Salisbury MH, Fountain DM (eds) *Exposed cross sections of the continental crust*. Kluwer, Dordrecht, pp 1–19

Differential Rotation of the Earth's Inner Core

Xiaodong Song

Department of Geology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Definition

Differential rotation of the inner core refers to the difference between the rotation of the Earth's solid inner core and the daily rotation of the Earth's crust and the mantle.

Background

Driven by gravity, the Earth differentiated into iron core and rocky mantle at the early stage of its formation. The central core is liquid because of high temperature; but at the center of the Earth, an inner core is formed and gradually grows as the Earth cools, and the liquid iron freezes under tremendous pressure (three million times the atmospheric pressure). The size of the inner core is slightly smaller than that of the moon. Separated from the solid mantle by the fluid outer core, the inner core is thus free to rotate independently under an external torque.

The early idea of inner core rotation came from studies of the geodynamo, which generated the Earth's magnetic field. Gubbins (1981) first suggested that electromagnetic forces between the electrically conducting inner core and the

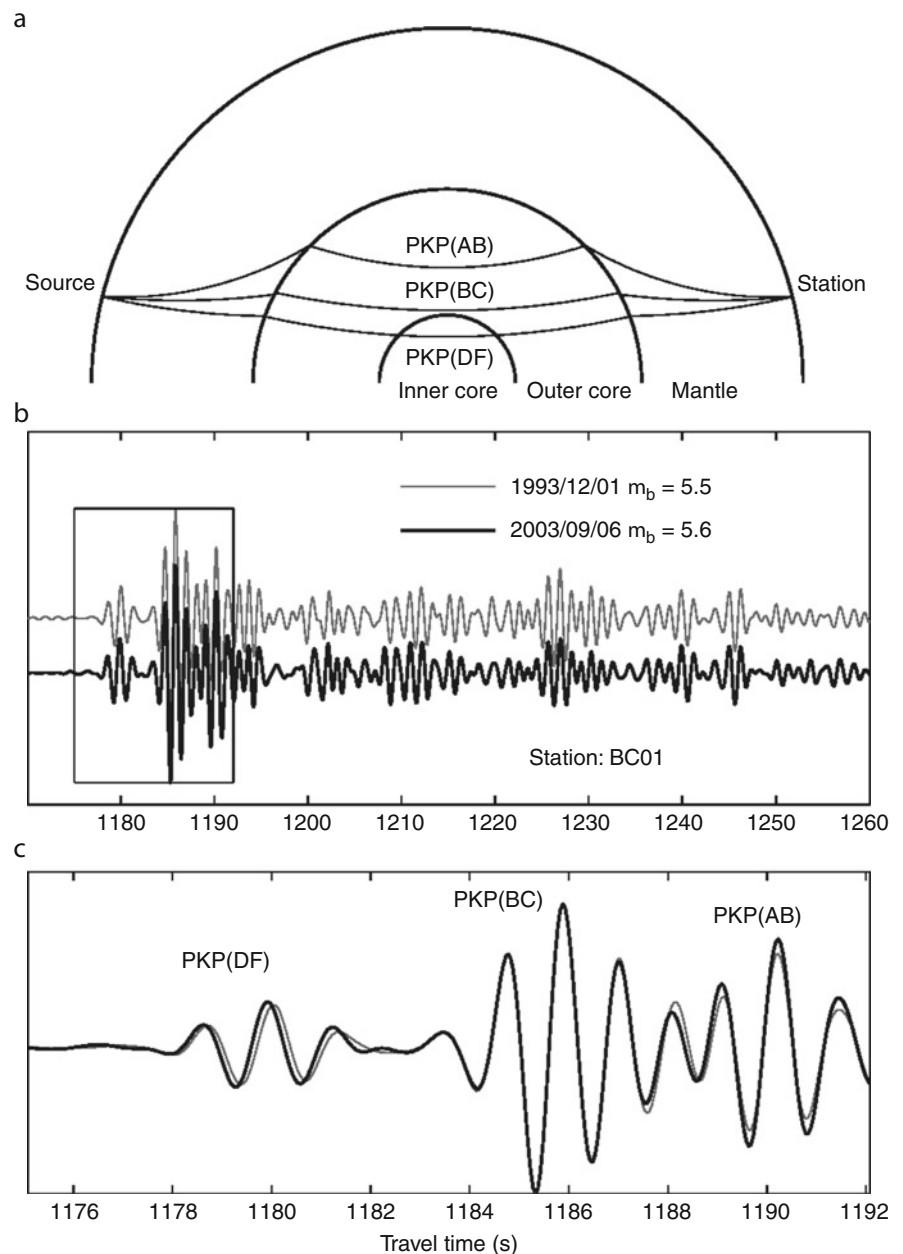
magnetic field generated in the fluid outer core would cause a differential rotation of the inner core. In a computer simulation of a three-dimensional, self-consistent geodynamo, Glatzmaier and Roberts (1995) predicted that the inner core is driven to rotate by electromagnetic coupling at a few degrees per year faster than the mantle and the crust.

Seismological Observations

Song and Richards (1996) reported first evidence for differential rotation of inner core from seismic observations. The

basic idea of their method is simple. They chose a fixed monitoring station and compare seismic waves that travel through the inner core from earthquakes at the same region but a few years apart. This is like the twinkling of a star, which is caused by disturbance of the light as it passes through the atmosphere (an analogy used by coauthor Paul Richards). The seismic waves used are short-period (about 1 s) longitudinal waves (P waves) that have propagated through the fluid core (generically called PKP waves) (Fig. 1). At the distance range of about 146–154°, there are three possible paths, i.e., PKP (DF) waves that traverse the inner core, PKP(BC) waves that turn at the bottom of the outer core, and PKP(AB) waves that

Differential Rotation of the Earth's Inner Core, Fig. 1 Ray paths of PKP waves and example of waveform doublet used to detect temporal change of travel times through the inner core, (a) Ray paths of three branches of PKP waves turning in the solid inner core (DF), the bottom of the fluid outer core (BC), and the mid-outer core (AB). (b) Highly similar waveforms recorded at one station in Alaska from a waveform doublet from South Sandwich Islands. The two events are 10 years apart, one in 1993 and the other in 2003. (c) Superimposed and enlarged PKP waveforms from the box in (b). The waves through the outer core (BC and AB) are aligned, but the wave through the inner core (DF) shows a small time shift (about 0.1 s). (From Zhang et al. 2005)



turn at the middle of the outer core. The basic data used in detecting the inner core rotation is the time difference between the PKP(BC) and PKP(DF) arrivals, abbreviated as the BC-DF time. Because these two waves travel very closely together in the crust and mantle and in most of the fluid core, the differential time BC-DF is relatively insensitive to uncertainty in source location and to three-dimensional heterogeneities of the crust and mantle along the ray paths. To remove influence of small difference in earthquake location and epicentral distance, the differential time residual is formed by subtracting the observed BC-DF time from the predicted difference for a standard earth model. In addition, residuals of the differential times between PKP(AB) and PKP(BC), which pass through the fluid core only, are used to examine possible systematic biases in source locations.

Song and Richards found that BC-DF residuals along certain paths have changed systematically with time over a few decades. In particular, the seismic waves that emanate from earthquakes in South Sandwich Islands in the South Atlantic travel through the inner core and reach College, Alaska seismic station. They have sped up systematically by about 0.3 s over 28 year-year time interval from 1967 to 1995. They interpreted the temporal change as evidence of an inner core rotation.

The estimate of the rotation rate depends on the underlying aspherical structure of the inner core that results the temporal change from a rotation. The inner core is known to be very anisotropic (i.e., seismic waves traveling through the inner core along different directions have different speeds) (Morelli et al. 1986; Woodhouse et al. 1986; see also review by Song 1997; Tromp 2001). The anisotropy is roughly axis-symmetric with the P-wave speed faster by some 3% along the Earth's rotation axis than along the equatorial plane. However, the anisotropy changes significantly laterally and with depth. The observed temporal change in BC-DF residuals was first interpreted as a change of the orientation of the fast axis of the inner core anisotropy, which yielded a rate of 1.1° per year (Song and Richards 1996), but it was later and preferably interpreted as a shift of lateral velocity gradient in the inner core caused by the inner core rotation (Creager 1997; Song 2000), which yielded a rate ranging from $0.2\text{--}0.3$ to $0.3\text{--}1.1^\circ$ per year.

Subsequent studies provided further support and most estimates of the rotation rate are a few tenths of a degree per year faster than the rotation of the Earth (a super-rotation) (see reviews by Tromp 2001; Song 2003; and Souriau et al. 2003). The methods used include BC-DF times for additional paths (Ovchinnikov et al. 1998; Song and Li 2000; Xu and Song 2003), inner core scattering waves (Vidale et al. 2000), so-called earthquake waveform doublets (Li and Richards 2003; Zhang et al. 2005; see below), and normal modes (Laske and Masters 2003). However, whether the inner core

rotation is real has also been hotly debated because of potential biases such as earthquake mislocation (e.g., Poupinet et al. 2000), failure to detect the motion (Souriau 1998), and discrepancy in the inferred rotation rate (e.g., Creager 1997; Vidale et al. 2000; Laske and Masters 2003).

The most important source of errors is systematic event mislocation. Because global stations used to locate the earthquakes are not exactly the same in different years, the temporal variation could potentially be an artifact of the changes of the station distributions (Poupinet et al. 2000; Souriau et al. 2003). To resolve the debate, the two sides (Dr. Poupinet and the author) worked directly in 2004–2007. A thorough presentation of the collaborative work was published (Sun et al. 2006). The basic conclusion is that earthquake mislocation is too small to explain the observed time shifts in the data, and that inner core rotation is still the best explanation.

Perhaps the strongest support comes from studies of earthquake waveform doublets (Zhang et al. 2005). A waveform doublet is a pair of earthquakes occurring at different times but at essentially the same spatial position, as evidenced by their highly similar waveforms at each station recording both events (Poupinet et al. 1984). Zhang et al. (2005) observed that the waves that traveled through the area outside the inner core, in the crust, the mantle, and the outer core are all the same. Only when they travel through the inner core at different times are they different (Fig. 1). The changes are two folds: first, and most prominently, the PKP(DF) waves (passing through the inner core) move faster in a systematic pattern over time, by about one tenth of a second per decade; second, the PKP(DF) waves themselves change in shape over time, an independent signature for motion of the inner core. The fact that the wave shapes are similar when they emanate from the source region allows them to measure precisely the small time shifts and to pin down precisely where the changes took place. The temporal change of BC-DF differential times for the South Sandwich Islands to College, Alaska path is constrained to be 0.0092 s/year with standard deviation of 0.0004 s/year. The best estimate of the rotation rate is $0.3\text{--}0.5^\circ$ per year faster than the Earth's mantle. Subsequent studies of earthquake doublets suggest that the waves bouncing off the inner core boundary also shows temporal variation (Wen 2006; Cao et al. 2007; Song and Dai 2008).

Concluding Remarks

The discovery of the inner core super-rotation has attracted attention of the academia and popular media. It has implications for understanding the geodynamo and angular momentum transfer in the interior of the Earth (Buffett and Glatzmaier 2000). The origin of the Earth's magnetic field has been regarded as one of the great unsolved problems in

modern physics. The observation of the inner core rotation provides a unique observational constraint on the geodynamo at the center of the earth.

The existence of the differential rotation of the inner core has now generally, although not universally, been accepted. However, major questions remain. What is the acceptable range of the rotation rate? Is the inner core rotation variable? Does the rotation change direction? Does the inner core oscillate within a certain range? The inner core possesses large hemispheric scale variations (Tanaka and Hamaguchi 1997; Niu and Wen 2001). This is difficult to reconcile with an inner core of a constant rotation, which may be expected to average out the lateral varying structure of the inner core over geological time as the inner core grows from the crystallization of liquid iron. If the rotation is variable or oscillating, what is the time scale? What is the driving force? How does it relate to the geodynamo processes? Recently, Lindner et al. (2010) proposed a nonparametric modeling method, which is able to separate mantle structure, inner core structure, and inner core motion. The study suggests an average rate of inner-core rotation of about 0.39° per year and that the rotation has accelerated from about 0.24° to about 0.56° per year within the last 55 years. The speed of the differential rotation, averaging about 10 km per year at the inner core equator, is by 50,000 times the fastest motion of the tectonic plates (20 cm per year) at the Earth's surface. The minimum torque acting on the inner core is estimated 1.19×10^{16} N m, which could easily result from the imbalance of much larger electromagnetic and gravitational torques. With the continuing accumulation of high quality seismic data, I believe we'll reach a new understanding on the inner core motion within next 50 years.

Cross-References

- Body Waves
- Core Dynamo
- Core-Mantle Coupling
- Earth Rotation
- Earth, Density Distribution
- Earth's Structure, Core
- Earth's Structure, Global
- Earthquakes, Location Techniques
- Geodynamics
- Geomagnetic Field, Global Pattern
- Geomagnetic Field, Theory
- Magnetic Anisotropy
- Mantle Convection
- Propagation of Elastic Waves: Fundamentals
- Radioactivity in Earth's Core
- Seismic Properties of Rocks

Bibliography

- Buffett BA, Glatzmaier GA (2000) Gravitational braking of inner-core rotation in geodynamo simulations. *Geophys Res Lett* 27:3125–3128
- Cao AM, Masson Y, Romanowicz B (2007) Short wave-length topography on the inner-core boundary. *Proc Natl Acad Sci U S A* 104(1):31–35
- Creager KC (1997) Inner core rotation rate from small-scale heterogeneity and time-varying travel times. *Science* 278:1284–1288
- Glatzmaier GA, Roberts PH (1995) A three-dimensional convective dynamo solution with rotating and finitely conducting inner core and mantle. *Phys Earth Planet Inter* 91:63–75
- Gubbins D (1981) Rotation of the inner core. *J Geophys Res* 86:11695–11699
- Laske G, Masters G (2003) The Earth's free oscillations and the differential rotation of the inner core. In: Dehant V et al (eds) *Earth's core: dynamics, structure, rotation*. Geodynamics series, vol 31. American Geophysical Union, Washington, DC, pp 5–21
- Li A, Richards PG (2003) Using earthquake doublets to study inner core rotation and seismicity catalog precision. *Geochem Geophys Geosyst* 4(9):1072. <https://doi.org/10.1029/2002GC000379>
- Lindner D, Song X, Ma P, Christensen DH (2010) Inner core rotation and its variability from nonparametric modeling. *J Geophys Res* 115: B04307. <https://doi.org/10.1029/2009JB006294>
- Morelli A, Dziewonski AM, Woodhouse JH (1986) Anisotropy of the inner core inferred from PKIKP travel times. *Geophys Res Lett* 13:1545–1548
- Niu FL, Wen LX (2001) Hemispherical variations in seismic velocity at the top of the Earth's inner core. *Nature* 410:1081–1084
- Ovchinnikov VM, Adushkin VV, An VA (1998) On the velocity of differential rotation of the Earth's inner core. *Dokl Akad Nauk* 362:683–686
- Poupinet G, Ellsworth WL, Frechet J (1984) Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras Fault, California. *J Geophys Res* 89:5719–5731
- Poupinet G, Souriau A, Coutant O (2000) The existence of an inner core super-rotation questioned by teleseismic doublets. *Phys Earth Planet Inter* 118:77–88
- Song XD (1997) Anisotropy of the earth's inner core. *Rev Geophys* 35:297–313
- Song XD (2000) Joint inversion for inner core rotation, inner core anisotropy, and mantle heterogeneity. *J Geophys Res* 105:7931–7943
- Song XD (2003) Three-dimensional structure and differential rotation of the inner core. In: Dehant VM, Creager KC, Zatman S, Karato S (eds) *Earth core: dynamics, structure, rotation*. Geodynamics series, vol 31. American Geophysical Union, Washington, DC, pp 45–63
- Song XD, Dai W (2008) Topography of Earth's inner core boundary from high-quality waveform doublets. *Geophys J Int* 175:386–399. <https://doi.org/10.1111/j.1365-246X.2008.03909.x>
- Song XD, Li AY (2000) Support for differential inner core superrotation from earthquakes in Alaska recorded at South Pole station. *J Geophys Res* 105:623–630
- Song XD, Richards PG (1996) Seismological evidence for differential rotation of the Earth's inner core. *Nature* 382:221–224
- Souriau A (1998) New seismological constraints on differential rotation of the inner core from Novaya Zemlya events recorded at DRV, Antarctica. *Geophys J Int* 134:F1–F5
- Souriau A, Garcia R, Poupinet G (2003) The seismological picture of the inner core: structure and rotation. *Compt Rendus Geosci* 335:51–63
- Sun XL, Poupinet G, Song XD (2006) Examination of systematic mislocation of South Sandwich Islands earthquakes using station pairs: implications for inner core rotation. *J Geophys Res* 111:B11305. <https://doi.org/10.1029/2005JB004175>

- Tanaka S, Hamaguchi H (1997) Degree one heterogeneity and hemispherical variation of anisotropy in the inner core from PKP(BC)-PKP(DF) times. *J Geophys Res* 102:2925–2938
- Tromp J (2001) Inner-core anisotropy and rotation. *Annu Rev Earth Planet Sci* 29:47–69
- Vidale JE, Dodge DA, Earle PS (2000) Slow differential rotation of the Earth's inner core indicated by temporal changes in scattering. *Nature* 405:445–448
- Wen L (2006) Localized temporal change of the Earth's inner core boundary. *Science* 314:967–970
- Woodhouse JH, Giardini D, Li X-D (1986) Evidence for inner core anisotropy from free oscillations. *Geophys Res Lett* 13:1549–1552
- Xu XX, Song XD (2003) Evidence for inner core super-rotation from time-dependent differential PKP travel times observed at Beijing Seismic Network. *Geophys J Int* 152:509–514
- Zhang J, Song XD, Li YC, Richards PG, Sun XL, Waldhauser F (2005) Inner core differential motion confirmed by earthquake waveform doublets. *Science* 309:1357–1360

Earth Rotation

Harald Schuh¹ and Sigrid Böhm²

¹Department 1 “Geodesy” at Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany

²Research Division Higher Geodesy, Department of Geodesy and Geoinformation, TU Wien, Vienna, Austria

Definition

Earth Solid Earth including oceans and atmosphere.
Earth Temporal variation of the orientation and the
rotation rotation speed of the Earth.

Introduction

The rotation of the Earth or Earth rotation, respectively, specifies the spatiotemporal change of the Earth rotation vector. The direction of the Earth rotation vector corresponds to the instantaneous rotation axis of the Earth and its absolute value equals the rotation speed. The Earth’s rotation is not uniform and is given in terms of Earth orientation parameters (EOP): *precession* and *nutation* are long-term and periodic changes of the direction of the Earth rotation vector with respect to a space-fixed reference system. *Polar motion* is the variation of the direction of the Earth rotation vector with respect to an Earth-fixed reference system (qv ► “[Geodesy, Networks, and Reference Systems](#)”). Changes in the Earth rotation speed are expressed as deviations of *Universal Time 1* (UT1) from the uniform atomic time (*Universal Time Coordinated*, UTC) $dUT1 = UT1 - UTC$ or as variations in the *length of day* (LOD). The subgroup of *polar motion* and $dUT1$ or LOD is called Earth rotation parameters (ERP). Fundamental information on Earth rotation theory and observation and about the relations of Earth rotation variations with geo-

physical processes is given in the seminal works of Munk and MacDonald (1960), Lambeck (1980), and Moritz and Mueller (1987).

While the existence of *precession* was already known to the Greek astronomer Hipparchus in the second century before Christ, *nutation* was not discovered before the eighteenth century by James Bradley. Observations of *polar motion* were taken for the first time by the Bonn astronomer Friedrich Küstner at the end of the nineteenth century by measuring latitude variations. From the 1970s to the twentieth century onward, space geodetic techniques like *Very Long Baseline Interferometry* (VLBI) (qv ► “[Very Long Baseline Interferometry](#)”), *Satellite Laser Ranging* (SLR), *Lunar Laser Ranging* (LLR), *Doppler Orbitography and Radio-positioning Integrated by Satellite* (DORIS), and the *Global Navigation Satellite Systems* (GNSS) like the *Global Positioning System* (GPS) (qv ► “[GPS, Data Acquisition, and Analysis](#)”) have been employed in addition to astronomical methods. Approximately since the year 2000, the latter are no longer routinely applied. Nowadays, the achievable accuracies of the measurements are better than 2×10^{-4} arcsec or less than 0.6 cm, if projected to the Earth’s surface. EOP can be determined from space geodetic observations within the respective parameter estimation process. The transformation between Earth-fixed and space-fixed system and thus the EOP are thereby components of the technique’s observation equation and can be solved for as unknowns. The most precise technique to observe *polar motion* is GNSS, whereas *precession/nutation* and $dUT1$ can be measured directly only by VLBI. Due to the correlation of *nutation* and $dUT1$ with the orbital elements of the satellites, satellite techniques are only sensitive to the time derivation of those parameters, i.e., *nutation rates* and LOD. Promising new devices for the observation of high-frequency variations of the instantaneous Earth rotation vector are large ring laser gyroscopes. These instruments can access absolute rotation by measuring the beat frequency of two laser beams rotating in opposite direction, the *Sagnac frequency* (Schreiber et al. 2004).

Information about the long-term behavior of the EOP is obtained from historical records of lunar and solar eclipses and from the analysis of sedimentary deposition. Evidence for a secular increase in LOD can be found for instance in paleontological studies of coral growth rate.

The observed EOP show a wide spectrum of variations. Its interpretation allows for drawing valuable conclusions about the structure of the Earth and the dynamical features of the Earth's interior, the atmosphere, hydrosphere, and cryosphere. Even anthropogenic influences on Earth rotation such as the impact of mass transfer due to an increased CO₂ emission are subject of scientific investigations (de Viron et al. 2002).

Medium and long-period as well as long-term mass displacements affect Earth rotation (cf. e.g., Gross (2007) or Seitz and Schuh (2010) for a comprehensive compilation and further references). According to the conservation of total angular momentum of the Earth in short terms, the rotation of the solid Earth undergoes variations, which are mirror image to changes in atmosphere and oceans. Any mass variation in one or more components of the system Earth that changes the Earth inertia tensor leads to a corresponding variation in the Earth rotation. One can also think of a mass movement which does not change the inertia tensor, because once a mass element moved away it is replaced by another, like in a circular ocean current. Such a mass transport causes a motion relative to the considered reference frame, which affects Earth rotation as well. In which way changes of the inertia tensor and relative motions can be related to variations in Earth rotation is shown in the successive section.

Mathematical Formulation of Earth Rotation

The dynamical equation of motion of a rotating body with respect to a space-fixed system is given by

$$\mathbf{L} = \frac{d\mathbf{H}}{dt} \quad (1)$$

relating the torque $\mathbf{L}$ acting on the body to the temporal change of its angular momentum $\mathbf{H}$. This is the basic relation for the development of (Newtonian) *precession/nutation* theories, since it describes the motion of the rotating body in a space-fixed reference system. To characterize *polar motion* and changes of the rotation rate of a rotating body, Eq. 1 has to be referred to body-fixed axes, rotating with the angular velocity $\boldsymbol{\omega}$:

$$\mathbf{L} = \frac{d\mathbf{H}}{dt} + \boldsymbol{\omega} \times \mathbf{H} \quad (2)$$

This is one form of *Euler's dynamical equations* for rigid body rotation referred to a body-fixed coordinate system. The angular momentum $\mathbf{H}$ can be expressed as the product of the tensor of inertia $\mathbf{I}$ and the angular velocity vector $\boldsymbol{\omega}$. The inertia tensor is a symmetric matrix containing the moments of inertia and the products of inertia of a rotating body and thus characterizing the mass distribution in the body. Since single particles do not move with respect to the body-fixed system, this tensor is invariant in the case of a rigid body. For a nonrigid (deformable) body, the inertia tensor becomes time variable, and the particles can move with respect to the body frame, thus allowing for relative motions which introduce relative angular momentum. The angular momentum $\mathbf{H}$ of a rotating deformable body is then written as

$$\mathbf{H} = \mathbf{I}\boldsymbol{\omega} + \mathbf{h} \quad (3)$$

with $\mathbf{h}$ denoting relative angular momentum. The first summand is often referred to as *mass* or *matter term*, and the second is called *motion term*. As the Earth is a nonrigid body, the equation of rotational motion 2 has to be extended considering the above stated differences to the motion of a rigid body, leading to

$$\mathbf{L} = \frac{d}{dt}(\mathbf{I}\boldsymbol{\omega} + \mathbf{h}) + \boldsymbol{\omega} \times (\mathbf{I}\boldsymbol{\omega} + \mathbf{h}) \quad (4)$$

These are the *Euler-Liouville equations* or simply *Liouville equations*. The deviations from uniform, rigid rotation are formulated as follows for the rotation vector $\boldsymbol{\omega}$:

$$\omega_1 = \Omega m_1, \omega_2 = \Omega m_2, \omega_3 = \Omega(1 + m_3) \quad (5)$$

where Ω is the mean angular velocity of the Earth. The m_i are small dimensionless quantities describing the excursions of the rotation vector from its uniform rotation due to time-variable changes in the mass distribution of the Earth system and relative particle motion. Assuming the body-fixed axes to be principal axes of inertia and the body to be rotationally symmetric, these mass changes can be taken into consideration by composing the tensor of inertia from constant and time-variable parts:

$$\begin{aligned} I_{11} &= A + \Delta I_{11} & I_{22} &= A + \Delta I_{22} & I_{33} &= C + \Delta I_{33} \\ I_{ij} &= \Delta I_{ij}, & i &\neq j \end{aligned} \quad (6)$$

where the constant parts are the polar moment of inertia C , and the equatorial moment of inertia A of the undeformed Earth and the time-variable components are the quantities ΔI_{ij} . If second-order terms are neglected and the relations 5 and 6 are introduced, the linearized equations of motion can be rewritten as

$$\begin{aligned}
m_1 - \frac{1}{\sigma_E} \dot{m}_2 &= \psi_1, \\
m_2 + \frac{1}{\sigma_E} \dot{m}_1 &= \psi_2, \\
m_3 &= \psi_3 + \text{const.}
\end{aligned} \tag{7}$$

with $\sigma_E = \left(\frac{C-A}{A}\right)\Omega$ being the *Euler frequency*, which would be the frequency of the resonance of a rigid Earth, corresponding to a circular motion of the rotation axis with a period of approximately 305 days. If the Earth is considered to be an elastic or at least deformable body, the *Euler frequency* is replaced by the complex *Chandler frequency* $\sigma_C = \frac{2\pi}{T_C} \left(1 + \frac{i}{2Q}\right)$, with $T_C \sim 433$ days denoting the observed period of the *Chandler wobble* and $Q = 30 \dots 200$ being a dimensionless dissipation factor (see next section for a brief discussion of the *Chandler wobble*). The ψ_i are called *excitation functions*. They contain the changes of the tensor of inertia and the relative angular momenta:

$$\begin{aligned}
\psi_1 &= \frac{1}{\Omega^2(C-A)} [\Omega^2 \Delta I_{13} + \Omega \dot{\Delta I}_{23} + \Omega h_1 + \dot{h}_2 - L_2] \\
\psi_2 &= \frac{1}{\Omega^2(C-A)} [\Omega^2 \Delta I_{23} - \Omega \dot{\Delta I}_{13} + \Omega h_2 - \dot{h}_1 + L_1] \\
\psi_3 &= \frac{-1}{\Omega^2 C} \left[\Omega^2 \Delta I_{33} + \Omega h_3 - \Omega \int_0^t L_3 dt \right]
\end{aligned} \tag{8}$$

The first two equations of 7 for m_1 and m_2 express *polar motion*, while the third equation for m_3 describes the magnitude of the rotation speed and hence LOD variations. Provided that all the time-variable parts of the *excitation functions* (changes of the inertia tensor ΔI_{ij} , relative angular momenta h_i , and external torques L_i) are introduced as known quantities from models or observations, the *Euler-Liouville equations* can be solved for ω . Earth rotation variations can be thus calculated or predicted, respectively, from observed or modeled changes in the components of the system Earth, such as atmosphere or oceans. In general, relative motions (relative angular momenta) contribute more to LOD variations while the major effect on *polar motion* comes from alterations of the inertia tensor. The *excitation functions* are in practice mostly replaced by so-called angular momentum functions, introduced by Barnes et al. (1983). Elaborate explications of the basic Earth rotation formalism can be found in the fundamental books mentioned in the introduction.

Definition and Observation of Earth Orientation Parameters

The elements which are used to perform the transformation between a space-fixed and an Earth-fixed reference system are

commonly referred to as EOP. The realizations of such reference systems by assigning coordinates to selected celestial objects or terrestrial sites, respectively, are called reference frames. The definition and maintenance of the reference frames is one major task of the *International Earth Rotation and Reference Systems Service* (IERS). Conventional reference frames are the space-fixed *International Celestial Reference Frame* (ICRF) and the Earth-fixed *International Terrestrial Reference Frame* (ITRF). The current version of the ICRF is its third realization, the ICRF3 (Charlot 2018). The ITRF2014 (Altamimi et al. 2016) has been adopted as the official terrestrial reference frame and will be followed by the ITRF2020 which is currently under construction. The definition of the EOP actually depends on the kind of the applied transformation method. The IERS recommends the transformation according to the IAU (*International Astronomical Union*) 2000 and 2006 Resolutions (Petit and Luzum 2010). The transformation procedure is defined from the *Geocentric Celestial Reference System* (GCRS) to the *International Terrestrial Reference System* (ITRS) or vice versa. The ICRS is not a geocentric system – its origin is located in the barycenter of the solar system. To transfer from a barycentric system to a geocentric system, effects, which are not directly related to Earth rotation, like aberration and parallaxes have to be taken into account. The orientation of the ICRS however corresponds to the orientation of the GCRS. The transition from the GCRS to the ITRS is described as a sequence of time-dependent rotation matrices:

$$[\text{ITRS}] = \mathbf{W}(t) \cdot \mathbf{R}(t) \cdot \mathbf{Q}(t) \cdot [\text{GCRS}] \tag{9}$$

$\mathbf{W}(t)$ (with “W” for wobble) designates the *polar motion* matrix. It contains the coordinates x_p and y_p of the reference pole CIP (*Celestial Intermediate Pole*) in the Earth-fixed system and the angle s' , which provides the position of the TIO (*Terrestrial Intermediate Origin*) on the equator of the CIP. The TIO and the *Celestial Intermediate Origin* (CIO) realize an instantaneous prime meridian in the respective system. These terms are part of the transformation concept using the *non-rotating origin* (NRO), which replaced the older transformation scheme with ecliptic (plane of the Earth’s orbit) and equator. The rotation between TIO and CIO is performed with $\mathbf{R}(t)$ using a quantity named *Earth Rotation Angle* (ERA). The ERA is directly related to UT1, which is relevant for Earth rotation research. *Precession* and *nutation* are represented by the matrix $\mathbf{Q}(t)$. It comprises rotations around the angles X and Y , the coordinates of the CIP in the celestial system and around the angle s , which locates the CIO on the equator of the CIP. In case of using the transformation according to IAU 2000 resolutions, the five quantities $\{x_p, y_p, dUT1, X, Y\}$ therefore represent the EOP. If

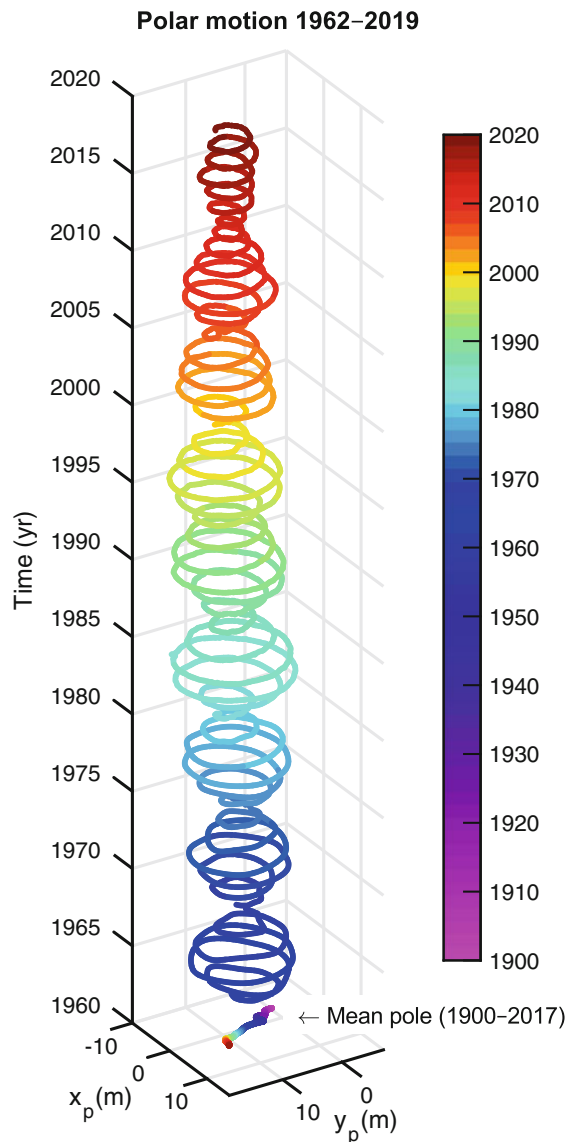
the older transformation concept based on ecliptic and equator is applied, X and Y are replaced by $\Delta\epsilon$ and $\Delta\psi$, *nutations* in obliquity and longitude. The before cited *Celestial Intermediate Pole* is the reference pole measurements of space geodetic techniques are related to. The CIP thus defines the observed axis. This is a pure convention, realized by an accordingly adapted *precession-nutation* theory. The direction toward the CIP does not correspond to any physically defined axis, like the rotation axis, the figure axis, or the angular momentum axis; nevertheless, it is possible to mathematically connect it to each of those axes. Casually, it is often said that Earth rotation observations represent the motion of the Earth rotation axis. Regarding the measurements of space geodetic techniques, this is actually not entirely correct, since they are not sensitive to the instantaneous Earth rotation vector but to the complete rotation matrix only. The CIP defines an intermediate pole, separating the motion of the pole of the ITRS in the GCRS into a celestial part and a terrestrial part. The celestial part (*precession* and *nutations*, $\{X, Y\}$) comprises all motions with periods greater than 2 days, as seen from space. This is equivalent to frequencies between -0.5 and $+0.5$ cycles per sidereal day (cpsd). With the minus sign labeling retrograde motions (opposite to the sense of Earth rotation) and the plus sign labeling prograde motions (in the sense of Earth rotation), all motions outside of the retrograde diurnal band in the Earth-fixed system, i.e., frequencies below -1.5 and above $+0.5$ cpsd, are allocated to the terrestrial part (*polar motion*, $\{x_p, y_p\}$).

The celestial motions, *precession* and *nutations*, are long-term and periodic changes of the direction of the Earth rotation axis or actually of the CIP axis, with respect to a space-fixed reference system, which is realized, for example, by the positions of extragalactic radio sources observed by VLBI. Due to *precession*, the Earth axis moves with respect to the space-fixed system on a cone with an aperture of 23.5° , corresponding to the angle between the Earth's equator plane and the ecliptic. The revolution period amounts to approximately 25,800 years. This time interval is also called Platonic year. The tidal forces of moon and sun are responsible for this steady motion. Since the Earth is not a sphere but can be characterized as an oblate spheroid (qv ► [“Geodesy, Figure of the Earth”](#)) and its rotation axis is inclined by 23.5° with respect to the ecliptic normal, the gravitational torques force the equatorial plane into the ecliptic. Because of its rotation, the Earth acts like a gyroscope and swerves by moving on the abovementioned cone, whose surface can be calculated very precisely.

The smaller periodic change in the direction of the Earth rotation axis that is superimposed to *precession* is called *nutations*. This comprises motions of the Earth rotation axis with respect to the space-fixed system with periods from a few days to 18.6 years, caused by gravitational influences of sun, moon, and the planets of our solar system. *Precession* and

nutations can be modeled and predicted precisely using time-dependent harmonic series expansions. The arguments of the individual harmonic terms are thereby calculated from combinations of five fundamental astronomical arguments. The currently most precise *precession-nutations* model adopted by IAU Resolutions (2000, 2006) is IAU 2006/2000A. In this model, effects of planets, ocean tides, mantle anelasticity, and electromagnetic coupling mechanisms between core and mantle as well as between inner and outer core are considered (Mathews et al. 2002). As for the planets, their direct effect on Earth as well as the indirect effect of the sun acting on a planet, which also causes subtle *nutations* terms, is taken into account. Remaining parts of the axis motion with respect to the space-fixed system that are not covered by the *precession-nutations* model can be measured by means of VLBI and are provided by the IERS as so-called celestial pole offsets. These residuals originate from still deficiently modeled and unpredictable effects, like the *free core nutations* (FCN). The FCN is a proper mode of the Earth, caused by a misalignment of the rotation axes of mantle and core. Whereas first theoretical estimations indicated a period of about 460 days for this mode, VLBI measurements show that the FCN period is more likely to be around 430 days with highly variable amplitude.

The terrestrial part of the change of the direction of the Earth axis or rather the CIP axis is designated *polar motion* (Fig. 1). *Polar motion* has an order of magnitude of several meters and is expressed in a two-dimensional coordinate system by the *pole coordinates* x_p and y_p . According to the definition of the IERS and its precedent organizations, the x -axis is oriented in the direction of the Greenwich meridian, and the y -axis is oriented positively toward 90° west longitude. Already in 1765, the Swiss mathematician Leonhard Euler calculated a circular motion of the pole (of an Earth at that time assumed to be rigid) with a period of around 304 days. Today, we know that *polar motion* is mainly composed from an annual variation and the *Chandler oscillation* or *Chandler wobble*, named after Seth Carlo Chandler who first detected a period of approximately 14 months when analyzing *polar motion* observations at the end of the nineteenth century. The *Chandler wobble* is another proper motion of the Earth with strongly varying amplitude. More precisely, this nearly circular motion is a damped oscillation, which would have vanished due to friction in the Earth's interior after a few decades if there was not a constantly revolving excitation. Although the *Chandler oscillation* is known for more than 100 years, its underlying excitation mechanism is still under investigation. Today, there is broad consensus that the necessary excitation energy emerges from irregular processes in the atmosphere-ocean system (Gross 2000). A radical change in the phase of the *Chandler wobble* around 1925 could also not yet fully be explained. Variations of the Earth's magnetic field are nowadays quoted as potential causes for the phase variations. The interference of the



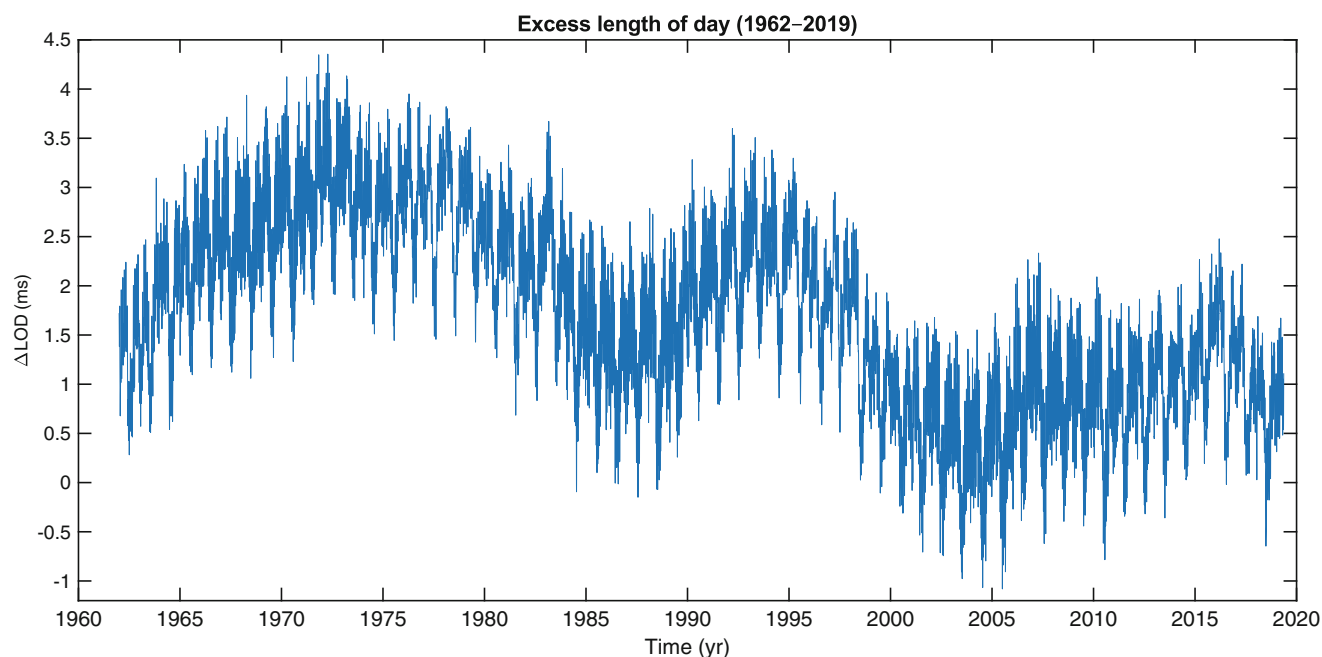
Earth Rotation, Fig. 1 Polar motion from the combined EOP series C04 14 of the IERS (1962–2019, daily resolution) and position of the mean pole from 1900 to 2017

Chandler wobble and the annual motion leads to a beat-like rising and ebbing of the *polar motion* amplitude to a maximum of 9 m with a beat period of about 6.3 years. The *polar motion* spectrum below 1 year is dominated by irregular variations appearing in 3 and 5 months intervals. The most important short-period variations are due to the ocean tides caused by sun and moon with essentially diurnal and semidiurnal periods (Karbon et al. 2019), albeit the total effect is only about 1/100 of the *Chandler wobble*. Analysis of long-term *polar motion* shows a distinct variation of approximately 11 years. It is usually associated with processes in the Earth's interior but could also be connected to a corresponding cycle of the solar activity. In addition there are other, decadal variations with periods around 30 years and between 70 and

80 years. These are assumed to be caused by geodynamic couplings between Earth's core and mantle. Recently, Adhikari and Ivins (2016) identified terrestrial water storage variability as the source of decadal-like changes in observed *polar motion* during their study period 2003–2015. Secular *polar motion*, i.e., the long-term linear trend in the position of the pole, amounts to about 10.5 cm/year toward Labrador, Canada (e.g., Schuh et al. 2001 or Adhikari et al. 2018). This effect is supposed to be predominantly due to melting of the polar ice masses and postglacial rebound, but also long-term mass movement due to mantle convection is suspected to play a role.

For an extensive review on precession, nutation, and wobble of the Earth, refer to Dehant and Mathews (2015).

LOD is used to express the speed of Earth rotation alternatively to the difference $dUT1$ between UT1, which is directly connected to the Earth rotation speed, and the regular atomic time UTC. The parameter, which is usually quoted to characterize changes in LOD, is actually ΔLOD , the so-called excess LOD which represents the deviation of the effective LOD from the nominal value of 86,400 s (Fig. 2). Variations of LOD can be assigned to different period ranges. In the short-period range with periods from around 1 day and half a day, the strongest influence emerges from the ocean tides caused by sun and moon (Karbon et al. 2018). Hydrodynamic models or altimetry (qv ► “Geoid Determination, Theory and Principles”) provide variations of ocean heights and velocities of the water masses, from which global changes of angular momentum of the world oceans are calculated. These are subsequently used to estimate the oceanic influence on the Earth rotation parameters. In this way high-frequency ERP variations can be predicted from tide and circulation models, and they can already be observed by the high-precision space geodetic techniques mentioned above. In the range of a few weeks to months, periods of ~ 14 and ~ 28 days, due to solid Earth tides (qv ► “Earth Tides”), are dominant. In addition, there are other strong variations between 40 and 90 days basically excited by zonal winds. Seasonal variations due to changes in the angular momentum of the atmosphere show semiannual and annual periods. The variation of LOD with a period of approximately 1 year is predominantly due to annually changing wind patterns. A definite amplification of the annual variation every 4–6 years is associated with large-scale climate signals related to the El Niño phenomenon by which an ocean circulation with a very characteristic pattern in the Southern Pacific is connected to variations of the meteorological parameters. The existence of decadal fluctuations of LOD at the ms level suggests an internal coupling between the Earth's core and mantle exhibiting torques on the order of 10^{18} Nm. Four coupling mechanisms and associated torques are commonly identified, the viscous, topographic, gravitational, and electromagnetic torques. Except for the viscous torque which seems too weak, any of the other



Earth Rotation, Fig. 2 Excess length of day from the combined EOP series C04 14 of the IERS (1962–2019, daily resolution)

torques, would be capable of explaining these LOD variations (qv ► “Core-Mantle Coupling”). Because of tidal friction and long-term mass variations, a secular prolongation of the day by about 1.8 ms in 100 years is observed as well (Stephenson et al. 2016). This is, of course, an average long-term trend deduced from eclipse data over the last 2,700 years. Besides the secular trend, there is also evidence of LOD fluctuations on a timescale of centuries and some indication of an around 1500 years oscillation.

IERS International Earth Rotation and Reference Systems Service

The IERS is an international service for the determination of EOP and their dissemination to interested users. The definition and realization of reference systems for geodesy and astronomy can be regarded as its superior task. According to the terms of reference (Dick and Thaller 2018), the primary objectives of the IERS are to serve the astronomical, geodetic, and geophysical communities by providing the following:

- The International Celestial Reference System (ICRS) and its realization, the International Celestial Reference Frame (ICRF).
- The International Terrestrial Reference System (ITRS) and its realization, the International Terrestrial Reference Frame (ITRF).

- EOP required to study Earth orientation variations and to transform between the ICRF and the ITRF.
- Geophysical data to interpret time/space variations in the ICRF, ITRF, or EOP, and to model such variations.
- Standards, constants, and models (i.e., conventions) encouraging international adherence.

In addition, the IERS collects, archives, and distributes several products such as reference frames, monthly Earth orientation data, daily rapid service estimates of near real-time Earth orientation data, and their predictions. The IERS also announces the differences between astronomical and civil time for time distribution by radio stations and leap seconds which – if necessary to keep the differences $dUT1$ smaller than 0.9 s – are added at midnight of July 31 or December 31. Further products are related to global geophysical fluids such as mass and angular momentum distribution, annual reports and technical notes on conventions and other topics, and long-term Earth orientation information. The IERS began operation on January 1, 1988, as common institution of the IAU and the *International Union of Geodesy and Geophysics* (IUGG) and replaced thereby the former *International Polar Motion Service* (IPMS) and the Earth rotation section of the *Bureau International de l’Heure* (BIH). The service consists, among other parts of a Central Bureau, combination centers of the space geodetic techniques and several product centers, for example, for the collection of data about geophysical influences on Earth rotation (*Global Geophysical Fluids Center*, GGFC). These are atmospheric

and oceanic variations, hydrodynamic effects like groundwater variations, and processes in the Earth's interior which lead to changes in the rotational behavior of the Earth. The Central Bureau is located at the BKG (*Bundesamt für Kartographie und Geodäsie*) in Frankfurt am Main, Germany. Apart from regular publication of EOP, the IERS also issues guidelines (IERS Conventions, Petit and Luzum (2010)), which contain models and standards recommended for the data processing of space geodetic techniques.

Summary

Earth rotation is conventionally described by the EOP, which represent the link between a space-fixed (celestial) and an Earth-fixed (terrestrial) coordinate system. Those EOP expressing the temporal variations of the orientation of the Earth correspond to *precession/nutation* (changes of the direction of Earth rotation axis with respect to a space-fixed reference frame) and *polar motion* (wobbling of the Earth with respect to its axis). Small variations of the speed of Earth rotation are expressed by UT1 minus the uniform atomic time (UTC) or as variations in the LOD. A precise knowledge of the Earth's attitude is needed for all positioning and navigation tasks on Earth and in space. It also gives fundamental information about the interactions between the various geophysical components of system Earth and allows deriving conclusions about phenomena of celestial mechanics. The EOP are nowadays monitored by space geodetic techniques and assembled and published by the IERS.

Cross-References

- Core-Mantle Coupling
- Earth Tides
- Geodesy, Figure of the Earth
- Geodesy, Networks, and Reference Systems
- Geoid Determination, Theory and Principles
- GPS, Data Acquisition, and Analysis
- Very Long Baseline Interferometry

Bibliography

- Adhikari S, Ivins ER (2016) Climate-driven polar motion: 2003–2015. *Sci Adv* 2(4):e1501693. <https://doi.org/10.1126/sciadv.1501693>
- Adhikari S, Caron L, Steinberger B, Reager JT, Kjeldsen KK, Marzeion B, Larour E, Ivins ER (2018) What drives 20th century polar motion? *Earth Planet Sci Lett* 502:126–132. <https://doi.org/10.1016/j.epsl.2018.08.059>
- Altamimi Z, Rebischung P, Métivier L, Collilieux X (2016) ITRF2014: a new release of the International Terrestrial Reference Frame modeling non-linear station motions. *J Geophys Res* 121:6109–6131. <https://doi.org/10.1002/2016JB013098>
- Barnes RTH, Hide R, White AA, Wilson CA (1983) Atmospheric angular momentum functions, length-of-day changes and polar motion. *Proc R Soc Lond A* 387:31–73
- Charlot P (2018) The third realization of the International Celestial Reference Frame. Presentation given at the IAU General Assembly 2018, Vienna. https://www.iau.org/static/science/scientific_bodies/divisions/a/2018/Charlot.pdf. Accessed May 2019
- de Viron O, Dehant V, Goosse H, Crucifix M, Participating CMIP Modeling Groups (2002) Effect of global warming on the length-of-day. *Geophys Res Lett* 29(7):1146. <https://doi.org/10.1029/2001GL013672>
- Dehant V, Mathews PM (2015) Precession, nutation and wobble of the earth. Cambridge University Press, Cambridge. <https://doi.org/10.1017/CBO9781316136133>
- Dick WR, Thaller D (eds) (2018) IERS annual report 2017. International Earth Rotation and Reference Systems Service, Central Bureau. Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt am Main
- Gross RS (2000) The excitation of the Chandler wobble. *Geophys Res Lett* 27(15):2329–2332
- Gross RS (2007) Earth rotation variations – long period. In: Herring TA (ed) *Physical geodesy. Treatise on geophysics*, vol 3. Elsevier, Amsterdam
- IAU Resolutions (2000) http://www.iau.org/static/resolutions/IAU2000_French.pdf
- IAU Resolutions (2006) http://www.iau.org/static/resolutions/IAU2006_French.pdf
- Karbon M, Balidakis K, Belda S, Nilsson T, Hagedoorn J, Schuh H (2018) Long-term evaluation of ocean tidal variation models of polar motion and UT1. *Pure Appl Geophys* 175(5):1611–1629. Springer International Publishing AG, part of Springer Nature. <https://doi.org/10.1007/s00024-018-1866-1>
- Karbon M, Balidakis K, Belda S, Nilsson T, Hagedoorn J, Schuh H (2019) Long-term evaluation of ocean tidal variation models of polar motion and UT1. In: Braitenberg C, Rossi G, *Geodynamics and Earth Tides* Editor (eds) *Geodynamics and Earth tides observations from global to micro scale. Pageoph topical volumes*. Birkhäuser, Basel, pp 17–35, eBook ISBN 978-3-319-96277-1, Softcover ISBN 978-3-319-96276-4, Buchreihen ISSN 2504-3625. <https://doi.org/10.1007/978-3-319-96277-1>
- Lambeck K (1980) The Earth's variable rotation, geophysical causes and consequences. Cambridge University Press, Cambridge
- Mathews PM, Herring TA, Buffett BA (2002) Modeling of nutation and precession: new nutation series for nonrigid Earth, and insights into the Earth's interior. *J Geophys Res* 107(B4). <https://doi.org/10.1029/2001JB000390>
- Moritz H, Mueller II (1987) *Earth rotation: theory and observation*. Ungar, New York
- Munk WH, MacDonald GJF (1960) *The rotation of the Earth: a geophysical discussion*. Cambridge University Press, Cambridge
- Petit G, Luzum B (eds) (2010) *IERS conventions (2010)*. IERS technical note 36. Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt am Main
- Schreiber U, Velikoseltsev A, Rothacher M, Klügel T, Stedman G, Wiltshire D (2004) Direct measurement of diurnal polar motion by ring laser gyroscopes. *J Geophys Res* 109(B6). <https://doi.org/10.1029/2003JB002803>
- Schuh H, Nagel S, Seitz T (2001) Linear drift and periodic variations observed in long time series of polar motion. *J Geod* 74:701–710
- Seitz F, Schuh H (2010) Earth rotation. In: Xu G (ed) *Sciences of geodesy*. Springer, Berlin
- Stephenson FR, Morrison LV, Hohenkerk CY (2016) Measurement of the Earth's rotation: 720 BC to AD 2015. *Proc R Soc A* 472: 20160404. <https://doi.org/10.1098/rspa.2016.0404>

Earth Tides

John M. Wahr
(Deceased)

Definition

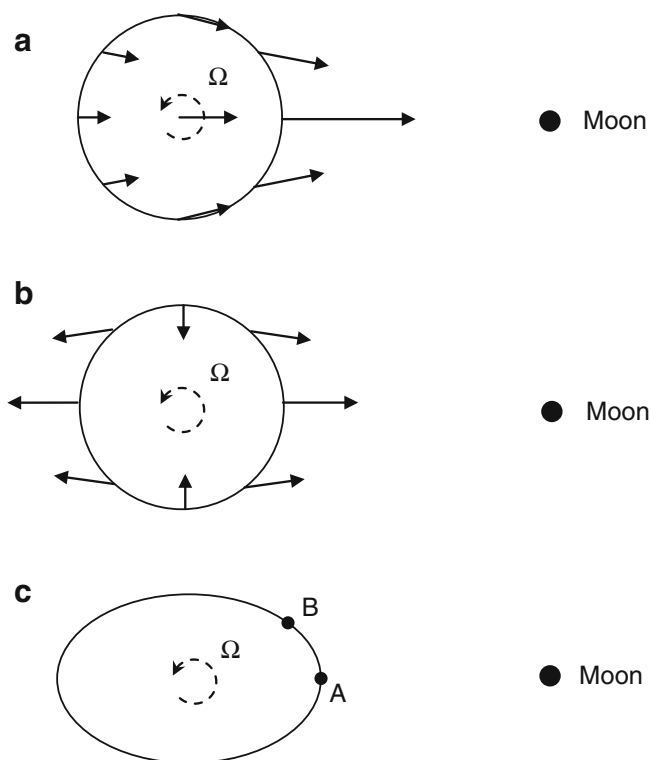
The Earth tide is the deformation of the solid Earth caused by the gravitational attraction of the Sun and moon. The most striking gravitational effects of the Sun and moon are the orbital motion of the Earth's center of mass. The moon causes the Earth to orbit the Earth–moon center of mass, and the Sun causes that center of mass to orbit the Sun.

The Tidal Force

But the Sun and moon also cause points within the Earth to be displaced relative to each other: they cause the Earth to deform. For example, the side of the Earth nearest the moon is attracted toward the moon more than is the center of the Earth. And the side farthest from the moon is attracted less than the center of the Earth. Consequently, both the far and near sides of the Earth are pulled radially outward away from the center. The regions of the Earth that are at right angles to the Earth–moon vector are pulled radially inward.

This deformation pattern is illustrated in Fig. 1. The total gravitational acceleration vectors due to the moon are shown in Fig. 1a. The vectors all point toward the moon, but are of unequal length and direction, although the differences are greatly exaggerated in the figure. The orbital acceleration is, to a high degree of accuracy, equal to the acceleration vector at the Earth's center of mass. Subtracting that vector from the other vectors in Fig. 1a results in the pattern of vectors shown in Fig. 1b. These residual vectors represent that portion of the lunar gravitational force tending to deform the Earth. The residual acceleration field multiplied by the local material density is defined as the lunar tidal force, and the deformation it induces is called the lunar tide.

Note from Fig. 1b that because of the Earth's diurnal rotation the tidal force at a fixed point in the Earth varies through two complete cycles in 1 day; a fixed point rotates through one outward bulge, then through an inward bulge, and then into the other outward bulge, all within a half a rotation period (i.e., within 12 h). This semidiurnal time dependence is split into many periodic terms with frequencies closely spaced about 2 cycles/day, due to the time variability of the orbital motion of the moon. Furthermore, because the moon is not always in the plane of the Earth's equator, there is also significant variability at frequencies



Earth Tides, Fig. 1 All three panels show the Earth as seen from above the north pole. The Earth's diurnal rotation is represented by Ω . The arrows in (a) illustrate the magnitude and direction of the gravitational acceleration toward the moon. The differences in the lengths and directions of the arrows are greatly exaggerated. The lunar tidal acceleration is defined by subtracting the acceleration vector at the Earth's center of mass from the other acceleration vectors. The resulting vectors are shown in (b). The tidal force (the local density times the acceleration) deforms the Earth into the elliptical shape shown greatly exaggerated in (c). Points A and B are used in the text to illustrate tidal strain and tilt

closely spaced about 1 cycle/day (imagine, e.g., that the Earth–moon vector is inclined at 45° to the Earth's rotation axis) and about 0 cycles/day (imagine that the moon is directly above the north pole). The solar tidal force and the solar tide are defined in a similar manner, and can also be decomposed into semidiurnal, diurnal, and long period terms.

For a quantitative description, it is useful to work with the tidal potential, denoted here as V_T and defined so that its gradient is the tidal force per unit mass. Define a nonrotating coordinate system with origin at the center of the Earth. The total lunar gravitational potential at the point $\mathbf{r}$ inside the Earth is (assuming the moon to be a point mass)

$$V = \frac{GM}{|\mathbf{r} - \mathbf{R}|} \quad (1)$$

where M is the mass of the moon, G is Newton's gravitational constant, and $\mathbf{R}$ is the position vector of the moon. Let r, θ, ϕ

be the radial distance, colatitude, and eastward longitude of the Earth-fixed point, $\mathbf{r}$, so that $\phi + \Omega t$ (where $\Omega = 1$ cycle/day is the Earth's angular velocity of rotation) is the azimuthal angle of $\mathbf{r}$ in nonrotating, inertial space. The factor $1/|\mathbf{r} - \mathbf{R}|$ can be expanded as a sum of complex spherical harmonics, $Y_n^m(\theta, \phi)$, so that

$$V = \text{Re} \left[\frac{GM}{R} \sum_{n=0}^{\infty} \sum_{m=0}^n \left[\frac{r}{R} \right]^n d_n^m Y_n^m(\theta, \phi) e^{im\Omega t} \right] \quad (2)$$

where Re denotes the real part (included in Eq. 2 so that the sum over m includes only $m \geq 0$), R is the radial coordinate of the moon, and d_n^m are coefficients that depend on the angular position of the moon relative to nonrotating inertial space. The coordinate R and the moon's angular position depend on time, due to the moon's orbital motion.

The $n = 0$ term in Eq. 2 is a spatial constant and so is physically meaningless. It can be ignored.

The $n = 1$ term represents a spatially constant force, equal to the lunar force at the Earth's center of mass. Consequently, the $n = 1$ term is removed from Eq. 2 when defining the tidal potential V_T .

Since $r \leq a$ where a is the Earth's radius, and since $a/R \approx 1/60$ for the moon (and is much smaller for the Sun), the factor $(r/R)^n$ in Eq. 2 causes the contributions to Eq. 2 to fall off rapidly with increasing n . For most purposes it is sufficient to keep only the $n = 2$ terms in V_T , so that V_T has the approximate form

$$V_T = \text{Re} \left[\left[\frac{r}{a} \right]^2 \sum_{m=0}^2 c_2^m Y_2^m(\theta, \phi) e^{im\Omega t} \right] \quad (3)$$

where $c_2^m (= GM/R(a/r)^2 d_2^m)$ are complex, time-varying coefficients that depend on the orbital coordinates of the moon.

The dominant time dependence in Eq. 3 comes from the $e^{im\Omega t}$ term. This term results in semidiurnal, diurnal, or long period tides, depending on whether m is 2, 1, or 0, respectively. The time-dependent c_2^m in Eq. 3 split the $e^{im\Omega t}$ time dependence into terms with frequencies closely spaced about $m\Omega$.

The Earth's Response

The tidal force causes ocean tides as well as Earth tides. The observed ocean tide is actually the difference between the response of the ocean and the response of the solid Earth. Ocean tides are complicated and difficult to model, for a number of reasons that need not concern us here. Open ocean tides are typically less than 1 m in height, although

ocean tides at some coastal locations can be as large as several meters peak-to-peak.

Earth tides are much easier to understand and model. The tidal force shown in Fig. 1b tends to deform the solid Earth into the elliptical shape shown greatly exaggerated in Fig. 1c. Tidal displacements of the solid Earth are typically several tens of centimeters. But, unlike ocean tides, Earth tides cannot be observed without sensitive instruments. The reason is that Earth tides cause both the ground and the observer to be displaced by the same amount.

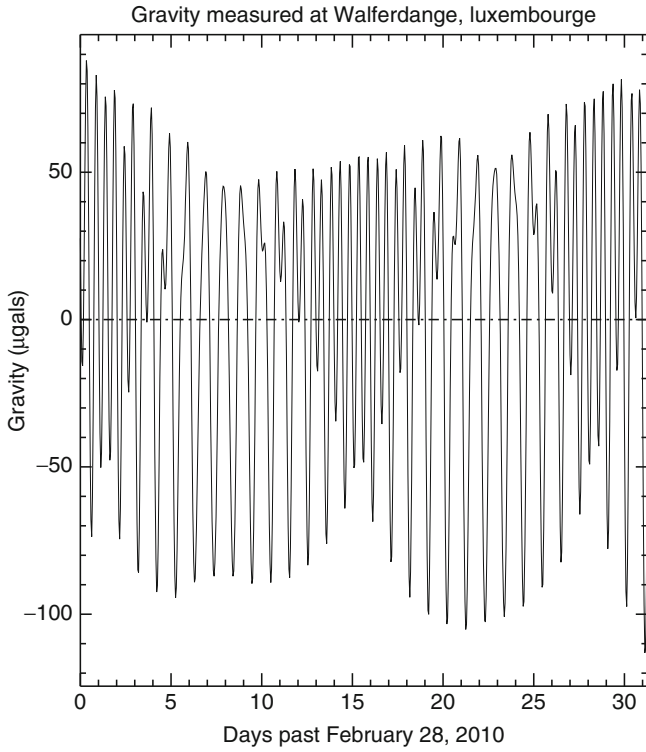
One way to detect Earth tides is to use a gravimeter. There are three contributions to the observed tidal variation in the gravitational acceleration: (1) the direct attraction of the Sun and moon, (2) the change in the Earth's gravity field due to tidal deformation within the Earth, and (3) the change in the gravitational acceleration at the gravimeter caused by the radial tidal displacement of the Earth's surface beneath the gravimeter (if the surface moves outward, gravity decreases, and vice versa). All three contributions are roughly of the same order.

The total tidal variation of the observed gravitational acceleration is on the order of 100 μgals or more (1 gal = 1 cm/s²). The unperturbed gravitational acceleration at the Earth's surface is about 10³ gals. So, tides affect gravity at about the 10⁻⁷ level. This is roughly the relative importance of tides on any physical observable, and can be traced to the fact that tidal displacements (a few tens of centimeters) are about 10⁻⁷ of the Earth's radius (about 6×10^8 cm).

A simple, idealized gravimeter is a mass hanging from a spring. The amount the spring is stretched at equilibrium is proportional to the local gravitational acceleration. So by continually monitoring the length of the spring, tidal variations in gravity can be determined. Although real gravimeters are more complicated than this, the basic idea is the same. Most instruments, in fact, do use springs.

To illustrate the effects of tides on gravity, Fig. 2 shows 1 month of hourly gravitational acceleration data as observed with a spring-type instrument at Walferdange, Luxembourg (data courtesy of Olivier Francis). The tides are the dominant signal in the data. The semidiurnal and diurnal tides are especially evident. The longer period (2 weeks and 1 month) amplitude modulations evident in the figure are caused by beating between frequencies separated by 1 cycle/13.7 days and between frequencies separated by 1 cycle/27.6 days. Long period gravity tides have small amplitudes at the latitude of Walferdange, and are hard to distinguish in Fig. 2.

One problem common to all relative gravimeters is calibration. For spring-type instruments, the proportionality



Earth Tides, Fig. 2 One month of hourly gravity data, taken with a spring-type gravimeter at Walferdange, Luxembourg. (Data courtesy of Olivier Francis). The tides are the dominant signal and are described further in the text

constant between the length of the spring and gravity depends on the spring constant, which can never be precisely known. This, for example, could cause errors in the vertical scale in Fig. 2.

Another method of detecting Earth tides is to use a strainmeter. This is an instrument that continually monitors the distance between two points on the Earth's surface, separated by anywhere from a few centimeters to a kilometer. The fact that Earth tides perturb the distance between points can be seen by considering points A and B in Fig. 1c. The distance between those points has been changed by the tidal deformation, because A has moved radially outward more than B. Also, not evident in the figure, A and B have been displaced horizontally by different amounts. Typical tidal variations in the distance between two points are on the order of 10^{-7} times the unperturbed distance.

A third way to detect tidal deformation is with tiltmeters. These instruments measure the tilt of the Earth's surface with respect to the direction of gravity. Tidal tilt of the surface can be understood from Fig. 1c by noting that the tidal bulge has affected the slope of the line between points A and B. The direction of gravity also tips during the deformation, but by a different amount. It is this difference that is observed. Typically, observed tidal tilts are a few tenths of a microradian.

Tidal deformation can also be observed using space geodetic techniques, including satellite laser ranging (SLR), very-long-baseline-interferometry (VLBI), and GPS positioning. These techniques are sensitive to displacements of the ground beneath the observing stations. In fact, one of the primary geophysical objectives of these techniques is to detect tectonic motion of the stations. The effects of Earth tides are evident in the data, but they cannot generally be determined with as much relative accuracy as can the tidal effects on surface gravity, strain, or tilt.

Satellite orbit tracking can also be used to detect tidal variations in the Earth's gravity field. Those variations affect the orbit of a satellite, and so can show up in the satellite tracking data. In fact, for lower-flying satellites the ranging data are usually more sensitive to these tidal gravity variations than they are to tidal displacements of the ground stations.

These various observable tidal effects can be conveniently described, mathematically, by using dimensionless parameters known as Love numbers, defined as follows. The deformation of a spherically symmetric, nonrotating Earth in response to a $Y_n^m(\theta, \lambda)$ applied external potential can be described with that same $Y_n^m(\theta, \lambda)$ angular dependence. Mathematically, this is because the Y_n^m separate spherically symmetric differential equations. The change in the Earth's gravitational potential at $r = a$ (the unperturbed surface of the Earth) and the radial displacement at $r = a$ are both proportional to Y_n^m , and the horizontal displacement vector at $r = a$ is proportional to ∇Y_n^m . The proportionality constants depend on n but are independent of m . Thus, the same constants are pertinent for all three Y_2^m terms in Eq. 3.

Let $\Phi(\theta, \lambda)$, $U_r(\theta, \lambda)$, $U_\theta(\theta, \lambda)$, and $U_\lambda(\theta, \lambda)$ denote tidal effects at the Earth's surface ($r = a$) on, respectively, the Earth's gravitational potential, and the radial, southward, and eastward displacements of the point (θ, λ) . Then, Φ , U_r , U_θ , and U_λ for our assumed spherical Earth have the form

$$\begin{aligned}\Phi(\theta, \lambda) &= kV_T^a(\theta, \lambda) \\ U_r(\theta, \lambda) &= \frac{h}{g}V_T^a(\theta, \lambda) \\ U_\theta(\theta, \lambda) &= \frac{l}{g}\partial_\theta V_T^a(\theta, \lambda) \\ U_\lambda(\theta, \lambda) &= \frac{l}{g\sin\theta}\partial_\lambda V_T^a(\theta, \lambda)\end{aligned}\quad (4)$$

where k , h , and l are the dimensionless Love numbers, g is the unperturbed gravitational acceleration at $r = a$, and

$$V_T^a(\theta, \lambda) = \text{Re} \left[\sum_{m=0}^{\infty} c_2^m Y_2^m(\theta, \lambda) e^{im\Omega t} \right] \quad (5)$$

is the tidal potential (Eq. 3) evaluated at $r = a$.

Tidal variations in gravity, strain, tilt, surface displacements, and satellite orbital motion can all be parameterized in terms of k , h , and l . As one example, consider the gravitational acceleration at a fixed point on the Earth's surface. There are three contributions to this acceleration, as discussed above. First, the direct gravitational acceleration from the moon (or Sun) in the inward radial direction is $-\partial_r V_T(r=a) = -\frac{2}{a} V_T^a$ (using the r^2 radial dependence of V_t shown in Eq. 3). Second, Φ (the tidal change in the Earth's gravitational potential at $r=a$) has an angular dependence Y_2^m and so, outside of the Earth, has radial dependence r^{-3} . Thus, the effect of Φ on the radially inward gravitational acceleration is $-\partial_r \Phi(r=a) = \frac{3}{a} \Phi = \frac{3}{a} k V_T^a$. Third, in the absence of tides, the surface gravitational acceleration, g , varies with radius as r^{-2} . So, if the surface point is displaced radially by U_r , the resulting perturbation in the gravitational acceleration at the displaced surface point is $U_r \partial_r g = -\frac{2}{a} g U_r = -\frac{2}{a} h V_T^a$.

Adding these three contributions together gives the total tidal effect on the observed acceleration as

$$\Delta g(\theta, \lambda) = -\delta \frac{2}{a} V_T^a(\theta, \lambda) \quad (6)$$

where $\delta = 1 - \frac{3}{2}k + h$ is called the gravimetric factor. Similar exercises show that tidal tilt is described by the diminishing factor $\gamma = 1 + k - h$, and that tidal strain, although more complicated than either tilt or gravity, depends on the Love numbers h and l .

The numerical values of the Love numbers depend on the Earth's internal properties. It is possible to learn about some of those properties by comparing tidal observations with predictions based on theoretical results for the Love numbers. For example, tidal observations have been used to place constraints on the Earth's anelastic properties. The effects of anelasticity are frequency- and stress-dependent, but are not well understood. Because tides sample the Earth in unique frequency and stress regimes (at lower frequencies and larger deep-earth stresses than seismology, e.g.), tidal studies have proven useful as complements to other types of anelastic observations.

The use of tides to study the Earth's internal structure can be difficult, for several reasons. One problem is instrument calibration, as described above for gravimeters. Any error in the calibration maps directly into a frequency independent, multiplicative error in the estimated Love numbers.

But there are also a number of other complicating effects, most related to the fact that the Earth is not really spherically symmetric and nonrotating. Some of those effects are potentially useful. For example, the Earth does rotate, and because of that rotation the Earth's internal properties are closer to being elliptically, rather than spherically,

symmetric. In this case, the results of Eq. 4 are still approximately valid, but the Love numbers k , h , and l are notably dependent on frequency near 1 cycle/day. This diurnal frequency dependence is particularly sensitive to the shape of the core-mantle boundary, and so diurnal tidal results can be used to help constrain that shape.

Another omission in the theory described above is the ocean. Tides in the ocean cause time-varying pressure loads on the surface of the solid Earth with the same frequencies as the Earth tides. These loads cause the Earth to deform, and this "load tide" (the tidal deformation that would occur in the absence of oceans is called the "body tide") affects all Earth tide observations to some extent. Tidal gravity observations, for example, can be perturbed by up to 10% near coasts, and typically a few percent in the interior of continents. Tidal tilt and strain can be perturbed by several hundred percent near coasts.

It is often difficult to model and remove the load tide well enough to use the remaining body tide to learn about the Earth's deep interior. There are uncertainties in ocean tide models. Furthermore, the Earth's response to the ocean tides, particularly the contributions to tilt and strain near coasts, can be sensitive to the local material properties that may or may not be adequately known. On the other hand, because of these possible uncertainties, people have sometimes been able to use tidal observations to help constrain the nearby ocean tide or the underlying material properties.

In fact, for tilt and strain local effects can be important on both the load tide and the body tide. By comparing observed tidal amplitudes from instruments in an array of tiltmeters or strainmeters, geophysicists can learn about the local geology and underlying structure. There have even been attempts to look for time-dependent variations in tidal amplitudes near active earthquake faults that might be caused by sudden changes in local material properties preceding an earthquake.

A related problem is that tilt and strain amplitudes are also affected by the local topography and by the shape and size of any cavity the instrument is placed in. (Tiltmeters and strainmeters are often placed in boreholes or tunnels to minimize the effects of the surface environment.) These effects are rarely interesting, and they cannot be observationally separated from the effects of local geology. Instead, they must be modeled separately and removed from the data.

Summary

The effects of Earth tides can be detected in several types of geophysical measurements. These tidal observations can be inverted to learn about large-scale material properties of the Earth's deep interior and about local geological structure.

In practice, a useful inversion is often difficult due to the large number of possible geophysical interpretations, and due to uncertainties associated with instrument calibration, topographic and cavity effects, and ocean tidal loading of the solid Earth.

Cross-References

- [Earth Rotation](#)
- [Earth, Density Distribution](#)
- [Geodesy, Figure of the Earth](#)
- [Geodesy, Networks, and Reference Systems](#)
- [Geodesy, Physical](#)
- [Geodynamics](#)
- [GPS, Tectonic Geodesy](#)
- [Gravity Method, Satellite](#)

Bibliography

- Agnew DC (2007) Earth tides. In: Schubert G, Herring T (eds) *Treatise on geophysics*, vol 3. Elsevier, Oxford, pp 163–195
- Baker TF (1984) Tidal deformations of the earth. *Science Progress* (Oxford) 69:197–233
- Harrison JC (1984) Earth tides (Benchmark papers in geology). Hutchinson Ross, Pennsylvania
- Melchior P (1983) *The tides of the planet earth*. Pergamon, Oxford
- Wilhelm H, Zürn W, Wenzel HG (1997) *Tidal phenomena*. Springer, Berlin

Earth, Density Distribution

Frank D. Stacey¹ and Paul M. Davis²

¹Division of Exploration and Mining, CSIRO, Kenmore, Australia

²Earth and Space Sciences, UCLA, Los Angeles, CA, USA

Definition

Except for iron meteorites, the Earth is the densest body in the solar system. As has been recognized for more than two centuries, it is, on average, about twice as dense as common surface rocks. But many of the early speculations on its internal structure were bizarre, ignoring the most basic physics and the simplest density information, and prompting a remark in one of the early papers on the subject with a sound scientific base (Oldham 1906): “Many theories of the earth have been propounded ... till geologists have turned in despair from the subject, and become inclined to confine their attention to the outermost crust of the earth, leaving the center as a playground for mathematicians.” Seismological studies,

for which Oldham was one of the pioneers, have changed that by giving us a clear picture of the interior. But debates on many details and improvements in the methods of investigating them continue.

The Mean Density

The first well-informed estimate of the Earth’s mean density was by Isaac Newton, who appealed to four pieces of evidence: the Earth’s surface gravity, the period of the Moon’s orbit, the size of Jupiter, and the orbital periods of its satellites. It was evident to him that the density of Jupiter is a quarter of the density of the Earth. Arguing that the density of Jupiter would not be lower than that of water (it is now known to be 1327 kg m^{-3}), and recognizing that the densest Earth materials would have sunk to the deep interior, he inferred that the mean Earth density is between 5000 and 6000 kg m^{-3} , a remarkable coincidence with the now established value $\bar{\rho} = 5513.4(6) \text{ kg m}^{-3}$. More precise estimates awaited the determination of the gravitational constant, G . Any astronomical or geophysical measurement of gravity, g , or, equivalently the acceleration of a body in a gravitational field, depends on the product of G and a mass. In the approximation of a spherical, nonrotating Earth, mass M , and radius R , as Newton showed, the surface gravity is $g = GM/R^2$ (the departure from this simple equation caused by rotation and ellipticity is about 0.5%). Since g and R have been reasonably well known since Newton’s time, a determination of G is equivalent to a determination of M , or $\bar{\rho}$, and was referred to as such by early investigators. Satellite measurements have given the most precise value of the product $GM = 3.986004415(8) \times 10^{14} \text{ m}^3 \text{ s}^{-2}$, where M includes the mass of the atmosphere, 8.84×10^{-7} of the total.

The first attempts to measure G were by M. Bouguer, who, in the 1730s, was conducting an astrogeodetic survey in the extreme topography of the Andes and noted a plumb line deflection by Mount Chimborazo, in Ecuador. He also compared the gravity on a plateau, at Quito, with observations on the nearby flood plain of the Esmeralda River, the first use of what we now know as the Bouguer method of gravity interpretation. In both cases Bouguer recognized that he was thwarted by heterogeneity of the geological structures and that his estimates of $\bar{\rho}$ were unsatisfactory, but in 1774 his plumb line deflection method was repeated more successfully by N. Maskelyne on the slopes of Mt. Schiehallion in Scotland. Maskelyne’s estimate was revised several times by others, using different assessments of the density of the mountain. This introduced a measure of uncertainty, so that geophysical measurements were generally disregarded for about 50 years after 1798, when H. Cavendish made a series of measurements of the gravitational force between laboratory masses, using a torsion balance developed for the

purpose by J. Michell. However, Cavendish's result was also a subject of doubt. Although the measurements were excellent, the final result was clouded by an arithmetic error, recognized in 1842 by F. Baily, leading to a revival of interest in geophysical methods. The most significant of these was by G. B. Airy, who, in 1856, measured the variation of gravity with depth in a coal mine near Newcastle, in northeast England. Although limited by inadequate measurements of rock density, Airy's experiment was well conceived. It did not require an absolute measurement of gravity but only the differences in its value at different levels in the mine, as indicated by the fractional differences between the periods of pendulums operated by observers who exchanged electrical signals to indicate their timing. The linked variations in gravity and density, recognized by Airy, are central to the development of earth models. The Airy method was used by others for the following 50 years, but with mixed results, and by 1900 the superiority of torsion balance measurements was clear. Now, it appears that conventional balance experiments, pioneered in the late 1800s by P. von Jolly and J. H. Poynting, are better still and the current best estimate of G derives primarily from them: $G = 6.6743(7) \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$. The uncertainty (1 standard deviation), 1 part in 10^4 , is far larger than the uncertainties in the other fundamental physical constants and translates to a similar uncertainty in M , and therefore $\bar{\rho}$, through the very precise value of (GM) .

The Moment of Inertia Constraint

The understanding of the Earth's density structure made little progress, beyond Newton's inference that denser materials would have sunk to the center, until the late 1800s. By then astronomical measurements of precession, combined with astrogeodetic measurements of ellipticity, had given an estimate of the moment of inertia of the Earth, C . This is expressed as a coefficient, f relating C to the mass, M , and equatorial radius, a

$$C = fMa^2 \quad (1)$$

For a uniform sphere $f = 0.4$ and a lower value is a measure of the concentration of mass toward the center. Using recent data the value for the Earth is $f = 0.330968(2)$, with the uncertainty in parenthesis. This number represents a quite strong central concentration of mass and prompted speculation on the Earth's deep composition. The cosmic abundance of iron, apparent from meteorites and the solar spectrum, invited the supposition that the Earth has an iron core. This idea was pursued by E. Wiechert who, in 1897, showed that both M and f could be matched by a model with an iron core and rocky mantle, with both materials approximating their familiar low pressure densities, if the core radius was about

0.78 a . But in 1906 R. D. Oldham reported seismological evidence for a smaller core, with a radius that he estimated to be about 0.4 a (it is now known to be 0.546 a), requiring higher densities for both the core and the mantle. Two developments were needed: recognition of the increase in density by the compression of deep burial and the introduction of a transition zone of increasing intrinsic density within the mantle. Both required more detailed seismological data than were available to Wiechert and Oldham.

Self-Compression

Williamson and Adams (1923) recognized that information about self-compression in a homogeneous layer was available from seismic wave velocities. Using the velocities of both compressional and shear waves

$$V_P = \{[K_S + (4/3)\mu]/\rho\}^{1/2}; V_S = \{\mu/\rho\}^{1/2} \quad (2)$$

they separated the adiabatic bulk modulus, K_S , from the rigidity modulus, μ , by writing

$$\Phi = V_P^2 - (4/3)V_S^2 = K_S/\rho = (\partial P/\partial \rho)_S \quad (3)$$

That is, the seismological parameter Φ gives directly the variation of pressure, P , with density, ρ , in a homogeneous region with an adiabatic temperature gradient. Since $dP/dz = \rho g$, g being gravity at depth z ,

$$d\rho/dz = \rho g/\Phi \quad (4)$$

This is the Williamson–Adams equation for the variation of density with depth in a homogeneous region. An additional term, $-\alpha\rho\tau$, was added by F. Birch to account for a temperature gradient differing from the adiabat

$$\tau = dT/dz - (dT/dz)_{\text{Adiabatic}} \quad (5)$$

where α is the volume expansion coefficient. However, this term is very small except at the limited depths of steep temperature gradients, the lithosphere, and the base of the mantle (D''). In the outer core, it is certainly negligible.

Another reason for imperfection of the Williamson–Adams equation is that, for composite materials, such as mantle rocks that are comprised of several minerals with different properties, there is a frequency dependence of elasticity. The unrelaxed elastic moduli that apply to rapid cycles of low stress, as in seismic waves, are slightly higher than the relaxed moduli applying to strong static compression. Averaged over the lower mantle, the difference is about 0.5% (Stacey and Davis 2008, Table 18.1). This means that $d\rho/dz$ is 0.5% greater than by the Williamson–Adams equation.

Interpreted as an error in the inferred temperature gradient, it amounts to ~ 100 K over the depth of the lower mantle, but earth models are not so accurate that this is regarded as serious.

When self-compression, as represented by the Williamson–Adams equation, is allowed for a simple two-layer model with the observed core size and a rocky mantle cannot match the moment of inertia. It is necessary to bring some mass up higher into the mantle. This was the vital next step to realistic earth models, introduced by K. E. Bullen and B. Gutenberg in the 1930s: the mantle transition zone. By then, seismological information was sufficiently detailed to indicate the depth range of the transition zone and the earth models that followed were quite close to our present understanding. As we now know, the transition zone is a combination of transitions at several levels in the depth range of 220–660 km, where mantle minerals are converted to denser phases. The transition at 660 km is conventionally regarded as the boundary between the upper and lower mantles. The lower mantle, which has two-thirds of the mantle volume and three-fourths of its mass, appears to be more or less homogeneous over almost the whole depth range to the core, at 2900 km depth, and conforms well to the Williamson–Adams equation with a density 18% higher than for the low pressure forms of the minerals. The upper mantle is more complicated. Bullen (1975) reviewed this stage in the development of our understanding of the Earth's density.

Free Oscillation Data

In the 1960s, observations of the Earth's modes of free oscillation introduced a new constraint on the density structure. Whereas body wave seismology gives the ratio K_S/ρ throughout the Earth but provides no formal separation of K_S and ρ , the spheroidal modes involve oscillations in the radial distribution of mass with consequent gravitational as well as elastic restoring forces. The equations of motion and the mode frequencies involve not just the ratios K_S/ρ and μ/ρ but also $G\rho$. With G a known constant, the $G\rho$ term allows an assessment of absolute density independent of elasticity. This is illustrated by considering the frequencies of two modes for the simple case of a uniform solid sphere. The spheroidal mode ${}_0S_2$ is an alternation between prolate and oblate deformations and, since there is radial motion, the $G\rho$ term has an important control on the frequency, f which is given by

$$f^2({}_0S_2) = (2/3\pi)G\rho + (5/3\pi^2)\mu/\rho R^2 \quad (6)$$

whereas the lowest toroidal mode, ${}_0T_2$, is an alternating twist between two hemispheres, with no radial motion and a frequency given by

$$f^2({}_0T_2) = (7/4\pi^2)\mu/\rho R^2 \quad (7)$$

so that

$$G\rho = (3\pi/2)f^2({}_0S_2) - (10\pi/7)f^2({}_0T_2) \quad (8)$$

The frequencies of more than 500 modes have been identified (Masters and Widmer 1995), providing a direct measure of global average properties, including the radial density profile. Fine details, such as the sharpness of boundaries, still require analysis of body wave data, but the broad scale structure is derived from the mode frequencies. The Preliminary Reference Earth Model (PREM) of Dziewonski and Anderson (1981) was produced in this way. It signaled a turning point in seismological studies, with less attention to global average properties and increasing emphasis on lateral heterogeneities, boundary structure, and anisotropy.

A powerful new use of free mode data to clarify the radial density structure was presented by Masters and Gubbins (2003), who selected a combination of modes sensitive to a restricted depth range, rather than inverting the whole data set. By concentrating on a region close to the inner core boundary, they obtained an estimate of the density difference between the inner and outer cores, 820 kg m^{-3} , compared with the PREM difference, 597 kg m^{-3} . The revision amounts to 1.8% of the density at that level, compared with the estimated 0.5% precision of the new technique, and may be interpreted as a measure of the revisions to PREM that may be expected in an updated model. From the splitting of mode frequencies, free oscillation data allow identification of lateral density variations independently of elasticity. This is referred to in a later paragraph on lateral heterogeneity.

A Density Gradient Anomaly

An interesting feature of PREM is the decrease in density with depth in the lithosphere and asthenosphere, over the depth range 24.4–220 km. The obvious interpretation is a steep temperature gradient, but this presents difficulty. The inverted density gradient is slight, little more than 0.1 kg m^{-3} per kilometer of depth, but the difficulty would arise even with zero gradients. For a uniform layer, density is a function only of pressure and temperature, so we can write

$$\begin{aligned} d\rho/dz &= (\partial\rho/\partial P)_T dP/dz + (\partial\rho/\partial T)_P dT/dz \\ &= -0.109 \text{ kg m}^{-3}/\text{km} \end{aligned}$$

and substitute the physical properties of mantle material in this depth range to obtain $dT/dz = 8.6 \text{ K/km}$. (For $d\rho/dz = 0$ this would be 7.9 K/km .) Over the depth range of 195.6 km,

the anomalous gradient gives a temperature range of 1680 K and since the temperature at 24.4 km (the averaged crust-mantle depth of PREM) cannot be taken as less than about 700 K, the implied temperature at 220 km is close to 2400 K. This is about 600 K too high; the temperature at 660 km depth required to meet the (P, T) conditions of the phase transition there is about 1900 K (Boehler 2000) and the 220 km temperature must be lower than that. The discrepancy can be understood if PREM underestimates the effect of anisotropy in the asthenosphere, causing the inverted density gradient as a modelling artifact.

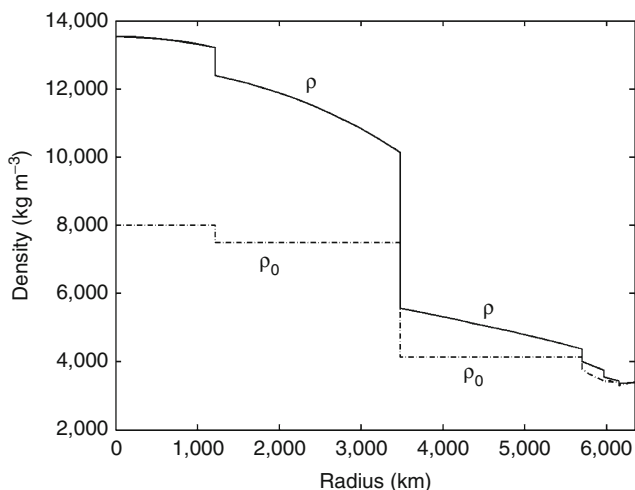
Extrapolation to Zero Pressure and Inferences About Composition

The density profile of PREM, with minor equation of state adjustments, is plotted in Fig. 1, together with equation of state extrapolations to zero pressure and a temperature of 300 K. The extrapolations use equations of state fitted to each of the ranges, assuming that the materials can be decompressed and cooled without phase changes. Thus, the broken line in the figure represents the densities that the high pressure phases would have under ambient conditions. For each of the inner and outer cores and lower mantle, the extrapolated densities are uniform, within the uncertainties of the calculation, and so indicate homogeneity. However, this is not true for the upper mantle, where the model indicates heterogeneity additional to the boundaries that mark recognized phase changes.

The density estimates provide the basic data needed for a discussion of composition. Two long-standing debates that

hang on this are the composition of the core, that is, the light elements that must be dissolved in iron-nickel alloy to reduce its density to that of the core, and the possibility of a significant compositional difference between the upper and lower mantles. As Birch (1952) recognized, there is no plausible alternative to iron as the major core constituent, but the outer core is about 10% less dense than pure iron and Poirier (1994) reviewed the arguments for rival candidate light solutes. At core pressures iron occurs as a hexagonal close-packed (ϵ) phase, for which the extrapolated density would be $\rho_{0,0} = 8352 \pm 23 \text{ kg m}^{-3}$ (Stacey and Davis 2004), compared with the core density, extrapolated to the same conditions and also in the ϵ phase, 7488 kg m^{-3} for the outer core and 7993 kg m^{-3} for the inner core. The effects on the iron density of each of the candidate light elements are known from metallurgical literature; although none of the observations were made on ϵ iron, they are taken as a sufficient guide to their potential contributions to the core density deficit. There are undoubtedly many elements dissolved in the core, but the favored important ones, with the mass percentages required in the outer core if each were the only light ingredient, are H (1.4%), C (10.6%), O (12.7%), Si (17.7%), and S (18.2%). The core densities can, in principle, be explained by any combination of these elements, but geochemical arguments restrict the range. The mass percentages of each, in the inner and outer core, respectively, favored by Stacey and Davis (2008, Table 2.5) are H (0.07%, 0.08%), C (0.45%, 0.50%), O (0.11%, 5.34%), Si (nil, nil), and S (8.02%, 8.44%).

There are two main arguments bearing on the similarity of the compositions of the upper and lower mantles, the seismological evidence for penetration of the 660-km boundary by subducting slabs, and the densities of high pressure phases of mantle minerals. The perovskite phase of $(\text{Mg,Fe})\text{SiO}_3$ is believed to be the dominant lower mantle mineral. It is produced by high pressure laboratory experiments and survives in a metastable state at $P = 0$ provided it is not heated. With no Fe the density at ambient conditions is 4107 kg m^{-3} . The second most important lower mantle mineral is almost certainly ferro-periclase (magnesiowustite), $(\text{Mg, Fe})\text{O}$, which may account for as much as 20% of the composition. It is less dense than the perovskite (3585 kg m^{-3} for MgO) but more readily accepts substitution of Fe for Mg and would be 5956 kg m^{-3} for FeO . The magnesium silicate perovskite can accept Al ions, but neither of these minerals readily accepts Ca, so a calcium perovskite is believed to occur also, but, unlike the magnesium perovskite and periclase, it does not survive decompression and is not as well studied. It is evident that, with a modest iron content, a mixture of these minerals can match the lower mantle density, $\rho_{0,0} = 4144 \text{ kg m}^{-3}$, and that, on the basis of density, there is no evidence for a compositional difference between the upper and lower mantles.



Earth, Density Distribution, Fig. 1 Density profile of the Earth based on PREM (solid line) and the profile of $\rho_{0,0}$, the density obtained by extrapolating to zero pressure and a temperature of 300 K (broken line)

Lateral Density Variations

Convection is ubiquitous in the Earth and this means that lateral density variations, with hot, dilated materials rising, and replaced by denser, cool materials. Heterogeneity of the mantle inferred from body wave tomography is generally consistent with the convective pattern, acknowledging that relics of earlier convective patterns are still apparent, but elastic moduli are more sensitive to temperature than is density and a thermodynamic analysis is needed for interpretation (Stacey and Davis 2008, Sect. 19.7). The splitting of mode frequencies presents evidence of lateral density variations that does not rely on such interpretation, but, in spite of several detailed studies, the difficulties are such that available results must be regarded as preliminary (Ishii and Tromp 1999; Masters et al. 2000; Romanowicz 2001). The general conclusion is that, in addition to temperature variations, compositional heterogeneity of the deep mantle must be invoked to explain the observations, in confirmation of a thermodynamic assessment (Forte and Mitrova 2001; Stacey and Davis 2008, p. 288) of results of body wave tomography (Robertson and Woodhouse 1996; Su and Dziewonski 1997).

With respect to the outer core, observable lateral heterogeneity must be discounted as quite implausible, in view of the speed of convective motion apparent from the geomagnetic secular variation. Compositional layering, with a stable low-density layer at the top of the core has been suggested (Braginsky 1999), but is not generally supported.

Summary

The Earth's density structure is now known in sufficient detail to provide a sound basis for discussions of internal composition. High pressure equations of state (Stacey and Davis 2004), the modern version of finite strain theory, provide reliable extrapolations of density to zero pressure, as well as distinguishing internal heterogeneities caused by temperature variations from compositional and mineralogical differences. Although fine details in the deep interior are not all resolvable, it is evident that the bulk of the Earth does not have very strong lateral heterogeneities, such as are obvious in the crust. Nevertheless, the heterogeneities that are seen are indications of the dynamics of the Earth, making them prime targets for continuing investigation.

Cross-References

- [Geodesy, Physical](#)
- [Gravity Measurements, Absolute](#)
- [Gravity, Global Models](#)
- [Seismic, Velocity, and Density Relationships](#)

Bibliography

- Birch F (1952) Elasticity and constitution of the Earth's interior. *J Geophys Res* 57:227–286
- Boehler R (2000) High pressure experiments and the phase diagram of lower mantle and core materials. *Rev Geophys* 38:221–245
- Braginsky SI (1999) Dynamics of the stably stratified ocean at the top of the core. *Phys Earth Planet Inter* 111:21–34
- Bullen KE (1975) *The Earth's density*. Chapman and Hall, London
- Dziewonski AM, Anderson DL (1981) Preliminary reference earth model. *Phys Earth Planet Inter* 25:297–356
- Forte AM, Mitrova JX (2001) Deep-mantle high viscosity flow and thermochemical structure inferred from seismic and geodynamic data. *Nature* 410:1049–1056
- Ishii M, Tromp J (1999) Normal-mode and free-air gravity constraints on lateral variations in velocity and density of Earth's mantle. *Science* 285:1231–1236
- Masters G, Gubbins D (2003) On the resolution of density within the earth. *Phys Earth Planet Inter* 140:159–167
- Masters G, Widmer R (1995) Free oscillations: frequencies and attenuation. In: Ahrens TJ (ed) *A handbook of physical constants 1: global earth physics*. American Geophysical Union, Washington, DC, pp 104–125
- Masters G, Laske G, Gilbert F (2000) Matrix autoregressive analysis of free oscillation coupling and splitting. *Geophys J Int* 143:478–489
- Oldham RD (1906) The constitution of the interior of the earth as revealed by earthquakes. *Q J Geol Soc Lond* 62:456–475
- Poirier J-P (1994) Light elements in the Earth's core: a critical review. *Phys Earth Planet Inter* 85:319–337
- Robertson GS, Woodhouse JH (1996) Ratio of relative S to P heterogeneity in the lower mantle. *J Geophys Res* 101:20041–20052
- Romanowicz B (2001) Can we resolve 3D density heterogeneity in the lower mantle? *Geophys Res Lett* 28(6):1107–1110
- Stacey FD, Davis PM (2004) High pressure equations of state with applications to the lower mantle and core. *Phys Earth Planet Inter* 142:137–184
- Stacey FD, Davis PM (2008) *Physics of the earth*, 4th edn. Cambridge University Press, Cambridge
- Su WJ, Dziewonski AM (1997) Simultaneous inversion for 3D variations in shear and bulk velocity in the mantle. *Phys Earth Planet Inter* 100:135–156
- Williamson ED, Adams LH (1923) Density distribution in the earth. *J Wash Acad Sci* 13:413–428

Earth's Structure, Core

Lianxing Wen

Department of Geosciences, State University of New York at Stony Brook, Stony Brook, NY, USA

Definition

Structure of Earth's core is referred to as the properties of velocity, density, attenuation, anisotropy, composition, and surface feature of Earth's core and how these properties change with depth, geographic location, and time.

Attenuation is a measure of material's ability to absorb energy as seismic waves pass through it.

Anisotropy is a material property that is directionally dependent.

Mushy zone is a layer of mixing solid inner core and liquid outer core materials.

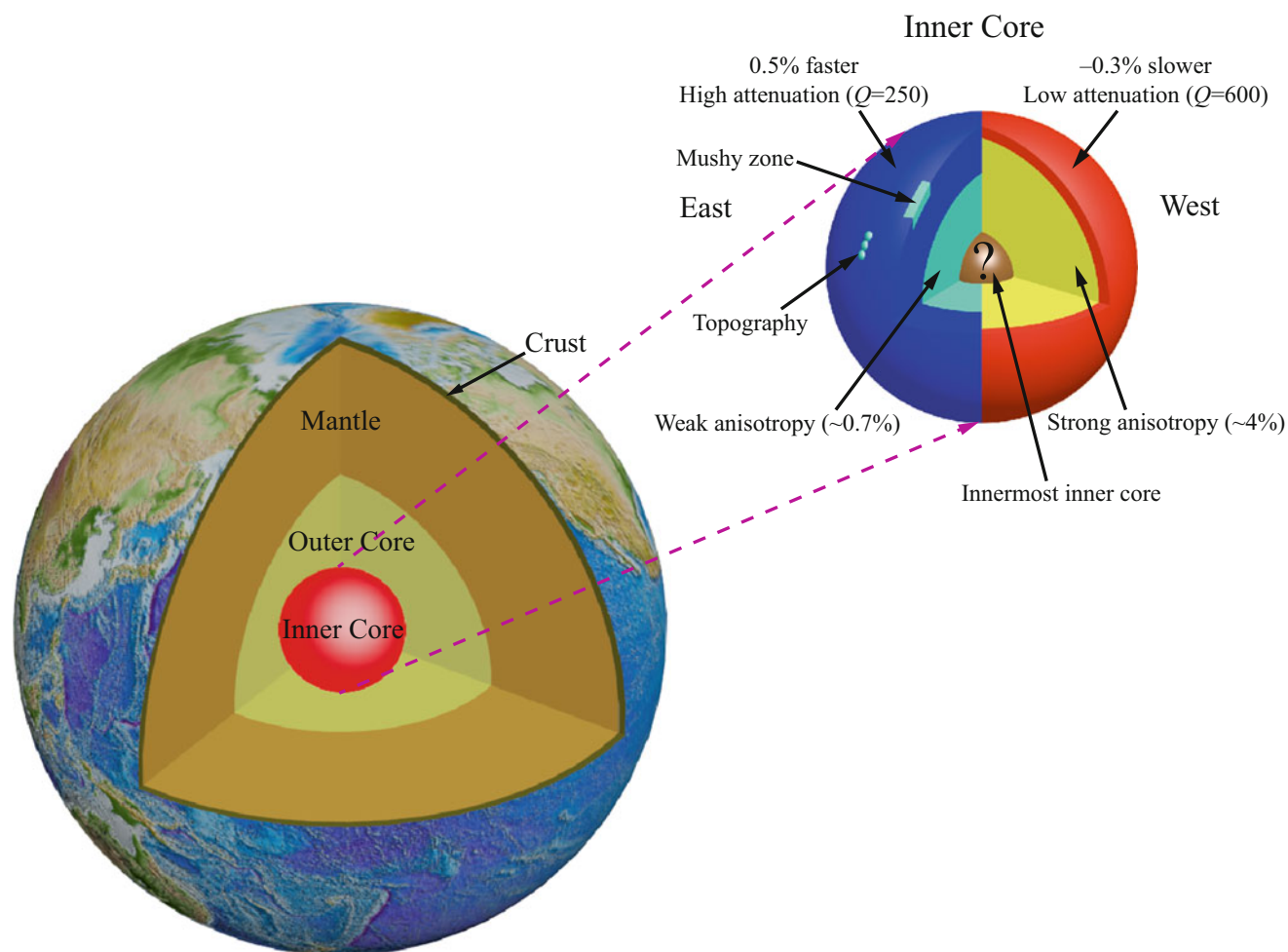
Overall Structure

The Earth's core occupies the center portion of Earth with a radius of 3480 km. It consists of a liquid outer core and a solid inner core. Earth's core is composed of Fe/Ni and some minor light elements, such as S, C, O, and Si. Those light elements are preferentially enriched in the outer core after the continuing solidification of outer core material to the inner core. Seismic compressional velocity and density exhibit jumps at the inner core boundary at a radius of about 1221 km. The inner core is anisotropic in both velocity and attenuation, with seismic waves exhibiting higher velocities and higher attenuation as they propagate along the polar paths than along the equatorial paths. In first order, the inner core exhibits an east-west hemispheric difference of various seismic properties, a

change of anisotropy with depth, an irregular surface, some localized mushy zones at the top, and temporal change of its surface in a time scale of days or months (Fig. 1).

Radial and Lateral Variations

With possible exception at the bottom and the top, the outer core is homogeneous due to vigorous mixing of the liquid iron/nickel. The inner core exhibits complex lateral and radial variations of structure, with the most prominent feature being the hemispheric difference in seismic velocity, attenuation, and anisotropy, between the "eastern" hemisphere (defined between longitudes from 40°E to 180°E) and the "western hemisphere" (defined between longitudes from 180°W to 40°E). The eastern hemisphere exhibits a higher velocity, higher attenuation, weaker anisotropy, and a larger transitional depth from top isotropy to deep anisotropy. The top 235 km of the eastern hemisphere also exhibits a flat velocity gradient with depth. At depth, the innermost 300 km of the inner core may have different forms of anisotropy from the rest of the inner core.



Earth's Structure, Core, Fig. 1 Cartoon summary of east-west hemispheric differences in seismic properties (velocity, attenuation, and anisotropy), surface topography, and mushy zone in the Earth's inner core

Irregular Surface

The inner core surface is irregular, despite that its global feature is not yet clear. In the regions that have been studied by seismology, most regions exhibit flat topography. But in some localized regions, the inner core surface exhibits at least two scales of topography: a height variation of 14 km changing within a lateral distance of no more than 6 km and a height variation of 4–8 km with a lateral length scale of 2–4 km.

Mushy Zone at the Top

A localized 4–8 km thick mushy zone is discovered across the inner core boundary beneath southwest Okhotsk Sea, with seismic properties intermediate between those of the inner and outer core and of a mushy zone. The discovered localized mushy zone is surrounded by a sharp inner core boundary nearby.

Temporal Changes

In some localized regions, mostly beneath Africa, South America, and central Pacific, the inner core surface changes its radius by several kilometers in a time scale of days or months. The temporal change of inner core surface appears episodic, rapidly migrating, and alternatively enlarged and shrunk.

Possible Physical Mechanisms

The proposed physical mechanisms for explaining the inner core structure include:

- (1) *For the inner core anisotropy*: solidification texturing due to the growth of inner core material and preferential alignment of anisotropic Fe crystals caused by inner core convection, magnetic force, and preferential growth of the inner core in the equator
- (2) *For the east-west hemispheric features*: a mantle-induced inner core solidification and a self-sustained translational inner core with the eastern hemisphere melting and the western hemisphere solidifying
- (3) *For the temporal change of inner core surface and inner core topography*: small-scale variation of temperature and/or composition near the inner core boundary and deformation of inner core surface by small-scale force near the inner core boundary
- (4) *For the mushy zone*: the current outer core composition being close to eutectic in most regions resulting in a sharp inner core boundary, but deviation in some localized regions generating a mushy zone

Cross-References

- [Differential Rotation of the Earth's Inner Core](#)
- [Earth, Density Distribution](#)

- [Earth's Structure, Global](#)
- [Magnetic Anisotropy](#)

Bibliography

- Alboussière T, Deguen R, Melzani M (2010) Melting-induced stratification above the Earth's inner core due to convective translation. *Nature* 466:744–747. <https://doi.org/10.1038/nature09257>
- Aubert J, Amit H, Hulot G et al (2008) Thermochemical flows couple the Earth's inner core growth to mantle heterogeneity. *Nature* 454:758–761. <https://doi.org/10.1038/nature07109>
- Dai Z, Wang W, Wen L (2012) Irregular topography at the Earth's inner core boundary. *Proc Natl Acad Sci U S A* 109:7654–7658. <https://doi.org/10.1073/pnas.1116342109>
- Ishii M, Dziewonski AM (2002) The innermost inner core of the Earth: evidence for a change in anisotropic behavior at the radius of about 300 km. *Proc Natl Acad Sci* 99:14026–14030
- Morelli A, Dziewonski AM, Woodhouse JH (1986) Anisotropy of the inner core inferred from PKIKP travel times. *Geophys Res Lett* 13:1545–1548
- Niu F, Wen L (2001) Hemispherical variations in seismic velocity at the top of the Earth's inner-core. *Nature* 410:1081–1084
- Tanaka S, Hamaguchi H (1997) Degree one heterogeneity and hemispherical variation of anisotropy in the inner core from PKP(BC) – PKP(DF) times. *J Geophys Res Solid Earth* 102:2925–2938
- Tian D, Wen L (2017) Seismological evidence for a localized mushy zone at the Earth's inner core boundary. *Nat Commun* 8:165. <https://doi.org/10.1038/s41467-017-00229-9>
- Wen L (2006) Localized temporal change of the Earth's inner core boundary. *Science* 314:967–970. <https://doi.org/10.1126/science.1131692>
- Yao J, Tian D, Sun L, Wen L (2019) Temporal change of seismic Earth's inner core phases: inner core differential rotation or temporal change of inner core surface? *J Geophys Res Solid Earth* 124:6720. <https://doi.org/10.1029/2019JB017532>

Earth's Structure, Global

Jean-Paul Montagner

Seismological Laboratory, Institut de Physique du Globe, University Paris-Diderot, UMR CNRS/7154, Paris, France

Definition

Structure of the Earth	The manner in which the Earth is constructed. It is characterized through physical and chemical parameters.
Crust	The outer layer of the solid Earth that lies above the Mohorovicic discontinuity.
Mantle	A thick layer of rock below the crust down to the core-mantle boundary CMB at 2,900 km depth.
Core	The central part of the Earth below CMB. It is composed of the fluid outer core (2,900

	down to 5,100 km) and the solid inner core below.
Anelasticity	Property of a substance in which there is no definite (linear) relation between stress and strain.
Anisotropy	Property of a substance being directionally dependent, as opposed to isotropy (identical properties in all directions).
Tomography	Imaging technique of 3D objects, developed in medical sciences, astrophysics, geophysics, etc. It is based on the use of large density of penetrating waves illuminating the object. The images are displayed as 2D cross sections (in Greek, “ <i>tomos</i> ” means slice) of the object.

Introduction

The Earth is the largest in size and mass of the Solar System's four terrestrial planets. Of these four planets, Earth also has the highest density, the highest surface gravity, the strongest magnetic field, and fastest rotation. It is also the third planet from the Sun and the densest and fifth largest of the eight planets in the Solar System. Earth is a terrestrial planet, meaning that it is a rocky body, rather than a gas giant like Jupiter. Earth is currently the only place where life is known to exist and the only terrestrial planet with active plate tectonics.

Earth Sciences in general, and Geophysics in particular, are built on the foundations of physics, mathematics, geology, and astronomy and are closely connected to other disciplines such as chemistry and biology. Its roots therefore go far back in history, but the science has blossomed only in the last century with the impressive increase in our ability to measure the properties of the Earth and the processes in play inside the Earth, on and above its surface. Geophysics includes the study of the Earth's solid layers, fluid envelopes, and near-space environment. However, in this contribution, the subject is narrowed to the solid Earth.

The interior structure of the Earth, like that of the other terrestrial planets, is divided into layers by their physical (rheological) or chemical properties. Scientific understanding of Earth's internal structure is based on geological, geochemical, and geophysical observations at the surface of the Earth including topography and bathymetry, observations of rock in outcrops, and samples brought to the surface from greater depths, primarily by volcanic activity. Geophysics can make use of measurements of different physical fields such as the gravity, magnetic fields, and the fields associated with the propagation of seismic waves passing through the Earth. The Earth's gravity field can be used to calculate its mass, its moment of inertia. By adding some additional

physical constraints, the density variations within the Earth can be calculated. The magnetic field of internal origin provides fundamental information on the deepest layer of the Earth, the core (see contributions by Foss, “Magnetic Data Enhancements and Depth Estimation” and Arora, “Magnetic Methods, Principles”). This interdisciplinary observational approach is complemented by laboratory experiments with minerals at pressures and temperatures characteristic of the Earth's deep interior.

The technological advances of the last century in laboratory and field instrumentation, computing, and satellite-based remote sensing are largely responsible for the explosive growth of geophysics. These advances enabled geophysicists to make more and more accurate measurements, to collect and to analyze enormous amounts of data, and to model more and more complex systems. This new view of how the Earth works enabled a fundamental understanding of structure of the Earth as well as earthquakes, volcanoes, and mountain building, indeed all of geological processes and objects.

1D Structure of the Earth

The global structure of Earth can be defined in two different ways, either by considering its physical properties or its chemical properties. Geophysics is the only Earth science discipline enabling to image the invisible deep inner structure of our planet by a direct measurement of its physical properties. To first order, the structure of the earth is dominated by its radial structure (spherically symmetric model). During the first part of the twentieth century, geophysicists determined the onion-like fine structure of the Earth. Seismic waves enable to investigate the elastic and to some extent anelastic properties of the whole Earth.

Characterizing the Interior of the Earth with Earthquake Waves

Global seismology started in 1889 when Ernst von Rebeur-Paschwitz in Potsdam (Germany) associated a disturbance recorded on a tiltmeter with the occurrence of a remote Japanese earthquake. Since then, seismologists modeled the propagation of seismic waves from teleseismic events by using the ray theory considering seismic waves as virtual particles following rays between earthquake focus and receivers, and their travel times are sensitive to the structure of the Earth.

During the first decades of the twentieth century (see Dziewonski and Romanowicz (2015) for details), the layering of Earth was indirectly inferred using the travel times of refracted and reflected seismic waves generated by earthquakes. The changes in seismic velocity of P-waves (compression waves) and S-waves (shear waves) between

different layers causes refraction owing to Snell-Descartes's law. Reflections are caused by a large increase in seismic velocity and are similar to light reflecting from a mirror. Seismologists were able to find and characterize the main seismic discontinuities (depth and velocity jump) within the different layers of the Earth (see Vinnik, this issue). First of all, the Mohorovicic discontinuity between crust and mantle in the depth range 10–60 km then the core-mantle boundary at 2,900 km depth by Oldham (1906). The core does not allow shear waves to pass through it, showing that its outer part is fluid. Finally, Inge Lehmann (1936) demonstrated the existence of a solid inner core.

Mantle itself can be divided into upper mantle and lower mantle by a discontinuity at 660 km depth. In the upper mantle, the finding of a discontinuity around 410 km depth defines the “upper transition zone” between 410 and 660 km depth. At the base of the mantle, about 200 km above the core-mantle boundary, a discontinuity was discovered, identifying the still mysterious D”-layer. There are some other candidates for entering the club of seismic discontinuities, the Hales discontinuity at 80 km depth, the Gutenberg discontinuity at the base of continental and oceanic lithosphere, the Lehmann discontinuity at 220 km, and also at 520 km and around 900 km depths, but their global nature is still debated. The layer between 660 km and 900 km is commonly named the “lower transition zone.” Most seismic discontinuities are interpreted as a mineralogical phase transition between minerals of mantle silicates (olivine-perovskite system), but some of them might be a boundary between different chemical compositions.

From its elastic properties, the different layers are presented in Table 1 and Fig. 1. The Earth can be divided into the outer silicate solid crust, the upper mantle (including lithosphere, asthenosphere, and upper transition zone), the lower mantle (with the lower transition zone and the

D”-layer at its base), the liquid outer core much less viscous than the mantle, and the solid inner core.

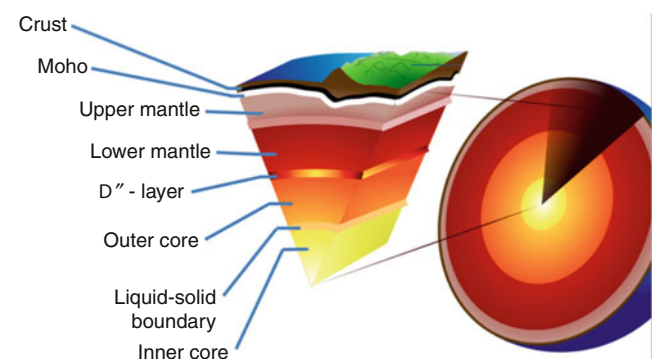
From 1D Models to 3D Tomographic Models

Many reference models (spherically symmetric models) have been developed during the last century, but the most popular ones are the PREM (Preliminary Reference Earth Model by Dziewonski and Anderson 1981) (Fig. 2) and IASP91 (Kennett and Engdahl 1991). PREM introduced for the first time an anisotropic layer (with the specific case of radial anisotropy with a vertical symmetry axis) in the uppermost 220 km of the mantle. Montagner and Kennett (1996) by trying to reconcile body wave and normal mode data showed that a significant amount of radial anisotropy is also necessary in the mantle transition zones between 410 and 900 km depth.

However, important deviations between observed and theoretical travel times of seismic waves with respect to 1D-reference models cannot be explained by these simple 1D models. These differences, though relatively small (<10%), are due to the existence of lateral heterogeneities of physical properties within the different layers of the Earth. Since the 1960s, the acceptance of the paradigm of plate tectonics and mantle convection leads the seismological community to a better appraisal of the coupling between lithosphere and the underlying mantle convection. Plate tectonics demonstrated that the internal dynamics is very active and that tectonic plates are the surface evidence of deeper convective phenomena. According to geodynamics calculations, large lateral variations in temperature must be associated with lateral variations of seismic velocities. This new approach of global geodynamics has been facilitated by the revolution of broadband seismology in the 1980s (see Engdahl, this issue), which came up with the rapid expansion of seismic networks, the free accessibility of digital data, and more and more powerful computers. Time was ripe for the new era of seismic tomography (imaging of 3D seismic anomalies) and its application to global geodynamics.

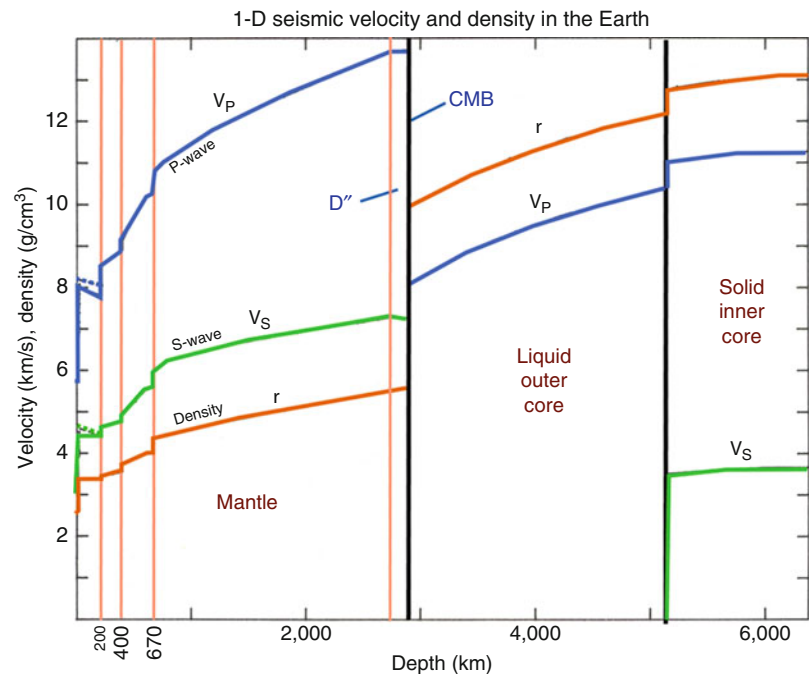
Earth's Structure, Global, Table 1 The 1D structure of the Earth

Layer (Kilometers)	
0–80	Lithosphere (locally varies between 5 and 200 km)
0–35	Crust (locally varies between 5 and 70 km)
35–80	Uppermost part of mantle
35–2,890	Mantle
35–660	Upper mantle
80–220	Asthenosphere
410–660	Upper transition zone
660–2,890	Lower mantle
660–900	Lower transition zone
2,740–2,890	D”-layer
2,890–5,150	Outer core
5,150–6,360	Inner core



Earth's Structure, Global, Fig. 1 The different layers of the Earth

Earth's Structure, Global,
Fig. 2 Preliminary Reference
 Earth Model (PREM Dziewonski
 and Anderson 1981)



"PREM", Dziewonski and Anderson (1981)

Mantle Convection Jigsaw

The investigation of mantle convection is the best example of the implementation of these new available tools in the framework of an interdisciplinary approach. Thermal convection in the Earth has been recognized for a long time (Holmes 1928; Pekeris 1935) as the cause of dynamic processes such as mountain building, volcanic and seismic activity, and, more recently, seafloor spreading. With the advent of plate tectonics, it has become possible to describe quantitatively the motions at the surface of the Earth. The plates, considered as rigid bodies, are the surface manifestation of deep-seated movements related to mantle convection (e.g., Turcotte and Schubert 2002). This evidence constitutes the first manifestation of peculiarities of mantle convection, for it demonstrates large variations in the rheological properties of Earth material between rigid plates and underlying mantle, which cannot be explained by 1D models. Meanwhile, hotspots were identified (Wilson 1963) and their fixity with respect to plate motions demonstrated (Morgan 1971). There is a good consensus on the origin of plumes in a thermal boundary layer, but the depth of origin of hotspots is still controversial. The rheology of the lithosphere and the existence of hotspots are two simple examples which illustrate the fact that mantle convection differs in many respects from "classical" convection (such as Benard convection) observed in usual viscous fluids.

The understanding of mantle dynamics is similar to a police investigation: each earth science field provides miscellaneous and often contradictory clues, which must be assembled like jigsaw puzzle pieces. Different approaches

have been used to characterize and understand mantle convection (see ► "Mantle Convection" in the chapter by Bercovici, this issue). None of them could provide a general and consistent explanation of simple features such as the shape and geometry of plates, the origin of hotspots, the number of layers in the mantle, etc. Many questions arose: How is surface geology (continents, plate boundaries, mountain ranges, hotspots) reflected at depth? How are interactions and the coupling between the different layers of the Earth characterized? How does deep mantle flow affect the surface geology, or conversely how do surface phenomena affect deep dynamics?

Another issue was the characterization of the spatial scales of heterogeneities at the global scale. Geology makes evident a multiplicity of scales. Some tectonic plates are gigantic (larger than 10,000 km), some others smaller (1,000 km). Continents have as well various sizes. The different plate boundaries (subduction zones, ridges, transform faults) have a very small lateral extent (a maximum of 100 km except continent-continent collision zones). Hotspots are even smaller and their role in global geodynamics cannot be explained by plate tectonics. This multiscale approach is also present for the temporal scales. Whereas times associated with plate tectonics count in tens of millions of years, catastrophic events such as hotspot birth, mantle avalanches, and seismic and volcanic cycles last only centuries. According to time scales, rheological properties change. Therefore, natural phenomena are by essence multiscale and make it necessary a multidisciplinary approach. The role of geophysicists consists in quantifying, understanding

processes, and proposing simple laws in spite of the complexity of processes.

One of the most powerful approaches to investigate mantle convection is the numerical approach. The basic equations of physics governing mantle convection are well known (Navier-Stokes equations and thermal equation of energy conservation). This system of equations is very complex and nonlinear, expressed in terms of adimensional numbers such as the Rayleigh number Ra . Actual estimates of Ra make it largely beyond the critical Rayleigh number, which means that mantle convection is thermally turbulent. In other words, mantle convection is chaotic at the geological time scales, meaning that the convective flow is not stationary in time and space. With the rapid expansion of computer facilities, it is now possible to numerically simulate the equation system. However, it turns out difficult to reproduce plate behavior with numerical experiments. Laboratory experiments complete the whole numerical battery, enabling to finely investigate some specific phenomena. Material science, by studying physical properties of earth minerals (rheology, anelastic and elastic properties) at very high pressure (by using diamond anvil cells and synchrotrons, etc.), also provide important information on mineralogical composition of the mantle. Most seismic discontinuities can be interpreted as pressure- and temperature-induced phase changes in the olivine system (Birch 1952), such as the 410 km and 660 km discontinuities. However, some composition changes might also occur around these depths and deeper in the lower mantle where thermochemical convection might be predominant. The recent and unexpected discovery of the post-perovskite (high-pressure phase of magnesium silicate) in 2004 (Murakami et al. 2004; Oganov et al. 2004), present in the lowermost mantle, has fundamental influence of the convective mixing in the deep mantle, particularly in the D''-layer.

In that context, the determination of 3D models of seismic parameters of the mantle (also coined tomographic models) can be considered as a pragmatic approach which consists in imaging the structure of convection in order to provide clues to the nature of driving mechanism and to investigate mantle convective processes, the engine of plate tectonics. Thanks to the intensive use of computers which can handle very large data sets, the techniques of seismic tomography (Nolet 2008) can image the 3D deep structure of the Earth and provide unavoidable constraints of possible models of mantle convection and makes it possible to test different competing hypotheses. However, the main limitation of seismic tomography is that it only provides an instantaneous image of our planet Earth, and it is necessary to confront seismic results with other Earth science fields providing information on the history of our planet.

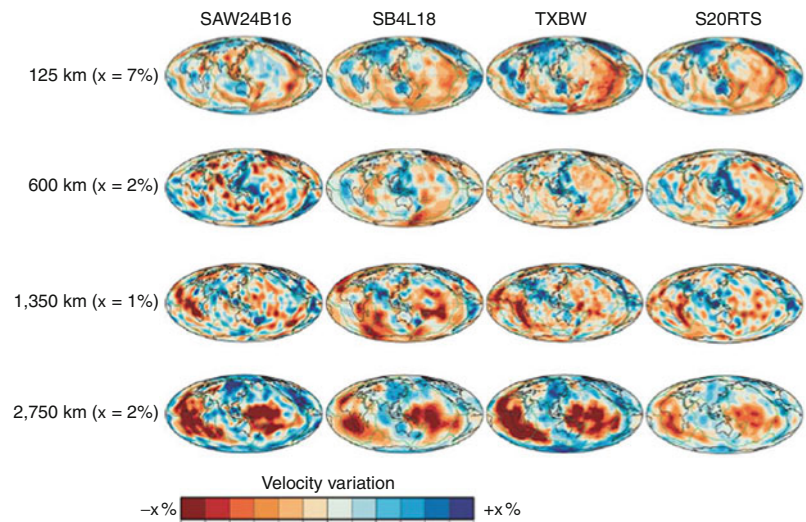
Seismic 3D Tomography: The Earth Scanner

Seismic wave propagation is very sensitive to temperature, chemical, and mineralogical anomalies. These anomalies reflect the active dynamics of the Earth which should be a dead planet without them. It was observed for many decades that seismic velocity below old continents is much larger than the velocity beneath young oceans. This difference between tectonic provinces can be explained by a different history: cratons older than one billion years had a lot of time to get colder, whereas below oceans, the younger and warmer lithosphere is permanently recycled within the mantle. An increase in temperature is usually translated into decrease in seismic velocity, whereas an increase in pressure is translated into velocity increase. At a given depth (approximately constant pressure), the lateral variations of seismic velocity can be primarily interpreted in terms of temperature variations. A large variety of tomographic techniques similar to ultrasonography or scanner has been developed for getting the 3D structure of the Earth, notwithstanding the geological nature of the surface.

The first 3D models of the Earth's mantle were proposed by the end of the 1970s (Dziewonski et al. 1977; Sengupta and Toksoz 1976), but the true revolution of seismic tomography took place in the mid-1980s when a complete set of tomographic models was proposed from the surface down to the center of the Earth (see Woodhouse and Dziewonski (1984); Nataf et al. (1994); Tanimoto and Anderson (1985) for the upper mantle; Dziewonski (1984) for the lower mantle). The present lateral resolution at global scale for the whole Earth is approximately 1,000 km. Tomographic models (Fig. 3) fundamentally renewed our vision of Earth dynamics and shed new light on deep processes.

At 100 km depth, the distribution of seismic velocity anomalies reflects the geological surface structure. Almost all plate boundaries are slow since they are often associated with active volcanism. In oceanic zones, seismic velocity increases with the age of the sea floor confirming the cooling of plates with age. This remarkable agreement between observed seismic structure and structure deduced from geology was the first proof of the validity of tomographic approach. At larger depths, the correlation between surface geology and seismic structure progressively vanishes and the amplitude of seismic anomalies decreases. The continental roots can be found down to 250–300 km, but mid-ocean ridges are no longer visible. In contrast, subducting slabs associated with high velocities can be traced down very deeply. The upper and lower transition zones (410–900 km) are still one of the less well-resolved depth range. It is characterized by weak heterogeneities, but the velocity distribution is dominated by a simple scheme, with high

Earth's Structure, Global,
Fig. 3 Maps through S-velocity models SAW24B16, SB4L18, TXBW, and S20RTS and P-velocity models P-MIT and FF-PRI at depths of 125, 600, 1,350, and 2,750 km. (From Thurber and Ritsema 2015)



velocities around the Pacific Ocean and slow velocities below central Pacific and Africa, the so-called “degree 2” pattern (Masters et al. 1982).

At the global scale, it is usual to expand the velocity anomalies in terms of spherical harmonics $Y_l^m(\theta, \phi)$, where θ, ϕ are the spherical geographical coordinates, l the angular degree, and m the azimuthal order. The spectral amplitude of individual degree l enables to assess the importance of different spatial wavelengths λ , since λ is inversely proportional to l . The “degree 2-pattern” predominant in the transition zone as well as in the D”-layer is associated with large-scale mantle flows, upwellings on both sides of Pacific and upwellings below central Pacific and Africa. The spectrum of heterogeneities is “red” (see Dziewonski and Romanowicz 2015 for a review), therefore dominated by very large wavelengths, but the role of smaller scale objects such as slabs and mantle plumes is still discussed.

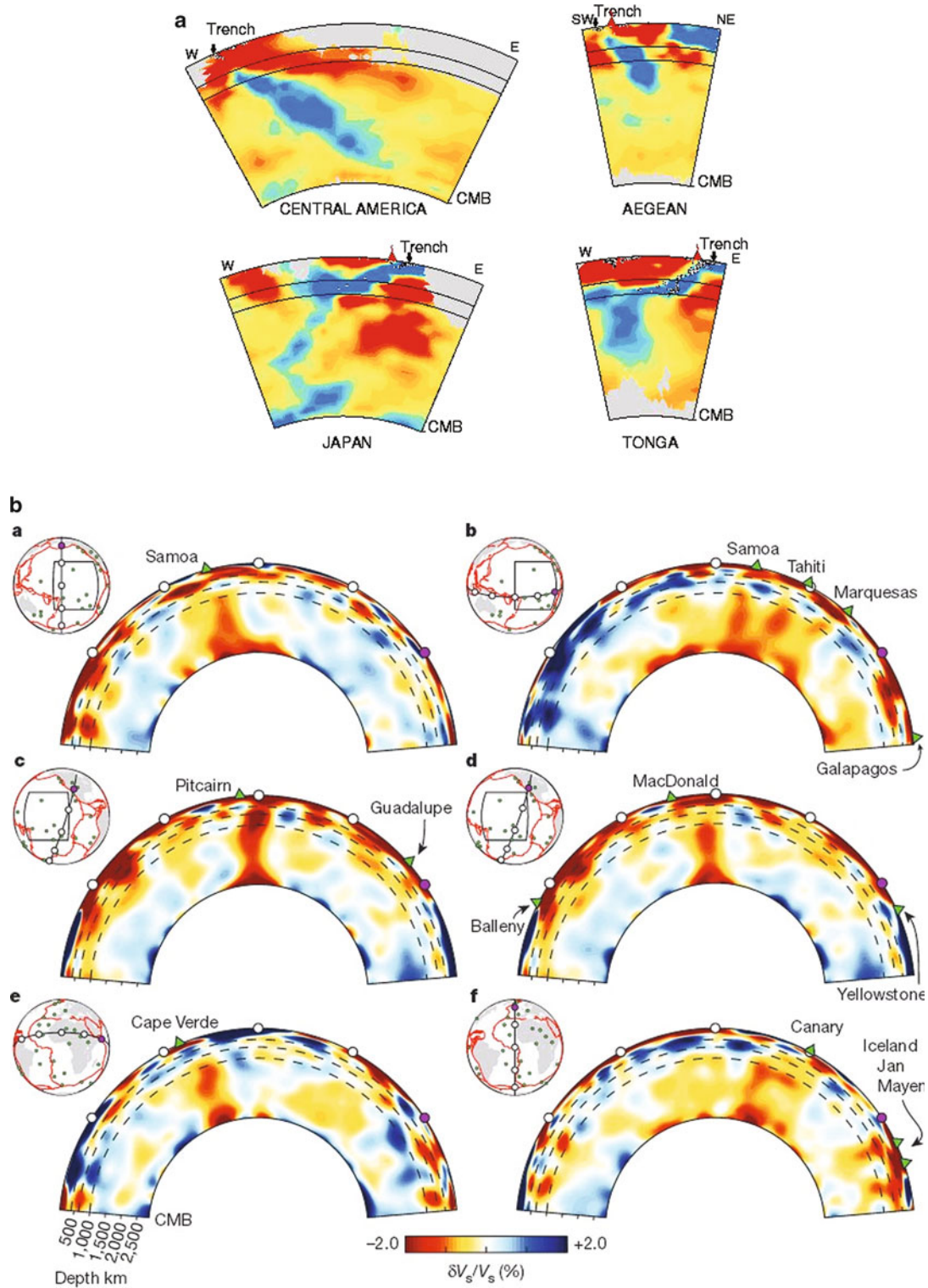
The behavior of slabs in regional studies does not seem to be uniform (Fig. 4a). Some slabs seem to fall down to the core-mantle boundary (van der Hilst et al. 1997), whereas others, e.g., beneath Japan (Fukao et al. 1992), are stagnant in the transition zone. Mantle plumes and the associated hotspots might initiate continental breakup, inducing the opening of new ridges of oceans, and might dramatically affect biodiversity. After plate tectonics, plume tectonics? There is still a controversy on the origin of plumes. It is likely that they do not have a single origin but can originate from the different boundary layers within the Earth, in the asthenosphere, transition zones, and D”-layer. There are still lively debates about the origin at depth of mantle plumes (Fig. 4b shows an example of plume imaging), and two opposite extreme models (plate model and plume model) are summarized in Fig. 5. At the base of the mantle, the D”-layer is still very mysterious since it is the probable

graveyard of slabs and a good candidate for being the matrix of some mantle plumes giving birth to hotspots at the surface of the Earth.

An Anisotropic and Anelastic Earth

Most earth minerals cannot be limited to their isotropic and elastic properties. When they are deformed over geological timescales, mineral grains and aggregates tend to align by crystal preferred orientation (CPO) which can lead to bulk anisotropy (see “► Magnetic Anisotropy”). Additional anisotropy can also come up when crack distribution (in the crust) or fluid inclusions (asthenosphere, inner core) are present within the material (shape preferred orientation, SPO). Water content might also play a key role in alignment mechanisms (Karato et al. 2008). In the upper mantle, the different minerals present are strongly anisotropic. The difference of velocity between the fast axis and the slow axis is larger than 20% for olivine, the main constituent of the upper mantle. Other important constituents such as orthopyroxene or clinopyroxene are anisotropic as well (>10%) (see Babuska and Cara (1991), Anderson (2007), and Mainprice (2015) for reviews). Consequently, the petrological models which are assemblages of different minerals are less anisotropic than pure olivine. The amount of anisotropy is largely dependent on the percentage of these different minerals and on the mechanisms aligning the crystallographic axes according to CPO. For example, the anisotropy of the pyrolytic model, mainly composed of olivine and orthopyroxene (Ringwood 1975), will depend on the relative orientation of the crystallographic axes of different constituents (Christensen and Lundquist 1982).

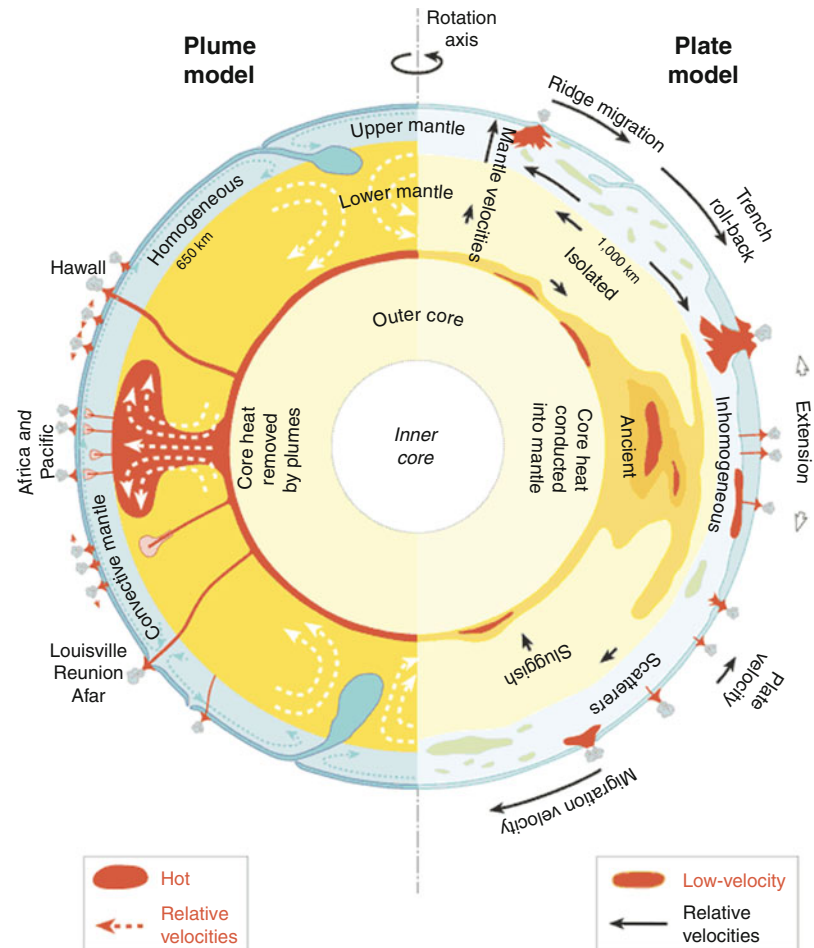
For the sake of mathematical simplicity, it was usually assumed that the propagating elastic medium is isotropic, in spite of the large amplitude of anisotropy in several depth



Earth's Structure, Global, Fig. 4 Tomographic images of slabs (Top, van der Hilst et al. 1997) and plumes (Bottom, French and Romanowicz 2015)

Earth's Structure, Global,

Fig. 5 Competing models for explaining the origin of hotspots (Left, Courtillot et al. 2003; Anderson 2001). (From Foulger et al. 2005)



ranges. And only 1D reference models of attenuation are incorporated in seismic modeling. Global tomographic models improved over the years not only by an increase in the number of data but also by more general parameterizations. They are now including anisotropy (radial anisotropy in Nataf et al. (1994); general slight anisotropy in Montagner and Tanimoto (1991)) and anelasticity (Romanowicz 1995). Whereas isotropic heterogeneities enable to map hot and cold regions within the Earth, seismic anisotropy gives access to the convective flow. Gaboret et al. (2003) and Becker et al. (2003, 2008) demonstrated that mantle circulation inferred from geodynamic models is in good agreement with radial and azimuthal anisotropy distributions (Fig. 6). Seismic anisotropy has many other applications such as the determination of the base of continental roots (Montagner 1994; Gaherty and Jordan 1995; Silver 1996; Gung et al. 2003; Debayle et al. 2005) and the investigation of small-scale convection at the base of the lithosphere and of convection boundary layers within the Earth, in the transition zones (Trampert and van Heijst 2002), D"-layer (see Lay and Garnero et al.), and even within the inner core (see Souriau and Calvet 2015 in structure of the inner core).

Because of its strong dependence on temperature, partial melting, and water content, mapping anelastic attenuation in the Earth has the potential to provide valuable information on Earth's three-dimensional (3D) structure and dynamics, in complement to what can be learned from mapping elastic isotropic velocities and anisotropic parameters. A significant challenge is to separate the effects of anelastic (or intrinsic) attenuation from those of elastic focusing and scattering (or extrinsic) due to propagation in elastic 3D structure.

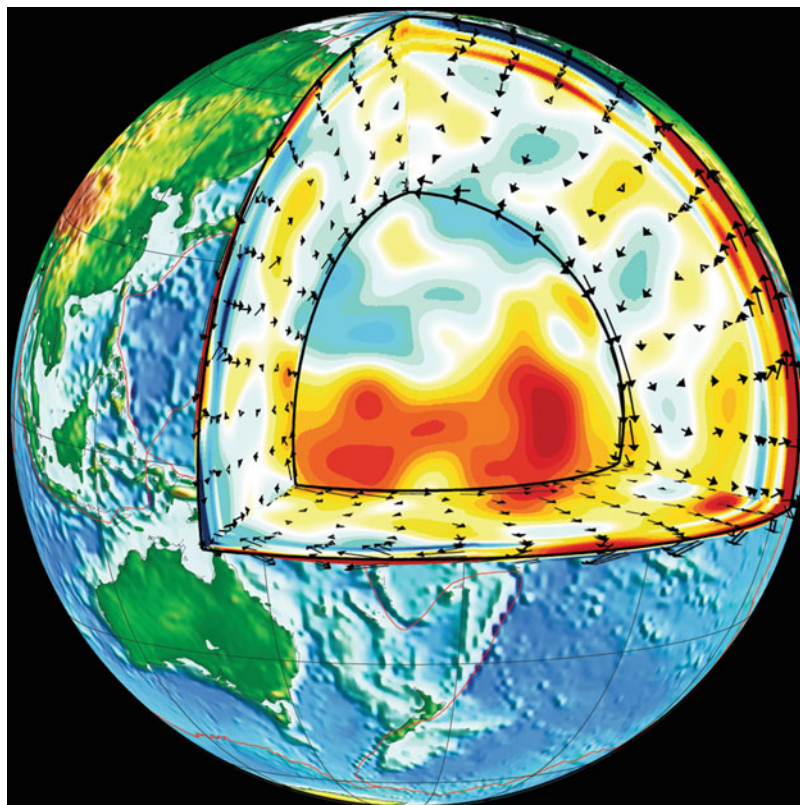
Seismic anisotropy and anelasticity provide new ways to investigate geodynamic processes within the different layers of the Earth.

Outlook: New Challenges

So far, global 3D-tomographic models are well resolved for very large spatial scales (>1,000 km). And the global flow circulation in most of the mantle is dominated by the "degree 2 pattern." However, smaller wavelengths are not so well resolved even though many geological objects with a lateral

Earth's Structure, Global,

Fig. 6 Numerical modeling of flow circulation superimposed to 3D tomographic model (Gaboret et al. 2003)



extent of the order of 100 km, such as slabs or mantle plumes, play a key role in global geodynamics.

New innovative numerical techniques for forward modeling (such as Spectral Element Method) and for inverse problems (adjoint tomography of Tarantola 1986; Tromp et al. 2005) are now available and make it possible to use the full waveforms of seismograms. The first images of different volcanic plumes including Hawaiian plume (Montelli et al. 2004) were first obtained by incorporating finite frequency effects (see ► “Seismic Tomography”). Therefore, a new revolution in global tomography is underway (see Fig. 4b for images of plumes).

Impressive progress through global tomography, geochemistry, and mineral physics has been made during the last 30 years in our understanding of global geodynamics, demonstrating how active and turbulent is our planet. All layers are heterogeneous, interact, and exchange matter. The geosciences community must improve the lateral resolution and the quality of the 3D images and incorporate on a routine basis anisotropy and anelasticity. There is a real need for a 3D seismic reference Earth model in agreement with geological and mineralogical data and fluid dynamics modeling.

This progress brings about the expansion of new instruments such as very dense seismic networks on land (e.g., US array, Hi-Net, Euro-Array, etc.) and on the sea floor, by exploring other planets and by incorporating new data

such as seismic noise (Shapiro et al. 2005; Nishida et al. 2009) implementing more powerful numerical techniques. Geophysics only provides some pieces for our puzzling global geodynamics, and a multidisciplinary effort is necessary to fully understand the spatiotemporal evolution of our planet. The exploration of other telluric planets is necessary for having other examples of solutions chosen by nature, and the seismometer within InSight mission successfully deployed in January 2019 on planet Mars will provide such answers.

Cross-References

- [Body Waves](#)
- [Core-Mantle Coupling](#)
- [Differential Rotation of the Earth's Inner Core](#)
- [Earth's Structure, Core](#)
- [Earth's Structure, Lower Mantle](#)
- [Earth's Structure, Upper Mantle](#)
- [Free Oscillations of the Earth](#)
- [Geodynamics](#)
- [Inverse Theory, Linear](#)
- [Lithosphere, Continental](#)
- [Magnetic Anisotropy](#)
- [Mantle Convection](#)
- [Mantle D'' Layer](#)

- Mantle Plumes
- Plate-Driving Forces
- Propagation of Elastic Waves: Fundamentals
- Seismic Discontinuities in the Transition Zone
- Seismic Imaging, Overview
- Seismic Instrumentation
- Seismic Ray Theory
- Seismic Structure at Mid-Ocean Ridges
- Seismic Tomography
- Seismic, Migration
- Subduction Zones
- Surface Waves

Bibliography

- Aki A, Richards PG (2002) Quantitative seismology, 2nd edn. University Science Books, Sausalito
- Anderson DL (2007) New theory of the Earth. Cambridge University Press, Cambridge
- Babuska V, Cara M (1991) Seismic anisotropy in the Earth. Kluwer, Dordrecht
- Dahlen FA, Tromp J (1998) Theoretical global seismology. Princeton University Press, Princeton
- Gutenberg B, Richter CF (1954) Seismicity of the Earth. Princeton University Press, Princeton
- Lay T, Wallace TC (1995) Modern global seismology. Academic, San Diego
- Nolet G (2008) A breviary of seismic tomography. Cambridge University Press, Cambridge, UK
- Poirier J-P (1991) Introduction to the physics of the earth interior. Cambridge University Press, Cambridge, UK
- Richter CF (1958) Elementary seismology. Freeman, San Francisco
- Treatise on Geophysics (2015) Gerald Schubert (Editor-in-Chief), Elsevier B.V.
- Turcotte DL, Schubert G (2002) Geodynamics, 2nd edn. Cambridge University Press, Cambridge
- Foulger GR, Natland JH, Presnall DC, Anderson DL (2005) Plates, plumes, and paradigms. Geological society of America special, vol 388, 881pp
- French S, Romanowicz B (2015) Broad plumes rooted at the base of the Earth's mantle beneath major hotspots. *Nature* 525:95–101
- Fukao Y, Widiyantoro S, Obayashi M (2001) Stagnant slabs in the upper and lower mantle transition region. *Rev Geophys* 39:291–323
- Gaboret C, Forte AM, Montagner J-P (2003) The unique dynamics of the Pacific hemisphere mantle and its signature on seismic anisotropy. *Earth Planet Sci Lett* 208:219–233
- Gaherty JB, Jordan TH (1995) Lehmann discontinuity as the base of an anisotropic layer beneath continents. *Science* 268:1468–1471
- Gung Y, Panning M, Romanowicz B (2003) Global anisotropy and the thickness of continents. *Nature* 422:707–711
- Holmes A (1928) Radioactivity and Earth movements. *Trans Geol Soc Glasgow* 18:559–606
- Karato S-I, Jung H, Katayama I, Skemer P (2008) Geodynamic significance of seismic anisotropy of the upper mantle: new insights from laboratory studies. *Annu Rev Earth Planet Sci* 36:59–95
- Kennett BLN, Engdahl ER (1991) Travel times for global earthquake location and phase identification. *Geophys J Int* 122:429–465
- Lehmann I (1936) Publications du Bureau Central Séismologique International, série A, Travaux Scientifiques 14: 87–115
- Mainprice D (2015) Seismic anisotropy of the deep Earth from a mineral and rock physics perspective. In: Schubert G, Bercovici D (eds) *Treatise on geophysics*, vol 2. Elsevier, pp 487–538
- Masters G, Jordan TH, Silver PG, Gilbert F (1982) Aspherical earth structure from fundamental spheroidal mode data. *Nature* 298:609–613
- Montagner J-P (1994) What can seismology tell us about mantle convection? *Rev Geophys* 32:115–137
- Montagner J-P, Kennett BLN (1996) How to reconcile body wave and normal-mode reference earth models? *Geophys J Int* 125:229–248
- Montagner J-P, Tanimoto T (1991) Global upper mantle tomography of seismic velocities and anisotropies. *J Geophys Res* 96:20337–20351
- Montelli R, Nolet G, Dahlen F, Masters G, Engdahl E, Hung S (2004) Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* 303:338–343
- Morgan WJ (1971) Convection plumes in the lower mantle. *Nature* 230:42–43
- Murakami M, Hirose K, Kawamura K, Sata N, Ohishi Y (2004) Post-perovskite phase transition in MgSiO₃. *Science* 304:5672
- Nataf H-C, Nakanishi I, Anderson DL (1994) Anisotropy and shear velocity heterogeneity in the upper mantle. *Geophys Res Lett* 11:109–112
- Nishida K, Montagner J-P, Kawakatsu H (2009) Global surface wave tomography using seismic hum. *Science* 326(5949):112. <https://doi.org/10.1126/science.1176389>
- Oldham RD (1906) The constitution of the Earth. *Q J Geol Soc Lond* 62:456–475
- Pekeris C (1935) Thermal convection in the interior of the Earth. *Geophys J* 3:343–367
- Romanowicz B (1995) A global tomographic model of shear attenuation in the upper mantle. *J Geophys Res* 100:12375–12394
- Romanowicz B, Mitchell BJ (2015) Deep earth structure – Q of the earth from crust to core. In: Romanowicz B, Dziewonski A (eds) *Treatise on geophysics*, vol 1. Elsevier, pp 789–828
- Sengupta MK, Toksoz MN (1976) Three dimensional model of seismic velocity variation in the Earth's mantle. *Geophys Res Lett* 3:84–86
- Shapiro NM, Campillo M, Stehly L, Ritzwoller MH (2005) High-resolution surface-wave tomography from ambient seismic noise. *Science* 307:1615–1618
- Silver PG (1996) Seismic anisotropy beneath the continents: probing the depths of geology. *Annu Rev Earth Planet Sci* 24:385–432

Further Reading

- Anderson DL (2001) Top-down tectonics? *Science* 293:2016–2018
- Becker TW, Kellogg JB, Ekström G, O'Connell RJ (2003) Comparison of azimuthal seismic anisotropy from surface waves and finite-strain from global mantle-circulation models. *Geophys J Int* 155:696–714
- Bercovici D, Karato S-I (2003) Whole-mantle convection and the transition-zone water filter. *Nature* 425:39–44
- Birch F (1952) Elasticity and constitution of the Earth's interior. *J Geophys Res* 57:227–286
- Courtillot V, Davaille A, Besse J, Stock J (2003) Three distinct types of hotspots in the Earth's mantle. *Earth Planet Sci Lett* 205:295–308
- Debaille E, Kennett BLN, Priestley K (2005) Global azimuthal anisotropy and the unique plate-motion deformation of Australia. *Nature* 433:509–512
- Dziewonski AM (1984) Mapping the lower mantle: determination of lateral heterogeneity in P velocity up to degree and order 6. *J Geophys Res* 89:5929–5952
- Dziewonski AM, Anderson DL (1981) Preliminary reference earth model. *Phys Earth Planet Inter* 25:297–356
- Dziewonski AM, Hager BH, O'Connell R (1977) Large-scale heterogeneities in the lower mantle. *J Geophys Res* 82:239–255
- Dziewonski AM, Romanowicz B (2015) Overview. In: Romanowicz B, Dziewonski A (eds) *Treatise on geophysics*, vol 1. Elsevier B.V., pp 1–28

- Souriau A, Calvet M (2015) The Earth's cores. In: Romanowicz B, Dziewonski A (eds) *Treatise on geophysics*, vol 1, Elsevier, pp 725–758
- Tanimoto T, Anderson DL (1985) Lateral heterogeneity and azimuthal anisotropy of the upper mantle: love and Rayleigh waves. 100–250s. *J Geophys Res* 90:1842–1858
- Thurber C, Ritsema J (2015) Theory and observations – seismic tomography and inverse methods. In: Romanowicz B, Dziewonski A (eds) *Treatise on geophysics*, vol 1. Elsevier, pp 307–338
- Trampert J, van Heijst HJ (2002) Global azimuthal anisotropy in the transition zone. *Science* 296:1297–1299
- Tromp J, Tape C, Liu Q (2005) Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophys J Int* 160:195–216
- van der Hilst RD, Widiyantoro S, Engdahl R (1997) *Nature* 386:578–584
- Vinnik L, Kosarev GL, Makeyeva LI (1984) Anisotropy of the lithosphere from the observations of SKS and SKKS phases. *Proc Acad Sci USSR* 278:1335–1339
- Wilson JT (1963) Evidence from islands on the spreading of ocean. *Nature* 197:536–538
- Woodhouse JH, Dziewonski AM (1984) Mapping the upper mantle; three-dimensional modeling of Earth structure by inversion of seismic waveforms. *J Geophys Res* 89:5953–5986

velocity province (LLSVP)
Ultra low velocity zone (ULVZ)

beneath the Atlantic Ocean and Africa. They are continental-sized in lateral dimension, and up to and over 1000 km high. A ULVZ is a thin (tens of kilometers thick) and isolated (100–1000 km laterally) structure with very low seismic wave speeds.

Introduction

Earth's interior has three fundamentally distinct divisions: (1) the relatively buoyant and thin outer crust, which averages roughly 6–7 km in thickness beneath oceans and 30–40 km in continents; (2) the mantle, which extends to nearly half way to Earth's center; and (3) the dominantly iron (Fe) core, which has a fluid outer part and a solid inner part. Figure 1 highlights these basic shells, along with a number of additional significant mantle subdivisions. The mantle is divided into an upper and lower part, with the boundary near 660 km depth. The

Earth's Structure, Lower Mantle

Edward J. Garnero¹, Allen K. McNamara² and James A. Tyburczy¹

¹School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA

²Department of Earth and Environmental Sciences, Michigan State University, East Lansing, MI, USA

Synonyms

Deep Earth; Deep mantle

Definition

Seismic Tomography An inverse method that utilizes seismic wave travel times and/or waveforms to estimate Earth's seismic velocity and density structure, either regionally or globally.

Forward Modeling The procedure involving the generation of models that predicts observations, which is frequently used in deep Earth seismology.

Phase Transition A crystallographic rearrangement of atoms in a mineral that occurs at high pressure.

D'' The depth shell of Earth occupying the lowermost 200–300 km of Earth's lower mantle. (Pronounced “D-double-prime”)

Large low shear Two LLSVPs exist at the base of the mantle, one beneath the Pacific Ocean, and one

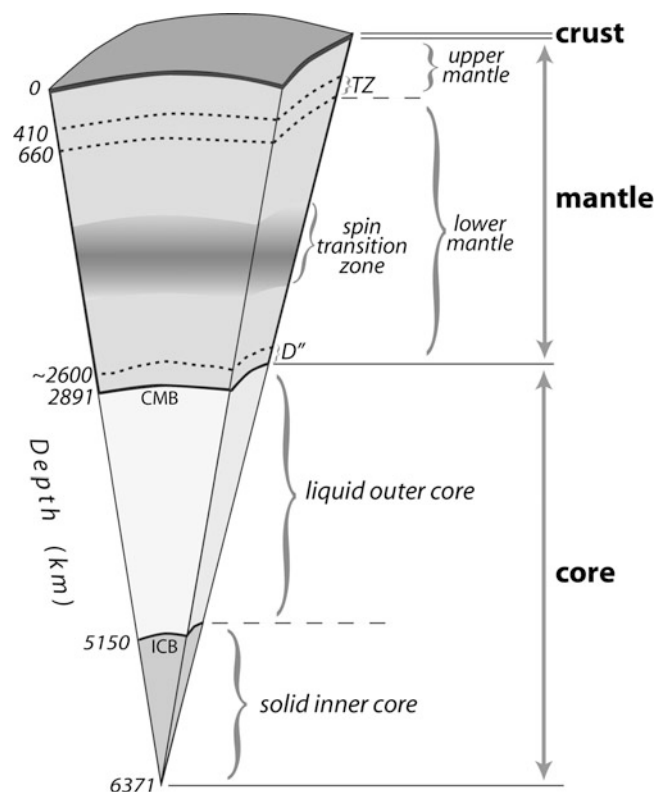


FIGURE 1
Garnero, McNamara, Tyburczy

Earth's Structure, Lower Mantle, Fig. 1 Major depth shells of Earth, along with a number of subdivisions. TZ, transition zone (the upper mantle between 410 and 660 km depth); CMB, core-mantle boundary; ICB, Inner core boundary

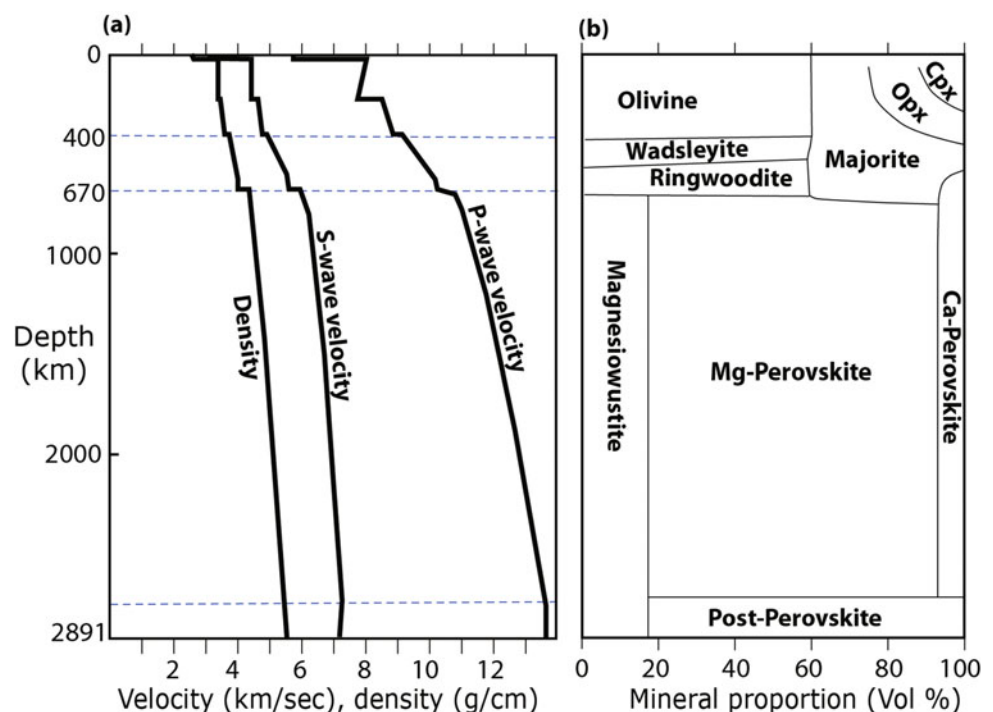
upper mantle has another distinct discontinuity near 410 km depth. As shown in Fig. 2, the dominant mineral in the upper mantle is olivine, a magnesium-iron silicate mineral with composition $(\text{Mg,Fe})_2[\text{SiO}_4]$, that undergoes phase transitions to higher density mineral forms (or polymorphs) near 410, 520, and 660 km depths. The region between 410 and 660 km is referred to as the mantle transition zone. Below 660 km, mantle silicates undergo another phase transformation, compacting into the perovskite structure (recently given the mineral name bridgmanite) with composition $(\text{Mg,Fe})[\text{SiO}_3]$, plus magnesium-iron oxides, $(\text{Mg,Fe})\text{O}$ (known as ferropericlase and also referred to as magnesiowüstite in earlier literature). In the lowermost several hundred kilometers of the mantle, a region referred to as D'' , perovskite has been shown to undergo an additional phase transition into the “post-perovskite” structure (Murakami et al. 2004), and depending on the temperature, post-perovskite might back-transform into perovskite at greater depth, remaining as perovskite down to the CMB (see *The D'' Layer* encyclopedia entry). The lower mantle depth shell accounts for nearly 56% of Earth's volume, and extends from 660 km depth down to the core-mantle boundary (CMB) at 2891 km depth (and thus includes the D'' layer). The major discontinuities throughout Earth's mantle are thus well-explained by phase transitions in the olivine-perovskite system.

While uncertainties are present in our understanding of the exact compositional makeup of Earth's mantle, it is generally believed that bridgmanite is the dominant mineral structure of the lower mantle, possibly accounting for ~80% of the lower

mantle, and hence is the most abundant mineral on Earth. This perovskite mineral structure is the same as that occurring in well-known electronic and piezoelectric materials, such as barium titanate. A fundamental difference between the silicate minerals of the upper mantle and lower mantle perovskite is that every silicon (Si) atom in perovskite is surrounded by six oxygen (O) atoms (silicon is octahedrally coordinated), whereas for upper mantle silicates the silicon is bonded to four O atoms (tetrahedral coordination). Thus a primary effect of the higher pressure in the lower mantle is that O atoms occupy a smaller volume, thus more can be organized around (coordinated with) Si. This more compact rearrangement of atoms in lower mantle phases means that they have higher density than upper mantle materials, as well as differences in other properties such as the material's stiffness and resistance to shearing. These changes affect the velocities of seismic waves; thus, at depths where phase transitions occur, there are jumps (i.e., discontinuous increases) in seismic compressional and shear wave speeds (referred to as P-wave and S-wave velocities, or V_P and V_S , respectively). Velocity and density depth distributions in Earth's mantle are presented in Fig. 2; these values represent an estimation of the globally averaged properties.

The exact depth of a phase transition in any given mineral group depends upon temperature and pressure (and chemical composition). Earlier conceptions of the lower mantle (Birch 1952) postulated that the lower mantle had a constant composition, and hence the temperature and velocity variations could be described by simple adiabatic compression.

Earth's Structure, Lower Mantle, Fig. 2 (a) Variations in seismic properties of the Earth's mantle with depth, from the surface to the core-mantle boundary, showing the major discontinuities in P-wave velocity, S-wave velocity, and density. (b) Mineralogical makeup of the mantle. Cpx, clinopyroxene; Opx, orthopyroxene



However, in light of more recent studies, it is interesting to consider the fact that lateral temperature variations are likely in the lower mantle, because the mantle is convecting, and hence phase boundaries are not expected to be perfectly flat (e.g., Lebedev et al. 2002). Rather, they should have relief that directly relates to the temperature field (or to perturbations in chemical composition, since minerals may transform at different pressures if there are variations or impurities in the mineral composition). The Earth presumably was significantly hotter in the distant geologic past (from heat generated during formation and differentiation of the mantle and core), and therefore we expect phase boundaries in the earliest Earth to have been at different depths than today.

Tools for Studying the Lower Mantle

Earth's mass and moment of inertia, determined from orbital and rotational dynamics, provide important information about the planet's internal density distribution. These data, combined with assumptions about starting materials for Earth's formation from analyses of meteorites, have long indicated a predominantly iron core with a silicate rock mantle (Carlson 2003). Seismic studies early in the twentieth century confirmed the existence of the core (outer and inner), as well as the fact that the outer core is liquid (from the absence of shear waves on the opposite side of the planet following an earthquake, because shear waves cannot propagate in fluids).

A number of Earth science disciplines contribute to our body of knowledge of the lower mantle (and the deep Earth in general). For example, the field of mineral physics includes high pressure laboratory experiments aimed at simulating the pressures and temperatures within the deep Earth. Predictions of the melting temperature of iron at the conditions of the boundary between the inner and outer core (at radius of ~1220 km), as well as measurements delineating the temperature of the olivine to bridgmanite plus ferropericlase phase transition at 660 km depth, provide two important temperature tie-in points that help to constrain Earth's geotherm; temperature must be extrapolated from these points toward the CMB, where a very large temperature jump occurs. Based on such studies, the top of the lower mantle at 660 km depth is at a pressure of about 23 GPa (1 Pa = 1 N m⁻² and 10⁵ Pa corresponds to approximately 1 atmosphere of pressure) and a temperature of about 1800 ± 200 K (e.g., Ito and Katsura 1989; Frost 2008). The CMB is at a pressure of about 130 GPa; the temperature, while less constrained, has been estimated to be about 4000 ± 500 K (Boehler 1996; Van der Hilst et al. 2007) but recently lower values have been proposed (Nomura et al. 2014). The temperature jump from the CMB to the outermost core may be as much as 1000–1500 K.

A number of notable uncertainties are present that currently preclude precise knowledge of lower mantle

temperature, including the amount of heat-producing radioactive materials, the vigor of mantle convection, the heat flux from the fluid core into the mantle, and important thermal properties of bridgmanite at lower mantle pressure and temperature conditions (e.g., thermal conductivity, heat capacity, and the coefficient of thermal expansion). However, using a diamond anvil cell high pressure-device to achieve pressures as great as those of the lower mantle and even the core, researchers have recently refined our understanding about two important characteristics of bridgmanite. As mentioned above, bridgmanite (*br*) can transform into a higher pressure (and density) structure termed “post-perovskite” (*ppv*) at D'' pressures (around 120 GPa). Thus, if the systematics of this phase transition can be well defined from laboratory experiments or theory, then seismic mapping of the *br*→*ppv* transition depth (as well as that of possible *ppv*→*br* back-transformation) adds an additional temperature tie-in point for the geotherm (Hernlund et al. 2005). Another discovery describes a change with increased pressure in the electronic spin state of iron in magnesiowüstite and perovskite (a change from a high spin configuration with a maximum number of unpaired d-shell electrons to a low spin configuration in which d-shell electron pairing is maximized). Experiments indicate that this will occur between the depths of 1000 and 2200 km (Lin et al. 2013; Badro 2014). Current research predicts that this spin transition may result in a softer material, which can thus be more compressed and increase the density of the deeper mantle, possibly affecting convective dynamics as well as seismic velocities, element partitioning, and oxidation state (Shim et al. 2017; Liu et al. 2018).

The vigor of mantle convection plays a central role in the cooling of Earth. The most commonly used parameter to determine if convection is occurring and to describe the strength of convection is the Rayleigh number, *Ra*:

$$Ra = \frac{\alpha g \rho \Delta T D^3}{\eta \kappa},$$

where α is the coefficient of thermal expansion, g is the acceleration due to gravity, ρ is density, ΔT is the non-adiabatic temperature change across the convecting layer (in this case, the whole mantle) of thickness D , η is the viscosity, and κ is the thermal diffusivity. *Ra* in the lower mantle is estimated to be on the order of 10⁷; values greater than ~10³ indicate that convection is occurring and values as high as 10⁷ indicate convection with motions on the order of cm per year, which is considered quite vigorous when considered over geologic timescales. Furthermore, factors such as the temperature dependence of viscosity can strongly affect the style of convection, causing warmer regions to undergo significantly more strain than cooler regions. Thus, numerical convection calculations incorporate our best understanding of Earth's properties and provide important information

about the dynamics and evolution of the mantle (and planet as a whole).

Erupted materials offer an opportunity to study mantle chemistry. Specifically, some hotspot volcanoes are thought to be caused by mantle plumes that originate in the lower mantle. However, it is possible that not all hotspots tap the lower mantle, and instead have different depths of plume source origin (Courtillot et al. 2003). It is widely accepted that lavas from hot spot volcanoes are enriched in incompatible elements and often have a higher $^3\text{He}/^4\text{He}$ ratio than lavas erupted at mid-ocean ridges. This has led to a view that the lower mantle contains a distinct reservoir (or multiple reservoirs) that has undergone only limited stirring with the upper mantle (Hofmann 1997). Subsequent and ongoing work continues to challenge and extend that perspective (Jackson et al. 2017; Stracke et al. 2019). We address geodynamic issues in the “Three-Dimensional Structure and Dynamics” section below.

Seismic imaging is the primary tool that gives us the ability to map lateral variations of mantle properties from the upper mantle to the CMB. Seismic tomography provides an image of the full three-dimensional heterogeneity field, albeit at relatively long horizontal wavelengths, on the order of 1000 km and greater (e.g., Kustowski et al. 2008; Ritsema et al. 2011; French and Romanowicz 2015). Variations in seismic velocity as well as density (e.g., Trampert et al. 2004; Ishii and Tromp 2004; Koelemeijer et al. 2017) have been mapped, and have important implications on convective flow in the interior.

Information about the small-scale heterogeneity field has been provided by a number of forward modeling seismic studies. For example, some studies have suggested that small-scale scatterers (~ 1 km) exist throughout the lower mantle (Hedlin and Shearer 2000). Other studies find evidence for discrete reflectors in some locations, particularly beneath subduction zones (e.g., Rost et al. 2008; Frost et al. 2018). The exact nature of such small-scale heterogeneity is unknown, but may indeed relate to an incompletely mixed mantle, or heterogeneities entrained from the top (dense things falling) or the bottom (buoyant things rising).

Three-Dimensional Structure and Dynamics

For decades it has been known that Earth's mantle contains significant lateral variations in seismic properties, which are due to variations in temperature and/or chemical composition compared to the local average properties. Results from seismic tomography depict V_P , V_S , and ρ changes throughout the mantle. Variations are neither expected nor detected in the fluid outer core; the extremely low viscosity and relatively fast convection promote homogenization (Stevenson 1987). Seismically detected mantle heterogeneity is the strongest in the outermost few hundred kilometers of the mantle (i.e., near

the planet's surface), due to the strong variations in temperature and composition associated with plate tectonic and mantle convection processes (Masters et al. 2000). The next most seismically heterogeneous depth shell is the D'' region at the base of the mantle (see Encyclopedia entry on *The D'' layer*). Figure 3 shows seismic velocity variations in the lower mantle. D'' silicate rock is juxtaposed against the fluid iron outer core, which represents roughly a 75% increase in density. Thus the mantle side of the CMB can accommodate long-lived stable structures over a wide range of densities (between the mantle and core density); it is possible that cumulates denser than average lowest mantle silicates can survive entrainment by the convective currents in the overlying mantle (e.g., McNamara and Zhong 2005; Garnero and McNamara 2008; Garnero et al. 2016). However, other possibilities are viable (Deschamps et al. 2007; Schubert et al. 2009; Davies et al. 2015).

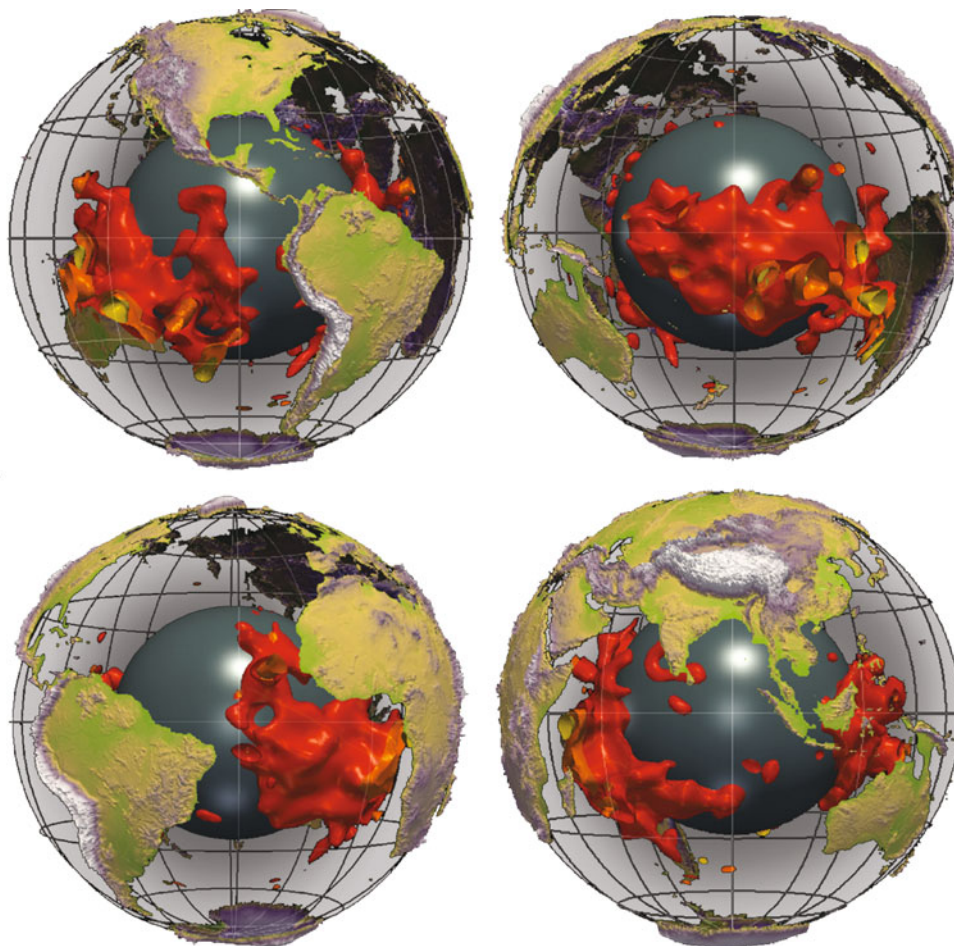
The lower mantle above D'' displays variability in seismic properties, but at a much weaker level than that imaged at the top and bottom of the mantle. In some regions it is plausible that low amplitudes of mid-mantle heterogeneity may reflect our limited ability to confidently image the mantle there (due to a lack of seismic information because earthquakes and seismic recorders are not uniformly distributed on Earth). Nonetheless, all seismic tomography studies to date agree that heterogeneity is greater at the top and bottom of the mantle relative to the middle of the mantle.

Plate tectonics involve the creation of plates at mid-ocean ridges, and destruction or consumption of plates at subduction zones, where the cold and dense oceanic crust and lithosphere (compared to the surrounding mantle) descend into the interior. Tomographic analyses have shown that lineations in high velocity wave speeds beneath subduction zones (e.g., Grand et al. 1997; Li et al. 2008; Obayashi et al. 2013) are consistent with cold material subducting deep into the lower mantle. While the penetration depth of subducting slabs is not adequately constrained beneath all subduction zones, it is widely accepted that many slabs descend through the 660 km discontinuity, in some cases reaching to the CMB (such as beneath the Caribbean). It is important to understand the large-scale long-term flux of material from Earth's surface to great depths in the interior because of the importance for long-term (geological time) effects on surface volatiles, the atmosphere, and global climate.

The amount of mass flux between the upper mantle down into the lower mantle depends on several poorly known quantities such as the viscosity of the descending slab and mantle, the viscosity increase at the 660 km discontinuity, a possible viscosity increase in the upper lower mantle near 1200 km depth (Rudolph et al. 2015), as well as the degree of viscous coupling between the oceanic lithosphere and underlying asthenosphere (before and after subduction). Time-integrated mass transfer into (and out of) the lower mantle

Earth's Structure, Lower

Mantle, Fig. 3 Tomographically derived shear velocity heterogeneity in the lower mantle (French and Romanowicz 2015) is shown for four positions of the globe. Heterogeneity has been iso-surfaced (red) in the lower half of Earth's mantle at the -0.75% level. The two large low shear velocity provinces (LLSVPs) are present



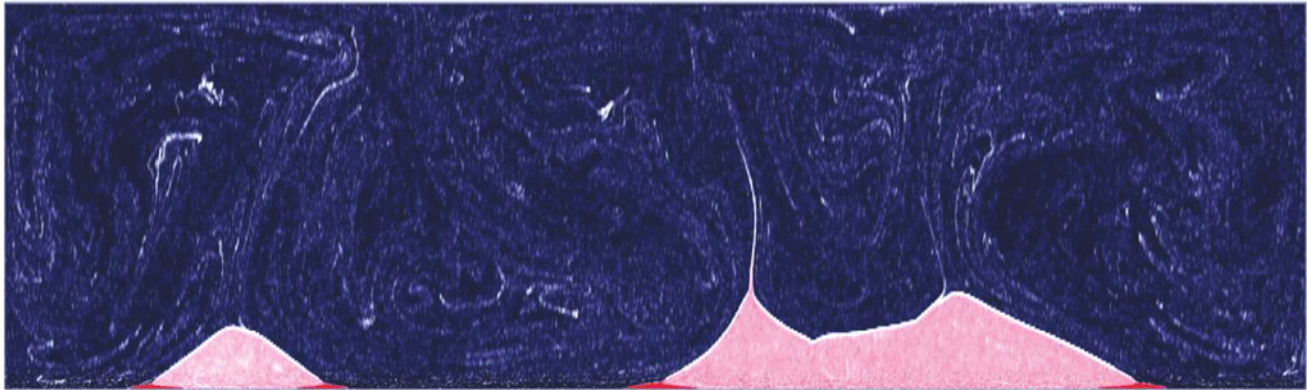
plays a role in long-term mixing and lower mantle residence time of deep mantle minerals. For example, weak viscous coupling between slabs and the surrounding mantle predicts lower flux rates into (and out of) the lower mantle, and hence less mixing between the upper and lower mantle. A strong coupling between the slab and mantle predicts the opposite. Uncertainties in current data and predictions result in difficulties in distinguishing between these possibilities.

An analogous dependency between mixing and convective parameters exists for the bottom of the mantle, where recent work argues for large low shear velocity provinces (referred to as LLSVPs) being compositionally distinct from the surrounding mantle (Cottaar and Lekic 2016; Garnero et al. 2016; Deschamps et al. 2019; Roy et al. 2019). Convective currents associated with subduction sweep basal dense material to regions away from downwellings, forming thermochemical “piles” (Fig. 4) (McNamara and Zhong 2005; McNamara et al. 2010). The contrast in properties (e.g., density, viscosity, temperature) between piles and surrounding mantle will determine the morphology and vertical extent of the pile material, as well as the degree of entrainment (and

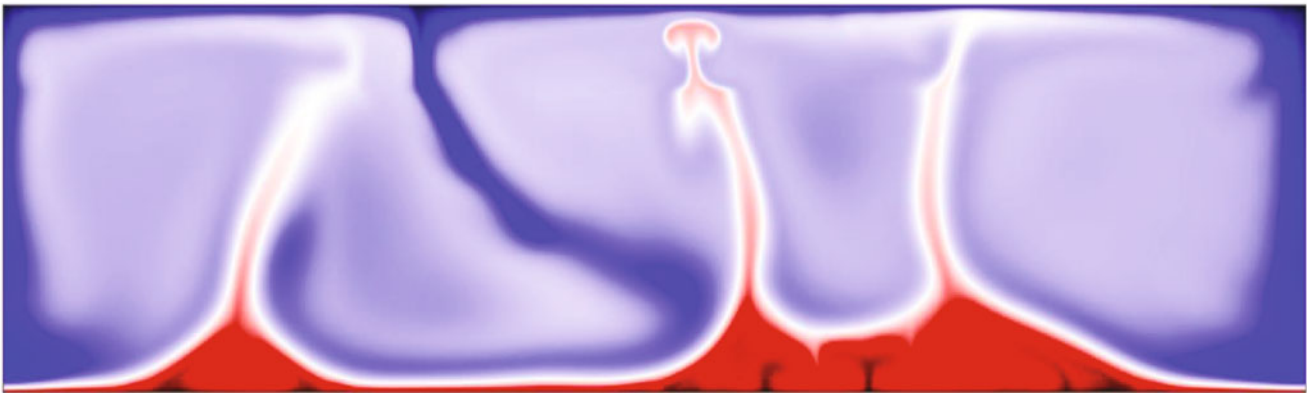
hence its evolutionary behavior). The two nearly antipodal large low shear velocity provinces, or “LLSVPs” (Cottaar and Lekic 2016; McNamara 2019), one beneath the central Pacific and the other beneath the southern Atlantic and Africa (Fig. 3), are accompanied (in certain areas) by low velocity conduits that are associated with mantle plumes (e.g., Montelli et al. 2004), which appears to correlate with the surface locations of hotspots and reconstructed origination locations of large igneous provinces (Burke et al. 2008; Torsvik et al. 2010). Thus the present-day large-scale structure and circulation pattern of the lower mantle represent a time-integrated effect of anomalous temperature and/or chemistry in Earth's lowermost and uppermost mantle boundary layers (D'' and the lithosphere, respectively).

At smaller scale lengths, seismic studies have found evidence for seismic wave scattering. The imaged size of heterogeneities in the lower mantle ranges from kilometer to tens of kilometers in scale (e.g., Vidale and Hedlin 1998; Wen and Helmberger 1998; Frost et al. 2017). This observation indicates chemical anomalies at a spectrum of physical scales and a mantle that is not fully mixed. Indeed, analyses of magmas

(a) Chemically distinct regions: mantle (blue), lowermantle reservoirs (pink), and ultra-low velocity zones at the core-mantle boundary (red)



(b) Temperature variations cooresponding to (a)



Earth's Structure, Lower Mantle, Fig. 4 Numerical geodynamics calculations for a model mantle consisting of three distinct chemistries are shown for the whole mantle, from Earth's surface to the core-mantle boundary. Panel (a) displays the chemical field, showing the bulk of the upper and lower mantle in blue, as one chemistry type, dense lower mantle piles (pink) as a second distinct chemistry, and thin ultra-low

velocity zones at the core mantle boundary (red) that accrue near the pile margins. (b) The temperature field associated with (a) is shown. Red and blue are hotter and colder regions, respectively. The dense piles can be stable for billions of years, though a low level of entrainment is constant, as evidenced by the white streaks throughout the mantle in panel (McNamara et al. 2010)

from hot spot volcanoes have yielded characteristic isotopic signatures distinct from those of mid-ocean ridge basalts, arguing for unique, possibly isolated deep mantle reservoirs that feed mantle plumes. Irregular convective patterns (e.g., see Fig. 4) are consistent with such heterogeneities spreading through the lower mantle as well as being entrained into mantle plumes, but not necessarily in a systematic fashion (since convection patterns appear multi-scaled and may be complex). Alternatively, scattering of seismic energy can occur from small pockets of partially molten mantle material. This scenario has been motivated by the detection of thin (a few kilometers) laterally discontinuous layers right at the CMB, called ultra-low velocity zones (ULVZs). ULVZs have been argued to contain some degree of partial melt, and hence offer a plausible source of scatters composed of melt pockets (whether or not they are of a distinct chemical composition). Subduction of oceanic crust into the deep mantle also

represents a viable source of lower mantle heterogeneity (e.g., Li et al. 2014; Mulyukova et al. 2015).

Future Progress

As seismic sensors are deployed in new regions across Earth's surface, new geographical regions will be sampled with greater certainty, and thus our understanding of global processes will greatly improve. As sensor deployments become denser (with smaller distances between sensors), Earth's deep structure can be imaged with a much greater resolution. Thus, small-scale structures and processes can be inferred with greater certainty and tied to estimations of global processes. The seismic work only maps present-day structure, which must be put in a dynamical and evolutionary framework (e.g., Hernlund and McNamara 2015) with the predictions

from geodynamical convection calculations, which in turn must first be informed by an understanding of the material properties. The latter comes from the field of mineral physics, both in laboratory experiments at high temperature and pressure and with computer simulations of material behavior at the atomic scale. The time evolution of Earth is also informed by geochemical analyses of erupted materials and meteorites. Thus, future work that advances our understanding of the structure, dynamics, and evolution of Earth's interior will be multidisciplinary in nature.

Summary

Earth's lower mantle represents the largest volume of any depth shell on Earth. It is unique in that its chemistry, structure, dynamics, and evolution represent a time-integrated effect of the chemistry and dynamics of the surface (lithosphere, asthenosphere) and lowermost mantle (D'', ULVZ, LLSVP, and CMB) boundary layers. The degree to which the lower mantle is recycled into the upper mantle depends upon many poorly known convective parameters, such as the viscosity and viscous coupling of descending slabs, as well as material properties, such as the nature and origin of plausibly dense deep mantle chemically distinct piles. Future work of improved seismic imaging coupled with continued advancements in other deep Earth disciplines, such as geodynamics, mineral physics, and geochemistry, will greatly help to reduce uncertainties in our understanding of Earth's evolutionary pathway, and present-day structure and dynamical state.

Cross-References

- [Core-Mantle Coupling](#)
- [Earth's Structure, Core](#)
- [Earth's Structure, Upper Mantle](#)
- [Mantle Convection](#)
- [Mantle D'' Layer](#)
- [Mantle Viscosity](#)

Bibliography

- Badro J (2014) Spin transitions in mantle minerals. *Annu Rev Earth Planet Sci* 42:231–248. <https://doi.org/10.1146/annurev-earth-042711-105304>
- Birch F (1952) Elasticity and constitution of the Earth's interior. *J Geophys Res* 57(2):227–286
- Boehler R (1996) Melting and temperature of the Earth's mantle and core. *Annu Rev Earth Planet Sci* 24:15–40
- Burke K, Steinberger B, Torsvik TH, Smethurst MA (2008) Plume generation zones at the margins of large low shear velocity provinces on the core–mantle boundary. *Earth Planet Sci Lett* 265:49–60
- Carlson RW (2003) In: Carlson RW (ed) *Treatise on geochemistry*, volume 2: Geochemistry of the mantle and core. Elsevier, 586pp
- Cottaar S, Lekic V (2016) Morphology of seismically slow lower-mantle structures. *Geophys J Int* 207:1122–1136. <https://doi.org/10.1093/gji/ggw324>
- Courtillot V, Davaille A, Baesje J, Stock J (2003) Three distinct types of hotspots in the Earth's mantle. *Earth Planet Sci Lett* 205:295–308
- Davies DR, Goes S, Sambridge M (2015) On the relationship between volcanic hotspot locations, the reconstructed eruption sites of large igneous provinces and deep mantle seismic structure. *Earth Planet Sci Lett* 411:121–130
- Deschamps F, Trampert J, Tackley PJ (2007) Thermo-chemical structure of the lower mantle: seismological evidence and consequences for geodynamics. In: Yuen DA et al (eds) *Superplume: beyond plate tectonics*. Springer, Dordrecht, pp 293–320
- Deschamps F, Konishi K, Fuji N, Cobden L (2019) Radial thermo-chemical structure beneath Western and Northern Pacific from seismic waveform inversion. *Earth Planet Sci Lett* 520:153–163. <https://doi.org/10.1016/j.epsl.2019.05.040>
- French SW, Romanowicz B (2015) Broad plumes rooted at the base of the Earth's mantle beneath major hotspots. *Nature* 525:95–99
- Frost DJ (2008) The upper mantle and transition zone. *Elements* 4:171–176
- Frost DA, Rost S, Garnero EJ, Li M (2017) Seismic evidence for Earth's crusty deep mantle. *Earth Planet Sci Lett* 470:54–63
- Frost DA, Garnero EJ, Rost S (2018) Dynamical links between small- and large-scale mantle heterogeneity: seismological evidence. *Earth Planet Sci Lett* 482:135–146
- Garnero EJ, McNamara AK (2008) Structure and dynamics of Earth's lower mantle. *Science* 320:626–628
- Garnero EJ, McNamara AK, Shim S-H (2016) Continent-sized anomalous zones with low seismic velocity at the base of Earth's mantle. *Nat Geosci* 9(7):481–489. <https://doi.org/10.1038/ngeo2733>
- Grand SP, van der Hilst RD, Widiyantoro S (1997) Global seismic tomography: a snapshot of convection in the Earth. *GSA Today* 7:1–7
- Hedlin MAH, Shearer PM (2000) An analysis of large scale variations in small-scale mantle heterogeneity using Global Seismic Network recordings of precursors to PKP. *J Geophys Res* 105:13,655–13,673
- Hernlund JW, McNamara AK (2015) The core-mantle boundary region. In: Schubert G (editor-in-chief) *Treatise on geophysics*, 2nd edn, vol 7. Elsevier, Oxford, pp 461–519
- Hernlund JW, Thomas C, Tackley PJ (2005) A doubling of the post-perovskite phase boundary and structure of the Earth's lowermost mantle. *Nature* 434:882–886
- Hofmann AW (1997) Mantle geochemistry: the message from oceanic volcanism. *Nature* 385:219–229
- Ishii M, Tromp J (2004) Constraining large-scale mantle heterogeneity using mantle and inner-core sensitive normal modes. *Phys Earth Planet Inter* 146:113–124
- Ito E, Katsura T (1989) A temperature profile of the mantle transition zone. *Geophys Res Lett* 16:425–428
- Jackson MG, Konter JG, Becker TW (2017) Primordial helium entrained by the hottest mantle plumes. *Nature* 42:340–343. <https://doi.org/10.1038/nature21023>
- Koelmeijer P, Deuss A, Ritsema J (2017) Density structure of Earth's lowermost mantle from Stoneley mode splitting observations. *Nat Commun* 8:15241. <https://doi.org/10.1038/ncomms15241>
- Kustowski B, Ekström G, Dziewonski AM (2008) Anisotropic shear-wave velocity structure of the Earth's mantle: a global model.

- J Geophys Res Solid Earth (1978–2012) 113(B6):B06306. <https://doi.org/10.1029/2007JB005169>
- Lebedev S, Chevrot S, Van der Hilst RD (2002) Seismic evidence for olivine phase changes at the 410- and 660-kilometer discontinuities. *Science* 296:1300–1302
- Li C, van der Hilst RD, Engdahl ER, Burdick S (2008) A new global model for P wave speed variations in Earth's mantle. *Geochim Geophys Geosyst* 9(5):Q05018
- Li M, McNamara AK, Garnero EJ (2014) Chemical complexity of hotspots caused by cycling oceanic crust through mantle reservoirs. *Nat Geosci* 7:366–370
- Lin J-F, Speziale S, Mao Z, Marquardt H (2013) Effects of electronic spin transitions of iron in lower mantle minerals: implications for deep mantle geophysics and geochemistry. *Rev Geophys* 51:244–275. <https://doi.org/10.1002/rog.20010>
- Liu J, Dorfman SM, Zhu F, Li J, Wang Y, Zhang D, Xiao Y, Bi W, Alp EE (2018) Valence and spin states of iron are invisible in Earth's lower mantle. *Nat Commun* 9:1284. <https://doi.org/10.1038/s41467-018-03671-5>
- Masters G, Laske G, Bolton H, Dziewonski AM (2000) The relative behavior of shear velocity, bulk sound speed, and compressional velocity in the mantle: implications for chemical and thermal structure. In: Karato S-I, Forte AM, Liebermann RC, Masters G, Stixrude L (eds) *Earth's deep interior: mineral physics and tomography from the atomic to the global scale*. AGU, Washington, DC, pp 63–87
- McNamara AK (2019) A review of large low shear velocity provinces and ultra low velocity zones. *Tectonophysics* 760:199–220
- McNamara AK, Zhong S (2005) Thermochemical structures beneath Africa and the Pacific Ocean. *Nature* 437:1136–1139
- McNamara AK, Garnero EJ, Rost S (2010) Tracking deep mantle reservoirs with ultra low velocity zones. *Earth Planet Sci Lett* 299:1–9
- Montelli R, Nolet G, Dahlen F, Masters G, Engdahl E, Hung S (2004) Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* 303:338–343
- Mulyukova E, Steinberger B, Dabrowski M, Sobolev SV (2015) Survival of LLSVPs for billions of years in a vigorously convecting mantle: replenishment and destruction of chemical anomaly. *J Geophys Res Solid Earth* 120:3824–3847. <https://doi.org/10.1002/2014JB011688>
- Murakami M, Hirose K, Kawamura K, Sata N, Ohishi Y (2004) Post-perovskite phase transition in MgSiO_3 . *Science* 304(5672)
- Nomura R, Hirose K, Uesugi K, Ohishi Y, Tsuchiyama A, Miyake A, Ueno Y (2014) Low core-mantle boundary temperature inferred from the solidus of pyrolite. *Science* 343:522–525
- Obayashi M, Yoshimitsu J, Nolet G et al (2013) Finite frequency whole mantle P wave tomography: improvement of subducted slab images. *Geophys Res Lett* 40:5652–5657
- Ritsema J, Deuss A, van Heijst HJ, Woodhouse JH (2011) S40RTS: a degree-40 shear-velocity model for the mantle from new Rayleigh wave dispersion, teleseismic traveltime and normal-mode splitting function measurements. *Geophys J Int* 184:1223–1236
- Rost S, Garnero EJ, Williams Q (2008) Seismic array detection of subducted oceanic crust in the lower mantle. *J Geophys Res* 113: B06303. <https://doi.org/10.1029/2007JB005263>
- Roy SK, Takeuchi N, Srinagesh D, Kumar MR, Kawakatsu H (2019) Topography of the western Pacific LLSVP constrained by S-wave multipathing. *Geophys J Int* 218(1):190–199. <https://doi.org/10.1093/gji/ggz149>
- Rudolph ML, Lekic V, Lithgow-Bertelloni C (2015) Viscosity jump in Earth's mid-mantle. *Science* 350(6266):1349–1352
- Schuberth BSA, Bunge H-P, Ritsema J (2009) Tomographic filtering of high-resolution mantle circulation models: can seismic heterogeneity be explained by temperature alone? *Geochim Geophys Geosyst* 10: Q05W03
- Shim S-H, Grocholski B, Ye Y, Alp EE, Xu S, Morgan D, Meng Y, Prakapenka VB (2017) Stability of ferrous-iron-rich bridgmanite under reducing midmantle conditions. *Proc Natl Acad Sci* 114(25):6468–6473. <https://doi.org/10.1073/pnas.1614036114>
- Stevenson DJ (1987) Limits on lateral density and velocity variations in the Earth's outer core. *Geophys J R Astron Soc* 88:311–319
- Stracke A, Genske F, Berndt J, Koornneef J (2019) Ubiquitous ultra-depleted domains in Earth's mantle. *Nat Geosci* 12:851–855. <https://doi.org/10.1038/s41561-019-0446-z>
- Torsvik TH, Burke K, Steinberger B, Webb SJ, Ashwal LD (2010) Diamonds sampled by plumes from the core-mantle boundary. *Nature* 466:352–355
- Trampert J, Deschamps F, Resovsky J, Yuen DA (2004) Probabilistic tomography maps chemical heterogeneities throughout the mantle. *Science* 306:853–856
- Van der Hilst RD, de Hoop MV, Wang P, Shim S-H, Ma P, Tenorio L (2007) Seismostratigraphy and thermal structure of Earth's core-mantle boundary region. *Science* 315:1813–1817
- Vidale JE, Hedlin MAH (1998) Evidence for partial melt at the core-mantle boundary north of Tonga from the strong scattering of seismic waves. *Nature* 391:682–685
- Wen L, Helmberger DV (1998) Ultra-low velocity zones near the core-mantle boundary from broadband PKP precursors. *Science* 279:1701–1703

Earth's Structure, Upper Mantle

Guust Nolet

IRD Geoazur and Université Côte d'Azur,
Sophia Antipolis, France

Definition

The upper mantle is defined as that part of the mantle between the crust and the phase transition of γ -olivine to perovskite. The total mass of the upper mantle is 1.06×10^{24} kg, about a quarter of the total mass of the mantle. Its volume, 2.95×10^{11} km³, is a third of the total volume of the mantle.

Until the discovery of a major transition near 660-km depth, the upper mantle was better known as Bullen's "layer B," extending to 400-km depth. A sharp gradient or discontinuity in seismic velocity causes a bend in the travel time curves of P-wave arrivals near 20°, such that the slowness drops from more than 12 s/deg. to 10 s/deg. This phenomenon was observed as early as 1931 by Jeffreys, who correctly adopted a suggestion by J.D. Bernal that it represents a high-pressure modification of olivine. The modern notion of the upper mantle actually extends into Bullen's next layer, "layer C," which has now been abandoned as a physically meaningful subdivision.

The depths to the top and the bottom of the upper mantle cannot be given precisely because they depend on the region. The Mohorovičić discontinuity marks the transition from the crust to the upper mantle, with a density jump of the order of 10%. It is found at very shallow depth near ocean ridges where oceanic crust is formed but extends to more than 70-km depth beneath the Himalayas and the Andes. The “transition zone” between upper and lower mantle is marked by the transition of olivine (Mg_2SiO_4) to its polymorph β -olivine (Wadsleyite) near 410-km depth to γ -olivine near 520-km depth and finally to perovskite (MgSiO_3), which occurs near a depth of 660 km, but with variations of ± 20 km depending on temperature and composition and possibly with variations more than twice that large for narrow areas inside hot plumes and cold slabs. A typical temperature estimate for the phase transition gives a temperature of about 1600 °C at 660-km depth. The upper mantle is one to two orders of magnitude less viscous than the lower mantle, and its electrical conductivity is much lower, with the possible exception of regions of partial melt.

The upper mantle itself is subdivided into several depth regions, characterized by their differences in density, mechanical properties, and/or seismic velocity:

Lithosphere It is a strong layer with a viscosity of the order of 10^{21} Pa s and a thickness typically about 80–150 km, though it may exceed 200 km beneath old cratons, representing the rigid “plates.” The temperature gradient is strongly superadiabatic. The geotherm in the lithosphere connects the low temperature at the bottom of the crust, 600–900 °C under continents but as low as 250 °C in the oceans, to the (approximately adiabatic) temperature in the asthenosphere (1200–1300 °C).

Gutenberg discontinuity The transition between lithosphere and asthenosphere, historically known as the Gutenberg discontinuity, is defined as the depth where shear velocity starts to decrease with depth. The most likely cause is that the weakening effect of temperature exceeds the effect of pressure at this depth. However, where it is sharp the rock composition, melt and/or the presence of water must also play a role. Sharp transitions may be mapped using receiver functions, in which case the boundary is often denoted as the lithosphere–asthenosphere boundary or LAB. Byrnes et al. (2015) discuss the various options for this LAB.

Asthenosphere It is a mechanically weak layer with low viscosity of the order of 10^{19} Pa s. This layer is present under the oceans and the younger part of continents, but its existence beneath the oldest cratons is uncertain. It may contain a small degree of partial melt.

Lehmann discontinuity It is a positive velocity jump named after Inge Lehmann who first observed this feature beneath

Europe and North America near 150-km depth. It is generally interpreted as a sharp boundary representing the bottom of the asthenosphere, but not globally observed (Gu et al. 2001).

“410-km” discontinuity It is a globally observed jump over a narrow depth interval in seismic velocity, marking the transition from the $\alpha \rightarrow \beta$ phase of olivine near 400-km depth. This transition occurs at a temperature of about 1400–1450 °C.

“520-km” discontinuity Predicted by the phase diagram for olivine, which transforms to the γ phase under certain temperature conditions, this gradient in seismic velocities is occasionally observed.

Reflections or refractions from other depths have been observed – such as the “Hales discontinuity” within the lithosphere – and are occasionally attributed to a global physical or chemical transition, but a lack of consistency between such observations makes it more likely that the observations should be explained as wave energy returned from three-dimensional structure rather than from a discontinuity that represents an omnipresent layering.

Early regionalized refraction studies of the upper mantle already showed a marked variability in P and S velocity with depth. Surface wave studies, which are the preferred tool to investigate layers of low velocity, indicate minimum S velocities as low as 4.05 km/s under the Western United States, 4.09 km/s under the East African Rift, and 4.14 km/s under the Pacific Ocean, but body waves show a minimum S velocity as high as 4.63 km/s under the North American craton.

The Lehmann discontinuity is observed in few regional models, but it is present in the Preliminary Reference Earth Model (PREM) (Dziewonski and Anderson 1981), which shows a major velocity increase in P velocity of 7.1% at 220 km. By way of compensation, PREM has only a minor increase of 2.6% from 8.90 to 9.13 km/s at the “410-km” discontinuity. This, however, disagrees with all regional observations and reveals a major shortcoming of the reference model PREM. In fact, the variability in the velocity jump at 410 km is not very pronounced in the early refraction models. Typically, the P velocity in such models jumps about 5%. An average over seven regional models gives a P velocity jump from 8.78 ± 0.06 km/s to 9.26 ± 0.03 km/s, where the uncertainty represents one standard deviation. The jump in P velocity at 660 km over the same set of models is roughly 4% but more variable than that at 410 km: from 10.22 ± 0.20 to 10.72 ± 0.10 km/s. Near 660-km depth model PREM is in agreement with regional studies.

Though the regional models based on refracted waves provide the most accurate estimates of *absolute* velocities, the velocity–depth relationships represent horizontal averages

over thousands of km. We know that velocity variations exist over length scales much smaller than that. New tomographic techniques are able to delineate such variations and allow us to image them.

Tomographic Techniques

Knowledge of the three-dimensional upper mantle structure comes from tomographic techniques, notably:

- Transmission tomography of body wave delay times for both P and S velocity structure.
- Phase or group delay tomography using surface waves, primarily sensitive to S velocity.
- Imaging of waves reflected or converted at discontinuities (“receiver functions”), sensitive to the depth location of such discontinuities.
- Waveform modelling is able to combine the information of the previous three wave types in one inversion, though at the expense of very large computation times.
- Inversion of the splitting of normal mode frequencies by lateral heterogeneity, for long wavelength heterogeneity in both P and S velocity.

For an extensive overview of transmission tomographic techniques, see ► [“Seismic Tomography”](#). Each of these techniques has seen important progress in recent times. Ray theory is being replaced with first-order scattering theory that allows for the modelling of the frequency-dependent dispersion in delay times (Sigloch et al. 2008) or in receiver functions (Deng and Zhou 2015). Single station (or “single channel”) stacking of converted or reflected phases is being replaced by multichannel techniques, in particular Kirchhoff-type stacking along diffraction hyperboles much as one does to migrate seismic sections in exploration seismics (see Bostock 2009).

Seismic tomography allows the mapping of “anomalies” or deviations from a regular layering of the upper mantle. The most remarkable anomalies are listed here:

Subduction zones and the associated low velocity structure in the mantle wedge can be very clearly imaged because of increased seismicity when actively subducting. The observed velocity anomalies are consistent with average (cold) temperature anomalies of several hundred degrees.

Oceanic ridges are mapped best with surface waves, which leaves the exact depth extent as well as the temperature anomaly somewhat uncertain.

Cratonic roots were originally detected through the large negative delay times they cause for S waves, consistent

with a cold lithosphere (as much as 400° colder than oceanic lithosphere), but the depth extent is mapped using teleseismic surface waves. In the tectosphere model of Jordan (1981), the high density induced by the cooling is compensated by a chemical composition that is depleted in basaltic components.

Upper mantle plumes are rather narrow anomalies in the upper mantle but can be imaged adequately using body wave delays as well as surface wave observations from field deployments of portable seismographs. The velocity anomalies are consistent with temperature anomalies of 100–300 °C inferred from petrological studies.

Topography of the “410”- and “660”-km discontinuity is globally imaged using underside reflections of waves such as SS. This involves some geographic averaging, but topographic anomalies of 20 km or more are observed.

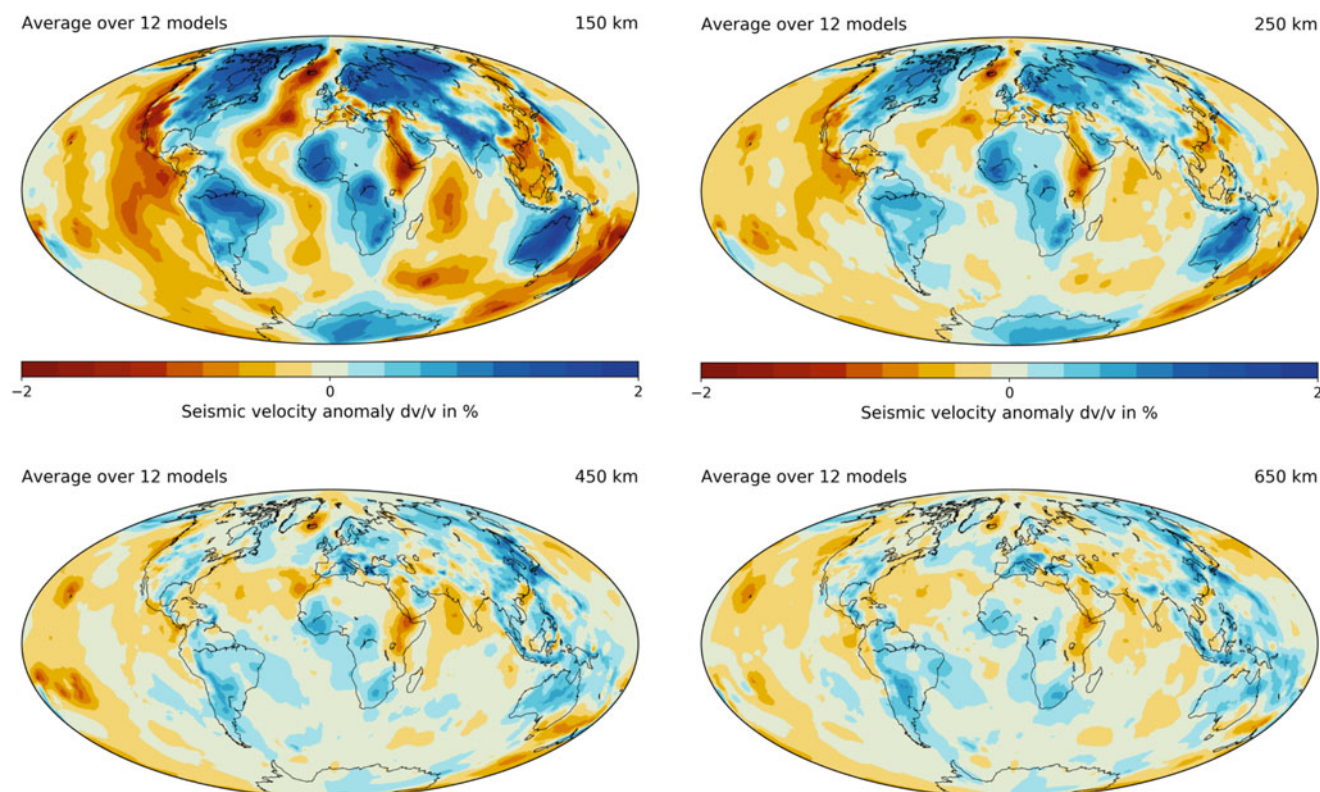
These anomalies are all visible in the variations of isotropic velocity variations. There is strong evidence, notably from the birefringence of SKS waves, that the seismic velocity in the shallow part of the upper mantle depends locally on the propagation direction of the wave; under favorable circumstances, this could in principle allow the mapping of mantle flow, since the crystal structure is assumed to align itself to the flow direction. However, at present there is little agreement between different tomographic models of anisotropy, partly because the SKS information is insensitive to depth, while the crustal correction, which is badly known, is in itself as large as the signal from anisotropy in surface waves (Ferreira et al. 2010).

Global Models

Schaeffer and Lebedev (2013) developed a detailed tomographic model of shear velocity variations in the upper mantle and compare this with five other models. At a length scale of thousands of km, all models are in agreement, but large differences still exist at shorter scales (500–1000 km). Very low velocities beneath oceanic ridges extend to about 100-km depth. The roots beneath the old cratons do not seem to extend far beyond 200-km depth.

Hosseini et al. (2018) collect 36 global models, not all independent. Figure 1 shows the upper mantle P-wave perturbations averaged over 12 of these models. Figure 2 shows the average S-wave perturbations averaged over 17 models. The ability to resolve depth is superior for S-wave models since these are partly based on surface wave information.

The averages displayed in Figs. 1 and 2 illustrate the global features that are present in all tomographic models: old continental shields (Canada, Scandinavia, Western



Earth's Structure, Upper Mantle, Fig. 1 An average over 12 tomographic models of P-wave velocity perturbations, plotted at 4 different depths in the upper mantle using the web-based toolbox of Hosseini et al. (2018)

Australia, West and South Africa) are characterized by high P and S velocities down to depths of about 250 km. The ridge in the North Atlantic interacts with the Iceland mantle plume and is still clearly visible at 250-km depth that shows low P and S anomalies. The oceans are generally characterized by low velocities throughout much of the upper mantle, with the exception of the South Atlantic and the western-most Pacific. Plume-like anomalies, with a high degree of vertical continuity in the upper mantle, are visible in the P velocity model beneath Hawaii, the Society Islands (Tahiti), and the East-African rift area (Afar). For a proper imaging of plumes, tomographic techniques that correct for the effects of diffraction around small or narrow anomalies are more suitable. Figure 3 shows models PRI-05 for P and S velocity in the upper mantle from Montelli et al. (2006), which clearly show a number of narrow upper mantle plumes.

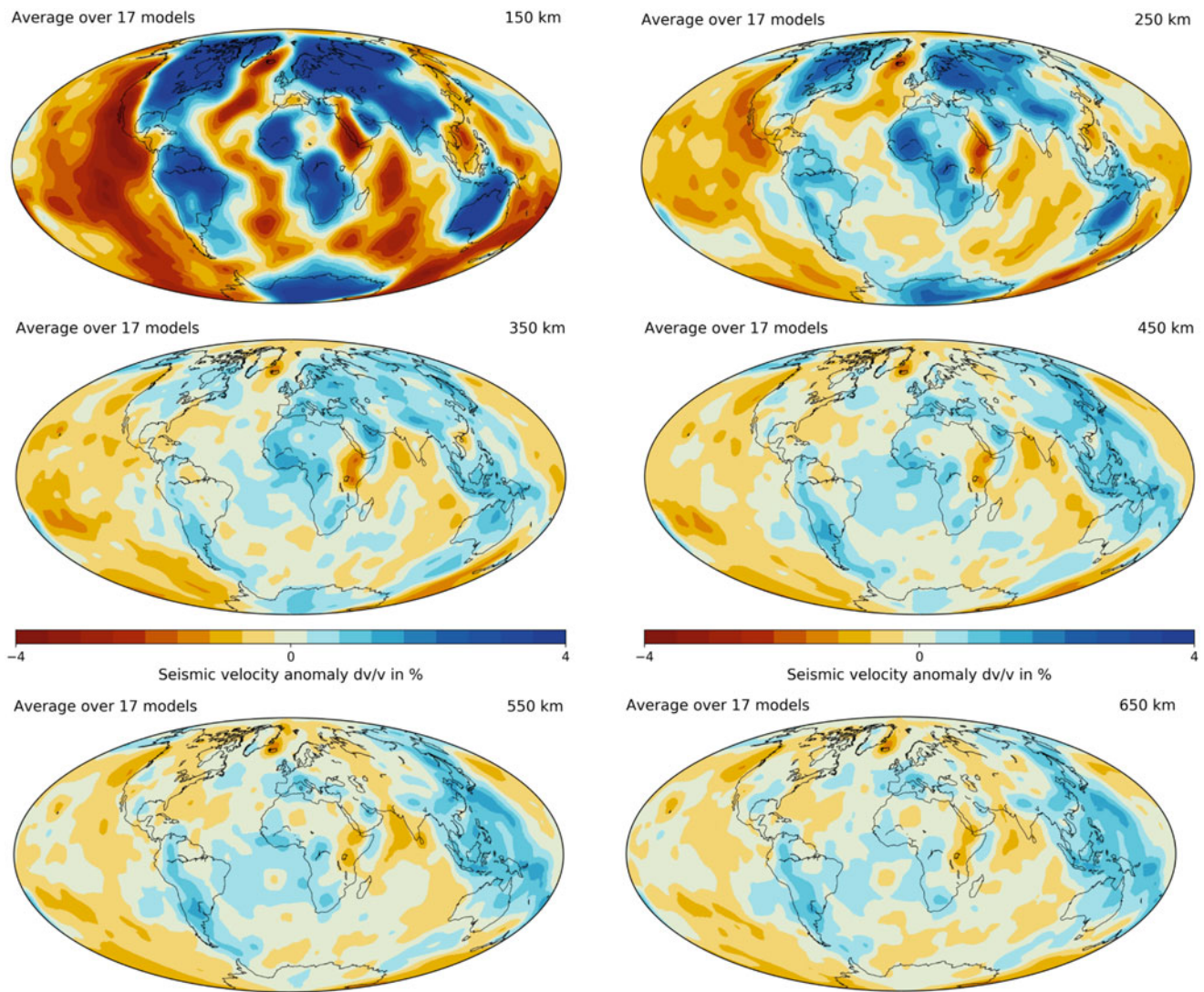
Underside reflections from the upper mantle discontinuities allow the depth variations to the “410-” and “660-km” discontinuity to be mapped globally. This has most recently been done by Houser et al. (2008) using precursors to SS waves (Fig. 4).

Regional Models

Regional studies are able to offer a better resolution than the global efforts, certainly when using data from large and dense seismic arrays such as Skippy (Australia), USArray (USA), or HiNet (Japan), and other, usually temporary, deployments of broadband seismographs. Efforts to image the upper mantle beneath them have led to very precise images of subduction zone structure in particular. Figure 5 shows the slab subducting in the Fiji–Tonga region.

Though this slab image is continuous down to 600-km depth, slab geometry in the upper mantle is very variable, whereas some slabs remain for at least some time lodged in the transition zone others have no trouble sinking through the phase transition barrier at 660-km depth, with a negative Clapeyron slope and an increase of as much as two orders of magnitude in viscosity (Fukao et al. 2001).

A very different fate of subducting slabs is shown for the now inactive Farallon subduction beneath western North America, where the former Farallon plate is slowly disintegrating, letting mantle rock flow through major ruptures such as the “slab gap” in Fig. 6 (Sigloch et al. 2008).



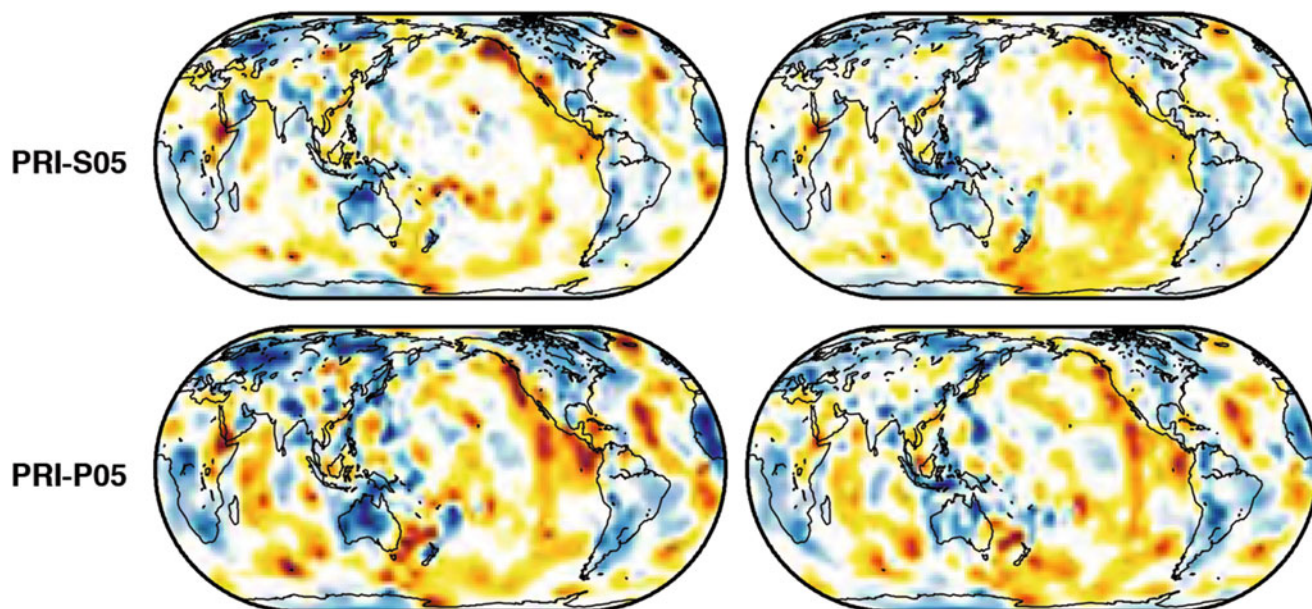
Earth's Structure, Upper Mantle, Fig. 2 An average over 17 tomographic models of S-wave velocity perturbations, plotted at 6 different depths in the upper mantle using the web-based toolbox of Hosseini et al. (2018)

Outlook

It is now clear that the simple, layered models of the upper mantle are inadequate to explain many of the dynamic processes we observe at the Earth's surface. The resolving power of tomographic techniques is still growing because of theoretical improvements and the densification of seismic networks. The coming decades will undoubtedly reduce the high degree of nonuniqueness that still plagues tomographic images and allows for more reliable estimates of temperature and anisotropy. Floating seismic robots (see ► [“Floating Seismographs \(MERMAIDS\)”](#)) are opening up the oceans for dense seismic observations.

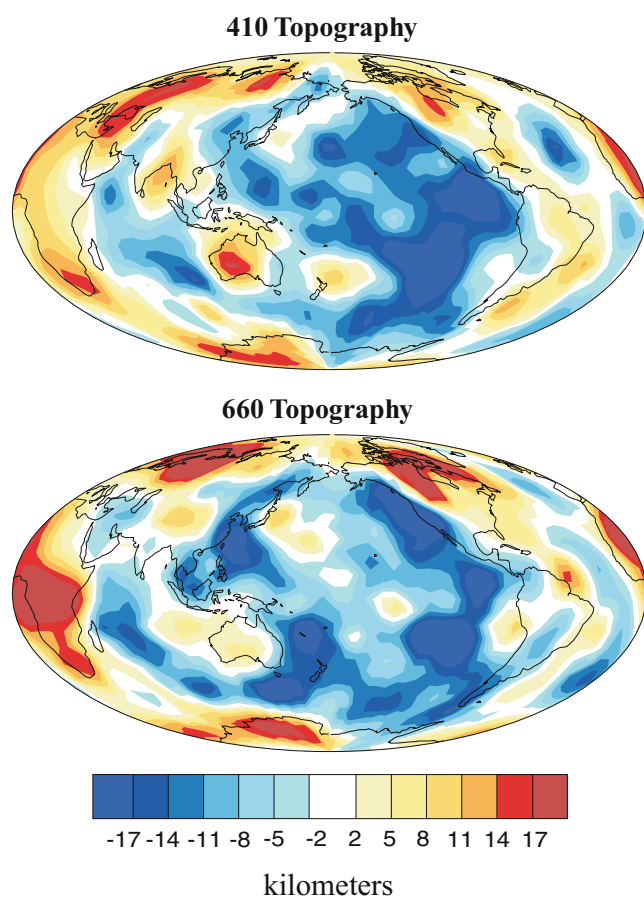
Cross-References

- [Body Waves](#)
- [Earth's Structure, Global](#)
- [Floating Seismographs \(MERMAIDS\)](#)
- [Geodynamics](#)
- [Gravity Field of the Earth](#)
- [Lithosphere, Continental](#)
- [Lithosphere, Oceanic: Thermal Structure](#)
- [Mantle Plumes](#)
- [Plate-Driving Forces](#)
- [Seismic Tomography](#)
- [Seismic, Receiver Function Technique](#)

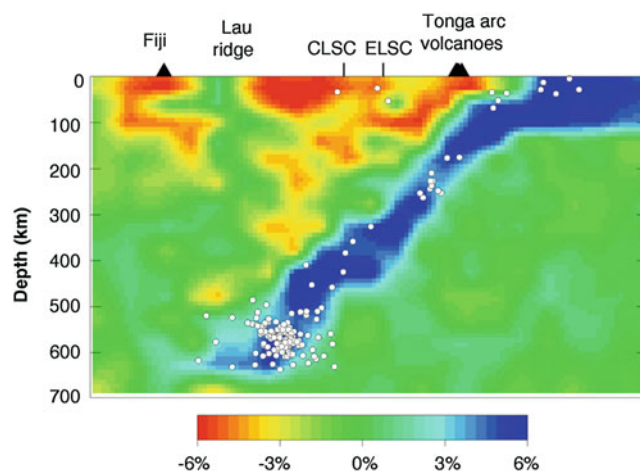


Earth's Structure, Upper Mantle, Fig. 3 Models PRI-P05 (P velocity) and PRI-S05 (S velocity anomaly) at a depth of 350 km (left) and 600 km (right). The color scale ranges from -1.5% (red) to

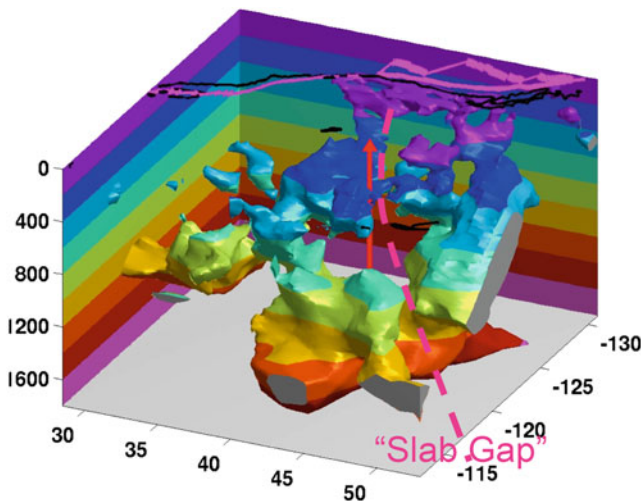
$+1.5\%$ (blue) in P velocity and $\pm 3\%$ in S velocity. Upper mantle plumes show as narrow quasi-circular slow anomalies in red. (From Montelli et al. (2006) with permission from the AGU)



Earth's Structure, Upper Mantle, Fig. 4 The depth variations to the "410-" and "660-km" discontinuities, with respect to 410- and 650-km depth, respectively. (From Houser et al. (2008), with permission from the AGU)



Earth's Structure, Upper Mantle, Fig. 5 A west-to-east cross section through the Tonga subduction zone crossing the trench at 19S, 173 W shows the subducting slab as a continuous feature down to 600-km depth (blue positive anomaly) and coinciding with the deep seismicity (circles). The negative velocity anomaly above the slab represents most likely the effect of dehydration. Active volcanism on Fiji and Tonga is indicated by black triangles. Also indicated are the Central and Eastern Lau Spreading Centers. (Source: Zhao et al. 1997)



Earth's Structure, Upper Mantle, Fig. 6 A view from the northeast at the subducted Farallon slab beneath western North America shows major ruptures that allow mantle rock to flow around it, such as the “slab gap” indicated by the broken line. The slab itself is delineated by a surface with an anomaly of +0.4% in P velocity (high velocity anomalies associated with the lithosphere are not mapped). The location of Yellowstone, another suspected passage of hot mantle rock toward the surface, is indicated by the arrow. Numbers along the axes indicate depth in km, latitude, and longitude (negative for west). Color shading indicates depth. The plate boundaries between the North America, Pacific, and Juan de Fuca plates are indicated as purple lines at the surface. (Figure courtesy Karen Sigloch)

- [Seismic, Waveform Modeling and Tomography](#)
- [Subduction Zones](#)
- [Surface Waves](#)

Bibliography

- Bostock M (2009) Teleseismic body-wave scattering and receiver-side structure. In: Romanowicz B, Dziewonski A (eds) *Seismology and structure of the Earth*. Elsevier, Amsterdam, pp 219–246
- Byrnes J, Hooft E, Toomey D, Villagomez D, Geist D, Solomon S (2015) An upper mantle seismic discontinuity beneath the Galápagos archipelago and its implications for studies of the lithosphere asthenosphere boundary. *Geochem Geophys Geosys* 16:1–19
- Deng K, Zhou Y (2015) Wave diffraction and resolution of mantle transition zone discontinuities in receiver function imaging. *Geophys J Int* 201:2008–2025
- Dziewonski A, Anderson D (1981) Preliminary reference Earth model. *Phys Earth Planet Inter* 25:297–356
- Ferreira A, Woodhouse J, Visser K, Trampert J (2010) On the robustness of global radially anisotropic surface wave tomography. *J Geophys Res* 115:B04313
- Fukao Y, Widiyantoro S, Obayashi M (2001) Stagnant slabs in the upper and lower mantle transition region. *Rev Geophys* 39:291–323
- Gu Y, Dziewonski A, Ekström G (2001) Preferential detection of the Lehmann discontinuity beneath continents. *Geophys Res Lett* 28:4655–4658
- Hosseini K, Matthews K, Sigloch K, Shepard G, Domeier M, Tseknistrenko M (2018) Submachine: web-based tools for exploring seismic tomography and other models of Earth's deep interior. *Geochem Geophys Geosys* 19:1464–1483
- Houser C, Masters G, Flanagan M, Shearer P (2008) Determination and analysis of long-wavelength transition zone structure using precursors. *Geophys J Int* 174:178–194
- Jordan T (1981) Continents as a chemical boundary layer. *Philos Trans R Soc Lond* 301A:359–373
- Montelli R, Nolet G, Dahlen F, Masters G (2006) A catalogue of deep mantle plumes: new results from finite-frequency tomography. *Geochem Geophys Geosys* 7:Q11007
- Schaeffer A, Lebedev S (2013) Global shear speed structure of the upper mantle and transition zone. *Geophys J Int* 194:417–449
- Sigloch K, McQuarrie N, Nolet G (2008) Two-stage subduction history under North America inferred from multiple-frequency tomography. *Nat Geosci* 1:458–462
- Zhao D, Xu Y, Wiens DA, Dorman L, Hildebrand J, Webb, S (1997) Depth extent of the Lau back-arc spreading center and its relation to subduction processes. *Science* 278:254–257

Earthquake Lights

John S. Derr¹, France St-Laurent², Friedemann T. Freund³ and Robert Thériault⁴

¹Tijeras, NM, USA

²LaSalle, QC, Canada

³Code SCR, NASA Ames Research Center/San Jose State University, Moffett Field, CA, USA

⁴Québec Ministry of Energy and Natural Resources, Québec, QC, Canada

Synonyms

Luminous phenomena associated with earthquakes

Definition

Earthquake lights (EQLs) are anomalous luminosities associated with and presumably caused by the accumulation of stress, most characteristically before or during a seismic event or during an aftershock sequence. There is some disagreement over whether the term “earthquake lights” should also be applied to similar luminous phenomena observed in earthquake-prone areas or along faults without immediately associated seismic activity. In these cases the EQLs may be due to locally high stress levels in the Earth's crust that wax and wane without catastrophic rock failure or by earthquakes which are too distant or late relative to the observed EQL.

Classification of Earthquake Lights

The earliest known report of earthquake lights dates to at least 373 BCE when the Greek cities Helice and Buris were destroyed by an earthquake accompanied by “immense

columns of fire” (Seneca). Ancient references will always be questionable, especially when the surviving accounts are written later, but such references indicate that the people at that time were aware that lights might accompany earthquakes. Many sightings from various countries around the world have since been reported. They have been summarized in a number of papers (see, e.g., Thériault et al. 2014; St-Laurent et al. 2006; Derr 1973). As a result of extensive observations associated with the Saguenay earthquake sequence from November 1988 to January 1989, St-Laurent (2000) examined and extended earlier EQL classifications (e.g. Montandon 1948). Six types of luminous phenomena were documented:

1. “Seismic lightning” may appear to be similar to ordinary lightning. The latter appears to be an almost instantaneous flash lasting typically 30 μ s, but seismic lightning is an electric discharge lasting typically 150 ms, bursting out of the ground, rising upward into the atmosphere, and illuminating large areas of the sky without thunder. “Seismic lightning” can be mistaken for “sheet” or “heat” lightning at night which is too distant for thunder to be heard.
2. Luminous bands in the atmosphere at indeterminate height, sometimes horizontally or vertically and sometimes in a bundle, similar in appearance to some polar aurorae although, at much lower altitude, possibly even extending from the ground level upward.
3. Globular incandescences (moving or static), sometimes attached to luminous bands, called “orbs,” “globes,” or “meteors.” They have the appearance of ball lightning, e.g., of luminous spheres floating in midair, sometimes coalescing, lasting for up to a few minutes.

4. Fire tongues, small “flames” flickering or creeping along or near the ground, or like ignis fatuus.
5. Seismic “flames” seen emerging from the ground like an evanescent gas flame but very rarely causing visible burns.
6. Coronal and point-discharge-like lights.

A possible seventh type might be “luminous clouds” like those filmed shortly before the M 7.9 Wenchuan, China, earthquake of 12 May 2008 (YouTube 2008). These are phenomena in the high atmosphere, similar to “fire rainbows,” possibly linked to processes at the upper edge of the atmosphere and/or to ionospheric perturbations.

Examples of EQL

While reliable pictures of EQL are scarce, one example of type 1 comes from the Matsushiro earthquake swarm from 1965 to 1967, Fig. 1 (Yasui 1968, 1971; discussed in Derr 1973). Video of type 1 can be found in Heraud and Lira (2011). An example of type 3 was photographed at Tagish Lake, Yukon Territory, Canada, Fig. 2 (Jasek 1998), around 1 July, probably 1972 or 1973. While it is not possible to say that this picture is definitely EQL, these orbs are typical of the most-frequently observed EQL, as described by Montandon (1948). The only known photo of “flames,” type 5, was taken during the aftershock sequence to the Vrancea, Romania, earthquake of 4 March 1977, discussed in St-Laurent (2000). A sketch of type 6 comes from the account by Joseph A. Dallaire during the Saguenay M 5.9 mainshock of

Earthquake Lights,

Fig. 1 Earthquake lights from Mt. Kimyo, Matsushiro area, Japan, 26 September 1966, 0325 (JST). Luminosity lasted 96 s (Derr 1973; Yasui 1968). (© Seismological Society of America)



Earthquake Lights,

Fig. 2 Earthquake lights from Tagish Lake, Yukon-Alaska border region, around 1 July, probably 1972 or 1973 (exact date unknown). Estimated size: 1 m diameter. Closest orbs slowly drifted up the mountain to join the more distant ones (Jasek 1998). (Photo credit: Jim Conacher, used with permission)



Earthquake Lights, Fig. 3 Drawing by artist from witness' description of a bright, fast-moving bluish-white light, preceded and accompanied by crackling noise emitted by the trees. Laterrière, Québec, 19 km N

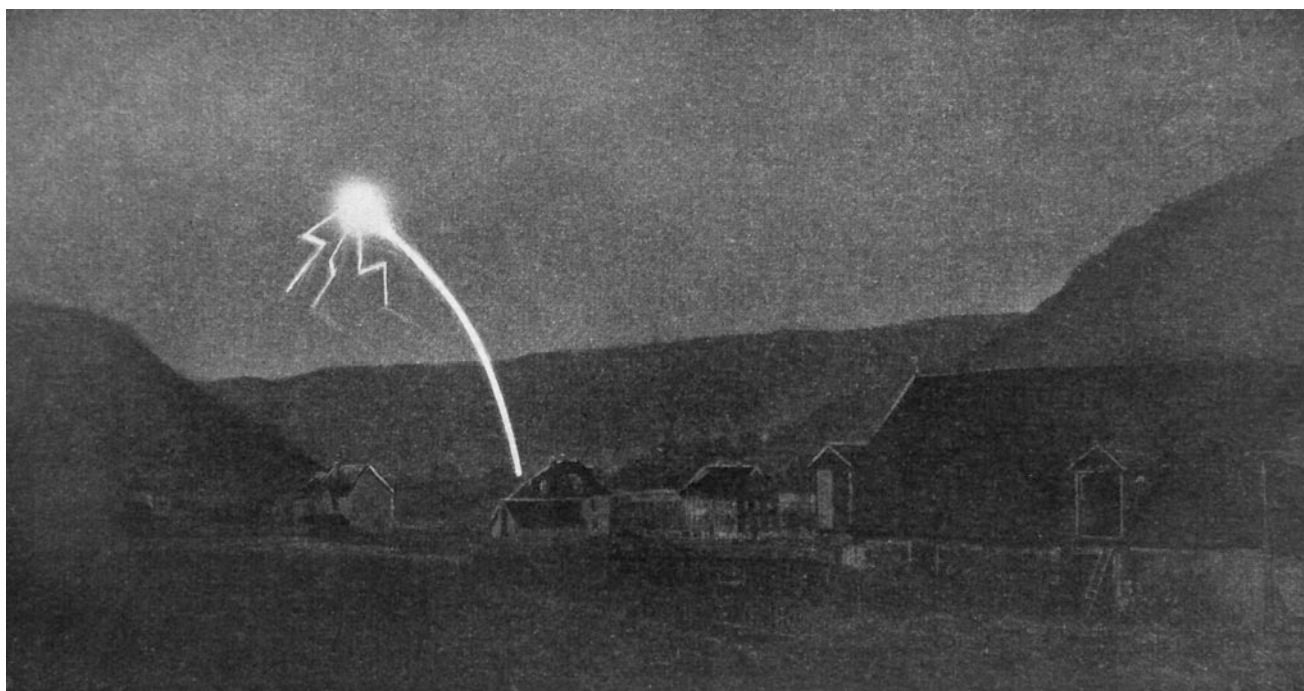
of the epicenter, observed a few seconds before the M 5.9 mainshock of 25 November 1988. Not to scale (St-Laurent 2000). (© Seismological Society of America)

25 November 1988, Fig. 3 (presented in St-Laurent 2000). These examples are notable for the differences they exhibit: illumination of a broad area of the sky, discrete glowing plasmas, evanescent “flames,” and an electric discharge bursting through the surface of the Earth. They are representative of a number of other sightings, although the last one, the rapidly moving electric discharge, is so far unique in the detail of its description.

Example of Type 3 EQL

Schmidt and Mack (1912; Translation by Steven Soter, personal communication) give a number of eye-witness descriptions of type 3 EQL. The most detailed was by Friedrich Konzelmann, foreman in an Ebingen factory:

On the evening of 16 November [1911] between 10 and 10:30, I went home with my wife. My house is about 200 m from the city



Earthquake Lights, Fig. 4 Earthquake lights over Ebingen, Germany, 16 November 1911, ~10 pm. Observed and drawn by Friedrich Konzelmann (Schmidt and Mack 1912)

on the national road to Strasbourg. As we were about 30 m from our house, I heard a distant noise like thunder from Sigmaringen, followed by a faint vibration. As I glanced in the direction from which the noise came, I suddenly noticed at some distance a bright flash from the ground, which then at a considerable height turned into a ball of light, about the size of 20 suns. This ball maintained its brightness about 2–3 seconds and then divided itself like lightning in the direction of Ebingen. In my opinion the earthquake came from southeast to northwest. The earthquake began with the lighting of the fireball. After the rolling went past, I looked backward to the city and saw above the Katzenbuckel a similar but somewhat smaller fireball. The whole surroundings were brightly illuminated. (Fig. 4)

Source of EQL and Associated Electromagnetic (EM) Phenomena

The leading theory of how these various light phenomena are generated, based on extensive laboratory experiments, attributes them to effects by mobile positive hole charge carriers, which become activated when rocks are subjected to high levels of stress (Freund 2002, 2010, 2019). Positive holes exist in common rocks in a dormant state as peroxy links, $\text{O}_3\text{Si}^{\text{O}\cdot}\text{SiO}_3$, many of which are located at grain boundaries or may straddle grain boundaries. When stresses are applied, grains are moving relative to each other causing the peroxy links to break. They thereby activate electrons and positive holes, which immediately begin to recombine, returning to the inactive peroxy state. The difference between the rate of activation and rate of recombination determines how many of the positive hole charge carriers are able to flow out of the stressed rock volume into

the surrounding less stressed or unstressed rocks at speeds up to about 100 m/s (360 km/h or 220 miles per hour). Those positive holes constitute an electric current with attendant magnetic field variations and electromagnetic (EM) emissions. As positive holes arrive at the Earth surface, they create electric microfields powerful enough to field-ionize air molecules and to produce positive airborne ions, corona discharges, and EQL. They also recombine at the surface, returning to the peroxy state and leading to a distinct IR emission, spectroscopically identified in laboratory experiments and known from satellite infrared images as thermal infrared (TIR) anomalies.

How many peroxy a given rock contains represents an important factor. Gabbro and diabase often form vertical sheets of magmatic rocks called dykes, which have deep roots in the Earth's crust, even the upper mantle. They are emplaced during periods of extensional tectonics and are believed to be particularly rich in peroxy. When such rocks are stressed very rapidly, within milliseconds, for instance, by the passage of an earthquake wave, the number density of activated positive holes can instantly reach such high values that the electrical properties of the rocks change dramatically and outburst of electric charges can occur at the speed of light (Freund 2019).

The positive hole theory accounts not only for EQL but also for other pre-earthquake phenomena such as:

1. Air ionization at the ground-to-air interface with either exclusively positive airborne ions or corona discharges producing positive plus negative ions that tend to rise through the atmosphere. They are suspected to be linked

- to widely reported anomalous pre-earthquake fog/haze/cloud formation.
- Changes in the electrical conductivity of the soil and geoelectric potentials, both as a function of lateral distances and as a function of depth.
 - Geomagnetic anomalies in the Earth's crust due to stress-activated telluric currents that flow through the rock column.
 - Ionospheric perturbations in response to air ionization processes in the low atmosphere and/or low to ultralow electromagnetic emission from the Earth.
 - Ultralow and extremely low-frequency (ULF/ELF) and radio-frequency (RF) emissions. Such signals are being used successfully to forecast earthquakes 2–3 weeks in advance in the Peru subduction zone (Heraud and Lira 2011; Heraud et al. 2017).
 - Anomalous thermal infrared (TIR) emissions from around future epicentral areas, occurring particularly from topographic highs.
 - Unusual animal behavior thought to be triggered by air ionization at the ground-to-air interface, in particular by the formation of positive airborne ions or by broadband ultralow frequency (ULF) electromagnetic (EM) emissions from within the Earth's crust, due to stress-activated fluctuating telluric currents, which interfere with physiologically important EM emissions from living organisms such as the alpha brain waves (Freund and Stolz 2013).

Conclusion

At present there is no consensus among seismologists on the phenomenology and physical mechanism of EQL and associated EM phenomena. However, if the positive hole theory is correct, the entire suite of pre-earthquake phenomena, including EQL, may become useful in recognizing the build-up of tectonic stresses in the Earth's crust and, hence, for forecasting earthquakes.

Cross-References

- [Earthquake Precursors and Prediction](#)
- [Electrical Properties of Rocks](#)
- [Electrical Resistivity Surveys and Data Interpretation](#)
- [Goelectromagnetism](#)
- [Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity](#)

Bibliography

- Derr J (1973) Earthquake lights: a review of observations and present theories. *Bull Seismol Soc Am* 63:2177–2187
- Freund F (2002) Charge generation and propagation in rocks. *J Geodyn* 33:545–572
- Freund F (2010) Toward a unified theory of pre-earthquake signals. *Acta Geophys* 58:719–766
- Freund F (2019) Co-seismic earthquake lights: the underlying mechanism. *Pure Appl Geophys*. <https://doi.org/10.1007/s00024-019-02142-2>
- Freund F, Stolz V (2013) Nature of pre-earthquake phenomena and their effects on living organisms. *Animals* 3:513–531
- Heraud J, Lira J (2011) Co-seismic luminescence in Lima, 150 km from the epicenter of the Pisco, Peru earthquake of 15 August 2007. *Nat Hazards Earth Syst Sci* 11:1025–1036. EQL Video at: <https://www.youtube.com/watch?v=f14pQakxXjc>. Last accessed 1 February 2020
- Heraud J, Centa V, Bleier T (2017) Images of the benioff zone in the lithosphere using electromagnetic energy released from stress in tectonic plates. 32nd URSI GASS, Montreal, 19–26 August 2017
- Jasek M (1998) Tagish Lake UFO photo report. <http://www.ufobc.ca/yukon/tagish.htm>. Last accessed 1 February 2020
- Montandon F (1948) Lueurs et malaises d'origine séismique. *Geogr Helv* 3(1):157–178
- Schmidt A, Mack K (1912) Das Süddeutsches erdbeben vom 16 November 1911. Abschnitt VII: Lichterscheinung. *Württembergische Jahrbücher für Statistik und Landeskunde*, Jahrg. part 1, 131–139
- Seneca, *Natural Questions*, VI, 26.2–4
- St-Laurent F (2000) The Saguenay, Québec, earthquake lights of November 1988–January 1989. *Seismol Res Lett* 71:160–174
- St-Laurent F, Derr J, Freund F (2006) Earthquake lights and the stress-activation of positive hole charge carriers in rocks. *Phys Chem Earth* 31:305–312
- Thériault R, St-Laurent F, Freund F, Derr J (2014) Prevalence of earthquake lights associated with rift environments. *Seismol Res Lett* 85(1):159–178
- Yasui Y (1968) A study of the luminous phenomena accompanied with earthquake (part 1). *Mem Kakioka Magn Obs* 13:25–61
- Yasui Y (1971) A study of the luminous phenomena accompanied with earthquake (part 2). *Mem Kakioka Magn Obs* 14:67–68
- YouTube (2008) <http://www.youtube.com/watch?v=KKMTSDzU1Z4>. Last accessed 1 February 2020

Earthquake Precursors and Prediction

Toshiyasu Nagao¹, Masashi Kamogawa² and Seiya Uyeda³

¹Institute of Oceanic Research and Development, Tokai University, Shizuoka, Japan

²Division for Earthquake Prediction Research, Global Center for Asian and Regional Research, University of Shizuoka, Shizuoka, Japan

³Japan Academy, Tokyo, Japan

Definitions and Introduction

Earthquake (EQ hereafter) prediction or forecast is to specify the source location, magnitude M , and occurrence time of EQ with certain accuracy before its occurrence. The terms prediction and forecast are often used with different nuances, the latter allowing for more stochastic nature. We will mainly use the former term in this entry. EQ prediction is often roughly classified by the length of concerned time into long term

($10^{1.5}$ – 10^1 years), intermediate term (10^1 – $10^{0.5}$ years), and short term ($<10^{0.5}$ year). EQ precursors are anomalous phenomena, which take place before EQs so that prediction is made on their information. This entry will focus on the precursors for short-term prediction. At this stage, however, the existence itself of the precursors has to be firmly established as a scientific fact, and further details of them must be thoroughly investigated.

Various EQ precursors have been noticed throughout human history until present day (Tributsch 1982). They are mostly short-term precursors on animals, plants, clouds, and sounds (see ► [“Earthquake Sounds”](#)). Precursors perceived without aid of scientific instruments are called macroscopic anomalies.

Precursors instrumentally recorded are classified into long-/intermediate- and short-term ones. Long-/intermediate-term precursors are active faults serve as geological precursors. Trenching of active seismic faults provides the dates and displacements of past large EQs, from which those of the next large EQ may be inferred. Other long-/intermediate-term precursors include the following:

1. Geodetic: land deformation, changes in sea-level, tilt, and strain
2. Seismological: seismic gaps (see ► [“Characteristic Earthquakes and Seismic Gaps”](#)), quiescence (see ► [“Seismic Quiescence and Activation”](#)), S wave splitting (see ► [“Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting”](#)), changes in micro-seismicity, seismic wave velocity, *b*-values, and characteristic EQs (see ► [“Characteristic Earthquakes and Seismic Gaps”](#))
3. Changes in electrical resistivity
4. Geochemical/hydrological: changes in radon, CO₂, and other gas discharge (see ► [“Isostasy, Thermal”](#)) and level and temperature of underground water (see ► [“Seismic Signals in Well Water Observations”](#))

Some, in particular, those in (3) and (4) serve for short-term EQ prediction also.

Short-term precursors which are of major concern of this article are mainly geochemical/hydrological and electric, magnetic, and electromagnetic (EM). The latter are changes in (1) geoelectric field (telluric current), geomagnetic field, and EM emissions at various frequencies (DC to VLF), (2) transmission properties of man-made EM waves of VLF to VHF range, (3) ionosphere (foF2 and total electron contents (TEC)), and (4) infrared (IR) or ground temperature. These changes are measured by apparatus on the ground or those mounted on artificial satellite.

Note that precursors and triggering effects are sometimes very difficult to discriminate. If the crustal state is very critical, small disturbances such as a passage of low atmospheric

pressures such as typhoons/cyclones/hurricanes may trigger earthquakes (e.g., Liu et al. 2009). The notable triggering effect is a large dam construction (see ► [“Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India”](#)). The case of Koyna dam in India is the most well-studied case (Gupta et al. 2011). Furthermore, water injection to deep well triggers earthquakes. Nowadays, shale gas retrieving is creating many induced earthquakes (e.g., Lei et al. 2017).

History

Zechariah of Israel (ca. 520 BC) is said to have issued the first long-term EQ forecast, based on the surface rupture of an EQ in 759 BC. Precursors were recorded as early as in 373 BC in Greece. Many followed from many parts of the world in the form of documents and folklore. In later years, even a larger number of macro-anomalies are kept accumulated (Tributsch 1982), now including such items as radio and TV noise, e.g., at 1995 M7.3 Kobe, 1999 M7.4 Izmit, and 1999 M7.7 Chichi EQs (Ikeya 2004). Most of macro-anomalies, however, have not been well studied systematically, although they apparently played significant roles, e.g., in the success of 1975 M7.3 Haicheng EQ prediction (Raleigh et al. 1977).

After 1906 M8.3 San Francisco EQ, through investigation of geodetic data on faults, the elastic rebound theory (see ► [“Earthquakes and Crustal Deformation”](#)) was developed, laying the theoretical foundation for EQ prediction. From the 1960s to the 1970s, countries including USSR, Japan, China, and the USA embarked on national scale EQ prediction projects. Projects consisted mainly of statistical approach for long-/intermediate-term prediction (see ► [“Statistical Seismology”](#)), but there was also deterministic approach to short-term prediction. In the early 1970s, the dilatancy-diffusion model (Sholtz et al. 1973) appeared. This model took advantage of results of field and laboratory studies obtained in various countries, like 10–20% Vp/Vs change in Garm, USSR (Semyenov 1969), and the USA (Aggarwal et al. 1973), crustal uplift in Japan (Danbara 1981), radon emission and electrical resistivity (Sadovsky et al. 1972), and so forth. All the quoted precursors, long and short, appeared to support the physical picture of the dilatancy-diffusion model. The famous prediction of 1975 M7.3 Haisheng EQ also was based on all kinds of precursors (Raleigh et al. 1977). Intermediate-term measures were taken in June 1974 and short-term in December 1974. Emergency alert to the public was issued in the early afternoon of February 4, 1975, several hours before the main shock to evacuate the public in time. These and other encouraging works in the 1960–1970s made the whole prediction community highly optimistic (e.g., Press 1975).

However, the period of optimism did not last long. First, later studies using controlled sources (McEvilly and Johnson 1974) did not confirm the significant change of V_p/V_s , which was the most critical background of the dilatancy-diffusion model. Second, for 1975 M7.8 Tangshan EQ, no timely alert was issued because there was no foreshock activity (see ► “Earthquake, Foreshocks”), despite many other precursors, causing the loss of many 100,000 lives (Chen et al. 1988).

In fact, the traditional seismological/geodetic methods were unable to short-term predict any great EQs (see ► “Great Earthquakes”) other than the Haisheng EQ. For none of these EQs, any pre-slip, the most legitimate short-term precursor from seismological point of view (e.g., Shibazaki and Matsu’ura (1995) and Kato and Hirasawa (1999)) has ever been detected. Yet, the main strategy of Japan Meteorological Agency (JMA) for short-term prediction of the Tokai EQ is to catch pre-slips as reported for 1944 M7.9 Tonankai EQ (Fig. 1; Mogi 1985). The Tokai EQ has been at a forecasted status since the late 1970s (Ishibashi 1981). Another well-publicized unsuccessful case is the Parkfield, where M6 class event was predicted in 1985 to come within 5 years, and 72 hours alert was issued in 1992 (Roeloffs 1994). It did not come until 2004, 12 years later (Bakun et al. 2005).

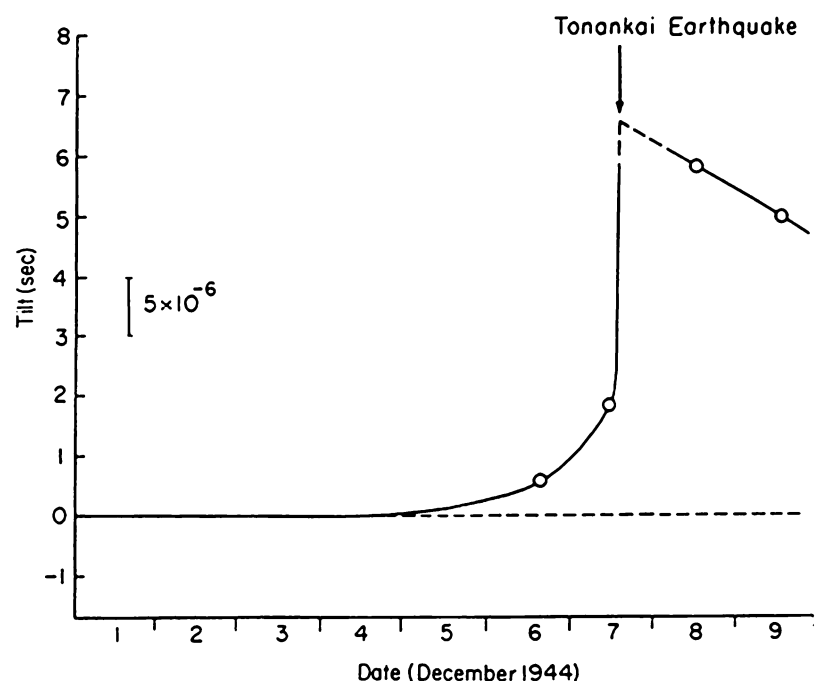
As a matter of fact, seismological precursors and methods, such as seismic gap (see ► “Characteristic Earthquakes and Seismic Gaps”), M8 algorithm (see ► “Earthquake Prediction, M8 Algorithm”), ZMAP (Wiemer and Wyss 1994), hotspots (Holliday et al. 2006), etc., were not aiming at short-term prediction. The new development and discoveries

in seismology such as the asperity models (Yamanaka and Kikuchi 2004), slow/silent EQs (see ► “Slow Earthquake”), and low-frequency non-volcanic tremors (Obara 2002) greatly contributed to the understanding of EQ mechanisms, but their direct relevance to short-term prediction has not been made clear. Characteristic EQ (see ► “Characteristic Earthquakes and Seismic Gaps”) may be the only case that can be quasi-short term predictable by seismic monitoring, but they are rare cases. By the late 1970s, general view of the main stream community became entirely pessimistic (e.g., Evernden 1982). This pessimism has persisted until now, making the word prediction almost “forbidden.”

It may be noted, however, that numerous undeniable short-term precursors were observed in non-seismic/geodetic (nonmechanical hereafter) fields. On top of many macro-anomalies, they were radon, CO_2 and other gas discharges, water level, and electric, magnetic, and EM anomalous changes.

Short-term EQ prediction may be realized only after the main stream strategy makes a change of course toward a nonmechanical approach, except for a few cases, e.g., GPS detected pre-seismic deformation from Kamchatka (Gordeev et al. 2001) and pre-seismic velocity changes at the Parkfield SAFOD drill site (Niu et al. 2008). However, thanks to the increase of the observation network and the progress of critical phenomena theory, even mechanical precursors recently began to be reported (e.g., Lippiello et al. 2012 (foreshock), Sarlis et al. 2015 (natural time analysis), Nanjo et al. 2012, 2016 (seismicity changes including b -value change), Tanaka 2012 (Tidal triggering)).

Earthquake Precursors and Prediction, Fig. 1 Precursory crustal deformation observed just before the M7.9 Tonankai EQ in 1944 (Mogi 1985)



Nonmechanical Short-Term Precursors

Difficulties Inherent to Short-Term Precursors

Short-term EQ prediction by standard mechanical approach seems to have scarce the chance except for clear foreshock activities because the instant the stress reaches critical state can hardly be pinpointed by the present seismic or geodetic observations. Nonmechanical precursors make use of some phenomena which necessarily occur at that instant as by-products of the same cause, the stress buildup. The only necessary condition to make them useful is that their critical stress is slightly smaller than that of EQ. When the state attains critical condition, EQ would occur soon after the detection of them, perhaps aided by the last push of some triggers.

There are two rates to evaluate precursors. One is success rate which is the number of successful predictions divided by the number of issued predictions, and the other is alarm rate which is the number of successfully predicted EQs divided by the number of target EQs. In applying these evaluating rates, criteria of success and failure, i.e., the allowable errors in the

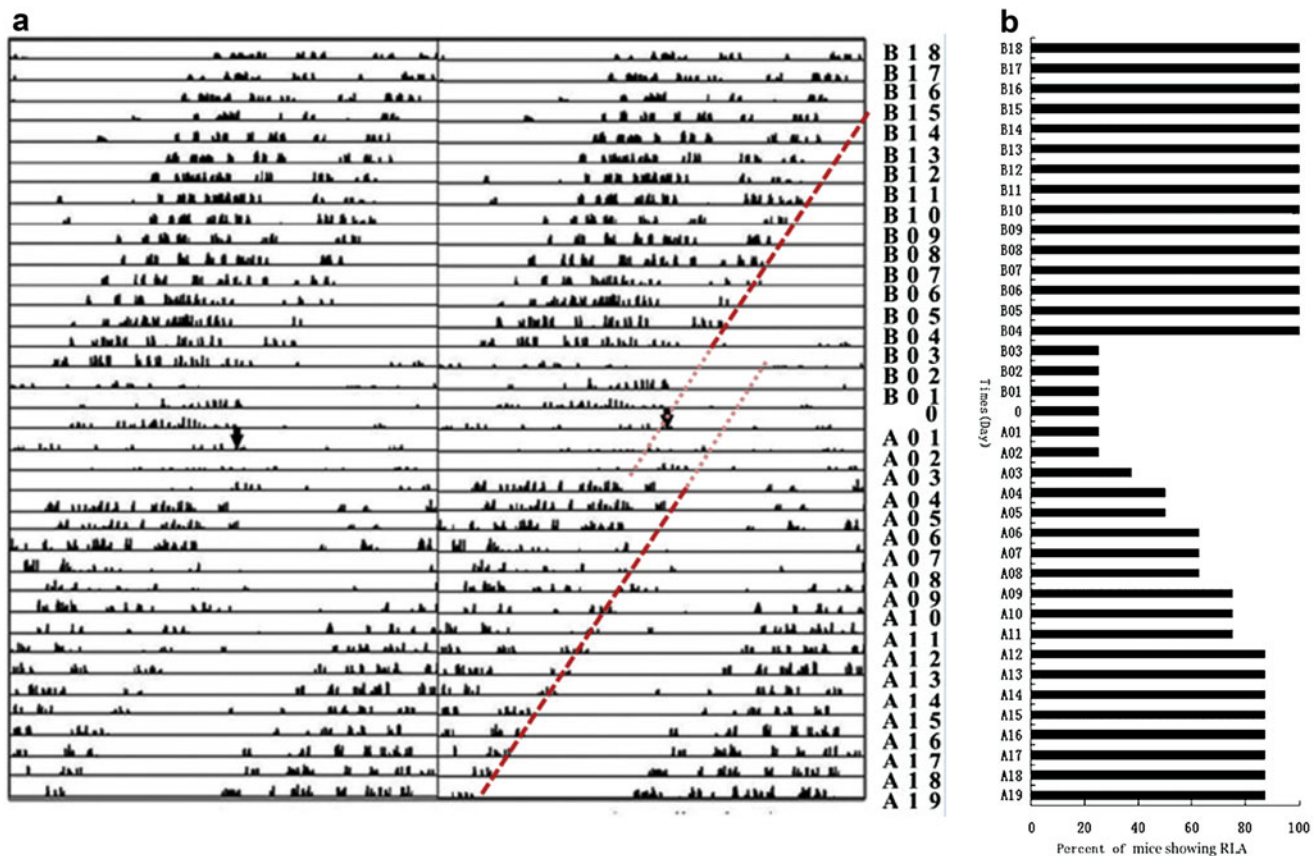
three prediction targets, the dates, location, and magnitude, must be clearly stated. When they are presented properly, however, criticizing them as causal relation is not clear, or there are no other supporting geophysical simultaneous phenomena that seem unreasonable.

Macro-anomalies

Macro-anomalies are numerous, and the well-known examples, after the Haisheng EQ, are 1995 M7.3 Kobe, 1999 M7.4 Izmit, 1999 M7.7 Chichi EQs (Ikeya 2004), and 2008 M8.0 Shuchuan EQ (Fig. 2; Li et al. 2009). The causes of modern macro-anomaly precursors, such as radio and TV noise and malfunctioning of electrical appliances, are most probably related to pre-seismic EM emissions. Recently, Martilelli (2018) documented that earthquake precursors existed even in BC.

Geochemical/Hydrological Precursors

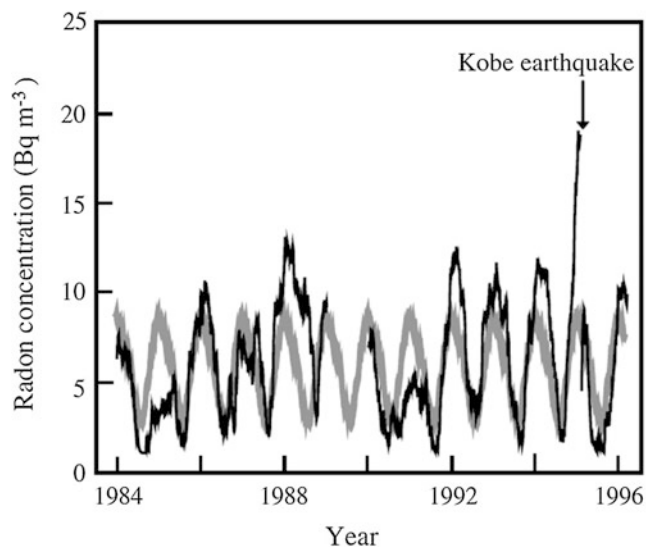
Geochemical/hydrological precursors are also diverse. Examples are discharges of radon, carbon dioxide, helium, H₂, and CH₄ (see Toutain and Baurbron 1999). Among them, radon



Earthquake Precursors and Prediction, Fig. 2 Abnormal animal behavior observed just before the M8.0 Wenchuan EQ in 2008 (Li et al. 2009). (a) Under free-running conditions, mouse locomotor activity presents circadian rhythm from B18 (18 days before EQ) to B04 (4 days before) and no rhythm from B03 (3 days before EQ) to A02

(2 days after EQ) including the day of EQ. Horizontal axis is 48 h (2 days). Diagonal solid lines mean the circadian rhythm does not coincide with 24 h. (b) The percentage of mice showing RLA (circadian rhythm of mouse locomotor activity) among the eight mice, indicating that some mice lost RLA around the EQ

emission has been most widely monitored, e.g., 1976 M7.1 Gazli EQ, Uzbekistan, 1985 M8.1 Mexico EQ, and 1995 M7.2 Kobe EQ shown in Fig. 3 (Yasuoka et al. 2009). Level changes of underground water are also potential precursors (e.g., Wang 1985 for Tangshan EQ, Igarashi et al. 1992 for 1992 M5.9 Tokyo Bay EQ.). Geochemical/hydrological precursors are often observed at localities long distance (100 km or more) away from epicenters. Perhaps their appearances are heavily controlled by local geological situations such as faults and poroelastic conditions (see ► “Seismicity, Intraplate”). For water level and temperature, coseismic changes are much more eminent and commonly observed than pre-seismic ones, in contrast to EM precursors (see later). In March 2011, the 2011 M9.0 off the Pacific coast of the Tohoku earthquake took place. This EQ brought us huge tsunami damage and severe nuclear power plant accident. This earthquake accompanied many precursors (Nagao



Earthquake Precursors and Prediction, Fig. 3 Time series of the daily radon concentration observed from January 1984 to February 1996. The daily minimum variation is shown by a black line and the normal variation by a gray line (Yasuoka et al. 2009). Clear enhancement of radon concentration can be seen before the M7.3 Kobe EQ

et al. 2014). Orihara et al. (2014) reported precursory water level change, which reached about 10 m for 2000 m deep well, about 3 months before the main shock. Furthermore, Tsunomori and Tanaka (2014) reported very large groundwater radon concentrations in Izu Peninsula. This change was the first for over 35 years of observation (Fig. 4). This well had been operated for Japanese EQ prediction research because of the well-recorded textbook-type radon change at the time of 1978 Izu-Oshima-Kinkai EQ (Wakita et al. 1980).

Satellite IR Images as Possible Surface Temperature

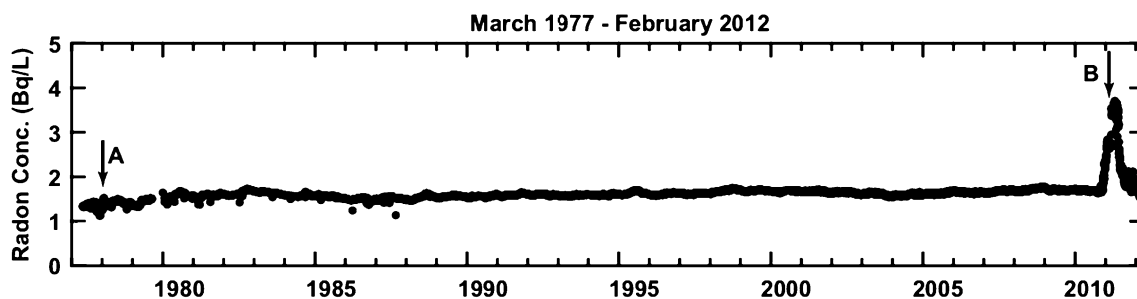
Anomalous satellite IR images as possible ground surface temperature have been reported over seismic regions (see ► “Isostasy, Thermal”). The meaning of satellite-observed IR anomalies, however, is not clear. There have been arguments on what they are and how they are formed. An alternative interpretation is that the satellite-observed IR image is not a temperature anomaly. When the p-holes (see later) arrive at the earth’s surface, they recombine to cause a spectroscopically distinct nonthermal IR emission (Freund 2009). Ouzounov et al. (2011) examined outgoing longwave radiation (OLR, infrared 10–13 μm) measured around troposphere before the 2011 Tohoku EQ. They claim that during 21–24 February and 7–10 March, transient OLR anomalous fields were observed near the epicentral area and over the major faults, with a confident level greater than 2 sigma.

Electric, Magnetic, and EM Precursors

Systematic research on EQ-related electric, magnetic, and EM phenomena was initiated more or less simultaneously in varied parts of the world in the 1980s. This relatively new branch of science is now called Seismo-Electromagnetics.

Telluric Current Precursors: The VAN Method

The VAN method, developed in Greece since the early 1980s, is named after initials of the founding scientists, P. Varotsos, K. Alexopoulos, and K. Nomikos (Varotsos and Alexopoulos 1984a, b). This has been the only working system for real short-term prediction for more than three decades. For $M \geq 5$ Greek EQs, the criteria for successful prediction imposed by



Earthquake Precursors and Prediction, Fig. 4 Radon concentration data from March 1977 to February 2012. A: onset time of 1978 Izu-Oshima-Kinkai EQ, B: onset time of 2011 Tohoku earthquake. (After Tsunomori and Tanaka 2014)

themselves are $<$ a few weeks in time, <0.7 units in M , and <100 km in epicentral location. SES (seismic electric signals) are transient DC geopotential variations observed before EQs by dipoles of buried electrodes at separate sites and continuously monitored at many stations (e.g., Varotsos 2005). By adoption of multiple dipole system, they accomplished successful noise rejection. Single SES precedes single EQ, whereas the so-called SES activity, which consists of a number of SES in a short time, is followed by a series of EQs.

VAN group made two major discoveries. One is the “Selectivity,” which means that only some sites are sensitive to SES, and a sensitive site is sensitive only to SES from some specific focal area(s). This information helps for estimating the epicenter area. The selectivity is considered to be originated from the inhomogeneity of the subterranean electrical structures, i.e., SES goes only through underground conductive channels to sensitive sites. The other is the “VAN-relation” which means that the focal distance r , EQ magnitude M , and the potential difference ΔV for dipole of length L are related by $\log(\Delta V/L \times r) = aM + b$ where a is a constant 0.34–0.37 and b is a site-dependent constant. Once the epicentral location is estimated from the selectivity

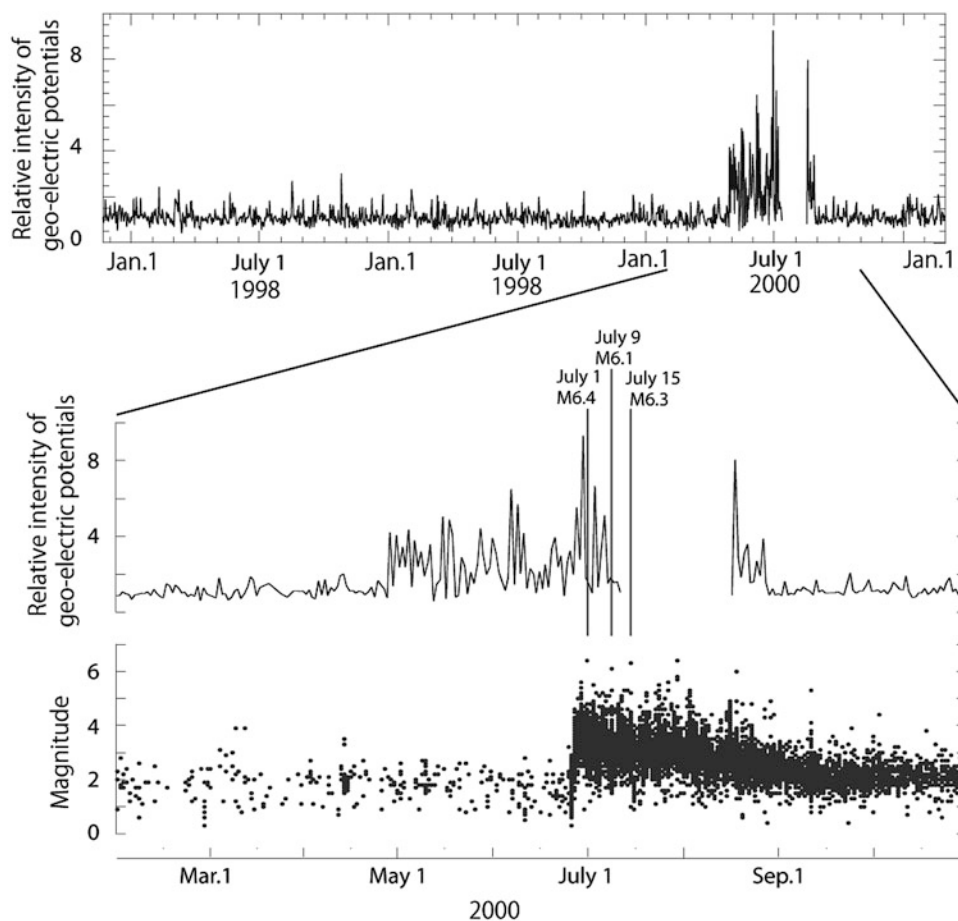
data, M can be assessed since both $\Delta V/L$ and r are known. The VAN method has been a contentious subject (e.g., Geller 1996; Lighthill 1996). However, it has well survived all the criticism.

In Japan, VAN-type monitoring has been tried since the late 1980s (Uyeda et al. 2000). Despite the high-level noise of DC trains, the existence of the VAN-type SES has been confirmed. In the year 2000, a 2-month long seismic swarm, with 7,000 $M \geq 3$ shocks and 5 $M \geq 6$ shocks, occurred in Izu Island region. For this swarm activity, significant pre-seismic electric disturbances were observed (Fig. 5, Uyeda et al. 2002).

Ultralow Frequency (ULF) Precursors

Well-known example is the case of the 1989 M7.1 Loma Prieta EQ (Fraser-Smith et al. 1990). The horizontal component measured at a site 7 km from the epicenter started anomalous enhancement about 2 weeks prior to and a sharp increase a few hours before the EQ. The disturbance lasted for about 3 months after the EQ. These have never been observed at any other time during the observation of more than 15 years. Reports of observing pre-seismic ULF geomagnetic anomalies have been made also for 1988 M6.9 Spitak

Earthquake Precursors and Prediction, Fig. 5 Anomalous geo-electric potential changes in ULF (0.01 Hz) band prior to 2000 seismic swarm activity in Izu island region, Japan. The bottom panel shows seismicity. (Modified from Uyeda et al. 2002)



(Armenia) EQ (Kopytenko et al. 1993) and 1993 M8.0 Guam (Marianas) EQ (Hayakawa et al. 1996). There are some discussions pointing out that these ULF signals might have been artifact or magnet-telluric origin (e.g., Thomas et al. 2009). Concerning ULF anomalies, some statistical works were made during last decade (e.g., Hattori et al. 2013; Han et al. 2014). Han et al. (2017) showed a correlation between the ultralow frequency (ULF) seismo-magnetic phenomena and local seismicity in the Kakioka region, Japan. They used Molchan's error diagram (see Molchan and Kagan 1992) for statistical evaluation. For instance, the results indicate that predictions based on ULF magnetic data in Kakioka, Japan, perform better than random prediction.

Higher-Frequency EM Emission

Gokhberg et al. (1982) was a pioneering observation of pre-seismic EM wave emission in the LF range. Asada and his group in Japan monitored the wave forms of two horizontal magnetic components of VLF EM waves, through which apparent incoming directions of VLF pulses were assessed (Asada et al. 2001). They found that, a few days before M5 class land EQs within 100 km of their stations, pulses with fixed incoming direction appeared and the EQ actually occurred in that direction, whereas the sources of overwhelmingly more numerous and stronger noises were moving along with lightning sources

Warwick et al. (1982) reported on the observation of high-frequency radio waves possibly related to the precursory activity of 1960 Great M9.5 Chilean EQ. This was probably the first report of this kind. Enomoto et al. (2006) recorded

two events of anomalous geo-electric current pulses (HF-band) at Erimo station, Hokkaido, Japan. One in 2000–2001 occurred before and during the volcanic activity of Mt. Usu (200 km away), and the second started 1 month before 2003 M8.0 Tokachi-Oki EQ (80 km away). These were the only anomalies during the 10-year observation period.

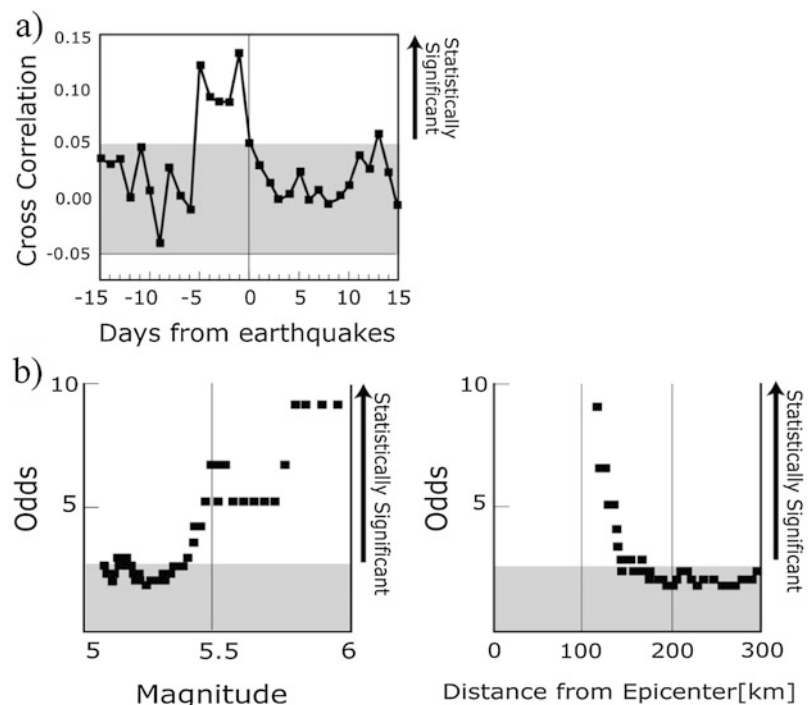
Lithosphere-Atmosphere-Ionosphere Coupling

Sub-ionospheric anomalies before large EQs were reported (e.g., Molchanov and Hayakawa 1998) through monitoring 10–20 kHz VLF ship-navigation waves, although Clilverd et al. (1999) did not obtain similar results. A decrease in electron density of ionosphere within 5 days before $M \geq 5$ class EQ was found in Taiwan by the ionosonde measurements of critical plasma frequency, foF2 (see Fig. 6; Liu et al. 2006). Fujiwara et al. (2004) found pre-seismic atmospheric anomalies causing FM radio transmission anomalies within 5 days before $M \geq 4.8$ EQs, which was pioneered by Kushida and Kushida (2002). Moriya et al. (2010) further promoted this study and obtained statistical relations between the total time length of anomalous reception and EQ magnitude. The concept of pre-seismic lithosphere-atmosphere-ionosphere coupling (LAI coupling) arisen from these investigations became an important issue in precursor studies (Pulinets and Boyarchuk 2005; Kamagowa 2006), because some causative factor should exist in the lithosphere, and there should be some mechanism to transport its energy to the high atmosphere.

Considering earthquake prediction research from space, the French satellite program was truly an epoch. The French

Earthquake Precursors and Prediction, Fig. 6

Statistical significance of pre-seismic ionospheric anomaly (PEIA) (after, Liu et al. 2006). (a) Cross-correlation coefficient between EQs and anomalies with time lag. The shaded area denotes the rejection bounds at significance level 0.01. (b) Odds of the shallow EQs with the PEIA against the magnitude and distance. Odds versus related mean magnitude (left), and odds versus related mean distance (right). The shaded areas in the two plots denote the rejection value of odds as 2.5 under significance level of 0.01

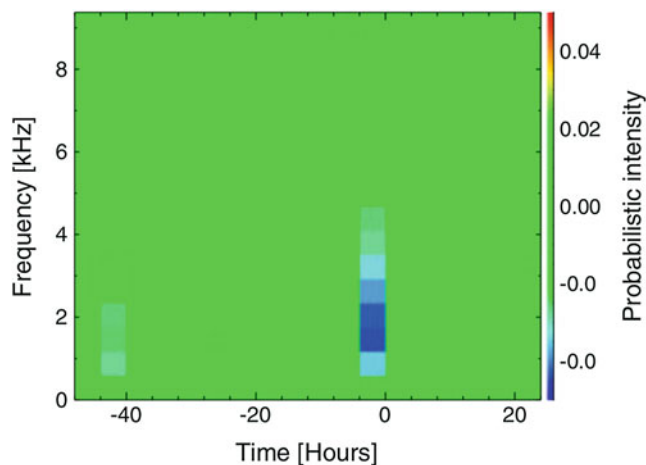


government launched the DEMETER (Detection of Electro-Magnetic Emissions Transmitted from Earthquake Regions) satellite in 2004 (see Parrot 2018). A lot of satellites measured many plasma parameters such as electron density, temperature, ULF-HF band electromagnetic waves, etc. As a result of the analysis, it was proposed that the most statistically significant precursor was a nighttime VLF electric field intensity decrease phenomenon that seems to be due to the D region fluctuation just before earthquakes (Fig. 7, Pisa et al. 2013).

As mentioned above, there are many reports that the ionosphere changed before earthquakes. Outstanding findings were reported in the 2011 Tohoku M9 earthquake. Total electron content (TEC) in the ionosphere relatively increased from 40 min before the main shock (Heki 2011). Although the pre-seismic TEC clearly increased, this phenomenon seemed an artifact change for Kamogawa and Kakinami (2013). In the very beginning, Heki claimed that “TEC increases,” but now the expression that “TEC curve” is “bent” just before the M8 class earthquake is used, and now Kamogawa agrees with this expression (Kamogawa, private communication). Now Heki and his colleagues confirmed that the same phenomenon occurred immediately before all eight earthquakes with a moment magnitude of 8.2 or larger. Note that there are many similar ionospheric changes without the earthquakes. Their further study supports this phenomenon using ionospheric three-dimensional tomography (e.g., He and Heki 2016) just before the huge earthquake. In addition, theoretical research for this phenomenon also comes to be very active (e.g., Kelley et al. 2017) (Fig. 8).

Possible Mechanism of Pre-seismic EM Emissions

It is often questioned why EM signals, such as SES, appear only pre-seismically and not coseismically. This point makes



Earthquake Precursors and Prediction, Fig. 7 Frequency-time dependence of the probabilistic intensity obtained from the nighttime electric field data measured within 440 km of the epicenters of earthquakes with magnitudes greater than >5.0 (Pisa et al. 2013)

scientific community dubious about EQ-related EM precursors in general because EQ is by far the most energetic event. Actually, coseismic signals are routinely observed, but they are not useful as precursors. Moreover, all the so-called coseismic EM signals observed so far were found to occur at the time of the arrival of seismic waves. They are, therefore, “coseismic wave” and not “true” coseismic. The fact that no true coseismic signals are observed may be an important clue in exploring the physical mechanism of generation of both EM signals and EQs themselves. In fact, the slow pre-seismic stress buildup and the instantaneous stress release at EQ are physically very different processes, and there is no compelling reason why they generate similar EM signals.

The generation mechanism is different for different frequencies. It may be the electrokinetic effects (e.g., Mizutani et al. 1976; Fitterman 1978) and/or pressure-stimulated current (PSC) effects for DC to low-frequency signals and piezoelectric effects and exo-electron emission (Enomoto and Hashimoto 1990) for higher-frequency signals. A mechanism involving the positive holes (p-holes) in rock forming minerals under stress has also been proposed.

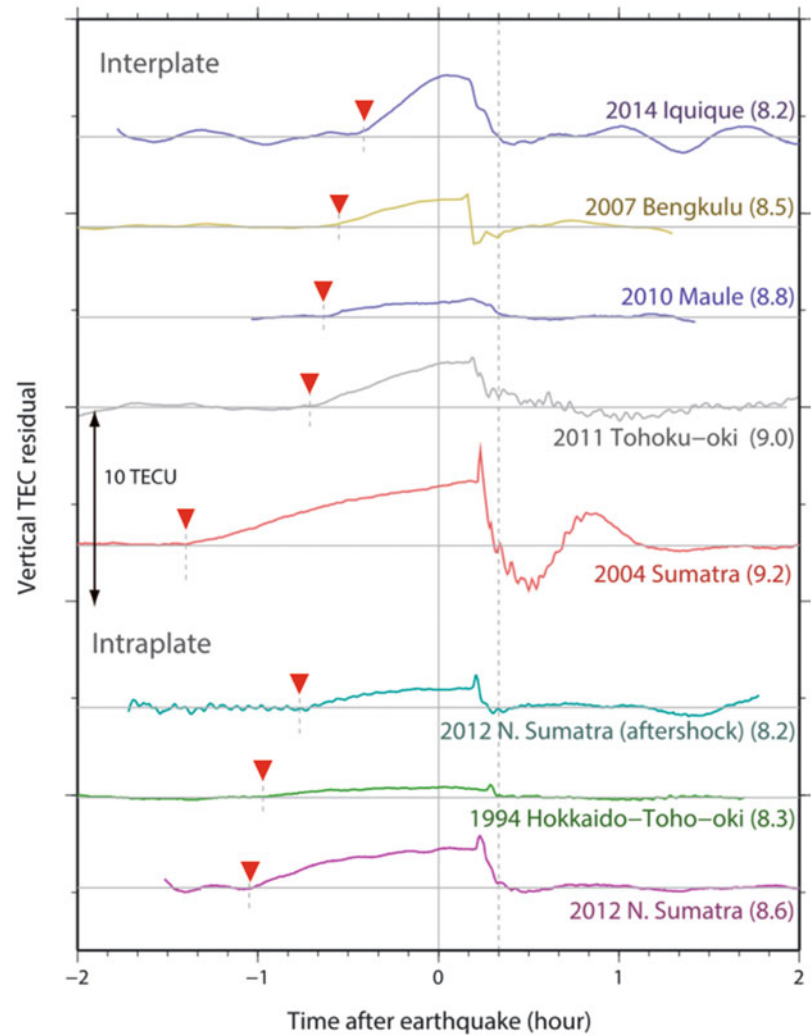
The PSC was proposed by Varotsos and Alexopoulos (1986). The impurities and vacancies form randomly oriented local electric dipoles, which align their orientation in avalanche following an activation process to generate the electric current. PSC model is unique in generating currents during gradual stress buildup without requiring any sudden change of stress. Freund and his colleagues are proposing a different mechanism for ULF electric signals (Freund et al. 2006). They discovered in laboratory that when a block of igneous rock is put under stress locally, the rock turns into a battery without external electric field. The charge carriers in this case are semiconductor-type positive holes (p-holes) which is oxygen in one-valence state. They attribute the satellite image of pre-seismic IR anomalies to the result of recombination of two p-hole charge carriers at earth’s surface as mentioned earlier (e.g., Freund 2009; Takeuchi and Nagao 2013).

Although models related to micro-cracking based on laboratory experiments such as discharge of screening charge of piezoelectric polarization (Ikeya 2004; Yoshida et al. 1997), electrification of fresh crack surfaces (Yamada et al. 1989), and exo-electron have been proposed, there have been no reliable field reports on the occurrence of pre-main shock micro-cracking. Furthermore, it might be pointed out that in all these models much stronger coseismic signals would be expected.

Critical Phenomena

EQ may be a critical phenomenon defined in statistical physics (e.g., Bak and Tang 1989) which means that the short time EQ prediction is synonym for identifying the approach to criticality. Although this is difficult to achieve (see Main 1995), it has been recently shown that by analyzing time

Earthquake Precursors and Prediction, Fig. 8 The residual plot of VTEC for the eight earthquakes is compared. Mw is indicated within the parentheses. Red rectangles indicate the times of positive breaks. (After Heki and Enomoto 2015)



series of seismicity in a newly introduced time domain “natural time,” the approach to the state of criticality can be clearly identified (Sarlis et al. 2008). This way, they seem to have succeeded in shortening the lead time of VAN prediction to only a few days (Uyeda and Kamogawa 2008). We thought that seismic data can play an amazing role of short-term precursor when combined with SES data. However, with the recent progress of the critical theory, seismic data alone can play important role for the short-term prediction (e.g., Sarlis et al. 2013; Kamogawa et al. 2019).

Summary and Future Outlook

Most of the precursors mentioned above are possibilities, but so far they were recognized or documented only after main shocks, and they gave scarce constraints on the source location and magnitude M (except VAN’s SES). Much more enhanced research is needed to make them practically useful precursors.

1. Some macro-anomaly short-term precursors are real. Scientific research on them is inadequate so far.
2. Seismic and geodetic precursors can be useful for long/intermediate EQ prediction.
3. Foreschock activity may be useful for short-term prediction.
4. Nonmechanical precursors, such as some geochemical, hydrological, and EM precursors, are plausible for short-term EQ prediction. For sooner realization of practical short-term prediction, more research on them must be promoted.

Cross-References

- ▶ [Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India](#)
- ▶ [Characteristic Earthquakes and Seismic Gaps](#)
- ▶ [Earthquake Prediction, M8 Algorithm](#)
- ▶ [Earthquake Sounds](#)

- [Earthquake, Foreshocks](#)
- [Earthquakes and Crustal Deformation](#)
- [Isostasy, Thermal](#)
- [Great Earthquakes](#)
- [Seismic Quiescence and Activation](#)
- [Seismic Signals in Well Water Observations](#)
- [Seismicity, Intraplate](#)
- [Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting](#)
- [Slow Earthquake](#)
- [Statistical Seismology](#)

Bibliography

- Aggarwal Y, Sykes L, Armbruster J, Sbar M (1973) Premonitory changes in seismic velocities and prediction of earthquakes. *Nature* 241(5385):101–104
- Asada T, Baba H, Kawazoe K, Sugiura M (2001) An attempt to delineate very low frequency electromagnetic signals associated with earthquakes. *Earth Planets Space* 53:55–62
- Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. *J Geophys Res* 94:15635–15637
- Bakun WH, Aagaard B, Dost B, Ellsworth WL, Hardebeck JL, Harris RA, Ji C, Johnston MJS, Langbein J, Lienkaemper JJ, Michael AJ, Murray JR, Nadeau RM, Reasenberg PA, Reichle MS, Roeloffs EA, Shakal A, Simpson RW, Waldhauser F (2005) Implications for prediction and hazard assessment from the 2004 Parkfield earthquake. *Nature* 437:969–974
- Chen Y, Tsoi K, Chen F, Gao Z, Zou Q, Chen Z (1988) The great Tangshan earthquake of 1976., 153pp. Pergamon, Tarrytown
- Clilverd MA, Rodger CJ, Thomson NR (1999) Investigating seismo-ionospheric effects on a long subionospheric path. *J Geophys Res* 104(A12):28171–28179
- Danbara T (1981) Geodesy and earthquake prediction. In: Rikitake T (ed) *Current research in earthquake prediction I*. Cent. Acad. Publishing Japan/D. Reidel Publishing Company, pp 167–220
- Enomoto Y, Hashimoto H (1990) Emission of charged particles from indentation fracture of rocks. *Nature* 346:641–643
- Enomoto Y, Hashimoto H, Shirai N, Murakami Y, Mogi T, Takada M, Kasahara M (2006) Anomalous geoelectric signals possibly related to the 2000 Mt. Usu eruption and 2003 Tokachi-oki earthquake. *Phys Chem Earth* 31:319–324
- Evernden JF (1982) Earthquake prediction: what we have learned and what we should do now. *Bull Seismol Soc Am* 72:343–349
- Fitterman DV (1978) Electrokinetic and magnetic anomalies associated with dilatant regions in a layered earth. *J Geophys Res* 83:5923–5928
- Fraser-Smith AC, Bernardi A, McGill PR, Ladd ME, Helliwell RA, Villard OG Jr (1990) Low-frequency magnetic field measurements near the epicenter of the Ms 7.1 Loma Prieta earthquake. *Geophys Res Lett* 17:1465–1468
- Freund F (2009) Stress-activated positive hole charge carriers in rocks and the generation of preearthquake signals. In: Hayakawa M (ed) *Electromagnetic phenomena associated with earthquakes*. pp 41–96. ISBN: 978-81-7895-297-0
- Freund F, Takeuchi A, Lau BES (2006) Electric currents streaming out of stressed igneous rocks – a step towards understanding pre-earthquake low frequency EM emissions. *Phys Chem Earth* 31:389–396
- Fujiwara H, Kamogawa M, Ikeda M, Liu JY, Sakata H, Chen YI, Ofuruton H, Muramatsu S, Chuo YJ, Ohtsuki YH (2004) Atmospheric anomalies observed during earthquake occurrences. *Geophys Res Lett* 31:L17110. <https://doi.org/10.1029/2004GL019865>
- Geller R (ed) (1996) Debate on VAN. *Geophys Res Lett* 23(Special Issue):11
- Gokhberg MB, Morgounov VA, Yoshino T, Tomizawa I (1982) Experimental measurement of electromagnetic emissions possibly related to earthquakes in Japan. *J Geophys Res* 87(B9):7824–7828
- Gordeev E, Gusev A, Levin V, Bakhtiarov V, Pavlov M, Chevrov V, Kasahara M (2001) Preliminary analysis of deformation at the Eurasia-Pacific-North America plate junction from GPS data. *Geophys J Int* 147:189–198
- Gupta H, Shashidhar D, Mallika K, Purnachandra Rao N, Srinagesh D, Satyanarayana HVS, Saha S, Naik RTB (2011) Short term earthquake forecasts at Koyna, India. *J Geol Soc India* 77:5–11. <https://doi.org/10.1007/s12594-011-0001-z>
- Han P, Hattori K, Hirokawa M, Zhuang J, Chen CH, Febriani F, Yamaguchi H, Yoshino C, Liu JY, Yoshida S (2014) Statistical analysis of ULF seismomagnetic phenomena at Kakioka, Japan, during 2001–2010. *J Geophys Res Space Physics* 119:4998–5011. <https://doi.org/10.1002/2014JA019789>
- Han P, Hattori K, Zhuang J, Chen CH, Liu JY, Yoshida S (2017) Evaluation of ULF seismo-magnetic phenomena in Kakioka, Japan by using Molchan's error diagram. *Geophys J Int* 208(1):482–490. <https://doi.org/10.1093/gji/ggw404>
- Hattori K, Han P, Yoshino C, Febriani F, Yamaguchi H, Chen CH (2013) Investigation of ULF seismo-magnetic phenomena in Kanto, Japan during 2000–2010: case studies and statistical studies. *Surv Geophys*. <https://doi.org/10.1007/s10712-012-9215-x>
- Hayakawa M, Kawate R, Molchanov OA, Yumoto K (1996) Results of ultra-low frequency magnetic field measurements during Guam earthquake of 8 August 1993. *Geophys Res Lett* 23:241–244
- He L, Heki K (2016) Three-dimensional distribution of ionospheric anomalies prior to three large earthquakes in Chile. *Geophys Res Lett* 43:7287–7293. <https://doi.org/10.1002/2016GL069863>
- Heki K (2011) Ionospheric electron enhancement preceding the 2011 Tohoku-Oki earthquake. *Geophys Res Lett* 38:L17312. <https://doi.org/10.1029/2011GL047908>
- Heki K, Enomoto Y (2015) Mw dependence of the preseismic ionospheric electron enhancements. *J Geophys Res Space Physics* 120:7006–7020. <https://doi.org/10.1002/2015JA021353>
- Holliday JR, Rundle JB, Turcotte DL, Kristy WK, Tiampo F, Donnellan A (2006) Space-time clustering and correlations of major earthquakes. *Phys Rev Lett* 97:238501
- Igarashi G, Wakita H, Sano T (1992) Precursory and coseismic anomalies in well water levels observed for February 2, 1992, Tokyo Bay earthquake. *Geophys Res Lett* 19:1583–1586
- Ikeya M (2004) *Earthquakes and animals: from folk legends to science*. World Scientific Publications Company Inc. ISBN-13: 978-9812385918
- Ishibashi K (1981) Specification of a soon-to-occur seismic faulting in the Tokai district, central Japan, based upon seismotectonics. In: Simpson DW, Richards PG (eds) *Earthquake prediction: an international review*, Maurice Ewing series, vol 4. AGU, Washington, DC, pp 297–332
- Kamogawa M (2006) Preseismic lithosphere-atmosphere-ionosphere coupling. *Eos* 87:417, 424
- Kamogawa M, Kakinami Y (2013) Is an ionospheric electron enhancement preceding the 2011 Tohoku-Oki earthquake a precursor? *J Geophys Res Space Physics* 118:1751–1754. <https://doi.org/10.1002/jgra.50118>
- Kamogawa M, Nanjo KZ, Izutsu J, Orihara Y, Nagao T, Uyeda S (2019) Nucleation and cascade features of earthquake mainshock

- statistically explored from foreshock seismicity. *Entropy* 21(4):421. <https://doi.org/10.3390/e21040421>
- Kato N, Hirasawa T (1999) A model for possible crustal deformation prior to a coming large interpolate earthquake in the Tokai district, central Japan. *Bull Seismol Soc Am* 89:1401–1417
- Kelley MC, Swartz WE, Heki K (2017) Apparent ionospheric total electron content variations prior to major earthquakes due to electric fields created by tectonic stresses. *J Geophys Res Space Physics* 122:6689–6695. <https://doi.org/10.1002/2016JA023601>
- Kopytenko YA, Matishvili TG, Voronov PM, kopytenko EA, Molchanov OA (1993) Detection of ultra-low-frequency emissions connected with the Spitak earthquake and its aftershock activity, based on geomagnetic pulsation data at Dusheti and Vardzia observatories. *Phys Earth Planet Inter* 77:85–95
- Kushida Y, Kushida R (2002) Possibility of earthquake forecast by radio observations in the VHF band. *J Atmos Electr* 22:239–255
- Lei X, Huang D, Su J, Jiang G, Wang X, Wang H, Guo X, Fu H (2017) *Sci Rep* 7:7971. <https://doi.org/10.1038/s41598-017-08557-y>
- Li Y, Liu Y, Jiang Z, Guan J, Yi G, Cheng S, Yang B, Fu T, Wang Z (2009) Behavioral change related to Wenchuan devastating earthquake in mice. *Bioelectromagnetics*. <https://doi.org/10.1002/bem.20520>
- Lighthill J (ed) (1996) A critical review of VAN – earthquake prediction from seismic electrical signals. World Scientific, Singapore. 376pp
- Lippiello E, Marzocchi W, Arcangelis L, Godano C (2012) Spatial organization of foreshocks as a tool to forecast large earthquakes. *Sci Rep* 2(846). <https://doi.org/10.1038/srep00846>
- Liu JY, Chen YI, Chuo YJ (2006) A statistical investigation of pre-earthquake ionospheric anomaly. *J Geophys Res* 111:A05304. <https://doi.org/10.1029/2005JA011333>
- Liu CC, Linde AT, Sacks IS (2009) Slow earthquakes triggered by typhoons. *Nature* 459:833–836
- Main I (1995) Statistical physics, seismogenesis, and seismic hazard. *Rev Geophys* 34:433–462
- Martilelli G (2018) Contribution to a history of earthquake prediction research. *Geophys Monogr* 234:67–76. Wiley
- McEvelly TV, Johnson LR (1974) Stability of P and S velocities from central California quarry blasts. *Bull Seismol Soc Am* 64:342–353
- Mizutani H, Ishido T, Yokokura T, Ohnishi S (1976) Electrokinetic phenomena associated with earthquakes. *Geophys Res Lett* 3:365–368
- Mogi K (1985) Temporal variation of crustal deformation during the days preceding a thrust-type great earthquake – the 1944 Tonankai earthquake of magnitude 8.1, Japan. *Pageoph* 122:765–780
- Molchan GM, Kagan YY (1992) Earthquake prediction and its optimization. *J Geophys Res* 97:4823–4838. <https://doi.org/10.1029/91JB03095>
- Molchanov OA, Hayakawa M (1998) Subionospheric VLF signal perturbations possibly related to earthquakes. *J Geophys Res* 100:1691–1712
- Moriya T, Mogi T, Takada M (2010) Anomalous pre-seismic transmission of VHF-band radio waves resulting from large earthquakes, and its statistical relationship to magnitude of impending earthquakes. *Geophys J Int* 180(2):858–870
- Nagao T, Orihara Y, Kamogawa M (2014) Precursory phenomena possibly related to the 2011M9.0 off the Pacific Coast of Tohoku Earthquake. *J Disaster R* 9(3):303–310
- Nanjo KZ, Hirata N, Obara K, Kasahara K (2012) Decade-scale decrease in b value prior to the M9-class 2011 Tohoku and 2004 Sumatra quakes. *Geophys Res Lett* 39:L20304. <https://doi.org/10.1029/2012GL052997>
- Nanjo KZ, Izutsu J, Orihara Y, Furuse N, Togo S, Nitta H, Okada T, Tanaka R, Kamogawa M, Nagao T (2016) Seismicity prior to the 2016 Kumamoto earthquakes. *Earth, Planets Space* 68(18). <https://doi.org/10.1186/s40623-016-0558-2>
- Němec F, Santolík O, Parrot M (2009) Decrease of intensity of ELF/VLF waves observed in the upper ionosphere close to earthquakes: a statistical study. *J Geophys Res* 114:A04303. <https://doi.org/10.1029/2008JA013972>
- Niu F, Silver P, Daley T, Cheng X, Majer E (2008) Preseismic velocity changes observed from active source monitoring at the Parkfield SAFOD drill site. *Nature* 454:204–208
- Obara K (2002) Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science* 296:1679–1681
- Orihara Y, Kamogawa M, Nagao T (2014) Preseismic changes of the level and temperature of confined groundwater related to the 2011 Tohoku Earthquake. *Sci Rep* 4(6907):1–6. <https://doi.org/10.1038/srep06907>
- Ouzounov D, Pulinet S, Romanov A, Romanov A, Tsybulya K, Davidenko D, Kafatos M, Taylor P (2011) Atmosphere-ionosphere response to the M9 Tohoku earthquake revealed by multi-instrument space-borne and ground observations: preliminary results. *Earthq Sci* 24:557. <https://doi.org/10.1007/s11589-011-0817-z>
- Parrot M (2018) DEMETER satellite and detection of earthquake signal. In: Singh R, Bartlett D (eds) *Natural hazards: earthquakes, volcanoes, and landslides*. CRC Press, Boca Raton, pp 115–138
- Pisa D, Nemec F, Santolík O, Parrot M, Rycroft M (2013) Additional attenuation of natural VLF electromagnetic waves observed by the DEMETER spacecraft resulting from preseismic activity. *J Geophys Res Space Physics* 118. <https://doi.org/10.1002/jgra.50469>
- Press F (1975) Earthquake prediction. *Sci Am* 232(5):14–23
- Pulinets S, Boyarchuk K (2005) *Ionospheric precursors of earthquakes*. Springer, Berlin. 316 pp
- Raleigh B, Benett D, Craig H, Hanks T, Molnar P, Nur A, Savage J, Scholz C, Turner R, Wu F (1977) Prediction of haisheng earthquake. *Eos* 58(5):2
- Roeloffs E (1994) The earthquake prediction experiment at Parkfield, California. *Rev Geophys* 32(3):315–336
- Sadovsky M, Nersesov I, Nigmatullaev S, Latynina L, Lukk A, Semenov A, Simbireva I, Ulmov V (1972) The processes preceding strong earthquakes in some regions of Middle Asia. *Tectonophysics* 14:295–307
- Sarlis NV, Skordas ES, Lazaridou MS, Varotsos PA (2008) Investigation of seismicity after the initiation of a seismic electric signal activity until the main shock. *Proc Jan Acad Ser B Phys Biol Sci* 84:331–343
- Sarlis NV, Skordas ES, Varotsos PA, Nagao T, Kamogawa M, Tanaka H, Uyeda S (2013) Minimum of the order parameter fluctuations of seismicity before major earthquakes in Japan. *Proc Natl Acad Sci* 110(34):13734–13738. <https://doi.org/10.1073/pnas.1312740110>
- Sarlis NV, Skordas S, Varotsos PA, Nagao T, Kamogawa M, Uyeda S (2015) Spatiotemporal variations of seismicity before major earthquakes in the Japanese area and their relation with the epicentral locations. *Proc Natl Acad Sci* 112(4):986–989. www.pnas.org/cgi/doi/10.1073/pnas.1422893112
- Semyenov AN (1969) Variation in the travel time of traverse and longitudinal waves before violent earthquakes. *Izv. Acad. Sci. USSR (Phys. Solid Earth)*, 4, 245–248 (English transl.)
- Shibazaki B, Matsuura M (1995) Foreshocks and pre-events associated with the nucleation of large earthquakes. *Geophys Res Lett* 22:1305–1308
- Sholtz C, Sykes LR, Aggarwal YP (1973) Earthquake prediction: a physical basis. *Science* 181:803–810
- Takeuchi A, Nagao T (2013) Activation of hole charge carriers and generation of electromotive force in gabbro blocks subjected to nonuniform loading. *J Geophys Res Solid Earth* 118:915–925. <https://doi.org/10.1002/jgrb.50111>

- Tanaka S (2012) Tidal triggering of earthquakes prior to the 2011 Tohoku-Oki earthquake (Mw9.1). *Geophys Res Lett* 39:L00G26. <https://doi.org/10.1029/2012GL051179>
- Thomas JN, Love JJ, Johnston MJS, Yumoto K (2009) On the reported magnetic precursor of the 1993 Guam earthquake. *Geophys Res Lett* 36:L16301. <https://doi.org/10.1029/2009GL039020>
- Toutain JP, Baurbron JC (1999) Gas geochemistry and seismotectonics: a review. *Tectonophysics* 304:1–27
- Tributsch H (1982) When the snakes awake: animals and earthquake prediction. MIT Press, Cambridge, MA
- Tsunomori F, Tanaka H (2014) Anomalous change of groundwater radon concentration monitored at Nakaizu well in 2011. *Radiat Meas* 60:35–41. <https://doi.org/10.1016/j.radmeas.2013.11.006>
- Uyeda S, Kamogawa M (2008) The prediction of two large earthquakes in Greece. *Eos Trans AGU* 89(39):363. <https://doi.org/10.1029/2008EO390002>
- Uyeda S, Nagao T, Orihara Y, Yamaguchi Y, Takahashi I (2000) Geoelectric potential changes: possible precursors to earthquakes in Japan. *Proc Natl Acad Sci U S A* 97:4561–4566
- Uyeda S, Hayakawa M, Nagao T, Molchanov O, Hattori K, Orihara Y, Gotoh K, Akinaga Y, Tanaka H (2002) Electric and magnetic phenomena observed before the volcano-seismic activity in 2000 in the Izu Island Region, Japan. *Proc Natl Acad Sci USA* 99:7352–7355
- Varotsos PA (2005) The physics of seismic electric signals. TerraPub, Tokyo. 358pp
- Varotsos P, Alexopoulos K (1984a) Physical properties of the variations of the electric field of the earth preceding earthquakes, I. *Tectonophysics* 110:73–98
- Varotsos P, Alexopoulos K (1984b) Physical properties of the variations of the electric field of the earth preceding earthquakes, II. *Tectonophysics* 110:99–125
- Varotsos P, Alexopoulos K (1986) Stimulated current emission in the earth and related geophysical aspects. In: Amelinckx S, Gevers R, Nihoul J (eds) *Thermodynamics of point defects and their relation with bulk properties*. North Holland, Amsterdam
- Varotsos P, Sarlis N, Skordas E (2002) Long range correlations in the electric signals that precede rupture. *Phys Rev E* 66(7):011902
- Wakita H, Nakamura Y, Notsu K, Noguchi M, Asada T (1980) Radon anomaly: a possible precursor of the 1978 Izu-Oshima-kinkai earthquake. *Science* 207(4433):882–883. <https://doi.org/10.1126/science.207.4433.882>
- Wang C (1985) Ground-water studies for earthquake prediction in China, *Pure and Applied Geophysics (PAGEOPH)* 122(2–4):215–217. <https://doi.org/10.1007/BF00874594>
- Warwick JW, Stoker C, Meyer TR (1982) Radio emission associated with rock fracture: possible application to the great Chilean earthquake of May 22, 1960. *J Geophys Res* 87:2851–2859
- Wiemer S, Wyss M (1994) Seismic quiescence before the Landers (M=7.5) and Big Bear (M=6.5) earthquakes. *Bull Seismol Soc Am* 84(3):900–916
- Yamada I, Masuda K, Mizutani H (1989) Electromagnetic and acoustic emission associated with rock fracture. *Phys Earth Planet Inter* 57:157–168
- Yamanaka Y, Kikuchi M (2004) Asperity map along the subduction zone in northeastern Japan inferred from regional seismic data. *J Geophys Res* 109:B07307. <https://doi.org/10.1029/2003JB002683>
- Yasuoka Y, Kawada Y, Nagahama H, Omori Y, Ishikawa T, Tokonami S, Shinogi M (2009) Preseismic changes in atmospheric radon concentration and crustal strain. *Phys Chem Earth* 34:431–434
- Yokoi S, Ikeya M, Yagi T, Nagai K (2003) Mouse circadian rhythm before the Kobe earthquake in 1995. *Bioelectromagnetics* 24(4):289–291
- Yoshida S, Uyeshima M, Nakatani M (1997) Electric potential changes associated with slip failure of granite: preseismic and coseismic signals. *J Geophys Res* 102:14883–14897

Earthquake Prediction, M8 Algorithm

Alik Ismail-Zadeh^{1,2} and Vladimir Kossobokov²

¹Institute of Applied Geosciences, Karlsruhe Institute of Technology, Karlsruhe, Germany

²Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Sciences (IEPT RAS), Moscow, Russia

Definition

An *earthquake* is a sudden movement within the Earth's crust or the upper mantle that releases tectonic stress and, usually, originates seismic waves. The hierarchy of movable lithospheric volumes composes a large nonlinear dynamical system. Prediction of such a system in a sense of extrapolation of trajectory into the future is futile. However, upon coarse-graining, the integral empirical regularities emerge opening possibilities of earthquake prediction (Keilis-Borok et al. 2001).

Earthquake prediction is a statement about future earthquake occurrence based on the information, data, and scientific methods that are available now. To predict an earthquake, someone “must specify the expected magnitude range, the geographical area within which it will occur, and the time interval within which it will happen with sufficient precision so that the ultimate success or failure of the prediction can readily be judged. Only by careful recording and analysis of failures as well as successes can the eventual success of the total effort be evaluated and future directions charted. Moreover, scientists should also assign a confidence level to each prediction” (National Research Council 1976).

Introduction

According to the definition of earthquake prediction, one can classify an earthquake prediction of certain magnitude range by duration of time interval and/or by territorial specificity. Four major stages can be distinguished in temporal earthquake predictions: (1) long-term, (2) intermediate-term, (3) short-term, and (4) immediate (see Table 1).

Long-term prediction is essentially based on the determination of probabilities of active fault segments to rupture for the next few decades (e.g., Working group on California earthquake probabilities 1999). This kind of prediction can guide engineering and emergency planning measures to mitigate the impact of a target earthquake.

An intermediate-term prediction for the next few years is an update of the long-term prediction by some indicators (e.g., an increase in background seismicity, clustering of events in

space and time, transformation of magnitude distribution, and some others). The M8 algorithm is designed for an intermediate-term prediction of earthquakes (Keilis-Borok and Kossobokov 1990).

In the 1980s, P. Varotsos, K. Alexopoulos, and K. Nomicos proposed the VAN method (named after the researchers' initials) of short-term earthquake prediction for the next few months/weeks, which is based on detection of characteristic changes in the geoelectric potential via a telemetric network of conductive metal rods inserted in the ground (Lighthill 1996).

Immediate earthquake prediction for the next few hours is sometimes mixed with "early warning," which is usually based on the first arrival of seismic waves and transmission of an electronic alert within a lead-time of seconds. Early warning is used (e.g., in Japan) to shut down gas and electricity grids and to stop high-speed trains in the event of a strong earthquake.

Following common perception, some investigators concentrate their efforts on predicting the "exact" fault segment to rupture, e.g., the Parkfield earthquake prediction experiment (Bakun and Lindh 1985), which is by far more difficult and might be an unsolvable problem (Main 1999). Being related to the rupture size L of the incipient earthquake, predictions could be summarized in a classification that distinguishes

several options of wider spatial ranges of certainty and not only "exact" mode of prediction (Table 1). The Gutenberg–Richter power law relation between the number of earthquakes and their magnitude suggests limiting magnitude range of prediction, $Mm + = [m, m + \Delta m]$, to about one unit or less (i.e., $\Delta m \leq 1$).

From a viewpoint of such a classification, the earthquake prediction problem might be approached by a hierarchical, step-by-step prediction technique, which accounts for multi-scale escalation of seismic activity to the main rupture (e.g., Keilis-Borok 1990; Kossobokov et al. 1999). It starts with term-less recognition of the earthquake-prone zones (Gorshkov et al. 2003) and then follows with determination of the areas of different spatial ranges, where long- and intermediate-term *times of increased probability* (TIP) of a target earthquake occurrence are estimated. Finally, a forecast may come out with short- or immediate-term alerts at specific sites.

The M8 Algorithm

The intermediate-term middle-range earthquake prediction method, called briefly the *M8 algorithm*, was designed by retroactive analysis of dynamics in seismicity preceding the great, magnitude 8.0+, earthquakes worldwide, hence its name. The exact definitions and computer code of the M8 algorithm can be found in Keilis-Borok and Kossobokov (1990), Healy et al. (1992), and Kossobokov (1997).

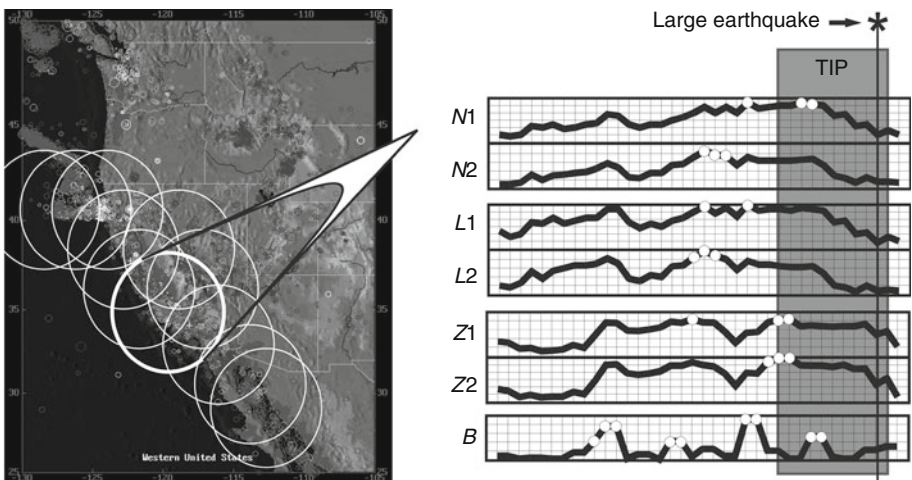
Prediction is aimed at the earthquakes of magnitude M , where $M \in [m_0, m_0 + \Delta m] = M_0 +$. The territory of the seismic region under study is analyzed by moving overlapping circles of the fixed size (by a factor of about 5–10 larger than a source size of target magnitude M_0 events) (Fig. 1). The sequence of earthquakes with aftershocks

Earthquake Prediction, M8 Algorithm, Table 1 Classification of earthquake predictions

Temporal, in years		Spatial, in source zone size L	
Long-term	10	Long-range	Up to $100 L$
Intermediate-term	1	Middle-range	$5\text{--}10 L$
Short-term	0.01–0.1	Narrow	$2\text{--}3 L$
Immediate	0.001	Exact	$1 L$

Spatial accuracy of predictions depends on the source size L of target earthquakes

Earthquake Prediction, M8 Algorithm, Fig. 1 General scheme of the M8 algorithm. *Left panel:* Circles of investigation overlay the targeted seismic region; seismic sequences in each circle give a description of seismic activity, which is then used to diagnose an alert (TIP). *Right panel:* presents the plots of the M8 algorithm's functions (explained below) with time. The interval of TIP is shaded by gray; the asterisk denotes the large event occurred; and the white dots show anomalously high values of the functions



removed is considered within each circle. Sequences in different circles are normalized to the annual number of earthquakes $\bar{N}$ by selecting the lower magnitude cutoff $\underline{M} = M_{\min}(\bar{N})$.

For this sequence, several functions are computed in the trailing time window $(t-s, t)$ and the magnitude range $(\underline{M} \leq \text{magnitude} < M_0)$. These functions include (1) the number of earthquakes $N(t)$ of magnitude M or greater; (2) the deviation of $N(t)$ from longer-term trend $L(t)$; (3) concentration $Z(t)$ estimated as the ratio of the average source diameter to the average distance between sources; and (4) the maximum number of aftershocks $B(t)$. Each of the functions N , L , and Z is calculated twice with $\underline{M} = M_{\min}(\bar{N})$ for $\bar{N} = 20$ and $\bar{N} = 10$. As a result, the earthquake sequence is given a robust averaged description by seven functions $N1$, $N2$, $L1$, $L2$, $Z1$, $Z2$, and B (Fig. 1). An alarm or a TIP is declared for 5 years when at least six out of seven functions, including B , become “anomalously large” within a time window of 3 years.

Retrospectively the M8 algorithm was applied to predict earthquakes from different magnitude ranges in a number of regions worldwide. Its modified versions targeting earthquakes from magnitude ranges down to M5.5+ were found efficient in regions of seismic activity lower than required by the original version (Kossobokov and Shebalin 2003; and references therein).

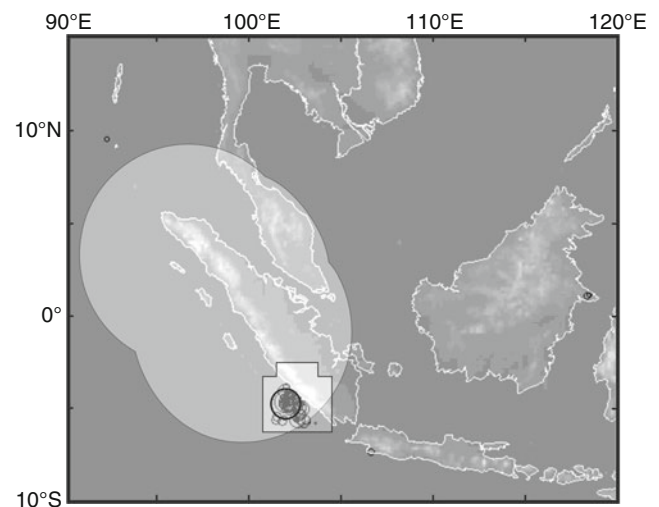
The MSc Algorithm

The second approximation prediction method, called briefly the *MSc algorithm* (Kossobokov et al. 1990), was designed by retroactive analysis of the detailed regional seismic catalog prior to the Eureka earthquake (1980, M7.2) near Cape Mendocino in California, hence its name, *Mendocino Scenario*, and an abbreviation. The MSc algorithm is applied to reduce the area of alarm by analyzing dynamics at lower magnitude levels of seismic hierarchy. Sometimes, the data is enough to get a near-perfect outline of the incipient large earthquake. More often, the catalog of earthquakes is just enough for the M8 analysis but insufficient for the MSc application.

Qualitatively, the MSc algorithm determines an area of the territory of alarm, where the activity recognized by the first approximation prediction algorithm (e.g., by the M8 algorithm) within a sufficient temporal and/or spatial span is continuously high and infrequently drops for a short time. The phenomenon, which is used in the MSc algorithm, might reflect the second (possibly, shorter-term and, definitely, narrow-range) stage of the premonitory rise of seismic activity near the incipient source of the main shock. The MSc algorithm reduces area of the M8 predictions much better (by at least a factor of 2) than a few simple alternatives like all earthquake-prone or the most active cells that contain

certain part of the recent seismic activity (Kossobokov et al. 1990). Fig. 2 shows, as an example of the M8-MSc forecast in the real time, the case history of June 4, 2000, M8.0 Sumatra earthquake.

The earthquake prediction algorithms M8 and MSc receive a fair amount of attention due to ongoing real-time experimental testing unprecedented in rigor and global coverage (Healy et al. 1992; Kossobokov et al. 1999). Relevant predictions are communicated, with due discretion, to several dozens of leading scientists and administrators in many countries. The accumulated statistics of this experiment confirm intermediate-term predictability of large earthquakes with middle- to exact-range of location (Kossobokov 2006). Table 2 summarizes the prediction outcomes for the great (magnitude 8.0+) earthquakes. For the last 34 years, 17 out of 27 great earthquakes were predicted by the M8 algorithm, while the M8-MSc algorithm narrows the area of alarm at the cost of additional six failures-to-predict. The ratio of the *space-time volume alerted by the algorithms* (i.e., alarms) to the total space-time volume is about 31% in the case of the M8 algorithm and 15% in the case of the M8-MSc algorithm. A confidence level estimate tells how sure one can be that the achieved performance of the algorithm is not arisen by chance. To bring the confidence level down to generally accepted value of 95%, the ongoing experimental testing has to encounter 11 failures-to-predict in a row for the M8 algorithm and 18 in the case of the M8-MSc narrow-range predictions, which appears absolutely unlikely to happen. The progress in hypocenter determination in the last decades resulted in a possibility of expanding the territory of testing the M8 and M8-MSc algorithms (Kossobokov 2017).



Earthquake Prediction, M8 Algorithm, Fig. 2 Prediction of June 4, 2000, M8.0 Sumatra earthquake. The highlighted overlapping circular areas of alarm determined by the M8 algorithm and the highlighted overlapping rectangular areas of alarm by the MSc algorithm. The main event and its aftershocks are marked by circles

Targeting Mega-earthquakes

The December 26, 2004, M9.2 Sumatra-Andaman (S-A), March 28, 2005, M8.5 Nias, February 27, 2010, M8.8 Chile, and March 11, 2011, M9.1 Tohoku mega-earthquakes, which ruptured the 1300-, 400-, 600-, and 500-km portions of subducting margins of the Indian and Pacific Oceans, are not included in Table 2 because of their rupture size. According to the M8 methodology, much larger circles of investigation (CI) with diameter of several thousand kilometers should then be employed to search for a precursory seismic activation in advance of mega-earthquakes. An application of the M8 algorithm targeting M9.0+ earthquakes could have predicted these earthquakes (see <http://www.mitp.ru>, section “Predictions”).

Actually, the retrospective application of the M8 algorithm (with the 6000-km diameter CI) showed that in July 2004 the precursory seismic activation was widespread over the most of the earthquake-prone areas of the Earth (Romashkova 2008). This activation reduced to four clusters of TIPs by 2008. The largest cluster includes S-A and adjacent major seismic zones from Kashmir in the North-West to Timor Islands in the East, where March 28, 2005, Nias, M8.7

mega-earthquake and coupled September 12, 2007, Southern Sumatra, M8.5 and M7.9 great earthquakes have already confirmed the M8 algorithm diagnosis (see <http://www.mitp.ru>; note that an access to the current predictions is available only for experts; contact volodya@mitp.ru for the access code). Two of the other three smaller clusters of TIPs, i.e., the circles centered at (31°S, 70°W) and (45°N, 150°E) got their confirmations with the occurrences of February 27, 2010, offshore Maule, Chile M8.8 and March 11, 2011, M9.1 Tohoku mega-earthquakes (Fig. 3).

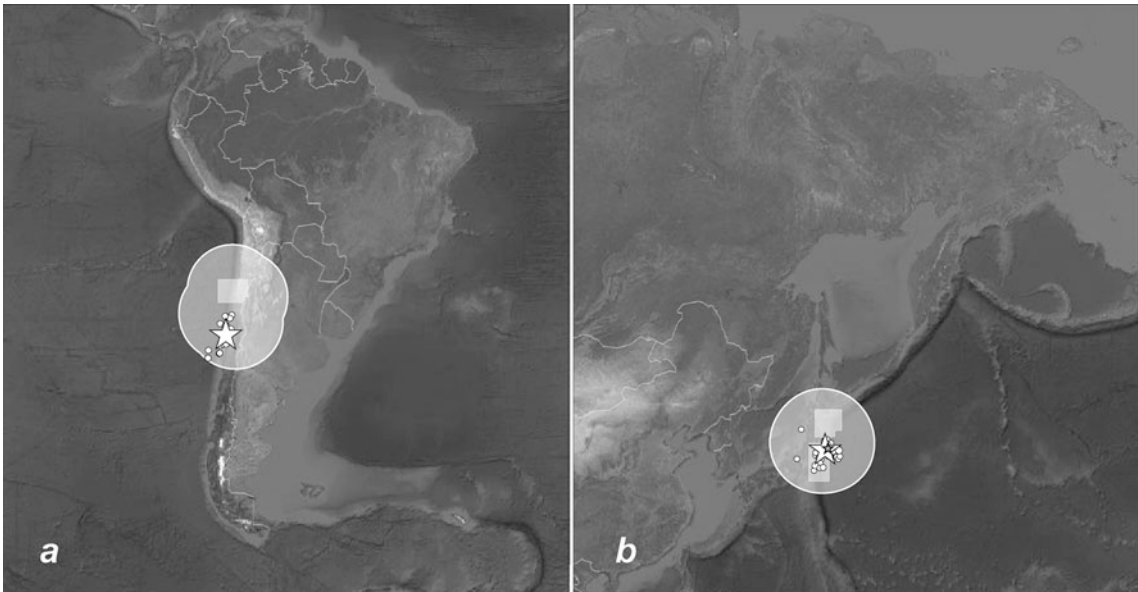
Figure 3 does not show these CIs, but the TIPs in smaller (667-km diameter) CIs of the regular M8 algorithm diagnosis aimed at M8.0+. Evidently, this provides a narrow-range reduction of the prediction uncertainty acting as the second approximation to diagnosis aimed at M9.0 + .

Conclusion

High statistical significance of intermediate-term earthquake prediction methods, achieved in the ongoing experimental testing worldwide, confirms the following paradigms:

Earthquake Prediction, M8 Algorithm, Table 2 Worldwide performance of earthquake prediction algorithms M8 and M8-MSc

Test period	Large earthquakes						
	Predicted by		Total	Alarms, %		Confidence level, %	
	M8	M8-MSc		M8	M8-MSc	M8	M8-MSc
1985–2019	17	11	27	30.92	14.59	99.94	99.91



Earthquake Prediction, M8 Algorithm, Fig. 3 Global testing of the M8-MSc predictions in advance of February 27, 2010, M8.8 offshore Maule, Chile (a) and March 11, 2011, M9.1 Tohoku (b) mega-earthquakes. The epicenters of the main shock (big star) and its after-shocks (white dots) fill the portion of the subduction zone, which is about

one half of the major seismic segment recognized (gray outline) as capable of producing a magnitude M8.0+ event in the ongoing real-time experimental testing. March 9, 2011, Sankiru-Okii M7.3 foreshock of the Tohoku mega-earthquake is marked by a small dark gray star

(1) precursory seismicity patterns exist; (2) the dimension of an area where precursory seismicity pattern appear is by far (about ten times) larger than that of source zone of the incipient large earthquake; (3) many precursory seismicity patterns are similar in regions of different seismic and tectonic environment; and (4) some precursory seismicity patterns are universal. Their analogs take place in advance extreme catastrophic events in other complex nonlinear systems.

Our knowledge of earthquake physics and earthquake dynamics is not sufficient to predict strong earthquakes with a high accuracy. We do not know well (1) how earthquakes, especially large events, originate; (2) when an earthquake starts, when it stops, and what magnitude could it be; (3) how and why earthquakes cluster; and (4) what were the initial conditions of stress state before a large event in terms of stress transfer.

The scientific community should use the full potential of mathematics, statistics, statistical physics, and computational modeling and the data derived from seismological (monitoring of physical parameters of earthquakes and tectonic stress, fluid migration, etc.), geodetic (GPS, InSAR, and other measurements of the crustal deformation), and geological (e.g., determination of the time intervals between large earthquakes using paleo-seismological tools) studies to improve intermediate- and short-term earthquake predictions (Ismail-Zadeh 2010).

Though the current accuracy of earthquake prediction is limited, any scientifically validated prediction can be useful for earthquake preparedness and disaster management, if the accuracy of the prediction is known, even though it is not high. An inexpensive low-key response to intermediate-term middle-range predictions described here is well justified, if even a little part of the total damage due to a strong event is prevented.

Cross-References

- [Earthquake Precursors and Prediction](#)
- [Earthquake, Aftershocks](#)
- [Earthquake, Foreshocks](#)
- [Earthquake, Magnitude](#)
- [Earthquakes, Early and Strong Motion Warning](#)
- [Earthquakes, Strong-Ground Motion](#)
- [Great Earthquakes](#)
- [Seismic Hazard](#)

Bibliography

- Bakun WH, Lindh AG (1985) The Parkfield, California, earthquake prediction experiment. *Science* 229:619–624
- Gorshkov AI, Kossobokov V, Soloviev A (2003) Recognition of earthquake-prone areas. In: Keilis-Borok VI, Soloviev AA (eds) *Nonlinear dynamics of the lithosphere and earthquake prediction*. Springer, Heidelberg, pp 239–310
- Healy JH, Kossobokov VG, Dewey JW (1992) A test to evaluate the earthquake prediction algorithm, M8. US Geol Surv Open-File Report 92–401
- Ismail-Zadeh A (2010) Computational geodynamics as a component of comprehensive seismic hazards analysis. In: Beer T (ed) *Geophysical hazards: minimizing risk and maximizing awareness*. Springer, Amsterdam, pp 161–178
- Keilis-Borok VI (1990) The lithosphere of the Earth as a nonlinear system with implications for earthquake prediction. *Rev Geophys* 28:19–34
- Keilis-Borok VI, Kossobokov VG (1990) Premonitory activation of earthquake flow: algorithm M8. *Phys Earth Planet Inter* 61:73–83
- Keilis-Borok VI, Ismail-Zadeh AT, Kossobokov VG, Shebalin PN (2001) Non-linear dynamics of the lithosphere and intermediate-term earthquake prediction. *Tectonophysics* 338(3–4):247–259
- Kossobokov V (1997) User manual for M8. In: Healy JH, Keilis-Borok VI, Lee WHK (eds) *Algorithms for earthquake statistics and prediction*. IASPEI software library, vol 6. Seismological Society of America, El Cerrito, pp 1–167
- Kossobokov V (2006) Quantitative earthquake prediction on global and regional scales. In: Ismail-Zadeh A (ed) *Recent geodynamics, georisk and sustainable development in the Black Sea to Caspian Sea region*. Conference proceedings, vol 825. American Institute of Physics, New York, pp 32–50
- Kossobokov VG (2017) Testing an earthquake prediction algorithm: the 2016 New Zealand and Chile earthquakes. *Pure Appl Geophys* 174(5):1845–1854
- Kossobokov V, Shebalin P (2003) Earthquake prediction. In: Keilis-Borok VI, Soloviev AA (eds) *Nonlinear dynamics of the lithosphere and earthquake prediction*. Springer, Heidelberg, pp 141–207
- Kossobokov VG, Keilis-Borok VI, Smith SW (1990) Localization of intermediate-term earthquake prediction. *J Geophys Res* 95(B12):19763–19772
- Kossobokov VG, Maeda K, Uyeda S (1999) Precursory activation of seismicity in advance of Kobe, 1995, M = 7.2 earthquake. *Pure Appl Geophys* 155:409–423
- Lighthill J (ed) (1996) *A critical review of VAN: earthquake prediction from seismic electric signals*. World Scientific Publication, Singapore
- Main I (1999) Is the reliable prediction of individual earthquakes a realistic scientific goal? *Nature debates* at <http://www.nature.com/nature/debates/earthquake>. Accessed 31 Mar 2019
- National Research Council (1976) *Predicting earthquakes: a scientific and technical evaluation, with implications for society*. The National Academies Press, Washington, DC. <https://doi.org/10.17226/18533>
- Romashkova LL (2008) Global-scale analysis of seismic activity prior to 2004 Sumatra-Andaman mega-earthquake. *Tectonophysics* 470:329–344
- Working Group on California Earthquake Probabilities (1999) *Earthquake probabilities in the San Francisco Bay region: 2000–2030 – a summary of findings*. US Geol Surv Open-File Report 99–517

Earthquake Rupture: The Inverse Problem

Shamita Das

Department of Earth Sciences, University of Oxford,
Oxford, UK

Definition

Forward problems are those in which given the cause we are required to find the effect. For example, in seismology, given the Earth structure, find the expected arrival times of different

seismic phases. Or, given the details of the earthquake rupture process, find the seismograms at different stations. Inverse problems are those in which we are given the effect and required to find the cause. For example, given the times of arrivals of, say, P-waves at different stations from an earthquake, find the hypocentral location. Or, given the travel time curves, find the Earth structure. In seismology, as in all branches of geophysics, it is inverse problems that we are most interested in, forward problems being very useful in elucidating the physical principles underlying the problem.

Forward problems	Inverse problems
Serve to elucidate underlying physical principles	Say, given some gravity anomaly, find shape and density of anomalous body
Solutions are unique	Solutions are non-unique, i.e., more than one density distribution can give same anomaly

Introduction

The inverse problem of earthquake source mechanics consists of analyzing seismograms to obtain detailed information on the earthquake rupturing process. In particular, some of the questions that seismologists are interested in are how does the motion develop over the fault area during the time period that the two sides of the fault are moving past one another, at what rupture speed does slip spread out over the fault, at what speed do the two sides of the fault slide past one another, and what is the resulting stress drop over the fault. Such information is essential for reliable seismic hazard assessment.

Definition of Some Common Terms Used in Inverse Theory

Stability is defined as the property that a solution is insensitive to small random errors in the data. Non-uniqueness is the property that more than one solution (almost) equally well-fitting the data exists. Robustness is the property that a solution is insensitive to a small number of large errors (“outliers”) in the data. As stated by Das and Kostrov (1990), from the computational point of view, this instability manifests itself as non-uniqueness of the solution, in numerical problems.

The solution of this problem is far from trivial, because it is unstable (Kostrov 1975; Kostrov and Das 1988), and from the computational point of view, this instability manifests itself as non-uniqueness of the solution, as mentioned above. Adding constraints to the solution of the relevant equations has been shown to stabilize the problem. Such constraints appear best when based on the physics of problem, that is, the physics of the earthquake faulting process. Unfortunately, our knowledge of the physics of the earthquake process is still limited.

So, seismologists have had to use those properties which are considered to be reliably known. For example, forward modeling showed that it is not likely that there is major back-slip on the fault. So, a constraint that is often used is the requirement is that the slip rate vector is positive, that is, no back-slip is permitted on the fault. While, in principle, slip in the opposite direction is possible due to interference of waves on the fault, especially at the fault edges, experience from three-dimensional forward modeling shows that such back-slip is likely to be very small. This constraint has been used by Olson and Apsel (1982), Hartzell and Heaton (1983), and Das and Kostrov (1990, 1994), among others, and has been shown to stabilize the solution. Other possible constraints, such as requiring the solution to be in agreement with the seismic moment obtained from the centroid moment tensor solution or from geodetic measurements, or, requiring the fault rupture speed to be limited by some wave velocity, can also be used. Further constraints such as requiring the maximum slip rate on the fault to be limited by some considerations from fracture mechanics could also be considered, though this has not yet been implemented in currently used methods. It has been found that minimizing the seismic moment often provides stable solutions. Laboratory experiments have recently played a crucial role in providing such constraints and will be discussed later.

As stated by Olson and Apsel (1982), in order for a particular rupture process to be an acceptable solution to the inverse problem, it must satisfy the following three conditions:

- 1. The solution must explain the data.
- 2. The solution must be physically reasonable (consistent with independent constraints).
- 3. If more than one solution fits the data equally well, additional information must be supplied to uniquely define which solution is being obtained.

The Method

The rupture process is solved for by determining the slip rate distribution over the fault in time, and the entire rupture process is reconstructed from it. For example, the final slip distribution is found by integrating the slip rate distribution over time. The rupture speed can be found by tracking the rupture front and so on. It should be noted that fault slip rate and fault rupture speed are very different quantities and should not be confused. Faults slip at rates of about m/s, whereas the rupture speed is generally about 90% or more, of the shear wave speed of the medium in which the fault is propagating, that is about 3 km/s or more.

To invert for the three-dimensional slip rate distribution (two spatial dimensions on the fault and time) is a computationally difficult task. Not surprisingly, simplified methods of

E

solving the problem have been used. For example, in order to reduce the number of unknowns, one often restricts for how long a point on the fault can slip. But, in reality, if some region of the fault re-slips, or has delayed slip, both shown to be feasible in forward models, this may not be found by the solution. This restriction would inhibit discovering fault complexity. The discretization itself may implicitly impose some constraint on the solution. It is thus important to ensure that the solution methods do not impose any implicit (and unknown) constraints on the solution, and if such constraints are unavoidable, they must be kept in mind when interpreting such solutions in geophysical terms.

To formulate the problem, we start from the representation theorem (e.g., Aki and Richards 1980), where the ground displacement at a station located at a point $\mathbf{x}$ on the Earth's surface can be expressed in terms of the slip distribution over a fault Σ as

$$u_k(\mathbf{x}_1, t_1) = \int_0^{t_1} dt \iint_{\Sigma} K_{ik}(\mathbf{x}_1, \mathbf{x}, t_1, t) a_i(\mathbf{x}, t) dS \quad (1)$$

where $i, k = 1, 2, 3$, $u_k(\mathbf{x}_1, t_1)$ are the components of the displacement vector, $a_i(x, t)$ are the components of the slip, and $K_{ik}(\mathbf{x}_1, \mathbf{x}, t_1, t)$ are the components of the impulse response of the medium at $(\mathbf{x}_1, t_1)$, due to a dislocation point source at $(\mathbf{x}, t)$. Assuming a planar fault and the slip direction on the fault to be constant, we can write, after some transformations:

$$S_j(t_1) = \int_0^{t_1} dt \iint_{\Sigma} W_j(\boldsymbol{\xi}, t_1 - t) \dot{a}(\boldsymbol{\xi}, t) dS \quad (2)$$

where j identifies the station and ground displacement components of the seismogram $S(t_1)$, $W_j(\boldsymbol{\xi}, t)$ is the impulse response at $(\boldsymbol{\xi}, t)$ corresponding to a fixed slip direction, $\dot{a}(\boldsymbol{\xi}, t)$ is the fault slip rate, the two-dimensional vector $\boldsymbol{\xi}$ gives the position on the fault relative to some reference point (e.g., the earthquake hypocenter), and t is the time measured from the origin time of the earthquake. With a continuous distribution of stations, Eq. (2) would represent an integral equation of the first kind (and such equations are known to be unstable).

The following additional constraints are often used:

$$\dot{a}(\boldsymbol{\xi}, t) \geq 0 \text{ for all } (\boldsymbol{\xi}, t) \quad (3a)$$

$$\dot{a}(\boldsymbol{\xi}, t) = 0 \text{ for } t < T(\boldsymbol{\xi}) \quad (3b)$$

where $t = T(\boldsymbol{\xi})$ gives the boundary of the area where slip is permitted (due to some chosen causality condition) at time t . Note that $t = T(\boldsymbol{\xi})$ is not necessarily the actual rupture front, which should not generally be assigned a priori, but shown to

be determined as part of the solution, as was done by Das and Kostrov (1990), Henry et al. (2000), and others studies using this method.

$$\int_0^\infty dt \iint_{\Sigma} \mu(\boldsymbol{\xi}) \dot{a}(\boldsymbol{\xi}, t) dS = M_0 \quad (3c)$$

where M_0 is the seismic moment and μ is the modulus of rigidity of the medium. The constraint (3a) is the “no-back-slip” constraint or the “positivity” constraint, and the constraint (3b) is a “causality constraint,” which could be taken as the P - or S -wave causal front, or any other chosen region. The constraint (3c) is the “seismic moment constraint.”

We restrict this discussion to long-period body waves at teleseismic distances (“far-field”) because they are less sensitive to the details of Earth structure. This problem has been looked at using ground motion data from seismograms located at regional distances from earthquakes, for example, for the 1985 Mexican earthquake (Mendoza and Hartzell 1989) and the 1992 Landers, California, earthquake (Cohee and Beroza 1994; Wald and Heaton 1994), but the problem has not yet been considered for the case of true “near-field” data where the higher order terms are kept in the Green function calculations. The Green functions W_j for the teleseismic problem can be calculated using, for example, the ray approximation and the Thomson-Haskell technique or ray-tracing methods to account for the layered crustal structure at the source and the station. These techniques take into account all surface reflections near the source and at the receiver, such as pP , sP , pS , and sS , as well as all reflections and transmissions at every layer of the crustal structures. For the numerical solution, the integrals in (2) must be discretized. For this, the area Σ where slip is permitted must be divided into a number of cells (usually square or rectangular) and $\dot{a}$ within each cell approximated by linear functions in time and along strike and dip (if the width of the cells are small, the variation along dip can be taken as constant). The W_j 's must then be integrated over each cell (analytically, if possible, as done by Das and Kostrov (1990, 1994)) and the integrals over the fault area replaced by sums. The time at the source is discretized by taking a fixed time step, Δt say, and it could be assumed that the slip rate $\dot{a}$ during the time step is a constant, or varies in some prescribed way, say, linearly with time. Both the fault area and the total source duration should be determined as part of the inversion process. One can use the aftershock area as starting area, and then refine it further during the inversion. The initial source duration cannot be longer than the longest record used, in other words, the time length of the seismograms used need to be long enough to study the entire source process. It is important to keep in mind that at many distances other phases, such as core reflections, arrive within this time window, and either the seismogram

used has to be terminated just before this time or these additional phases correctly modeled. Unfortunately, often, neither is done in some studies, based on the incorrect assumption that core reflections are always small compared to the direct arrivals. The seismograms are sampled at some chosen time step Δt_1 , not inconsistent with the spatial cell size of the fault and time step of the source duration. Then Eq. (2) takes the form

$$A\mathbf{x} \approx \mathbf{b} \quad (4)$$

where A is the matrix obtained by integration of W_j , each column of A being a set of Green function seismograms for all stations corresponding to different cells and time instants of the source duration, ordered in the same way as the observed seismograms. The arrangement of Eq. (4) is shown in Fig. 1.

Different stations and components can be weighted differently in Eq. (2). This is very useful when inverting P- and S-waves simultaneously.

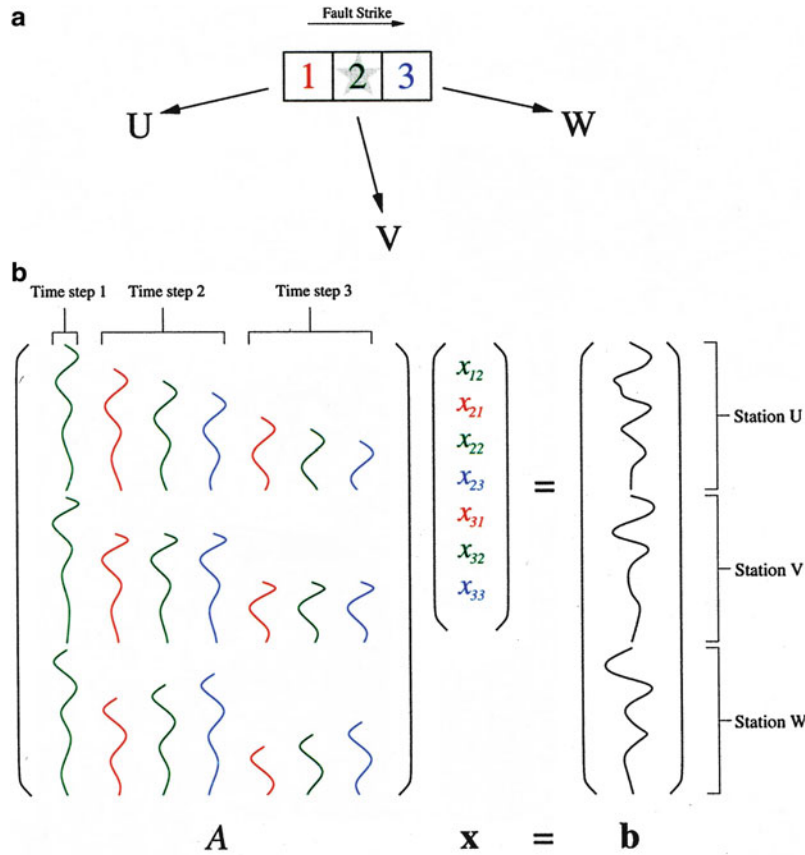
For the discrete problem, the positivity constraint (3a) becomes:

$$\mathbf{x} \geq 0, \quad (5a)$$

the inequality meaning that every component of $\mathbf{x}$ is non-negative. The causality constraint (3b) can be replaced by

$$x_i = 0 \quad (5b)$$

for those i corresponding to cells and time samples outside the “rupture front” and the seismic moment constraint (3c) becomes:



Earthquake Rupture: The Inverse Problem, Fig. 1 A simplified example of the matrix A for the inverse problem. (a) Schematic map of the unilateral fault used in the inversion. It has three numbered spatial cells along the fault strike and one cell along its down-dip width. The rupture initiates at time-step 1, in cell 2, and proceeds to the left, breaking cell 2 and then cell 3. The azimuths to three stations (U, V, W) are shown by arrows. (b) Schematic showing the arrangement of the linear system $A\mathbf{x} = \mathbf{b}$ for this problem. In the matrix A , containing the concatenated Green functions, portions relating to the cells 1, 2, and 3 drawn in red,

green, and blue, respectively. The unknown vector $\mathbf{x}$ is the slip rate in each cell 1, 2, or 3, at time-steps 1, 2, and 3 and, in this simple example, has seven elements. In reality, there will be many more time steps, and A will have many more columns. Since only cell 1 is allowed to slip at the first time step, by invoking the causality condition the unknown slip rates x_{11} and x_{13} are zero, and hence excluded from the linear system. The vector of seismograms $\mathbf{b}$ consists of the seismograms at the stations, placed one after another, but having the appropriate sample length consistent with the l.h.s. of the equation

$$\sum c_i x_i \approx M_0 \quad (5c)$$

where c_i is time-independent and for each cell is equal to the product of the average rigidity multiplied by the area of the cell. So, the inverse problem has been reduced to the solution of the linear system (4) under one or more of the constraints (5a), (5b), (5c). In the system (4), the number of equations, m , is equal to the total number of samples taken from all the records involved and the number of unknowns, n , is equal to the number of cells multiplied by the number of time steps at the source. We shall take $m > n$ to reduce the influence of the noise contained in the observations $\mathbf{b}$ on the solution. Then, the system (4) is overdetermined and we can only obtain a solution $\mathbf{x}$ which provides a best fit to the observations, under constraints (5). It is important to note that, in theory, for the constrained system, it is not necessary for $m > n$, though it has been so in all actual studies of earthquakes.

It is now well understood that the matrix A can be ill-conditioned which implies that the system (4) admits more than one solution, equally well fitting the observations. The constraints (5) are introduced for the purpose of reducing the set of permissible (feasible) solutions. When a “best-fitting” solution is obtained, other solutions almost satisfying the equations are also of great interest. This is because the data used in geophysical applications often contains noise and the models used are themselves approximations to reality, so that the “best-fitting” solution may not represent actuality. In addition, though the different solutions may be very close together in data-space, it is possible that they may be somewhat different in model-space, that is, they may physically represent rather different solutions. Thus, more than one solution must be found and the set of (almost) equally well-fitting solutions examined to find features common to the set of solutions, as was done by Das and Kostrov (1990), Henry et al. (2000), and others studies using this method. Das and Kostrov (1990, 1994) also suggested additional tests to determine the robustness of such features. Only features clearly required by the solutions should be interpreted in a geophysical context, as was done in the specific examples discussed below.

For the system of equations (4) together with the constraints (5) to comprise a complete mathematical problem, it is necessary to formulate the exact form of what the “best-fit” to observations means. We include here only constraint (5a) in the mathematical formulation for the sake of simplicity, the inclusion (5b) and (5c) being simple, and is discussed, for example, by Das and Kostrov (1990, 1994). We have to minimize the vector of residuals:

$$\mathbf{r} = \mathbf{b} - A\mathbf{x} \quad (6)$$

For this purpose, some norm of the vector $\mathbf{r}$ must be adopted. One may choose to minimize the ℓ_1 , the ℓ_2 or the

ℓ_∞ norm, all three being equivalent in the sense that they tend to zero simultaneously. Hartzell and Heaton (1983) used the ℓ_2 norm to solve the optimization problem with the positivity constraint, employing computer programs in Lawson and Hanson (1974). Das and Kostrov (1990, 1994) have used the linear programming method (Press et al. 1986) to solve the system (4) and minimize the ℓ_1 norm subject to the condition (5a). The ℓ_1 norm is considered more robust (Tarantola 1987), and Hartzell et al. (1991) have investigated the effect of minimizing different norms on the solution and have shown that the most robust features are properly obtained in the solutions, whichever norm is used. Das and Kostrov (1990, 1994) have checked that when the ℓ_1 norm is minimized, the ℓ_2 and the ℓ_∞ norms for the solution are also small.

In reality, the actual scales at which the rupture propagation takes place in the Earth is microscopic whereas the inverse problem must be solved using finite grids. Therefore, it is important to consider how well the inverse solution obtained gives a picture of the actual solution. To do this, it is necessary to decide how large the cells in the inversion should be. Clearly, the smaller the cells, the better the approximation to the integral equation being solved (Eq. 1) but the larger the number of unknowns. On the other hand, though the approximation to the integral equation improves by going to smaller and smaller cells, the condition number of the matrix A increases (due to columns becoming very similar) making the problem more and more unstable. As pointed out by Backus and Gilbert (1967), since the amount of data is finite, the problem of finding a continuous function from it is indeterminate. Das and Suhadolc (1996) and Das et al. (1996) have suggested that synthetic tests using artificially created data should be used to best estimate the spatial and temporal grid sizes and that the spatial and temporal grid sizes should be consistent. Henry et al. (2000) have also compared the Green functions obtained using various grid sizes to aid in the grid size selection. Sarao et al. (1998) have studied the effect of non-uniform seismic station distribution around the earthquake epicenter.

Constraints from Laboratory Experiments

As mentioned earlier, advanced technological developments has led to major improvements in laboratory experiments on rupture propagation. One important result is the demonstration that ruptures can not only exceed the shear wave speed of the medium but can even reach its compressional wave speed value. Though super-shear ruptures

speed had been found in laboratory experiments as far back as the 1970s (Wu et al. 1972), since these experiments were conducted on plastic polymer, under very low normal stresses, this was considered unrealistic for real earthquakes – both the material and the low normal stress – and the results were ignored. In the late 1990s to the early 2000s, a group at Caltech, led by Rosakis, measured earthquake speeds in the laboratory, not only exceeding the shear wave speed (Rosakis et al. 1999; Xia et al. 2004) but even reaching the compressional wave speed (Xia et al. 2005). The validity of such experiments was confirmed by Mello et al. (2010), who compared velocity and acceleration records from the 2002 M_w 7.9 Denali, Alaska, earthquake at a station close to the fault with those found in the laboratory experiments. Though these experiments were carried out on man-made material (Homalite) under normal temperature and pressure, they revolutionized our way of thinking.

To mimic more realistic conditions, a group led by Schubnel at the École Normale Supérieure (ENS) in Paris has carried out experiments on rock samples (Westerly granite) and obtained super-shear rupture speeds (Passelègue et al. 2013). The rupture front position was obtained by analysis of acoustic high-frequency recordings on a multistation array. This is clearly very close to the situation in seismology, where the rupture details are obtained by seismogram (time-series) analysis, as discussed earlier.

In the real Earth, earthquake ruptures propagate through material at higher temperatures and pressures than those in these experiments, causing lower crustal, intermediate, and deep earthquakes. The same group at ENS has started looking at this problem in the laboratory (Schubnel et al. 2013) by simulating the nucleation of acoustic emission events that resemble deep earthquakes. These events are caused by an instantaneous phase transition from olivine to spinel, which would occur at the same depth and result in large stress releases observed for some deep earthquakes. Now this group has installed laboratory equipment which can reach pressures of 4GPa and temperatures of 1100 °C. For the larger samples, this would simulate conditions in the Earth approximately at 15–20 km depth. On smaller size samples, this would cover the 20–100 km depth range. Thus, in future we will be able to better simulate ruptures at the depths where earthquakes actually occur, thereby improving our understanding of such earthquakes. From the practical point of view, the idea that only crustal earthquakes cause damage to the built environment is not true. The 1970 M_w 7.8 Peru earthquake occurred at a depth of 45 km (International Seismological Centre Bulletin), and killed more than 66,000 people and completely devastated the city of Huaraz, Peru. It remains till today the worst natural disaster in the

western and southern hemispheres of the Earth. Thus, study of sub-crustal earthquakes also has important societal implications.

Discussion of Results of Inversion

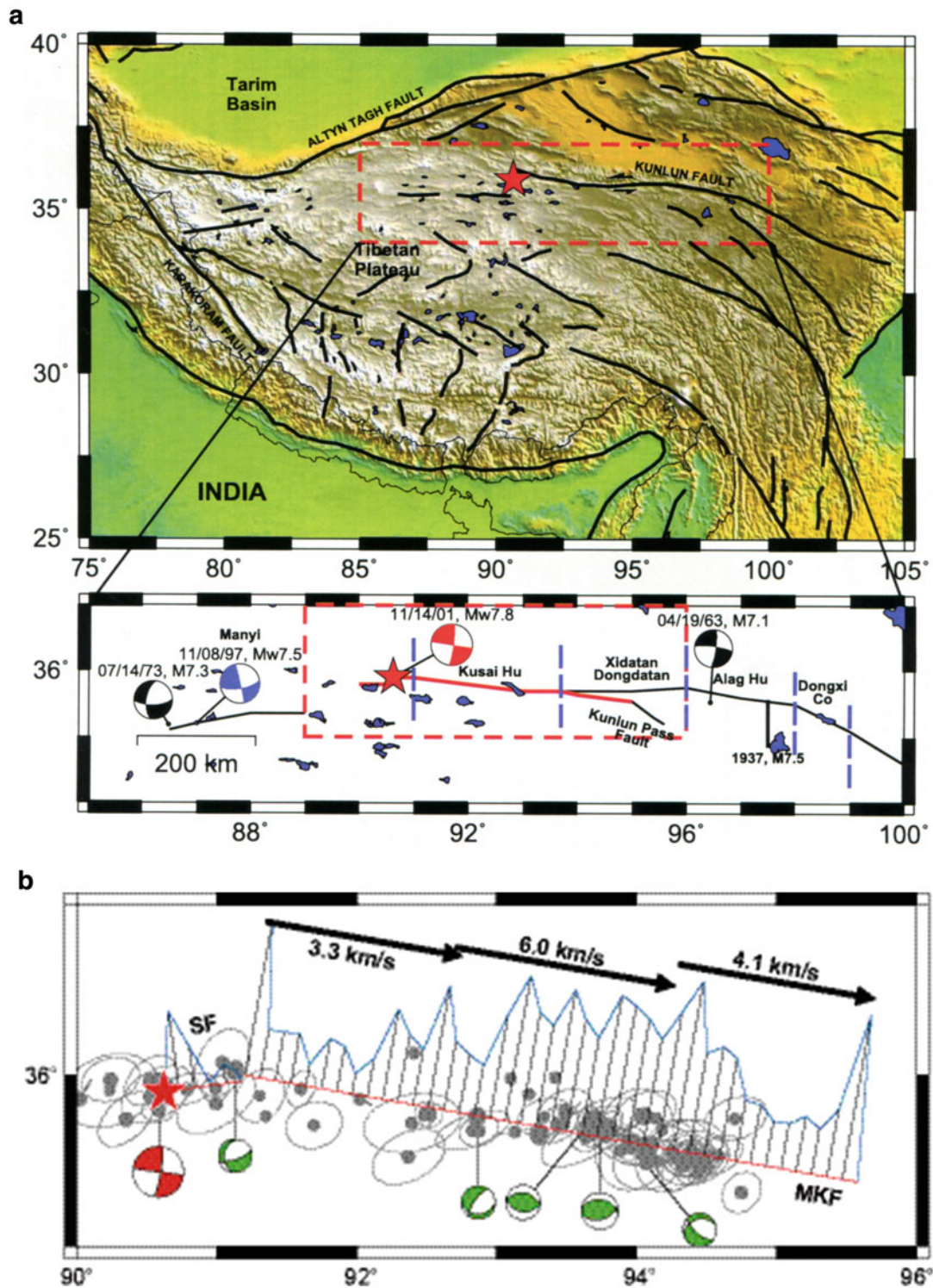
Finally, we discuss some examples of what we have learned regarding the mechanics of the earthquake source from the solution of such inverse problems. The first great earthquake which for which the slip history and distribution was obtained was the great 1985 M_w 8.0 Michoacan, Mexico, earthquake (Mendoza and Hartzell 1988). It was found that the main slip took place in two patches on the fault and that aftershocks occurred in the regions of low fault slip or regions of transition from high to low slip, that is, where stresses were increased due to the earthquake. Some studies since then have been summarized by Das and Henry (2003), and many earthquakes have taught us important, and often unexpected, features of earthquake mechanics.

Super-Shear Wave Rupture Speeds

One very important outcome of the study of the inverse problem of earthquake source mechanics has been the result that actual earthquakes can have rupture speeds exceeding the shear wave speed and can even approach the compressional wave speed of the medium. Robinson et al. (2006a) found this when studying the >400 km long strike-slip 2001 M_w 7.8 Kunlun, Tibet, earthquake, which propagated for nearly 100 km at this very high speed. The fault geometry was shown to be a controlling factor for the changes in rupture speed (Robinson et al. 2006a), with the longest, straight portion of the fault having the fastest rupture speed (Fig. 2). The off-fault regions of the super-shear portions of this earthquake had great damage (Bhat et al. 2007), and this again has obvious important implications for earthquake hazard mitigation. By inspecting all long strike-slip faults in the world, and noting their long straight portions, Robinson et al. (2010) made a list of all faults likely to have sustained super-shear rupture and called them “earthquake fault superhighways.” Das (2015) compiled a list of earthquakes known to rupture at super-shear speeds and discussed its importance for seismic hazard assessment and for earthquake engineering applications. Table 1 gives an updated list of crustal earthquakes known to have reached these speeds, and Table 2 the list of earthquake fault superhighways.

Earthquake Rupture Complexity

Another result that has come from the study of inverse problems is the complexity of earthquake ruptures. Forward modeling had shown that such complexity can exist (see,



Earthquake Rupture: The Inverse Problem, Fig. 2 The 14 November 2001 M_w 7.8 Kunlun, Tibet, earthquake. (a) (top) Tectonic setting and topography. The star indicates the epicenter. (bottom) Segmentation of the Kunlun fault is shown by vertical blue dashed lines, together with the CMT solution (red) obtained by Robinson et al. (2006a). The Harvard CMT solution for the 8 November 1997 Manyi earthquake is shown in

blue and the focal mechanisms of two earlier earthquakes in the region with magnitudes >7 are shown in black. The faulting extent in the 2001 Kunlun earthquake is shown in red. (b) Relocated aftershocks for the 6-month period following the earthquake are shown by gray dots, together with their 90% confidence error ellipses. The two faults on which rupture occurred are marked (in red) as MKF (Main Kunlun

e.g., Das and Kostrov 1988). Detailed analyses of body wave seismograms have now shown that such complexity does exist in real earthquakes (Das and Henry 2003). For example, Henry et al. (2000) showed that the 1998 M_w 8.1 Antarctic Plate earthquake jumped over a barrier of width as large as 70–100 km, with a time delay of about 40s, and then

continued propagating for about another 60 km, this second earthquake itself having a M_w 7.6–7.8 (Fig. 3). Both this complexity and the pattern of aftershocks is controlled by preexisting fracture zones on the ocean floor. The great 2001 M_w 8.4 Peru earthquake (Fig. 4) showed how a barrier on the fault could stall the rupture process (Robinson et al.

Earthquake Rupture: The Inverse Problem, Table 1 Strike-slip earthquakes known to have reached super-shear rupture speeds

Year	M_w	Location	Super-shear segment length	Data used	Reference
1979	6.5	Imperial Valley, California	35 km	Strong ground motion	Archuleta (1984); Spudich and Canswick (1984)
1999	7.6	Izmit, Turkey	45 km	''	Bouchon et al. (2000)
1999	7.2	Duzce, Turkey	40 km	''	Bouchon et al. (2001)
2001	7.8	Kunlun, Tibet	>400 km	Teleseismic	Robinson et al. (2006a)
2002	7.9	Denali, Alaska	340 km	''	Walker and Shearer (2009)
2012	8.6	N. Sumatra	200, 400, 400 km	''	Wang et al. (2012)
2013	7.5	Craig, Alaska	100 km	''	Yue et al. (2013)
2018	7.5	Palu, Indonesia	30 km	''	Socquet et al. (2019)

Earthquake Rupture: The Inverse Problem, Table 2 Earthquake fault superhighways

	Fault system and location	Total (km)	Segment lengths ^a	Affected population (millions) ^b	Size and dates of past earthquakes ^c
1	Red River, Vietnam/China	1000	280, 230, 290	25.7	7.7 (1733), 8.0 (1833)
2	San Andreas, California	1050	160, 230	13.1	7.9 (1857), 7.9 (1906)
3	Sagaing, Burma	1000	700	9.1	N.D.(1839), 7.3 (1930), 7.3 (1930), 7.6 (1931), 7.5 (1936), 7.4 (1946)
4	Great Sumatra	1600	100, 160, 220, 200	6.7	7.7 (1892), 7.6 (1909), 7.5 (1933), 7.4 (1943), 7.6 (1943)
5	Dead Sea, Jordan/Israel	1000	100, 125	5.2	N.D. (1068), N.D. (1170), N.D. (1202)
6	Chaman/Herat,	1100	170, 320, 210	2.5	N.D. (1892)
7	Luzon, Philippines Pakistan/Afghanistan	1600	130	2.1	7.8 (1990)
8	Kunlun, Tibet	1600	270, 130, 180, 100	0.15	7.5 (1937), 7.8 (2001)
9	Altyn Tagh, Tibet	1200	100, 100, 150	0.062	7.6 (1932)
10	Bulnay, Mongolia	300	100, 200	0.020	7.8, 8.2 (1905)
11	Denali, Alaska	1400	130	Negligible	7.8 (2002)

^aLengths of straight segments, identified as superhighways, listed from south to north

^bCurrent population, in millions, within 50 km of the superhighways, this being the region expected to be most damaged by earthquake rupture along the superhighways

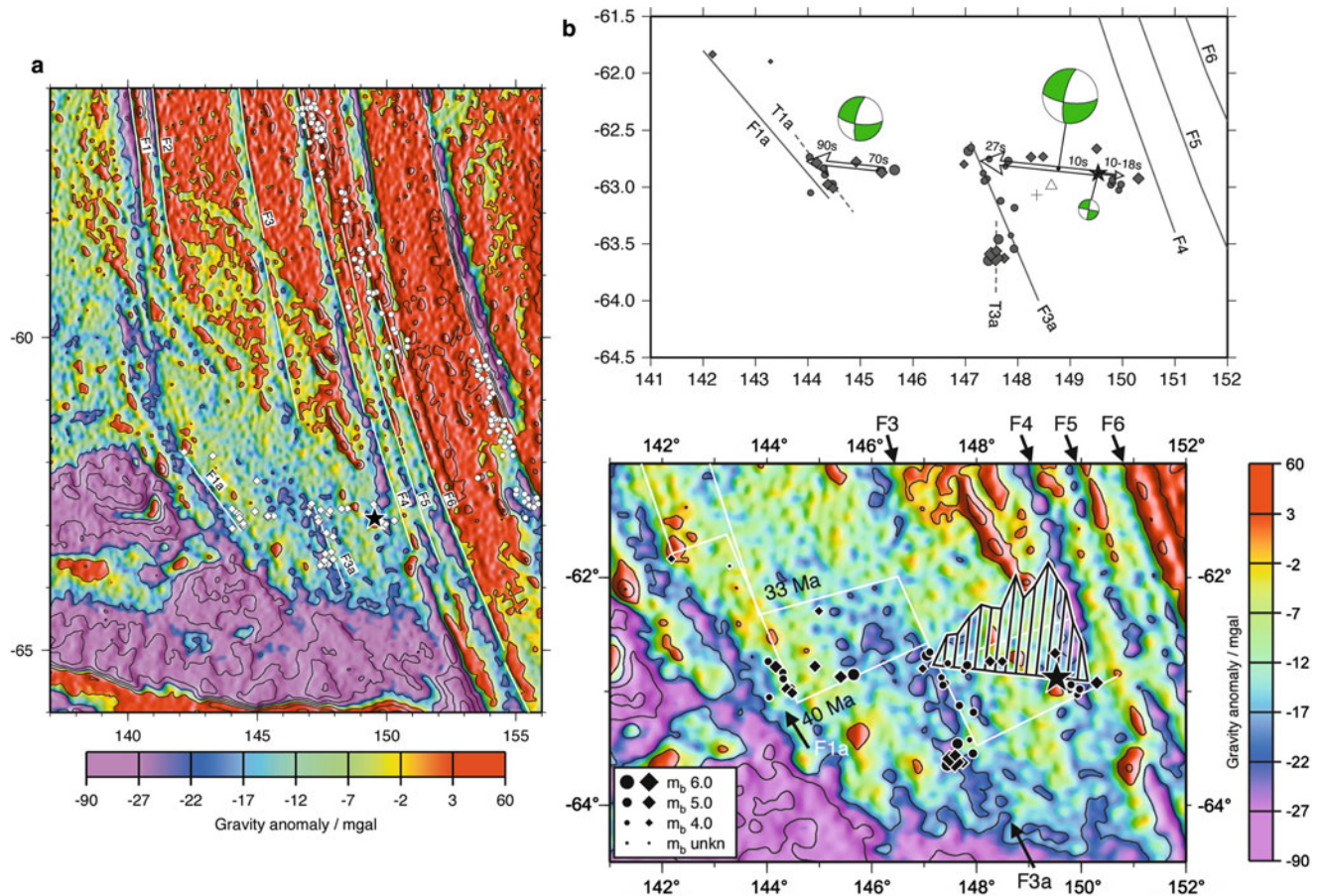
^cMagnitude of old earthquakes are surface wave magnitude or moment-magnitude, as available; N.D. if unknown

Earthquake Rupture: The Inverse Problem, Fig. 2 (continued) Fault) and SF (Secondary Fault). The star shows the rupture initiation region. Available CMT solutions are shown in green. Schematic graphs showing the final slip on the two faults are shown as blue lines, with the

rupture speeds in the three phases of rupture on the MKF indicated at the top. The central portion is where rupture reached the compressional wave speed of ≈ 6 km/s. (Figure based on Robinson et al. (2006a))

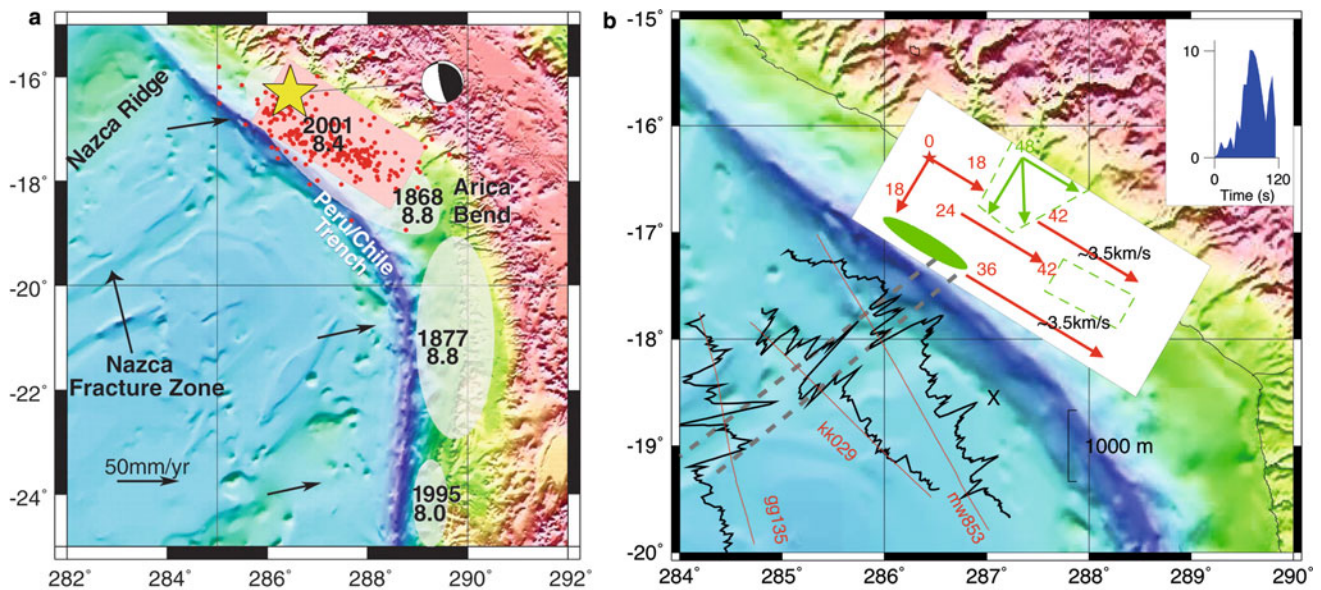
2006b). Based on marine geophysical data from the portion of the Nazca plate that was subducting in this region, this barrier was identified as being due to bathymetry (such as seamounts located on the inner corner of a fracture zone) related to a subducting fracture zone. Robinson and Cheung (2010) found that barriers limited the faulting extent of the 2003 M_w 8.3 Tokachi-Oki, Japan, earthquake and that the

regions of high slip (“asperities”) aligned with other seamounts on the subducting plate about to enter the trench (Fig. 5). The fact that all the earthquakes mentioned in this section are oceanic is simply due to the fact that many such earthquakes have occurred in the last three decades, and does not imply that earthquakes on land do not have complexity!



Earthquake Rupture: The Inverse Problem, Fig. 3 (a) Tectonic setting of the 25 March 1998 M_w 8.1 Antarctic Plate earthquake. Marine gravity anomalies are from an updated version of Sandwell and Smith (1997), illuminated from the east, with contours every 20 mGal. The relocated aftershocks for a one-year period following the earthquake are shown as white diamonds, the main shock epicenter being the black star. Background seismicity for the period January 1, 1964, to July 31, 1997, is shown as white circles. Selected linear gravity features are identified by white lines and are labeled F1–F6. F1a is the southward continuation of F1. (b) (Top) Principal features of the main shock rupture process. Arrows show location and directivity of the two sub-events and are labeled with start and end times of rupture segments, obtained by Henry et al. (2000). Focal mechanisms are shown in green for the initiation, the first sub-event plotted at the centroid location obtained in

the study of Henry et al. (2000), and for the second sub-event. The open triangle shows the Harvard CMT centroid location. Epicenters shown in (a) which fall in the region of this map are shown in gray. Linear gravity features F1a, F3a, and F4–F6 are shown as solid gray lines. T1a and T3a (dashed lines) are the locations of tectonic features inferred by Henry et al. (2000). (Bottom) Final moment distribution for the preferred solution of Henry et al. (2000) shown as black line graph, with the baseline of the graph being the physical location of the fault. The peak moment density is $1.25 \times 10^{19} \text{ N m km}^{-1}$. Regions of the fault with <15% of this maximum value are excluded in this plot. Same gravity anomalies, same linear features, and same epicenters as Fig. 3(a), except that now only earthquakes which are located with the semi-major axis of the 90% epicentral confidence ellipse <20 km are shown. The white lines show known isochrons



Earthquake Rupture: The Inverse Problem, Fig. 4 (a) Tectonic setting of the 23 June 2001 M_w 8.4 Peru earthquake, together with bathymetry constructed from a GEBCO (2003) $1' \times 1'$ grid. The rupture zone of the earthquake is shown as a pink rectangle and the epicenter by a yellow star. Rupture areas, magnitudes, and years of earlier great earthquakes in the region are indicated by ellipses. (b) The rupture process of this earthquake. The red and green lines show the rupture process, the

numbers indicating the seconds after rupture initiation. The green ellipse shows the region where large slip was induced by delayed rupture of the trapezoid-shaped barrier, shown by green dashed lines. A second rectangular barrier to the southeast is shown by green dashed lines. The moment-rate function is shown on the upper right in blue, with its abscissa having units of 10^{19} Nm/s. (Based on Robinson et al. (2006b))

Large Earthquakes on Transform Faults

Another result from the analysis of seismograms by solving the inverse problem is that transform faults which were thought to be relatively simple can actually be quite complex. Das (2019; ► “Earthquakes on Oceanic Transform Faults, Unexpectedly Large”) has discussed the occurrence of unusually large earthquakes on transform faults. Such faults were believed in the past to have few, small earthquakes. Robinson (2011) studied the 2004 M_w 8.1 Tasman Sea earthquake, a surprising example of a great earthquake on a fossil fracture zone, and showed that due to tectonic changes, such fossil fracture zones can be reactivated by a new stress regime to which the former fossil fracture zone is now subjected.

Another interesting example of an earthquake on a transform fault was the 2000 M_w 7.8 Wharton Basin earthquake in the Indian Ocean. The rupture occurred simultaneously on both nodal planes (Robinson et al. 2001), so the transform fault as well as the ridge both ruptured (see ► “Earthquakes on Oceanic Transform Faults, Unexpectedly Large” for more details).

Summary

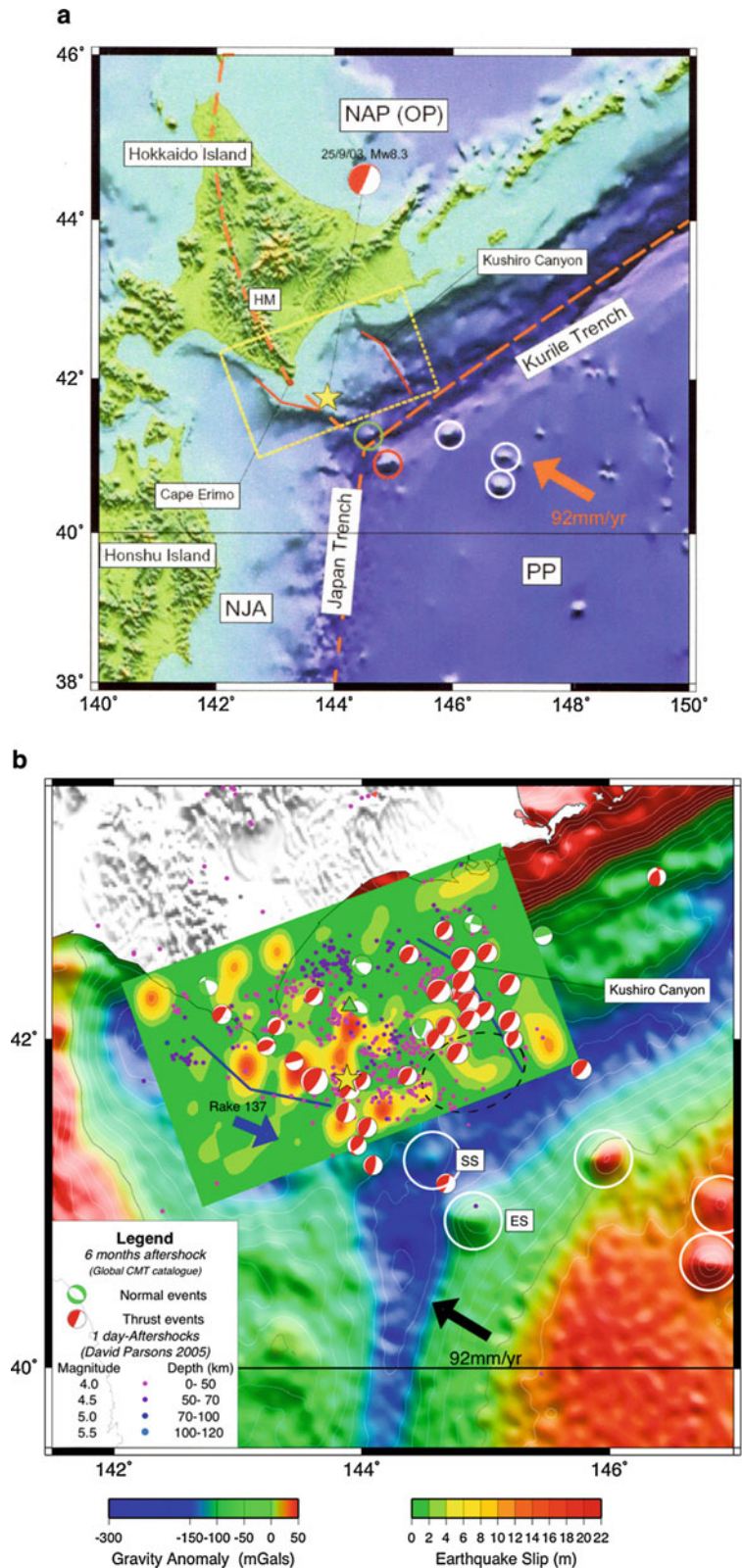
In the last three decades, seismologists have solved the “inverse problem” of using broad band digitally recorded seismograms to infer the details of the dynamics of the earthquake rupturing process, including the rupture speed, the slip-rate, and the slip on the fault. The theory and some examples of particular earthquakes showing rupture complexity are given.

Cross-References

- [Body Waves](#)
- [Earthquake Source Theory](#)
- [Earthquake, Aftershocks](#)
- [Earthquake, Focal Mechanism](#)
- [Great Earthquakes](#)
- [Time-Dependent Seismic Hazard Assessment](#)

Earthquake Rupture: The Inverse Problem, Fig. 5 (a)

Tectonic setting of the 25 September 2003 Mw 8.3 Tokachi-Oki, Japan, earthquake. The sea-floor topography is based on GEBCO $1' \times 1'$ grid. PP – Pacific Plate; NJA – Northeast Japan arc; HM – Hidaka mountains; NAP – North American Plate. The orange arrow shows the rate and direction of subduction of the PP under NAP. Orange dashed lines indicate plate boundaries, and the yellow star the epicenter of this earthquake. The Global CMT solution for this earthquake is shown in red. The modeled rectangular region for the dynamic rupture process is shown by the yellow dashed line. Red lines mark the Kuroshio Canyon and the break in the continental slope offshore of Cape Erimo. The circles mark the seamounts on the Pacific Plate, the red one being the Erimo seamount. (b) Final slip distribution with an overlay of 6 months of aftershocks, relocated by Parsons (2006). (See Robinson and Cheung (2010) for the 24 h distribution.) Seamounts are marked as circles, with SS indicating a subducted seamount and ES the Erimo seamount. The rake of the earthquake is marked in blue. The high slip region of the 1952 earthquake is marked by the black dashed ellipse



Bibliography

- Aki K, Richards PG (1980) Quantitative seismology: theory and methods. W.H. Freeman, New York
- Archuleta R (1984) Faulting model for the 1979 Imperial Valley earthquake. *J Geophys Res* 89:4559–4585
- Backus G, Gilbert F (1967) Numerical applications of a formalism for geophysical inverse problems. *Geophys J R Astron Soc* 13:247–276
- Backus G, Gilbert F (1968) The resolving power of gross earth data. *Geophys J R Astron Soc* 16:169–205
- Bhat HS, Dmowska R, King GCP, Klinger Y, Rice JR (2007) Off-fault damage patterns due to supershear ruptures with application to the 2001 M_w 8.1 Kokoxili (Kunlun) Tibet earthquake. *J Geophys Res* 112:B06301
- Bouchon M, Bouin MP, Karabulut H, Toksöz MN, Dietrich M, Rosakis AJ (2001) How fast is rupture during an earthquake? New insights from the 1999 Turkey earthquakes. *Geophys Res Lett* 28:2723–2726
- Bouchon M, Toksöz MN, Karabulut H, Bouin MP, Dietrich M, Aktar M, Edie M (2000) Seismic imaging of the Izmit rupture inferred from the near-fault recordings. *Geophys Res Lett* 27:3013–3016
- Cohee BP, Beroza GC (1994) Slip distribution of the 1992 Landers earthquake and its implications for earthquake source mechanism. *Bull Seismol Soc Am* 84:692–712
- Das S (2007) The need to study speed. *Science* 317(5840):905–906
- Das S (2015) Chapter 1: Supershear earthquake ruptures – theory, methods, laboratory experiments and fault superhighways. In: Perspectives on European earthquake engineering and seismology, vol 2. Springer, pp 1–20
- Das S (2019) Unusual large earthquakes on oceanic transform faults. In: Bizzarri A, Das S, Petri A (eds) Proceedings of the international school “Enrico Fermi”, course 202. IOS/SIF, Amsterdam/Bologna, pp 33–52. <https://doi.org/10.3254/978-1-61499-979-9-33>
- Das S, Henry C (2003) Spatial relation between main earthquake slip and its aftershock distribution. *Rev Geophys* 41(3):1013. <https://doi.org/10.1029/2002RG000119>
- Das S, Kostrov BV (1988) An investigation of the complexity of the earthquake source time function using dynamic models. *J Geophys Res* 93:8035–8050
- Das S, Kostrov BV (1990) Inversion for seismic slip rate and distribution with stabilizing constraints: application to the 1986 Andreanof Islands earthquake. *J Geophys Res* 95:6899–6913
- Das S, Kostrov BV (1994) Diversity of solutions of the problem of earthquake faulting inversion. Application to SH waves for the great 1989 Macquarie Ridge earthquake. *Phys Earth Planet Inter* 85:293–318
- Das S, Suhadolc P (1996) On the inverse problem for earthquake rupture. The Haskell-type source model. *J Geophys Res* 101:5725–5738
- Das S, Suhadolc P, Kostrov BV (1996) Realistic inversions to obtain gross properties of the earthquake faulting process. Special issue entitled Seismic source parameters: from microearthquakes to large events, ed. C. Trifu. *Tectonophysics* 261:165–177
- GEBCO (2003) The GEBCO digital atlas, centenary edition (CD-ROM). British Oceanographic Data Centre, Liverpool
- Hartzell SH, Heaton TH (1983) Inversion of strong-ground motion and teleseismic waveform data for the fault rupture history of the 1979 Imperial Valley, California earthquake. *Bull Seismol Soc Am* 73:1553–1583
- Hartzell S, Stewart GS, Mendoza C (1991) Comparison of L_1 and L_2 norms in a teleseismic waveform inversion for the slip history of the Loma Prieta, California, earthquake. *Bull Seismol Soc Am* 81:1518–1539
- Henry C, Das S, Woodhouse JH (2000) The great March 25, 1998 Antarctic plate earthquake: moment tensor and rupture history. *J Geophys Res* 105:16,097–16,119
- Kostrov BV (1975) Mechanics of the tectonic earthquake focus, (in Russian). Nauka, Moscow
- Kostrov BV, Das S (1988) Principles of earthquake source mechanics. Applied mathematics and mechanics Series. Cambridge University Press, New York
- Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs
- Mendoza C, Hartzell SH (1988) Inversion for slip distribution using teleseismic P waveforms: North Palm Springs, Borah Peak, and Michoacan earthquakes. *Bull Seismol Soc Am* 78:1092–1111
- Mendoza C, Hartzell SH (1989) Slip distribution of the 19 September 1985, Michoacan, Mexico, earthquake: near-source and teleseismic constraints. *Bull Seismol Soc Am* 79:655–669
- Mello M, Bhat HS, Rosakis AJ, Kanamori H (2010) Identifying the unique ground motion signatures of supershear earthquakes: theory and experiments. *Tectonophysics* 493:297–310
- Olson AH, Apsel RJ (1982) Finite faults and inverse theory with applications to the 1979 Imperial Valley earthquake. *Bull Seismol Soc Am* 72:1969–2001
- Passelègue FX, Schubnel A, Nielsen S, Bhat HS, Madariaga R (2013) From sub-Rayleigh to supershear ruptures during stick-slip experiments on crustal rock. *Science* 340(6137):1208–1211
- Parsons D (2006) Rupture areas of the 1994 M_w 8.3 Shikotan and the 2003 M_w 8.3 Tokachi-Oki, Japa, earthquakes. MEarthSci 4th year extended project. Oxford University, Oxford, UK
- Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1986) Numerical recipes: the art of scientific computing. Cambridge University Press, New York
- Robinson DP (2011) A rare great earthquake on an oceanic fossil fracture zone. *Geophys J Int* 186:1121–1134
- Robinson DP, Cheung LT (2010) Source process of the M_w 8.3 Tokachi-Oki, Japan earthquake and its aftershocks. *Geophys J Int* 181:334–342. <https://doi.org/10.1111/j.1365-246X.2010.04513.x>
- Robinson DP, Brough C, Das S (2006a) The M_w 7.8 2001 Kunlunshan earthquake: extreme rupture speed variability and effect of fault geometry. *J Geophys Res* 111:B08303. <https://doi.org/10.1029/2005JB004137>
- Robinson DP, Das S, Searle MP (2010) Earthquake fault superhighways. *Tectonophysics* 493:236243
- Robinson DP, Das S, Watts AB (2006b) Earthquake rupture stalled by subducting fracture zone. *Science* 312(5777):1203–1205
- Robinson DP, Henry C, Das S, Woodhouse JH (2001) Simultaneous rupture along two conjugate planes of the Wharton Basin earthquake. *Science* 292:1145–1148. (See also: Deplus C (2001) Perspectives: Indian Ocean actively deforms. *Science* 292:1850–1851)
- Rosakis AJ, Samudrala O, Coker D (1999) Cracks faster than the shear wave speed. *Science* 284:1337–1340
- Sandwell DT, Smith WHF (1997) Marine gravity anomaly from Geosat and ERS 1 satellite altimetry. *J Geophys Res* 102:10039–10054
- Sarao A, Das S, Suhadolc P (1998) Effect of non-uniform station coverage on the inversion for seismic moment release history and distribution for a Haskell-type rupture model. *J Seismol* 2:1–25
- Schubnel A, Brunet F, Hilairet N, Gasc J, Wang Y, Green HW (2013) Deep-focus earthquake analogs recorded at high pressure and temperature in the laboratory. *Science* 341(6152):1377–1380. <https://doi.org/10.1126/science.1240206>
- Spudich P, Cranswick E (1984) Direct observation of rupture propagation during the 1979 Imperial Valley earthquake using a short baseline accelerometer array. *Bull Seismol Soc Am* 74:2083–2114
- Socquet A, Hollingsworth J, Pathier E, Bouchon M (2019) Evidence of supershear during the 2018 magnitude 7.5 Palu earthquake from space geodesy. *Nat Geosci* 12:192–199
- Tarantola A (1987) Inverse problem theory. Methods for data fitting and model parameter estimation. Elsevier, New York

- Wald DJ, Heaton TH (1994) Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. *Bull Seismol Soc Am* 84:668–691
- Walker KT, Shearer PM (2009) Illuminating the near-sonic rupture velocities of the intracontinental Kokoxili M_w 7.8 and Denali fault M_w 7.9 strike-slip earthquakes with global P wave back projection imaging. *J Geophys Res* 114:B02304
- Wang D, Mori J, Uchide T (2012) Supershear rupture on multiple faults for the M_w 8.6 off northern Sumatra, Indonesia earthquake. *Geophys Res Lett* 39:L21307
- Wu FT, Thomson KC, Kuenzler H (1972) Stick-slip propagation velocity and seismic source mechanism. *Bull Seismol Soc Am* 62:1621–1628
- Xia K, Rosakis AJ, Kanamori H (2004) Laboratory earthquakes: the sub-Rayleigh-to-supershear transition. *Science* 303:1859–1861
- Xia K, Rosakis AJ, Kanamori H, Rice JR (2005) Laboratory earthquakes along inhomogeneous faults: directionality and supershear. *Science* 308:681–684
- Yue H, Lay T, Freymuller JT, Ding K, Rivera L, Ruppert NA, Koper KD (2013) Supershear rupture of the 5 January 2013 Craig, Alaska (M_w 7.5) earthquake. *J Geophys Res* 118:5903–5919

Earthquake Sounds

Andrew J. Michael
USGS, Moffett Field, CA, USA

Definition

Earthquake sounds. Atmospheric pressure waves associated with earthquakes, especially those audible to humans.

Introduction

“...previous to an earthquake, a roaring is usually heard,” wrote Lucius Annaeus Seneca, around 65 C.E., in the sixth volume of *Naturales Quaestiones* (translation by Clarke and Geike 1910). While Seneca’s explanations for earthquakes now seem fanciful, this statement remains accurate, and historical records of earthquake sounds have been joined over the last few decades by instrumental studies of this phenomenon.

A particularly interesting case of earthquake sounds concerns the long-reported noises near Moodus, Connecticut, whose town name is derived from a Wangunk word meaning “place of noises” (Brigham 1871). Reports of these noises predate the arrival of European settlers who described them as accompanying shaking “as though in an earthquake” (Davis 1897). Seismographic studies of the earthquakes that accompany the Moodus Noises demonstrate that these sounds are caused by swarms of magnitude -2 to 2.1 events that occur at shallow depths (less than 2.4 km) (Ebel 1982). Similarly, explosive sounds similar to cannon fire have been

documented during earthquake swarms in the lower Rhone Valley of France since 1772 and have now been explained by events with ultra-shallow focal depths of only 200 m. With magnitudes as small as -0.7 , many of these events are heard even if the shaking cannot be felt (Thouvenot et al. 2009).

Because human hearing is an effective spectral analyzer, transforming seismograms to make them audible provides a useful educational tool. And because sound is commonly associated with earthquakes, audio can be an effective element of artistic works about earthquakes. This entry reviews the role of earthquake sounds in science, education, and art.

Records of Sounds

People have reported hearing earthquakes for as long as we have earthquake reports. *The Earthquake Catalogue of the British Association* (Mallet and Mallet 1858) lists reports of earthquake sounds back to 122 B.C.E., although the earliest of these are referenced to histories written by Julius Obsequens in the fourth century C.E.

Catalogs of intensity data are good sources of reports on earthquake sounds. Davison (1938) explored a global set of catalogs, including the Mallet and Mallet catalog mentioned above, and the references in his paper provide a guide to these classic sources. Most of his data came from Britain, and he classified about 20,000 descriptions of sounds into 7 classifications that are largely consistent with low rumbling sounds but also include some descriptions of impulsive sounds which are heard more often at short epicentral distances. A more recent intensity catalog with special emphasis on sounds was collected by Pierre Stahl in the French Pyrenees (Sylvander and Mogos 2005) and the US Geological Survey “Did You Feel It?” catalog of Internet-collected intensity data includes information on sounds in its 1.25 million observations (Wald et al. 1999).

A problem with sounds described in intensity catalogs is that they depend on peoples’ ability to describe an ephemeral experience in words. In the 1979 $M4.0$ earthquake near Bath, Maine, sounds were widely heard and mostly described as high-frequency booms (Pulli et al. 1980). Many respondents described the sound in terms of the explosion of a heating system boiler. One exception was a resident who had actually heard a boiler explode and said the earthquake sounded nothing like that. Regardless of the words used to describe the sound, some residents who had previously lived in Southern California explained that they were different than the longer-lasting, low-frequency rumbles they remembered from earthquakes there. Those descriptions are consistent with Hill et al. (1976) who described earthquake sounds heard in Southern California as “ranging from the rumble of distant thunder to the rushing of a wind.” The high-frequency content of the sounds in Maine could be due to high-frequency ground

motions created by a combination of earthquakes at shallow depths, low crustal attenuation, and relatively high stress drops in that region (Pulli et al. 1980). The Moodus area shares similar characteristics, and thus the shaking should also have substantial high-frequency energy. However, many observers there described the noises in 1981 as “distant thunder” (Ebel 1982). Thus, the subjective observations in intensity catalogs can make interpretation difficult. Because of the vague and varied descriptions of these sounds, it can be difficult to differentiate between sounds caused by earthquakes and other sources such as storm waves, tsunamis, natural and accidental gas explosions, booming sands, and meteorites (Hill 2011).

Interpreting earthquake sounds is also complicated when observers are indoors and may largely hear or record sounds created by the building. For instance, the first known recording of earthquake sounds is from 1954 when a recording that was being made inside a two-story wood-frame dwelling in Eureka, California, accidentally captured an M6.6 earthquake (Steinbrugge 1974). The catalog of earthquake sound recordings compiled by Steinbrugge is available from the Seismological Society of America at https://www.seismosoc.org/inside/earthquake_sounds/. All of these recordings mix together natural earthquake sounds with those produced by man-made structures and human reactions to the shaking. A notable entry in this catalog is a 5-minute recording done by Robert Pate during the 1964 M9.2 Alaska earthquake, including his narration of several minutes of felt shaking.

The first purposeful recordings of earthquake sounds away from man-made structures were made in studies of the 1965–1967 Matsushiro swarm in Japan (Japan Meteorological Agency 1968) and the 1975 earthquake swarm near Brawley, California (Hill et al. 1976). Hill et al. recorded the audio on one stereo channel of an audio cassette tape and the seismic sensor on the other channel. Unfortunately, the seismograms recorded by this analog system were clipped, and that made some of their analysis more uncertain. Sylvander et al. (2007) made digital audio and seismic recordings in the Pyrenees that stayed on scale.

How Are Sounds Produced?

The existence of earthquake sounds is hardly surprising given that seismic waves from the solid earth will refract, albeit weakly, into the atmosphere. This was first demonstrated by Ewing, who considered the simple case of a P-wave in a body of water being transmitted into the atmosphere (Ewing 1883).

As described by Seneca and in the observations cataloged by Mallet and Mallet (1858) and Davison (1938), earthquake sounds are often heard slightly before shaking is felt. Because people often feel the stronger, but later arriving S-wave, this

suggests that the sounds are produced by the P-wave. The instrumental studies referenced above used collocated seismometers and microphones to demonstrate that the sounds coincide with the arrival of the P-wave. The coincidence of the sounds with the arrival of the P-wave demonstrates that the onset of the sounds is generated by transmission of the seismic waves into the atmosphere very close to the listener. This is consistent with Davison’s conclusion that the audible sound and the felt shaking travel at the same velocity. Hill et al. (1976) extended the simple analysis done by Ewing and explored the transmission of sound from a realistic shallow crustal structure to the atmosphere, the frequency content of the shaking from the seismic source, and the sensitivity of human hearing. That analysis showed that the higher frequencies of the seismic shaking overlap with the lower range of human hearing (down to 20 Hz) and are transmitted into the atmosphere with sufficient amplitude to be heard. Both Hill et al. (1976) and Sylvander et al. (2007) observed audio frequencies up to 70 Hz.

Thus, what people hear is in the frequency range of 20 to 70 Hz. These low-frequency sound waves have wavelengths much longer than the size of a human head, and therefore people cannot localize the source of such sounds. The description of earthquake sounds as distant is based on association with other low-frequency sounds that actually are distant, despite that the earthquake sounds are generated close to the listener. It is noteworthy that because earthquake sounds are at such low frequencies, humans are at least as sensitive to these sounds as many animals (Hill 1976).

At Brawley, the later and stronger S-wave arrival was not associated with an audio signal, while at Matsushiro and in the Pyrenees, the S-wave was coincident with sound in some cases. The relative rarity of sound associated with the stronger S-wave shaking can be explained by the attenuation of higher seismic, and therefore audible, frequencies in the S-wave (Hill et al. 1976).

Applications of Sounds

Mallet’s classic study of the Great Neapolitan earthquake of 1857 includes many reports of people hearing the earthquake, and he devotes Part III, Chap. VIII, titled “Of the Sounds that Attended the Shock,” to discussing these observations and attempting to interpret them in terms of a finite source model by drawing an analogy to a long line of troops firing their rifles in succession (Mallet 1862).

Davison’s (1938) study of earthquake sounds attempted to relate his seven classifications to the maximum intensity on the Rossi-Forrel scale, the distance from the event, and the region within which the event occurred. He found that earthquakes were heard over the entire felt area for small

earthquakes but only half of the felt area for larger events and that at greater distances the sound was described as being smoother and more monotonous. The smoother and more monotonous sound with increasing distance is consistent with the attenuation of high-frequency shaking at greater distances from the event. Due to this attenuation, low-frequency motion from large earthquakes is felt at distances beyond the point where shaking in the higher-frequency human hearing range is transmitted. Thus, for large earthquakes, the area of felt shaking is larger than the area where the earthquake can be heard.

More recently, Sylvander and Mogos (2005) and Tosi et al. (2012) developed relationships between the locations where sounds are heard, earthquake magnitude, and epicentral distance and considered sound as a factor in human detection of small earthquakes. Souriau (2006) also explored relationships between heard sounds and shaking and points out that sound may be responsible for waking people and thus could contribute to intensity values at the low end of the scales. That builds on Hill et al.'s (1976) observation that the sound can be audible even when the P-wave is not felt. Tosi et al. (2000) examined whether records of sounds could be used to infer focal mechanisms (the orientation of faults and slip during earthquakes). Partially because the sound observations do not include polarity information, they concluded that the uncertainties were too large to use such data to infer unknown focal mechanisms.

Despite these efforts, it is not clear that earthquake sounds make a significant contribution to the analysis of pre-instrumental earthquakes beyond helping us understand how such sounds contribute to observations of lower intensities. As noted above, there are problems with interpreting subjective reports of sounds that mix together both the natural transmission of shaking into the atmosphere and the shaking of man-made objects. These objections could be removed for future earthquakes by making audio recordings which could easily be integrated into existing digital seismic acquisition systems (Sylvander et al. 2007). Unfortunately, audio recordings may not be a very useful addition to the collocated seismograms because little of the seismic energy is transmitted into the atmosphere and because audio captures only a scalar recording of the three-component ground motion.

While microphones may not be an effective way to record seismic data, seismometers can produce useful records of acoustic waves in the atmosphere that are strong enough to shake the ground. For instance, seismic records of acoustic waves have contributed to forensic seismology investigations into the 1995 Oklahoma City bombing (Holzer et al. 1996), the 1998 truck bombing of the American embassy in Nairobi (Koper et al. 1999), a chemical plant explosion in Nevada (Ichinose et al. 1999), and a pipeline explosion in New Mexico (Koper et al. 2003).

Infrasound

Very-low-frequency acoustic waves known as infrasound can travel great distances through the atmosphere, although with frequencies below 20 Hz, these waves are not audible to humans. Bolt (1964) demonstrated that micropressure recordings after the 1964 Mw 9.2 Alaska earthquake included acoustic waves with periods of 20 to 40 s (frequencies of 0.025 to 0.050 Hz) generated near the earthquake and recorded 3100 km away at Berkeley, California. Donn and Posmentier (1964) further explored these recordings and showed that earlier infrasound arrivals were generated near the recording station by the passing Rayleigh waves. Away from the earthquake, infrasound can also be generated when surface waves travel through mountainous regions. A good review of these mechanisms is Mutschlecner and Whitaker (2005) and the references therein. More recently, Le Pichon et al. (2006) suggested that infrasound could be used to study the amplification of ground displacement by topography, especially in regions with inadequate seismic networks, and Green et al. (2009) demonstrated that coastal cliffs could also be a source of infrasound. An unusual observation is infrasound associated with the tsunami generated by the M9.0 Sumatra earthquake of 2004 (Le Pichon et al. 2005). While infrasound arrays have recorded many earthquakes, their primary applications continue to be the study of phenomena closer to the Earth's surface or in the atmosphere such as volcanoes (Fee and Matoza 2013; Johnson and Ripepe 2011), meteorites (Evers and Haak 2001), and the study of explosions whether chemical or nuclear tests.

Sonification, Education, and Art

Sonification is the process of producing sound from non-auditory data. Most seismic data contains frequencies below human hearing and the simplest way to sonify seismograms is to speed them up, which increases the frequency content, and play them back through an audio system. Benioff mentions doing this with seismograms recorded at slow speed on audio tape (Anonymous 1951), and in 1953 he and Cook Laboratories (Stamford, Connecticut) used this method to produce an LP record with one side called "Earthquakes of this World" (Benioff 1953). Many seismologists have followed in Benioff's footsteps, and this has become particularly easy now that digital seismograms can be converted to audio files using common software such as MatLab. The SeisSound product at the Incorporated Research Institutions for Seismology provides sonified seismograms for many earthquakes (Kilb et al. 2012) and is available online at <https://ds.iris.edu/ds/products/seissound/>.

The human ear is an excellent spectral analyzer, and most people are more used to listening to music than looking at

seismograms. Therefore, sonification of seismograms can be an effective educational tool. Michael (1997) (<https://earthquake.usgs.gov/education/listen/> or <https://doi.org/10.5066/P90ILS3Z>) developed a set of audio playbacks of seismograms that explores how the frequency of shaking varies with source dimension, epicentral distance, and site geology. This web site includes a recording of the 1992 M7.3 Landers earthquake recorded at Mammoth Lakes where local earthquakes were triggered by the passing seismic waves. The local earthquakes are easily heard in the raw sonified seismogram but required filtering to be detected visually in the plotted seismograms (Hill et al. 1993). This demonstrates that audio can be a powerful educational tool.

Sound is a common part of experiencing earthquakes, and so naturally it is part of artistic explorations of seismic events. The seismograms from the above site were used as the basis for my Earthquake Quartet #1 for Trombone, Cello, Voice, and Seismograms (Michael 2000) which is available online at <https://earthquake.usgs.gov/education/listen/music/> or <https://doi.org/10.5066/P9LO2UU2>. Earthquake Quartet #1 starts by describing the earthquake cycle and then explores the idea that society and culture, including music, take place with the earthquakes as an often-ignored backdrop. The earthquakes play the role of percussion because they have abrupt beginnings and broad, non-harmonic, spectra.

Another musical approach was taken by Loos and Scherbaum (1999) in a CD titled “Inner Earth, a seismosonic symphony” which used many different methods of shifting, converting, and transposing seismograms to make them audible. They produced music with harmonic properties that is purely based on seismograms but sounds nothing like the ones that have simply been sped up.

Overlaying a seismogram on a musical staff provides a form of graphical musical notation that can then be played back using synthesized instruments. Some seismograms from the Mt. Etna volcano can be heard as piano music at <http://grid.ct.infn.it/etnasound/page4/page8/page8.html>. The goal of that project was to provide a way to detect changes in the behavior of the volcano by hearing changes in the sonified seismograms (<http://grid.ct.infn.it/etnasound/>), and it also extended into the arts.

The “Great California ShakeOut” earthquake preparedness drills (Jones et al. 2008) use audio combining synthesized earthquake sounds with sounds of damage occurring and sirens to simulate an earthquake experience at the start of the drills. These files can be downloaded from <http://www.shakeout.org/drill/broadcast/index.html>.

“Memento Mori: an internet based earthwork” by Goldberg et al. (1999) used live seismic data to trigger non-seismic audio in a dark installation that also included a visual display of the seismic particle motion. This approach was later used for “Ballet Mori: a ballet conducted by the earth” performed by Muriel Maffre at the San Francisco Ballet to commemorate

the centennial of the 1906 earthquake (Goldberg et al. 2006). Video of the dance can be seen at <http://goldberg.berkeley.edu/art/Ballet-Mori/>.

In 2009, the Parkfield Interventional Earthquake Fieldwork (PIEQF) (Rogers 2010), and online at <http://pieqf.allshookup.org/>, consisted of a large shake table whose motion was triggered by earthquakes throughout the state of California. Both the table itself and a set of rods mounted on the table created noise when an earthquake occurred. This mechanical amplification of the shaking replaced the natural “roaring” discussed by Seneca with a clatter that allowed the viewers to hear distant seismic activity in a way that is not possible through natural shaking and sounds.

Summary

Earthquake sounds form an important part of the earthquake experience and thus are a large part of artistic explorations of these events, but their study forms a small niche within seismology. The last few decades have produced the data necessary to finally understand this phenomenon along with its role in the study of earthquakes and explosions.

Cross-References

- ▶ [Earthquake, Focal Mechanism](#)
- ▶ [Earthquakes, Intensity](#)
- ▶ [Propagation of Elastic Waves: Fundamentals](#)
- ▶ [Seismic Instrumentation](#)
- ▶ [Seismic Signals in Well Water Observations](#)
- ▶ [Seismic, Waveform Modeling and Tomography](#)
- ▶ [Seismology, Monitoring of CTBT](#)
- ▶ [Surface Waves](#)

Acknowledgments This article greatly benefited from reviews by Madison Goldberg, Jeanne Hardebeck, David Hill, Carol Mark, Stephanie Ross, and Matthieu Sylvander.

Bibliography

- Anonymous (1951) Earthquakes – recorded on tape, engineering and science monthly, California Institute of Technology, Nov 1951
- Benioff H (1953) Earthquakes around the world. In: *Out of this world*. Cook Laboratories, Stamford
- Bolt BA (1964) Seismic air waves from the great 1964 Alaskan earthquake. *Nature* 202:1095–1096. <https://doi.org/10.1038/2021095a0>
- Brigham WT (1871) Volcanic manifestations in New England: being an enumeration of the principal earthquakes from 1638 to 1869. In: *Memoirs of the Boston society of natural history*. Boston Society of Natural History, Boston, pp 1–28
- Clarke J, Geike A (1910) Physical science in the time of Nero being a translation of the Quaestiones Naturales of Seneca. MacMillan, London, p 368

- Davis WM (1897) Current notes on physiography. *Science* 6:834–835
- Davison C (1938) Earthquake sounds. *Bull Seismol Soc Am* 28:147–161
- Donn WL, Posmentier ES (1964) Ground-coupled air waves from the great Alaskan earthquake. *J Geophys Res* 69:5357–5361
- Ebel JE (1982) The 1981 microearthquake swarm near Moodus, Connecticut. *Geophys Res Lett* 9:397–400
- Evers LG, Haak HW (2001) Listening to sounds from an exploding meteor and oceanic waves. *Geophys Res Lett* 28:41–44. <https://doi.org/10.1029/2000gl011859>
- Ewing JA (1883) Earthquake measurement. *Tokio Daigaku*, Tokyo, p 125
- Fee D, Matoza RS (2013) An overview of volcano infrasound: from hawaiian to plinian, local to global. *J Volcanol Geotherm Res* 249:123–139. <https://doi.org/10.1016/j.jvolgeores.2012.09.002>
- Goldberg K, Packer R, Kunh G, Matusik W (1999) Memento Mori: an internet-based earthwork, Berkeley. <http://goldberg.berkeley.edu/art/mori/>
- Goldberg K, Maffre M, Packer R, Possokhov Y (2006) Ballet Mori: a ballet conducted by the earth, San Francisco. <http://goldberg.berkeley.edu/art/Ballet-Mori/>
- Green DN, Guilbert J, Le Pichon A, Sebe O, Bowers D (2009) Modelling ground-to-air coupling of the shallow ml 4.3 Folkestone, United Kingdom, earthquake of 28 April 2007. *Bull Seismol Soc Am* 99:2541–2551. <https://doi.org/10.1785/0120080236>
- Hill DP (1976) Physical constraints on sounds generated by very small earthquakes, U.S. Geological Survey open file report, p 27
- Hill DP (2011) What is that mysterious booming sound? *Seismol Res Lett* 82:619–622. <https://doi.org/10.1785/gssrl.82.5.619>
- Hill DP, Fischer FG, Lahr KM, Coakley JM (1976) Earthquake sounds generated by body-wave ground motion. *Bull Seismol Soc Am* 66:1159–1172
- Hill DP, Reasenberg PA, Michael AJ, Arabasz WJ, Beroza G, Brumbaugh D, Brune JN, Castro R, Davis S, dePollo D, Ellsworth WL, Gombert J, Harmsen SC, House L, Jackson SM, Johnston MJS, Jones L, Keller R, Malone S, Munguia L, Nava S, Pechmann JC, Sanford A, Simpson RW, Smith RB, Stark M, Stickney M, Vidal A, Walter S, Wong V, Zollweg J (1993) Seismicity remotely triggered by the magnitude 7.3 landers, California, earthquake. *Science* 260:1617–1623
- Holzer TL, Fletcher JB, Fuis GS, Ryberg T, Brocher TM, Dietel CM (1996) Seismograms offer insight into Oklahoma city bombing. *Eos* 77(393):398–399
- Ichinose GA, Smith KD, Anderson JG (1999) Seismic analysis of the 7 January 1998 chemical plant explosion at Kean canyon, Nevada. *Bull Seismol Soc Am* 89:938–945
- Japan Meteorological Agency (1968) Report on the Matsushiro earthquake swarm, Aug 1965–Dec 1967, technical report of the Japan meteorological agency, pp 125–137
- Johnson JB, Ripepe M (2011) Volcano infrasound: a review. *J Volcanol Geotherm Res* 206:61–69. <https://doi.org/10.1016/j.jvolgeores.2011.06.006>
- Jones LM, Bernknopf R, Cox D, Goltz J, Hudnut KW, Mileti D, Perry S, Ponti D, Porter K, Reichle M, Seligson H, Shoaf K, Treiman JA, Wein A (2008) The ShakeOut scenario, U.S. Geological Survey open-file report 2008–1150, p 312
- Kilb D, Peng Z, Simpson D, Michael A, Fisher M, Rohrlick D (2012) Listen, watch, learn: SeisSound video products. *Seismol Res Lett* 83:281–286. <https://doi.org/10.1785/gssrl.83.2.281>
- Koper KD, Wallace TC, Holnack D (1999) Seismic analysis of the 7 August 1998 truck-bomb blast and the American embassy in Nairobi, Kenya. *Seismol Res Lett* 70:512–521
- Koper KD, Wallace TC, Aster RC (2003) Seismic recordings of the Carlsbad, New Mexico, pipeline explosion of 19 August 2000. *Bull Seismol Soc Am* 93:1427–1432
- Le Pichon A, Mialle P, Gulbert J, Vergoz J (2006) Multistation infrasonic observations of the Chilean earthquake of 2005 June 13. *Geophys J Int* 167:838–844. <https://doi.org/10.1111/j.1365-246X.2006.03190.x>
- Loos W, Scherbaum F (1999) Inner earth, a seismosonic symphony. Trautman, Berlin
- Mallet CE (1862) Great Neapolitan earthquake of 1857, vol 2. Chapman and Hall, London, p 399
- Mallet CE, Mallet JW (1858) The Earthquake catalogue of the British association with the discussion, curves, and maps, Etc. The British Association for the Advancement of Science, London, p 768
- Michael AJ (1997) Listening to earthquakes, USGS. <https://earthquake.usgs.gov/education/listen/> or [10.5066/P90ILS3Z](https://doi.org/10.5066/P90ILS3Z)
- Michael AJ (2000) Earthquake Quartet #1, USGS <https://earthquake.usgs.gov/education/listen/music/> or <https://doi.org/10.5066/P9LO2U2>
- Mutschlecner JP, Whitaker RW (2005) Infrasound from earthquakes. *J Geophys Res* 110:D01108. <https://doi.org/10.1029/2004JD005067>
- Le Pichon A, Herry P, Mialle P, Vergoz J, Brachet N, Garcés M, Drob D, Ceranna L (2005) Infrasound associated with 2004–2005 large Sumatra earthquakes and tsunamis. *Geophys Res Lett* 32:L19802. <https://doi.org/10.1029/2005gl023893>
- Pulli JJ, Stewart RR, Johnston JC, Tubman KM, Michael AJ (1980) Field investigation and fault plane solution of the Bath, Maine, earthquake of April 18, 1979. *Earthq Notes* 51:39–46
- Rogers DV (2010) Parkfield interventional EQ fieldwork (PIEQF), masters thesis, University of New South Wales, Sydney
- Souriau A (2006) Quantifying felt events: a joint analysis of intensities, accelerations and dominant frequencies. *J Seismol* 10:23–38. <https://doi.org/10.1007/s10950-006-2843-1>
- Steinbrugge KV (1974) A catalog of earthquake related sounds. *Bull Seismol Soc Am* 64:1409–1418. Sound files at https://www.seismosoc.org/inside/earthquake_sounds/
- Sylvander M, Mogos DG (2005) The sounds of small earthquakes: quantitative results from a study of regional macroseismic bulletins. *Bull Seismol Soc Am* 95:1510–1515
- Sylvander M, Ponsolles C, Benahmed S, Fels JF (2007) Seismoacoustic recordings of small earthquakes in the Pyrenees: experimental results. *Bull Seismol Soc Am* 97:294–304
- Thouvenot F, Jenatton L, Gratier J-P (2009) 200-m-deep earthquake swarm in Tricastin (lower Rhodane valley, France) accounts for noisy seismicity over past centuries. *Terra Nova* 21:203–210. <https://doi.org/10.1111/j.1365-3121.2009.00875.x>
- Tosi P, De Rubeis V, Tertulliani A, Gasparini C (2000) Spatial patterns of earthquake sounds and seismic source geometry. *Geophys Res Lett* 27:2749–2752
- Tosi P, Sbarra P, De Rubeis V (2012) Earthquake sound perception. *Geophys Res Lett* 39:L24301. <https://doi.org/10.1029/2012gl054382>
- Wald DJ, Quitoriano V, Dengler L, Dewey J (1999) Utilization of the internet for rapid community intensity maps. *Seismol Res Lett* 70:680–697

Earthquake Source Theory

Raul Madariaga

Ecole Normale Supérieure, Université PSL, Paris, France

Definition

Seismic moment

A measure of the size of an earthquake. In the simplest situation, it represents the moment of the couples of forces that make up a dipolar source. In more general cases, it is a 3 by

	3 symmetric tensor of elementary force couples. For planar faults it is the product of rigidity by slip times the source area.
Seismic radiation	The seismic waves emitted by a seismic source. For point sources these are spherical P and S waves emitted by the point source.
Seismic spectrum	Fourier transform of the seismic waves in the far field from the source. In general, when it is not noted otherwise, it refers to the amplitude of the Fourier transform of the far field displacement.
Kinematic earthquake model	A model of slip history of rupture on a fault determined from near and far field observations. In kinematic models the propagation of rupture is arbitrarily specified.
Dynamic earthquake model	A model of rupture propagation on an earthquake fault based on fracture mechanics.

Introduction

Earthquake source theory studies the dynamics of seismic sources in the Earth. The simplest possible model of a seismic source is that of a point source buried in an elastic medium. A proper understanding of the exact nature of the seismic source took more than 50 years since the first efforts by Nakano and colleagues in the 1920s in Japan. Earthquakes were initially modeled as simple explosions, then as point forces, and, finally, as the result of fast transformational strains inside a sphere. In the early 1950s, it was recognized that P waves radiated by earthquakes presented a spatial amplitude distribution similar to the waves produced by single couples of forces, but it was very soon recognized that this type of source could not explain S wave radiation. The next level of complexity was to introduce a double-couple source, a source without resultant force or moment. The physical origin of the double-couple model was established in the early 1960s. Soon after that finite source models derived from dislocation theory were introduced as a simple way to model observations of seismic radiation. Dislocation models present some severe mechanical problems that we will discuss later. They were soon replaced by more mechanically sound models based on dynamic shear cracks. These are the two kinds of models that are currently being used to model earthquakes over a broad range of frequencies and length scales.

Point Source Models

The radiation from a point force or Green's function is the fundamental solution of the equation of elastodynamics. With few rare exceptions, seismic sources are due to fast

internal deformation in the Earth, for instance, faulting or fast phase changes on localized volumes. For a seismic source to be of internal origin, it has to have zero net force and zero net moment. It is not difficult to imagine seismic sources that satisfy these two conditions. The simplest such source is an explosion or implosion, but they do not produce shear waves, at least in uniform media. The next group of possible sources is dipoles, quadrupoles, or combinations thereof. Experimental observation has shown that simple dipoles are inadequate models of seismic sources. It is possible however to combine three orthogonal linear dipoles in order to form a general seismic source. These three dipoles represent the principal directions of a symmetric tensor of rank 2 that we call the seismic moment tensor:

$$\mathbf{M} = \begin{bmatrix} M_{xx} & M_{xy} & M_{xz} \\ M_{xy} & M_{yy} & M_{yz} \\ M_{xz} & M_{yz} & M_{zz} \end{bmatrix} \quad (1)$$

This moment tensor has a structure that is identical to that of a stress tensor, but it is not of elastic origin. The off-diagonal elements of the moment represent the torque produced by two point forces of direction i separated by an infinitesimal distance in direction j . The diagonal terms represent linear dipoles, two opposite forces separated by a very small distance.

Radiation produced by a point moment tensor source located at point $\mathbf{r}_0$:

$$\mathbf{M}(\mathbf{r}, t) = \mathbf{M}_0(t) \delta(\mathbf{r} - \mathbf{r}_0) \quad (2)$$

form a set of Green's functions. In a homogeneous, linear elastic medium, these Green's functions are well-known and contain both near-field and far-field terms. The most important are the far-field terms, the radiation observed at large distance from the source. These are:

$$u_i^P(\mathbf{r}, t) = \frac{1}{4\pi\rho\alpha^3} \frac{1}{R} \sum_{jk} R_{ijk}^P(\theta, \varphi) \dot{M}_{jk}(t - R/\alpha) \quad (3a)$$

for P waves and

$$u_i^S(\mathbf{r}, t) = \frac{1}{4\pi\rho\beta^3} \frac{1}{r} \sum_{jk} R_{ijk}^S(\theta, \varphi) \dot{M}_{jk}(t - R/\beta) \quad (3b)$$

for S waves. Here R is the distance from the source to the observer. R^P and R^S are the radiation patterns of P and S waves, respectively; and ρ , α and β denote the density and the P and S wave speeds of the elastic medium. The radiation patterns are different for every element of the moment tensor with its own set of symmetries and nodal planes. We observe

that the far-field signal carried by both P and S waves is the time derivative of the seismic moment components, so that far-field seismic waves are proportional to the moment rate of the source.

Very often in seismology, it is assumed that the geometry of the source can be separated from its time variation, so that the moment tensor can be written in the simpler form:

$$\mathbf{M}_0(t) = \mathbf{M}_0 \Omega(t) \quad (4)$$

where $\mathbf{M}_0$ is a time-invariant tensor that describes the geometry of the source and $\Omega(t)$ is the time variation of the moment, the source time function determined by seismologists, often called also the scalar seismic moment $M_0(t)$. In this case we can now write a simpler form of (3):

$$u_i^c(\mathbf{r}, t) = \frac{1}{4\pi\rho c^3} \frac{1}{R} R^c(\theta, \varphi) \Omega(t - R/c) \quad (5)$$

where c stands for either the P or the S wave velocity. For P waves u_c is the radial component; for S waves it is the appropriate transverse component for SH or SV waves. We observe that the signal emitted by the source in the far field is the derivative of the scalar moment time function $\Omega(t)$. The term $R^c(\theta, \phi)$ is the radiation pattern, a function of the takeoff angle of the ray at the source.

With small modifications designed to take into account smooth variations of elastic wave speeds in the earth, these expressions are widely used to generate synthetic seismograms in the far field. The main changes that are needed are the use of travel time instead of R/c in the waveform (5) and a more accurate geometrical spreading to replace the $1/R$ term. The complete expressions are given by Aki and Richards (2002) and Udias et al. (2014). In most work with local earthquakes, the approximation (5) is frequently used with a simple correction for free surface response.

What does a seismic moment represent? A number of mechanical interpretations are possible. In the previous sections, we introduced it as a simple mechanical model of double couples and linear dipoles. Other authors (Backus and Mulcahy 1976) have explained them in terms of the distribution of inelastic stresses (sometimes called stress “glut”). Let us first notice that a very general distribution of force of internal origin must necessarily derive from a symmetrical seismic moment density of the form:

$$\mathbf{f}(\mathbf{r}, t) = \nabla \cdot \mathbf{M}_0(\mathbf{r}, t) \quad (6)$$

where $\mathbf{M}_0$ is the moment tensor density per unit volume. It is not difficult to prove that such a force distribution has neither net force nor torque. The moment tensor can be understood as stress tensor generated by internal forces as

proposed by Eshelby (1957). From the moment tensor, one can define the associated inelastic strain by:

$$\mathbf{M}_0 = \lambda \theta_I \mathbf{I} + 2\mu \boldsymbol{\varepsilon}_I \quad (7)$$

where $\mathbf{I}$ is the identity matrix in 3D, λ and μ are the elastic constants, and $\boldsymbol{\varepsilon}_I$ represents the inelastic strain at the source. This is the strain that would occur in the body in the absence of elastic forces. We can now define $\boldsymbol{\varepsilon}_I$ for a fault segment. Let the source be reduced to a two-dimensional fault element of area dS and thickness h . Then the inelastic strain that occurs across the fault element dS is approximately $\frac{1}{2} D/h$ the ratio of the displacement discontinuity or slip D across the fault by the thickness of the fault. Thus $\varepsilon_I h \rightarrow D/2$ as h decreases to zero fault width. For a planar moment distribution, we get a two-dimensional seismic moment distribution:

$$dm_0 = \mu D dS \quad (8)$$

That is, the moment density of a shear fault is just the product of the elastic constant by slip, a classical result obtained by Burridge and Knopoff (1964) using a different argument based on the representation theorem.

Seismic radiation in the spectral domain. In seismically active areas, the waves radiated by earthquakes become quite complex because of multipathing, scattering, attenuation, site response, etc., so that seismograms resemble the source time function $\Omega(t)$ only at long periods. Complexities in wave propagation affect much less the spectral amplitudes in the Fourier-transformed domain. Radiation from a simple point moment tensor source can be obtained from Eq. (5) by straightforward Fourier transformation:

$$u_i^c(\mathbf{r}, \omega) = \frac{1}{4\pi\rho c^3} \frac{1}{R} R^c(\theta, \varphi) \tilde{\Omega}(\omega) \exp\left(-\frac{i\omega R}{c}\right) \quad (9)$$

where $\tilde{\Omega}(\omega)$ is the Fourier transform of the source time function $\Omega(t)$.

From the scaling of moment with earthquake size, Aki (1967) and, from spectral observations, Brune (1970) concluded that the seismic spectra decayed as ω^{-2} at high frequencies. A simple model of the spectral amplitude is:

$$\Omega(\omega) = \frac{M_0}{1 + \omega^2/\omega_0^2} \quad (10)$$

where ω_0 is the corner frequency of the earthquake, the intersection between the low- and high-frequency asymptotes of expression (10). This is the so-called omega-squared model of seismic radiation. In this model seismic sources are characterized by only two independent scalar parameters: the seismic moment and the corner frequency.

As mentioned earlier, not all earthquakes have displacement spectra as simple as this, but the omega-squared model is a good starting point for understanding seismic radiation.

Finite Source Models

Earthquakes are due to the fast propagation of slip on a fault, or fault system, at speeds of the order of that of shear waves. There is evidence that a few earthquakes have propagated at speeds faster than that of shear waves, but these are exceptional events. For most others, the speed of rupture propagation is close to that of Rayleigh waves on the fault. Traditionally, finite source earthquake models are either kinematic or dynamic. In the first class, the distribution of seismic moment and rupture time are specified in order to satisfy certain observational constraints on the source. The most common way is to specify the rupture process based on simple models of rupture propagation, for instance, at constant speed, and then determine the distribution of moment (slip) on the fault from the waveforms observed in different seismic stations. As the data has improved, kinematic inversion methods have become increasingly sophisticated, but the basic problem of how does a rupture grow cannot be determined from basic mechanical principles. The alternative is to use dynamic models where rupture is computed numerically from the specification of the stresses that prevail on the fault and the friction law that acts on the fault surface.

Kinematic Models

One of the most widely used kinematic models was introduced by Haskell (1964). In this model a uniform displacement discontinuity spreads at constant rupture velocity inside a rectangular fault. At low frequencies, for wavelengths longer than the size of the fault, this model is a reasonable approximation to the rupture of a strike slip fault. In Haskell's model at time $t = 0$, a line dislocation of width W appears suddenly and propagates along the fault length at a constant rupture velocity v until a region of length L of the fault has been broken. As the dislocation moves, it leaves behind a zone of constant slip D . Assuming that the fault lies on a plane of coordinates (x, y) , the slip function can be written as:

$$D(x, y, t) = D s(t - x/v) \text{ for } -W < y < W \text{ and } 0 < x < L \quad (11)$$

where $s(t)$ is the source time function that, in the simplest version of Haskell's model, is invariant with position on the fault. The most important feature of this model is the propagation of rupture implicit in the time delay of rupture

x/v , where v is the rupture velocity. An obvious unphysical feature of this model is that rupture appears instantaneously in the y -direction; the other is that near the borders of the fault, slip suddenly jumps from D to zero. This violates material continuity and produces very strong unphysical stress concentrations. In spite of these two obvious shortcomings, Haskell's model gives a simple, first-order approximation to seismic slip, fault finiteness, and finite rupture speed. The seismic moment of Haskell's model is just $M_0 = \mu DLW$. This model has been extensively used to invert for seismic source parameters both in the near and far field from seismic and geodetic data. Some of the limitations of Haskell's model can be removed by admitting that slip is variable on the fault, so that D in Eq. (11) is no longer assumed to be constant but a function of position on the fault. Slip can be tapered as suggested by dynamic models of rupture. Rupture speed can also be made variable.

Dynamical Fault Models

The most frequently used source models are dynamic models, based on some basic concepts of fracture mechanics introduced by seismologists in the middle 1960s. The simplest of these dynamic models is a circular shear crack. This model was introduced by several authors in order to relate fault slip to stress drop inside a fault. A natural approach to model earthquakes was to use well-known results from fracture mechanics. Among dynamic models, we consider here a circular shear crack, assuming that rupture starts from a point and then propagates self-similarly until it finally stops at a certain source radius. This model was carefully studied in the 1970s, and a complete understanding of it is available without the need of getting into the details of the numerical modeling of earthquake sources.

Kostrov's self-similar circular crack. The simplest model of how an earthquake starts is a circular-shaped rupture front that starts from a point and spreads at constant rupture speed v without ever stopping. Slip on this fault is driven by stress drop inside the fault. The solution for slip across the fault, found by Kostrov in 1964, is surprisingly simple:

$$D(r, t) = C(v) \frac{\Delta\sigma}{\mu} \sqrt{v^2 t^2 - r^2} \quad (12)$$

where r is the radius of the fault measured from the point of rupture initiation. vt is the instantaneous radius of the rupture at time t . $\Delta\sigma$ is the constant stress drop inside the rupture zone, μ is the elastic shear modulus, and $C(v)$ is a very slowly varying function of the rupture velocity. For most practical purposes, $C \sim 1$. Inside the fault, for $r < vt$, the stress change produced by the slip function is constant and equal to $\Delta\sigma$. Thus for constant stress drop, slip inside the fault scales with the "ratio of stress

drop over rigidity ‘times’ the instantaneous radius of the fault.” As rupture propagates, displacement around the fault scales with the instantaneous radius of the rupture zone. The dynamic model (12) describes rupture propagation, but it has no mechanism to stop the rupture.

A finite fault model can be developed from Eq. (12) by assuming that rupture suddenly stops after the fault has grown to a certain fixed radius. This simple dynamic model was proposed by Madariaga (1976). The circular crack problem is posed in terms of stresses not of slip, but the rupture process is fixed in advance so that rupture does not develop spontaneously. This is the only unrealistic feature of this model, and it is the reason why it is qualified as quasi-dynamic, that is, rupture is kinematically defined, but slip is computed by solving the elastodynamic equations.

The static circular shear crack. We start by a quick study of a simple circular crack from which we derive some of the most fundamental properties of dynamic source models. Let us consider a static circular (“penny shaped”) crack of radius a lying on the x, y plane. Assuming that the fault is loaded by an initial shear stress σ_0 and that the stress drop $\Delta\sigma$ is uniform inside the fault, slip on the fault is given by:

$$D(r) = \frac{24}{7\pi} \frac{\Delta\sigma}{\mu} \sqrt{a^2 - r^2} \quad (13)$$

where r is the radial distance from the center of the fault, a is the radius of the shear crack, and μ is the elastic rigidity of the medium surrounding the crack. Slip in this model has

the typical elliptical shape that we associate with cracks. The taper of slip near the edges of the crack is the hallmark of dynamic source models. The slip distribution near the edges has a parabolic shape that is closely related to stress concentrations outside the fault zone.

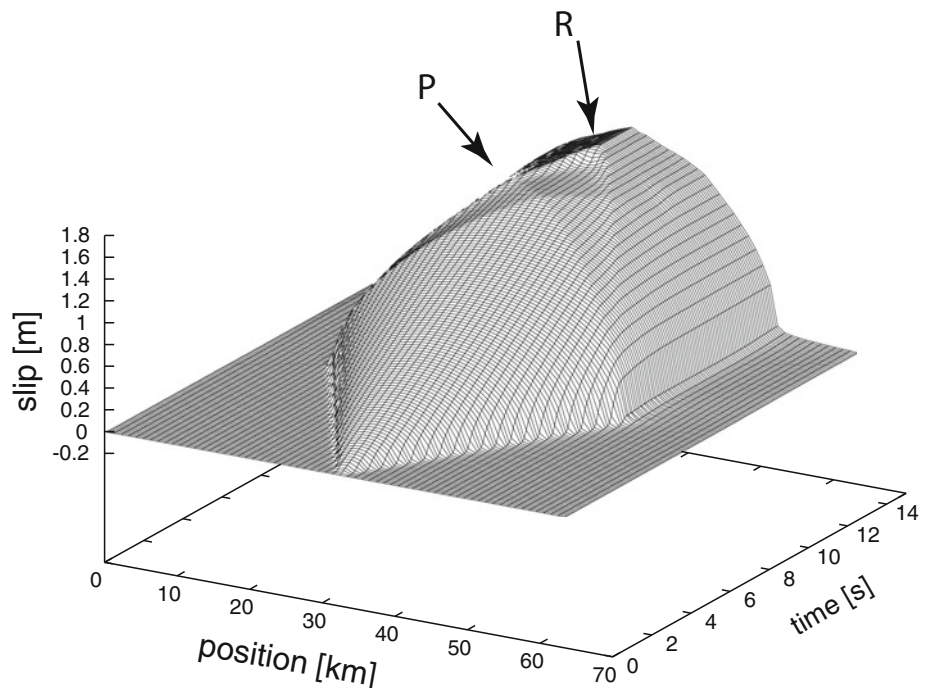
From Eq. (13) we can determine the scalar seismic moment for a circular fault, $M_0 = 16/7 \Delta\sigma a^3$,

so that the moment is the product of the stress drop times the cube of the fault size. This simple relation is the basis of the seismic scaling law proposed by Aki (1967). The circular crack model has been used to quantify numerous small earthquakes for which the moment was estimated from the amplitude of seismic waves and the source radius was estimated from corner frequencies, aftershock distribution, etc.; the result is that for shallow earthquakes in crustal seismogenic zones like the San Andreas fault or the North Anatolian fault in Turkey, stress drops are of the order of 1–10 MPa. For deeper events in subduction zones, stress drops can reach several tens of MPa. Thus, in earthquakes stresses do not change much, at most a couple of order of magnitudes, while source radius varies over several orders of magnitudes from meters to 100 km or more. It is only in this sense that the usual assertion “stress drop in earthquakes is constant” should be taken; it actually changes but much less than the other parameters in the scaling law.

The quasi-dynamic circular crack. There are no simple analytical solutions like (13) for finite dynamic cracks. We are forced to use numerical solutions that are actually very simple to compute. The full solution to the circular crack problem is shown in Fig. 1. Initially, until the sudden arrest

Earthquake Source Theory,

Fig. 1 Slip distribution as a function of time and position for Madariaga’s (1976) quasi-dynamic circular crack model



of rupture at the final radius a , the slip distribution can be accurately modeled using Kostrov's self-similar solution (12). The stopping of rupture generates strong healing waves that propagate inward from the rim of the fault. These waves are of three types: P, S, and Rayleigh waves. Soon after the passage of the Rayleigh waves, slip rate inside the fault decreases to zero and the fault heals. After healing, we assume that frictional forces are sufficiently strong that no slip will occur until the fault is reloaded. As observed in Fig. 1, it is clear that slip and rise time are functions of position on the fault, where the rise time being longer near the center and slip is also larger near the edges of the fault where slip is clamped. Finally, let us note that the slip after healing is very similar to that of a static circular crack, except that there is a slight overshoot of slip with respect to the static solution (13). The overshoot is of course a function of the rupture speed, but its maximum value is of the order of 15% for a rupture speed of 0.75.

Future Work

Seismic observations have steadily improved in the last 20 years due to the wide availability of digital, continuously recording accelerograms that can be integrated to obtain velocity and band-limited displacement fields near active faults. Additional information about rupture is provided by field studies of faulting associated with earthquakes and from geodetic observations. Modeling and inversion of such data have been usually done using kinematic models derived from the original Haskell model. This approach gives solutions that are very non-unique because the rupture process (the rupture velocity) is difficult to estimate. Rupture velocity actually controls seismic radiation, so that poor velocity control leads to poor seismic inversions. There is hope that dynamic models, where rupture is inverted as part of the inversion process, will provide better estimates of the stress field around major fault zones and provide some essential parameters, like the friction law that controls rupture propagation.

Bibliography

- Aki K (1967) Scaling law of seismic spectrum. *J Geophys Res* 73:5359–5376
- Aki K, Richards PG (2002) *Quantitative seismology*, 2nd edn. University Science Books, Sausalito
- Backus G, Mulcahy M (1976) Moment tensors and other phenomenological descriptions of seismic sources. I: continuous displacements. *Geophys J Roy Astr Soc* 46:321–361
- Brune J (1970) Tectonic stress and the spectra of seismic shear waves from earthquakes. *J Geophys Res* 75:4997–5009
- Burridge R, Knopoff L (1964) Body force equivalents for seismic dislocations. *Bull Seismol Soc Am* 54:1875–1878

- Eshelby JD (1957) The elastic field of an ellipsoid inclusion and related problems. *Proc R Soc Lond A* 241:376–396
- Haskell NA (1964) Total energy and energy spectral density of elastic wave radiation from propagating faults. *Bull Seismol Soc Am* 54:1811–1841
- Kostrov BV (1964) Self-similar problems of propagation of 1379 shear cracks. *J Appl Math Mech* 28:1077–1087
- Madariaga R (1976) Dynamics of an expanding circular fault. *Bull Seismol Soc Am* 66:639–667
- Udias A, Madariaga R, Buforn E (2014) *Source mechanisms of earthquakes, theory and practice*. Cambridge University Press, Cambridge, UK. ISBN: 9781107040274

Earthquake, Aftershocks

Mian Liu¹ and Seth Stein²

¹Department of Geological Sciences, University of Missouri, Columbia, MO, USA

²Department of Earth and Planetary Sciences, Northwestern University, Evanston, IL, USA

Definition

Aftershocks Earthquakes that follow a large earthquake (the mainshock) in an earthquake sequence.

Introduction

Earthquakes typically occur in sequences that may include foreshocks, the mainshock (the largest event or events), and aftershocks. Earthquake sequences without a clear mainshock are called swarms.

Aftershocks generally refer to the smaller earthquakes that follow a mainshock within certain spatial and temporal windows. However, the criteria for choosing these windows are somewhat arbitrary. Typically, aftershocks are defined within an area around the mainshock's source region (i.e., the ruptured fault segment, which is about 50–100 km long for a magnitude 7.0 earthquake). Most aftershocks occur on the main rupture surface; hence, they are often used to define the complex geometry of the rupture plane. However, in many cases, especially for earthquakes in subduction zones, the aftershock area increases significantly following the mainshock (Tajima and Kanamori 1985). Some aftershocks also occur off the main rupture surface. Distant earthquakes triggered by the mainshock's seismic waves (see ► [“Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India”](#)) may also be regarded as aftershocks (Gomberg and Johnson 2005). Temporally, aftershocks are defined as elevated seismicity above the background activity following the main shock. This definition

also leaves room for ambiguity, as the background seismic activity can be difficult to define in some places. Aftershocks usually count for less than 5% of the total seismic energy release of the entire seismic sequence (Scholz 2002).

Scaling Laws for Aftershocks

A large earthquake is usually followed by many smaller events (aftershocks); their occurrence rate decreases with time, typically following a pattern known as the Omori's law, named for Fusakichi Omori's observation of the 1891 Nobi earthquake in Japan (Omori 1894) and later modified by Utsu (1961):

$$n(t) = \frac{K}{(c+t)^p}$$

where $n(t)$ is the number of aftershocks by time t after the mainshock; K , p , and c are constants. The value of c is typically positive close to zero, and p close to 1. Hence, the number of aftershocks drops hyperbolically with time (Fig. 1). The values of these constants are obtained by fitting to the data for each aftershock sequence.

It has been observed that the largest aftershock is usually about one magnitude unit smaller than the mainshock independent of the mainshock magnitude. This is known as Bath's law (Bath 1965; Richter 1958). However, because the data selection is retrospective and subjective, the size of aftershocks can vary substantially for different earthquake sequences.

Like all earthquakes, the size distribution of aftershocks follow the Gutenberg–Richter law (Gutenberg and Richter 1954):

$$\log_{10} N(M) = a - bM$$

where $N(M)$ is the number of earthquakes of magnitude $\geq M$, a and b are constants. This relationship plots as a straight line with slope b , whose value, often referred to as the b -value, is typically in the range of 0.8–1.2. For $b = 1$, the number of earthquakes increases by a factor of 10 for every unit drop. The b -value varies in time and space even for the same earthquake sequence. Some studies have suggested that the b -value can be used as a stress indicator, with lower b -values often associated with higher stresses (Schorlemmer et al. 2005).

Causes of Aftershocks

Aftershocks are ubiquitous; they are observed following nearly all sizable earthquakes near the rupture zones of the mainshocks. Such close spatiotemporal relationship with

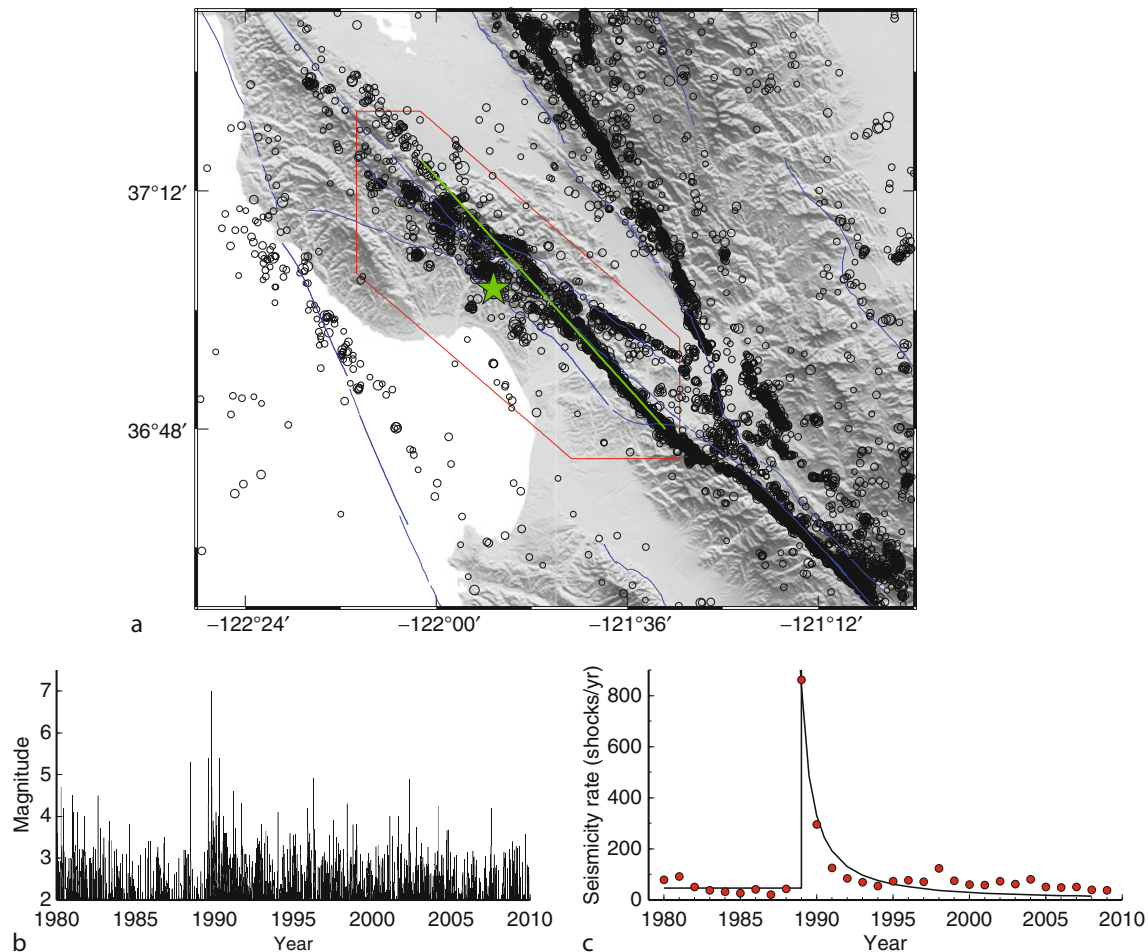
mainshock indicates that aftershocks result from incomplete rupture and heterogeneous slip (e.g., Bullen and Bolt 1947). This notion is consistent with some case studies that found aftershocks concentrated around the perimeter of the rupture zone or around asperities within the rupture zone where most of the coseismic slip occurred (Beroza 1991; Scholz 2002; Wetzler et al. 2018).

The main causes of aftershocks are thought to include mainshock-induced changes of frictional properties of the fault zone and stress perturbations. Laboratory experiments indicate that frictional properties change with the duration of stationary contact and the instantaneous sliding velocity (Scholz 2002). A constitutive relationship based on such results, known as the rate-and-state law (Dieterich 1979; Ruina 1983), can explain the time-dependent changes of seismic rates that are consistent with the Omori's law (Dieterich 1994). Stress perturbations may arise from creep recovery of rocks in the immediate areas of the fault (Benioff 1951), viscous relaxation from the lower crustal and upper mantle (Lieber and Braslau 1965; Stein and Liu 2009), after-slip (Koper et al. 2018; Ross et al. 2017), and pore-elastic effects (Scholz 2002). In particular, the changes of the Coulomb static stress resulted from the mainshock have been shown to cause spatial migration of seismicity and trigger earthquakes (Parsons 2002; Stein 1999). Other causes may include dynamic triggering (Gomberg and Johnson 2005; Fan and Shearer 2016) and pore fluid flow (Miller et al. 2004; Nur and Booker 1972).

Aftershocks and Earthquake Hazard

Aftershocks can be large and damaging (e.g., Wiemer et al. 2002). The durations of aftershock sequences for plate-boundary earthquakes are typically less than 10 years (Parsons 2002) (Fig. 1). However, aftershock sequences away from plate-boundary faults tend to last longer. In the broadly deforming western United States, for example, aftershocks are found to last 50–100 years. Within the stable continental interior, aftershock sequences may last hundreds of years or even longer (Ebel et al. 2000; Stein and Liu 2009). Such long aftershock sequences in continental interiors, where large earthquakes are infrequent and historic records are often incomplete, could bias hazard assessment.

Further complication may arise from the fact that earthquakes tend to cluster in time and space. Studies of large shallow earthquakes in the world show that these large events often occur in pairs or groups (called doublets or multiplets); the later events, which could be regarded as aftershocks, can be bigger than the triggering event (the mainshock) (Kagan and Jackson 1999). Such clustering patterns are consistent with the view that all earthquakes can trigger their own earthquakes that in turn trigger more quakes,



Earthquake, Aftershocks, Fig. 1 (a) Topography and seismicity (circles, size proportional to magnitude) around the epicenter (star) of the 1989 Loma Prieta earthquake (Mw 6.9) on the San Andreas Fault. Green line marks the fault rupture; red lines enclose the spatial window

for the aftershocks. (b) Earthquake occurrence between 1980 and 2010 within the spatial window. (c) Plot of seismicity showing the t^{-1} decay of aftershocks

and the triggered events may be bigger than the triggering event (Helmstetter et al. 2003; Ogata 1998). Because the triggered events (the aftershocks) statistically follow the Omori's law, statistical models can be developed to assess time-dependent probability of future damaging earthquakes after each event (Gerstenberger et al. 2005; Jones 1985; Page et al. 2016; Reasenberg and Jones 1989).

Summary

Aftershocks are earthquakes that follow the mainshock in an earthquake sequence. They are typically smaller than the mainshock and occur on or near the rupture plane of the mainshock, resulting from changes of stress and frictional properties of the fault zone caused by the mainshock. The duration of aftershock sequences is typically a few years for earthquakes at plate boundaries, but can last much longer within stable continental interiors.

Cross-References

- [Earthquake Source Theory](#)
- [Earthquake, Foreshocks](#)

Bibliography

- Bath M (1965) Lateral inhomogeneities in the upper mantle. *Tectonophysics* 2:483–514
- Benioff H (1951) Earthquakes and rock creep. *Bull Seismol Soc Am* 41:31–62
- Beroza GC (1991) Near-source modeling of the Loma Prieta earthquake: evidence for heterogeneous slip and implications for earthquake hazard. *Bull Seismol Soc Am* 81:1603–1621
- Bullen KE, Bolt BA (1947) An introduction to the theory of seismology. Cambridge University Press, New York, p 499
- Dieterich J (1979) Modeling of rock friction: 1. Experimental results and constitutive equations. *J Geophys Res* 84:2161–2168
- Dieterich J (1994) A constitutive law for rate of earthquake production and its application to earthquake clustering. *J Geophys Res* 99:2601–2618

- Ebel JE, Bonjer K-P, Oncescu MC (2000) Paleoseismicity: seismicity evidence for past large earthquakes. *Seismol Res Lett* 71:283–294
- Fan W, Shearer PM (2016) Local near instantaneously dynamically triggered aftershocks of large earthquakes. *Science* 353:1133–1136
- Gerstenberger MC, Wiemer S, Jones LM, Reasenber PA (2005) Real-time forecasts of tomorrow's earthquakes in California. *Nature* 435:328–331
- Gomberg J, Johnson P (2005) Seismology: dynamic triggering of earthquakes. *Nature* 437:830
- Gutenberg B, Richter CF (1954) Seismicity of the earth and associated phenomena, vol ix. Princeton University Press, Princeton, p 310
- Helmstetter A, Sornette D, Grasso J-R (2003) Mainshocks are aftershocks of conditional foreshocks: how do foreshock statistical properties emerge from aftershock laws. *J Geophys Res* 108:2046
- Jones LM (1985) Foreshocks and time-dependent earthquake hazard assessment in southern California. *Bull Seismol Soc Am* 75:1667–1679
- Kagan YY, Jackson DD (1999) Worldwide doublets of large shallow earthquakes. *Bull Seismol Soc Am* 89:1147–1155
- Koper KD, Pankow KL, Pechmann JC, Hale JM, Burlacu R, Yeck WL, Benz HM, Herrmann RB, Trugman DT, Shearer PM (2018) Afterslip enhanced aftershock activity during the 2017 earthquake sequence near Sulphur Peak, Idaho. *Geophys Res Lett* 45:5352–5361
- Lieber P, Braslau D (1965) On an earthquake and aftershock mechanism relating to a model of the crust and mantle. Report Am-65-8. Office of Research Services, University of California, Berkeley, p 141
- Miller SA, Collettini C, Chiaraluce L, Cocco M, Barchi M, Kaus BJP (2004) Aftershocks driven by a high-pressure CO₂ source at depth. *Nature* 427:724–727
- Nur A, Booker JR (1972) Aftershocks caused by pore fluid flow? *Science* 175:885–887
- Ogata Y (1998) Space-time point-process models for earthquake occurrences. *Ann Inst Stat Math* 50:379–402
- Omori F (1894) On the aftershocks of earthquakes. *J Coll Sci Imperial UnivTokyo* 7:111–200
- Page MT, Van Der Elst N, Hardebeck J, Felzer K, Michael AJ (2016) Three ingredients for improved global aftershock forecasts: tectonic region, time-dependent catalog incompleteness, and intersequence variability. *Bull Seismol Soc Am* 106:2290–2301
- Parsons T (2002) Global Omori law decay of triggered earthquakes: large aftershocks outside the classical aftershock zone. *J Geophys Res* 107(B9):2199. <https://doi.org/10.1029/2001JB000646>
- Reasenber PA, Jones LM (1989) Earthquake hazard after a mainshock in California. *Science* 243:1173–1176
- Richter CF (1958) Elementary seismology. W.H. Freeman, San Francisco, 768 pp
- Ross ZE, Rollins C, Cochran ES, Hauksson E, Avouac JP, Ben-Zion Y (2017) Aftershocks driven by afterslip and fluid pressure sweeping through a fault-fracture mesh. *Geophys Res Lett* 44:8260–8267
- Ruina A (1983) Slip instability and state variable friction laws. *J Geophys Res* 88:10359–10370
- Scholz CH (2002) The mechanics of earthquakes and faulting. Cambridge University Press, Cambridge/New York, 471 pp
- Schorlemmer D, Wiemer S, Wyss M (2005) Variations in earthquake-size distribution across different stress regimes. *Nature* 437:539–542
- Stein RS (1999) The role of stress transfer in earthquake occurrence. *Nature* 402:605–609
- Stein S, Liu M (2009) Long aftershock sequences within continents and implications for earthquake hazard assessment. *Nature* 462:87–89
- Tajima F, Kanamori H (1985) Global survey of aftershock area expansion patterns. *Phys Earth Planet Inter* 40:77–134
- Utsu T (1961) A statistical study of the occurrence of aftershocks. *Geophys Mag* 30:521–605
- Utsu T (2002) Statistical features of seismicity. In: Lee WHK (ed) International handbook of earthquake & engineering seismology, Part A. Academic, San Diego, pp 719–732
- Wetzler N, Lay T, Brodsky EE, Kanamori H (2018) Systematic deficiency of aftershocks in areas of high coseismic slip for large subduction zone earthquakes. *Sci Adv* 4. <https://doi.org/10.1126/sciadv.aao3225>
- Wiemer S, Gerstenberger M, Hauksson E (2002) Properties of the aftershock sequence of the 1999 Mw 7.1 Hector mine earthquake: implications for aftershock hazard. *Bull Seismol Soc Am* 92:1227–1240

Earthquake, Archaeoseismology

Klaus-G. Hinzen

Department of Geosciences, University of Cologne,
Cologne, Germany

Synonyms

Earthquake archaeology

Definition

The study of pre-instrumental earthquakes that, by affecting locations of human occupation and their environments, have left their mark in ancient structures uncovered by means of archaeological excavations or pertaining to the monumental heritage (Buck and Stewart 2000; Galadini et al. 2006).

Introduction

Ever since man-made structures have been erected, earthquakes have left their marks on these constructions. However, damage in archaeologically excavated buildings or continuously preserved monuments is often hard to unravel in terms of the causative effects. The use of archaeological data to investigate unknown or poorly known historical earthquakes and descriptions of earthquake effects was recorded in the archaeological heritage started in the nineteenth and early twentieth centuries (e.g., De Rossi 1874; Schliemann 1881; Evans 1928). Since the 1980s, the increased interest in the subject led to the publication of special volumes and articles in seismological and geological journals (e.g., Guidoboni 1989; Stiros and Jones 1996; McGuire et al. 2000; Galadini et al. 2006; Reicherter et al. 2009; Jusseret and Sintubin 2017). While earlier investigations were dominated by the qualitative description of damage records, more recent studies take a quantitative approach (e.g., Hinzen et al. 2018). The main questions to be answered by archaeoseismic investigations are the following: (1) How probable are seismic ground motions, or secondary earthquake effects, as the cause of

damage observed in man-made structures from the past? (2) When did the damaging ground motion occur? (3) What can be deduced about the nature of the causing earthquake?

Archaeoseismic Observations

The marks in ancient structures relevant for archaeoseismology fall into four main categories: (1) displacements along shear planes directly linked to the earthquake fault plane or side branches of it. In particular, earthquakes with a strike slip mechanism can leave distinctive traces in buildings and lifelines such as aqueducts, roads, and sewer systems. Case studies (e.g., Ellenblum et al. 1998; Meghraoui et al. 2003) show that under favorable conditions, the amount of slip and in case of repeated events the slip rate of the faults can be revealed. (2) Off-fault shaking effects including fractured building elements, tilted walls, shift of building elements, lateral warping, breaking and overthrowing of walls, and rotations of vertically oriented objects (tomb stones, columns, monuments). For most of these features, a seismogenic origin is not the only possible interpretation. Therefore, alternative causes must be taken into account in the damage analysis. (3) The secondary shaking effects lateral spreading and cyclic mobility as a consequence of liquefaction of the subsurface. Liquefaction always requires a certain level of dynamic excitation. So secondary damages in buildings and monuments due to liquefaction help exclude alternative causes from a damage scenario. (4) Archaeologically detected abandonment of a site and evidence of repair and rebuilding. These observations are mainly of interest in connection with the reasons mentioned before, because as a single observation in general, they do not give enough evidence for the conclusion of a seismogenic cause (Galadini et al. 2006).

While the first category is limited directly to the linear features of active faults, off-fault shaking affects a much larger area and is more common but harder to prove. Ancient structures show deformations related to seismic shaking similar to those observed in recent earthquakes. Typical earthquake effects on masonry walls are (1) cross fissures, due to stress concentrations often nucleating at corners of doors and windows, driven by shear forces; (2) corner expulsion of walls caused by differential movements in orthogonal directions; (3) horizontal and independent lateral and rotational shift of wall blocks, best visible in uncemented walls made of rectangular blocks; (5) spall of block corners due to stress concentrations; (6) height reduction by vertical crashing; (7) movement of keystones and rupture of arch piers; (8) rotation of vertically oriented objects; and (9) domino-like toppling of structured columns and walls. Examples are shown in Fig. 1. Following the term *seismite*, used for seismically disturbed sediment layers, the abovementioned deformations

may be called *archaeoseismites*. Several of these deformations may also originate without dynamic earthquake excitation. In these cases, a single piece of evidence, or at a single edifice only, cannot be considered a conclusive *archaeoseismite*. If damage is identified as *archaeoseismites*, intensities can be derived following the classical macroseismic methods. For arches affected by earthquake ground motions, Hinzen et al. (2016) introduced a specific rating scheme.

Methods

Archaeoseismology requires the integration of several scientific disciplines. Figure 2 shows a scheme of workflow in a quantitative archaeoseismic study. Besides archaeological and seismological methods, input from geodesy, pure and applied geology, and civil engineering is needed (Ambraseys 2006). Physical absolute dating techniques including radiometric methods, luminescence methods, and dendrochronology are necessary to determine a time window for the damaging event. In case of multiple events, the rough dates are crucial for seismicity models and hazard analysis.

The first step in studying field cases is the documentation and analysis of damage to the archaeologically excavated buildings or objects. In addition to the important traditional excavation techniques and classical measurements and drawings, 3D imaging techniques have proved their usefulness in archaeoseismic field cases. Photogrammetry and laser scanning are fast in the field and allow detailed analysis and post processing, particularly a quantification of deformations, fractures, etc. In cases where building parts are fragile and cannot be conserved in the original finding situation, successive laser scans during the excavation offer the possibility to build virtual 3D models. In ideal cases, the seismological relevant data can be acquired during an ongoing excavation. Important is the complete documentation of the stratigraphy of debris layers, which can reveal the decay rate of structures (Galadini et al. 2006). Building collapse may be sudden, progressive, or a combination of both. The progressive deterioration of walls results in accumulations having maximum thickness close to the wall feeding the material. The thickness decreases at a short distance from the wall. In case of height reduction of walls or buildings due to earthquake effects, the debris is more distributed. Sometimes even complete walls topple to one side; in this case the impact pattern can help to constrain ground motion amplitudes. However, if the findings of previous excavations are archaeoseismologically analyzed, the current situation and available excavation reports form the database. Possible alterations of the excavated objects since the first excavation and shortcomings in the documentation, especially if earthquake actions were not considered at the time of the original excavation, might hinder the interpretation (Hinzen et al. 2018).



Earthquake, Archaeoseismology, Fig. 1 Examples of deformations and damage which possibly are earthquake effects: (a) horizontally deformed wall of a crusader fortress build on top of the Dead Sea Transform Fault in the Jordan Valley; (b) deformed vault of a Roman sewer in Cologne, Germany; (c) toppled columns of a Byzantine church in Sussita located above the Sea of Galilee; (d) toppled column of the great palace in Patra, Jordan; (e) moved block in an arch of the Nimrod

fortress in the Golan Heights; (f) shifted blocks of an analemma of a Roman theater in Pinara, SW Turkey; (g) moved blocks of a corner wall of a Roman monument in Patara, SW Turkey; (h) shifted blocks of a Roman grave house in Pinara, SW Turkey; (i) spall of block corners, same object as in (g); (j) broken and horizontally displaced fortification wall of the Roman Tolbiacum (Zülpich, Germany); (k) rotated Lycien sarcophagus in Pinara, SW Turkey

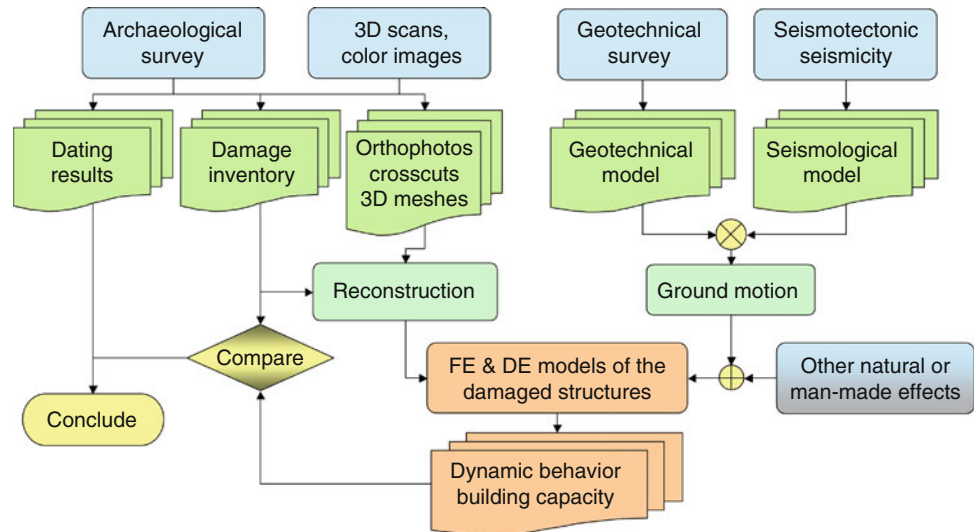
Depending on site conditions, exploration and modeling of the properties of the local geology at the site are an important second step. Several of the abovementioned deformations can also be caused by slow, quasi-static deformations of the subsoil, and ground amplification due to shallow soft

sediments can significantly amplify seismic shaking. If the latter is not taken into account, the size of the damaging earthquake can easily be overestimated.

Only on rare occasions, site-specific strong motion records from recent earthquakes will be available and help to prove or

Earthquake, Archaeoseismology,

Fig. 2 Schematic workflow in a quantitative multidisciplinary archaeoseismic study. The dynamic behavior is studied by finite element (FE) or discrete element (DE) methods



disprove proposed damage scenarios. To overcome this lack of information, the calculation of synthetic seismograms based on models of the capable faults and local geology has been used to better ascertain the probability of a seismogenic cause of the damages and name faults on which the damaging earthquake occurred. When other natural forces (flooding, storm, mass movement) or even human action also come into consideration, their action on the excavated objects has to be tested as well.

Reconstructions of the damaged archaeological objects are the basis for determining the dynamic behavior and the vulnerability of the ancient buildings. Common earthquake engineering techniques including finite element and discrete element modeling are used to analyze the reaction of the buildings to the estimated seismic loading functions. Results from such models are compared to the damages observed in situ.

The most challenging task of an archaeoseismic study is the determination of source parameters of the earthquake, which caused the analyzed damages. While timing of the earthquake has to be done by archaeological techniques including classical archaeological stratigraphy and the abovementioned physical dating, location and strength of the earthquake can only be narrowed down with seismological models. A territorial approach (Galadini et al. 2006) in which the collected information from the surroundings of an archaeological site are taken into account is the most promising tool to determine earthquake source parameters.

Recent studies have shown that precariously balanced archaeological structures can be treated in the same way as precariously balanced rock, which can give an estimate of ground motion levels at a site that have not been reached or exceeded. Schweppe et al. (2017) explored the potential

of this method by applying it to the remains of a Roman temple in Israel.

Summary

In the past decades, archaeoseismology has evolved as a new branch of seismological sciences. While palaeoseismology focuses on the marks left by earthquakes in the geological record, archaeoseismologists analyze and interpret marks in the archaeological record and/or in preserved monuments.

The use of quantitative methods in archaeoseismology, including calculation of synthetic site-specific strong motion seismograms, modeling of natural non-earthquake-related forces, anthropogenic forces, and finite or discrete element models of structures, supports conclusive discrimination between potential damage scenarios. However, if model parameters cannot be well constrained, modeling result uncertainties might still be too large to draw definite conclusions. Common sense interpretations of archaeoseismites as solitary evidence are generally too vague to complement earthquake catalogs for a seismic hazard analysis. Recent advances in ground motion simulation methods and computational possibilities promise to refine quantitative archaeoseismological methods and establish them at levels equal to historical and palaeoseismological methods. Finally, even if an archaeoseismic study does not deliver the often-requested improvement of hazard determination, it can still advance our picture of the past by attempting to answer open archaeological, historical, and geologic questions in a scientific manner. And following Ambraseys (2006), surveying and mapping an archaeological site is an art; verifying the cause of damage is science.

Cross-References

- [Absolute Age Determinations: Radiometric](#)
- [Earthquakes, Intensity](#)
- [Earthquakes, Strong-Ground Motion](#)
- [Numerical Methods, Finite Element](#)
- [Paleoseismology](#)
- [Seismic Hazard](#)

Bibliography

- Ambraseys NN (2006) Earthquakes and archaeology. *J Archaeol Sci* 33:108–1016
- Buck V, Stewart I (2000) A critical reappraisal of classical and archaeological evidence for earthquakes in the Atalanti region, central mainland Greece. In: The Archaeology of Geological Catastrophes, McGuire WG, Griffith DR, Hancock PL, Stewart I (eds) Geological society, vol 171. Special Publications, London, pp 33–44
- De Rossi MS (1874) La basilica di Santa Petronilla presso Roma, testé scoperta, caduta per terremoto. *Bull Vulc It* 1:62–65
- Ellenblum R, Marco S, Agnon A, Rockwell T, Boas A (1998) Crusader castle torn apart by earthquake at dawn. *Geology* 26:303–306
- Evans A (1928) The palace of Minos, part II, MacMillan and Co., London, p 844
- Galadini F, Hinzen K-G, Stiros S (2006) Archaeoseismology: methodological issues and procedure. *J Seismol* 10:395–414
- Guidoboni E (ed) (1989) I terremoti prima del Mille in Italia e nell'area mediterranea. *Storia, Archeologia, Sismologia*, vol 765. ING-SGA, Bologna
- Hinzen K-G, Schwellenbach I, Schweppe G, Marco S (2016) Quantifying earthquake effects on ancient arches, example: the Kalat nimrod fortress, Dead Sea fault zone. *Seismol Res Lett* 78:14. <https://doi.org/10.1785/0220150282>
- Hinzen K-G, Maran J, Hinojosa-Prieto H, Damm-Meinhardt U, Reamer SK, Tzislakis J, Kemna K, Schweppe G, Fleischer C, Demakopoulou K (2018) Reassessing the Mycenaean earthquake hypothesis: results of the HERACLES project from Tiryns and Midea, Greece. *Bull Seismol Soc Am*. <https://doi.org/10.1785/0120170348>
- Jusseret S, Sintubin M (2017) Minoan earthquakes: breaking the myth through interdisciplinarity. Leuven University Press, Leuven, p 408
- McGuire WJ, Griffith DR, Hancock PL, Stewart IS (eds) (2000) The archaeology of geological catastrophes, vol 171. Geological Society Spec. Publ, London. 417 pp
- Meghraoui M, Gomez F, Sbeinati R, Van der Woerd J, Mounty M, Darkal AN, Radwan Y, Layyous I, Al-Najjar H, Darawcheh R, Hijazi F, Al-Ghazzi R, Barazangi M (2003) Evidence for 830 years of seismic quiescence from palaeoseismology, archaeoseismology, and historical seismicity along the Dead Sea fault in Syria. *Earth Planet Sci Lett* 210:35–52
- Reicherter K, Michetti AM, Silva PG (eds) (2009). Palaeoseismology: historical and prehistorical records of earthquake ground effects for seismic hazard assessment, vol 316. The Geological Society Special Publications, London, 324 pp
- Schliemann H (1881) Autobiographie des Verfassers und Geschichte seiner Arbeiten in Troja. In: Ilios Stadt und Land der Trojaner, Forschungen und Entdeckungen in der Troas und besonders auf der Baustelle von Troja, vol 78, F.A. Brockhaus, Leipzig, p 26
- Stiros SC, Jones RE (eds) (1996) Archaeoseismology. Fitch laboratory occasional paper, vol 7. British school at Athens, Athens, 268 pp
- Schweppe G, Hinzen K-G, Reamer SK, Fisher M, Marco S (2017) The ruin of the roman temple of Kedesh, Israel; example of a precariously balanced archaeological structure used as a seismoscope. *Ann Geophys* 60. <https://doi.org/10.4401/ag-7152>

Earthquake, Focal Mechanism

Emile A. Okal

Department of Earth and Planetary Sciences, Northwestern University, Evanston, IL, USA

Definition

Focal mechanisms are geometrical or mathematical representations of the faulting during an earthquake. In very simple terms, the latter consists of the relative motions of two blocks of Earth called *walls* along a planar surface called *fault*. Figure 1 shows that the description of an earthquake rupture needs three angles. The *strike* angle ϕ is the azimuth (with respect to North) of the trace of the fault on a horizontal plane such as the Earth's surface; the *dip* angle δ characterizes the steepness of the fault, and the *rake* or *slip* angle λ , the direction of motion, within the fault plane and relative to the horizontal, of the hanging wall relative to the foot wall. A full description of the earthquake rupture requires an additional scalar, related to the length of slip of the fault walls, and thus to the size or *magnitude* of the event.

Representation by a Double-Couple

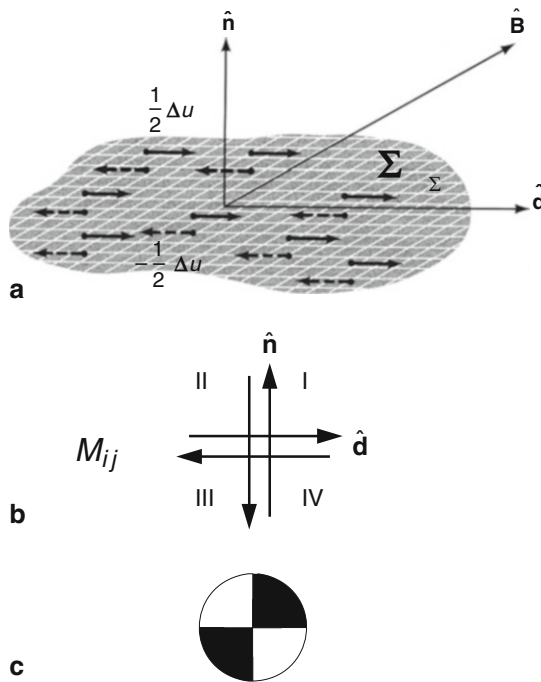
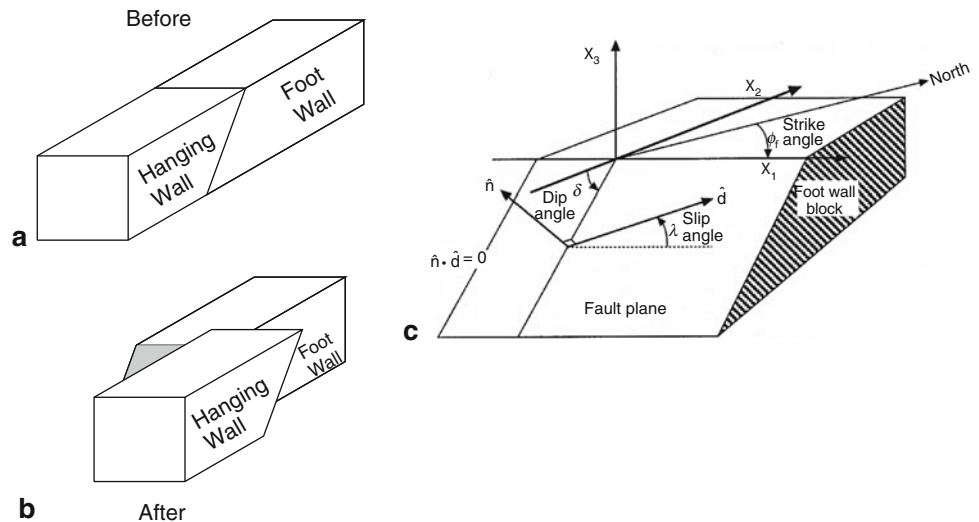
As such, an earthquake source, described by three angles and a scalar, is a more complex mathematical entity than a simple vector (described by two angles and a scalar) and in particular cannot be represented physically by a single force. Vvedenskaya (1956), followed in the West by Knopoff and Gilbert (1959), introduced the concept of the double-couple, a system of forces shown on Fig. 2 and consisting of two opposing torques of equal moment M_0 ; this system has no dynamic effect (translation or rotation) on a rigid body but results in the deformation of a medium of finite rigidity.

The representation theorem (e.g., Aki and Richards 1980, pp. 38–41) considers an elastic medium of rigidity μ and states that a dislocation with slip Δu (in the direction $\hat{\mathbf{d}}$) along a fault of surface Σ cut into the medium (Fig. 2a) will generate the same field of dynamic deformation (expressed as seismic waves) as a double-couple of moment $M_0 = \mu \Sigma \Delta u$, embedded in the medium in the absence of the cut, with the direction of forces and levers in Fig. 2b along those of the slip, $\hat{\mathbf{d}}$, and of the normal to the fault plane, $\hat{\mathbf{n}}$.

The double-couple shown in Fig. 2b separates the space around it into four quadrants containing respectively the heads and tails of the arrows. They are delineated by the fault plane and the plane normal to the slip $\hat{\mathbf{d}}$. In quadrants I and III, the seismic motion will be “*anaseismic*,” or away from the source, while in II and IV, it will be “*kataseismic*” or

Earthquake, Focal Mechanism,

Fig. 1 (a, b) Before and after sketches of the displacement of fault blocks during an earthquake. The fault plane is shaded in (b). (c) Definition of the strike, dip, and slip angles, ϕ , δ , λ , respectively, in the geometry of the rupture represented at left. The unit vector $\hat{\mathbf{d}}$ represents the direction of motion of the hanging wall (not represented to prevent clutter) with respect to the foot wall. (c: After Stein and Wysession 2003)



Earthquake, Focal Mechanism, Fig. 2 (a) Seismic slip along a fault (represented as the hatched area of surface Σ). The total slip Δu is the difference between the displacements of the top wall (shown as solid arrows) and of the bottom wall (dashed arrows). (b) “Double-couple” system of forces, equivalent, in the unruptured medium, to the dislocation shown in (a). (c) “Beachball” representation of the quadrants defined by the orientation of the double-couple. (a: After Aki and Richards 1980)

toward the source. The quadrants are accordingly shaded black and white in Fig. 2c. Note that in Fig. 2b, c, the source is approximated as a *point* in space, hence the concept of a “point-source double-couple.”

Focal Mechanisms from P-Wave First Motions

Under the high-frequency approximation of geometrical optics, a most remarkable property of *P* waves is the conservation of the ana- or kataseismic character along the entire seismic ray, all the way to a receiving seismic station at the Earth’s surface, where an anaseismic wave results in an initial *upward* motion of the ground, while a kataseismic one gives a *downward* first motion.

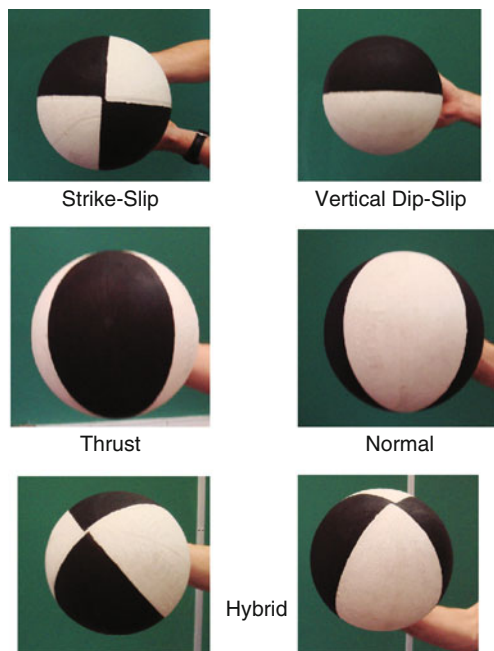
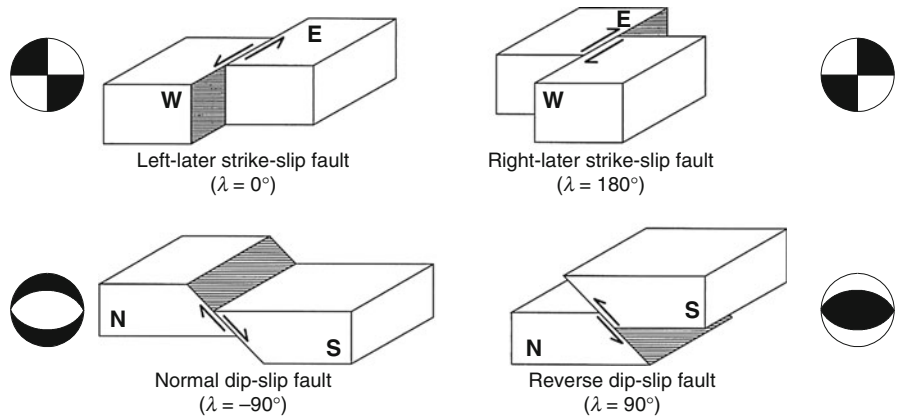
This property has been used extensively to obtain earthquake focal mechanisms by gathering first-motion *P*-wave data at large datasets of local and distant stations and backtracking this information along the relevant rays into a mapping of the focal sphere at the source. This amounts to resolving the orientation in space of the diagram on Fig. 2c, known traditionally as a “beachball.” Because of the difficulty of representing a full sphere on a planar figure (long known to world cartographers), and thanks to the symmetry of the beachball, it is convenient and sufficient to represent only a stereographic projection of the *lower focal hemisphere* separated at the horizontal plane and involving rays leaving down into the deep Earth, and eventually emerging at large distances. This constitutes the standard representation of earthquake focal mechanisms. Figure 3 illustrates examples of the most common types of mechanisms.

Note that a change of focal mechanism results in a mere *solid rotation* in space of the beachball pattern but does not change its nature (Fig. 4).

A common feature of the double-couple concept and of its beachball representation is their high level of symmetry. The two planes defining the quadrants on Fig. 2b are conceptually interchangeable, and as a result, there exists an inherent indeterminacy between two possible physical solutions

Earthquake, Focal Mechanism,

Fig. 3 Examples of focal geometries with corresponding beachball patterns. (Adapted from Stein and Wyssession 2003)



Earthquake, Focal Mechanism, Fig. 4 A basketball has been divided into four quadrants alternately painted black and white. By rotating the basketball in space, one can obtain a representation of all possible focal orientations of the double-couple shown in Fig. 2b. This experiment demonstrates that earthquake focal geometries differ only by a solid rotation in space

obtained from any field of seismic data generated by a point source double-couple. For example, a left-lateral strike-slip earthquake on a vertical fault striking North–South is seismically indistinguishable from a right-lateral one on a fault striking East–West. This property, rooted in the symmetry of the double-couple, applies to all seismic waves, including P and S waves, surface waves, and the free oscillations of the Earth. The indeterminacy can be lifted only by considering a source extended in space, as opposed to a point source. In practice, this is done either by using field methods to recognize the orientation of the fault through its expression

on the ground, or by mapping the aftershocks of the event which are expected to occur along the direction of faulting, or by considering the so-called source-finiteness effects in the seismograms, which result from the interference of the elementary sources, distributed both in time and space, as the rupture propagates over the faulting area. In all cases, note again that the fault can no longer be represented as a point source.

Stress Release

In another interpretation of focal mechanisms, one can expand the double-couple as a two-dimensional deviatoric and symmetric tensor:

$$M_{ij} = M_0 \left(\hat{d}_i \hat{n}_j + \hat{n}_i \hat{d}_j \right) \quad (1)$$

where $\hat{\mathbf{d}}$ and $\hat{\mathbf{n}}$ are the unit vectors along the direction of slip, and perpendicular to the fault, respectively. These components and the principal directions of the moment tensor are directly related to those of the stresses released during the earthquake.

In particular, the bisector $\mathbf{T} = \frac{1}{\sqrt{2}} (\hat{\mathbf{d}} + \hat{\mathbf{n}})$ represents the axis of maximum tension and the direction of maximum positive amplitude of P waves; it plots at the center of the shaded quadrants on the beachball. The direction $\mathbf{P} = \frac{1}{\sqrt{2}} (\hat{\mathbf{d}} - \hat{\mathbf{n}})$ is the axis of maximum compression (often and improperly called “pressure axis”) and that of maximum negative amplitude of P waves; it plots at the center of the open quadrants. The direction $\mathbf{B} = \hat{\mathbf{d}} \times \hat{\mathbf{n}} = \mathbf{T} \times \mathbf{P}$ is the hinge of the fault planes; it is a direction where both P and S amplitudes vanish. For a given beachball, the $\mathbf{T}$, $\mathbf{P}$, $\mathbf{B}$ axes are independent of the choice between the two possible fault planes.

The interpretation of focal mechanisms in terms of stress release is particularly suited to the study of intraplate

earthquakes, which express the release of stresses accumulated inside the plates by tectonic forces, while interplate events, which are controlled by the relative displacements of two of the Earth's plates at a common boundary, are more readily interpreted in terms of a slip vector on a fault plane.

Fundamental Results

The most important application of focal mechanisms was the verification of the fundamental concepts of continental drift, seafloor spreading, and plate tectonics, in the mid-1960s. These ideas were derived using the spatial distribution of earthquake epicenters but without the knowledge of the geometries of their ruptures. The latter could then be used as an independent verification of the concepts, following the development of the beachball model, and the deployment of the World-Wide Standardized Seismograph Network in the early 1960s. Sykes (1967) and Isacks et al. (1968) achieved a remarkable endorsement of the mechanisms predicted by plate motions (normal faulting expressing extension at the mid-oceanic ridges, underthrusting expressing convergence at the subduction zones), and most spectacularly, verified the polarity of strike-slip faulting at the oceanic transform faults, thus upholding the concept proposed by Wilson (1965).

Focal mechanisms inside subducting slabs were compiled by Isacks and Molnar (1971), who documented a universal mechanism of down-dip compression for events deeper than 450 km, expressing the difficulty of penetration by the slab of the lower mantle at the bottom of the transition zone. By contrast, a wide diversity of mechanisms is found in the upper parts of the slabs, where local parameters such as rate and steepness of subduction, as well as possible bending stresses, control the stress field accumulated inside the slab during the subduction process.

A most remarkable aspect of the focal mechanisms of deep earthquakes is the fact that, without exception, they can be modeled by a double-couple (e.g., using its beachball representation), even though the concept of brittle rupture should not apply at the relevant pressure and temperature conditions inside the slabs. In particular, modern analysis of major deep earthquakes (Kawakatsu 1991; Okal 1996) has ruled out the possibility of an isotropic (implosive) source even as a mere component of their mechanism, as initially envisioned by Bridgman (1945) and proposed by Gilbert and Dziewonski (1975). Deep events represent faulting on a planar dislocation in the framework of Fig. 1, and differ from their shallow counterparts only in their usually higher level of stress drop (Antolik et al. 1999), but not in the geometrical or mechanical nature of their process. This observation supports the hypothesis that deep earthquakes may be induced by transformational faulting in volumes of metastable olivine persisting

inside the slabs at greater depths than predicted by equilibrium thermodynamics (Kirby et al. 1996).

Modern Focal Solutions: Inversion of the Moment Tensor

The double-couple used to represent a seismic source can be used, through its components given by Eq. (1), as a forcing term in the wave equations expressing the fundamental law of dynamics (a variation of " $\mathbf{F} = m \mathbf{a}$ "), to compute simulations of the Earth's response in space and time to the seismic disturbance. These are known as *synthetic seismograms*.

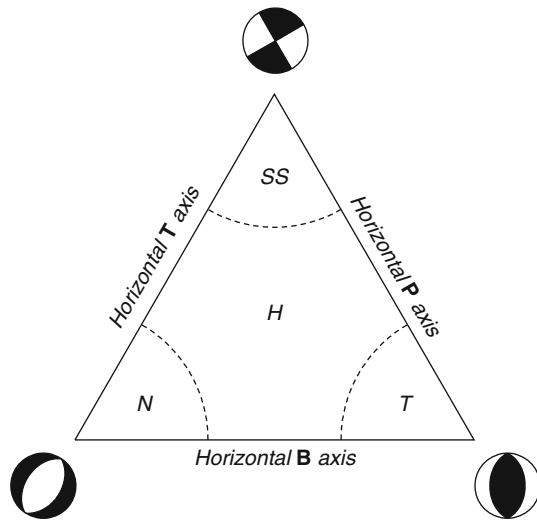
Conversely, and since the 1970s, inversion techniques have been used to derive the individual components of earthquake moment tensors directly from the waveforms of their recorded seismograms. Essentially, a best-fitting algorithm is used to express a large dataset of observed seismograms as a linear combination of synthetic seismograms computed for each individual component of the moment tensor. In addition, and for sufficiently large datasets, the inversion can resolve the best location in space of the source, and its result is known as a "centroid moment tensor" or CMT. These techniques, which can be applied on body waves, surface waves, or normal modes, as well as on regional phases in the near field, can be carried out automatically on digital data, and eliminate the occasional difficulty of picking a first motion polarity under noisy or emergent conditions. They have replaced the compilation of *P*-wave first motions as the primary source of earthquake focal mechanisms.

In particular, the Global CMT (*ex*-Harvard CMT) project, described initially by Dziewonski et al. (1981), has now inverted more than 30,000 earthquake sources since 1976, through a uniform, comprehensive procedure, into a catalog that has become the backbone of most modern seismotectonic studies.

Evaluation Tools for Double-Couple Solutions

Because of the relative complexity of the double-couple source that cannot be represented by a vector, a number of applications have been developed to illustrate their properties in a user-friendly representation. Among them, Frohlich and Apperson (1992) have proposed to plot focal mechanisms in a ternary diagram whose apices are the "*pure*" mechanisms corresponding to vertical stress axes (**B**, strike-slip; **T**, thrust; or **P**, normal); using barycentric coordinates within the triangle, the position of the mechanism indicates at a glance the nature of faulting (Fig. 5). Note that it remains independent of the azimuth ϕ of the fault strike.

In order to help compare different earthquakes, Kagan (1991) has offered an algorithm to quantify the proximity of



Earthquake, Focal Mechanism, Fig. 5 Ternary representation of focal mechanisms (Frohlich and Apperson 1992). This approach uses barycentric coordinates with respect to end members with one vertical principal axis. Mechanisms close to apices are either predominantly strike-slip (SS), thrust (T) or normal (N). Hybrid mechanisms (H) plot in the center of the diagram. By contrast, the sides of the diagram represent geometries with one horizontal axis. This representation is useful when studying trends among large families of mechanisms, such as intraplate earthquakes, irrespective of their particular strike angles

two focal solutions given by their angles ($\phi_1, \delta_1, \lambda_1$) and ($\phi_2, \delta_2, \lambda_2$). It computes the angle ω of the minimum solid rotation bringing one of the beachballs onto the other. Because of their symmetry, the maximum value of ω is 120° .

Non-double-Couple Solutions

The double-couple is a four-dimensional mathematical entity which constitutes a subset (featuring a zero-eigenvalue) but not a sub-vector space, of the five-dimensional vector space of deviatoric (i.e., zero-trace), symmetric second-order tensors. Consequently, linear inversions are carried out in the full five-dimensional space. The quality of fit of the inverted solution to a double-couple is assessed through a so-called compensated linear vector dipole (CLVD) parameter ϵ , expressing the ratio of the intermediate eigenvalue of the moment tensor to its largest one (in absolute value). In principle, ϵ could range from -0.5 to 0.5 and should be 0 for a perfect double-couple. The average value of $|\epsilon|$ for the whole CMT catalog is 0.12 , which constitutes a good verification of the concept of the double-couple, especially since it falls to 0.08 for earthquakes with moments $M_0 \geq 10^{26}$ dyn cm. Occasionally, larger values of $|\epsilon|$ (which can reach 0.4) indicate a deviation from the model of a double-couple, which may result from a number of scenarios. For example, earthquakes with complex ruptures consisting of multiple events with differing geometries will feature an overall deviatoric moment tensor, which however

cannot be described in terms of a lone double-couple. A recent example is the Samoa event of 29 September 2009 ($M_0 \approx 1.8 \times 10^{28}$ dyn cm; $\epsilon = 0.37$), which was modeled as a main outer rise normal faulting event triggering a smaller underthrusting shock on the nearby interplate contact (Li et al. 2009).

Other examples of non-double-couple solutions would include dyke injections during volcanic events (e.g., Tori-Shima Volcano, 13 June 1984; $\epsilon = 0.33$; Kanamori et al. 1993), rupture on nonplanar faults, again in a volcanic environment (Ekström 1994), and certain cryoseismic events believed to involve calving at the front of glaciers (e.g., 15 August 2008, $\epsilon = 0.33$; Nettles and Ekström 2009).

Conclusion

Focal mechanisms provide a mathematical or illustrative representation of the geometry of the rupture process during an earthquake. Their compilation forms the basis of our interpretation of earthquakes in the framework of global tectonics.

Cross-References

- [Earthquake Rupture: The Inverse Problem](#)
- [Earthquake Source Theory](#)

Bibliography

- Aki K, Richards PG (1980) Quantitative seismology. W. H. Freeman, New York, p 932
- Antolik M, Dreger DS, Romanowicz B (1999) Rupture processes of large deep-focus earthquakes from inversion of moment rate functions. *J Geophys Res* 104:863–894
- Bridgman PW (1945) Polymorphic transition and geological phenomenon. *Am J Sci* 243A:90–97
- Dziewonski AM, Chou A-T, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J Geophys Res* 86:2825–2852
- Ekström G (1994) Anomalous earthquakes on volcano ring-fault structures. *Earth Planet Sci Lett* 128:707–712
- Frohlich C, Apperson KD (1992) Earthquake focal mechanisms, moment tensors and the consistency of seismic activity near plate boundaries. *Tectonics* 11:279–296
- Gilbert JF, Dziewonski AM (1975) An application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra. *Philos Trans R Soc Lond* 278A:187–269
- Isacks BL, Molnar P (1971) Distribution of stresses in the descending lithosphere from a global survey of focal-mechanism solutions of mantle earthquakes. *Rev Geophys Space Phys* 9:103–174
- Isacks BL, Oliver J, Sykes LR (1968) Seismology and the new global tectonics. *J Geophys Res* 73:5855–5899
- Kagan YY (1991) 3-D rotation of double-couple earthquake sources. *Geophys J Int* 106:709–716
- Kanamori H, Ekström G, Dziewoński AM, Barker JS, Sipkin SA (1993) Seismic radiation by magma injection: an anomalous seismic event near Tori Shima, Japan. *J Geophys Res* 98:6511–6522

- Kawakatsu H (1991) Insignificant isotropic components in the moment tensor of deep earthquakes. *Nature* 351:50–53
- Kirby SH, Stein S, Okal EA, Rubie D (1996) Deep earthquakes and metastable mantle phase transformations in subducting oceanic lithosphere. *Rev Geophys Space Phys* 34:261–306
- Knopoff L, Gilbert JF (1959) Radiation from a strike-slip earthquake. *Bull Seismol Soc Am* 49:163–178
- Li X, Shao G, Ji C (2009) Rupture process of the $M_w = 8.1$ Samoa earthquake constrained by joint inverting teleseismic body, surface waves and local strong motion. *EOS Trans Am Geophys Union* 90(52):U21D–03. [Abstract]
- Nettles M, Ekström G (2009) Ice loss and glacial earthquakes in Greenland and Antarctica. *EOS Trans Am Geophys Union* 90(52):G52B–01. [Abstract]
- Okal EA (1996) Radial modes from the great 1994 Bolivian earthquake: no evidence of an isotropic component to the source. *Geophys Res Lett* 23:431–434
- Stein S, Wyssession ME (2003) An introduction to seismology, earthquakes and Earth structure. Blackwell, Malden, 498pp
- Sykes LR (1967) Mechanisms of earthquakes and nature of faulting on the mid-oceanic ridges. *J Geophys Res* 72:2131–2153
- Vvedenskaya AV (1956) Opredelenie polej smeshchenii pri zemletryasenyakh s pomoshchyu teorii dislokatsii. *Izv Akad Nauk SSSR Ser Geofiz* 6:277–284. [In Russian]
- Wilson JT (1965) A new class of faults and their bearing on continental drift. *Nature* 207:343–347

Earthquake, Foreshocks

Mian Liu

Department of Geological Sciences, University of Missouri, Columbia, MO, USA

Synonyms

Preshocks

Definition

Foreshocks. Smaller earthquakes preceding a large earthquake (the mainshock) in an earthquake sequence.

Introduction

Earthquakes typically occur in sequences that may include foreshocks, the mainshock (the largest event or events), and aftershock (► “[Earthquake, Aftershocks](#)”). Earthquake sequences without a clear mainshock are called swarms.

Foreshocks are the most obvious premonitory phenomenon of earthquakes. They are thought to indicate earthquake nucleation and hence may have the potential for short-term earthquake prediction (Scholz 2002). However, proof of the physical link between foreshocks and earthquake nucleation

is inconclusive, and earthquake prediction using foreshocks has not been reliable.

The problem begins with their recognition – foreshocks are earthquakes that are called foreshocks retrospectively, when a large event (the mainshock) followed. To be considered as foreshocks, these events need to occur within certain spatial and temporal windows, and need to be distinguished from background seismicity. Such spatiotemporal windows vary in different studies, and selecting foreshock events can be difficult, depending on the completeness of the earthquake catalogs available and other factors (Mignan 2014). Foreshocks are typically located within the mainshock’s source area with a dimension roughly the length of the fault rupture, which is in the order of 50–100 km for a magnitude 7.0 earthquake. Temporally, most foreshocks occur a few days to hours before the mainshock (Jones and Molnar 1979). In some cases, seismicity increases a few months or years before the mainshock. Some workers consider such increased seismicity foreshocks, others do not (Scholz 2002). Various studies suggest that 10–70% of large earthquakes may be preceded by foreshocks (Abercrombie and Mori 1996; Chen and Shearer 2016; Jones and Molnar 1979). A recent study found higher percentage of mainshocks associated with foreshock activity, using a high-resolution catalog in southern California that includes much more events, mostly small ones ($-2 < M < 0$), than previous catalogs (Trugman and Ross 2019). While the percentage varies with different studies and in different regions, it is important to note that not all earthquakes are preceded by foreshocks.

Characteristics of Foreshocks

Foreshocks, when they do occur, are typically few in comparison with aftershocks, sometimes consisting of only a few or a single event, hence are difficult for quantitative analysis. When considered collectively, they are shown to exhibit a temporal pattern similar to that of aftershocks (Jones and Molnar 1979; Shaw 1993). Most foreshocks occur within 10 days before the mainshock, and their numbers accelerate till the occurrence of the mainshock. The collective time sequence of foreshocks seems to fit an inverse Omori’s law:

$n(t) = at^{-q}$ where $n(t)$ is the number of foreshocks by time t before the origin of the mainshock, and a and q are constants, with q close to 1 (Kagan and Knopoff 1978; Papazachos 1975). Thus, the number of foreshocks tends to increase hyperbolically as time approaches the origin of the mainshock. Some studies suggest that foreshocks may have lower b -values (see entry ► “[Earthquake, Aftershocks](#)”) relative to aftershock sequences and background seismicity (Scholz 2002; Utsu 1966).

It is not clear why some earthquakes have foreshocks while others do not. Studies of global and regional earthquake catalogs have found that neither the time sequences nor the largest

magnitude of foreshocks correlate to the magnitude of the mainshock (Jones and Molnar 1979), suggesting either the size of the mainshock is independent of the earthquake nucleation process, or foreshocks are not part of earthquake nucleation (Abercrombie and Mori 1996). The incidence of foreshocks is found to be higher for shallower mainshocks and regions of higher heat flow (Abercrombie and Mori 1996; Chen and Shearer 2016; Trugman and Ross 2019). These results may suggest that foreshocks are related to small-scale heterogeneities in Earth's crust (Abercrombie and Mori 1996).

Causes of Foreshocks

The causes of foreshocks remain uncertain. Many workers view foreshocks as events related to earthquake nucleation. Evidence favoring this view includes laboratory experiments, which commonly show precursory slip events analogous to foreshocks (Bolton et al. 2019; Johnson et al. 2013). Theoretically, fault ruptures only when slip has occurred over a fault patch of some critical radius. Ohnaka (1992) has proposed a theoretical model that attributes the generation of foreshock to failure of asperities loaded by accelerating premonitory creep. This model predicts numerous features found in foreshocks: their occurrence typically a few days before the mainshock, their accelerated rates as time approaches that of the mainshock, and their occurrence in the immediate vicinity of the hypocenter of the mainshock. Some earthquake sequences in California show evidence of nucleation (Dodge et al. 1995). However, not all foreshocks readily fit the nucleation model. Aside from the obvious problem that many earthquakes do not have foreshocks, the nucleation model has difficulties explaining foreshocks that occur away from the rupture plane or have significantly different focal mechanisms from that of the mainshock (Jones et al. 1982).

Others think that foreshocks are fundamentally the same as other earthquakes (Felzer et al. 2004; Helmstetter et al. 2003). In this view, foreshocks, the mainshock, and aftershocks are just different names for earthquakes within the same sequence of cascading ruptures triggered by a common mechanism. The progression of a series of foreshock ruptures towards the mainshock hypocenter of the 1999 Izmit earthquake (Mw 7.6) in Turkey supports this view (Ellsworth and Bulut 2018).

Can Foreshocks Be Used to Predict Earthquakes?

Foreshocks are the most obvious precursor of earthquakes. If they are associated with earthquake nucleation, then they are potentially useful for short-term earthquake prediction. The best example is perhaps the 1975 Haicheng earthquake (magnitude 7.3) in China, which was predicted largely based on foreshocks, more than 500 of which were recorded within

4 days before the mainshock (Jones et al. 1982). However, an investigation by Wang et al. (2006) has revealed that the role of foreshocks in this prediction was more psychological than scientific: the jolts and damages from increased seismicity in the preceding months stressed people's minds, and the intensified foreshocks in the last day before the mainshock prompted some local officials to order an evacuation. In other places where official orders were not issued, the increased seismicity caused many residents to evacuate voluntarily.

Aside from the problem that many earthquakes do not have foreshocks, short-term earthquake prediction using foreshocks is challenging in theory and in practice. Even if foreshocks are physically linked to the process leading to the mainshock, the complexity of the linking processes in nature means that foreshocks probably will not help much for short-term earthquake prediction. On the other hand, if foreshocks are fundamentally the same as other earthquakes (Felzer et al. 2004; Helmstetter et al. 2003), then foreshocks cannot be predictors of mainshocks. In practice, we have yet to find any reliable ways to recognize foreshocks from background seismicity before the occurrence of the mainshock.

Nonetheless, there must be some symptoms associated with the stress buildup and physical property changes in and around the fault leading to a major rupture, and foreshocks are the most obvious symptoms known to us. Further studies of foreshocks will improve our understanding of the mechanics of earthquakes. There is increasing evidence of various creep slips before a mainshock (McGuire et al. 2005; Tape et al. 2018). On the other hand, by treating every earthquake as a potential foreshock that may be followed by a mainshock, statistical models based on the clustering properties of earthquake sequences can be developed to assess the real-time probability of damaging earthquakes, as has been done in California (Field et al. 2016; Gerstenberger et al. 2005; Jones 1985; Reasenber and Jones 1989).

Summary

Foreshocks are smaller earthquakes preceding a mainshock. They are thought to manifest the nucleation of the mainshock, hence have the potential for short-term earthquake prediction. However, not all large earthquakes are preceded by foreshocks, and they are difficult to recognize from the background seismicity before the occurrence of the mainshock. Foreshocks may also be fundamentally the same as other earthquakes, if so they cannot be used as earthquake predictors.

Cross-References

► [Earthquake, Aftershocks](#)

Bibliography

- Abercrombie RE, Mori J (1996) Occurrence patterns of foreshocks to large earthquakes in the western United States. *Nature* 381:303–307
- Bolton DC, Shokouhi P, Rouet-Leduc B, Hulbert C, Rivière J, Marone C, Johnson PA (2019) Characterizing acoustic signals and searching for precursors during the laboratory seismic cycle using unsupervised machine learning. *Seismol Res Lett* 90:1088–1098
- Chen X, Shearer PM (2016) Analysis of foreshock sequences in California and implications for earthquake triggering. *Pure Appl Geophys* 173:133–152
- Dodge DA, Beroza GC, Ellsworth WL (1995) Foreshock sequence of the 1992 landers, California, earthquake and its implications for earthquake nucleation. *J Geophys Res* 100:9865–9880
- Ellsworth WL, Bulut F (2018) Nucleation of the 1999 Izmit earthquake by a triggered cascade of foreshocks. *Nat Geosci* 11:531–535
- Felzer KR, Abercrombie RE, Ekström G (2004) A common origin for aftershocks, foreshocks, and multiplets. *Bullet Seismolog Soc Am* 94:88–98
- Field EH, Jordan TH, Jones LM, Michael AJ, Blanpied ML, Participants W (2016) The potential uses of operational earthquake forecasting. *Seismol Res Lett* 87:313–322
- Gerstenberger MC, Wiemer S, Jones LM, Reasenberg PA (2005) Real-time forecasts of tomorrow's earthquakes in California. *Nature* 435:328–331
- Helmstetter A, Sornette D, Grasso J-R (2003) Mainshocks are aftershocks of conditional foreshocks: how do foreshock statistical properties emerge from aftershock laws. *J Geophys Res* 108:2046
- Johnson P, Ferdowsi B, Kaproth B, Scuderi M, Griffa M, Carmeliet J, Guyer R, Le Bas PY, Trugman D, Marone C (2013) Acoustic emission and microslip precursors to stick-slip failure in sheared granular material. *Geophys Res Lett* 40:5627–5631
- Jones LM (1985) Foreshocks and time-dependent earthquake hazard assessment in southern California. *Bullet Seismolog Soci Am* 75:1667–1679
- Jones LM, Molnar P (1979) Some characteristics of foreshocks and their possible relationship to earthquake prediction and premonitory slip on faults. *J Geophys Res* 84:3596–3608
- Jones LM, Wang B, Xu S, Fitch TJ (1982) The foreshock sequence of the February 4, 1975, Haicheng earthquake ($M = 7.3$). *J Geophys Res* 87:4575–4584
- Kagan YY, Knopoff L (1978) Statistical study of the occurrence of shallow earthquakes. *Geophys J R Astron Soc* 55:67–86
- McGuire JJ, Boettcher MS, Jordan TH (2005) Foreshock sequences and short-term earthquake predictability on East Pacific rise transform faults. *Nature* 434:457–461
- Mignan A (2014) The debate on the prognostic value of earthquake foreshocks: a meta-analysis. *Sci Rep* 4:4099. <https://doi.org/10.1038/srep04099>
- Ohnaka M (1992) Earthquake source nucleation: a physical model for short-term precursors. *Tectonophysics* 211:149–178
- Papazachos BC (1975) Foreshocks and earthquake prediction. *Tectonophysics* 28:213–226
- Reasenberg PA (1999) Foreshock occurrence before large earthquakes. *J Geophys Res* 104:4755–4768
- Reasenberg PA, Jones LM (1989) Earthquake hazard after a mainshock in California. *Science* 243:1173–1176
- Scholz CH (2002) The mechanics of earthquakes and faulting, vol xxiv. Cambridge University Press, Cambridge/New York, p 471
- Shaw BE (1993) Generalized Omori law for aftershocks and foreshocks from a simple dynamics. *Geophys Res Lett* 20:907–910
- Tape C, Holtkamp S, Silwal V, Hawthorne J, Kaneko Y, Ampuero JP, Ji C, Ruppert N, Smith K, West ME (2018) Earthquake nucleation and fault slip complexity in the lower crust of Central Alaska. *Nat Geosci* 11:536–541
- Trugman DT, Ross ZE (2019) Pervasive foreshock activity across southern California. *Geophys Res Lett* 46(15):8772–8781
- Utsu T (1966) A statistical significance test of the difference in b-value between two earthquake groups. *J Phys Earth* 14:37–40
- Wang K, Chen QF, Sun S, Wang A (2006) Predicting the 1975 Haicheng earthquake. *Bullet Seismolog Soc Am* 96:757–795

Earthquake, Magnitude

Peter Borman
(Deceased)

Synonyms

Richter scale

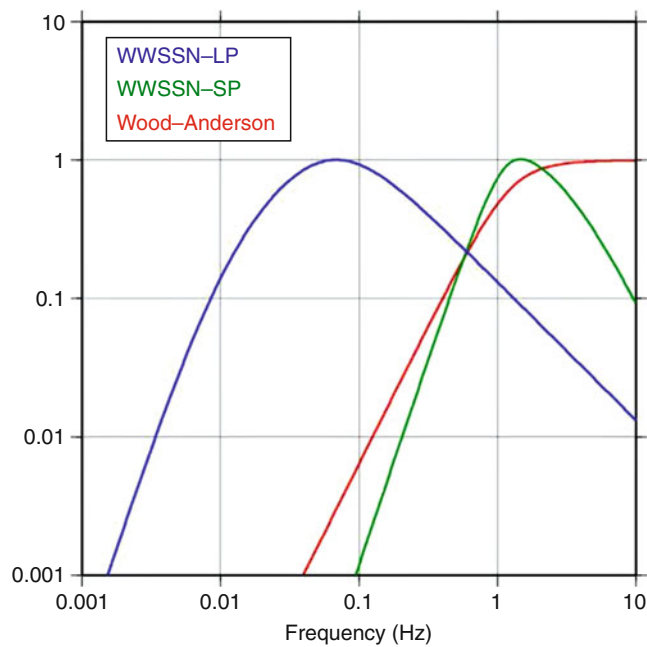
Definition

The *magnitude* of an earthquake is a number that characterizes the relative size or amount of elastic energy released by such an event (see ► [“Earthquakes, Energy”](#)). It is usually based on measurement of the maximum ground motion recorded by a seismograph (sometimes for a particular wave type and frequency) and corrected for the decay of amplitudes with epicentral distance and source depth due to geometric spreading and attenuation during wave propagation (see ► [“Propagation of Elastic Waves: Fundamentals”](#)).

Original Definition of Earthquake Magnitude

The first paper on earthquake magnitude was published by Charles F. Richter (1935), titled *An instrumental earthquake magnitude scale*. Therefore, often reference is made to the size or strength of earthquakes as measured on the *Richter scale*. Yet, this may be incorrect, especially for large earthquakes (magnitude saturation). Richter (1935, p. 1) wrote: “In the course of historical or statistical study of earthquakes in any given region it is frequently desirable to have a scale for rating these shocks in terms of their original energy, independently of the effects which may be produced at any particular point of observation.” The term *magnitude* scale was recommended to Richter by H.O. Wood in distinction to the name *intensity* scale (see ► [“Earthquakes, Intensity”](#)), which classifies the severity of earthquakes mainly on the basis of felt shaking or damage manifestations at different localities.

Richter compared records of a standard type of short-period horizontal component Wood-Anderson (WA) torsion

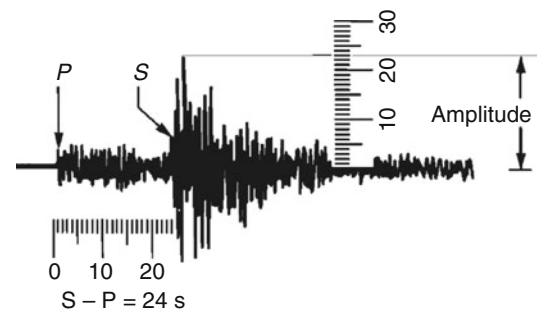


Earthquake, Magnitude, Fig. 1 Normalized amplitude– frequency responses of standard types of seismographs (or of their simulation filters) required for measuring several common types of magnitudes in agreement with recent IASPEI (2005) recommendations: Wood-Anderson seismograph and the short-estimated via the arrival period (SP) and long-period (LP) seismographs used in the former US World-Wide Standard Seismograph Network (WWSSN) (Seismic Instrumentation). (Reprint from IASPEI 2005. With copyright granted by IASPEI)

seismograph at different distances and azimuths from the earthquake source. All seismographs had an identical frequency response (red curve in Fig. 1) and a static magnification of (said to be) 2800. Many stations in Southern California were equipped with WA seismometers. Figure 2 shows a typical record of a near earthquake. Plotting for each event the logarithm of the recorded maximum trace amplitudes $A_{\max}$ over the epicentral distance Δ (see ► “Earthquakes, Location Techniques”) Richter derived an average attenuation law for a reference amplitude A_0 up to agreement with the above original definition $\Delta = 600$ km. He then corrected amplitudes measured at different stations by the $-\log A_0(\Delta)$ values. Thus measured amplitudes became comparable for a given event and differences of corrected amplitudes for different events indicative for difference in event size. In conclusion Richter (1935, p. 31) states:

The magnitude of a shock is defined as the logarithm of the calculated maximum trace amplitude, expressed in microns, with which the standard short-period torsion seismometer ($T_0 = 0.8$, $V = 2,800$, $h = 0.8$) would register that shock at an epicentral distance of 100 km.

Since then several other magnitude scales have been developed. They are based on different types of seismic waves (Bormann et al. 2002a; ► “Body Waves” and ► “Surface Waves”), recorded by seismographs with different frequency responses (e.g., those shown in Fig. 1) and measured at



Earthquake, Magnitude, Fig. 2 Record of a short-period Wood-Anderson seismograph of a local earthquake. P marks the onset of the first arriving longitudinal wave, and S that of the shear wave. Note the long tail of scattered coda-waves following S. From the time difference $S - P = 24$ s follows a distance of about 190 km. The largest amplitude is $A_{\max} = 23$ mm. With $-\log A_0(190 \text{ km}) = 3.45$ (Richter 1958) Eq. 1 yields $M_L = 4.8$. (Modified from Fig. 1 in Bormann and Saul 2009a. With copyright granted by Springer)

different periods and source distances. Yet, all had to be mutually scaled so that at certain magnitudes they yield values in agreement with the above original definition.

Local Magnitude Scales

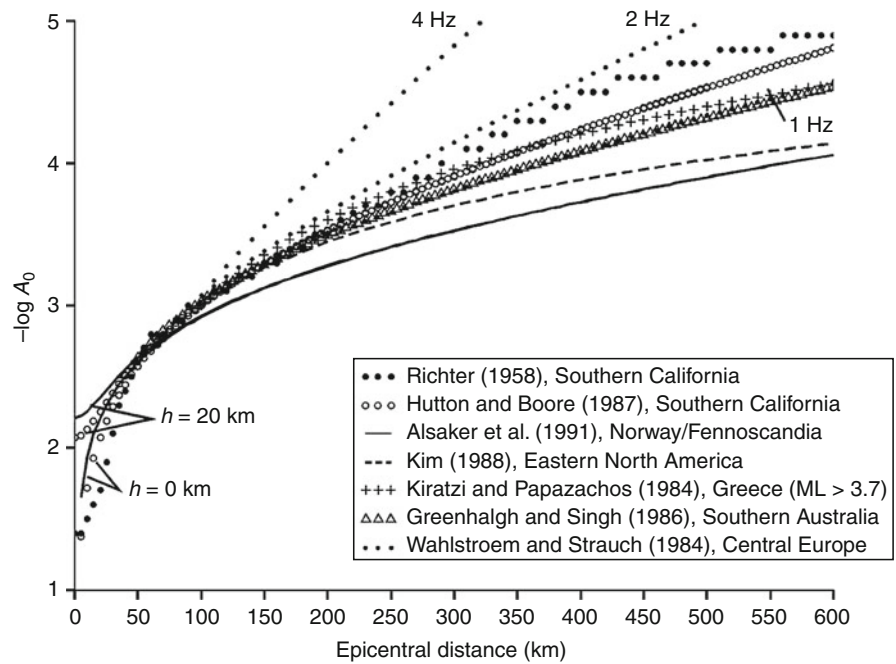
The *Richter scale* is termed M_L and applicable only to earthquakes recorded at local to near regional distances, typically less than 1,000 km distance. Its formula reads:

$$M_L = \log_{10} A_{\max} - \log A_0(\Delta). \quad (1)$$

Figure 2 shows how the trace amplitude $A_{\max}$ on a WA record is measured and the distance Δ estimated via the arrival-time difference between the P- and S-waves. Figure 3 is a compilation of different $-\log A_0(\Delta)$ calibration curves for different seismotectonic regions of the world. The Richter values for Southern California form the dot-step curve. Tabulated Richter values and calibration formulas for the other curves are given by Bormann (2002).

Figure 3 illustrates that M_L calibration functions may strongly differ from region to region because of differences in crustal structure, thickness, age, heat flow and related wave propagation and attenuation properties. Therefore, in order to assure compatibility of M_L values measured in different regions one has to (a) investigate the regional amplitude-attenuation conditions, (b) derive regional calibration functions, and (c) scale them so as to yield for identical amplitudes measured at 100 km *epicentral* distance the same M_L as with the Richter $-\log A_0(\Delta)$ value at this distance. However, in order to assure a better comparison of earthquakes in regions having very different attenuation of waves within the first 100 km, Hutton and Boore (1987) recommend to define local scales such that $M_L = 3$ corresponds to 10 mm trace

Earthquake, Magnitude,
Fig. 3 Calibration functions for M_L determination in different regions. (Reprint of Fig. 3.12 in Bormann et al. 2002b. With copyright granted by IASPEI)



amplitude on a Wood-Anderson instrument at 17 km *hypo-central* distance rather than 1 mm at 100 km *epicentral* distance. Further, empirical measurements proved that the static magnification of the WA seismographs is not 2,800 but around $2,080 \pm 60$ (Uhrhammer and Collins 1990).

Moreover, gain and response of modern seismographs may differ significantly from those of WA seismographs. Therefore, the Working Group on Magnitude Measurements of the International Association of Seismology and Physics of the Earth's Interior (IASPEI 2005) recommends for crustal earthquakes in regions with attenuative properties similar to those of Southern California the following standard formula:

$$M_L = \log_{10} A_{\max} + 1.11 \log_{10} R + 0.00189 * R - 2.09 \quad (2)$$

where $A_{\max}$ = maximum trace amplitude in nm that is measured on the output from a horizontal-component instrument that is filtered so that the response of the seismograph/filter system replicates that of a Wood-Anderson standard seismograph but with a static magnification of 1, and R = hypocentral distance in km, typically less than 1,000 km.

For crustal earthquakes in regions with different attenuative properties and for measuring magnitudes with vertical-component seismographs, the standard equation is of the form:

$$M_L = \log_{10}(A_{\max}) + C(R) + D \quad (3)$$

where $A_{\max}$ and R are defined as in Eq. 2, except that A may be measured from a vertical-component instrument, and where $C(R)$ and D have been calibrated to adjust for the different regional attenuation and for any systematic differences

between amplitudes measured on vertical instead of horizontal components. Besides M_L scales that measure $A_{\max}$, typically of relatively weak motion, also M_L^{SM} scales have been developed, which are based on peak ground accelerations (e.g., Lee et al. 1990; Hatzidimitriou et al. 1993; *Earthquakes, Strong-Ground Motion*). Also the duration d of the event record or of its coda length c after the S onset can be used for estimating magnitude from locally recorded data (Bormann et al. 2002b). Further, there exist regression relationships which relate magnitude values to other observed parameter data such as (see ► “Earthquakes, Intensity”). But any such alternative local scale has to be developed individually for different seismic networks or regions because of regionally variable attenuation, scattering, signal-to-noise and other conditions. Also, for assuring the compatibility of their values these scales should be calibrated with amplitude-based M_L .

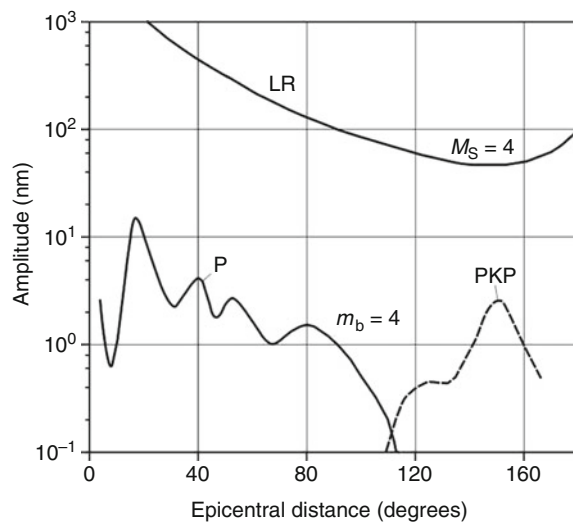
Teleseismic Magnitude Scales

Surface-Wave Magnitudes

Beno Gutenberg modified the Richter scale for application in the teleseismic range ($\Delta \geq 20^\circ$; $1^\circ = 111.22$ km) by using both body and surface waves. For shallow earthquakes the latter have by far the largest amplitudes in seismic records (Fig. 4). The Gutenberg (1945a) formula for surface wave magnitude is

$$M_s = \log A_{H\max} + 1.656 \log \Delta + 1.818 \quad (4)$$

with $A_{H\max}$ = vectorially combined maximum horizontal surface-waves displacement amplitude in micrometer



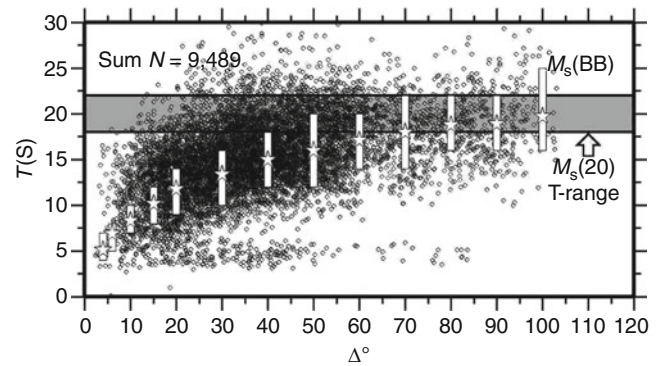
Earthquake, Magnitude, Fig. 4 Amplitude decay with distance for short-period (1 Hz) P- and PKP-waves as well as long-period (≈ 20 s) Rayleigh surface waves LR radiated by an earthquake of magnitude 4. (Reprint of Fig. 3.13 in Bormann et al. 2002a. With copyright granted by IASPEI)

measured at periods T around 20 s and the term $1.656 \log \Delta$ which compensates for the amplitude decay with distance in the range $20^\circ \leq \Delta < 140^\circ$. Equation 4 is best scaled for surface waves with oceanic travel paths.

Later, a group of Czech and Russian seismologists investigated the propagation of surface waves with dominantly continental travel paths, using records of seismographs of type Kirnos that are proportional to ground displacement in a wide range of periods. They found that surface waves may have their largest amplitudes in a much wider range of periods between about 3 and 30 s, as confirmed by Bormann et al. (2009); (see Fig. 5). Moreover, estimates of M_s proved to be more stable in a wider range of epicentral distances between 2° and 160° if one measures instead of $A_{\max}$ the ratio $(A/T)_{\max}$ (Vaněk et al. 1962). This ratio relates to ground velocity $V_{\max} = 2p(A/T)_{\max}$ and thus to the energy contained in the respective wave group. Therefore, these authors proposed a modified M_s formula, which is IASPEI standard for M_s determination since 1967:

$$M_s = \log (A/T)_{\max} + 1.66 \log \Delta + 3.3. \quad (5)$$

National and global seismic networks and data centers run or initiated by the USA accepted Eq. 5. However, with the introduction of the WWSSN they did not avail of medium- to long-period broadband displacement records that were standard equipment at first-rate stations over most of the Eurasian territory. Also, they continued to measure $A_{\max}$ only at periods around 20 s rather than true $(A/T)_{\max}$ in the whole surface-wave train. But then the use of Eq. 5 results in significantly distance-dependent M_s estimates (see discussion in Bormann



Earthquake, Magnitude, Fig. 5 Variability with distance Δ of periods associated with $(A/T)_{\max}$ in vertical component surface waves records. The gray shaded stripe marks the period range between 18 and 22 s, the light-gray columns mark the observed range of average periods as published by Vaněk et al. (1962), and the stars the average periods for different increments of the data cloud. (Reprint of Fig. 8 by Bormann et al. 2009, p. 1874. With copyright granted by the Seismological Society of America)

et al. 2009). Proposed new calibration functions for 20 s surface-waves, e.g., by von Seggern (1977), Herak and Herak (1993) and Rezapour and Pearce (1998) have not yet been accepted as global standards. In contrast, Kirnos-based M_s measurements in accordance with formula 5 have ever since been carried out in the former Soviet Union and its allied countries, in the now Commonwealth of Independent States (CIS), and China.

Thus, two traditions of M_s measurements have been practiced for half a century, producing a host of valuable but not fully compatible data. Yet, the generic nomenclature M_s does not allow to discriminate between these two kinds of data. Moreover, M_s is measured by different agencies still in sometimes different ranges of source-depth, on different record components of different types of seismographs and sometimes even using different calibration functions, e.g., the International Data Centre (IDC) of the Comprehensive Test-Ban Treaty Organization (see ► “Seismology, Monitoring of CTBT”). This may result in procedure-dependent errors that have to be reduced by standardization of both nomenclature and procedures. This task is eased by the fact that modern digital broadband records allow to accurately synthesize records of classical seismographs, on which the original definition of magnitude scales often rests. Moreover, modern broadband records are mostly proportional to velocity and allow to measure $(A/T)_{\max}$ directly and unambiguously on unfiltered traces. Therefore, IASPEI (2005) recommends for shallow earthquakes with source depth $h < 60$ km in future two surface-wave magnitude standards, namely, (a) a purely teleseismic one in the range $20^\circ \leq \Delta \leq 160^\circ$, termed $M_s(20)$, or when written in the International Seismological Format (ISF):

$$M_{S_20} = \log_{10}(A/T) + 1.66 \log_{10}\Delta + 0.3 \quad (6)$$

where A = vertical-component ground displacement in nanometer measured from the maximum surface-wave trace-amplitude having a period between 18 and 22 s on a waveform that has been filtered so as to replicate the record of the WWSSN-LP seismograph (Fig. 1), and (b) another broadband (BB) surface-wave magnitude that is measured in a wider distance range of $2^\circ \leq \Delta \leq 160^\circ$, termed $M_s(BB)$, or in the ISF format:

$$M_{s_BB} = \log_{10}(A/T)_{\max} + 1.66 \log_{10}\Delta + 0.3 \quad (7)$$

where $(A/T)_{\max} = (V_{\max}/2\pi)$, with $V_{\max}$ = ground velocity in nm/s associated with the maximum trace-amplitude in the surface-wave train recorded on a vertical-component seismogram that is proportional to velocity, and where the period T is between $3 \text{ s} < T < 60 \text{ s}$. $M_s(BB)$ generally agrees well with $M_s(20)$, however, for weaker regional and local earthquakes it may yield up to about 0.5 magnitude units (m.u.) larger values than $M_s(20)$ (Bormann et al. 2009). Besides these two standard M_s there exist other surface-wave magnitude scales. E.g., Bonner et al. (2006) proposed a largely theory-based $M_s(V_{\max})$ that requires multi-bandpass filtering and frequency-dependent attenuation corrections. Another scale, termed Mm (Okal and Talandier 1989), uses very long-period

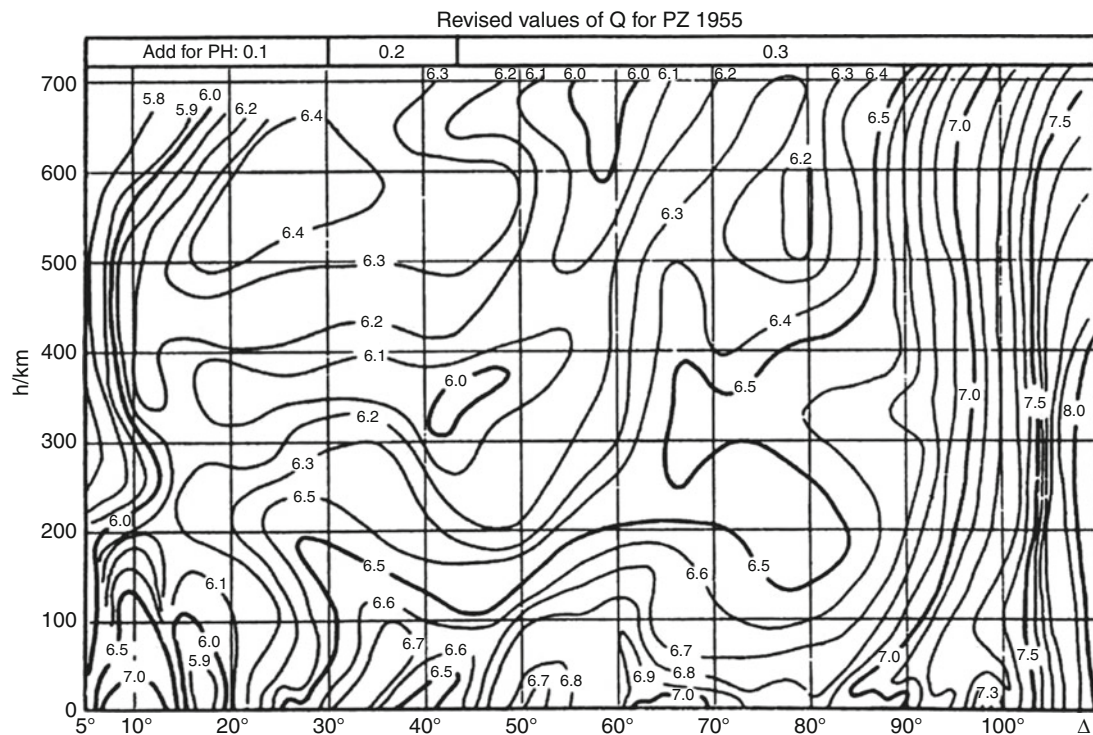
surface-wave amplitudes with periods between 60 and 400 s, which penetrate into the earth mantle. Mm does not saturate (magnitude saturation) and thus is well suited for tsunami warning (e.g., Weinstein and Okal 2005).

Body-Wave Magnitudes

Gutenberg (1945b, c) teleseismic magnitude scales for body-waves are applicable also to deep earthquakes down to source depths $h \approx 700 \text{ km}$. Diagrams and tables with calibration values $Q(\Delta, h)$ for P, PP, and S waves (see ► “Seismogram Interpretation”), covering the distance range between about 5° and 170° , were published by Gutenberg and Richter (1956) (see also Bormann 2002). Figure 6 shows the Q -values for vertical component P waves. It is the only classical body-wave scale which is still widely used. The general body-wave magnitude formula, applicable in the range $20^\circ \leq \Delta \leq 100^\circ$ reads, when A is measured in micrometer:

$$m_B = \log_{10}(A/T)_{\max} + Q(\Delta, h). \quad (8)$$

Gutenberg derived the Q -functions by analyzing mostly medium- to long-period body-wave amplitudes with periods $2 \text{ s} < T < 30 \text{ s}$ (Abe 1981). Therefore he did not account for frequency-dependent attenuation. Yet, the latter becomes significant for P-waves with $T < 4 \text{ s}$ (Di Giacomo et al. 2008). Further, Gutenberg measured $(A/T)_{\max}$ within the whole



Earthquake, Magnitude, Fig. 6 Calibration values $Q(\Delta, h)$ for vertical (Z) component P-wave amplitudes as a function of epicentral distance Δ in degree and source depth h in kilometer. They are used for the

calculation of body-wave magnitudes m_B and m_b . (Figure redrawn from Gutenberg and Richter 1956. Reprint of Fig. 1a in Bormann 2002. With copyright granted by IASPEI)

P-wave train and not within a fixed short time window after the P-wave onset. This was not taken into account when introducing with the WWSSN in the 1960s a short-period P-wave magnitude m_b . It measured $A_{\max}$ on narrow-band WWSSN-SP records (Fig. 1) at periods < 3 s (mostly around 1 s) at fixed time-windows. The main advantage was the improved signal-to-noise ratio. It allowed teleseismic magnitudes to be determined down to $m_b \approx 4$. m_b became standard in all western countries, which discontinued to measure m_B . Yet, on average for values above 5.5, m_b tends to be smaller than other kinds of magnitudes (see Fig. 7). Reasons are discussed in the section on “Magnitude “Saturation””.

Measurement procedures for m_b have been rather inconsistent, using different frequency responses, measuring the amplitudes within different time windows, applying different calibration functions (e.g., the IDC). The detrimental effect of such inconsistencies became strikingly evident after the great $M_w 9.3$ Sumatra earthquake of 26 December 2004. The IDC reported $m_b = 5.7$, CENC 6.3, NEIC 7.0 and the German Regional Seismic Network (GRSN) 7.5. The IDC and CENC measured $A_{\max}$ within 5 s, the NEIC within 60 s, and the GRSN some 90 s after the first P onset. Such differences for the same event on a logarithmic scale are not tolerable and necessitate standardization of procedures. Broadband $m_B = 8.3$ was still larger than the largest m_b value.

Based on such experience IASPEI (2005) proposes two complementary teleseismic body-wave magnitude standards:

$$m_b = \log_{10}(A/T) + Q(\Delta, h) - 3.0 \quad (9)$$

where: A = P-wave ground amplitude in nm, calculated from the maximum trace-amplitude in the entire P-phase train (time spanned by P, pP, sP, possibly PcP and their codas, and ending preferably before PP; *Seismogram Interpretation*); period $T < 3$ s, Δ in the range $20^\circ \leq \Delta \leq 100^\circ$ and h between 0 and 700 km. T and A are measured on the output of a vertical-component instrument that is filtered so that the frequency response of the seismograph/filter system replicates that of a WWSSNSP seismograph (Fig. 1). This new definition of m_b yields for great earthquakes significantly larger values than common procedures. None the less, short-period m_b should – at least for $m_b > 5.5$ – be complemented by broadband $m_B(BB)$, which reads in ISF format:

$$m_{B-BB} = \log_{10}(A/T)_{\max} + Q(\Delta, h) - 3.0, \quad (10)$$

where: $(A/T)_{\max} = (V_{\max}/2p)$, with $V_{\max}$ = ground velocity in nm/s associated with the maximum trace-amplitude in the entire P-phase train (see m_b) as recorded on a vertical-component seismogram that is proportional to velocity at least in the period-range $0.2 \text{ s} < T < 30 \text{ s}$; $Q(\Delta, h)$, Δ and h ranges are as for m_b .

Besides these two body-wave standards there exist several other broadband P-wave magnitudes. Most of them aim at fast estimates of the moment magnitude M_w (non-saturating magnitude scales).

Relationships Between Magnitude Scales and Released Seismic Energy E_s

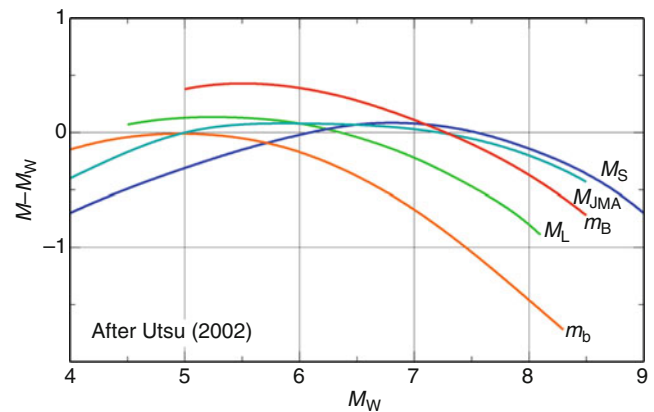
There is no “true” single magnitude parameter which can be sufficiently reliable and completely characterize both size and shaking strength of earthquakes and thus related *Seismic hazard*. Depending on the measured kinematic parameters and their frequency range the magnitudes relate to different geometric, kinematic, or dynamic aspects of the (see ► “Earthquake Rupture: The Inverse Problem”). Accordingly, the scales discussed so far do not scale 1:1, agree in some value ranges but differ in others. Although Fig. 7 hints to sometimes nonlinear relationships, the most important ones are reasonably linear within the range of data scatter, e.g., that of Gutenberg and Richter (1956).

$$m_B = 0.63M_s + 2.5. \quad (11)$$

Equation 11 is well reproduced by the new standards $m_B(BB)$ and $M_s(BB)$ (Bormann et al. 2009) but differs strongly from the $m_b - M_s(20)$ relationship (Gordon 1971):

$$m_b = 0.47M_s(20) + 2.79. \quad (12)$$

Despite the shortcomings of narrow-band m_b and $M_s(20)$ they are particularly useful for discriminating earthquakes and Nuclear Explosions.



Earthquake, Magnitude, Fig. 7 Average relationships of different common types of magnitude scales with the moment magnitude M_w (non-saturating magnitude scales). M_{JMA} is determined by the Japanese Meteorological Agency. (Modified from Fig. 1 of Utsu (2002). Reproduction of Fig. 4 in Bormann and Saul (2009a). With copyright granted by Springer)

Gutenberg and Richter (1956) published also a semiempirical relationship between m_B and seismic energy E_s :

$$\log E_s = 2.4m_B - 1.2 \text{ (with } E_s \text{ in units of Joule)}. \quad (13)$$

When inserting Eqs. 11 into 13 one gets

$$\log E_s = 1.5M_s + 4.8. \quad (14)$$

M_s in all these formulas means $M_s(20)$. But, 20 s displacement amplitudes are not very representative for the radiated seismic energy, especially not for smaller earthquakes. Another relationship has been derived by Kanamori et al. (1993) between M_L and $\log E_s$ for Southern California, valid in the range $1.5 < M_L < 6.0$:

$$\log E_s = 1.96M_L + 2.05. \quad (15)$$

But any such relationship allows only rough estimates of E_s within about an order of magnitude. This uncertainty can nowadays be reduced, down to a factor of about 3–5, by integrating squared velocity broadband records with good signal-to-noise ratio in the teleseismic range and thus measuring directly (see ► “Earthquakes, Energy”).

Magnitude “Saturation”

Magnitude saturation is a widely but not uniquely used term in seismological literature. Originally, one assumed that classical magnitudes do not increase further beyond some value of a reference magnitude of earthquake size such as M_w (non-saturating magnitude scales) (e.g., Geller 1976; Kanamori 1983). However, such a complete saturation is not supported by Figs. 7 and 8 and recently published data (e.g., Bormann et al. 2009). We therefore may speak of magnitude “saturation,” or more correct underestimation of earthquake size, also if one magnitude M_x increases slowly with respect to increases in M_w (see Fig. 7). Such saturation may occur when the earthquake rupture duration is much longer than the periods at which $A_{\max}$ or $(A/T)_{\max}$ are measured.

This is also illustrated by Fig. 8, calculated for a common ω^{-2} source model (Aki 1967), assuming an average rupture velocity, stress drop in the source volume and ratio E_s/M_0 . For details see Bormann et al. (2009). The spectra have been plotted together with the periods, respectively period ranges, at which m_b , $m_B(BB)$, $M_s(20)$ and $M_s(BB)$ are measured. The left panel is scaled to $M_0 = \mu D F$ (where F = rupture area, D = average displacement over the fault and μ = average rigidity of the rock material in the source volume) and the right panel to the time derivative of M_0 (moment rate), which is proportional to far-field ground motion velocity. M_0 can only be estimated correctly when displacement amplitudes are

measured at frequencies less than the corner frequency f_c (left panel) and seismic energy, which is proportional to the squared ground motion velocity, only when velocity spectral amplitudes are measured sufficiently wide around f_c . Accordingly, m_b measured around 1 Hz necessarily begins (on average) to underestimate both M_0 and E_s for $M_w > 5$ to 5.5, in contrast to the broadband magnitudes, which are more closely related to M_0 and E_s up to M_w 7.5–8. The spectral reasons for underestimating earthquake size or energy release may be aggravated by too short a measurement time window, which was rather common for m_b . Figure 9 illustrates both the spectral and timewindow component of magnitude saturation.

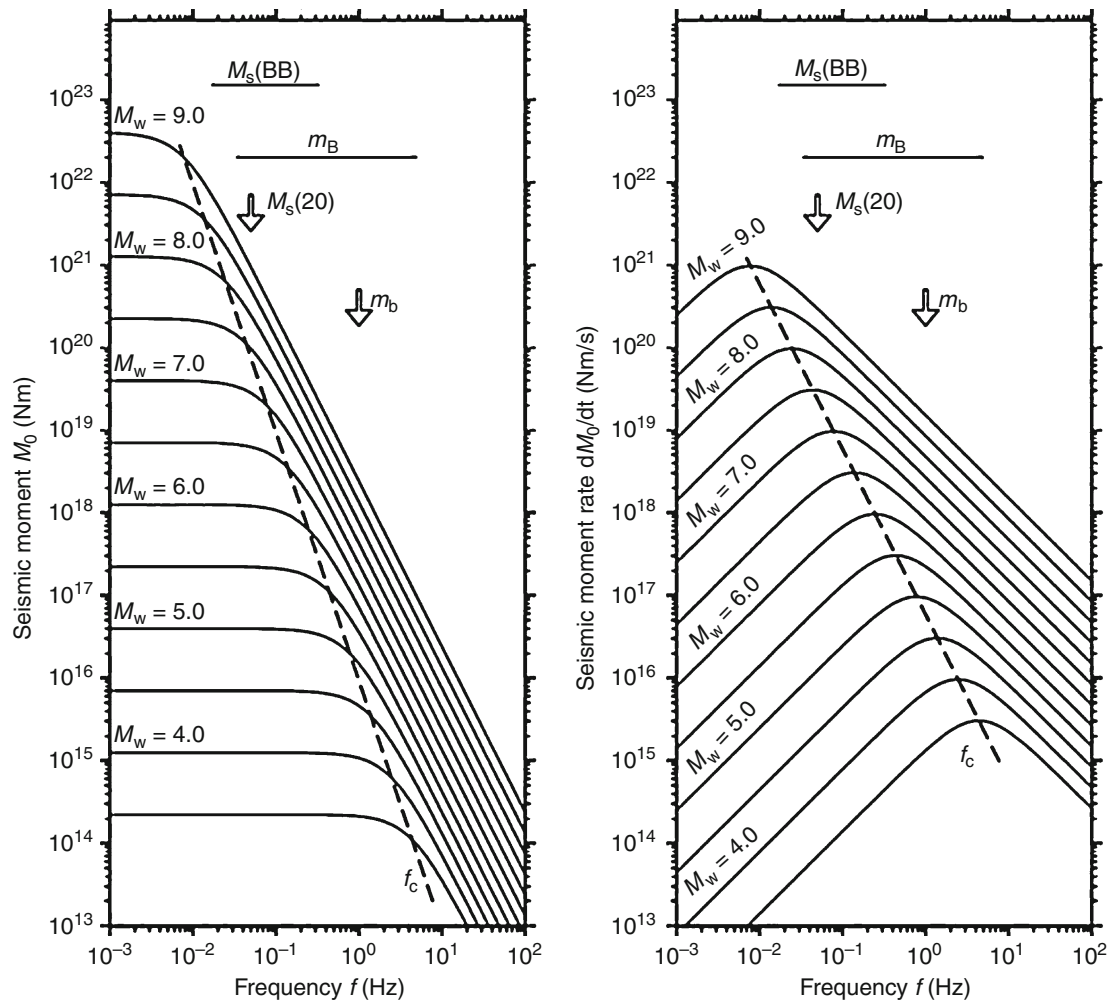
Non-Saturating Magnitude Scales

Kanamori (1977), Purcaru and Berckhemer (1978) and Hanks and Kanamori (1979) proposed a non-saturating magnitude scale for very large earthquakes, based on M_0 data. Kanamori termed it M_w . For converting M_0 into M_w Kanamori (1977) replaced in Eq. 14 M_s by M_w and assumed an average ratio $E_s/M_0 = 5 \times 10^{-5} = \text{constant}$. This led to a relationship, which reads in the IASPEI (2005) recommended standard form with M_0 in units of Nm:

$$M_w = (2/3)(\log_{10} M_0 - 9.1). \quad (16)$$

M_0 calculations are nowadays routine at major seismological centers, either by measuring the displacement amplitude u_0 on the long-period asymptote (plateau) of the wave spectrum, which is proportional to M_0 after correcting for the source radiation and wave propagation effects (as done in Fig. 8), or by fitting long-period synthetic seismograms for different source models best to long-period filtered real seismic records, using wavelengths that are larger than the rupture length (Dziewonski et al. 1981). However, if this condition is not fulfilled, as sometimes in routine computations, then also M_w may be underestimated, e.g., for the great 2004 Sumatra earthquake values between 8.2 and 9.3 were reported. The last value resulted from using either extremely long-period free oscillation data (Stein and Okal 2005) or from multiple CMT source analysis (Tsai et al. 2005). For all major earthquakes with $M_w > 5.5$ the Global Centroid Moment Tensor (GCMT) project publishes moment tensor and M_w solutions usually within a few hours after the event (<http://www.globalcmt.org/CMTsearch.html>). Regional centers, such as the Swiss Earthquake Service, may provide – with larger time delay – such data down to $M_w \approx 3.0$ when dense quality networks close to the source are available (Braunmiller et al. 2002).

Besides calculation of M_w via direct measurements of M_0 there exist several methods of fast M_w proxy estimates, e.g., M_{wp} in the context of tsunami early warning (Tsuboi et al. 1999). M_{wp} scales nearly 1:1 with broadband m_B (Bormann



Earthquake, Magnitude, Fig. 8 “Source spectra” of ground displacement amplitudes A for an “average” seismic shear source as a function of frequency f , scaled to M_0 and M_w . *Right*: The same as left, but for ground motion velocity amplitudes V , scaled to seismic moment rate and M_w . The maximum of seismic energy $E_s \sim V^2$ is radiated around f_c . The open arrows point to the center frequencies on the abscissa at which m_b and

$M_s(20)$ are measured. The horizontal bars mark the frequency range within which $V_{\max}$ of P- and Rayleigh-waves for $mB(BB)$ and $M_s(BB)$ should be measured. (Courtesy of Di Giacomo. Reproduced from Fig. 1 of Bormann et al. 2009, p. 1870. With copyright granted by the Seismological Society of America)

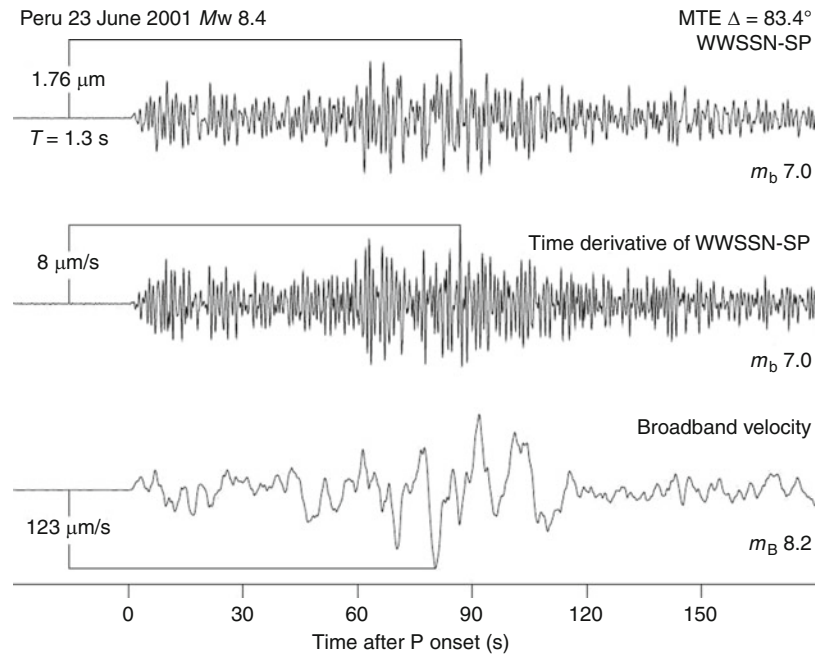
and Saul 2009a) but also tends to underestimate very great and (see ► “Slow Earthquake”). More advanced fast and non-saturating versions of broadband P-wave magnitudes are M_{wpd} by Lomax and Michelini (2009), the (unspecified) M by Hara (2007), and m_{Bc} , a modified “cumulative” m_B procedure for great multiple-rupture earthquakes (Bormann and Saul 2009b). All these procedures account additionally for rupture duration, estimated in real time. Another very perspective promising new method for rapid non-saturating moment tensor and M_w estimates uses very long-period W-phase observations between the P-wave and S-wave arrivals (Kanamori and Rivera 2008).

However, both M_0 and M_w are only long-period or quasi-static measures of earthquake size and thus of the tectonic earthquake effect. More relevant for assessing seismic hazard

in terms of the earthquake potential for causing shaking damage are estimates of the amount of released seismic energy E_s and of the corner frequency f_c . These two parameters largely control the relative amount of radiated high-frequency velocity amplitudes. This necessitates, complementary M_w , also an energy-scaled magnitude M_e , which better relates to the kinematics and dynamics of the earthquake rupture, such as rupture velocity and stress drop. Velocity-broadband records allow nowadays to estimate (see ► “Earthquakes, Energy”) directly by integrating squared velocity amplitudes over a wide range of periods around the peak of the velocity spectrum (see Fig. 8) and correcting them for wave propagation effects. Comparing so determined E_s values for hundreds of shallow earthquakes with their NEIC M_s magnitudes in the range $5.5 < M_s < 8.5$ Choy and Boatwright (1995) found

Earthquake, Magnitude,

Fig. 9 Plot of vertical-component P-wave records of an $M_w 8.4$ Peru earthquake at station MTE at $\Delta = 83.4^\circ$. It illustrates the difference between broadband m_B and short-period m_b for large earthquakes. *Top trace*: m_b based on measuring $A_{\max}$ and the related period T from the simulated WWSSN_SP record. *Second trace*: Time derivative of the top trace for making $V_{\max} = 8 \mu\text{m/s}$ for m_b , comparable with $V_{\max} = 123 \mu\text{m/s}$ on the *bottom trace*. The latter yields $m_B = 8.2$, in contrast to IASPEI standard $m_b = 7.0$, or 6.4 , when $A_{\max}$ is measured within the first 5 s. (Copy of Fig. 1 in Bormann and Saul 2008, p. 699. With copyright granted by the Seismological Society of America)



$$\log E_s = 1.5M_s + 4.4. \quad (17)$$

Equation 17 differs from Eq. 14 that was used for deriving the M_w relationship 16 only in the constant. Replacing in Eq. 17 M_s by M_e and resolving it for M_e yields

$$M_e = (2/3)(\log E_s - 4.4). \quad (18)$$

Di Giacomo et al. (2010b) showed that their $M_e(\text{GFZ})$ and the related $M_w(\text{GCMT})$ agree on average in the range $5.5 < M_w < 9$ within 0.1 m.u., with a standard deviation $SD = 0.25$. The respective deviations between $M_w(\text{GCMT})$ and $M_e(\text{NEIC})$ are larger. The reason is that $M_e(\text{NEIC})$ is calculated from E_s values to which theoretically derived source mechanism corrections have been applied. In contrast, $M_e(\text{GFZ})$ follows the tradition of classical magnitudes with no such corrections made to measured $A_{\max}$ or $V_{\max}$. The pros and cons of such corrections and of their size under real Earth conditions of wave radiation and propagation are still a matter of debate and require further research.

M_e and M_w may differ for the same earthquakes up to about 1 m.u. This is mainly due to the fact that earthquake stress drop and the related ratio E_s/M_0 are not constant but may vary more than two to three orders. Examples of striking differences in M_e and M_w have been published, e.g., by Choy and Kirby (2004) and by Di Giacomo et al. (2010a). Suitable seismotectonic setting provided, $M_e \ll M_w$ may be indicative for high tsunamigenic potential, and vice versa, $M_e \gg M_w$ for high shaking damage potential (see also Fig. 11 and related discussion).

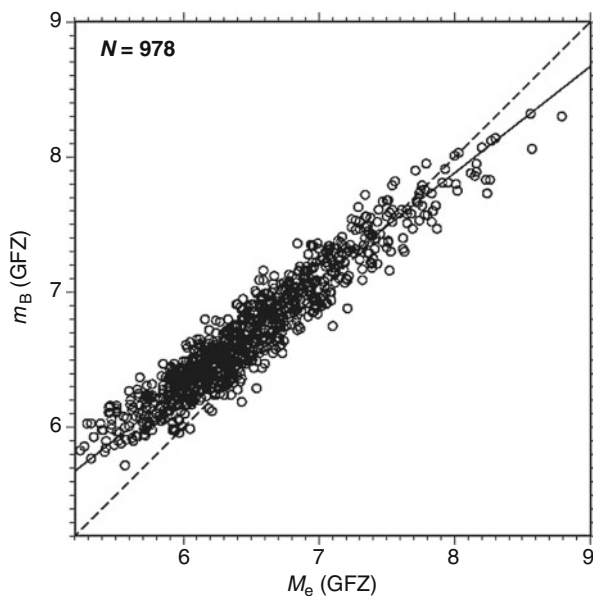
Relationship Between M_w , M_e , and Classical Magnitudes

Scordilis (2006) published global regression relationships, based on many thousands of NEIC and ISC (International Seismological Centre) m_b and M_s values, with M_w values of the US Geological Survey and of Harvard. First regression relationships between IASPEI standard magnitudes with non-saturating M_w , M_e , and m_{Bc} have been derived by Bormann and Saul (2008, 2009a, b) and by Bormann et al. (2009). From the classical magnitudes, M_s relates for strong earthquakes best with M_w , tends to be smaller, however, for $M_w < 6.2$ – 6.5 and > 7.8 – 8.5 (see Fig. 7; Scordilis 2006; Ekström and Dziewonski 1988). Yet, $M_s(\text{BB})$ reduces to half the systematic underestimation of M_w by $M_s(20)$ towards smaller magnitudes (Bormann et al. 2009). The $M_s(20)$ bias had been predicted already theoretically by Kanamori and Anderson (1975).

The first magnitude-energy Eq. 13 had been defined with m_B . Automatically determined $m_B(\text{GFZ})$, which agrees with IASPEI standard $m_B(\text{BB})$, correlates very well with $M_e(\text{GFZ})$ (Fig. 10), allowing estimates of M_e via simple m_B measurements with a $SD = \pm 0.18$ m.u. Both magnitudes are based on ground motion velocity and therefore reflect – when compared with M_w – in a similar way differences in source properties.

Figure 11 presents three records of equal M_w (within 0.1 m.u.), however very different rupture durations and m_B values. The Indian intra-plate Budj earthquake had the shortest rupture duration of about 30 s, the most impulsive waveform and largest $m_B = 8.2$, 0.6 m.u. larger than M_w .

This event caused tremendous shaking damage. In contrast, the two very slow Nicaragua and Java earthquakes with rupture durations around 200 s were (see ► [“Tsunami”](#)) earthquakes. Their m_B values were 0.5 and 0.7 m.u. smaller than M_w . Both earthquakes were not or only weakly felt at the relatively nearby coasts.



Earthquake, Magnitude, Fig. 10 Relationship between automatically determined values of m_B and M_e according to procedures developed at the GFZ German Research Centre for Geosciences by Bormann and Saul (2008) and Di Giacomo et al. (2010). The broken line is the 1:1 and the solid line the orthogonal regression relation $m_B = 0.79M_e + 1.58$. (Courtesy of D. Di Giacomo and J. Saul)

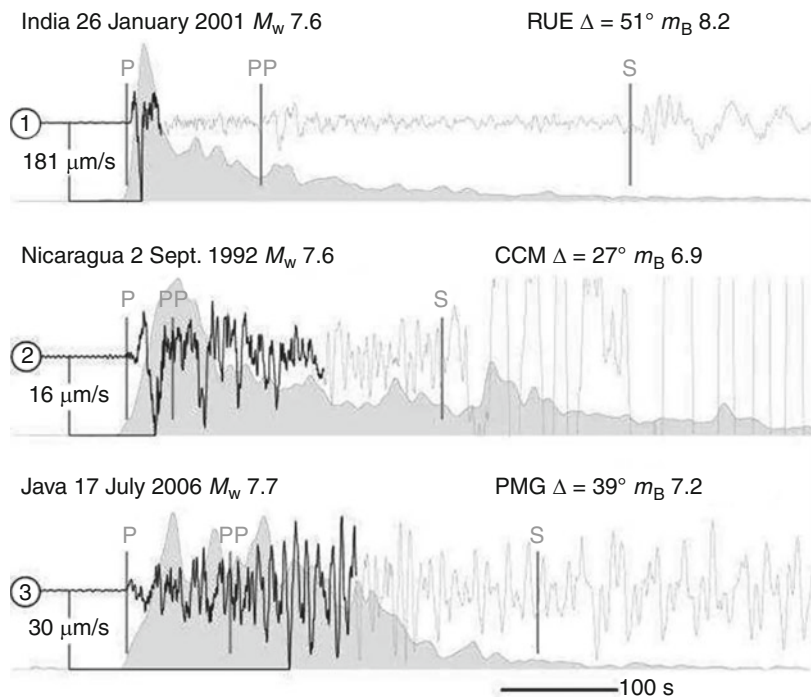
Earthquake, Magnitude, Fig. 11 Velocity broadband records of three earthquakes with similar M_w but exceptionally different rupture durations, m_B values and earthquake effects. The rupture duration D is estimated from the envelope of high-frequency P-wave radiation (gray shaded area). D corresponds approximately to the time window from the P-wave onset until the envelop amplitude drops to 40% of its maximum value (black record trace). (Copy of Fig. 2 in Bormann and Saul 2008, p. 700. With copyright granted by the Seismological Society of America)

Summary and Conclusions

The aim and historical development of magnitude scales have been sketched, systematic inconsistencies between current measurement procedures and their sometimes detrimental effects been outlined, the need for international standardization highlighted, and the recently proposed IASPEI standard procedures introduced. Reduction of magnitude errors due to incompatible measurement practices is a precondition for improving the usefulness of magnitude data for a more realistic quick assessment of differences in source physics and associated actual earthquake risk. No single magnitude value can describe the peculiarity of individual earthquakes both in terms of their tectonic size and shaking strength sufficiently reliable. A multi-magnitude documentation in earthquake catalogs and a deeper understanding of the physics behind different magnitude data is indispensable also for upgrading procedures aimed at improved medium- to long-term seismic hazards assessment.

Cross-References

- [Body Waves](#)
- [Earthquake Rupture: The Inverse Problem](#)
- [Earthquake, Focal Mechanism](#)
- [Earthquakes, Energy](#)
- [Earthquakes, Intensity](#)
- [Earthquakes, Location Techniques](#)
- [Earthquakes, Strong-Ground Motion](#)
- [Propagation of Elastic Waves: Fundamentals](#)



- Seismic Hazard
- Seismic Instrumentation
- Seismic Monitoring of Nuclear Explosions
- Seismogram Interpretation
- Slow Earthquake
- Surface Waves
- Tsunami

Bibliography

- Abe K (1981) Magnitudes of large shallow earthquakes from 1904 to 1980. *Phys Earth Planet Inter* 27:72–92
- Aki K (1967) Scaling law of seismic spectrum. *J Geophys Res* 72:1217–1231
- Bonner JL, Russell DR, Harkrider DG, Reiter DT, Herrmann RB (2006) Development of a time-domain, variable-period surface-wave magnitude measurement procedure for application at regional and teleseismic distances, part II: application and M_s - m_b performance. *Bull Seismol Soc Am* 96:678–696
- Bormann P (2002) Magnitude calibration functions and complementary data. In: Bormann P (ed) IASPEI new manual of seismological observatory practice, vol 2., DS3.1. GeoForschungs Zentrum Potsdam, Potsdam, pp 1–7
- Bormann P, Saul J (2008) The new IASPEI standard broadband magnitude m_B . *Seismol Res Lett* 79:698–705
- Bormann P, Saul J (2009a) Earthquake magnitude. In: Meyers A (ed) Encyclopedia of complexity and systems science, vol 3. Springer, New York, pp 2473–2496
- Bormann P, Saul J (2009b) A fast, non-saturating magnitude estimator for great earthquakes. *Seismol Res Lett* 80:808–816
- Bormann P, Engdahl ER, Kind R (2002a) Seismic wave propagation and Earth models, Chapter 2. In: Bormann P (ed) IASPEI new manual of seismological observatory practice, vol 1. GeoForschungs Zentrum Potsdam, Potsdam, pp 1–70
- Bormann P, Baumbach M, Bock G, Grosser H, Choy GL, Boatwright J (2002b) Seismic sources and source parameters, Chapter 3. In: Bormann P (ed) IASPEI new manual of seismological observatory practice, vol 1. GeoForschungs Zentrum Potsdam, Potsdam, pp 1–94
- Bormann P, Liu R, Xu Z, Ren K, Zhang L, Wendt S (2009) First application of the new IASPEI teleseismic magnitude standards to data of the China National Seismographic Network. *Bull Seismol Soc Am* 99:1868–1891
- Braunmiller J, Kradolfer U, Baer M, Giardini D (2002) Regional moment tensor determinations in the European-Mediterranean area – initial results. *Tectonophysics* 356:5–22
- Choy GL, Boatwright J (1995) Global patterns of radiated seismic energy and apparent stress. *J Geophys Res* 100:18205–18228
- Choy GL, Kirby S (2004) Apparent stress, fault maturity and seismic hazard for normal-fault earthquakes at subduction zones. *Geophys J Int* 159:991–1012
- Di Giacomo D, Grosser H, Parolai S, Bormann P, Wang R (2008) Rapid determination of M_e for strong to great shallow earthquakes. *Geophys Res Lett* 35:L10308. <https://doi.org/10.1029/2008GL033505>
- Di Giacomo D, Parolai S, Bormann P, Grosser H, Saul J, Wang R, Zschau J (2010a) Suitability of rapid energy magnitude determinations for rapid response purposes. *Geophys J Int* 180:361–374
- Di Giacomo D, Parolai S, Bormann P, Grosser H, Saul J, Wang R, Zschau J (2010b) Erratum. *Geophys J Int* 181:1725–1726
- Dziewonski AM, Chou TA, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J Geophys Res* 86:2825–2852
- Ekström G, Dziewonski AM (1988) Evidence of bias in estimations of earthquake size. *Nature* 332:319–323
- Geller RJ (1976) Scaling relations for earthquake source parameters and magnitudes. *Bull Seismol Soc Am* 66:1501–1523
- Gordon DW (1971) Surface-wave versus body-wave magnitude. *Earth Notes* 42:20–28
- Gutenberg B (1945a) Amplitude of surface waves and magnitude of shallow earthquakes. *Bull Seismol Soc Am* 35:3–12
- Gutenberg B (1945b) Amplitudes of P, PP, and S and magnitude of shallow earthquakes. *Bull Seismol Soc Am* 35:57–69
- Gutenberg B (1945c) Magnitude determination of deep-focus earthquakes. *Bull Seismol Soc Am* 35:117–130
- Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. *Ann Geofis* 9:1–15
- Hanks C, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84:2348–2350
- Hara T (2007) Measurement of duration of high-frequency energy radiation and its application to determination of magnitudes of large shallow earthquakes. *Earth Planets Space* 59:227–231
- Hatzidimitriou P, Papazachos C, Kiratzi A, Theodulidis N (1993) Estimation of attenuation structure and local earthquake magnitude based on acceleration records in Greece. *Tectonophysics* 217:243–253
- Herak M, Herak D (1993) Distance dependence of M_s and calibrating function for 20 second Rayleigh waves. *Bull Seismol Soc Am* 83:1881–1892
- Hutton LK, Boore DM (1987) The M_L scale in Southern California. *Bull Seismol Soc Am* 77:2074–2094
- IASPEI (2005) Summary of Magnitude Working Group recommendations on standard procedures for determining earthquake magnitudes from digital data. Available online at <http://www.iaspei.org/commissions/CSOI.html>
- Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2981–2987
- Kanamori H (1983) Magnitude scale and quantification of earthquakes. *Tectonophysics* 93:185–199
- Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. *Bull Seismol Soc Am* 65:1073–1095
- Kanamori H, Rivera L (2008) Inversion of W phase: speeding up seismic tsunami warning. *Geophys J Int* 175:222–238
- Kanamori H, Mori J, Hauksson E, Heaton TH, Hutton LK, Jones LM (1993) Determination of earthquake energy release and M_L using TERRASCOPE. *Bull Seismol Soc Am* 83:330–346
- Lee V, Trifunac M, Herak M, Živčić M, Herak D (1990) M_L^{SM} computed from strong motion accelerograms recorded in Yugoslavia. *Earth Eng Struct Dyn* 19:1167–1179
- Lomax A, Michelini A (2009) M_{wpd} : a duration-amplitude procedure for rapid determination of earthquake magnitude and tsunamigenic potential from P waveforms. *Geophys J Int* 176:200–214
- Okal EA, Talandier J (1989) M_m : a variable-period mantle magnitude. *J Geophys Res* 94:4169–4193
- Purcaru G, Berckhemer H (1978) A magnitude scale for very large earthquakes. *Tectonophysics* 49:189–198
- Rezapour M, Pearce RG (1998) Bias in surface-wave magnitude M_s due to inadequate distance corrections. *Bull Seismol Soc Am* 88:43–61
- Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seismol Soc Am* 25:1–32
- Richter C (1958) Elementary seismology. W.H. Freeman, San Francisco/London
- Scordilis EM (2006) Empirical global relations converting M_s and m_b to moment magnitude. *J Seismol* 10:225–236
- Stein S, Okal EA (2005) Speed and size of the Sumatra earthquake. *Nature* 434:581–582. <https://doi.org/10.1038/434581a>
- Tsai VC, Nettles M, Ekström G, Dziewonski A (2005) Multiple CMT source analysis of the 2004 Sumatra earthquake. *Geophys Res Lett* 32:L17304. <https://doi.org/10.1029/2005GL023813>

- Tsuboi S, Whitmore PH, Sokolowski TJ (1999) Application of M_{WP} to deep and teleseismic earthquakes. *Bull Seismol Soc Am* 89:1345–1351
- Uhrhammer RA, Collins ER (1990) Synthesis of Wood-Anderson seismograms from broadband digital records. *Bull Seismol Soc Am* 80:702–716
- Utsu T (2002) Relationships between magnitude scales. In: Lee A, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*. Academic, Part A, Amsterdam, pp 733–746
- Vaněk J, Zátpek A, Kárník V, Kondorskaya NV, Riznichenko YV, Savarenski EF, Soloviev SL, Shebalin NV (1962) Standardization of magnitude scales. *Izv Akad Nauk SSSR, Ser Geogr* 2:153–157. (with English translation)
- von Seggern D (1977) Amplitude-distance relation for 20-second Rayleigh waves. *Bull Seismol Soc Am* 67:405–411
- Weinstein SA, Okal EA (2005) The mantle wave magnitude M_m and the slowness parameter Θ : five years of real-time use in the context of tsunami warning. *Bull Seismol Soc Am* 95:779–799

Earthquakes and Crustal Deformation

Robert McCaffrey

Department of Geology, Portland State University, Portland, OR, USA

Definition

The term earthquake is commonly used only to describe sudden slip on a fault within the Earth that produces seismic waves in the frequency bands that we can either feel or observe with seismometers. However, other sources that produce deformation of the Earth's surface do not generate seismic waves but are instead observable with sensitive geodetic instruments that measure the crustal deformation directly. Because of the great advances over the past few decades in observing and understanding them, in this chapter we broaden the definition of earthquake to include these “quiet” sources of Earth deformation. Other chapters deal with some of these sources in more detail.

With this enhanced definition, earthquakes encompass a wide range of phenomena. The most common type we often think of is rapid slip within the Earth. The motion across a fault is typically of a shearing type but seismic disturbances can include volumetric or planar expansion in some settings, most notably near volcanoes. The classes of “earthquakes” now observed can be categorized by a rate of slip from very fast, kilometers per second, to slow, millimeters per year, covering ~ 12 orders of magnitude in slip rate. Figure 1, taken from Ide et al. (2007), shows the classes of earthquakes broken down by rate of the deformation. Because of the varying speeds of the motions, different types of instrumentation are required to cover specific bands.

Types of Earthquakes and Related Phenomena Defined

Here the types of earthquakes that produce crustal deformation are briefly defined. Many of them are described in detail in other chapters.

Tectonic earthquakes – Tectonic earthquakes are those that occur in response to plate motions or other predominantly shearing sources. They result primarily from the stick-slip behavior of faults. (Deep earthquakes occur by other mechanisms but are not discussed here as they do not visibly deform the crust.) As stress builds up across a fault surface, friction on the fault prevents the surface from sliding until the strength (some maximum shear stress) of the fault is exceeded. Then the two sides of the fault snap back (see “[Elastic Rebound Theory](#)” below), releasing some or all of the elastic strain built up. This slip occurs at several kilometers per second, releasing high-frequency seismic waves. These are by far the most common type of earthquakes and the most destructive.

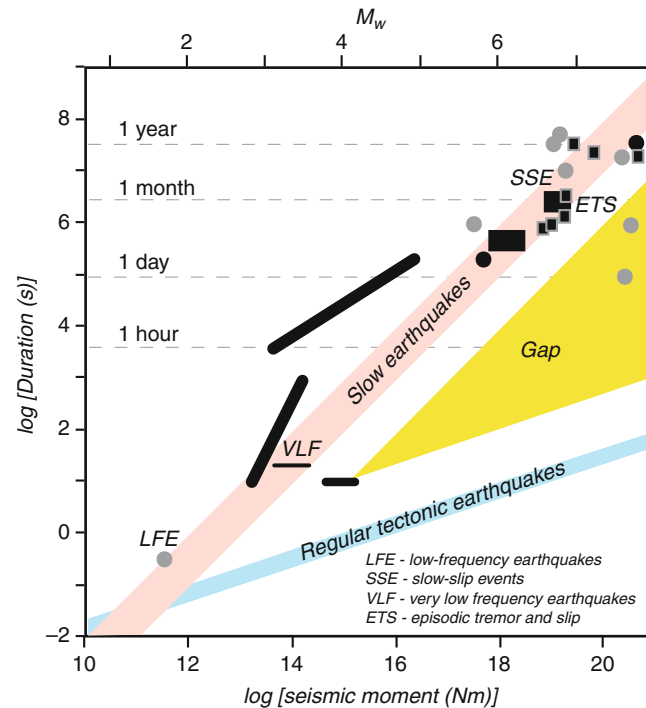
Tectonic tremors – These are a new type of seismic disturbance first recognized in Japan in 2002 (Obara 2002). Originally they were called non-volcanic tremors due to some similarities to volcanic tremors. Whereas tectonic earthquakes are due to finite duration slip (on the order of seconds to minutes) on the fault, tectonic tremors appear as low amplitude, non-harmonic background disturbances lasting days to weeks. They are comprised in part of very small earthquakes but most of the signals are incoherent. Often, but not always, they accompany slow-slip events and are thought to be some aspect of the same process (Rubenstein et al. 2010).

Volcanic earthquakes – Volcanic earthquakes are similar to tectonic earthquakes in that they are due to the failure of rock along faults. However, their size is limited due to the relatively small fault surface area available near volcanoes.

Volcanic tremors – These earth disturbances are likely caused by fluid (water or magma) moving inside the earth. The signals are harmonic and can last a long time (relative to tectonic earthquakes) but are generally of low amplitude and generate little crustal deformation.

Slow earthquakes – Some earthquakes occur at slip speeds that are slower than typical tectonic earthquakes but fast enough to excite seismic waves. They occur when slip is in materials of low rigidity such as accretionary prisms of subduction zones. In such cases, they can produce tsunamis that are larger than expected for the earthquake's magnitude. Examples are earthquakes in Nicaragua (Kanamori and Kikuchi 1993) and New Zealand (Doser and Webb 2003).

Low-frequency earthquakes – Such earthquakes, called LFEs, have been observed in volcanic settings and more recently associated with tectonic tremors at subduction zones. As the name indicates they generate much lower



Earthquakes and Crustal Deformation, Fig. 1 Durations and moments of various types of crustal deformation events, modified from Ide et al. (2007). Black and gray circles and elongated bars show silent earthquakes, afterslip and slowslip events from Italy, Nankai, Kilauea, San Andreas, and Sanriku (Ide et al. 2007). Squares are slowslip events

frequencies than normal tectonic earthquakes. In both settings the low frequencies may be attributable to fluid interactions (Shelly et al. 2006).

Slow-slip events – In this class of events, called SSEs, the rupture speed is on the order of a few millimeters per day. Thus the rupture propagates at 11 orders of magnitude slower than a tectonic earthquake but only about 10 times faster than tectonic creep at plate motion rates. At these slow-slip speeds, seismic waves are not generated and slow-slip events are not detected by seismometers. To date, SSEs have been detected by geodetic instruments at subduction zones, Kilauea and on the San Andreas fault.

Post-seismic slip events – Another type of crustal motion that is often observed following large tectonic earthquakes is due to afterslip, thought to arise from continued slip on the fault at slow speeds. Afterslip can in small part be explained by aftershocks but the majority of the slip occurs without producing seismic waves. Such slip can last years and produce as much total slip as the main earthquake itself. The afterslip events often decay in accordance with Omori's Law.

Mantle relaxation – Following large earthquakes, the redistribution of the stress in the crust and mantle can cause the mantle to flow in a viscous manner, which then produces a long-period deformation of the crust above. This signal can often be confused with post-seismic slip.

from New Zealand (McCaffrey et al. 2008) and Guerro, Mexico (Larson et al. 2007). These follow a scaling relation of $\sim M_0/t$. Typical scaling relation for regular shallow interplate earthquakes is shown by the blue line. (See Ide et al. (2007) for additional information)

Induced earthquakes – Earthquakes are induced by the activities of humans that modify the stress at shallow depths in the Earth. Examples of such activities are mining, drilling, and water impoundment behind dams. Induced earthquakes tend to be of significant hazard because they are very shallow and close to infrastructures (see ▶ “Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyana, India”).

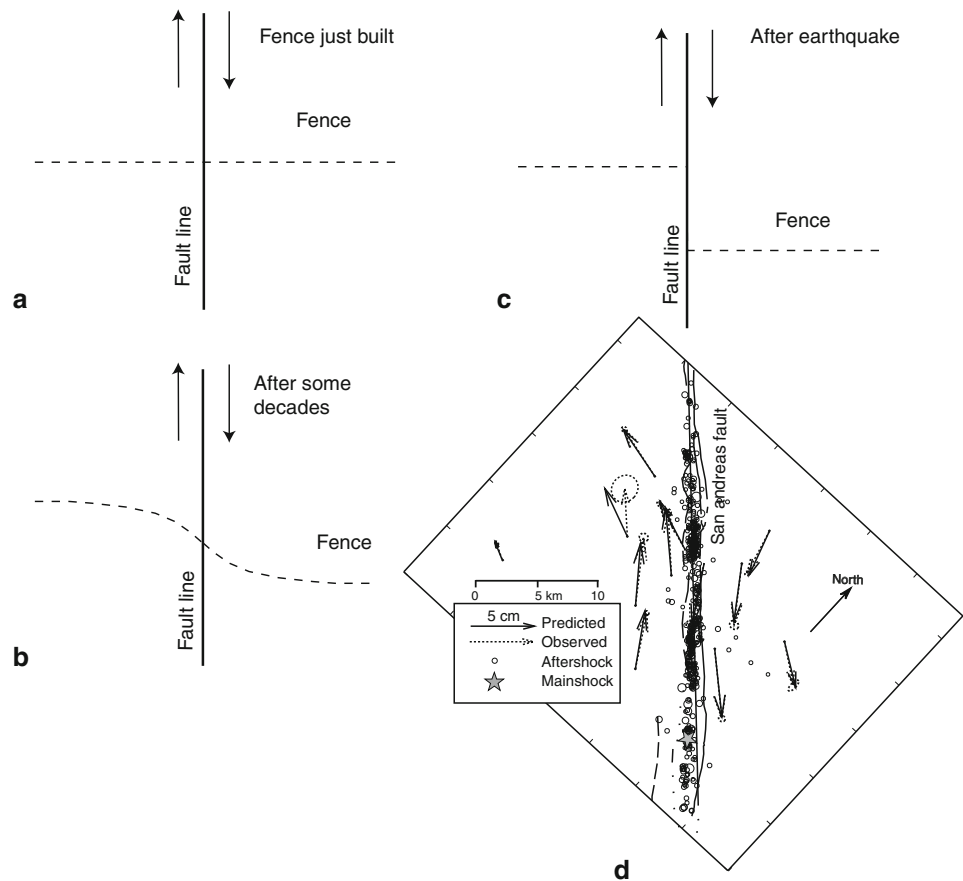
Fault morphology – Most deformation of the crust is accommodated by faulting. Crustal faults are the lasting expressions of past earthquakes and provide a means via geologic investigations to characterize the earthquakes. Compilations of estimates of slip rates and fault lengths are used to calculate long-term (Holocene, for example) deformation rates in continental regions.

Elastic Rebound Theory

Large earthquakes, as noted above, are caused by slip on crustal faults and exhibit stages of strain build-up and release known as the earthquake cycle. A simple form of this cycle is elastic rebound. The concept of elastic rebound as relating to the earthquake cycle was first introduced by Harry Reid (Reid 1910) following earthquake observations in Sumatra and California. (See the interesting, in-depth description of the history of elastic rebound by Segall 2010.) In this description,

Earthquakes and Crustal Deformation, Fig. 2

Illustration of the elastic rebound effect. The fence originally built in a *straight line* across the fault (a) is deformed by the strain around the fault in the period leading up to an earthquake (b). In (c) the earthquake has offset the fence but also straightened it again by the release of the strain. (d) Measured and modeled co-seismic offsets near Parkfield CA resulting from the 2004 M6 earthquake. (Redrawn with permission from Barbot et al. 2009)



the crustal blocks adjacent to a fault are stuck together by friction on the fault surface and their relative movement causes strain in the blocks (Fig. 2a, b). The material comprising the upper crust is thought to be linearly elastic, so the strain increases over time in a steady fashion as long as the blocks move at the same speed. At some point, the stress on the fault reaches its strength, and the fault slips suddenly in an earthquake (Fig. 2c). The strain accumulated over decades to centuries is relieved in a matter of seconds. The slip on the fault causes the stored strain energy in the crust to be converted to kinetic energy which in turn is emitted as seismic waves and heat and the stress on the fault drops to a lower level. The use of the word “rebound” in this case refers to the idea that any feature on the surface, such as a fence crossing the fault, will return to its original shape (i.e., free of strain) after the earthquake (Figs. 2 and 3). The earthquake in a broader sense produces permanent strain on a regional scale by displacing the crustal blocks along the fault (and offsetting the fence). Many observations of earthquake deformations have validated the elastic rebound model – the example of the 2004 Parkfield, California, earthquake rebound is shown in Fig. 2d (Barbot et al. 2009).

We have discovered in the past few years that elastic rebound pertains not only to earthquakes of the type Reid

observed, but also to earthquake afterslip (Heki et al. 1997) and so-called “slowslip events” (Dragert et al. 2001). In these phenomena, illustrated by GPS time series shown in Fig. 4, slip on faults can occur at much slower rates than it does during earthquakes. In these time series showing the position of a point on the ground, the steady motion of the linear earthquake cycle is interrupted not only by earthquakes (abrupt offsets in Fig. 4a) but also by afterslip that decays with time. In slowslip events, the linear site motion is interrupted for weeks to months (Fig. 4b) in Cascadia and New Zealand, for example. The spatial pattern of the slip and our knowledge of the rebound cycle guide our interpretation of these anomalous signals as slip on the fault surface.

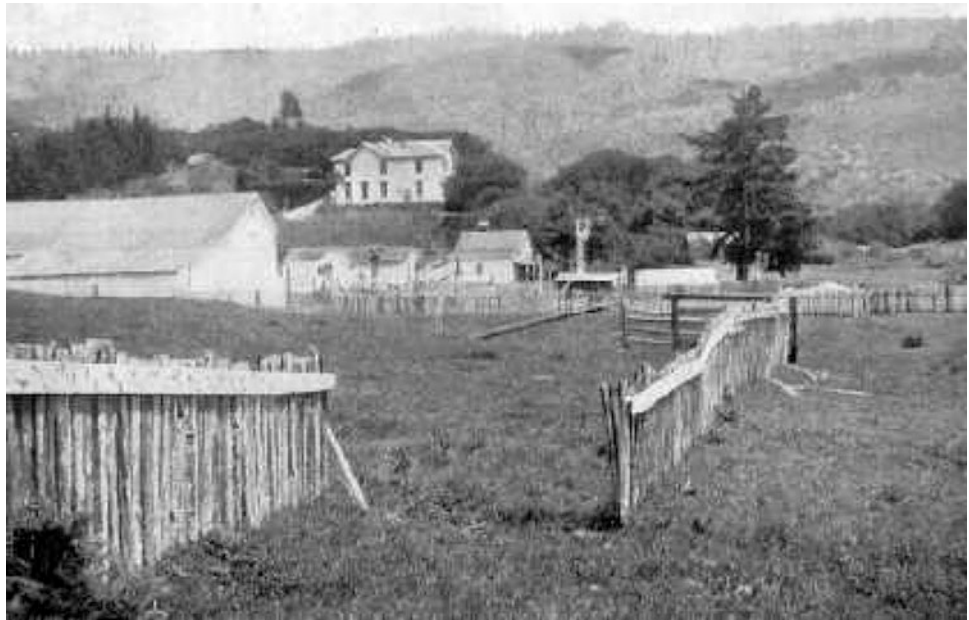
The simple picture of elastic rebound theory is a good conceptual view of the earthquake cycle but in detail we see that it is much more complex. As more and better observations are made we see that the earthquake cycle includes time-dependent deformations on many time scales. These are discussed next.

Pre-seismic Crustal Deformation

The deformation of the Earth’s surface leading up to earthquakes has been of great interest to scientists for a long time

Earthquakes and Crustal Deformation, Fig. 3

Picture taken in Marin County shows a fence that was offset about 10 ft. along the trace of the San Andreas fault by the 1906 earthquake. (Photo attributed to G.K. Gilbert; <http://america.hurrah.com/USGS/26.htm>)



due to the quest to identify precursors for prediction purposes. Such deformation can be grouped into long-term and short-term signals. For the long-term, the Earth deforms in the vicinity of faults largely in roughly the elastic rebound manner described above. Mapping out the temporal changes in crustal strain near faults is being used extensively to identify which parts of the fault are being loaded, or being subject to stress increases. In a simple sense, the parts of the fault being loaded may help delineate the future rupture region but the details are far from clear. For example, the current rate of loading of the Cascadia subduction thrust is inferred from the results of 15 years of GPS observations by McCaffrey et al. (2007) (Fig. 5) and has the important implications for how close to populated regions the rupture may come during a large subduction earthquake. Such studies have been conducted for most subduction zones and continental transform faults globally and the relationship of such loading to seismic slip is just starting to be understood.

A now common approach to understanding the crustal deformation between earthquakes (also called the *interseismic* phase) is through the use of block models (Matsu'ura et al. 1986; McCaffrey 2002; and many others). This method uses the idea that the crust or lithosphere is comprised of discrete entities (blocks) that are bounded on their lateral sides by other blocks and at their base by a stress-free region (asthenosphere). The lateral boundaries between the blocks are described by their frictional properties (i.e., a stuck fault). The blocks themselves rotate on the Earth's surface and so the motion can be described by an angular velocity (Euler pole) with its origin at the center of the Earth. Frictional interactions with adjacent blocks cause elastic strains to build in the blocks. Hence, new GPS observations

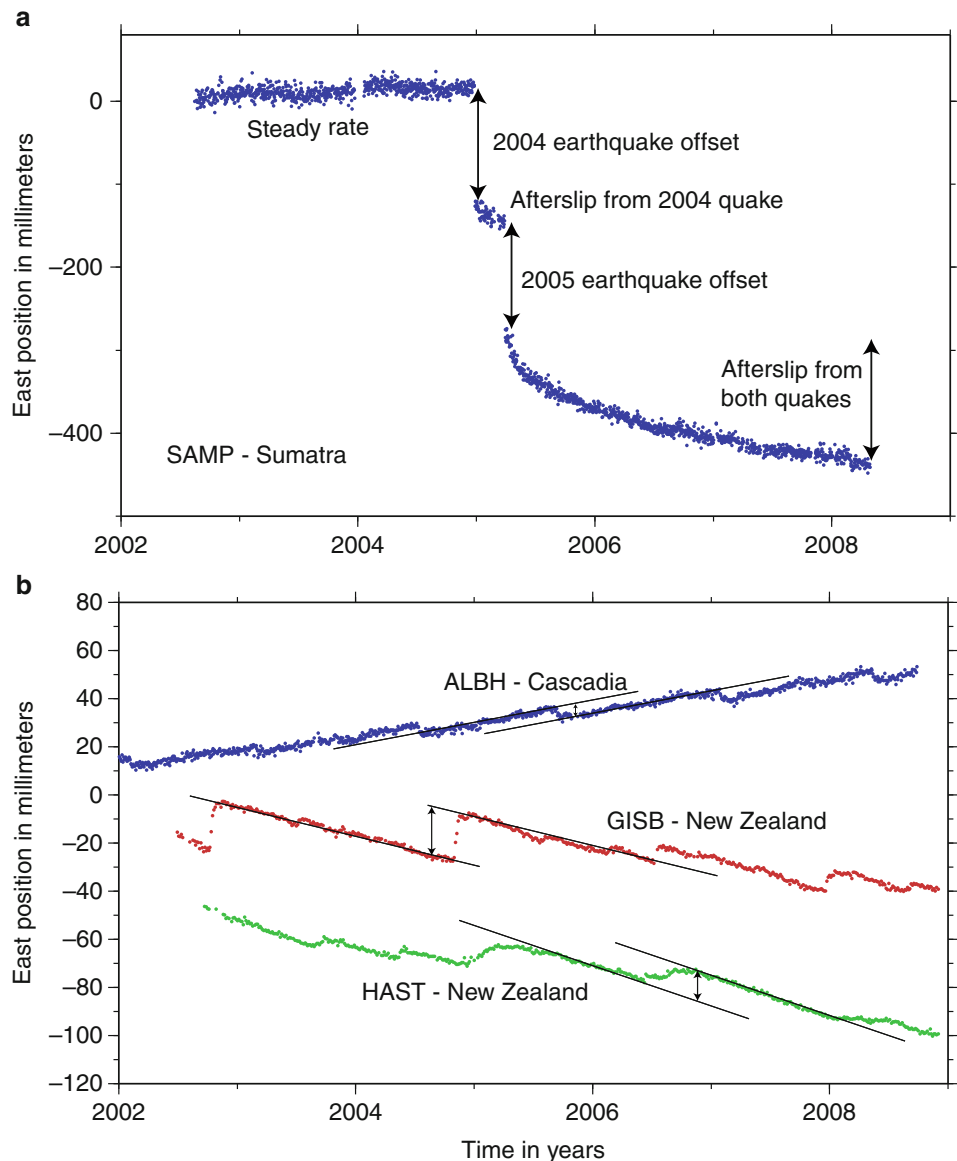
contain information about both the angular velocities of the blocks (and slip rates of their bounding faults) and the frictional properties of the bounding faults. A recent example of this is the January 2010 M 7.0 Haiti earthquake – block modeling by Manaker et al. (2008) had shown that a M ~7.2 was possible for the Enriquillo Fault in Haiti (of course such modeling does not tell when such an event will occur).

Another approach to understanding crustal deformation is to assume that at some depth below the surface, the deformation field is continuous and that the surface faulting is only a vague reflection of that motion (England and McKenzie 1982), in a similar fashion that floating ice blocks may show the general, but not specific, flow pattern of the water beneath. In this case, the surface faulting can be related to the continuous deformation field only in a statistical manner and the deformation field cannot be used to predict slip on unknown faults. On the other hand, the continuous deformation approach can be more readily related to the forces driving deformation through an assumption that the deformation is controlled by a viscous, rather than elastic, Earth. (The assumption of a viscous medium is required to relate our observations of deformation rates to forces.) For example, Jones et al. (1996) relate crustal deformation in the western USA to gravitational forces arising from spatial variations in crustal thickness acting on a viscous medium.

Short-term crustal deformation around faults leading up to earthquakes is much more elusive. To date there have been few or no convincing cases of isolated changes in deformation in the days to months prior to a large earthquake. Mogi (1985) shows evidence for accelerated uplift at tide gages in the 5 years preceding the 1983 Sea of Japan earthquake (see also Scholz 2002). This evidence must be considered less

Earthquakes and Crustal Deformation, Fig. 4

(a) The east position of continuous GPS site SAMP in northern Sumatra leading up to and following the December 2004 and March 2005 great earthquakes at the Sumatra subduction zone. The earthquakes each caused large sudden offsets of about 100 mm. The afterslip, which decays in a logarithmic fashion, produced over 100 mm of additional motion in the 3 years following the earthquakes. (b) GPS time series from Cascadia and New Zealand showing the geodetic signature of “slowslip” events. These events produce offsets of the time series but, unlike earthquakes, the offsets occur over weeks to months, rather than minutes. The *solid lines* show the linear trends (steady site velocities) between the slip events

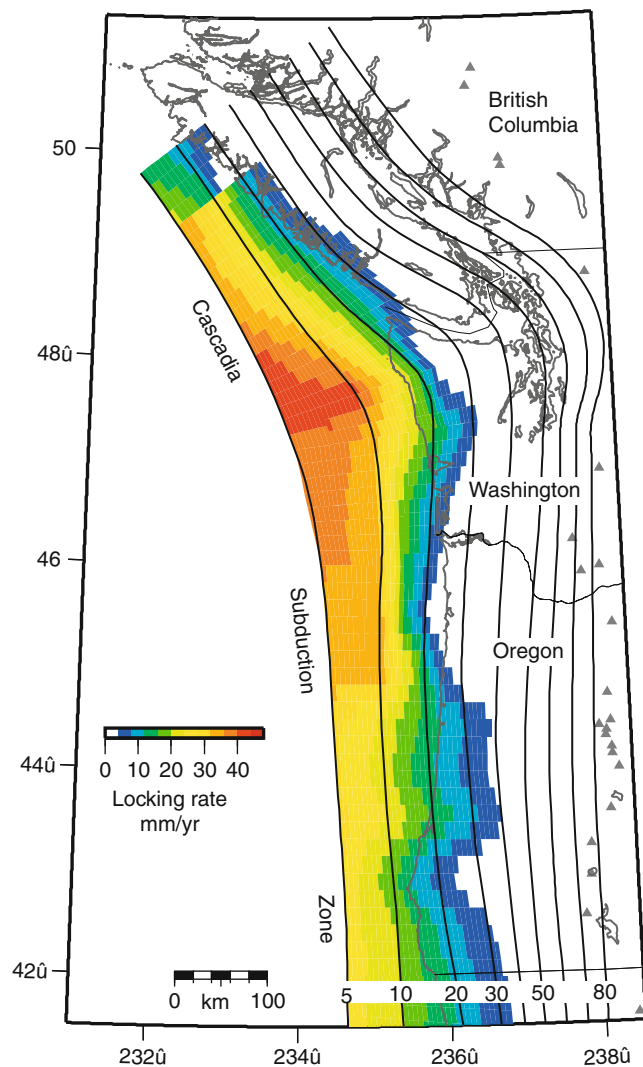


certain than it was a few years ago before we knew of slowslip events. Although no SSEs have been observed in the past 10 years on modern GPS equipment in the vicinity of the 1983 earthquake, we cannot rule out that the pre-earthquake uplift was due to slowslip on the fault that subsequently ruptured (or some other fault) rather than a useful precursory change prior to the quake. (Because many SSEs have occurred without triggering tectonic earthquakes, we know that slowslip events are NOT short-term precursors to earthquakes.)

Co-seismic Crustal Deformation

The term co-seismic refers to the time right around the earthquake when the fault rupture is still propagating and seismic waves are being generated. Generally this amounts to a few

seconds for a magnitude 6–7 earthquake and a minute or more for $M > 8$. The 2005 M9.3 Sumatra earthquake is thought to have ruptured for about 5 min – this is the time it takes to rupture such a long fault (~1300 km) at typical rupture speeds (~3–4 km/s). Earthquakes generate both dynamic displacements, those that propagate but return to their original positions after the waves pass, and static displacements that are permanent. Seismometers can record the dynamic motions within their pass band but cannot measure the static displacement which requires geodetic instrumentation. Both the dynamic and static displacements induce stress changes in the crust and mantle that can lead to other forms of crustal deformation, such as triggered earthquakes, afterslip, slowslip and mantle relaxation. While GPS observations can measure the passing waves for very large earthquakes (as can seismographs), we are more concerned here with the static displacements.



Earthquakes and Crustal Deformation, Fig. 5 Frictional loading of the Cascadia subduction zone inferred from crustal deformation of the coastal regions measured with 15 years of GPS observations (McCaffrey et al. 2007). The locking rate is the rate at which the subducting plate pushes the base of the overriding plate landward. This rate, integrated over time, provides an estimate, according to elastic rebound theory, of how much slip might occur in a large strain-relieving earthquake. Contours are of the depth to the plate boundary interface, in kilometers

The static displacements from a large earthquake provide much information on the faulting process and can be measured by comparing the position of ground points before and after the earthquake (Fig. 2d). However, because the points are moving constantly due to tectonic plate motions and fault-induced elastic and anelastic strain, as discussed above, it is not always easy to know where the points were just prior to the earthquake. Consequently, in many active regions of the world, GPS data are collected continuously (called continuous GPS, or cGPS) which provide a direct measure of the prequake position. However, such observatories are expensive and generally not sufficiently dense to characterize the earthquake details when one happens. The geodetic characterizations of most earthquakes are now based on a combination of

cGPS and survey-mode GPS, where measurements are made intermittently.

Simple physical representations of the Earth have been quite successful in characterizing the slip during large earthquakes based on the observed static displacements. (Seismic waves are also used to estimate the slip distributions of earthquakes – see ▶ “Earthquake Source Theory”). The most commonly used structure and source representation is the elastic, half-space dislocation (see Segall 2010). The popularity of this method owes to the simplicity of the algorithm published by Okada (1992, 1995) for use in three dimensions and its success in reproducing the observations. More times than not, the geodetic and seismologic representations of an earthquake source agree within the levels of accuracy. In fact, when they do not agree, the blame is often placed on the complexity of the source rather than on the techniques.

Such studies of the distributions of slip on the fault during earthquakes have shown us just how complex the slip can be. It remains enigmatic why the earthquake slip distributions (fault unloading) are so much more heterogeneous than the loading distribution (as estimated from interseismic studies as discussed above). This is one of the challenges facing us if we want to use contemporary fault loading measurements to understand the real earthquake hazards.

Post-seismic Crustal Deformation

Unlike pre-seismic deformation, many striking examples of post-seismic deformation have been documented for large- and moderate-magnitude earthquakes. The mechanisms of this deformation are most likely either continued slowslip on the fault or adjustment of the viscous mantle or lower crust to the changes in stress brought about by the earthquake.

It has been known for a long time that large, shallow earthquakes produce aftershocks that decay with time (t) in a $1/t$ fashion. This aftershock decay was explained by the changes in friction on the fault as the slip rate decayed, known as Omori’s Law. Continued slip on faults after large earthquakes sometimes shows a similar decay and this decay follows Omori’s Law as well (though the total slip is many times greater than what can be explained by aftershocks alone). The first clear example of significant afterslip was for the 1994 Japan Trench earthquake observed by Heki et al. (1997) where the relatively aseismic afterslip exceeded the co-seismic slip magnitude in the year following the earthquake. The 2004 Parkfield earthquake was followed by aseismic afterslip that was about three times the co-seismic slip (Johanson et al. 2006).

We now describe the magnitude of an earthquake or any other slip event with the moment-magnitude scale. For a slip event on a fault, the moment M_o is defined as the average slip magnitude times the area of that slip times the rigidity modulus

of the rocks around the fault, and the magnitude is $M_w = 2/3 \log(M_o) - 9.5$ (Hanks and Kanamori 1979). Hence moment can be based on either seismic or geodetic measures of the slip and area and these can be compared. By making such comparisons we now understand that aseismic afterslip (or slowslip) contributes as much or more to the total slip budget on some faults as do earthquakes. This discovery has positive implications for earthquake hazards in the sense that seismic events of a given size may be less frequent than previously estimated.

Observational Advances and New Discoveries

Our knowledge of the existence of the various types of earthquakes and their structures and our understanding of them rely on our ability to detect them and explain the observations. In the past 20 years, great advances in instrumentation have led directly to the detection of some new types of deformation sources. They, along with new analytical algorithms, have greatly advanced our understanding of earthquakes and crustal deformation. Here, we describe some of these new methods and the important discoveries that they have allowed.

Possibly the greatest advance in the past couple of decades has been in the use of remote sensing in earthquake and crustal deformation studies. These new methods include the Global Positioning System (GPS), Interferometric Synthetic Aperture Radar (InSAR), and Light Detection and Ranging (LIDAR). The first two of these allow us to measure displacements of the Earth's surface using information obtained from satellites orbiting the Earth. With repeated measurements we can watch the temporal variations in surface deformation with great accuracy. LIDAR is allowing unprecedented resolution of the topography of the Earth's surface beneath the biological layer and has delineated geologic structures (faults) in regions that are too remote or forest-covered to find otherwise.

GPS, in particular when used in the differential mode, has resulted in the detection of the large, slow movements along faults – slowslip, post-seismic, afterslip, and mantle relaxation described above. Today, thousands of GPS instruments operate continuously at sites around the globe providing daily estimates of the site's three-dimensional position, which can be combined to form the equivalent of a displacement seismogram. The positional accuracy achieved for any given day is on the order of a few millimeters. The signal can also be sampled at a smaller time interval, such as seconds, to record the passing seismic waves in addition to the total static displacement produced by the earthquake. Such a resampling is possible only for very large earthquakes where the ground motions exceed a few millimeters. However, the greatest contribution of GPS has been in the discovery of crustal deformation sources that do not produce seismic waves.

The knowledge that fault movement could occur at rates somewhere between fast slip (km/s) and creep (mm/year) has

been with us for years – such motions were observed on strain meters near the San Andreas fault in California (e.g., Linde et al. 1996). However, these slip events were very small compared to sizes of earthquakes on the fault. As described above, the observation by Heki et al. (1997) that post-seismic slip exceeded the slip during the earthquake itself was remarkable as it changed our ideas about how often earthquakes should occur on a given fault. In other words, we could no longer expect that the total slip on a fault in a given time period would be accompanied by earthquakes, an idea that held its roots in elastic rebound theory. Instead, some large portion of fault could slip without earthquake shaking.

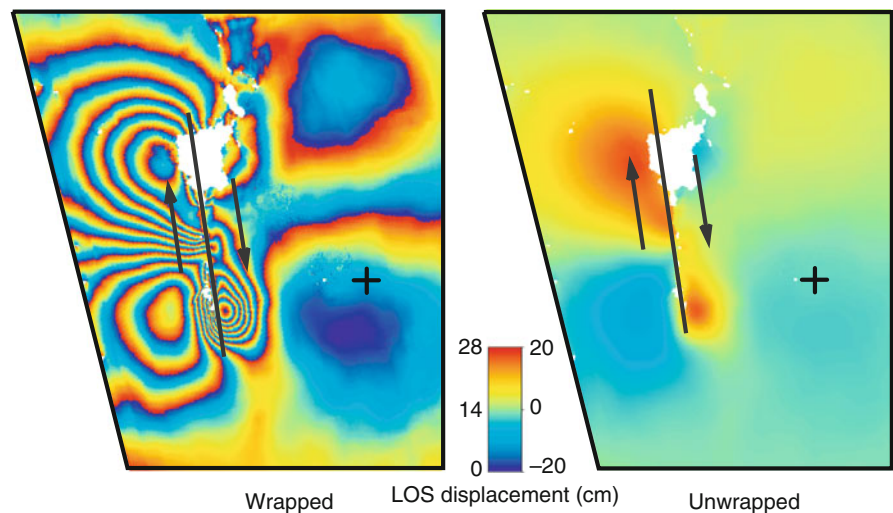
As remarkable as that discovery was, in 1999 Dragert et al. (2001) recorded in the Western Canada Deformation Array (continuous GPS network) a very slow deformation of the ground that looked just like a $M \sim 6.7$ earthquake, except that it occurred over 2 weeks time instead of just a few seconds. The maximum surface displacement was only 5 mm and at the time no seismic waves were detected. The slow surface deformation was interpreted by them to be due to slip on the Cascadia subduction fault directly beneath the GPS network. The slip on the fault was very slow, amounting to only 2 cm over the 2 weeks – this was 12 orders of magnitude slower than slip during a normal earthquake (~ 3 km/s).

Around the same time, Obara (2002) discovered the presence of low-level incoherent but intermittent seismic signals, called *tectonic tremor* above, coming from the subduction fault beneath Japan. Importantly, these signals appeared to originate on the fault just below the deepest slip extent from recent large earthquakes. Initial suggestions were that the tremor coincided in space and time with slowslip (Rogers and Dragert 2003) but additional studies of both phenomena have not borne that out. Further detailed studies of tectonic tremor requiring new seismological techniques revealed that it is likely in part due to numerous small thrust earthquakes (LFE), on the subduction thrust boundary (Shelly et al. 2006). However, when they occur together, the total slip that can be accounted for with these small quakes is far less than what is seen in the slowslip events. Hence, it appears that the majority of the slowslip is indeed seismically silent.

While GPS provides a fine temporal sampling of the deformation, it is generally sparse in the spatial dimension because continuous GPS installations are expensive. The use of differential satellite interferometry (InSAR) has helped fill in the spatial gaps in pictures of the deformation of the Earth's surface by earthquakes and volcanoes. The InSAR method uses the interference patterns between two or more radar images of the ground that have been obtained from satellites. Comparing the images allows estimation of the displacement of the ground in the direction of the satellite (the LOS or line-of-sight component). The method relies on the ground surface changing at long wavelengths only, so that the images have strong

Earthquakes and Crustal Deformation,

Fig. 6 Interferogram showing the ground deformation observed during the December 2003 M6.6 Bam, Iran, earthquake (http://neic.usgs.gov/neis/eq_depot/2003/eq_031226/). On *left*, the “wrapped” image shows line-of-sight (LOS) displacements by the phase of the difference in distance to the satellite; on the *right* the image has been “unwrapped,” that is, converted to distance changes. In both cases the four-lobe displacement field of the strike-slip earthquake is evident. (Image with permission from Funning et al. 2005)



correlation (coherence) and can be matched. InSAR provides a scalar observation of the deformation but at high spatial density (Fig. 6) whereas GPS provides a vector of the deformation at high temporal sampling. Many studies use GPS and InSAR together to characterize the deformation, taking advantage of the strengths of both types of observations.

The LIDAR method is extremely useful in detecting faults that intersect the ground surface. Because it uses shorter wavelengths than radar, it allows us to see finer features. Very often, faults produce only subtle lineations on the ground surface, which can now be detected at the sub-meter level. While LIDAR does not give details of the fault, like the slip rate, it can pinpoint targets for additional field study. An example of the utility of the LIDAR method is the delineation of the Seattle Fault in northwestern Washington state (Haugerud et al. 2003). This fault, which crosses the highly populated Puget Sound region, produced a large earthquake about 1,100 years ago. The fault scarp is large and obvious in some places but subtle and covered in forests in others.

The advance of the new geodetic techniques has had such a positive impact on geophysics that major organizational steps have been taken to utilize them. The consortium of universities called UNAVCO (University Navstar Consortium; unavco.org) was developed to support the global research community in the use of geodetic tools. In several countries, permanent dense geodetic networks and similar consortia have been established to monitor earthquakes and their associated phenomena.

Perspectives

Our ideas about earthquakes *sensu strictu* have undergone drastic changes in the past few years. Elastic rebound theory, while still a sensible conceptual model, does not cover the

large range of observed phenomena that make up the earthquake cycle. A clear challenge that still faces us is to tie together the various deformation processes we observe throughout the cycle into a comprehensive earthquake theory. Some of the more important goals are to understand the observations we make in terms of future earthquake risk. For example, although we can measure current loading of faults quite well by its elastic strain near the fault, we do not know what it means for the slip distribution in future events. Geodetic methods of measuring crustal deformation have a short history relative to the times between large earthquakes, so geologic methods including paleoseismology (see ► “Paleoseismology”) and LIDAR that extend the observation history are crucial. Continued study and understanding of the aseismic slip events and tectonic tremor on faults may pay big dividends toward delineating the parts of the faults that are prone to seismic slip during large destructive earthquakes. The acquisition of deformation data at high resolution in both time and space will assist greatly in understanding processes operating at depth within the Earth.

Cross-References

- Earthquake Precursors and Prediction
- Earthquake, Aftershocks
- Earthquakes, Volcanogenic
- GPS, Tectonic Geodesy
- Paleoseismology
- SAR Interferometry
- Seismic Hazard
- Seismicity, Intraplate
- Seismicity, Subduction Zone
- Slow Earthquake

Bibliography

- Barbot S, Fialko Y, Bock Y (2009) Postseismic deformation due to the Mw 6.0 2004 Parkfield earthquake: stress-driven creep on a fault with spatially variable rate-and-state friction parameters. *J Geophys Res* 114:B07405. <https://doi.org/10.1029/2008JB005748>
- Doser DI, Webb TH (2003) Source parameters of large historical (1917–1961) earthquakes, North Island, New Zealand. *Geophys J Int* 152:795–832
- Dragert H, Wang K, James TS (2001) A silent slip event on the deeper Cascadia subduction interface. *Science* 292:1525–1528
- England P, McKenzie D (1982) A thin viscous sheet model for continental deformation. *Geophys J R Astron Soc* 70:295–321
- Funning GJ, Parsons B, Wright TJ, Jackson JA, Fielding EJ (2005) Surface displacements and source parameters of the 2003 Bam (Iran) earthquake from Envisat advanced synthetic aperture radar imagery. *J Geophys Res* 110:B09406. <https://doi.org/10.1029/2004JB003338>
- Hanks TC, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84:2348–2350
- Haugerud RA, Harding DJ, Johnson SY, Harless JL, Weaver CS, Sherrod BL (2003) High-resolution Lidar topography of the Puget Lowland. Washington. *GSA Today* 13:4–10
- Heki K, Miyazaki S, Tsuji H (1997) Silent fault slip following an interplate thrust earthquake at the Japan Trench. *Nature* 386:595–598
- Ide S, Beroza GC, Shelly DR, Uchide T (2007) A scaling law for slow earthquakes. *Nature* 447:76–79. <https://doi.org/10.1038/nature05780>
- Johanson IA, Fielding EJ, Rolandone F, Bürgmann R (2006) Coseismic and postseismic slip of the 2004 Parkfield earthquake from space-geodetic data. *Bull Seismol Soc Am* 96(4b):S269–S282
- Jones CH, Unruh JR, Sonder LJ (1996) The role of gravitational potential energy in active deformation in the southwestern US. *Nature* 381:37–41
- Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake – a slow tsunami earthquake associated with subducted sediments. *Nature* 361:714–716
- Larson KM, Kostoglodov V, Miyazaki S, Santiago JAS (2007) The 2006 aseismic slow slip event in Guerrero, Mexico: new results from GPS. *Geophys Res Lett* 34:L13309. <https://doi.org/10.1029/2007GL029912>
- Linde AT, Gladwin MT, Johnston MJS, Gwyther RL (1996) A slow earthquake sequence on the San Andreas fault. *Nature* 383:65–68
- Manaker DM, Calais E, Freed AM, Ali ST, Przybylski P, Mattioli G, Jansma P, Prepetit C, de Chabaliér JB (2008) Interseismic Plate coupling and strain partitioning in the Northeastern Caribbean. *Geophys J Int* 174:889–903. <https://doi.org/10.1111/j.1365-246X.2008.03819.x>
- Matsu'ura M, Jackson DD, Cheng A (1986) Dislocation model for aseismic crustal deformation at Hollister, California. *J Geophys Res* 91:12661–12674
- McCaffrey R (2002) Crustal block rotations and plate coupling. In: Stein S, Freymueller J (eds) *Plate boundary zones*. AGU geodynamics series, vol 30. American Geophysical Union, Washington, DC, pp 101–122
- McCaffrey R, Qamar AI, King RW, Wells R, Ning Z, Williams CA, Stevens CW, Vollick JJ, Zwick PC (2007) Plate locking, block rotation and crustal deformation in the Pacific Northwest. *Geophys J Int*. <https://doi.org/10.1111/j.1365-246X.2007.03371.x>
- McCaffrey R, Wallace LM, Beavan J (2008) Slow slip and frictional transition at low temperature at the Hikurangi subduction zone. *Nat Geosci* 1:316–320
- Mogi K (1985) *Earthquake prediction*. Academic, Tokyo
- Obara K (2002) Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science* 296:1679. <https://doi.org/10.1126/science.1070378>
- Okada Y (1992) Internal deformation due to shear and tensile faults in a half-space. *Bull Seismol Soc Am* 82:1018–1040
- Reid HF (1910) *The mechanics of the earthquake*, The California earthquake of April 18, 1906, Report of the state investigation commission, vol 2. Carnegie Institution of Washington, Washington, DC
- Rogers G, Dragert H (2003) Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Scienceexpress*. <https://doi.org/10.1126/science.1084783>
- Rubenstein JL, Shelley DR, Ellsworth WL (2010) Nonvolcanic tremor: a window into the roots of fault zones. In: Cloetingh S, Negendank J (eds) *New frontiers in integrated solid earth sciences*. International Year of Planet Earth. https://doi.org/10.1007/978-90-481-2737-5_8
- Scholz C (2002) *The mechanics of earthquakes and faulting*, 2nd edn. Cambridge University Press, Cambridge, p 471
- Segall P (2010) *Earthquake and volcano deformation*. Princeton University Press, Princeton
- Shelly DR, Beroza GC, Ide S, Nakamura S (2006) Low frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. *Nature* 442:188–191

Earthquakes in the Himalaya

Dibyashakti Panda¹, Bhaskar Kundu¹ and Vineet K. Gahalaut²

¹Department of Earth and Atmospheric Sciences, NIT Rourkela, Rourkela, India

²National Geophysical Research Institute (CSIR), Hyderabad, India

Definition

Earthquake: Sudden release of elastic strain energy along a fault.

Himalayan arc: About 2500 km long mountain chain running from Afghanistan, Pakistan, India, Nepal, China, and Myanmar, formed due to collision/subduction of Indian and Eurasian plate.

Main Himalayan thrust: The contact surface between the underthrusting Indian plate, and the overlying Himalayan wedge rocks.

Introduction

With the advent of modern concept of plate tectonics “*Himalaya-Tibetan orogenic system*,” a classical subduction-collision plate boundary zone has attracted the attention of geoscientists globally, who have considered it as a “*geodynamic black box*.” Himalayan geodynamics, its genesis, earthquake occurrence process, and associated lithospheric deformation have largely been considered within the framework of the collision of the Indian and Eurasian plates at some period during the Eocene (~50 Ma), resulting in the subduction of the Indian plate below Tibet. This has resulted in intense crustal shortening and deformation, taken up by crustal stacking along a system of intracontinental thrusts

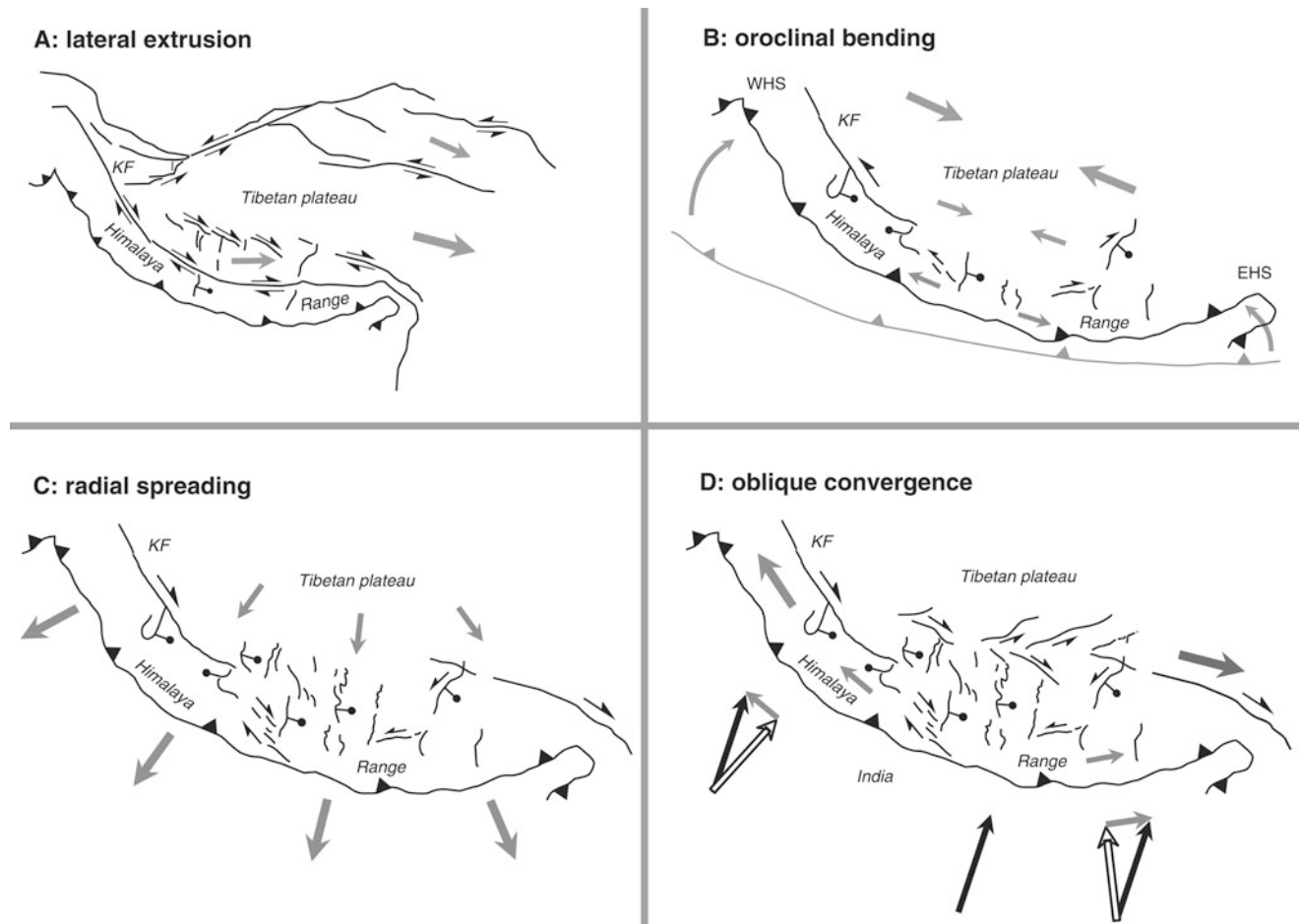
and internal deformation of adjacent Tibet and Indian plate (Murphy et al. 2009; Styron et al. 2011).

In the present entry, we have tried to focus on Himalayan earthquakes, in view of advances in knowledge about the past ~1000 years of earthquakes' occurrence processes, constrained by geodetic, historic, and seismological data, along the Himalayan arc in order to objectively assess seismic hazard of the region and adjoining Indo-Gangetic plains (~250 million hectares fertile plains), which is one of the most densely populated regions of the Southeast Asia accommodating ~40% of the India's population. In addition to that, significant high relief of the Himalayan Mountains blocks the northward monsoonal winds from the Indian Ocean during June–September, resulting in intense orography-controlled precipitation, and associated seasonal water mass accumulation and mass redistribution by the Himalayan river system, which dominate the regional surface loading process along the arc. Further, this seasonal non-tectonic deformation also influences the tectonic deformation along the Himalayan convergent plate boundary, making Himalayan earthquake

occurrence and associated deformation process much more complicated. Rather than repeating what has already been documented in a few excellent reviews (e.g., Bilham et al. 2017; Bilham 2019), we have attempted to add more information from the perspective of crustal deformation.

Geodynamic Models of Himalaya-Tibetan Orogenic System

The Himalaya-Tibetan orogeny has always remained a key domain for scientific community in terms of geodynamic evolution and tectonic complexity. Deformation pattern along this subduction-collision plate boundary has been explored widely (Copley and McKenzie 2007; Murphy et al. 2009; Styron et al. 2011); however, there are several issues which are still unexplored. Several geodynamic models have been proposed in order to explain the geometry and active tectonics along the Himalaya-Tibetan orogen (Styron et al. 2011). These models are (Fig. 1): (a) *Lateral extrusion*



Earthquakes in the Himalaya, Fig. 1 Geodynamic models for Himalayan and south Tibetan deformation. (After Styron et al. 2011). The gray arrows represent motion of the Tibetan crust for different models. The

black arrow in oblique convergence model represents the overall India-Southern Tibet convergence. KF: Karakorum Fault; EHS: Eastern Himalayan Syntaxis; WHS: Western Himalayan Syntaxis

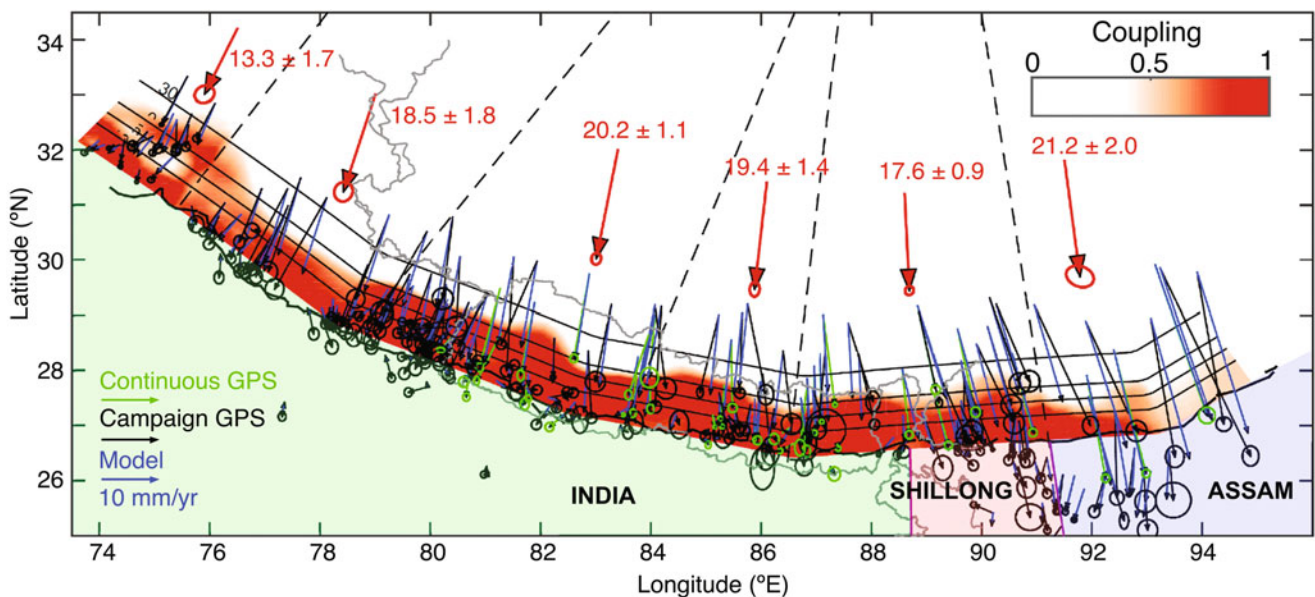
model that deals with the ductile flow and associated distributed deformation of the Tibetan crust; (b) *Oroclinal bending* which implies the curvature of the Himalayan arc and deformation associated with it; (c) *Lateral spreading* that is concerned with the outward radial expansion of the Tibetan Plateau; and finally (d) *Oblique convergence model* that deals with the variable obliquity in the convergence and deformation along the Himalayan arc (Styron et al. 2011; Li and Yin 2008; Murphy et al. 2009). Out of these four, there are two popular models that have been widely accepted in recent past, namely, lateral extrusion model and oblique convergence model.

Following the oblique convergence model in NW-Himalaya, Kundu et al. (2015) proposed that the India-southern Tibet oblique motion is being partitioning in the backarc as dextral motion along the Karakorum fault system, thereby developing the NW-Himalayan sliver. Similarly, upon applying the oblique convergence model in NE-Himalaya, one should expect a sliver with left lateral motion in the backarc. However, geomorphic offsets, geodetic estimates and the available earthquake data do not support the presence of any active left-lateral fault system in the region (Woerd et al. 2009). Thus, it appears that the oblique convergence model and lateral extrusion model are mutually contradictory in view of the tectonics and distributed deformation (or slip partitioning) in the NE-Himalaya. In order to resolve this issue, Panda et al. (2018a) proposed that instead of partitioning the backarc, the NE-Himalaya has developed an active sliver along the Assam-Brahmaputra

valley in the outer deformation front to accommodate the oblique motion through distributed deformation. This suggests the complex deformation behavior along the Himalayan arc and indicates that there is no unified model (or mechanism) that can explain the whole deformation process along the arc.

Crustal Deformation and Strain Buildup Along Himalayan Arc

In the past few decades with the advancement of geodetic techniques and relatively good spatial coverage of GPS measurements spanning the Himalayan arc, more insight has been gained into the deformation behavior and associated seismic potential along this subduction-collision plate boundary. GPS observations of surface displacement indicate that the Himalaya accommodates about ~ 20 mm/yr. of convergence rate between India and Eurasia (Stevens and Avouac 2015) (Fig. 2). However, the overall convergence rate decreases in the North-Western Himalaya to 13 mm/yr. (Kundu et al. 2014; Schiffman et al. 2013) and North-Eastern Himalaya to 16 mm/yr. (Panda et al. 2018a; Vernant et al. 2014). All along the Himalayan arc, the Main Himalayan Thrust (MHT) appears to be fairly uniformly locked from the surface to beneath the front of the high ranges of the Higher Himalaya over a width of 100 ± 20 km, which suggests high seismic hazard in the frontal ranges (Stevens and Avouac 2015). The Himalayan range has indeed produced a number of great earthquakes recently (e.g.,



Earthquakes in the Himalaya, Fig. 2 Interseismic coupling model (marked by red) of the Himalayan arc derived from GPS measurements. A coupling value of 1 means the portion is fully locked, while 0 means the region is fully creeping. Black and blue arrows represent observed

and modelled GPS velocities with respect to fixed India reference frame. Red arrows are the averaged velocities of the particular segment along the arc, which are separated by dashed lines. (After Stevens and Avouac 2015)

M7.8 Gorkha 2015, M6.9 Sikkim 2011, and M7.6 Kashmir 2005) and even larger historical earthquakes (e.g., M8.5 Assam 1950; Chen and Molnar 1990 and M8.2 1934 Nepal-Bihar earthquakes; Sapkota et al. 2013). However, still more than two-thirds of the Himalayan arc has not experienced a great earthquake in the past 200 years (Bilham and Wallace 2005) and most of the unruptured Himalayan segments may have the potential to generate great earthquakes (Bilham et al. 2001).

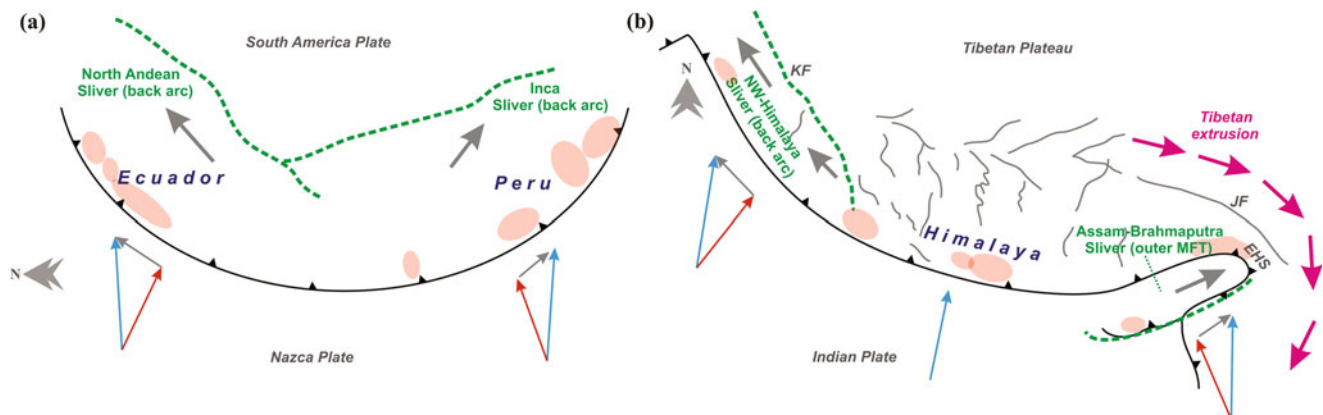
In case of central Himalaya (i.e., Nepal Himalaya), all the ~ 20 mm/yr. of convergence is accommodated through thrust motion and such high strain rate is consistent with higher seismic hazards as compared to the adjacent NE and NW-Himalaya (Stevens and Avouac 2015). In case of NW-Himalaya (i.e., Kashmir Himalaya), GPS observations indicate the overall convergence between India-southern Tibet to be ~ 13 mm/yr. and the remaining oblique motion is accommodated through dextral strike slip motion along the Karakorum fault system in the backarc forming the NW-Himalayan sliver (Kundu et al. 2014) (Fig. 3). It appears that the convergence rate in the Kashmir frontal Himalaya is about 25% less than that in the Nepal Himalayan region and so is the seismic potential, assuming that the entire convergence occurs through stick-and-slip motion without any significant aseismic motion. Similarly, in the NE-Himalaya (i.e., Bhutan and Arunachal region), the overall convergence rate accommodated in the Himalayan arc is about 20–25% less than that in the neighboring central Himalaya (Panda et al. 2018a; Vernant et al. 2014).

Panda et al. (2018a) explained such low convergence rate in NE-Himalaya due to the presence of Assam-Brahmaputra valley sliver in the outer deformation front (Fig. 3). Presence of

such sliver in the outer deformation front may be attributed to the low strength of the crustal material and therefore it tentatively implies a lower seismic hazard. However, tectonics and earthquake occurrence processes of the two Himalayan syntaxes are complex and relatively less understood, due to poor accessibility and limited investigations on crustal deformation patterns. GPS measurements of crustal deformation across the eastern Himalayan syntaxis (EHS) indicate ~ 20 mm/yr. of convergence across it due to India-Eurasia oblique convergence (Devachandra et al. 2014). Convergence across the EHS is accommodated through stick-and-slip process, where the 65 km wide main thrust zone of the plate boundary interface is locked, which significantly increases the seismic vulnerability of this region.

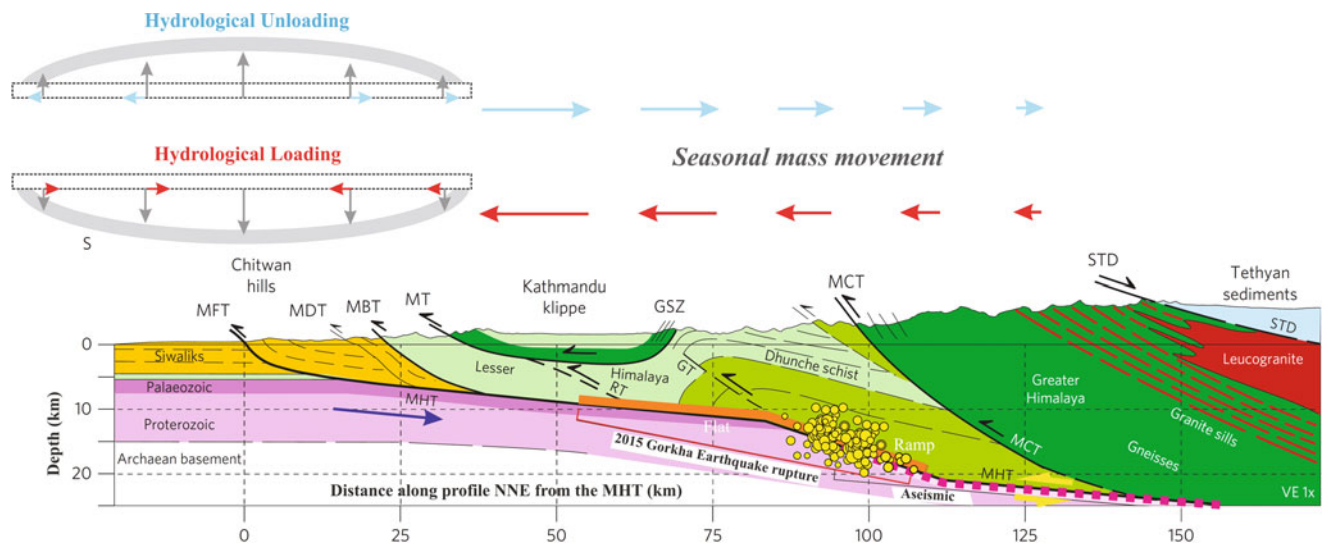
Mid-Crustal Seismicity and Associated Seismic Swarm Along Main Himalayan Thrust

Interseismic strain accumulation process across the Himalaya (Lave and Avouac 2000; Ader et al. 2012; Stevens and Avouac 2015) is also associated with the micro-seismicity (also called as mid-crustal seismicity), which occurs at a depth of 10–20 km on the Main Himalayan Thrust, MHT (Fig. 4). These smaller magnitude earthquakes (M 2–4) are considered to occur on and above the base of seismogenic MHT (Bollinger et al. 2007). However, the very accurately located aftershocks of the 2015 Gorkha earthquake provided evidence that they are mainly associated with the duplex structure (Mendoza et al. 2019). Earthquake focal mechanism solutions, which mainly describe the nature of faulting and orientation of the fault planes,



Earthquakes in the Himalaya, Fig. 3 Comparison between backarc and outer deformation-front strain partition in two different convergent plate margins. (a) In South America, the Nazca plate is subducting beneath the South American plate. Due to the arcuate geometry of the arc, the overall motion (blue arrow) is partitioned into arc parallel (gray arrow) and arc perpendicular (red arrow) component. The arc perpendicular component contributes to convergence, while the arc parallel component partitioned in the backarc on either side of the arc to produce two largest continental slivers (i.e., Ecuador and Peru sliver) on the

backarc. (b) Indian plate is subducting beneath the Eurasia plate and due to the shape of the Himalayan arc, the motion should partition in the backarc on either side of the arc. In NW-Himalaya, the oblique motion is partitioned on the backarc as dextral motion along the Karakorum fault system, thereby creating the NW-Himalayan sliver. However, in NE-Himalaya, due to presence of strong eastward flow of the Tibetan crust in the backarc, the motion is partitioned on the outer deformation front along the Assam-Brahmaputra valley. KF: Karakorum Fault; JF: Jiali Fault; EHS: Eastern Himalayan Syntaxis. (After Panda et al. 2018a)



Earthquakes in the Himalaya, Fig. 4 Conceptual model describing the effect of the hydrological mass oscillation along the North-South profile of Nepal-Himalaya region. (After Panda et al. 2018b). Yellow symbols represent mid-crustal seismicity at the base of the seismogenic zone (i.e., mid-crustal ramp). GSZ: Galchi shear zone; GT: Gorkha blind

Thrust; MHT: Main Himalayan Thrust; STD: South Tibetan Detachment; MCT: Main Central Thrust; MT: Munsiri/Mahabharat Thrust; RT: Ramgarh Thrust; MBT: Main Boundary Thrust; MDT: Main Dum Thrust; MFT: Main Frontal Thrust

indicate predominantly thrust motion on the shallow dipping MHT (Rajaure et al. 2013). This narrow and spatially distinguishable seismicity belt roughly coincides with the topographic front of the Himalaya in Nepal (i.e., 3.5 km contour), also considered as the down-dip edge of the locked portion of the MHT (Avouac 2003; Bollinger et al. 2004). A few magnetotelluric studies confirm the presence of fluids at the base of seismogenic MHT, where the micro-seismicity occurs (e.g., Lemonnier et al. 1999; Rawat et al. 2014; Hoste-Colomer et al. 2018). A few earthquake swarms have been reported from Nepal Himalaya (Hoste-Colomer et al. 2017) and Garhwal-Kumaun Himalaya (Paul and Sharma 2011; Singh et al. 2018), which also imply the presence of fluids in these regions.

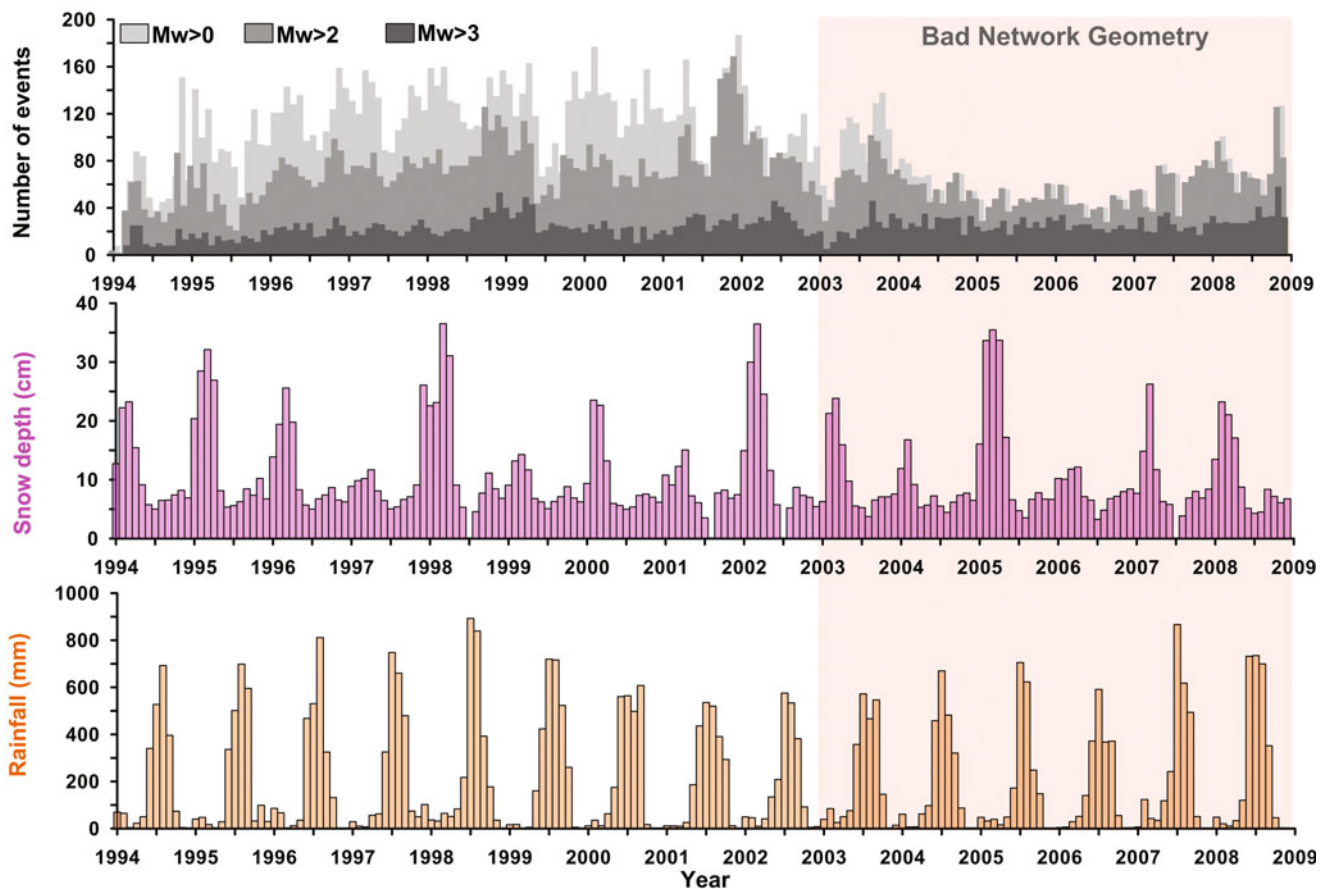
As discussed above, the micro-seismicity (or mid-crustal seismicity) belt along the Himalaya is located on the base of the seismogenic MHT and on the duplex structure above it, which is considered to be the transition between the velocity-weakening (favoring earthquakes) to velocity-strengthening (suppressing earthquakes) domain or conditionally stable frictional regimes. This segment is very much sensitive to external stress perturbations by climatic loading. So, there is a possibility that these micro-earthquakes may be modulated by external periodic load perturbations. Bettinelli et al. (2008) proposed seasonal modulation of micro-seismicity in the Himalaya, which are controlled by surface water load variations. The seismicity rate is twice as high in the winter as in the summer, and correlates with stress rate variations due to continental water storage oscillations (Bollinger et al. 2007; Bettinelli et al. 2008). Although the stress change due to such external surface load variations are very small, but presence of high pore pressure at seismogenic depth may cause seismicity

modulation that is controlled by seasonal variations. Similarly, Kundu et al. (2017) found strong annual periodicity in the mid-crustal seismicity, in response to the annual stress variation of hydrological loads. They reported strong annual periodicity in the eastern Nepal as compared to that in the western Nepal, with prominent annual periodicity in the small magnitude earthquakes (Kundu et al. 2017). Rainfall and seismicity cross correlation is also significantly higher in the eastern Nepal region as compared to that in the western Nepal with a prominent phase lag of 5–6 months with respect to the occurrence of monsoonal rainfall (Kundu et al. 2017) (Fig. 5). These studies confirm that during the post-monsoon period (i.e., during winter), significant component of horizontal compression is added to the interseismic contraction at the MHT, which is the main driving mechanism for the seasonal modulation of micro-seismicity in Nepal Himalaya (Kundu et al. 2017).

Factors Influencing Himalayan Earthquakes

Role of Subducted Ridge

It has been reported extensively in several literatures that subduction of diverse type of bathymetric features (e.g., seamounts, aseismic ridges, fractures, banks and rises, etc.) influences the location and frequency of great earthquakes in subduction zones and also significantly alters the arc morphology (Rosenbaum and Mo 2011; Robinson et al. 2006; Bilek 2007; Das and Watts 2009; Watts et al. 2010). In this context, it appears that spatial extents of historic and recent great Himalayan earthquakes are also influenced by at least three major subsurface ridges (e.g.,



Earthquakes in the Himalaya, Fig. 5 Temporal correlation between mid-crustal seismicity, snow depths in Higher Himalaya and rainfall in the surrounding region. Different gray bars represent different magnitudes of mid-crustal seismicity. Note that the seismicity peaks correlate

with the peak rainfall with a phase lag of ~ 6 months. Light shade region represents the period of bad network geometry, which is excluded from the analysis. (After Kundu et al. 2017)

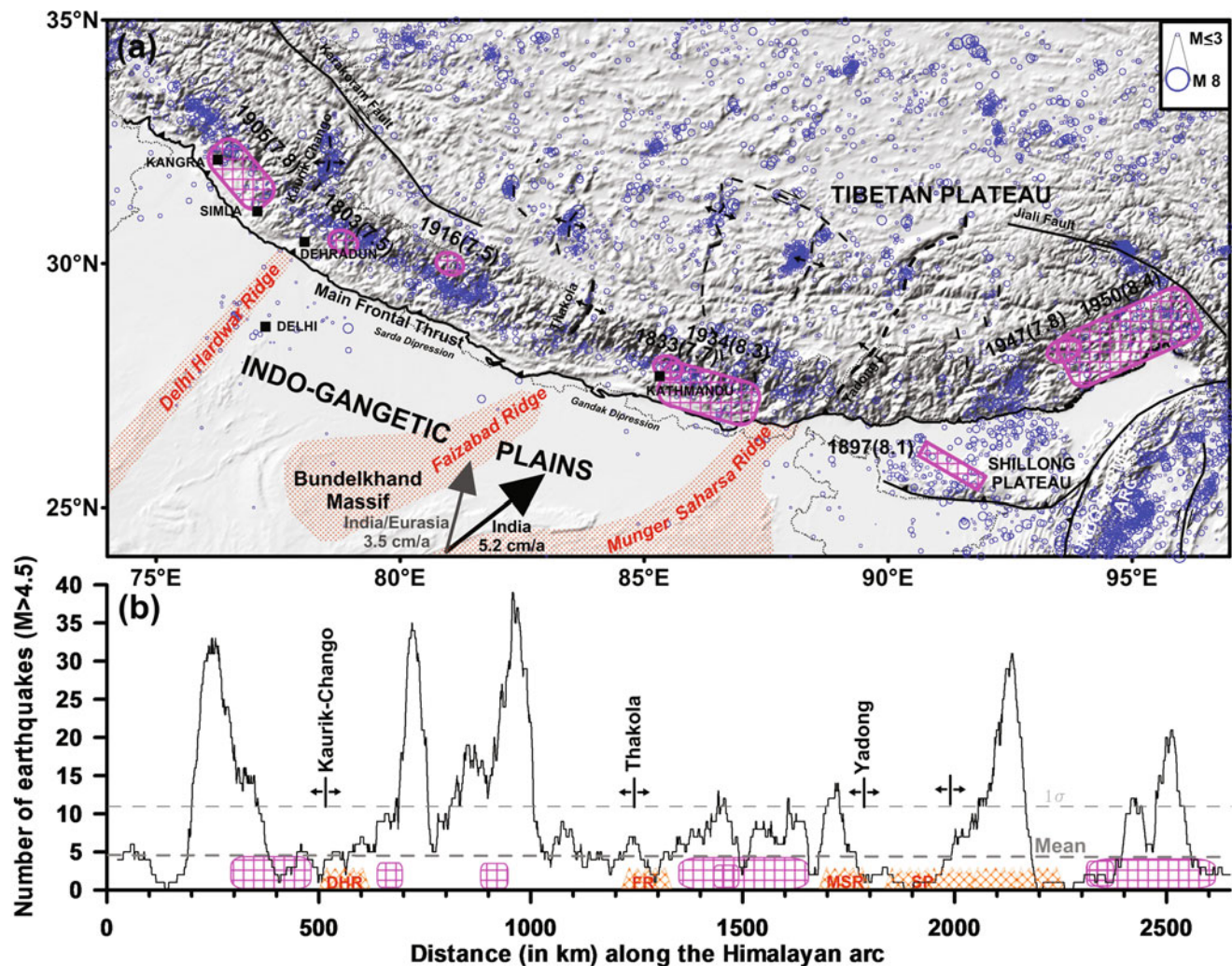
from east to west Monghyr Saharsa ridge, Faizabad ridge, Delhi Hardwar ridge) that have been mapped on the underthrusting Indian plate beneath the Indo-Gangetic plains (Valdiya 1976; Gahalaut and Kundu 2012) (Fig. 6). It has been proposed that the subduction of the two most prominent ridges namely the Chagos-Laccadive-Deccan ridge and the 90°E Ridge contributed to the development of the arcuate shape of the Himalayan arc. The interaction and subduction of the other subsurface ridges probably influenced the Himalayan arc morphology by causing a localized cusp in the frontal topography and further these ridges probably acted as barriers to the ruptures of the major and great Himalayan earthquakes (Gahalaut and Kundu 2012). These barriers are considered to be the regions of low friction which do not allow the rupture to propagate further, thereby arresting the rupture propagation.

Stress Triggering Process

Generally, earthquakes have potential to modify the stress regime of the critically stressed fault system, extending into brittle crust predominantly over a variety of spatiotemporal

scales by three partially overlapping categories: (a) static stress (stress change decays rapidly with distance (d) from the epicenter as d^{-3}); (b) quasi-static stress (decay as d^{-2}); and (c) dynamic stress (decay as $d^{-1.5}$ for surface waves) (Prejean and Hill 2014).

Such earthquake triggering phenomenon is also reported from the Himalayan arc in a few studies (Gahalaut et al. 2011; Mendoza et al. 2016). It has been argued that the seismicity of the Bhutan Himalaya region is generally low as compared to its adjacent Himalayan segments. It has been suggested that the lower seismicity in the Bhutan Himalaya may be due to static stress change by the great 1897 Shillong Plateau earthquake (Gahalaut et al. 2011). The low seismicity region of the Bhutan Himalaya clearly coincides with the static stress shadow of the 1897 Shillong Plateau earthquake. (Fig. 7a, b) In addition to that, seismically active Kopili Fault segment in Assam-Brahmaputra valley and some of the dextrally slipping transverse faults system of Himalayan region coincides with the positive static stress change, caused by the 1897 earthquake rupture (Fig. 7c, d). Therefore, it appears that



Earthquakes in the Himalaya, Fig. 6 Seismicity of the Himalayan arc, general geology, rifts in the Tibetan plateau and subsurface ridges in the Indo-Gangetic plains. (After Gahalaut and Kundu 2012). (a) Blue dots represent seismicity along the Himalaya and surrounding region. Different ridges are marked by light shaded color. Gray arrow represents

India-Eurasia convergence. (b) Variation of seismicity ($M > 4.5$) of the Himalayan Seismic Belt (HSB) in an overlapping and sliding 50 km window. DHR: Delhi Hardwar ridge; FR: Faizabad ridge; MSR: Munger-Saharsa ridge; MFT: main frontal thrust; MBT: main boundary thrust; MCT: main central thrust

1897 Shillong Plateau earthquake has altered stress regime of the Himalaya and adjacent segments up to some extent.

Besides the static stress triggering effect, Mendoza et al. (2016) reported the observation of remote dynamic triggering of local micro-earthquakes in the central Himalaya, caused by the teleseismic waves from the M8.5, 2007 Sumatra earthquake which occurred ~4500 km away from this region. It has been argued that the Himalayan fault systems in the central Himalaya region are extremely sensitive to small stress changes of about ~9 kPa of peak dynamic stresses imparted by the teleseismic waves (Mendoza et al. 2016). We are hopeful to detect more examples of dynamic triggering with better seismic station coverage in the Himalayan arc.

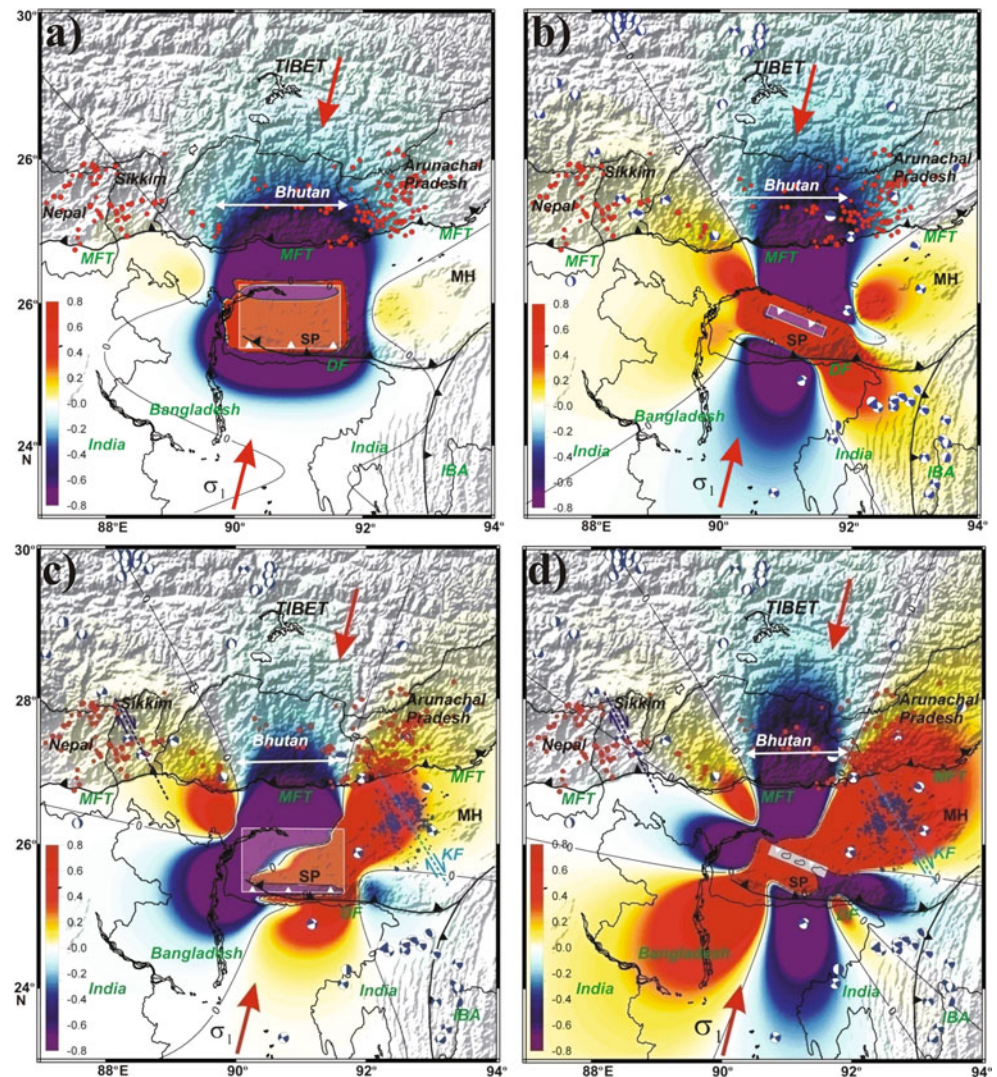
Non-tectonic Deformation

Besides tectonic process, there are several non-tectonic processes that can also significantly deform Earth's crust and may

cause seismicity modulation. These non-tectonic processes may include continental water storage (Blewitt et al. 2001; Bettinelli et al. 2008; Steckler et al. 2010; Fu et al. 2013; Chanard et al. 2014), tidal loading (Agnew 1996, 1997), ice and snow load (Matsuo and Heki 2010; Jiang et al. 2010), heavy precipitation (Hainzl et al. 2006), atmospheric pressure and disturbances (Kaniuth and Vetter 2006; Liu et al. 2009), sediment unloading by river incision (Calais et al. 2010), reservoir impoundment (Gupta 2002), etc.

Considering the continental water storage (or hydrological load) variation, the Amazon basin experiences the largest seasonal mass variation because of heavy rainfall, followed by the Southeast Asia where the hydrological load variations are controlled by strong monsoonal rainfall. The hydrological load variations are measured by Gravity Recovery and Climate Experiment (GRACE) and are expressed in terms of equivalent water height (EWH). Strong seasonal horizontal

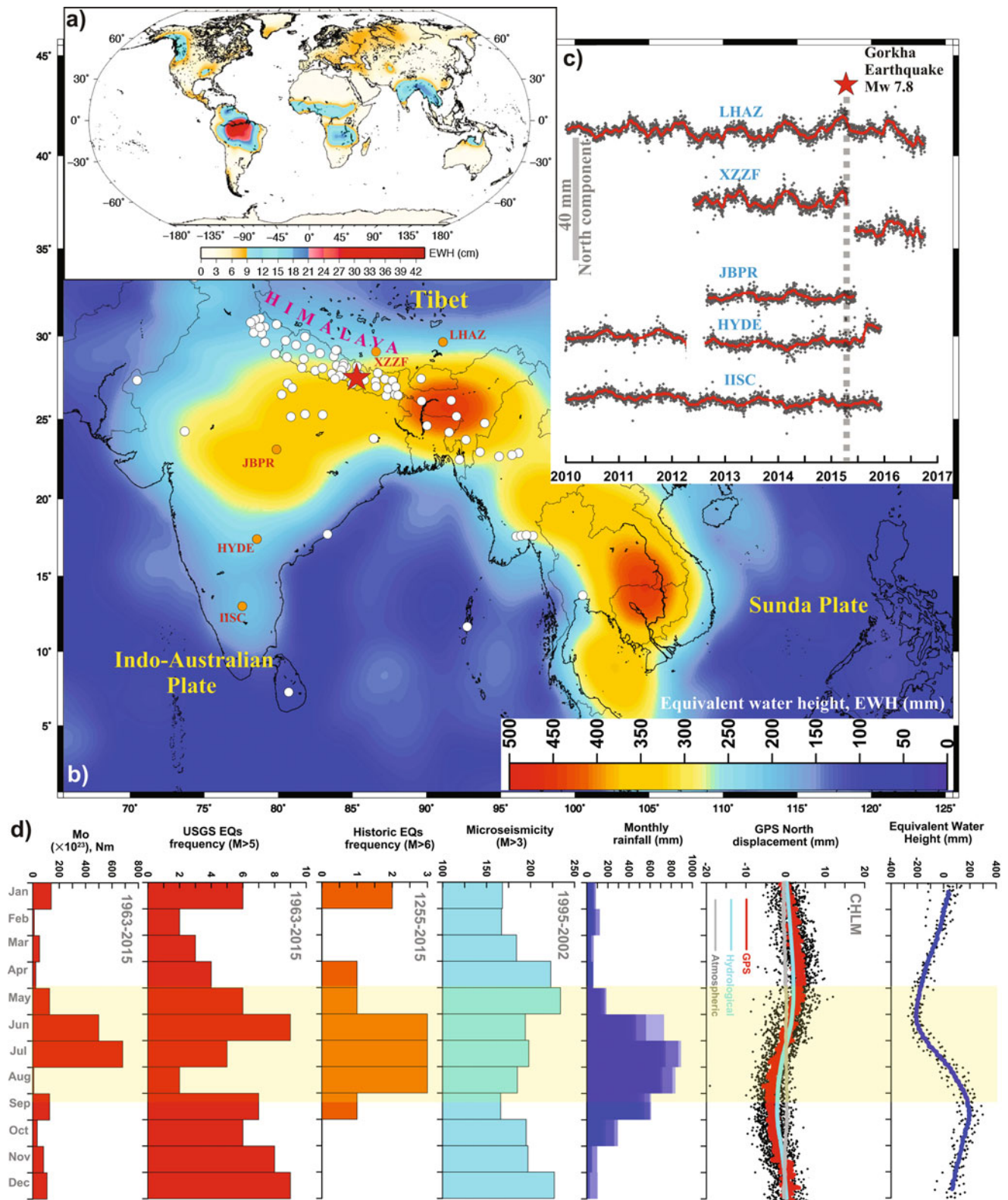
Earthquakes in the Himalaya, Fig. 7 Coulomb failure stress change (Δ CFS, in bar) on Himalayan detachment (~10 km depth) caused by 1897 Shillong Plateau (SP) earthquake. (After Gahalaut et al. 2011). (a) Δ CFS on optimally oriented planes using earthquake rupture model of Gahalaut and Chander (1992). (b) Δ CFS on optimally oriented thrust plane using rupture model of Bilham and England (2001). (c, d) Coulomb failure stress change (Δ CFS, in bar) on dextrally slipping Kopili Fault (KF) (~25 km depth) caused by 1897 Shillong Plateau (SP) earthquake



crustal motions are observed by GPS measurements in case of Southeast Asia (Fu et al. 2013) (Fig. 8). However, there are some discrepancies between the GRACE and GPS observations in terms of amplitude and phase of seasonal transients at a few sites in southwest Nepal and Kumaun-Garhwal Himalaya due to local effects of mountain valley, hydrology, groundwater irrigation, and reservoir impoundments (Fu et al. 2013; Panda et al. 2018b; Gahalaut et al. 2017; Gautam et al. 2017).

Hydrological load over Southeast Asia, which is the second largest one after the Amazon basin, lies close and above the Himalayan plate boundary region and may significantly influence the tectonic process along it (Fig. 8). However, the interaction between seasonally induced non-tectonic and tectonic deformation along it remains debated. Panda et al. (2018b) proposed that tectonic deformation along this plate boundary can be significantly influenced by the deformation induced by the non-tectonic hydrological loading cycles. They reported substantially higher transient displacements

in GPS time series that are located above the base of the seismogenic zone, i.e., mid-crustal ramp on the MHT (Panda et al. 2018b). Such higher seasonal transients in the GPS time series indicate change in aseismic slip rate (or slow-slip) on the deep megathrust that may be controlled by seasonal hydrological loading (Panda et al. 2018b). Further, they invoked the fault resonance process in order to explain the higher transient slip at the base of the seismogenic zone (i.e., mid-crustal ramp). It appeared that slip velocity perturbation is several folds higher on the mid-crustal ramp than on the flat during hydrological loading episodes. This suggests that the Himalayan ramp is more sensitive to seasonal modulation than the updip flat segment. Therefore, micro-seismicity and abnormally higher transient in GPS displacement are favored near the transition from velocity-weakening to velocity-strengthening behavior at the base of the Himalayan seismogenic zone (i.e., on the mid-crustal ramp), which are modulated by the seasonal stress changes (Panda et al. 2018b).



Earthquakes in the Himalaya, Fig. 8 Seasonal deformation of Southeast Asia and its influence over tectonic deformation and seismicity along the Himalaya. (After Panda et al. 2018b). (a) Global seasonal mass variation of equivalent water height (EWH, in cm) measured by GRACE. (b) Average peak-to-peak annual surface load variations over Southeast Asia, in terms of equivalent water height (EWH, in mm)

measured by GRACE for the period 2002–2012. (c) Small inset represents detrended GPS time series (north component) of five representative sites from China and India. (d) Correlation between equivalent water height (EWH), rainfall, GPS displacement, and different seismicity catalogues

Human-Induced Deformation

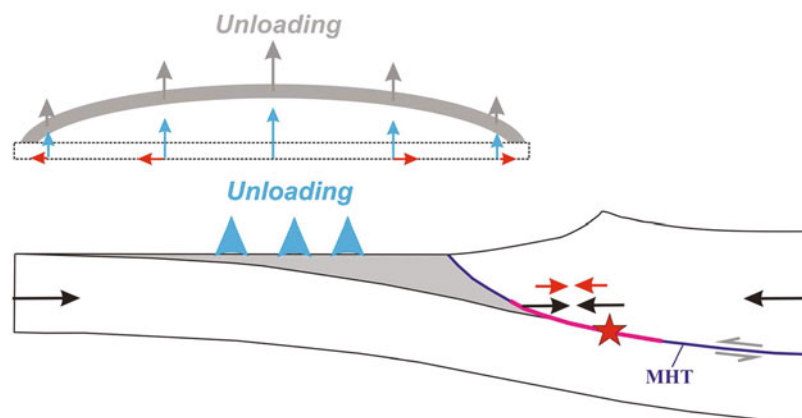
Although, mechanism of faulting and earthquake occurrence process along the Himalayan arc is predominantly tectonically driven process, it has been suggested that the anthropogenic crustal (un-)loading process can also drive long-term fault slip or modulate seismicity (Kundu et al. 2015). In fact, in the past few decades' identification of human-induced earthquakes and associated deformation have become focal points in sociopolitical and scientific discussion (Gonzalez et al. 2012; Kundu et al. 2015; Foulger et al. 2017). More than 700 cases have been documented in the human-induced earthquake database, HiQuake (Foulger et al. 2017; Wilson et al. 2017) and the 25 April 2015, M7.8 Gorkha earthquake, Nepal too has been considered as a human-induced earthquake in that database (Kundu et al. 2015). Using Gravity Recovery and Climate Experiment data and hydrological models, Kundu et al. (2015) have claimed that the thrust earthquakes on the Main Himalayan Thrust (MHT), including the 2015 Mw 7.8 Gorkha, Nepal earthquake, are probably influenced by the steady decline of annual groundwater level over the Indo-Gangetic plain at the rate of rate of $17.7 \pm 4.5 \text{ km}^3/\text{yr}$ over the Indian states of Rajasthan, Punjab, and Haryana (Rodell et al. 2009) (Fig. 9).

Impoundments of reservoir for hydropower production, due to its substantial volume and associated water load-induced deformation (mechanical reservoir load and pore-elastic diffusion process), can also significantly alter the stress regime of critically stressed faults system in the vicinity. They have been considered to have triggered strong earthquakes. The impoundment of Lake Mead and Hoover dam near Boulder City, Nevada, in 1935 and the seismicity in its vicinity is probably the first reported case of reservoir-triggered seismicity (RTS), and since then several cases have been reported worldwide (Gupta 1992). However, no

reservoir-triggered seismicity has been reported so far from the Himalayan region, which consists of at least nine reservoirs. However, theoretical study (Chander and Kalpna 1996) indicates that even a thrust fault can be destabilized by the impoundment of reservoir and the possibility of RTS cannot be ruled out completely from the Himalaya. Recently, a new promising case of reservoir triggered seismicity of the Garhwal region of NW Himalaya, influenced by Tehri reservoir load has been reported (Gahalaut et al. 2017, 2018a). Combined GPS and InSAR measurements indicate that the seasonal water level variations in the Tehri reservoir cause crustal deformation (Gahalaut et al. 2017). Precisely relocated small earthquakes in the Garhwal region, reservoir water level variation at the dam site, and simulation of reservoir-induced stress incorporating pore-pressure diffusion confirm that it is yet another case of RTS in an active thrust faulting environments of the NW-Himalaya (Gahalaut et al. 2018b).

Future Directions: Key Questions to Be Answered

Spatiotemporal variations in seismicity and crustal deformation along the Himalayan arc and adjacent domains are primarily controlled by variation in convergence rate, lithology, rheological properties, thermal structure, topographic features, and structural aspect between the two plates (Gahalaut and Kundu 2012; Arora et al. 2008). Spatiotemporal variations in the mid-crustal seismicity production and crustal movements occur on annual, seasonal, or diurnal basis and can be induced by monsoon, ice-snow load, atmospheric pressure, tidal loading, continental water storage and its fluctuation, and anthropogenic process (Blewitt et al. 2001; Bettinelli et al. 2008; Flouzat et al. 2009; Chanard et al.



Earthquakes in the Himalaya, Fig. 9 Schematic diagram representing the effect of unloading (by anthropogenic groundwater loss) in the Indo-Gangetic plains on the MHT (Bettinelli et al. 2008; Kundu et al. 2015). Crustal unloading in the Indo-Gangetic plains will induce some component of horizontal compression (red arrows) that

adds to the secular interseismic contraction (black arrows) at seismogenic depth on the MHT. Red star represents hypocenter of the 2015 Gorkha, Nepal earthquake and the associated rupture is shown by the pink line along the MHT. (After Kundu et al. 2015)

2014; Ellsworth 2013). Specifically, the regional loading process along the Himalayan arc and adjacent Indo-Ganga plains are complicated by several factors and quantification of tectonic vs. non-tectonic deformation process in the Himalayan arc and adjacent region is a challenge.

Displacement time series observed by the continuous Global Positioning System (cGPS) have revealed that surface load fluctuation can induce surface strain (Wahr et al. 2013; Heki 2004; Bevis et al. 2005; van Dam et al. 2007; Amos et al. 2014). A significant variation of surface loading can be the result of various processes, which influence the seismicity (see above). All of these processes cause the earth lithosphere to deform and can induce significant surface strain. Therefore, cGPS measurements for surface deformation can be used as a proxy for surficial mass distribution. Seasonal hydrological effects are the main candidates for global mass fluctuation process and cause periodic crustal deformation (Bevis et al. 2005; van Dam et al. 2007). In addition, space geodetic techniques can quantify the surface hydrological mass fluctuation with the Gravity Recovery and Climate experiment (GRACE). cGPS-derived vertical seasonal fluctuation have been investigated extensively and compared with GRACE-derived seasonal loading deformation (Fu et al. 2013). Strong seasonal variations of horizontal and vertical components are noticed on cGPS time series in elastic response to heavy rain fall and monsoonal variation in Southeast Asia.

The regional loading process along the Himalayan arc and adjacent Indo-Ganga plains are complicated by several factors. The phase relationship of loading to precipitation is not simple, as suggested by many authors (Wahr et al. 2013; Fu et al. 2013). For the Amazon basin, much of the water mass there is temporarily absorbed into groundwater, which delays its transport out of the system (in the Amazon, by as much as 3 months). GRACE data may also be difficult to relate directly to the cGPS observations, because it averages over such large scale that we could not be able to distinguish between northward accumulating mass and southward mass-transit.

Therefore, in future, we must focus on the nature of deformation process along the Himalayan arc and adjacent Indo-Ganga plains region through the combined geodetic and seismic measurements with the following key objectives (Fig. 8):

1. Understand the role of the Indo-Ganga plains (i.e., surface hydrological load) in the Himalayan deformation process and mid-crustal seismicity generation.
2. Understand the complicated nature of Himalayan loading process (e.g., orographic effect, monsoonal effect, river transportation effect, and flexural effect due to surface hydrological variation in the Indo-Ganga plains).
3. How can we separate the non-tectonic signals from the geodetic and seismic observations? It will also help us to

understand the effect of strong seasonal effects on the cGPS and seismic time series.

4. Does the MHT experience tremor and slow-slip events? Does it get modulated by climatic load? How does the slow-slip event affect the adjacent locked mega-thrust zones?

The explorations of these questions will help us in understanding the earthquake occurrence and deformation process, slip rate, recurrence interval of mega-thrust earthquakes, and strain budget along the Himalayan arc region.

Cross-References

- Earthquakes and Crustal Deformation
- GPS, Tectonic Geodesy
- Orogenic Belts

Bibliography

- Ader T, Avouac JP, Liu-Zeng J, Lyon-Caen H, Bollinger L, Galetzka J, Genrich J, Thomas M, Chanard K, Sapkota SN, Rajaure S, Shrestha P, Ding L, Flouzat M (2012) Convergence rate across the Nepal Himalaya and interseismic coupling on the Main Himalayan thrust: implications for seismic hazard. *J Geophys Res Sol Earth* 117: B04403. <https://doi.org/10.1029/2011JB009071>
- Agnew D (1996) SPOTL: Some Programs For Ocean-Tide Loading, SIO Reference Ser:96–98
- Agnew D (1997) NLOADF: a program for computing ocean-tide loading. *J Geophys Res* 102:5109–5110
- Amos CB, Audet P, Hammond WC, Bürgmann R, Johanson IA, Blewitt G (2014) Contemporary uplift and seismicity in Central California driven by groundwater depletion. *Nature* 509:483–486
- Arora BR, Kumar A, Rawat G, Kumar N, Choubey VM (2008) First observations of free oscillations of the earth (FOE) from indian superconducting gravimeter in Himalaya. *Curr Sci* 95(11):1611–1617
- Avouac JP (2003) Mountain building, erosion and the seismic cycle in the Nepal Himalaya. *Adv Geophys* 46. [https://doi.org/10.1016/S0065-2687\(03\)46001-9](https://doi.org/10.1016/S0065-2687(03)46001-9)
- Bettinelli P, Avouac JP, Flouzat M, Bollinger L, Ramillien G, Rajaure S, Sapkota S (2008) Seasonal variations of seismicity and geodetic strain in the Himalaya induced by surface hydrology. *Earth Planet Sci Lett* 266:332–344
- Bevis M, Alsdorf D, Kendrick E, Fortes LP, Forsberg B, Smalley R, Becker J (2005) Seasonal fluctuations in the mass of the Amazon River system and Earth's elastic response. *Geophys Res Lett* 32: L16308. <https://doi.org/10.1029/2005GL023491>
- Bilek SL (2007) Influence of subducting topography on earthquake rupture. In: Dixon T, Moore C (eds) *The Seismogenic Zone of Subduction Thrust Faults*. Columbia University Press, pp 123–146
- Bilham R (2019) Himalayan earthquakes: a review of historical seismicity and early 21st century slip potential. In: Treloar PJ, Searle MP (eds) *Himalayan tectonics: a modern synthesis*. Geological Society, Special publications, vol 483, London. <https://doi.org/10.1144/SP483.16>
- Bilham R, England P (2001) Plateau pop-up in the 1897 Assam earthquake. *Nature* 410:806–809

- Bilham R, Wallace K (2005) Future $M_w > 8$ earthquakes in the Himalaya: implications from the 26 Dec 2004 $M_w = 9.0$ earthquake on India's eastern plate margin. In: Kangra 1905 earthquake centenary seminar. Geological Survey of India, Special publication, vol 85, Palampur, pp 1–14
- Bilham R, Gaur VK, Molnar P (2001) Himalayan seismic hazard. *Science* 293:1442–1444
- Bilham R, Mencin D, Bendick R, Bürgmann R (2017) Implications for elastic energy storage in the Himalaya from the Gorkha 2015 earthquake and other incomplete ruptures of the Main Himalayan thrust. *Quat Int.* <https://doi.org/10.1016/j.quaint.2016.09.055>
- Blewitt G, Lavallee D, Clarke P, Nurutdinov K (2001) A new global mode of earth deformation: seasonal cycle detected. *Science* 294:2342–2345
- Bollinger L, Avouac JP, Cattin R, Pandey MR (2004) Stress build up in the Himalaya. *J Geophys Res* 109:B11405. <https://doi.org/10.1029/2003JB002911>
- Bollinger L, Perrier F, Avouac JP, Sapkota S, Gautam U, Tiwari DR (2007) Seasonal modulation of seismicity in the Himalaya of Nepal. *Geophys Res Lett* 34. <https://doi.org/10.1029/2006GL029192>
- Calais E, Freed AM, Van Arsdale R, Stein S (2010) Triggering of New Madrid seismicity by late Pleistocene erosion. *Nature* 466:608–611
- Chanard K, Avouac JP, Ramillien G, Genrich J (2014) Modeling deformation induced by seasonal variations of continental water in the Himalaya region: sensitivity to Earth elastic structure. *J Geophys Res Solid Earth* 119:5097–5113
- Chander R, Kalpna C (1996) Probable influence of Tehri reservoir load on earthquakes of the Garhwal Himalaya. *Current Sci* 70:291–299
- Chen WP, Molnar P (1990) Source parameters of earthquakes and intraplate deformation beneath the Shillong plateau and the northern Indoburman ranges. *J Geophys Res* 95(B8):12,527–12,552
- Copley A, McKenzie D (2007) Models of crustal flow in the India-Asia collision zone. *Geophys J Int* 169:683–698
- Das S, Watts AB (2009) Effect of subducting seafloor topography on the rupture characteristics of great subduction zone earthquakes. In: Lallemand S, Funicello F (eds) *Subduction zone geodynamics*. Springer, Berlin/Heidelberg, pp 103–118. <https://doi.org/10.1007/978-3-540-87974-9>
- Devachandra M, Kundu B, Catherine J, Kumar A, Gahalaut VK (2014) Global positioning system (GPS) measurements of crustal deformation across the frontal eastern Himalayan syntaxis and seismic-hazard assessment. *Bull Seismol Soc Am* 104(3):1518–1524
- Ellsworth WL (2013) Injection-induced earthquakes. *Science* 341(6142). <https://doi.org/10.1126/science.1225942>
- Flouzat M, Bettinelli P, Willis P, Avouac JP, Heritier T, Gautam U (2009) Investigating tropospheric effects and seasonal position variations in GPS and DORIS time-series from the Nepal Himalaya. *Geophys J Int* 178:1246–1259
- Foulger GR, Wilson MP, Gluyas JG, Julian BR, Davies RJ (2017) Global review of human-induced earthquakes. *Earth Sci Rev* 178:438–514
- Fu Y, Argus DF, Freymueller JT, Heflin MB (2013) Horizontal motion in elastic response to seasonal loading of rain water in the Amazon Basin and monsoon water in Southeast Asia observed by GPS and inferred from GRACE. *Geophys Res Solid Earth* 40:6048–6053
- Gahalaut VK, Chander R (1992) A rupture model for the great earthquake of 1897. Northeast India. *Tectonophysics* 204:163–174
- Gahalaut VK, Kundu B (2012) Possible influence of subducting ridges on the Himalayan arc and on the ruptures of great and major Himalayan earthquakes. *Gondwana Res* 21:1080–1088
- Gahalaut VK, Rajput S, Kundu B (2011) Low seismicity in the Bhutan Himalaya and the stress shadow of the 1897 Shillong plateau earthquake. *Phys Earth Planet Int* 186:97–102
- Gahalaut VK, Yadav RK, Sreejith KM, Gahalaut K, Bürgmann R, Agrawal R, Sati SP, Bansal A (2017) InSAR and GPS measurements of crustal deformation due to seasonal loading of Tehri reservoir in Garhwal Himalaya, India. *Geophys J Int* 209(1):425–433
- Gahalaut K, Gupta S, Gahalaut VK, Mahesh P (2018a) Influence of Tehri reservoir impoundment on local seismicity of northwest Himalaya. *Bull Seismol Soc Am* 108:3119–3125
- Gahalaut K, Gupta S, Gahalaut VK, Mahesh P (2018b) Influence of Tehri reservoir impoundment on local seismicity of NW Himalaya. *Bull Seismol Soc Am.* <https://doi.org/10.1785/0120180077>
- Gautam PK, Gahalaut VK, Prajapati SK, Kumar N, Yadav RK, Rana N, Dabral CP (2017) Continuous GPS measurements of crustal deformation in Garhwal-Kumaun Himalaya. *Quat Int.* <https://doi.org/10.1016/j.quaint.2017.05.043>
- Gonzalez PJ, Tiampo KF, Palano M, Cannavò F, Fernández J (2012) The 2011 Lorca earthquake slip distribution controlled by groundwater crustal unloading. *Nat Geosci* 5:821–825
- Gupta HK (1992) *Reservoir-induced earthquakes*. Elsevier, Amsterdam. 364 pp. ISBN: 9780444597359
- Gupta H (2002) A review of recent studies of triggered earthquakes by artificial water reservoirs with special emphasis on earthquakes in Koyna, India. *Earth Sci Rev* 58:279–310
- Hainzl S, Kraft T, Wassermann J, Igel H, Schmedes E (2006) Evidence for rainfall-triggered earthquake activity. *Geophys Res Lett* 33: L19303
- Heki K (2004) Dense GPS array as a new sensor of seasonal changes of surface loads, in the state of the planet. *Frontiers and Challenges in Geophysics. Geophys Monogr Ser* 150:177–196
- Hoste-Colomer R, Bollinger L, Lyon-Caen H, Burtin A, Adhikari LB (2017) Lateral structure variations and transient swarm revealed by seismicity along the Main Himalayan thrust north of Kathmandu. *Tectonophysics* 714:107–116
- Hoste-Colomer R, Bollinger L, Lyon-Caen H, Adhikari LB, Baillard C, Benoit A, Bhattarai M, Gupta RM, Jacques E, Kandel T, Koirala BP (2018) Lateral variations of the midcrustal seismicity in western Nepal: seismotectonic implications. *Earth Planet Sci Lett* 504:115–125
- Jiang Y, Dixon T, Wdowski S (2010) Accelerating uplift in the North Atlantic region as an indicator of ice loss. *Nat Geosci* 3(6):404–407
- Kaniuth K, Vetter S (2006) Estimating atmospheric pressure loading regression coefficients from GPS observations. *GPS Solut* 10(2):126–134
- Kundu B, Yadav RK, Bali BS, Chowdhury S, Gahalaut VK (2014) Oblique convergence and slip partitioning in the NW Himalaya: implications from GPS measurements. *Tectonics* 33:2013–2024
- Kundu B, Vissa NK, Gahalaut VK (2015) Influence of anthropogenic groundwater unloading in indo-Gangetic plains on the 25 April 2015 M_w 7.8 Gorkha, Nepal earthquake. *Geophys Res Lett* 42:10,607–10,613
- Kundu B et al (2017) Influence of a meteorological cycle in mid-crustal seismicity of the Nepal Himalaya. *J Asian Earth Sci* 146:317–325
- Lave J, Avouac JP (2000) Active folding of abandoned fluvial terraces across the Siwalik Hills (Nepal). *J Geophys Res* 105:5735–5770
- Lemonnier C, Marquis G, Perrier F, Avouac JP, Chitrakar G, Kafle B, Sapkota S, Gautam U, Tiwari D, Bano M (1999) Electrical structure of the Himalaya of Central Nepal: high conductivity around the mid-crustal ramp along the MHT. *Geophys Res Lett* 26:3261–3264
- Li D, Yin A (2008) Orogen-parallel, active left-slip faults in the eastern Himalaya: implications for the growth mechanism of the Himalayan Arc. *Earth Planet Sci Lett* 274:258–267
- Liu CC, Linde AT, Sacks IS (2009) Slow earthquakes triggered by typhoons. *Nature* 459:833–836
- Matsuo K, Heki K (2010) Time-variable ice loss in Asian high mountains from satellite gravimetry. *Earth Planet Sci Lett* 290:30–36
- Mendoza MM, Ghosh A, Rai SS (2016) Dynamic triggering of small local earthquakes in the central Himalaya. *Geophys Res Lett* 43:9581–9587
- Mendoza MM, Ghosh A, Karplus MS, Klemperer SL, Sapkota SN, Adhikari LB, Velasco A (2019) Duplex in the main Himalayan thrust

- illuminated by aftershocks of the 2015 Mw 7.8 Gorkha earthquake. *Nat Geosci* 12(12):1018–1022
- Murphy MA, Saylor JE, Ling D (2009) Landscape evolution of South-west Tibet based on integrated paleo elevation reconstructions and structural history. *Earth Planet Sci Lett* 282:1–9
- Panda D, Kundu B, Santosh M (2018a) Oblique convergence and strain partitioning in the outer deformation front of NE Himalaya. *Sci Rep* 8(1):10564
- Panda D, Kundu B, Gahalaut VK, Bürgmann R, Jha B, Asaithambi R, Yadav RK, Vissa NK, Bansal AK (2018b) Seasonal modulation of deep slow-slip and earthquakes on the Main Himalayan thrust. *Nat Commun* 9(1):4140
- Paul A, Sharma ML (2011) Recent earthquake swarms in Garhwal Himalaya: a precursor to moderate to great earthquakes in the region. *J Asian Earth Sci* 42(6):1179–1186
- Prejean SG, Hill DP (2014) Earthquakes, dynamic triggering of, in encyclopedia of complexity and systems science. Springer, New York, pp 1–36
- Rajaure S et al (2013) Double difference relocation of local earthquakes in the Nepal Himalaya. *J Nepal Geol Soc* 46:133–142
- Rawat G, Arora BR, Gupta PK (2014) Electrical resistivity cross-section across the Garhwal Himalaya: proxy to fluid-seismicity linkage. *Tectonophysics* 637:68–79
- Robinson DP, Das S, Watts AB (2006) Earthquake rupture stalled by a subducting fracture zone. *Science* 321:1203
- Rodell M, Velicogna I, Famiglietti JS (2009) Satellite-based estimates of groundwater depletion in India. *Nature* 460:999–1003
- Rosenbaum G, Mo W (2011) Tectonic and magmatic responses to the subduction of high bathymetric relief. *Gondwana Res* 19:571–582
- Sapkota SN, Bollinger L, Klinger Y, Tapponnier P, Gaudemer Y, Tiwari D (2013) Primary surface ruptures of the great Himalayan earthquakes in 1934 and 1255. *Nat Geosci* 6:71–76
- Schiffman C, Bali BS, Szeliga W, Bilham R (2013) Seismic slip deficit in the Kashmir Himalaya from GPS observations. *Geophys Res Lett* 40:5642–5645
- Singh R, Prasath RA, Paul A, Kumar N (2018) Earthquake swarm of Himachal Pradesh in northwest Himalaya and its seismotectonic implications. *Phys Earth Planet Int* 275:44–55
- Steckler MS, Nooner SL, Akhter SH, Chowdhury SK, Bettadpur S, Seeber L, Kogan MG (2010) Modeling earth deformation from monsoonal flooding in Bangladesh using hydrographic, GPS, and gravity recovery and climate experiment (GRACE) data. *J Geophys Res* 115:B08407. <https://doi.org/10.1029/2009JB007018>
- Stevens VL, Avouac JP (2015) Interseismic coupling on the main Himalayan thrust. *Geophys Res Lett* 42(14):5828–5837
- Styron RH, Taylor MH, Murphy MA (2011) Oblique convergence, arc-parallel extension, and the role of strike-slip faulting in the high Himalaya. *Geosphere* 7:582–596
- Valdiya KS (1976) Himalayan transverse faults and folds and their parallelism with subsurface structures of north Indian planes. *Tectonophysics* 32:353–386
- van Dam T, Wahr J, Lavallée D (2007) A comparison of annual vertical crustal displacements from GPS and gravity recovery and climate experiment (GRACE) over Europe. *J Geophys Res* 112:B03404. <https://doi.org/10.1029/2006JB004335>
- Vernant P, Bilham R, Szeliga W, Drupka D, Kalita S, Bhattacharyya AK, Gaur VK, Pelgay P, Cattin R, Berthet T (2014) Clockwise rotation of the Brahmaputra Valley relative to India: tectonic convergence in the eastern Himalaya, Naga Hills, and Shillong plateau. *J Geophys Res Solid Earth* 119(8):6558–6571
- Wahr J et al (2013) The use of GPS horizontals for loading studies, with applications to northern California and Southeast Greenland. *J Geophys Res* 118:1795–1806
- Watts AB, Koppers AAP, Robinson D (2010) Seamount subduction and earthquakes. *Oceanography* 23:166–173
- Wilson MP, Foulger GR, Gluyas JG, Davies RJ, Julian BR (2017) HiQuake: the human-induced earthquake database. *Seismol Res Lett* 88:1560–1565
- Woerd JV, Leloup PH, Zeng JL, Lascassin R, Tapponnier P (2009) A comment on “Orogen-parallel, active left-slip faults in the eastern Himalaya: implications for the growth mechanism of the Himalayan arc” by Li and Yin. *Earth Planet Sci Lett* 274:258–267

Earthquakes on Oceanic Transform Faults, Unexpectedly Large

Shamita Das

Department of Earth Sciences, University of Oxford, Oxford, UK

Definition

Transform faults are conservative plate boundaries, where material is neither created nor destroyed. The most prominent of these worldwide are oceanic transform faults such as the mid-ocean ridges.

Introduction

One of the key pieces of evidence that supported the theory of plate tectonics in the 1960s was the explanation of how plates move along mid-oceanic transform faults (Wilson 1965). The six possible types of dextral (right-lateral) transform faults are shown in Fig. 1a. Six sinistral (left-lateral) transform faults, which are mirror images of the ones plotted, also can exist. The appearance of the right-lateral systems after a period of growth is shown in Fig. 1b. Note that length of the active transform fault is growing in three cases, decreasing in one case, and indeterminate in two cases, as indicated in Fig. 1b.

The most common transform fault has a ridge-ridge configuration (see Fig. 1a top left-most), in which the oceanic transform fault is bounded at both ends by spreading ridges, with the portion of the previously active transform fault beyond the ridges, called “fracture zones” being practically aseismic. A beautiful example of such transform faults is seen along the Mid-Atlantic ridge. In regions where the spreading ridge is no longer active, the fracture zones are termed “fossil fracture zones.”

Explanation of the focal mechanisms along the ridge and transform faults along mid-ocean ridges was first given by Sykes (1967), and a schematic is shown in Fig. 2. Very occasionally earthquakes do occur on the fracture zone, and have focal mechanisms with reversed polarity from the nearby oceanic transform fault (Engeln et al. 1986). This very useful fact can be used to identify such earthquakes.

Earthquakes on Oceanic Transform Faults, Unexpectedly Large, Fig. 1

(a) Illustration of the six kinds of right-lateral transform faults, with well-known examples indicated on some more common ones. The three kinds of such faults that can have faulting on conjugate planes are located inside the box indicated by the green dashed lines. These are mainly discussed here. Three other possible transform faulting settings are also shown, for completeness. (b) The same as (a) but after some period of growth

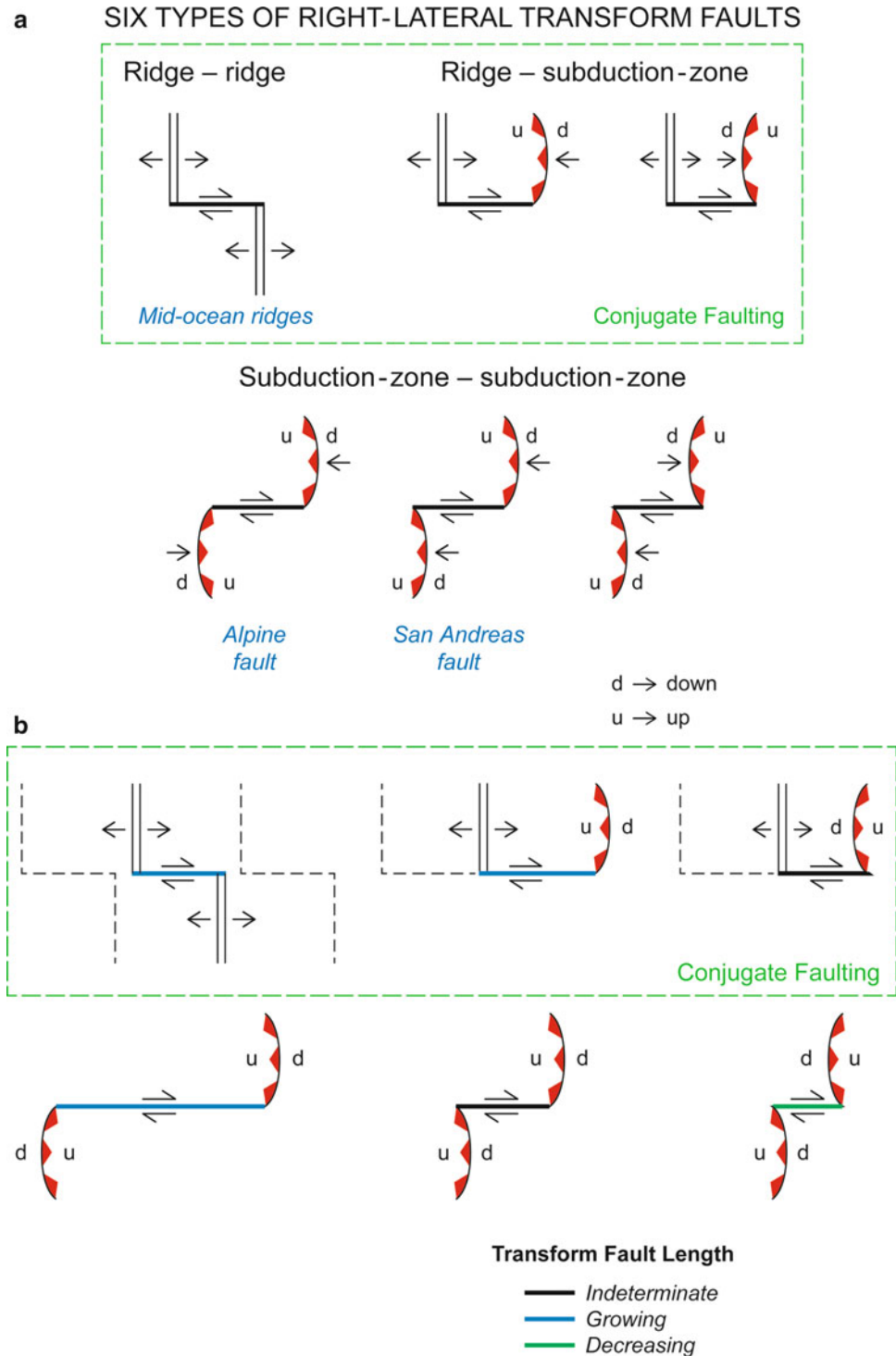


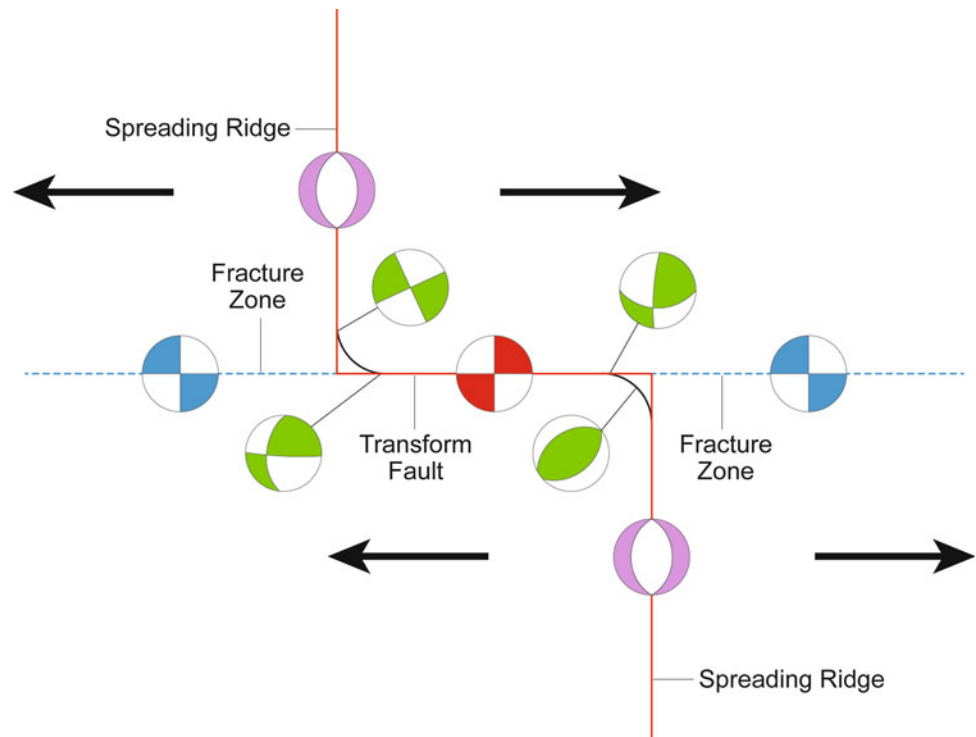
Figure 2 also shows that the situation is even more complex with other kinds of mechanisms occurring at the “inner corners.”

That fracture zones are essentially aseismic can be seen by examining earthquakes large enough to have a centroid moment-tensor solution (CMT can be accessed at [http://](http://www.globalcmt.org/CMTsearch.html)

www.globalcmt.org/CMTsearch.html). By counting such earthquakes located within ~1000 km of (as an example) the Mid-Atlantic Ridge between the Equator and 70°N (just north of Iceland), for the 34 year period from 1976 to 2009 inclusive, Robinson (2011) showed that of the 1002 CMT solutions reported, 266 are strike-slip. Of these 266, only 8 are

Earthquakes on Oceanic Transform Faults, Unexpectedly Large,

Fig. 2 Schematic representation of the classical picture at mid-ocean (spreading) ridges, showing the focal mechanisms. Black arrows indicate relative plate motions and red lines the active plate boundary. The red focal mechanism shows the typical ones expected on the (active) transform fault, and the purple ones the typical ones on the spreading ridges. The dashed blue lines are the continuations of the transform fault that are usually visible in bathymetric surveys of the seafloor. The blue focal mechanisms are the ones for the rare earthquakes on these and have strike-slip mechanism reverse to that on the active transform. Both strike-slip and thrust earthquakes are seen in the inner corners and are in green



unambiguously identified as having occurred on the fracture zone by virtue of having focal mechanisms with reversed polarity relative to the nearest transform fault. These rare fracture zone earthquakes had negligible seismic moment compared to the strike-slip earthquakes on the nearby transform faults.

Several unusual cases of large or great earthquakes on oceanic transform faults, including a M_w 8.1 on a fossil fracture zone and a M_w 8.0 pure strike-slip earthquake, which are nearly normal to the existing oceanic transform faults (i.e., fractures along the preexisting ridge fabric) as well as several large mid-oceanic earthquakes with simultaneous conjugate faulting have occurred in the last ≈ 30 years. Table 1 lists all such earthquakes, showing that though the classical theory of plate tectonics says that oceanic transform faults do not usually have “large” or “great” earthquakes, they do occur both on the active transform, on the fracture zone, and, in one case, even on a fossil fracture zone.

The major features of the earthquakes are summarized in Table 1. For a more detailed description of some of these earthquakes, with color figures showing the tectonic settings and rupture processes, see Das (2019). These earthquakes suggest that the oceanic lithosphere is breaking up along the ridge-transform conjugate fault systems under tectonic stresses, that is, the oceanic plates are not rigid as was believed in the early days of plate tectonics.

Rupture on Conjugate Faults in an Oceanic Setting

The intra-plate oceanic sequence of the 1987–1992 Gulf of Alaska earthquakes occurred in a region with no known earthquakes since 1964. The first one occurred on an E-W trending fault (Pegler and Das 1996), with the second one, the largest of the sequence, having its epicenter near the eastern end of the fault zone of the first one. However, its rupture occurred on a N-S trending fault plane. The third one occurred at the southern end of the fault of the second one. Pegler and Das (1996) identified the plane of rupture as having the same N-S trend as the second one. The last, and smallest, earthquake of the sequence occurred at the southernmost end of the N-S fault mentioned above. The aftershocks of the second and third earthquakes, in addition to lying on the main N-S fault plane, activated aftershocks on nearly conjugate NE and SW planes. Using body-wave analysis, magnetic anomaly, and bathymetric data, Pegler and Das (1996) interpreted the sequence as reactivation of preexisting nearly conjugate planes of weakness, related to a fossil ridge-transform system of faults, on the Pacific plate near where it subducts under Alaska. The 2018 Gulf of Alaska earthquake occurred to the west of this sequence, and shows a cross-hatched pattern (United States Geological Survey website) similar to the 1987–1992 sequence, again suggesting a similar interpretation.

The great 1989 Macquarie Ridge earthquake was studied by Das (1992, 1993) using body-wave analysis and the

Earthquakes on Oceanic Transform Faults, Unexpectedly Large, Table 1 Significant earthquakes on oceanic transform faults

	Date (month/day/year)	Location	M_w	Faulting mode
1.	1987–1992	Gulf of Alaska	7.2, 7.8, 7.7, 6.8	Conjugate faults
2.	05/23/1989	Macquarie Ridge	8.2	Conjugate fault was reactivated by aftershocks
3.	03/25/1998	Antarctic Plate	8.0	Cuts across preexisting ridge fabric
4.	06/18/2000	Wharton Basin	7.8	Simultaneous rupture on transform fault and ridge fabric
5.	12/23/2004	Tasman Sea	8.1	Great earthquake on a fossil fracture
6.	04/11/2012	Wharton Basin	8.6, 8.2	Same as 4
7.	01/23/2018	Gulf of Alaska	7.9	Conjugate faults reactivated by aftershocks

inversion technique of Das and Kostrov (see Das (2020), for details of the method). It was shown that the main shock reactivated a dormant conjugate fault, which is seen in the local bathymetry and magnetic anomaly data.

Rupture on a Fault Normal to the Preexisting Zones of Weakness

At the time of its occurrence, the great 1998 Antarctic Plate earthquake was the largest crustal submarine strike-slip earthquake ever recorded. It took place in a region where no seismicity has previously been recorded and was the largest known on the entire Antarctic Plate. Its occurrence is still puzzling as its rupture zone does not coincide with any known feature on the ocean floor, all of which are nearly N-S trending in this region. Even more puzzling, the rupture is nearly E-W trending and actually cuts across these features at right angles, that is, close to the ridge fabric (Henry et al. 2000). A full discussion on the rupture of this earthquake is given by Das (2020, this encyclopedia).

Simultaneous Earthquake Rupture on Conjugate Faults in an Oceanic Setting

The 2000 Wharton Basin earthquake was the first such example known to us. Analysis of broadband teleseismic data and relocated seismicity was used by Robinson et al. (2001) to show that this intraplate earthquake, in the region of diffuse deformation separating the Indian and Australian plates, consisted of two subevents that simultaneously ruptured two near-conjugate planes. This mode of rupture accommodates shortening by a mechanism different from that previously known elsewhere in the region. The larger subevent occurred on a N-S fossil fracture zone, with a relatively high stress drop of about 20 megapascals, showing that large stresses can accumulate in such regions of distributed deformation. The second sub-earthquake occurred on the E-W preexisting ridge fabric. The twin 2012 Wharton Basin earthquakes in a nearby region may have behaved in a similar way, with simultaneous rupture on multiple conjugate planes (Meng et al. 2012).

A Great Earthquake on a Fossil Fracture Zone

Reactivation of fossil fracture zones by nearby large earthquakes had generally been found to produce earthquakes of smaller magnitudes ($M_w < 6$), as shown earlier in the entry. So the M_w 8.0 2004 South Tasman Sea earthquake, away from the Australia-Pacific plate boundary, was very surprising. It was also the fifth largest strike-slip earthquake worldwide since 1900 (see Table 1 of Robinson 2011), to which 2012 M_w 8.6 Sumatra earthquake of Table 1 of this entry now needs to be appended at the top). Inversion of broad-band P- and SH-wave data (Robinson 2011) showed that rupture occurred on a curved fault, with differing strike, slip, and rake on the portions lying to the north and south of the epicenter. The occurrence of this earthquake shows that dead fracture zones can be reactivated in regions of very complex tectonic history with large internal deformation of the plate, to produce stresses required to generate great earthquakes.

Summary

In the classical theory of plate tectonics, oceanic transform faults do not usually have large earthquakes. However, the recent record of the occurrence of several such earthquakes shows that it does. Some of these earthquakes have fractured both conjugate nodal planes. Others have occurred on the fracture zone, and one on a fossil fracture zone. This shows that the oceanic crust is crumbling by fracture of nearly normal conjugate fault planes – the old planes of weakness being related to the ridge fabric and transform faults – even in regions where the transform fault is no longer active. Clearly oceanic plates are not as rigid as was believed in the early days of plate tectonics.

Cross-References

- [Continental Rifts](#)
- [Earthquake Rupture: The Inverse Problem](#)
- [Earthquake, Focal Mechanism](#)
- [Great Earthquakes](#)
- [Ocean, Spreading Center](#)
- [Seafloor Spreading](#)

Bibliography

- Das S (1992) Reactivation of an oceanic fracture by the Macquarie Ridge earthquake of 1989. *Nature* 357:150–153
- Das S (1993) The Macquarie Ridge earthquake of 1989. *Geophys J Int* 115:778–798
- Das S (2019) Unusual large earthquakes on oceanic transform faults. In: Bizzarri A, Das S, Petri A (eds) *Proceedings of the International School “Enrico Fermi”, Course 202*. IOS/SIF, Amsterdam/Bologna, pp 33–52. <https://doi.org/10.3254/978-1-61499-979-9-33>
- Das S (2020) Earthquake rupture: the inverse problem, *Enc Solid Earth Geophys*, in press
- Engeln JF, Wiens DA, Stein S (1986) Mechanisms and depths of Atlantic transform earthquakes. *J Geophys Res* 91:548–577
- Henry C, Das S, Woodhouse JH (2000) The great March 25, 1998 Antarctic Plate earthquake: moment tensor and rupture history. *J Geophys Res* 105:16,097–16,119
- Meng L, Ampuero J-P, Stock J, Duputel Z, Tsai VC (2012) Earthquakes in a maze: compressional rupture branching during the 2012 M_w 8.6 Sumatra earthquake. *Science* 337:724–726
- Pegler G, Das S (1996) The 1987–92 Gulf of Alaska earthquakes. *Tectonophysics* 257:111–136
- Robinson DP (2011) A rare great earthquake on an oceanic fossil fracture zone. *Geophys J Int* 186:1121–1134
- Robinson DP, Henry C, Das S, Woodhouse JH (2001) Simultaneous rupture along two conjugate planes of the Wharton Basin earthquake. *Science* 292:1145–1148
- Sykes LR (1967) Mechanism of earthquakes and nature of faulting on the mid-ocean ridges. *J Geophys Res* 72:2131–2153
- Wilson JT (1965) A new class of faults and their bearing on continental drift. *Nature* 207:343–347

Earthquakes, Did You Feel It?

David J. Wald, Vincent Quitoriano and James W. Dewey
U.S. Geological Survey, Denver Federal Center, Lakewood,
CO, USA

Definition and Introduction

The US Geological Survey (USGS) “Did You Feel It?”[®] (DYFI) system is an automated system for rapidly collecting macroseismic intensity data from Internet users’ shaking and damage reports and generating intensity maps immediately following earthquakes.

Although the collection and assignment of DYFI-based Macroscopic Intensity (MI) data depart from traditional assignments, they are made more quickly, provide more complete coverage at higher spatial resolution, offer citizen input and interaction, and allow data collection at rates and quantities that were not previously possible. These aspects of Internet-based data collection, in turn, allow for data analyses, graphics, and ways to communicate with the public, opportunities that were not feasible with traditional data-collection approaches. DYFI decimal MI assignments are based on regression against traditional MI, and geocoding allows for

spatial averaging; both of these strategies facilitate quantitative data analyses. DYFI can be found online at the website <http://earthquake.usgs.gov/dyfi/>.

Historical Context

Over the past two decades, the DYFI system has collected shaking and damage reports from Internet users immediately following earthquakes. The USGS has been operating DYFI since 1999 in California, since the early 2000s for the rest of the USA, and since the mid-2000s globally (Wald et al. 2011). DYFI has become vital for collecting macroseismic data for all felt seismic events in the USA, and it is also one of the most popular interactive websites within the Federal Government. For earthquakes occurring outside the USA, the worldwide DYFI data rapidly indicate or confirm earthquake occurrence for seismic analysts and scientists at the USGS National Earthquake Information Center (NEIC), giving a quick indication of the occurrence, extent, and nature of shaking effects. The global intensity data from DYFI, and in some cases its international counterparts, can be automatically used as constraints in the USGS ShakeMap system (Wald et al. 1999b), which is the hazard input for the USGS Prompt Assessment of Global Earthquakes for Response system (PAGER, Wald et al. 2006). USGS also allows users who experienced older events to enter DYFI questionnaire responses for important events such as the 1994 Northridge and 1989 Loma Prieta, California, earthquakes via the DYFI Web portal and earthquake archives.

Traditionally, intensities are assigned integer values by a classification process or, less frequently, a range of integer values such as “6–7” or “>6” for less-certain assignments (Musson et al. 2010). The DYFI process of assigning numerical values to macroseismic observations and then calculating real-valued intensities thus represents a philosophical as well as a procedural departure from traditional intensity-assigning methods. Over the years, analyses of the DYFI data strongly suggest that assigning intensities at higher precision is warranted (e.g., Atkinson and Wald 2007; Wald et al. 2011). USGS DYFI and ShakeMap intensities are now precise to one decimal place, and the discrete (I–X) intensity scale is replaced with a continuous, real-valued, scale. When DYFI and similar decimal intensities are plotted as a function of distance, the retention of information to tenths of an intensity unit results in lower scatter than is the case when intensities are assigned by classification and reported to the nearest integer (Wald et al. 2011). For map representation of DYFI decimal intensities, a continuous palette of colors is chosen using algorithms that automatically interpolate between discrete color values, using a similar technique to the one used by ShakeMap (Wald et al. 1999b). This allows us to present more subtle variations of intensity than previously achievable.

The evolution from manual/postal, to emailed, to Web-based macroseismic surveys has been accomplished in many regions of the world. Many countries either maintain the manual approach as the primary strategy or reserve the option to augment their web-based approaches with traditional assignments. Several very successful Internet-based macroseismic survey systems are now implemented in many countries or regions (see summaries in Wald et al. 2011; Goded et al. 2018).

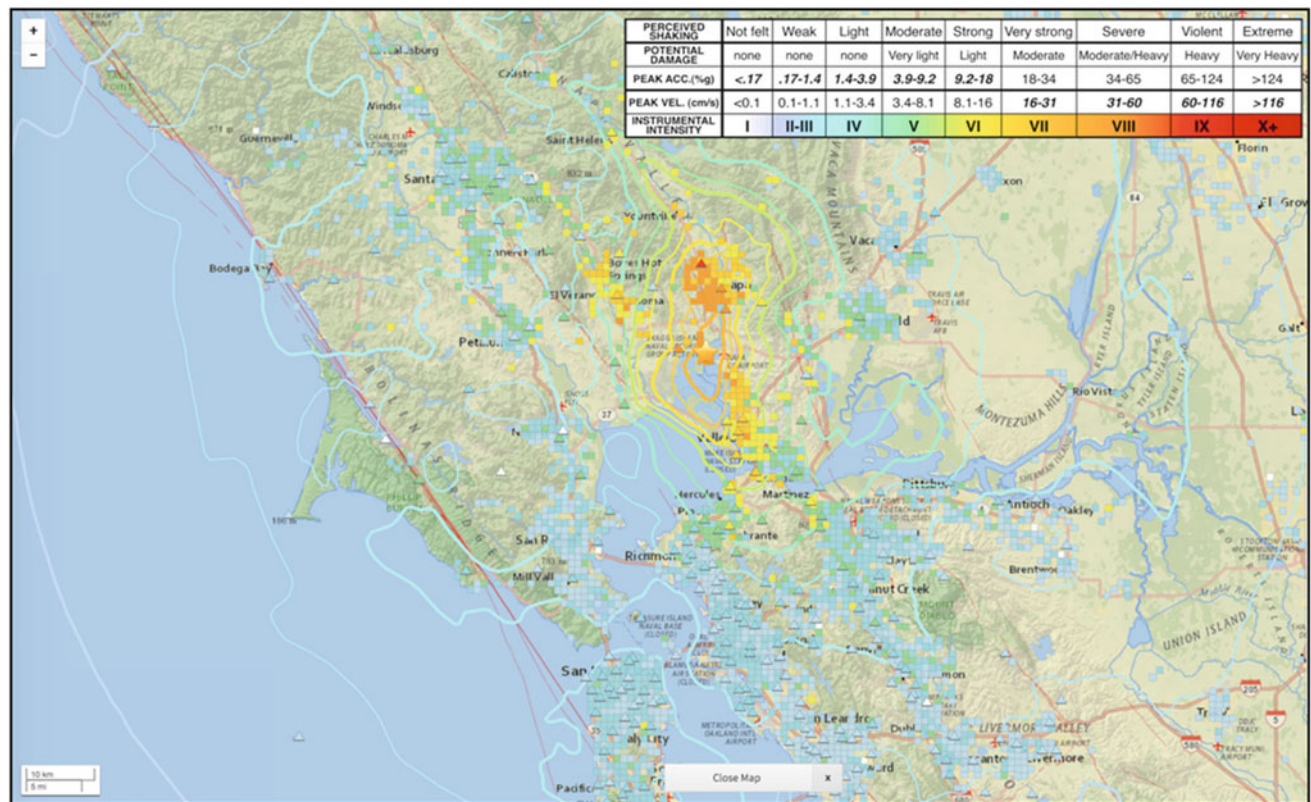
The DYFI System

The DYFI software is fully open source, written in Python, and available publicly through GitHub. Incoming entries from multiple web servers are processed and aggregated over both 1-km and 10-km geocoded boxes for each earthquake and are then used to make interactive event page maps and plots (Fig. 1). A natural challenge with a system such as DYFI is accommodating the post-earthquake deluge of input data and Web traffic. As DYFI has grown in popularity, the USGS has continuously improved capacity via both hardware and software improvements in order to handle the spike in Internet traffic following a widely felt earthquake. While new

responses come in continuously, the backend processes only refresh the online maps every 5 min to ensure that products are updated and transferred to the web before a new run begins. Response rate of DYFI report entries for the 2014 M6.0 South Napa, California, earthquake are shown in Fig. 2. Nearly 26,000 were received within the first hour and a total of 44,000 were ultimately received. Typically, about 60–90% of the entries are received within the first hour of an earthquake. For the largest earthquakes, response times peak at more than 30 responses per second. Figure 3 provides an indication of the variability of the data and its distribution with distance from the epicenter.

Data for widely felt earthquakes come in at a rate that allows confirmation of the earthquake’s occurrence within minutes. The DYFI maps evolve rapidly to informative macroseismic intensity distributions that are useful to the public, media, scientists, and even emergency managers, typically within tens of minutes. Statistics attest to the abundance and rapid availability of these Internet-based macroseismic data: nearly five million entries have been amassed over two decades, there are 64 events with more than 10,000 responses, and 550 events have over 1000 entries. The greatest number of responses for an earthquake is more than 146,000, for the 2011 M5.8 Mineral, Virginia event, which was felt by more

E

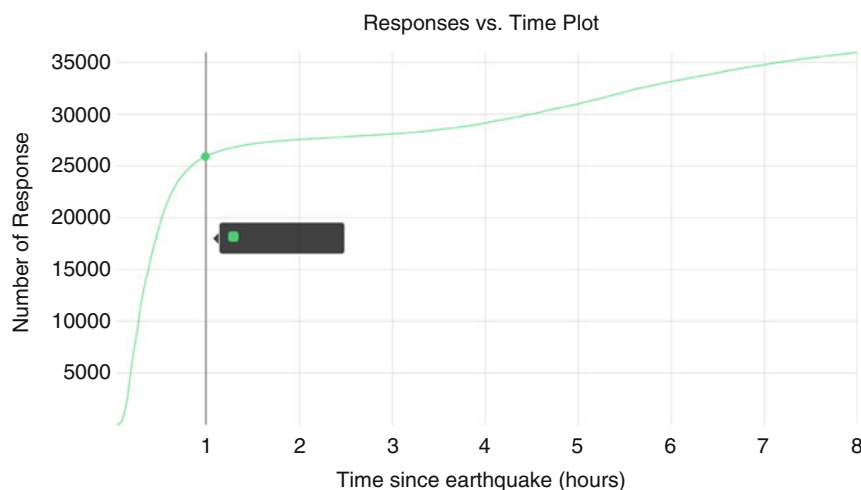


Earthquakes, Did You Feel It?, Fig. 1 Interactive map from the USGS event web pages for the 2014 magnitude 6.0 South Napa Valley, California, earthquake. DYFI data are displayed in 1-km square boxes color-coded to

the inset legend used by DYFI and ShakeMap to show the community decimal intensity (Wald et al. 2011). Independently derived ShakeMap intensity contours and seismic stations (triangles) are shown for comparison

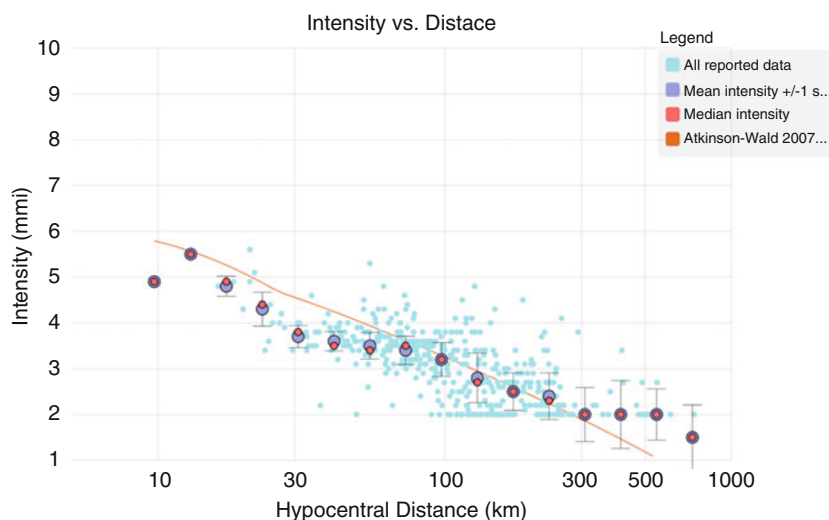
Earthquakes, Did You Feel It?,

Fig. 2 Response rate of DYFI report entries for the M6.0 South Napa earthquake



Earthquakes, Did You Feel It?,

Fig. 3 Plot of intensity versus distance from the epicenter for the M6.0 South Napa earthquake. The colored line is the predicted intensities based on the equations of Atkinson and Wald 2007, for magnitude 6.0. Small circles are individual geocode box intensity values; larger circles are averages over distance bins equal in log space

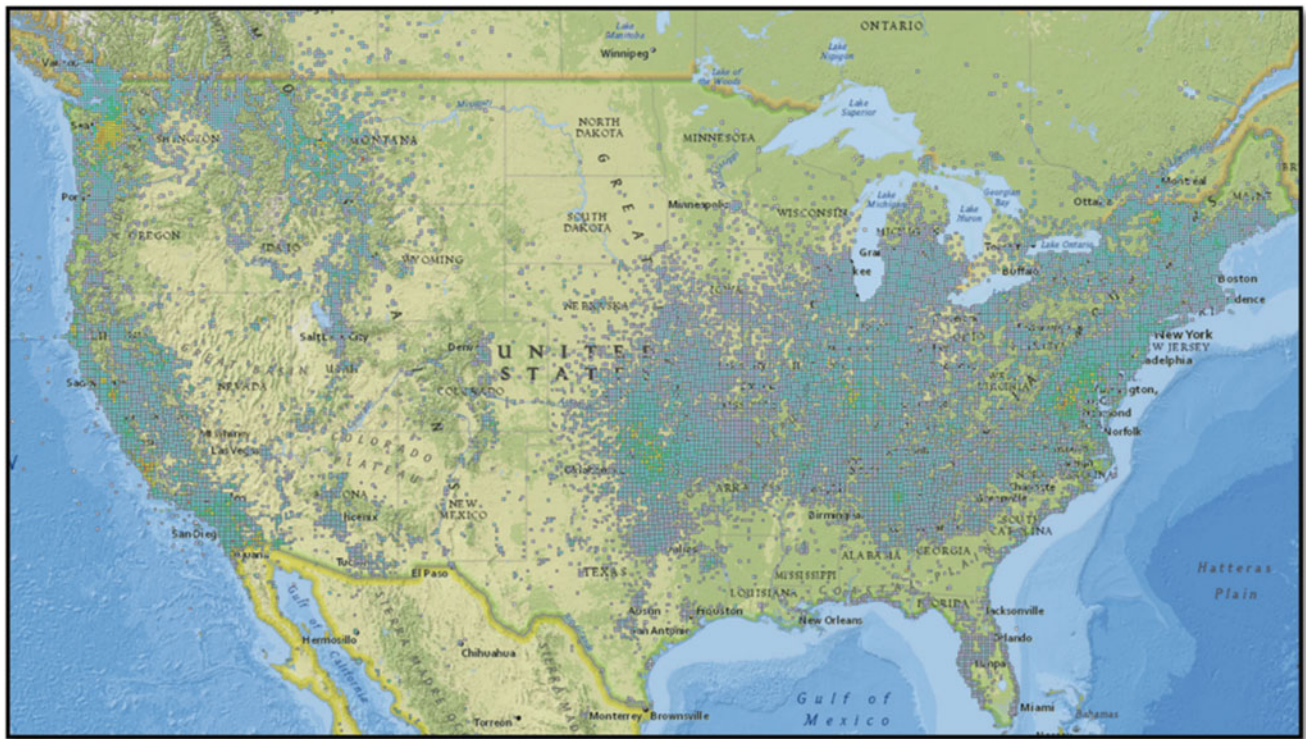


Americans than any other in history. Questionnaire response rates have reached 62,000 per hour (approx. 1000 per min) for another event, requiring substantial web resource allocation and capacity.

Citizen Science Seismology

The DYFI portal provides a two-way information conduit, since citizens coming to the USGS for information are empowered to become data providers themselves by contributing valuable observations that benefit the USGS as well as the observers, their local communities, and earthquake responders. DYFI also provides an important human perspective on earthquakes, providing sociological documentation of

the way people behave and respond, and how they perceive risk (e.g., Celsi et al. 2005). DYFI also seems to provide emotional help to citizens who have just had a frightening or even traumatic experience, by allowing citizens to share their experiences and enabling them to contribute their observations towards a general public understanding of the phenomenon they have experienced, DYFI provides many with a form of catharsis at an opportune time, and in a rapid fashion. Often users describe the desire to confirm their experience with others, or in the case of DYFI, with the collective community. The DYFI system also educates the public on oft-misunderstood seismological concepts like the spatial variations of shaking intensity, and it provides a basis for understanding the difference between macroseismic intensity and instrumental magnitude.



Earthquakes, Did You Feel It?, Fig. 4 Cumulative maximum DYFI intensities, 10-km geocoded, in the USA from 1999 to 2018. The same intensity color-coding is used as in Fig. 1

With two decades of DYFI reports, we can now map out the maximum intensity of shaking reported over the entire U.S. during that time period (Fig. 4). With DYFI, nearly every felt earthquake in the USA is or can be reported now. Thus, this map represents an extraordinary contribution of the DYFI system and something we have never been able to show before: the actual distribution of shaking intensity over the entire nation for nearly two decades. One can further examine the overall extent and intensity of DYFI contributions nationwide on the interactive DYFI cumulative and annual summary maps provided on a separate USGS ArcGIS platform link found on the USGS DYFI homepage. These data have been particularly useful in determining the level and extent of shaking for induced earthquakes in the Central USA (Fig. 4). The DYFI induced database has over 200,000 observations in the past decade with 22,000 at distances less than 20 km; we provide individual responses and aggregated event-associated intensities in 1-km and 10-km boxes. The data are available online in USGS ScienceBase data archives (<https://doi.org/10.5066/F7WM1BPC>).

Scientifically, DYFI data have been widely utilized in seismological and earthquake engineering studies (for a partial summary, see Wald et al. 2011). For example, Worden et al. (2012) employed DYFI data to derive new relations among peak ground motion parameters and intensity data.

The development of these relations sets a new standard for ground-motion-to-intensity relations in that the DYFI intensity data used are decimal intensities and reverse relations are provided. Similarly, DYFI data have recently been used to derive direct intensity prediction equations or IPEs (e.g., Atkinson and Wald 2007); these equations are important for robust ShakeMap generation (Worden et al. 2018) and hazard evaluations from historical macroseismic observations, as well as for improved DYFI filtering. DYFI data have been widely used in attempts to distinguish shaking levels expected from induced earthquakes from naturally occurring ones (e.g., Atkinson et al. 2018).

Outlook

DYFI data have been made accessible via the NEIC's ComCat earthquake archives and database. Users can now replicate, filter, and update MI datasets via USGS web services, though of course individual responses are not accessible given that they contain personally identifiable information (PII). Ongoing DYFI developments include voice activated (smart-speaker) response entries and an open API for third-party entries. Further, since higher macroseismic intensity values (typically, VIII and greater) primarily describe

observed structural damage to buildings (e.g., Musson et al. 2010), and the usual DYFI contributor generally lacks the necessary engineering expertise, one could argue that assigning intensities to intensity VIII and higher should be done formally by professionals. We are investigating ways to augment DYFI at these higher intensities by using crowd-source engineering experience for events that have damage, though nearly 95% of our data collection is at intensity 6 and lower. Likewise, efforts are ongoing to employ more uniform data collection strategies, with the goal of harmonizing data collection around the globe (e.g., Goded et al. 2018).

Cross-References

- [Earthquakes, Intensity](#)
- [Earthquakes, ShakeMap](#)

Bibliography

- Atkinson GM, Wald DJ (2007) “Did you feel it?” intensity data: a surprisingly good measure of earthquake ground motion. *Seis Res Lett* 78:362–368
- Atkinson GM, Wald DJ, Worden CB, Quitoriano V (2018) The intensity signature of induced seismicity. *Bull Seism Soc Am* 108:1080–1086
- Celsi R, Wolfenbarger M, Wald DJ (2005) The effects of magnitude anchoring, earthquake attenuation estimation, measure complexity, hubris, and experience inflation on individuals’ perceptions of felt earthquake experience and perceptions of earthquake risk. *Earthq Spectra* 21:987–1008
- Goded T, Horspool N, Canessa S, Lewis A, Geraghty K, Jeffrey A, Gerstenberger M (2018) New macroseismic intensity assessment method for New Zealand web questionnaires. *Seismol Res Lett* 89(2A):640–652. <https://doi.org/10.1785/0220170163>
- Musson RMW, Grünthal G, Stucchi M (2010) The comparison of macroseismic intensity scales. *J Seismol* 14:413–428. <https://doi.org/10.1007/s10950-009-9172-0>
- Wald DJ, Quitoriano V, Dengler L, Dewey JW (1999a) Utilization of the internet for rapid community intensity maps. *Seismol Res Lett* 70(6):680–697
- Wald DJ, Quitoriano V, Heaton TH, Kanamori H, Scrivner CW, Worden CB (1999b) TriNet “ShakeMaps”: rapid generation of peak ground motion and intensity maps for earthquakes in southern California. *Earthq Spectra* 15(3):537–556
- Wald DJ, Jaiswal KS, Marano KD, Bausch DB, Hearne MG (2010) PAGER – rapid assessment of an earthquake’s impact (U.S. Geological Survey Fact Sheet No. 2010–3036)
- Wald DJ, Quitoriano V, Worden CB, Hopper M, Dewey JW (2011) USGS “did you feel it?” internet-based macroseismic intensity maps. *Ann Geophys* 39:688–709
- Worden CB, Gerstenberger MC, Rhoades DA, Wald DJ (2012) Probabilistic relationships between peak ground motion and modified Mercalli intensity. *Bull Seism Soc Am* 102(1):204–221
- Worden CB, Thompson EM, Baker JW, Bradley BA, Luco N, Wald DJ (2018) Spatial and spectral interpolation of ground motion intensity measure observations. *Bull Seism Soc Am* 108(2):866–875. <https://doi.org/10.1785/0120170201IP-092580>

Earthquakes, Early and Strong Motion Warning

Richard M. Allen

Berkeley Seismological Laboratory, Department of Earth and Planetary Science, University of California, Berkeley, CA, USA

Synopsis

Strong motion earthquake warning (also known as earthquake early warning) is the rapid detection of an earthquake underway, the estimation of the ground shaking likely to result, and the issuance of an alert to people in the region where the ground shaking is likely to be hazardous. The last decade has seen rapid development of methodologies for early warning, which include the use of strong motion observations close to the epicenter to provide warnings at greater distances, the use of P-wave observations across a network to locate the event and map the distribution of likely strong ground motion, and the use of a P-wave detection at a single station to estimate the likely shaking intensity at the same station. These approaches have been listed in the order of increasing warning time, but the additional warning time comes at the expense of certainty. The earlier a warning is provided the greater the likelihood of false and missed alarms and the greater the uncertainty in the estimated shaking intensity. Typical warning times range from a few seconds to tens of seconds; the upper limit is about 1 min. Identified applications include personal protective measures where individuals move to a safe-zone within a few feet; automated mechanical response including stopping trains and isolating sensitive and hazardous machinery and chemicals; and situation awareness by large organizations that can help prevent cascading failures and be available before shaking disrupts communications. Large-scale public warning systems are currently operational in Mexico and Japan, and smaller systems are used in Romania, Turkey, and Taiwan. Testing is underway by seismic networks in many other countries; however, implementation will require significant financial, political, and sociological hurdles to be crossed.

Introduction

Strong motion earthquake warning, more commonly referred to as earthquake early warning (EEW), is the rapid detection of an earthquake underway, estimation of the intensity of ground shaking likely to result, and issuance of a warning in the area likely to experience damaging strong motion. It provides high-probability warnings of shaking likely to occur

within seconds to tens of seconds. The ability to provide warnings is relatively new and due to recent scientific development of methodologies to rapidly estimate the size of earthquakes underway, coupled with modern engineering of rapid communications to collect seismic data and distribute warnings, and a political will to operate state-of-the-art geophysical networks for the purpose of both scientific discovery and societal benefit.

EEW is one of three categories of earthquake warnings. Long-term warnings provide relatively low probabilities of earthquakes over long time periods, typically decades, and are usually referred to as earthquake forecasts. On these long timescales, earthquake likelihood can be estimated based on historic earthquake recurrence intervals and plate motion rates, linked together with deformation models incorporating fault geometries. For example, the probability of an $M > 6.7$ earthquake on the Hayward-Rodgers Creek fault that crosses the UC Berkeley campus is estimated to be 31% over 30 years (Working Group on California Earthquake Probabilities 2007). These forecasts are also translated into estimates of ground shaking intensity that can be used to define building codes designed to prevent new buildings from collapse in the forecast of earthquakes.

The other end-member is earthquake prediction. The term prediction usually means a high probability of a large, damaging earthquake over a short time interval. For example, predicting an $M > 7$ earthquake on the Hayward Fault next week (Allen 1976). Earthquake prediction is currently not possible, as we have not been able to identify a precursory signal that consistently occurs before earthquakes, and most seismologists agree that prediction will not be possible in the near future. EEW is intermediary in that it is possible to provide warning before shaking is felt, however, the warnings are only a few seconds to tens of seconds.

Uses of Early Warning

Uses of EEW that have been identified to date fall into three broad categories (Allen et al. 2009d): personal protection, automated mechanical actions, and situation awareness. The first use that people think about when introduced to early warning is personal protection. For most people, given that there will only be a few to a few tens of seconds of warning, the appropriate action is to move to a “safe-zone” within a few meters of your current location – under a sturdy table, away from falling hazards including windows, bookshelves, etc. These safe-zones need to be identified ahead of time in the locations that people spend most of their time, that is, in their homes and offices or workplaces. In most cases there is not sufficient time to evacuate buildings, and EEW is no substitute for building standards that prevent building collapse. It currently takes schoolchildren in Japan and Mexico

~5 s to get under their desks in response to audible warnings. Outdoors, people should move away from buildings, masonry walls, and other falling hazards including broken glass, street signs, etc. In hazardous work sites including construction sites, manufacturing and chemical facilities, the same principle applies: move to predefined safe-zones.

The second category is automated mechanical control. Within the transportation sector, implemented warning systems are currently used to slow and stop trains, stop airplanes from landing by issuing a go-around command, and to bring traffic to a halt. They could also be used to turn metering or tollbooth lights to red at the entrance to vulnerable road bridges, tunnels, and overpasses. Within buildings, warnings are now used to stop elevators at the next floor and open the doors, open emergency exit doors and window blinds, and turn on lights. Industrial facilities use the warning to place machinery or product into a safe mode. For example, a chip manufacturer in Japan places sensitive micro-robots into a hold mode to reduce damage, and isolates hazardous chemical systems to minimize any resulting leaks. In Japan, engineers are also experimenting with using warnings within automated structural control systems that change the mechanical properties of the building to better withstand the expected shaking. This is an area where there are a myriad of possible applications, most of which are very specific to individual users.

Finally, situation awareness is very important to organizations with responsibilities for large infrastructures and/or people. These include government agencies, utility, and transportation companies. While these groups may take early warning actions, they also have a need to rapidly understand why part or all of their system is failing. For example, an electrical power grid may start to fail in one area, and operators need to rapidly understand why it is failing in order to reduce cascading failures throughout the system. EEW can provide this information more rapidly than any of the existing post-earthquake information systems. EEW also has the advantage that much of the data needed to map out the likely shaking is collected before significant shaking has occurred. This means that the information can be transmitted across data-networks and out to users before strong shaking has the opportunity to disrupt communications.

One of the main concerns expressed in response to the early warning concept is panic: will people panic when they receive the warning and injure themselves or others as a result. Social scientists have long studied this issue and find that it is not the source of concern that many seismologists think it is (Quarantelli 1956; Goltz 2002). The experience that we now have from public warnings in Mexico and Japan provides no evidence that warnings result in panic (Espinosa Aranda et al. 2009; Kamigaichi et al. 2009). There is no evidence for traffic accidents resulting from warnings and no evidence of public stampedes. Implementation of EEW does require an educational campaign to educate

people about what EEW can do and help them identify the appropriate actions in their locations (Kamigaichi et al. 2009).

Approaches to Early Warning

All EEW methodologies must use just a few seconds of data to detect an earthquake underway and rapidly assess its intensity. There is generally a trade-off between the amount of warning time and the accuracy or certainty of the warnings. Greater certainty requires more data from one or more stations, but waiting for that data reduces the warning time. The challenge when building an early warning system is to determine whether additional accuracy is worth the increasing delay.

Strong-motion-based warning The first warning methodologies used this approach. When strong ground shaking is observed at one location, close to the earthquake epicenter, the shaking intensity at greater distances can be estimated using ground motion prediction equations and a warning transmitted ahead of the ground shaking. This approach is called front detection and is used in Mexico City where the earthquake source zone is ~300 km from the city. Figure 1 shows the time at which peak ground shaking was observed at all available stations recording nine $M \geq 6$ earthquakes. A strong-motion-based warning would be possible at some point between 10 s and 20 s after the origin time for these events, depending on the proximity of the closest station(s). By 20 s, the onset of strong shaking is ~60 km from the epicenter meaning that warnings based on peak-shaking observations near the epicenter are only available at distances greater than ~60 km. The region within ~60 km of the epicenter would get no warning and is referred to as the “blind zone.”

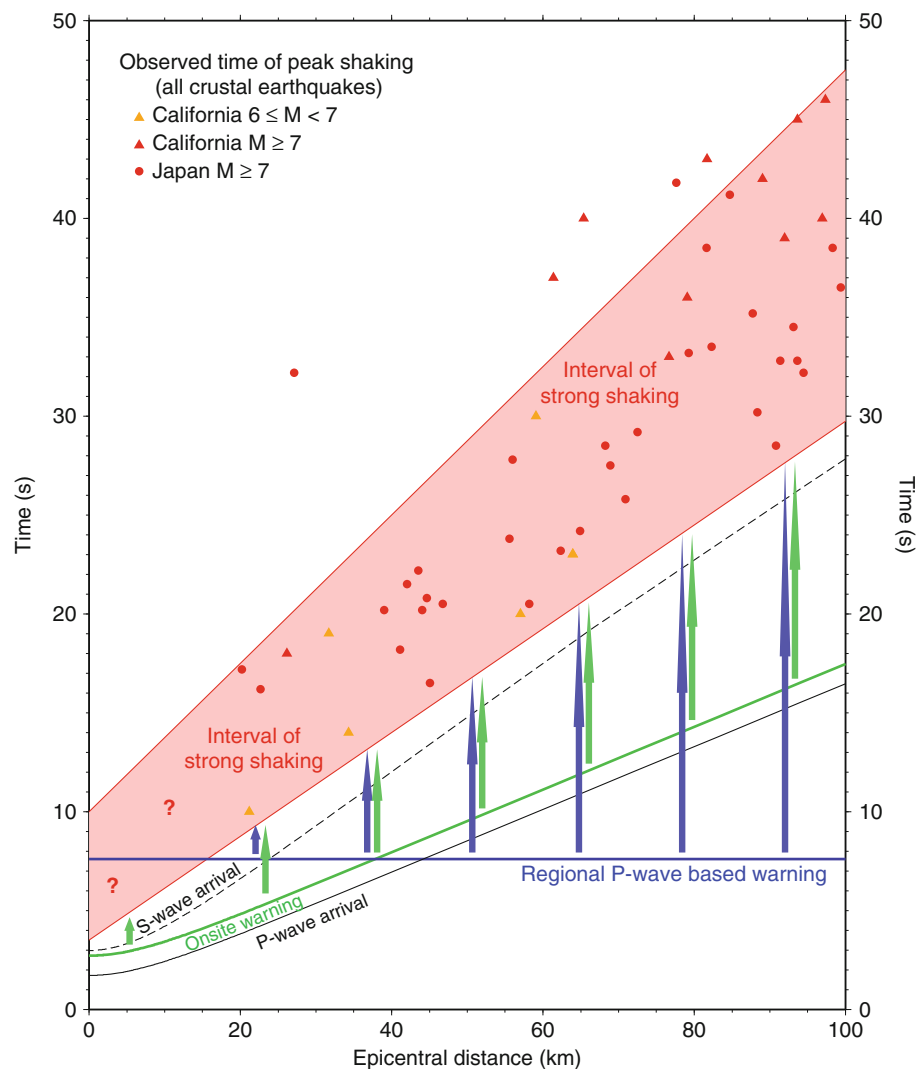
P-wave based warning. To reduce the size of the blind zone, P-wave based methodologies are used. These approaches use a few seconds of the P-wave to estimate either the magnitude or the peak-shaking amplitude. In order to estimate the magnitude, a variety of parameters have been developed that are measures of the frequency content (Nakamura 1988; Allen and Kanamori 2003; Kanamori 2005) or the amplitude (Kamigaichi 2004; Wu and Kanamori 2005b; Zollo et al. 2006; Cua and Heaton 2007; Wurman et al. 2007; Böse et al. 2008; Wu and Kanamori 2008b; Köhler et al. 2009). In order to estimate magnitude, amplitude-based parameters need to be corrected for epicentral distance. However, the frequency-based parameters have been found to be insensitive to epicentral distance provided the observations are made within ~100 km of the epicenter. For all of these parameters empirical scaling relations between the observation and earthquake magnitude have been developed so that magnitude (plus or minus some uncertainty) can be estimated from an observation at one or more stations.

Regional warning. When a regional network is available, P-wave based estimates of magnitude can be combined with an estimate of the earthquake location to then map the distribution of strong ground motion using ground motion prediction equations. In their simplest form these maps are a “bullseye” centered on the epicenter. Corrections can also be made for site amplification effects. Examples of this approach include ElarmS (Allen 2007; Wurman et al. 2007; Allen et al. 2009a; Brown et al. 2009), Presto (Zollo et al. 2009), the Virtual Seismologist (Cua and Heaton 2007; Cua et al. 2009), and the NIED approach in Japan (Horiuchi et al. 2005).

When using a regional network, P-wave observations at multiple stations can be combined in order to verify that an earthquake is underway and improve the accuracy of the magnitude estimate. Generally, the more stations used, the better the accuracy of the warnings and the fewer false alarms. The time at which a warning is available therefore depends on the number of station detections required and the number of seconds of P-wave data required at each station. Figure 1 illustrates the amount of warning time available using this approach. The horizontal blue line at 7.6 s represents the time at which a warning is available for a particular scenario (Fig. 1). In this case, the size of the blind zone is ~15 km and the amount of warning time increases linearly with distance beyond 15 km. There would be ~9 s warning at 50 km and ~22 s warning at 100 km.

Onsite warning. Onsite warning provides the most rapid approach to earthquake warning by removing the need for communications. The delay in issuing a warning is minimized by using a seismic station at the location where the warning is needed. Onsite methodologies are intended to estimate the intensity of ground shaking at the same location that the P-wave is detected (Odaka et al. 2003; Wu and Kanamori 2005a, b, 2008a, b; Nakamura and Saita 2007a, b; Wu et al. 2007; Böse et al. 2009a, b). The challenge with onsite warning is to provide enough accuracy that the warnings are useful. The character of P-wave arrivals is variable at different locations for the same earthquake due to variability in source radiation and also local site effects. Using a single station (or several stations within a few hundred meters) does not allow for averaging of this variability as in the case of regional methodologies. However, the onsite approach is faster.

Figure 1 illustrates the best-case scenario for onsite warning in which the warning is available using 1 s of P-wave data and the earthquake is at a sufficient depth that there is more than 1 s between the P-wave arrival and the onset of peak shaking at the epicenter. This means that there is no blind zone and a warning is available at the epicenter. The warning time increases with distance as the time-separation of the P-wave and peak shaking increases. Typical warning times are ~2 s at 10 km, ~7 s at 50 km, and ~12 s at 100 km. The onsite approach provides the most warning close to the epicenter.



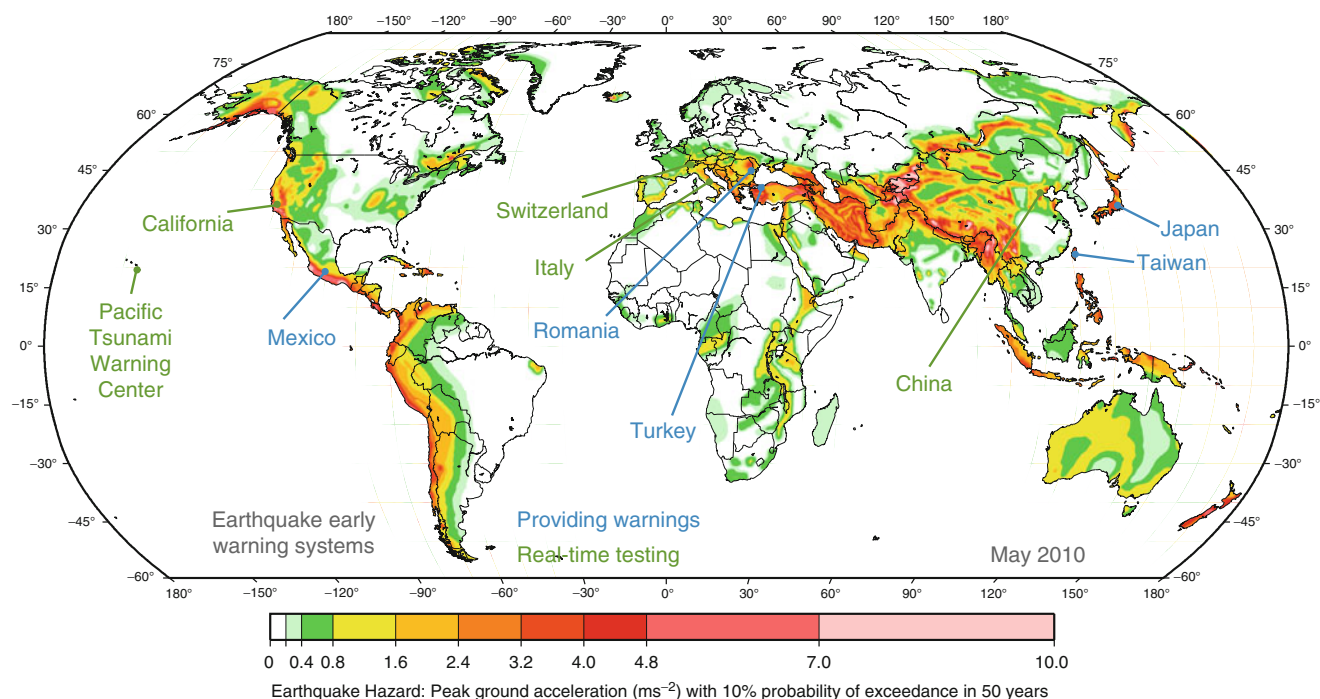
Earthquakes, Early and Strong Motion Warning, Fig. 1 Plot showing the time of observed peak shaking for large magnitude events and illustrating the time at which warnings could be available using various methods. The time of peak shaking was determined for all available stations in nine $M \geq 6$ earthquakes from California (triangles, magnitudes 6.0, 6.5, 6.7, 7.1, 7.1, and 7.5) and Japan (circles, magnitudes 7.0, 7.1, and 7.3) and is plotted as a function of epicentral distance. All events are crustal earthquakes with epicentral depths of 21 km or less. The red shaded region illustrates the approximate interval of strong shaking when most of the peak-shaking observations are made. Note that there are few observations close to the epicenter due to the stations spacing of networks in California and Japan. The P-wave (solid black line) and S-wave (dashed black line) arrival times for a 10 km depth earthquake are also shown. The blue line and arrows illustrate available warning times for a regional P-wave based warning system. The horizontal line is the time at which a warning is available. It assumes that the required number of stations for a warning have detected a P-wave once the

P-wave front reaches 25 km (at 4.6 s), that 1 s of P-wave is required to estimate magnitude (available at 5.6 s), and that an additional 2 s is needed for communications. The purpose-built seismic network in Japan provides data to the network centers within 1 s, and we assume that an additional 1 s is required to get the warning out to users. This means that the warning reaches users at 7.6 s after the origin time. The onset of strong shaking has reached ~15 km from the epicenter at this time. The available warning time increases linearly (shown by blue arrows) from 0 s at 15 km to ~22 s at 100 km. The green line and arrows illustrate the available warning time using the onsite P-wave based warning approach. It is assumed that 1 s of P-wave data is needed and the time of warning (green line) is therefore 1 s after the P-wave arrival. As shown, it is possible to provide a warning at the epicenter if the earthquake is more than ~10 km deep meaning there could be no blind zone. The warning time then increases with distance as the time between the P-wave arrival and peak shaking increases. Warning times (green arrows) range from ~2 s at 10 km to ~12 s at 100 km

However, for locations at greater distances, regional warning provides more warning time than onsite warning as alerts are transmitted electronically ahead of ground motion. In the scenarios shown in Fig. 1, the crossover when the regional method provides more warning than onsite is at ~35 km.

Implementations of Early Warning

There are currently two large-scale implementations of early warning in Mexico and Japan (Fig. 2). The first public system started to provide warnings to Mexico City in 1993. It was



Earthquakes, Early and Strong Motion Warning, Fig. 2 Map of seismic hazard around the world (pink represents the greatest hazard) showing regions where earthquake early warnings are currently being

provided (blue labels) and the regions where real-time testing on existing seismic networks is underway (green labels). (After Giardini 1999)

built following the 1985 M8.1 earthquake that killed 10,000 people in the city. The system uses a line of seismometers deployed near the coastline to detect offshore earthquakes and transmit the warning 300 km to the city ahead of strong shaking (Espinosa Aranda et al. 1995). The offset between the source region and the city allows the use of a front detection approach and still provides more than 60 s of warning time. These warnings are broadcast by 58 radio stations and 6 TV channels in the city. There are also more than 250 users who receive the warning through dedicated radio links including schools and universities, emergency and civil protection agencies, government buildings, and the subway system.

Japan Railways first began to experiment with early warning in the 1960s. These systems became increasingly sophisticated including the first P-wave based application of UrEDAS in the 1980s (Nakamura 1988). With the growth of computerized communication systems in the 1990s automated remote sensing of earthquakes expanded. Following the 1995 Kobe earthquake that killed 6,000 people, Japan installed dense seismic networks across the entire country (Okada et al. 2004). This network led to the development and implementation of a nationwide regional warning system that became operational in 2007 and is operated by the Japan Meteorological Agency (JMA). These warnings are now broadcast by the majority of TV and radio stations in Japan, through the government operated J-Alert system that issues warnings to municipalities (many of which relay the

warning to loudspeaker systems in public spaces), by cell phone companies, and by a myriad of private providers and purpose-built consumer devices (Kamigaichi et al. 2009). The various Japanese systems use both onsite and regional warning systems. In the 2004 M_w 7.2 Niigata earthquake the emergency brakes of one of the bullet trains were triggered by the onsite UrEDAS system (Nakamura and Saita 2007a), and the regional warning system operated by JMA has issued warnings for multiple earthquakes including the 2008 M_w 7.2 Iwate-Miyagi Nairiku earthquake (Kamigaichi et al. 2009).

Smaller scale and user-specific warning systems have also been implemented in various locations around the world. In Bucharest, warnings are provided to a nuclear research facility (Ionescu et al. 2007). Istanbul has warning systems for a power plant and a large banking building, and a more general use warning system has also been implemented in the last few years using a new strong motion network (Alcik et al. 2009). In Taipei, a test warning system provided warnings to the rail system, the university, and a hospital (Wu and Teng 2002). Finally, in the US, a temporary warning system was implemented following the 1989 Loma Prieta earthquake (Bakun et al. 1994).

Testing of early warning methodologies is now underway on an increasing number of regional seismic networks around the world. This includes networks in California, China, Italy, at the Pacific Tsunami Warning Center, and in Switzerland (Fig. 2). In many of these cases the technical feasibility of

providing warnings has been demonstrated. Whether these test systems will transition to implementation of public warning systems will therefore be largely determined by financial, political, and sociological factors.

Summary

The last decade has seen rapid development of earthquake early warning methodologies (e.g., Gasparini et al. 2007; Allen et al. 2009b, c) by multiple research groups around the world. A range of methodologies has been developed that generate warnings based on observations of strong shaking near the epicenter, estimation of shaking hazard maps using P-wave detections across a network, and single-station detections of P-waves to estimate hazard at the same location. Warning times typically range from a few seconds to tens of seconds with an upper limit of around 1 min. With all these methodologies there is a trade-off between the accuracy of the warning and the amount of warning time. The challenge is to identify the optimal threshold at which to provide a warning.

There are now several implementations of public warning systems around the world, most notably in Mexico and Japan. These provide insight to the potential uses of earthquake alerts, which fall into three categories. Personal protection is about actions of individuals to protect themselves. For individuals it is important to identify “safe-zones” at home and work. These are spaces that are reachable within a few seconds where they will be protected from falling hazards. Automated mechanical control can be used to further reduce the impacts of earthquakes. This category includes slowing trains, opening elevators at the nearest floor, and isolating sensitive or hazardous machinery or chemicals. Finally, early warning can provide situational awareness information to large organizations that may be used to limit cascading failures and initiate response. This information could be available before much of the shaking, and therefore increases the likelihood that the information can be transmitted before communications are lost.

While 5 years ago there were many technical questions about the feasibility of earthquake early warning, today the primary hurdles to implementation are financial, political, and sociological. The existing public warning systems in Mexico and Japan were built following major destructive earthquakes. The challenge for the hazard mitigation communities in other countries is to generate the necessary will to implement these systems before the next earthquake.

Cross-References

- Earthquake Precursors and Prediction
- Earthquakes, Intensity

- Earthquakes, ShakeMap
- Earthquakes, Strong-Ground Motion
- Seismic Hazard

Bibliography

- Alcik H, Ozel O, Apaydin N, Erdik M (2009) A study on warning algorithms for Istanbul earthquake early warning system. *Geophys Res Lett* 36:L00B05
- Allen CR (1976) Responsibilities in earthquake prediction. *Bull Seismol Soc Am* 66:2069–2074
- Allen RM (2007) The ElarmS earthquake warning methodology and application across California. In: Gasparini P, Zschau J (eds) *Seismic early warning*. Springer, New York, pp 21–44
- Allen RM, Kanamori H (2003) The potential for earthquake early warning in southern California. *Science* 300:786–789
- Allen RM, Brown H, Hellweg M, Khainovski O, Lombard P, Neuhauser D (2009a) Real-time earthquake detection and hazard assessment by ElarmS across California. *Geophys Res Lett* 36:L00B08. <https://doi.org/10.1029/2008GL036766>
- Allen RM, Gasparini P, Kamigaichi O (eds) (2009b) Earthquake early warning. *Seismol Res Lett* 80:682–782
- Allen RM, Gasparini P, Kamigaichi O (eds) (2009c) New methods and applications of earthquake early warning. *Geophys Res Lett* 36:L00B05, L00B08, L00B03, L00B04, L00B02, L00B06, L00B01, L00B07
- Allen RM, Gasparini P, Kamigaichi O, Böse M (2009d) The status of earthquake early warning around the world: an introductory overview. *Seismol Res Lett* 80:682–693. <https://doi.org/10.1785/gssrl.80.5.682>
- Bakun WH, Fischer FG, Jensen EG, Vanschaack J (1994) Early warning system for aftershocks. *Bull Seismol Soc Am* 84:359–365
- Böse M, Wenzel F, Erdik M (2008) PreSEIS: a neural network-based approach to earthquake early warning for finite faults. *Bull Seismol Soc Am* 98:366–382
- Böse M, Hauksson E, Solanki K, Kanamori H, Heaton TH (2009a) Real-time testing of the on-site warning algorithm in southern California and its performance during the July 29 2008 Mw5.4 Chino Hills earthquake. *Geophys Res Lett* 36:L00B03. <https://doi.org/10.1029/2008GL036366>
- Böse M, Hauksson E, Solanki K, Kanamori H, Wu Y-M, Heaton T (2009b) A new trigger criterion for improved realtime performance of on-site early warning in southern California. *Bull Seismol Soc Am* 99:897–905
- Brown H, Allen RM, Grasso VF (2009) Testing ElarmS in Japan. *Seismol Res Lett* 80:727–739. <https://doi.org/10.1785/gssrl.80.5.727>
- Cua G, Heaton T (2007) The Virtual Seismologist (VS) method: a Bayesian approach to earthquake early warning. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake early warning systems*. Springer, Berlin, pp 97–132
- Cua G, Fischer M, Heaton T, Wiemer S, Giardini D (2009) Real-time performance of the virtual seismologist method in southern California. *Seismol Res Lett* 80:740–748. <https://doi.org/10.1785/gssrl.80.5.740>
- Espinosa Aranda JM, Jimenez A, Ibarrola G, Alcantar F, Aguilar A, Inostroza M, Maldonado S (1995) Mexico city seismic alert system. *Seismol Res Lett* 66:42–52
- Espinosa Aranda JM, Cuellar A, Garcia A, Ibarrola G, Islas R, Maldonado S (2009) Evolution of the Mexican Seismic Alert System (SASMEX). *Seismol Res Lett* 80:694–706. <https://doi.org/10.1785/gssrl.80.5.694>
- Gasparini P, Manfredi G, Zschau J (2007) *Earthquake early warning systems*. Springer, Berlin, p 350

- Giardini D (1999) The global seismic hazard assessment program (GSHAP) – 1992/1999. *Ann Geofis* 42:957–974
- Goltz JD (2002) Introducing earthquake early warning in California: a summary of social science and public policy issues. Report of California Office of Emergency Services by the Caltech Seismological Laboratory, Disaster Assistance Division, Pasadena
- Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An automatic processing system for broadcasting earthquake alarms. *Bull Seismol Soc Am* 95:708–718
- Ionescu C, Böes M, Wenzel F, Marmureanu A, Grigore A, Marmureanu G (2007) An early warning system for deep Vrancea (Romania) earthquakes. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake early warning systems*. Springer, Berlin, pp 343–349
- Kamigaichi O (2004) JMA earthquake early warning. *J Jpn Assoc Earthquake Eng* 4:134–137
- Kamigaichi O, Saito M, Doi K, Matsumori T, Tsukada S, Takeda K, Shimoyama T, Nakamura K, Kiyomoto M, Watanabe Y (2009) Earthquake early warning in Japan – warning the general public and future prospects. *Seismol Res Lett* 80:717–726. <https://doi.org/10.1785/gssrl.80.5.717>
- Kanamori H (2005) Real-time seismology and earthquake damage mitigation. *Annu Rev Earth Planet Sci* 33:195–214
- Köhler N, Cua G, Wenzel F, Böse M (2009) Rapid source parameter estimations of southern California earthquakes using PreSEIS. *Seismol Res Lett* 80:748–754. <https://doi.org/10.1785/gssrl.80.5.748>
- Nakamura Y (1988) On the urgent earthquake detection and alarm system (UrEDAS). In: *Proceedings of the 9th world conference on earthquake engineering*, vol 7, pp 673–678
- Nakamura Y, Saita J (2007a) UrEDAS, the earthquake warning system: today and tomorrow. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake early warning systems*. Springer, Berlin/Heidelberg, pp 249–282
- Nakamura Y, Saita J (2007b) FREQL and AcCo for a quick response to earthquakes. In: Manfredi G, Zschau J, Gasparini P (eds) *Earthquake early warning systems*. Springer, Berlin/Heidelberg, pp 307–324
- Odaka T, Ashiya K, Tsukada S, Sato S, Ohtake K, Nozaka D (2003) A new method of quickly estimating epicentral distance and magnitude from a single seismic record. *Bull Seismol Soc Am* 93:526–532
- Okada Y, Kasahara K, Hori S, Obara K, Sekiguchi S, Fujiwara H, Yamamoto A (2004) Recent progress of seismic observation networks in Japan – Hi-net, F-net, K-NET, KiK-net. *Earth Planets Space* 56:xv–xxviii
- Quarantelli H (1956) *The condition of panic*. University of Chicago, Chicago
- Working Group on California Earthquake Probabilities (2007) *The Uniform California earthquake rupture forecast, version 2 (UCERF 2)*, US Geological Survey, 104 p
- Wu Y-M, Kanamori H (2005a) Experiment on an onsite early warning method for the Taiwan early warning system. *Bull Seismol Soc Am* 95:347–353
- Wu Y-M, Kanamori H (2005b) Rapid assessment of damage potential of earthquakes in Taiwan from the beginning of P waves. *Bull Seismol Soc Am* 95:1181–1185
- Wu Y-M, Kanamori H (2008a) Exploring the feasibility of on-site earthquake early warning using close-in records of the 2007 Noto Hanto earthquake. *Earth Planets Space* 60:155–160
- Wu Y-M, Kanamori H (2008b) Development of an earthquake early warning system using real-time strong motion signals. *Sensors* 8:1–9
- Wu Y-M, Teng T-L (2002) A virtual subnetwork approach to earthquake early warning. *Bull Seismol Soc Am* 92:2008–2018
- Wu Y-M, Kanamori H, Allen RM, Hauksson E (2007) Experiment using the tau-c and Pd method for earthquake early warning in southern California. *Geophys J Int* 170:711–717
- Wurman G, Allen RM, Lombard P (2007) Toward earthquake early warning in northern California. *J Geophys Res* 112:B08311. <https://doi.org/10.1029/2006JB004830>
- Zollo A, Lancieri M, Nielsen S (2006) Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records. *Geophys Res Lett* 33:L23312
- Zollo A, Iannaccone G, Lancieri M, Cantore L, Convertito V, Emolo A, Festa G, Gallović F, Vassallo M, Martino C, Satriano C, Gasparini P (2009) Earthquake early warning system in southern Italy: methodologies and performance evaluation. *Geophys Res Lett* 36: L00B07

Earthquakes, Energy

Domenico Di Giacomo and Peter Borman (Deceased)
Department 2: Physics of the Earth, GFZ German Research Center for Geosciences, Potsdam, Germany

Definition and Calculation

During the earthquake process, the strain energy W available for rupture is divided into fracture energy E_G , thermal energy E_H , and radiated seismic energy E_S . The stress conditions around the fault and the rheological and elastic characteristics of the Earth materials being ruptured determine the partitioning of the overall energy budget involved in the earthquake source process. In particular, E_G and E_H are the energies dissipated mechanically and as frictional heat during faulting, respectively, and E_S is the energy fraction that goes into elastic seismic waves energy. The latter being of great importance for seismologists in evaluating an earthquake's shaking potential, in the following the relationship between the strain energy and radiated seismic energy is outlined.

After Knopoff (1958), the change in strain energy before and after an earthquake can be obtained from:

$$\Delta W = S \bar{D} \bar{\sigma}, \quad (1)$$

where S = fault area, D = average displacement over the fault, and $\bar{\sigma} = (\sigma_0 + \sigma_1)/2$ is the mean stress (with σ_0 being the initial stress and σ_1 the final one). This energy represents the total work associated to the earthquake rupture process. The small energy fraction converted to E_S is (e.g., Wyss and Molnar 1972):

$$E_S = \frac{1}{2} (\sigma_0 + \sigma_1 - 2\sigma_f) \cdot S \bar{D}, \quad (2)$$

where σ_f is the average frictional stress during faulting. It is also useful to relate E_S and ΔW via the radiation efficiency $\eta_R = E_S/(E_S + E_G)$ (see, e.g., Wyss 1970; Husseini 1977):

$$E_S = \eta_R \Delta W. \quad (3)$$

The radiation efficiency η_R , spanning between 0 and 1, expresses the amount of work dissipated mechanically, therefore if an earthquake is effective in radiating E_S then η_R approaches 1 (that is to say about all the energy budget is converted into the seismic waves energy), whereas if η_R is very small or = 0 then the energy is dissipated and no E_S is radiated.

Assuming that the final stress σ_f and the frictional stress σ_f are identical (that is to say the stress drop $\Delta\sigma = \sigma_0 - \sigma_f$ is complete, Orowan 1960), it has been shown by Kanamori (1977) that a measure of ΔW can be obtained considering another important physical parameter of the seismic source, that is the seismic moment M_0 :

$$\Delta W = \left(\frac{\Delta\sigma}{2\mu} \right) M_0, \quad (4)$$

with $M_0 = \mu S \bar{D}$ (where μ = rigidity of the medium in the source area). M_0 is a measure of the overall static displacement caused by the fault rupture. It is a well determined seismological parameter (normally within a factor of 2) since it must be calculated from long and/or very long periods (usually >100 s for large earthquakes) of the recorded seismograms (Dziwonski et al. 1981), which are less affected than the short periods by the small-scale Earth heterogeneities along the path from the source to the receivers.

By assuming that the stress drop is complete and nearly constant (between some 20 and 60 bars = $2-6 \cdot 10^7$ dyn/cm² = 2–6 MPa) for very large earthquakes and that $\mu = 3-6 \cdot 10^{11}$ dyn/cm² = 3–6 $\cdot 10^4$ MPa in the source area under average crust-upper mantle conditions, Kanamori (1977) showed that the ratio

$$E_S/M_0 = 5 \times 10^{-5} = \text{constant}. \quad (5)$$

Therefore, under these conditions (adopted also for introducing the moment magnitude M_w , see ► “Earthquake, Magnitude”), one could easily have an indirect estimation of E_S from known M_0 .

However, M_0 and E_S are two distinct physical parameters. Indeed, M_0 is a static parameter that is not sensible to the details of the rupture history and does not bring any direct information about the frequency content radiated by the seismic source, whereas E_S depends on the dynamic characteristics of the seismic source and, being related to velocity, it is more important than M_0 for evaluating the earthquake damage potential over frequencies of engineering interest (between about 0.1 and 10 Hz). Therefore, direct and independent measurements of E_S are fundamentals in seismic source studies, also considering that many authors pointed

out how the assumption of constant stress drop may not be fulfilled (see, e.g., Purcaru and Berckhemer 1978). Indeed, E_S is sensible to the stress conditions during the fault rupture, as shown by the following formulation (e.g., Rudnicki and Freund 1981):

$$E_S = \int_{-\infty}^{\infty} \int_S [-\sigma_{ij} \gamma_j \dot{u}_i] dS dt, \quad (6)$$

where $\dot{u}_i$ = particle velocity, σ_{ij} = stress change associated with the fault displacement u_i , γ_j = unit vector normal to the surface S and pointing outward from the source, and t = time. This definition holds in the far field, that is, the radius of the surface S is much larger than the dimension of the earthquake source. It can be shown that Eq. 6 is equivalent to another one often used in seismology and proposed by Haskell (1964):

$$E_S = \int_{-\infty}^{\infty} \int_S \rho [\alpha \dot{u}_\alpha^2 + \beta \dot{u}_\beta^2] dS dt, \quad (7)$$

where ρ = density, α and β = P - and S -wave velocities in the medium, $\dot{u}_\alpha$ and $\dot{u}_\beta$ = far-field velocity records of P - and S -waves, respectively. Thus, it would be straightforward to compute E_S by analyzing the recorded waveforms. However, with respect to the determination of M_0 , for computing E_S it must be considered not only the low-frequency part but especially the high-frequency content of the seismograms, which is more sensible to the fine heterogeneities of the Earth structure. Since the seismic recordings must be corrected for the energy loss due to the geometrical spreading and anelastic attenuation of body waves during their travel from the seismic source to the receivers (see ► “Propagation of Elastic Waves: Fundamentals”), it follows that the computation of E_S is more difficult and less stable than M_0 . However, thanks to the worldwide deployment of high-quality digital broadband seismic station networks in the past few decades, much progress has been done and the calculation of E_S has started on a routine basis. Here a brief description of some techniques to calculate E_S is reported.

Since the correction for the high frequencies (above 1–2 Hz) is difficult to perform, many approaches to compute E_S use teleseismic recordings (usually between 30° and 90° epicentral distance Δ , although some procedure for rapid response purposes uses also recordings down to 20° or 25°, see Newman and Okal 1998; Di Giacomo et al. 2010a) for global earthquakes (magnitude above ~5) so that the effects of the Earth filter can be reasonably modeled by assuming an average 1-D layered structure. Moreover, since the P -waves are less affected by attenuation than S -waves, for which the correction for propagation effects at frequencies $f > 0.5$ Hz

may be difficult on a routine basis, the analysis is often performed (1) by considering the P -wave train of the seismograms and (2) in the frequency domain by exploiting the Parseval's identity:

$$\int_{-\infty}^{\infty} |\hat{s}(\omega)|^2 d\omega = 2\pi \int_{-\infty}^{\infty} |s(t)|^2 dt, \quad (8)$$

where $\hat{s}(\omega)$ is the Fourier transform of the time signal $s(t)$ and $\omega = 2\pi f$ the angular frequency.

Currently E_S determinations at the USGS are obtained from the approach proposed by Boatwright and Choy (1986) and are available at the SOPAR database (<http://neic.usgs.gov/neis/sopar/>). This method calculates the energy flux ε^* from the velocity spectrum of P -waves $|\dot{u}(\omega)|$ as:

$$\varepsilon^* = \frac{\rho\alpha}{\pi} \int_0^{\infty} |\dot{u}(\omega)|^2 e^{(\omega t_x^*(\omega))} d\omega, \quad (9)$$

where the symbol * indicates that the energy flux is corrected for anelastic attenuation by the function $t_x^*(\omega)$ (Choy and Cormier 1986). Then, the computation of the average E_S is obtained from:

$$E_S = 4\pi \langle F \rangle^2 \frac{\sum R_i^2 \varepsilon_i^*}{\sum F_i^2}, \quad (10)$$

where $\langle F \rangle$ is the average radiation pattern coefficient for P -waves over the focal sphere, ε_i^* the energy flux at a given station i , R_i the geometrical spreading, and F_i the specific radiation pattern for the station i (Aki and Richards 1980).

Other authors calculate E_S from integration of the moment rate spectral function (an average theoretical curve is shown in Fig. 8 of the entry ► “Earthquake, Magnitude”). The moment rate spectrum $|\hat{M}(f)|$ can be obtained from the teleseismic P -waves spectrum $|\hat{u}(f)|$ as (e.g., Houston and Kanamori 1986):

$$|\hat{M}(f)| = \frac{4\pi\rho\alpha^3 R_E e^{(\pi f t^*)} |\hat{u}(f)|}{g(\Delta)R(\theta, \phi)C|\hat{I}(f)|}, \quad (11)$$

where, R_E is the Earth radius, $g(\Delta)$ the geometrical spreading factor, $R(\theta, \phi)$ the radiation pattern, C the free surface coefficient, and $|\hat{I}(f)|$ the instrument response. Recalling that E_S is proportional to velocity, and, therefore, to the derivative of the function obtained from Eq. 11, E_S is obtained from integration over frequency f as (see, e.g., Venkataraman and Kanamori 2004):

$$\begin{aligned} E_S &= \left[\frac{8\pi}{15\rho\alpha^5} + \frac{8\pi}{10\rho\beta^5} \right] \int_0^{\infty} f^2 |\hat{M}(f)|^2 df \\ &= \left[\frac{2}{15\pi\rho\alpha^5} + \frac{1}{10\pi\rho\beta^5} \right] \int_0^{\infty} |\hat{\dot{M}}(f)|^2 df, \end{aligned} \quad (12)$$

which holds for point sources, that is, sources smaller than the analyzed wavelengths (see ► “Earthquake Source Theory”) and neglecting directivity effects.

In sum, the calculation of E_S requires, after correction for propagation effects, to integrate the squared velocity spectrum recorded at seismic stations over a broad frequency range. Correction for the propagation effects is one of the most challenging tasks in seismological practice, since – in contrast to seismic moment determination (see ► “Earthquake, Magnitude”) – the high-frequency part of the source spectrum must also be considered. Correcting frequencies >1 Hz requires detailed knowledge of the Earth fine structure and attenuation properties, also below the seismic station, which is often not available. Consequently, E_S determinations from local and/or regional recordings are more problematic and are, anyway, more reliable only in areas with a high density of seismic stations (like California and Japan, see, e.g., the recent works by Izutani and Kanamori 2001; Venkataraman et al. 2002; Boatwright et al. 2002; Oth et al. 2010) where a better knowledge of the local Earth structure is available.

Summary and Outlook

The energy budget involved in the earthquake process is partitioned in different forms. In seismology it is of great importance to quantify two macroscopic parameters of the source, namely the seismic moment M_0 and the radiated seismic energy E_S . While the former (and the related magnitude M_w , see ► “Earthquake, Magnitude”) is a static measure of the seismic source and fundamental for the tsunami potential evaluation of an earthquake, the latter (and the related magnitude M_e) represents the fraction of the strain energy converted in elastic seismic waves energy and fundamental for the evaluation of the shaking potential.

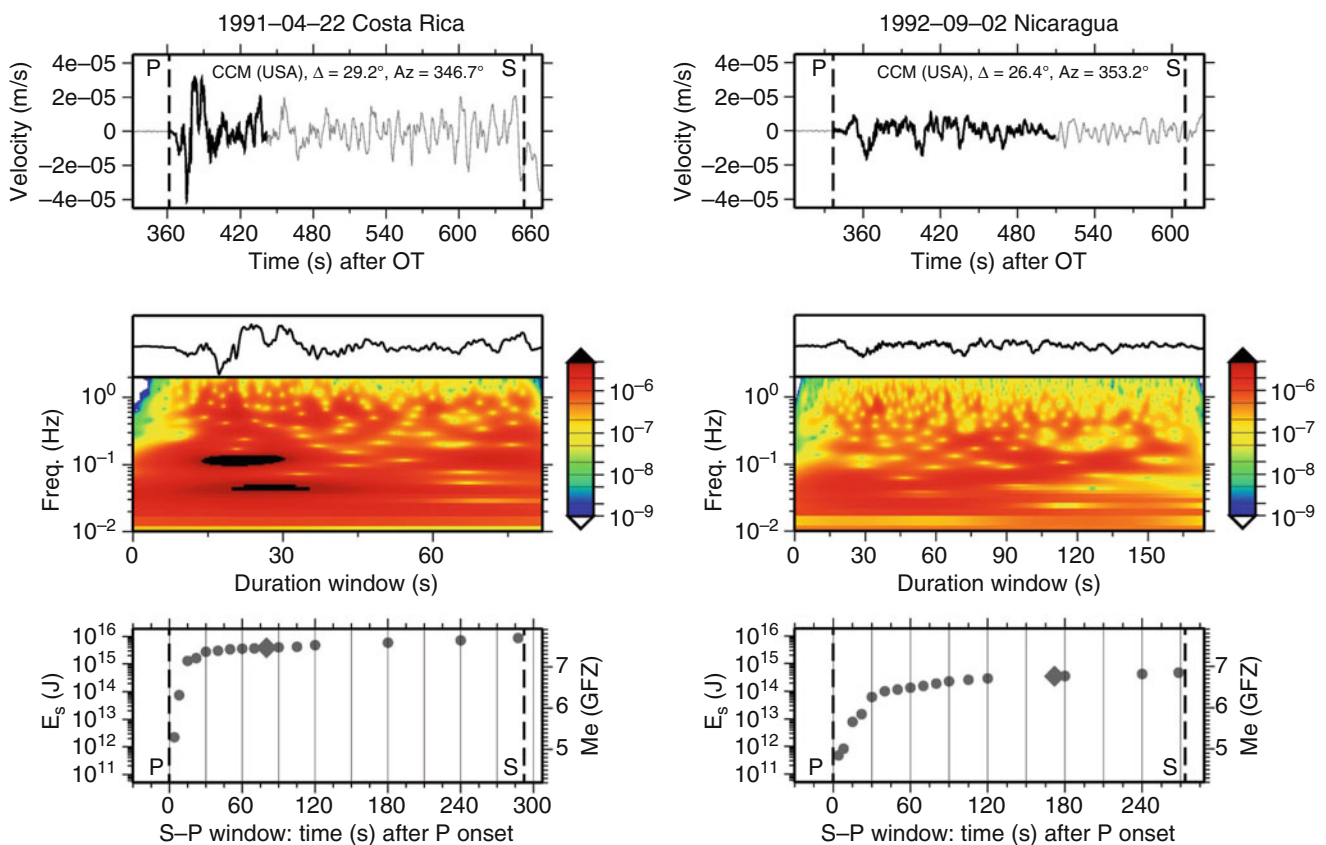
However, only since the early 1990s procedures to calculate E_S can be implemented on a routine basis at regional and global seismological centers. The approach of Boatwright and Choy (1986) allowed systematic and direct measurements of E_S from digital data, and later Choy and Boatwright (1995) demonstrated the importance of such measurements showing the significant variability of E_S for a given value of M_0 . Therefore, the indirect measurement of E_S , based on the assumptions leading to Eq. 5, may be affected by large errors.

Of course, a better picture of the source characteristics can be obtained and help end users in the evaluation of the

tsunami and shaking potential of an earthquake by complementing the information provided by direct measurements of E_S and M_0 and their related magnitudes. To this regard it is remarkable the case of two earthquakes that occurred about 500 km apart from each other in Central America, namely the 1991 Costa Rica earthquake and the 1992 Nicaragua earthquake. Both had the same moment magnitude $M_w = 7.6$ (i.e., about the same M_0). Because of their small difference in epicenter location, one can assume very similar propagation paths to teleseismic stations. However, the broadband waveforms of both earthquakes, recorded at identical stations, differ strongly. As an example, Fig. 1 illustrates the P -wave windows, their time-frequency analysis performed with the S -transform (Stockwell et al. 1996), and the corresponding E_S - Me (see Di Giacomo et al. 2010a, b) values at station CCM for both earthquakes. The 1991 Costa Rica earthquake has much larger amplitudes, especially at higher frequencies, than the 1992 Nicaragua event. Such large differences can be only ascribed to significant

differences in the medium-to-high-frequency parts of the source spectra of these two earthquakes with about the same M_0 . Accordingly, E_S (and the corresponding Me) for the Costa Rica earthquake is much larger than for the Nicaragua earthquake. The latter caused a damaging *Tsunami* with 8 m run-up height, in contrast to severe damages caused by strong ground shaking of *intensity* I = X in the epicentral area of the Costa Rica earthquake (see ► “Earthquakes, Intensity”). This shows that M_0 alone (and a single type of magnitude as M_w) may not describe accurately enough the kind of seismic hazard and damage potential posed by a given earthquake and that a combined use of M_0 and E_S (and a multiple magnitude approach, see ► “Earthquake, Magnitude”) is preferable.

Thus, for improving seismic hazard assessment, future studies must focus on combining in a systematic way E_S and M_0 determinations in order to fully harness the information provided by these two important physical parameters of the seismic source.



Earthquakes, Energy, Fig. 1 Upper panels: S - P time windows (see ► “Body Waves”) of the vertical component velocity seismograms recorded at station CCM for the 1991 Costa Rica (left) and the 1992 Nicaragua earthquake (right). Vertical broken lines mark the theoretical arrival times of the P and S phases. Middle panels: time-frequency analysis via the S -transform (Stockwell et al. 1996) calculated over the P -wave duration window (black record traces in the upper panels that are

also shown enlarged just above each middle panel); the color scales represent the amplitude of the S -transform in m/s. Lower panels: E_S and Me values as a function of time for different cumulative P -wave windows. The diamonds mark the end of the P -wave groups (see black record traces in the upper panels) that have been used for the single station E_S - Me estimates

Cross-References

- [Body Waves](#)
- [Earthquake, Magnitude](#)
- [Earthquakes, Intensity](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Tsunami](#)

Bibliography

- Aki K, Richards P (1980) Quantitative seismology. Theory and methods. Freeman, San Francisco, p 932
- Boatwright J, Choy GL (1986) Teleseismic estimates of the radiated energy by shallow earthquakes. *J Geophys Res* 91(B2):2095–2112
- Boatwright J, Choy GL, Seekins LC (2002) Regional estimates of radiated seismic energy. *Bull Seismol Soc Am* 92(4):1241–1255
- Choy GL, Boatwright J (1995) Global patterns of radiated seismic energy and apparent stress. *J Geophys Res* 100(B9):18205–18228
- Choy GL, Cormier VF (1986) Direct measurements of the mantle attenuation operator from broadband P and S waveforms. *J Geophys Res* 91(B7):7326–7342
- Di Giacomo D, Parolai S, Bormann P, Grosser H, Saul J, Wang R, Zschau J (2010a) Suitability of rapid energy magnitude determinations for rapid response purposes. *Geophys J Int* 180:361–374
- Di Giacomo D, Parolai S, Bormann P, Grosser H, Saul J, Wang R, Zschau J (2010b) Erratum to “Suitability of rapid energy magnitude estimations for emergency response purposes”. *Geophys J Int* 181:1725–1726. <https://doi.org/10.1111/j.1365-246X.2010.04610.x>
- Dziewonski AM, Chou TA, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J Geophys Res* 86(B4):2825–2852
- Haskell NA (1964) Total energy and energy spectral density of elastic wave radiation from propagating faults. *Bull Seismol Soc Am* 54(6):1811–1841
- Houston H, Kanamori H (1986) Source spectra of great earthquakes: teleseismic constraints on rupture process and strong motion. *Bull Seismol Soc Am* 76(1):19–42
- Hussein MI (1977) Energy balance for formation along a fault. *Geophys J R Astron Soc* 49:699–714
- Izutani Y, Kanamori H (2001) Scale-dependence of seismic energy-to-moment ratio for strike-slip earthquakes in Japan. *Geophys Res Lett* 28(20):4007–4010
- Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82(20):2981–2987
- Knopoff L (1958) Energy release in earthquakes. *Geophys J* 1:44–52
- Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: the E/M_0 discriminant for tsunami earthquakes. *J Geophys Res* 103:26885–26897
- Orowan E (1960) Mechanisms of seismic faulting in rock deformation: a symposium. *Geol Soc Am Mem* 79:323–345
- Oth A, Bindi D, Parolai S, Di Giacomo D (2010) Earthquake scaling characteristics and the scale-(in)dependence of seismic energy-to-moment ratio: insights from KiK-net data in Japan. *Geophys Res Lett* 37:L19304. <https://doi.org/10.1029/2020GL044572>
- Purcaru G, Berckhemer H (1978) A magnitude scale for very large earthquakes. *Tectonophysics* 49:189–198
- Rudnicki JW, Freund LB (1981) On energy radiation from seismic sources. *Bull Seismol Soc Am* 71(3):583–595
- Stockwell RG, Mansinha L, Lowe RP (1996) Localization of the complex spectrum: the S transform. *IEEE Trans Signal Process* 44:998–1001
- Venkataraman A, Kanamori H (2004) Effect of directivity on estimates of radiated seismic energy. *J Geophys Res* 109:B04301. <https://doi.org/10.1029/2003JB002548>
- Venkataraman A, Rivera L, Kanamori H (2002) Radiated energy from the 16 October 1999 Hector mine earthquake: regional and teleseismic estimates. *Bull Seismol Soc Am* 92(4):1256–1265
- Wyss M (1970) Stress estimates for South American shallow and deep earthquakes. *J Geophys Res* 74(8):1529–1544
- Wyss M, Molnar P (1972) Efficiency, stress drop, apparent stress, effective stress, and frictional stress of Denver, Colorado, earthquakes. *J Geophys Res* 77(8):1433–1438

Earthquakes, Intensity

Gottfried Grünthal¹ and Roger M. W. Musson^{2,3}

¹Helmholtz Centre Potsdam, German Research Centre for Geosciences GFZ, Potsdam, Germany

²British Geological Survey, Edinburgh, UK

³School of Geosciences, University of Edinburgh, Edinburgh, UK

Synonyms

Macroseismic intensity

Definition

The *intensity*, or *macroseismic intensity*, represents a classification of the severity of ground-motion shaking during an earthquake on the basis of observed effects at a given place (Grünthal et al. 1998). The word “macroseismic” refers to perceptible effects of earthquakes as opposed to instrumental observations.

Macroseismic Scales and Intensity Assignment

Earthquake intensities are defined in macroseismic scales. Since the early twentieth century, they usually contain 12 intensity degrees with the prominent exception of the seven degree Japanese Meteorological Agency (JMA) scale (later upgraded to ten degrees). Earthquakes generate different intensities at different places, generally decreasing with distance from the epicenter. The epicentral intensity can be used in combination with the hypocentral depth as a classification of earthquake strength – with restrictions for large earthquakes ($M \geq 7$) with extensive fault planes.

Although not a physical parameter as such, intensity is a measure of strength of shaking, which can be related to physical ground-shaking parameters. It is a shorthand

description of common observations from an earthquake, for example, instead of writing “the event was felt indoors by most, outdoors by few, and china and glasses clattered together,” one summarizes “intensity 5.”

Intensity data are capable of constraining the magnitude of an event with the same order of uncertainty as individual instrumental magnitude readings (Johnston 1996) and are essential in the study of historical earthquakes. Intensity data of an earthquake are usually presented in the form of a map showing for each intensity point either a corresponding color code or symbol. Contour lines in such maps separating intensities of different degrees are called *isoseismals*. The epicentral intensity is usually the largest observed intensity near to the epicenter or the barycenter of the highest intensity values. It can only be given for onshore events.

Intensity is an integer quantity. Uncertain intensity assignments can be expressed as, for example, 8–9 (meaning either 8 or 9), which may be used in numerical procedures as 8.5, without claim of precision. A macroseismic scale defines each intensity degree with a series of diagnostics. Each diagnostic describes typical effects of ground vibrations related to different sensors. For the lower intensities, the main sensors are people and objects indoors, whereas for higher intensities they are buildings, which suffer progressively more severe damage with increased intensities. The type of building, for example, masonry or reinforced concrete, its condition and workmanship, as well as the level of earthquake-resistant design influence its strength considerably.

Intensity values are often written using Roman numerals to emphasize the integer nature of intensity. However, as such numbers are hard to process in computer files, the practice is dying out.

The most recent scale in general use, the European Macroseismic Scale (EMS-98), fully considers the varying strength of buildings in the form of six vulnerability classes, five damage grades for both masonry and reinforced concrete structures, as well as differentiating structural and non-structural damage (Grünthal et al. 1998). Another element of the EMS-98 associated with the diagnostics is the relative frequency with quantitative definitions of the qualitative terms “few, many, most.” It is the only intensity scale supplemented by comprehensive guidelines and background materials, and this contributes, together with the other mentioned elements, to a reduction of the subjectivity, which often is associated with intensity assignments.

The EMS-98 is easily adaptable for use to the building stock anywhere in the world. Table 1 gives the short form of the EMS-98 as a very simplified and generalized description. The long, full version should be used for intensity assignments.

The EMS-98 should be the basis for intensity evaluation in European countries and has also been applied in a number of countries outside Europe. Some South European countries,

like Italy, still deploy Sieberg’s (1932) Mercalli-Cancani-Sieberg scale (MCS). The basic intensity scale used in North America is the Modified Mercalli scale (MMI). EMS-98 has also been applied in countries elsewhere. In China, the China seismic intensity scale (CSIS) has been used since 1999. This scale is related to the test version of the EMS-98, that is, the EMS-92 (Grünthal et al. 1993). While all these scales have 12 intensity degrees, the 1996 version of the JMA scale, introduced after the 1995 Great Hanshin earthquake, has been upgraded from originally seven degrees to ten degrees.

Nature of Sensors

Since intensity is a measure of earthquake shaking, it is important that scale diagnostics are effects that actually relate to shaking, such as the movement of objects and damage to buildings. Objects, buildings, and people make reliable sensors for deducing intensity; the natural environment does not provide good sensors. Especially, older intensity scales tend to include effects that may have been observed in an earthquake, but were not shaking related or not reliably associated with intensity degrees. Thus in Richter (1958; see below) intensity 11 is defined only by the bending of rail lines and damage to underground pipes. No amount of shaking will bend rail lines – such an effect is due to either fault rupture or ground failure. Ground failure and landslides are related to pre-existing slope vulnerability, which is unknown. Changes in well water level are influenced more by the condition of the water table than the actual strength of shaking (Vogt et al. 1994). Therefore, effects on natural surroundings as scale diagnostics are hard to use to advantage. This has the unfortunate result that it can be difficult if not impossible to assess intensity reliably away from settlements.

History of Intensity Scales

Historically, the use of intensities is important because it does not require instrumental measurements of earthquakes. In the last quarter of the nineteenth century, that is, before the establishment of instrumental seismology, macroseismic intensity became a widespread parameter. The first internationally introduced scale was the ten degree Rossi-Forrel scale of 1883, the use of which in some countries persisted up to the 1960s. This points to the relevance of knowing the different scales that have been used in different countries during different time periods. Here, only a short summary can be given. For a comprehensive compilation, see Musson et al. (2010) and Musson and Cčić (2002).

Earthquakes, Intensity, Table 1 Short form of the EMS-98

EMS intensity	Definition	Description of typical observed effects (abstracted)
I	Not felt	Not felt.
II	Scarcely felt	Felt only by very few individuals at rest in houses.
III	Weak	Felt indoors by a few people. People at rest feel a swaying or light trembling.
IV	Largely observed	Felt indoors by many people, outdoors by very few. A few people are awakened. Windows, doors, and dishes rattle.
V	Strong	Felt indoors by most, outdoors by few. Many sleeping people awake. A few are frightened. Buildings tremble throughout. Hanging objects swing considerably. Small objects are shifted. Doors and windows swing open or shut.
VI	Slightly damaging	Many people are frightened and run outdoors. Some objects fall. Many houses suffer slight nonstructural damage like hairline cracks and fall of small pieces of plaster.
VII	Damaging	Most people are frightened and run outdoors. Furniture is shifted and objects fall from shelves in large numbers. Many well-built ordinary buildings suffer moderate damage: Small cracks in walls, fall of plaster, parts of chimneys fall down; older buildings may show large cracks in walls and failure of fill-in walls.
VIII	Heavily damaging	Many people find it difficult to stand. Many houses have large cracks in walls. A few well-built ordinary buildings show serious failure of walls, while weak older structures may collapse.
IX	Destructive	General panic. Many weak constructions collapse. Even well-built ordinary buildings show very heavy damage: Serious failure of walls and partial structural failure.
X	Very destructive	Many ordinary well-built buildings collapse.
XI	Devastating	Most ordinary well-built buildings collapse; even some with good earthquake-resistant design are destroyed.
XII	Completely devastating	Almost all buildings are destroyed.

A later popular scale is Mercalli's version (with ten degree) from 1902. The first twelve degree scale is the one by Cancani (1904) with the focus on associating peak ground acceleration with single intensities. Based on the macroseismic observations from the 1911 $M_L = 6.1$ central European earthquake, Sieberg (1912) provided a completely revised twelve degree version with descriptions of the intensity degrees given in great detail. This scale, which overwhelmingly influenced all modern intensity scales, later became known as the Mercalli-Cancani-Sieberg (MSC) scale. Sieberg's (1923) minor revision of this scale was translated into English and published by Wood and Neumann (1931) in a largely condensed form, at the cost of clarity. The title is also inappropriate: the Modified Mercalli scale, MM-31. Sieberg's (1932) further revised version, with some major innovations, is still in use in parts of southern Europe as the MCS scale.

The MM-31 scale was completely overhauled in 1956 by Richter (1958). He refrained from associating his name with the new version to avoid confusion with the so-called Richter magnitude scale M_L , and he suggested it was to be known as the "Modified Mercalli scale of 1956" (MM-56).

Today's macroseismic evaluation procedures at the USGS rely basically on a version of the MM-31 scale by Stover and Coffman (1993) but in a newly edited and almost undocumented variant, that is, lacking differentiations in the terminology used, and in which extra rules and conventions are being applied (see Musson et al. 2010; Musson and Cčić 2002). Related intensity data sets are usually denoted only as MM or MMI.

In Europe, Medvedev, Sponheuer, and Kárník developed the MSK scale, introduced as MSK-64 (Medvedev et al.

1965a, b), on the basis of the MCS, the Russian GEOFIAN scale, and the MM-56. It is more systematically arranged than the forerunners and was widely adopted in Europe.

In 1988, the European Seismological Commission launched the updating of the MSK scale to bring it in line with modern earthquake-resistant building types and engineering requirements. After the new scale had been constructed, it was obvious that it contained so many new elements with respect to the MSK scale that it should bear a new name, and so the test version EMS-92 and the EMS-98 were established. The English original of the EMS-98 can be downloaded from <http://www.gfz-potsdam.de/EMS98> and also translations of the full scale into French, Italian, Spanish and Chinese. Translations of the core part and of the short form are also available in a total of altogether 30 languages. The availability in different languages is important since persons, not always fluent in English, act as observers and sensors in macroseismology.

An aspect connected with the history of intensity scales is the decline of interest in macroseismology in the middle of the twentieth century when instrumental monitoring was progressing. However, the late 1970s saw a revival of the subject since intensity assignments are fundamental for parametrizing historical seismicity and developing intensity-based ground-motion models essential in seismic hazard (see ► "Seismic Hazard") and risk assessments.

The design of the EMS-98 allows also its inverse use as risk predictive tool; i.e., the expected frequency of damage grades of buildings of different vulnerability classes can be assigned for a given intensity expected in a municipality for a certain occurrence probability (Musson 2000; Tyagunov et al. 2006).

Conversion of Intensity Scales

The conversion between intensity scales is, due to their integer nature, not possible in the way scalar parameters like different magnitude types can be converted. Ideally, conversions between intensities according to different scales should be avoided, and the original data should be reassessed using one scale. This is frequently not possible. Therefore, conversion diagrams have been developed in which the intensities according to different scales are represented as a series of “intensity ranges.” One scale, taken as a standard with equally large ranges, is, as it is perceived by different authors, compared with other scales with different ranges for the intensities. Musson et al. (2010) elaborate in more detail on this issue and present tables suggesting direct numerical correspondence. Among the twelve degree scales, Musson et al. (2010) conclude equality between EMS-98 and MCS, MMI, and MSK for the intensities 1–10. For intensities 11 and 12, the definitions in MCS, MSK, and MMI relate to phenomena, such as surface faulting, which do not represent strength of shaking or to a saturation of classification criteria in the scale where, for example, total damage refers to total damage to buildings without antiseismic design. Therefore, the intensities 11 and 12 do not have an equivalent in the EMS-98. The relation of the JMA-96 (JMA 1996) scale with the EMS-98 is given in Table 2.

Macroseismic Surveys

After an earthquake is felt, there are several ways to gather macroseismic data. The main source of data collection was in the past from questionnaires designed according to the diagnostics of the intensity-degree definitions. Questionnaires were especially used to compile data for the vast majority of observations in the range of intensities ≤ 6 , although many questionnaires were designed for diagnostics up to intensity 7 or 8.

Earthquakes, Intensity, Table 2 Conversion from the JMA-96 scale to the EMS-98. (After Musson et al. (2010))

JMA-96	EMS-98
0	1
1	2 or 3
2	4
3	4 or 5
4	5
5 Lower	6
5 Upper	7
6 Lower	8
6 Upper	9 or 10
7	11

Information on earthquake damage to buildings and other constructions need to be collected in the field. Special forms are in use in several high seismicity countries to optimize this procedure with respect to certain intensity scales and the building stock in a region.

Printed questionnaires used in the past have today been almost entirely replaced by Internet web-based versions. The “Did You Feel It?” (DYFI) program by the USGS (Wald et al. 1999a) is the basis of similar current activities worldwide. Musson (2006) describes the DYFI procedure for EMS-98 intensities in detail. The idea of such DYFI programs is that persons who have observed an earthquake use an Internet website to report the macroseismic effects through answering a multiple-choice questionnaire. Such DYFI programs are often not restricted to a certain country but can be applied internationally (see ► “Earthquakes, ShakeMap”).

The macroseismic data accumulated by the DYFI program represent a vast new source of engineering seismology data. The DYFI intensity data “make up in quantity what they may lack in quality. . . They are also robust and of surprisingly high utility” (Atkinson and Wald 2007).

Intensity Attenuation with Distance

The rate of decay of shaking strength with distance from the epicenter, that is, the intensity attenuation, is strongly dependent upon the focal depth h and to a lesser extent on the regionally dependent absorption of energy. Assuming a logarithmic proportionality between intensity and ground acceleration, an isotropic intensity-attenuation model results (Kövesligethy 1906), rewritten by Jánosi (1907) as

$$I_0 - I_i = 3 \log \left(\frac{r}{h} \right) + 3\alpha \log e(r - h) \tag{1}$$

with $r = \sqrt{R_i^2 + h^2}$, where R_i is the radius of the isoseismal intensity I_i , I_0 the epicentral intensity, α the absorption coefficient, and e is Euler’s constant. The same structure as Eq. (1) results in a logarithmic proportionality between intensity and seismic energy density (Howell and Schultz 1975; ► “Earthquakes, Energy”).

The first and, for a long time, the most important use of intensity attenuation was focused on depth determinations. For this purpose, graphical schemes by Sponheuer (1960) were frequently applied in the pre-computer era. In the case of good macroseismic data, the calculated depths coincide very well with instrumental depth determinations, at least for shallow (5–25 km) events with small source dimensions (Musson 1993). Intensity-attenuation relations of the type of Eq. (1) were determined from a few mean isoseismal radii. Only within the last two decades have modern statistical procedures been applied, which rely directly on a huge amount of

intensity data points. One of these relations based on Central European data after Stromeyer and Grünthal (2009) combined with Eq. (3) reads

$$I = 1.5M_w - 0.45 \log(h) + 0.15 - 2.95 \log\left(\frac{r}{h}\right) - 0.0025(r - h) \quad (2)$$

where I is the intensity at epicentral distance r . Gomez Capera (2006) compiled several intensity-attenuation relations of this type.

Well-established models of intensity attenuation with distance are especially required for operating ShakeMap programs, i.e., rapid post-earthquake information systems (see ► “Earthquakes, ShakeMap”), for earthquake early warning systems (see ► “Earthquakes, Early and Strong Motion Warning”), and for other quantitative scientific studies like intensity-based seismic hazard and risk assessments.

Conversion of Intensities to Magnitudes

Intensity data give surprisingly robust measures of earthquake magnitudes (see ► “Earthquake, Magnitude”). This important fact qualifies the macroseismic method to be used for magnitude determinations of historical earthquakes occurring in early or pre-instrumental times. Seismic hazard analysis greatly benefits from long-term earthquake catalogues. There are three ways of macroseismic magnitude determinations based on the following: (1) the epicentral intensity, (2) the areas affected by certain intensity levels, and (3) the entire field of intensity data points. These ways are described below.

1. A few years after the introduction of the earthquake magnitude, the first empirical relation with the epicentral intensity I_0 was given (Gutenberg and Richter 1942, 1956). The maximum or epicentral intensity is in many cases the only information at hand for historical earthquakes. I_0 is strongly affected by the focal depth h . Therefore, h should be taken into consideration, or the seismogenic depth in an area should be narrowly constrained at the conversion to magnitude. A relation based on well-constrained moment magnitude M_w , I_0 , and h of European earthquakes given by Grünthal et al. (2009) reads.

$$M_w = 0.667 I_0 + 0.30 \log(h) - 0.10 \quad (3)$$

or omitting h

$$M_w = 0.682 I_0 + 0.16 \quad (4)$$

These relations are valid for the ranges $5 < I_0 < 9$, $3.0 < M_w < 6.4$, $5 \leq h[\text{km}] < 22$.

2. For earthquakes with known total felt area or known area enclosed by one or more of the n (preferably) outer isoseismals (An), the magnitude can be determined as a function of $\log An$. Johnston (1996) derived regression relations for MMI isoseismal areas for stable continental regions to predict seismic moments M_0 according to a functional regression form proposed by Frankel (1994).

$$\log M_0 = a + b \log An + c\sqrt{An} \quad (5)$$

From the published relations for different isoseismal areas, the coefficients for the areas of intensity 3 and 4 read

	a	b	c
A3	17.59	1.020	0.00139
A4	18.10	0.971	0.00194

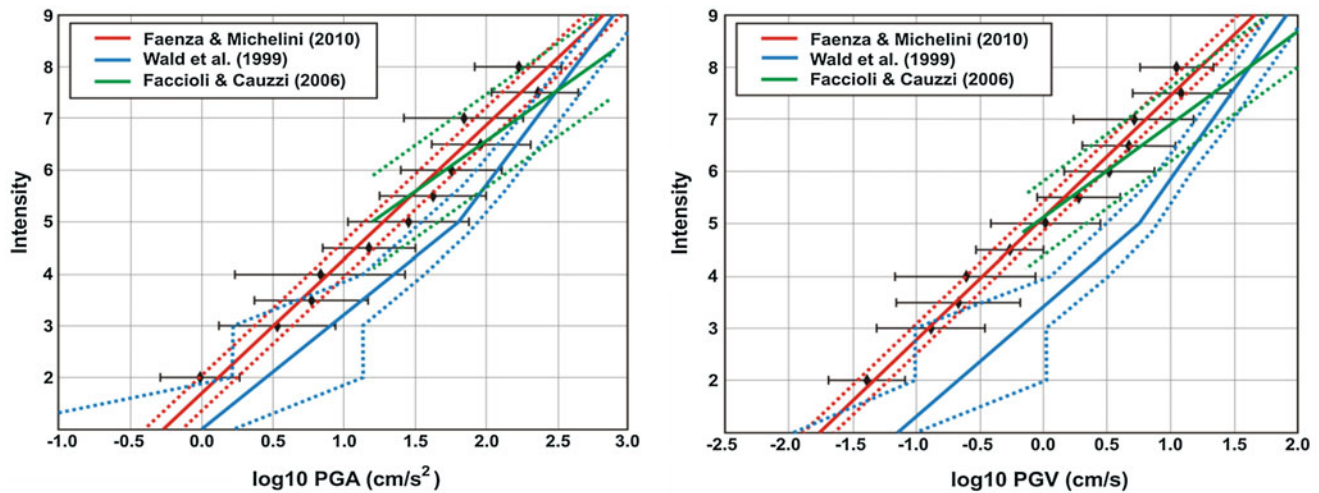
The uncertainties in the derived moment magnitudes can be halved by applying weighted averaging of a suite of isoseismal areas. The uncertainties are then in the range of $\pm 0.15 \dots 0.23 M_w$ units.

3. The most advanced methods to derive magnitude from macroseismic data use the intensity data points of an event to make a joint determination of epicenter location (see ► “Earthquakes, Location Techniques”) and magnitude. These methods need a well-calibrated intensity-attenuation model valid for a certain region. The most important methods of this type are by Bakun and Wentworth (1997), originally developed for Californian earthquakes $4.4 < M_w < 6.9$ and by Gasperini et al. (1999) for Italian earthquakes with $M_S \geq 5.5$. The latter additionally provides the fault azimuth, essentially derived from the direction of maximum elongation of the highest intensity data points.

Relation of Intensities to Recorded Strong Ground Motion

Attempts to express the strength of an earthquake and its observed effects in the form of physical parameters date back to the last quarter of the nineteenth century (cf. summary in Grünthal 1984). Omori (1900) and Cancani (1904) were the first who related intensity degrees to peak ground acceleration (PGA) (see ► “Earthquakes, Strong-Ground Motion”). Kövesligethy (1906) and Jánosi (1907) invented a functional form that is still used today – a linear regression between intensity I and the logarithm of Cancani’s PGA ($I = a + b \log(PGA)$). Grünthal (1984) compiled 30 published relations of intensity to peak ground motion, which showed a larger scatter in PGA compared to peak ground velocity (PGV).

For the implementation of the USGS ShakeMap software package (see ► “Earthquakes, ShakeMap”; Wald et al. 1999b),



Earthquakes, Intensity, Fig. 1 Relations between intensity versus PGA (left) and PGV (right) after Wald et al. (1999c), Faccioli and Cauzzi (2006), and Faenza and Michelini (2010). Although MMI values and

MCS values are used in these studies, the relations are given in a common scheme according to their de facto equivalence (Musson et al. 2010). (Redrawn after Faenza and Michelini (2010))

representing shaking through maps of PGA, PGV, and ground-motion shaking intensity, it was necessary to derive “instrumental intensity” by converting peak ground motion to MMI intensity values (Wald et al. 1999c; Fig. 1). In Europe, notably Italy, well-established relations for MCS intensities have been developed by Faenza and Michelini (2010) and references therein. Their orthogonal relations (Fig. 1) read

$$I = 1.68 + 2.58 \log PGA \text{ [cm/s}^2\text{]} \quad (6)$$

$$I = 5.11 + 2.35 \log PGV \text{ [cm/s]} \quad (7)$$

The densest network to generate ShakeMaps in terms of instrumental intensity exists in Japan. Here, the instrumental intensity according to the JMA scale (I_{JMA}) is computed from Fourier acceleration spectra multiplied with different filter functions (Kuwata and Takada 2002; Karim and Yamazaki 2002). Empirical relationships between I_{JMA} and different ground-motion parameters are given by Karim and Yamazaki (2002). Their relation between I_{JMA} and PGA was derived as

$$I_{JMA} = 1.58 + 1.38 \log SI + 0.59 \log PGA + 0.02 M \quad (8)$$

where SI is the “spectrum intensity” (cm/s^2) introduced by Housner (1961) as the integral over the velocity response spectrum in the period range from 0.1 to 2.5 s, and PGA is in cm/s^2 . However, correspondence between “instrumental intensity” and the wording of the macroseismic scale cannot be taken for granted.

The limitation of all these relationships is that intensity is not a function of either PGA or PGV, contrary to the

assumption firstly made by Omori (1900) and Cancani (1904), and the correlation is purely an indirect one. The duration of strong shaking is critical in determining earthquake damage (see, for instance, Gibson et al. 1995), and hence equations of the form of (6), (7), and (8) always have enormous scatter.

Conclusions

Macroseismic intensity is a robust measure of perceptible and potentially damaging ground motion and is essential in engineering seismology. Modern developments of macroseismic scales, defining intensity degrees with persons and buildings as sensors, fully consider the range of vulnerability of buildings and differentiate structural and nonstructural damage patterns. Intensity provides the principal means by which one can parameterize historical or pre-instrumental earthquakes, that is, to determine their location, magnitude, and focal depth. Good quality macroseismic data enables a similar range of uncertainty in the derived focal parameters as standard instrumental seismology. The increased relevance of the parameter macroseismic intensity is strongly related to the enhanced needs in reliable seismic hazard and risk assessments. With the Internet-based platforms ShakeMap and “Did You Feel It?”, macroseismic intensity has become omnipresent.

Cross-References

- [Earthquake, Magnitude](#)
- [Earthquakes, Early and Strong Motion Warning](#)
- [Earthquakes, Energy](#)

- Earthquakes, Location Techniques
- Earthquakes, ShakeMap
- Earthquakes, Strong-Ground Motion
- Seismic Hazard

Bibliography

- Atkinson GM, Wald DJ (2007) "Did you feel it?" intensity data: a surprisingly good measure of earthquake ground motion. *Seismol Res Lett* 78(3):362–368
- Bakun WH, Wentworth CM (1997) Estimating earthquake location and magnitude from seismic intensity data. *Bull Seismol Soc Am* 87(6):1502–1521
- Cancani A (1904) Sur l'emploi d'une double échelle sismique des intensités, empirique et absolute. *Beiträge zur Geophysik, Ergänzungsband (supplementary volume) 2, Verhandlungen der zweiten internationalen seismologischen Konferenz, Annex A10*, 2:281–283
- Faccioli E, Cauzzi C (2006) Macroseismic intensities for seismic scenarios, estimated from instrumentally based correlations. In: *Abstract Book 1st ECEES*, p 125. http://www.ecees.org/abstracts_book.pdf
- Faenza L, Michelini A (2010) Regression analysis of MCS intensity and ground motion parameters in Italy and its application in ShakeMap. *Geophys J Int* 180(3):1138–1152
- Frankel A (1994) Implications of felt area-magnitude relations for earthquake scaling and the average frequency of perceptible ground motion. *Bull Seismol Soc Am* 84(2):462–465
- Gasparini P, Bernardini F, Valensise G, Boschi E (1999) Defining seismogenic sources from historical earthquake felt reports. *Bull Seismol Soc Am* 89(1):94–110
- Gibson G, Wesson V, Jones T (1995) Strong motion from shallow intraplate earthquakes, Pacific Conference on Earthquake Engineering, 20–22 Nov 1995. *Proceedings*, pp 185–193
- Gomez Copera AA (2006) Seismic hazard map for the Italian territory using macroseismic data. *Earth Sci Res J* 10(2):67–90
- Grünthal G (1984) Seismische Gefährdung. In: Hurtig E, Stiller H (eds) *Erdbeben und Erdbebengefährdung*. Akademie Verlag, Berlin, pp 169–238
- Grünthal G (ed), Musson RMW, Schwarz J, Stucchi M (assoc eds) (1993) *European Macroseismic Scale 1992 (up-dated MSK scale)*. Cahiers du Centre Européen de Géodynamique et de Séismologie 7, Centre Européen de Géodynamique et de Séismologie, Luxembourg, 77 pp
- Grünthal G (ed), Musson RMW, Schwarz J, Stucchi M (assoc eds) (1998) *European Macroseismic Scale 1998 (EMS-98)*. Cahiers du Centre Européen de Géodynamique et de Séismologie 15, Centre Européen de Géodynamique et de Séismologie, Luxembourg, 99 pp. <https://doi.org/10.2312/EMS-98>. <http://www.gfz-potsdam.de/EMS98>
- Grünthal G, Wahlström R, Stromeyer D (2009) Harmonization check of M_w within the central, northern, and northwestern European earthquake catalogue (CENEC). *J Seismol* 13(4):613–632
- Gutenberg B, Richter CF (1942) Earthquake magnitude, intensity, energy and acceleration. *Bull Seismol Soc Am* 32:163–191
- Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy and acceleration (Second Paper). *Bull Seismol Soc Am* 46:105–145
- Housner GW (1961) *Vibration of structures induced by seismic waves, shock and vibration handbook*. McGraw Hill, New York, pp 1–32
- Howell BF Jr, Schultz TR (1975) Attenuation of modified Mercalli intensity with distance from the epicenter. *Bull Seismol Soc Am* 65(3):651–655
- Jánosi v I (1907) *Bearbeitung der makroseismischen Erdbeben auf Grund der "Cancanischen Gleichung."* Technical report, K. u. K. Reichsanstalt für Meteorologie und Erdmagnetismus, Budapest
- Japanese Meteorological Agency (1996) Explanation table of JMA seismic intensity scale. <http://www.jma.go.jp/jma/kishou/known/shindo/explane.html>
- Johnston AC (1996) Seismic moment assessment of earthquakes in stable continental regions – II. Historical seismicity. *Geophys J Int* 125(3):639–678
- Karim KR, Yamazaki F (2002) Correlation of JMA instrumental seismic intensity with strong motion parameters. *Earthq Eng Struct Dyn* 31(5):1191–1212
- Kövesligethy R (1906) A makroszeizmikus rengések feldolgozása. *Math és Természettudományi Értesítő* 24:349–368
- Kuwata Y, Takada S (2002) Instantaneous instrumental seismic intensity and evacuation. *J Nat Dis Sci* 24(1):35–42
- Medvedev SV, Sponheuer W, Kárník V (1965a) Seismic intensity scale MSK-64. UNESCO digital library, NS/SEISM/28, WS/0565.43 AVS, Paris. <https://unesco.unesco.org/ark:/48223/pf0000154508>
- Medvedev SV, Sponheuer W, Kárník V (1965b) Seismische Intensitätsskala MSK 1964. In: Sponheuer W (ed) *Bericht über die Weiterentwicklung der seismischen Skala*, vol 8. Veröff. Institut für Geodynamik Jena, Berlin, pp 12–21
- Musson RWM (1993) Macroscopic magnitude and depth for British earthquakes. BGS Global Seismology Report no. WL/ 93/13
- Musson RMW (2000) Intensity-based seismic risk assessment. *Soil Dyn Earthq Eng* 20:353–360
- Musson RMW (2006) Automatic assessment of EMS-98 intensities. British Geological Survey, Seismology and Geomagnetic Programme, Internal Report IR/06/048, 16 pp
- Musson RMW, Cécic I (2002) Macroscopic seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*. Academic, San Diego, pp 807–822
- Musson RWM, Grünthal G, Stucchi M (2010) The comparison of macroseismic intensity scales. *J Seismol* 14(2):413–428
- Omori F (1900) Seismic experiments on the fracturing and overturning of columns. *Publ Earthq Invest Comm Foreign Lang* 4:69–141
- Richter CF (1958) *Elementary seismology. A series of books in geology*. Freeman, San Francisco. 768 pp
- Sieberg A (1912) Über die makroseismische Bestimmung der Erdbebenstärke. *Gerlands Beiträge zur Geophysik* 11:227–239
- Sieberg A (1923) *Geologische, physikalische und angewandte Erdbebenkunde*. G. Fischer, Jena
- Sieberg A (1932) *Geologie der Erdbeben. Handbuch der Geophysik* 2(4):550–555
- Sponheuer W (1960) Methoden zur Herdtiefenbestimmung in der Makroseismik. *Freiberger Forschungshefte C* 88, Geophysik. Akademie-Verlag Berlin, 117 pp
- Stover CW, Coffman JL (1993) *Seismicity of the United States, 1568–1989 (revised)*, U.S. Geological Survey, Reston, Prof. Pap. 1527, 418 pp
- Stromeyer D, Grünthal G (2009) Attenuation relationship of macroseismic intensities in Central Europe. *Bull Seismol Soc Am* 99(2A):554–565
- Tyagunov S, Grünthal G, Wahlström R, Stempniewski L, Zschau J (2006) Seismic risk mapping for Germany. *Nat Hazards Earth Syst Sci* 6(4):573–586
- Vogt J, Musson RMW, Stucchi M (1994) Seismological and hydrological criteria for the new European macroseismic scale (MSK-92). *Nat Hazards* 10:1–6
- Wald DF, Quiroriano V, Heaton TH, Kanamori H (1999a) Relationships between peak ground acceleration, peak ground velocity, and modified Mercalli intensity in California. *Earthquake Spectra* 15(3):557–564

- Wald DF, Quitoriano V, Dengler LA, Dewey JW (1999b) Utilization of the internet for rapid community intensity maps. *Seismol Res Lett* 70(1):680–697
- Wald DF, Quitoriano V, Heaton TH, Kanamori H, Scrivner CW, Worden CB (1999c) TriNet “shake maps”: rapid generation of peak ground motion and intensity maps for earthquake in southern California. *Earthquake Spectra* 15(3):537–556
- Wood HO, Neumann F (1931) Modified Mercalli intensity scale of 1931. *Bull Seismol Soc Am* 21:277–283

Earthquakes, Intraplate

Prantik Mandal
 Earthquake Seismology, CSIR-National Geophysical
 Research Institute, Hyderabad, Telangana, India

Synonyms

Stable continental earthquakes

Definition

The plate tectonics revolution of the mid-twentieth century provided Earth scientists with the fundamental framework to understand why most earthquakes occur within geographically limited active belts around the world along the plate boundaries. However, large devastating continental earthquakes do also occur in the plate interior far away from plate boundaries, which are known as intraplate earthquakes.

Mechanism of Intraplate Earthquakes

Moderate to large size devastating continental intraplate earthquakes have occurred at many locations in the world, e.g., 1811–1812 New Madrid (USA); 2008 Wenchuan (China); 2016 Christchurch (New Zealand); 1988 Tennant Creek (Australia); 1989 Ungava (Canada) and 2001 Kachchh, Koyana; and 1993 Latur, India. The occurrences of these earthquakes account for only 0.5% of the average annual global energy release by earthquakes (Johnston and Kanter 1990). The locations and mechanisms of $M \geq 6$ intraplate earthquakes during 1900–2019 are listed in Table 1, with 46% thrust, 46% strike-slip, and 8% normal mechanisms. Thus, in general, these earthquakes occur in compressive stress regime. Nevertheless, these earthquakes are rare and can occur near to the thickly populated areas, causing large number of deaths/injuries as well as financial losses. The

recent occurrences of continental earthquakes have caused substantial damages and deaths due to the growing population and poor constructions. The 2001 Bhuj (India) of $M_w 7.7$ had claimed ~20,000 human lives, while the most devastating 2008 Wenchuan (China) earthquake of $M_w 7.9$ claimed ~87,000 human lives.

The generation of intraplate earthquakes has been attributed to the sudden movement along the pre-existing weak zones due to local stress perturbation related to the regional plate boundary forces, topography and crustal density inhomogeneities, reservoir loading, presence of fluids, etc. (Campbell 1978; Mareschal and Kuang 1986; Johnston and Kanter 1990). There are numerous examples of intraplate earthquake nucleation at the intersections of weak zones or along old rifts in response to the regional plate boundary forces (Talwani and Rajendran 1991; Johnston and Kanter 1990). In 1978, Campbell proposed that crustal intrusive bodies may also induce significant localized stresses in a stable continental setup. However, the intermediate to deep crustal earthquakes of moderate-to-large magnitudes have occurred in several continental rift zones like the Amazon (Brazil), East Africa (Africa), Baikal (Russia), Rio Grande (USA), Kachchh (India), Narmada (India), and Reelfoot (New Madrid, USA), which have been explained in terms of failure of old rift due to local stress perturbations associated with substantial crustal as well as asthenospheric thinning, intruded magmatic material from the underlying mantle and deepening of the brittle-ductile transition depth (Nyblade and Langston 1995; Johnston and Kanter 1990; Pollitz et al. 2002). Various local stress sources have been explored so far to explain the seismicity associated with the New Madrid (USA) and Kachchh (India) rift zones (Johnston 1996; Mandal 2019). For instance, New Madrid seismicity has been attributed to stress changes induced by melting of the Laurentide ice sheet or stresses generated by sinking of an ancient high-density mafic body or from a sudden weakening of the lower crust (Pollitz et al. 2002). Similarly, the seismicity in Kachchh (India) has been attributed to the stresses caused by mafic crustal intrusive bodies and in-plane plate compression, which has further been facilitated by the presence of aqueous fluids or volatile CO_2 released from the metamorphism of olivine-rich lower crustal rocks (Mandal 2019).

In typical continental rift zones, like New Madrid (USA) and Amazon (Brazil), the intraplate earthquake occurrences are attributed to localized stress generation due to density heterogeneity in the deep lower crust (Zoback and Richardson 1996). It is observed that the local tectonic forces, such as topography, and crustal density heterogeneities can contribute a sizable amount of differential stresses of the order of 100 MPa to the lithospheric stress regime (Mareschal and Kuang 1986). These stresses were also well correlated with

Earthquakes, Intraplate, Table 1 List of $M \geq 6$ SCR earthquakes during 1900–2019

SN	Date	OT	LAT	LONG	DEP	M	MECH	Locality
1	19250301	02:19:00	47.76 N	69.84 W	10 $\pm$ 2	6.4	TH	Charlevoix Quebec
2	19291118	20:32:00	44.69 N	56.01 W	20	7.3	SS	Grand Banks, N. Scotia
3	19331120	23:21:32	73.23 N	70.02 W	10	7.4	TH	Baffin Bay, Canada
4	19351101	06:03:34	48.89 N	79.00 W	10 $\pm$ 2	6.2	TH	Timiskaming, Ontario
5	19350419	15:23:24	31.23 N	15.31E	21 $\pm$ 3	6.8	SS	Sirt Basin, Libya
6	19350419	20:31:39	30.93 N	15.47E	15 $\pm$ 3	6.4	SS	Sirt Basin, Libya
7	19350420	05:10:56	30.93 N	15.56E	19 $\pm$ 3	6.4	SS	Sirt Basin, Libya
8	19390622	19:19:31	5.18 N	00.13 W	14 $\pm$ 4	6.5	SS	Accra Ghana
9	19410429	01:35:41	26.8S	116.1E	32	6.7	–	Meeberrie, WA
10	19450912	00:51:20	2.28 N	15.9 E	11 $\pm$ 9	6.1	SS	Cameroon
11	19550912	06:09:24	32.2 N	29.6 E	15 $\pm$ 3	6.4	TH	Egypt Offshore
12	19560721	15:32:30	23.3 N	70.0E	15	6.0	TH	Anjar, Kachchh, India
13	19570502	03:55:35	72.15 N	67.95 W	24 $\pm$ 4	6.2	TH	Baffin Bay, Canada
14	19620318	20:18:53	23.72 N	114.67E	7 $\pm$ 2	6.0	SS	Heyuan Kwantung
15	19630904	13:32:11	71.25 N	72.94 W	7	6.2	NS	Baffin Island, Canada
16	19671210	22:51:19	17.54 N	73.84E	4.5 $\pm$ 1.5	6.3	SN	Koyna, India
17	19681014	02:58:51	31.62S	116.98E	3 $\pm$ 2	6.6	TH	Meckering, WA
18	19690929	20:03:26	33.09 N	19.52E	11	6.4	SS	Ceres, South Africa
19	19700324	10:35:17	22.08S	126.65E	8 $\pm$ 3	6.0	TH	Canning Basin, WA
20	19721121	10:06:30	76.58 N	106.02 W	29 $\pm$ 4	6.0	NS	Sverdrup basin, Canada
21	19721227	22:59:26	76.80 N	106.49 W	11 $\pm$ 3	6.3	SS	Sverdrup basin, Canada
22	19740923	19:28:17	0.28 S	12.83E	9 $\pm$ 5	6.0	SS	Gabon
23	19740927	04:09:02	2.72 N	71.37 W	6	6.3	–	Eastern Columbia
24	19790423	05:45:10	16.66S	120.16E	26 $\pm$ 3	6.1	SS	Offshore, NW Australia
25	19831222	04:11:29	11.98 N	13.54 W	8 $\pm$ 1.5	6.2	NS	Guinea
26	19840521	15:38:59	32.67 N	121.51E	12 $\pm$ 4	6.1	ST	South Yellow Sea
27	19880122	00:35:58	19.83S	133.93E	6.5 $\pm$ 1	6.3	TH	Tennant Creek Australia
28	19880122	03:57:25	19.81S	133.92E	3.5 $\pm$ 0.5	6.4	TS	Tennant Creek Australia
29	19880122	12:05:00	19.85S	133.95E	4.5 $\pm$ 0.5	6.6	TH	Tennant Creek Australia
30	19891225	14:24:33	60.12 N	73.60 W	2.5 $\pm$ 2.5	6.0	TH	Ungava, N. Quebec
31	19930929	22:25:49	18.07 N	76.45E	5	6.2	TH	Latur, India
32	19940916	06:20:18	22.55 N	118.74E	12	6.8	NO	Taiwan Straits
33	20010126	03:16:40	23.42 N	70.23E	23	7.7	TH	Bhuj, Kachchh, India
34	20080512	06:28:01	31.02 N	103.37E	19	7.9	TH	Sichuan, China
35	20100903	16:35:46	43.55S	172.18E	10	7.1	SS	Canterbury, New Zealand
36	20110222	23:51:42	43.58S	172.68E	5	6.2	TH	Christchurch, New Zealand
37	20161113	11:02:56	42.76S	173.1E	15	7.8	TH	South Island, New Zealand

Type of faulting as defined in Johnston (1994):

TH: Thrust

TS: thrust with strike-slip component

NO: Normal

NS: normal with strike-slip component

SS: Strike-slip

ST: strike-slip with thrust component

SN: strike-slip with normal component

the geographical distribution of intraplate earthquakes in some regions of the USA, Canada, and India (Mareschal and Kuang 1986; Mandal et al. 1997). A good agreement between the occurrences of intraplate earthquakes with the change in lithospheric elastic thickness owing to rheology and local buoyancy forces has been noticed in the Western USA that shows the importance of local tectonic forces over the plate tectonic forces in causing intraplate seismic activity (Lowry and Smith 1995).

Summary and Conclusions

Intraplate earthquake generally occurs in the strong continental crust; thus, their seismogenesis is not well understood. Multidisciplinary geophysical and geological studies are required to understand their causative mechanism. In general, it is observed that large intraplate earthquakes can cause significant devastation to mankind; thus, the hazards concerning the occurrences of these earthquakes are very

important to access for taking proper preventive measures to create an earthquake-resilient society.

Cross-References

- [Earthquake Source Theory](#)
- [Lithosphere, Continental](#)
- [Plate Tectonics, Precambrian](#)

Bibliography

- Campbell DL (1978) Investigation of the stress-concentration mechanism for intraplate earthquakes. *Geophys Res Lett* 5:477–479
- Johnston AC (1996) Seismic moment assessment of earthquakes in stable continental regions-I. Instrumental seismicity. *Geophys J Int* 124:381–414
- Johnston AC, Kanter LR (1990) Earthquakes in stable continental crust. *Sci Am* 262:54–68
- Lowry AR, Smith RB (1995) Strength and rheology of the western U.S. Cordillera. *J Geophys Res* 100:947–917, 963
- Mandal P (2019) A possible origin of intraplate earthquakes in the Kachchh rift zone, India, since the 2001 M_w 7.7 Bhuj earthquake. *J Asian Earth Sci* 170:56–72
- Mandal P, Manglik A, Singh RN (1997) Intraplate stress distribution beneath the Killari region, India. *J Geophys Res* 102:11719–11729
- Mareschal JC, Kuang J (1986) Intraplate stresses and seismicity: the role of topography and density heterogeneities. *Tectonophysics* 132:153–162
- Nyblade AA, Langston CA (1995) East African earthquakes below 20 km depth and their implications for crustal structure. *Geophys J Int* 121:49–62
- Pollitz FF, Kellogg L, Burgmann R (2002) Sinking mafic body in a reactivated lower crust: a mechanism for stress concentration at the New Madrid Seismic zone. *Bull Seismol Soc Am* 91:1882–1897
- Talwani P, Rajendran K (1991) Some seismological and geometric features of intraplate earthquakes. *Tectonophysics* 186:19–41
- Zoback ML, Richardson RM (1996) Stress perturbation associated with Amazonas and other ancient continental rifts. *J Geophys Res* 101:5459–5475

Earthquakes, Location Techniques

Clifford H. Thurber
Department of Geoscience, University of Wisconsin-Madison, Madison, WI, USA

Synonyms

Hypocenter determination

Definition

Earthquake location. Methods for determining the latitude, longitude, depth, and time of origin of a seismic event (earthquake, explosion, etc.) using the arrival times of seismic waves.

Introduction

Knowing the location of an earthquake (its latitude, longitude, depth, and origin time) is an essential starting point for the vast majority of quantitative seismological analyses. In the context of this entry, earthquake location will be taken to mean determining the initiation point of fault rupture, i.e., the hypocenter. After briefly tracing some of the early history of methods for the quantitative determination of earthquake locations, the focus of discussion turns to modern methods for single-event location and uncertainty estimation. The final section covers multiple-event location, with an emphasis on the substantial to – in some cases – profound improvement in relative location accuracy that can be achieved by combining waveform cross-correlation with relative location techniques.

Early History

John “Earthquake” Milne devised some of the first quantitatively based methods for earthquake location, known as the methods of circles, hyperbolas, and coordinates (Milne 1886). For the method of circles, the difference in arrival time between later observing stations i ($i = 1, N$) and the first observing station 0, $t_i - t_0 = dt_i$, is used to construct circles of radii $V dt_i$, where V is the velocity of “shock” (i.e., seismic wave) propagation, and the center of the circle through station 0 that is roughly tangent to all the other circles defines the earthquake epicenter. The scale factor V is determined by trial and error. The commonly taught S minus P time location method (Bolt 2006) is similar in concept. The method of hyperbolas is a variant on the method of circles, whereby differential times between pairs of stations are used to define hyperbolas of possible locations and the intersection of hyperbolas defines the earthquake epicenter. The method of coordinates goes a step further by setting up a system of four equations (defined in Cartesian coordinates centered on the last observing station) relating differential times to “shock” velocity and the (x,y,z) coordinates of the earthquake. With four equations and four unknowns, the solution is unique. These methods were employed for many years, until slowly the method of Geiger (1910, 1912), discussed below, began to supplant it. Interestingly, Milne’s methods have been revived recently for use in locating earthquakes in an early warning system (Satriano et al. 2008) and in tectonic (Zhou 1994; Font et al. 2004) and volcanic (Sumiejski et al. 2009) settings.

Single-Event Location

The predominant method for earthquake location is based on what is commonly known as Geiger’s method, which was developed early in the twentieth century (Geiger 1910, 1912). This method involves deriving the least-squares (L2 norm)

solution to a set of equations, for a trial hypocenter, of the form

$$r_i = \sum_{j=1}^3 \frac{\partial T_i}{\partial x_j} \Delta x_j + \Delta t_0 \quad (1)$$

where r_i is the residual at station i , defined as the observed arrival time of a seismic phase t_i minus the calculated arrival time given by the origin time t_0 plus the calculated travel time T_i , the x_j 's are the hypocenter coordinates, and Δ indicates the (unknown) parameter perturbations. The spatial derivatives $\partial T_i / \partial x_j$ have a simple relationship to the seismic velocity V and the ray direction at the source,

$$\frac{\partial T_i}{\partial x_j} = \frac{-1}{V} \frac{dx_j}{ds} \quad (2)$$

where s is the path length parameter. Equation 1 represents a nonlinear parameter estimation problem (Aster et al. 2019), so it is solved iteratively, and some convergence criteria must be defined. Mathematically, this problem can be expressed as

$$\mathbf{G}\Delta\mathbf{m} = \mathbf{r} \quad (3)$$

where the matrix $\mathbf{G}$ contains the partial derivatives, $\Delta\mathbf{m}$ is the vector of hypocenter perturbations (position and origin time), and $\mathbf{r}$ is the residual vector. Equation 3 is typically solved using iterative least squares, generally with some type of regularization (Aster et al. 2019). For a layered model, Geiger's method can be extended from this simple linearized form to a nonlinear form (Thurber 1985), but that is of more theoretical than practical interest.

The seismic phases used for location vary depending in part on the epicentral distance domain – local (0° to 5°), regional (5° to 30°), and teleseismic ($>30^\circ$) (note that these breakpoints are relatively standard ones, although that for the local-regional is somewhat more arbitrary). At local distances, first-arriving P and S phases are used almost to exclusion. At teleseismic distances, first-P and to a lesser degree first-S remain critically important, albeit with such complications as shadow zones and triplications, but a virtual cornucopia of other phases is also available. Among them, the depth phases (solid-surface reflections pP and sP and the water-surface reflections pwP and swP), which have travel-time derivatives for depth opposite in sign to the direct phases, have tremendous value in constraining event depth, although misidentification problems can be significant (Engdahl et al. 1998; 2006). The outer core reflection PcP has also proven quite useful (Schöffel and Das 1999).

At regional distances, several factors combine to make the location situation complex. The mantle refractions Pn and Sn are the first-arriving body-wave phases, and their amplitudes

can be small relative to crustal arrivals. Crustal thickness variations at a regional scale make accurate prediction of regional travel times difficult. These problems have an enormous impact on nuclear explosion monitoring efforts, where the goal is to be able to locate relatively small events often with only sparse regional data. The use of later-arriving phases (crustal body-wave phases Pg, Sg, and PmP, and the surface-wave phase Lg) and also arrival azimuth is relatively common for this monitoring domain (Bratt and Bache 1988). Arrival azimuth can be determined at high precision with a seismic array or by a single three-component station with much lower precision.

There are some alternatives to Geiger's method, most of which fall into the general category of nonlinear optimization (Gill et al. 1981). One is the conceptually simple grid search approach (Rowlett and Forsyth 1984; Sambridge and Kennett 1986; Shearer 1997; Lomax et al. 2000; Rodi 2006). Others include the simplex method (Rabinowitz 1988), genetic algorithms (Sambridge and Gallagher 1993), and simulated annealing (Billings et al. 1994). Because these methods do not involve least squares and the calculation of derivatives (Eqs. 2 and 3), they are readily amenable to the use of the more robust L1 norm (Aster et al. 2019), and some can directly provide comprehensive uncertainty information, a topic we turn to next.

Location Uncertainty

In some circumstances, knowing the uncertainty of a seismic event's location may be nearly as important as the location itself. The discussion of this topic can be separated into two parts, precision and accuracy. Precision is covered first, including a brief discussion of arrival-time estimation (phase picking). Accuracy is a far more subtle issue, with significant limitations on our ability to assess it thoroughly. Husen and Hardebeck (2010) discuss these topics in detail.

The standard approach for the evaluation of precision is the confidence ellipsoid (Flinn 1965), which can be obtained from the model covariance matrix (Aster et al. 2019) for Eq. 3,

$$\mathbf{C}_m = \sigma_d^2 (\mathbf{G}^T \mathbf{G})^{-1} \left[(\mathbf{G}^T \mathbf{G})^{-1} \right]^T \quad (4)$$

where for now we assume that the data errors have a uniform standard deviation σ_d , often estimated from the data misfit, and are uncorrelated (independent). Given the desired confidence level p (e.g., 0.95), the corresponding confidence ellipsoid is obtained for $\mathbf{m}$ satisfying

$$(\mathbf{m} - \mathbf{m}^*)^T \mathbf{C}_m^{-1} (\mathbf{m} - \mathbf{m}^*) = \kappa_p^2 \quad (5)$$

where $\mathbf{m}^*$ is the location solution and κ_p^2 is given by

$$\kappa_p^2 = 4\sigma_d^2 [F_p(4, N-4)] / (N-4) \quad (6)$$

with F_p the F statistic at confidence level p and N the number of arrival-time data (Draper and Smith 1981). Some limitations of this approach to uncertainty estimation are (1) that the misfit-based estimate of σ_d becomes unreliable for small numbers of observations (Evernden 1969), (2) that it neglects the effects of velocity model inaccuracy and correlated errors, and (3) that the linear approximation itself is inadequate in some circumstances.

Buland (1976) noted that data uncertainty estimated from either the misfit or a priori experience is generally too small due to the inaccuracy of the travel-time model. He suggested that the σ_d value should simply be increased to compensate for this problem. Jordan and Sverdrup (1981) developed an elegant statistical approach for making this compensation. Their philosophy is to replace the usual a posteriori estimate of the data variance, s_d^2 , based on the data misfit for an individual earthquake,

$$s_d^2 = \|\mathbf{r}\|^2 / (N-4) \quad (7)$$

with an estimate based on a combination of a priori and a posteriori information:

$$s_d^2 = \frac{K s_0^2 + \|\mathbf{r}\|^2}{(K + N - 4)} \quad (8)$$

where the a priori variance estimate s_0 is given a relative weight of K . Thus the a priori information is treated as contributing K observations to the estimate of s_d^2 and the data for the particular event contribute N observations. For self-consistency, given a large number of events, the mean of the estimated variances should approach s_0 . Jordan and Sverdrup (1981) also show that the distribution of the estimated standard deviation needs to be consistent with the adopted value of K . In the limit of no prior information ($K = 0$), their estimate reverts to the standard formula (Flinn 1965), whereas for infinite prior information ($K = \infty$), their estimate becomes equivalent to that of Evernden (1969).

One limitation of the Jordan and Sverdrup (1981) approach is that it assumes an uncorrelated, Gaussian error distribution with a constant data variance for all observations. Varying data uncertainty estimates can be incorporated simply by including a weighting matrix $\mathbf{W}$ in the standard linearized analysis (Aster et al. 2019):

$$\mathbf{W} \mathbf{G} \Delta \mathbf{m} = \mathbf{W} \mathbf{r} \quad (9)$$

(e.g., Bolt 1960) and then solving the weighted system as before. Weighting is generally done based on reported arrival-time reading precision (Engdahl et al. 1998), reported arrival quality (Buland 1976), phase variance as a function of

distance (Engdahl 2006), residual size (Bolt 1960), station distance (Klein 1978), and/or azimuthal distribution of stations (Buland 1976). However, the Gaussian data error assumption is generally inappropriate, as residual outliers are common in arrival-time data (Bolt 1960; Buland 1986). Discussions of correlated errors and their treatment can be found in Chang et al. (1983), Rodi and Myers (2007), and Bondár and McLaughlin (2009).

The “uniform reduction” method (Jeffreys 1939) is effective for treating the situation of normally distributed errors plus a background of outliers. Bolt (1960) assumed a residual frequency distribution of the form

$$f(r) = a s + \frac{(1 \pm s)}{\sigma \sqrt{2\pi}} \exp \left(\pm \frac{(r \pm m)^2}{2\sigma^2} \right) \quad (10)$$

where a and s are constants and m and σ are the mean residual and its standard deviation. This corresponds to a residual weighting function of the form

$$w(r) = \left[1 + \mu \exp \left(\pm \frac{(r \pm m)^2}{2\sigma^2} \right) \right]^{\pm 1} \quad (11)$$

with $\mu = \sigma (2\pi)^{1/2} a s / (1 - s)$.

It is important to point out that estimating phase arrival times by itself is a difficult problem, particularly for S waves and other secondary phases. Many seismic networks use automated pickers, at least for preliminary locations. Simple auto pickers use a threshold of the ratio of the short-term average (STA) to the long-term average (LTA) of the signal amplitude (Allen 1978, 1982; Ruud and Husebye 1992). There are a wide variety of other (generally more computationally demanding) approaches. Among these are energy analysis (Earle and Shearer 1994), polarization analysis (Vidale 1986; Cichowicz 1993), autoregressive techniques (Maeda 1985; Sleeman and Eck 1999; Leonard and Kennett 1999; Leonard 2000), wavelet analysis (Zhang et al. 2003), or some combination (Wang and Teng 1997; Diehl et al. 2009).

As noted above, constraints on event locations can be improved markedly via the inclusion of secondary arrivals. Gomberg et al. (1990) and Wilcock and Toomey (1991) examine the importance of even a single S arrival in reducing location uncertainty. A “chicken and egg” problem with the use of later phases is that their correct identification often requires knowledge of the event depth and distance. Recognizing this problem, Kennett et al. (1995) and Engdahl et al. (1998) included a phase reassociation step in their location procedure. In their approach, catalog phase arrivals were reidentified after each relocation iteration using a statistically based association algorithm. This approach can be implemented by comparing probability density functions (PDF’s) for relevant phases, centered on their theoretical relative travel times for a given hypocenter, to the observed phase arrivals. When PDF’s overlap for a particular

phase, the idea is to assign a phase identification in a probabilistic manner based on the relevant PDF values, making sure not to assign the same phase to two different arrivals.

The above discussion focuses almost exclusively on location precision. There are two general approaches for evaluating location accuracy: theoretical and empirical. Pavlis (1986) carried out a thorough theoretical analysis of both precision and accuracy aspects for local earthquakes. He identified and evaluated three sources of hypocenter error: (1) measurement errors, (2) velocity model errors, and (3) nonlinearity. He treated the first via standard confidence ellipsoids, but the other two required the establishment of error bound estimates. Billings et al. (1994) explored aspects of (1) and (2) for the teleseismic case using a combination of Monte Carlo analysis and varying sets of observations, although their analyses involved real data, not synthetic data. One general conclusion of both these studies is that velocity-model-induced errors can be as significant as measurement errors.

“Ground truth” (GT) events provide a direct, empirical means to explore absolute location accuracy (also termed location bias). Having GT events allows one to evaluate location and uncertainty estimation strategies empirically that, for example, produce 95% confidence ellipses that overlap the true locations 95% of the time. This avenue of research has been motivated primarily by nuclear explosion monitoring needs (see ► “[Seismic Monitoring of Nuclear Explosions](#)”) but also is an important aspect of research on one-dimensional (1D) and three-dimensional (3D) earth models (see ► “[Seismic Tomography](#)”). See Bondár and McLaughlin (2009) and references therein for recent work on absolute location accuracy and Thurber and Engdahl (2000) for a review of the development of global 1D models.

The San Andreas Fault Observatory at Depth (SAFOD) provided a unique test of absolute location accuracy. One of SAFOD’s goals (unfortunately not yet realized) was to penetrate the rupture zone of a magnitude ~2 repeating earthquake. This required determining a borehole trajectory that would pass within 100 m of the target event. Using data from a dense temporary array, permanent networks, deep borehole instruments, and active-source experiments, Thurber et al. (2004) carried out 3D tomography and used “virtual earthquakes” (receiver gathers of shots recorded on borehole instruments, effectively providing sources of known location at depth) to assess absolute location accuracy. Their work succeeded, as subsequent borehole recordings of target earthquake aftershocks showed the target event to lie within 100 m of the borehole (Zoback et al. 2011).

Multiple-Event Location

There are two distinct but complementary aspects to this topic that will be discussed. One is the suite of methods used to

infer the relative locations of events, including joint hypocenter determination (JHD), hypocentroidal decomposition (HDC), and double-difference (DD) methods. Note that local earthquake tomography takes multiple-event location a step further to determine seismic velocity structure simultaneously with the locations of many earthquakes (see ► “[Seismic Tomography](#)”). The other is the use of waveform cross-correlation (WCC) methods to determine differential times (DT’s) and/or pick adjustments.

The use of catalog DT’s for determining relative locations of earthquakes (or explosions) has been a fruitful field of work for many decades (e.g., Douglas 1967). The precision of WCC DT’s allows for spectacular relative location results to be achieved over a broad range of spatial scales when DT’s from WCC are used in combination with joint location techniques. These and other studies have demonstrated the substantial improvement in the delineation of seismogenic features or in the accuracy of relative locations of GT events that is possible using multiple-event location methods in concert with high-precision arrival-time or differential-time data. WCC is covered first to streamline the subsequent discussion of the relative location techniques.

Waveform Cross-Correlation

For a pair of earthquakes that are located close to each other (relative to the predominant wavelength of the seismic signal from them) and have similar source mechanisms, the waveforms from them observed at a particular seismic station will be similar. The work by Poupinet et al. (1984) to use (frequency domain) WCC on the coda of similar earthquakes (“doublets”) to detect temporal changes in seismic velocity followed by the study of Fremont and Malone (1987) using the same WCC method to determine high-precision relative arrival times and earthquake locations at Mt. St. Helens ultimately led to an explosion in the use of WCC and location techniques in the 1990s and beyond.

The time-domain method for WCC is relatively simple although computationally demanding for large sets of events. Windowed waveform segments spanning the time of the relevant arrival (P or S) for a pair of earthquakes at a particular station are cross-correlated over a range of lag times to find the time lag corresponding to the best alignment of the waveforms (i.e., the maximum CC value). Waveforms are typically filtered either with a simple bandpass filter or with coherency weighting (Rowe et al. 2002a). For very similar events (those with a high CC value), it can be worthwhile to apply a frequency-domain technique such as cross-spectral phase to determine the lag to a precision below the sample rate. This procedure takes advantage of the fact that, as an example, for two frequencies that are a factor of two different, the phase difference at the higher frequency should be exactly double that at the lower frequency. The lag is determined by calculating the phase difference as a function of frequency and

fitting the result with a straight line through the origin at zero frequency.

Most of the studies using WCC adopt a threshold value (typically 0.7) below which DT's are discarded. This can result in good data being lost (correct lag value but low CC value – e.g., due to a noisy record) as well as bad data being used (incorrect lag value but high CC value – e.g., due to a cycle slip). Du et al. (2004a) developed a procedure to provide independent lag estimates for verification of WCC results. They used the bispectrum (BS) method, which works in the third-order spectral domain, to check the reliability of the CC-determined time delay. They calculated two time delays with the BS method, one using the bandpass-filtered waveforms and the other with the raw data, and used them to verify (select or reject) the CC estimate computed with the filtered waveforms. This approach appears to be particularly effective in reducing cycle-slip errors.

Relative Location

As noted above, the JHD method dates back to the work of Douglas (1967). In its basic form, the method takes the basic equation for earthquake location, Eq. 1, adds a “station correction” term Δs_i , and sets up a large set of equations for a suite of events, $k = 1, K$:

$$r_{ik} = \sum_{j=1}^3 \frac{\partial T_{ik}}{\partial x_{jk}} \Delta x_{jk} + \Delta t_{ok} + \Delta s_i \quad (12)$$

As noted by Wolfe (2002), JHD procedures assume that if the hypocentral separations among a set of earthquakes are small compared to the event-station distance and the scale length of heterogeneity, then the ray paths between the source region and a station are similar, and the travel-time error introduced by an incorrect velocity model has a nearly constant value for a given station, thus justifying the linear term Δs_i . Pujol (2000) provides a review of JHD methods mostly aimed at local-scale studies. Pavlis (1992) presents a comprehensive analysis of relative location errors.

Wolfe (2002) thoroughly analyzed the HDC and DD methods. Her analysis shows that the solutions for both operators are similar, with the double-difference operator weighting fits to travel times on stations that record more earthquakes more heavily. The HDC method was introduced by Jordan and Sverdrup (1981). As the name indicates, it involves decomposing the system Eq. 12 via projection operators into one involving just the locations of the events relative to the (unknown) average location (i.e., the hypocentroid). As Wolfe (2002) notes, the projection is in fact a travel-time differencing operation. The DD method of Waldhauser et al. (1999) and Waldhauser and Ellsworth (2000) constructs the event-pair difference equations directly, but it is distinguished from the original Jordan and Sverdrup

(1981) method by evaluating the spatial derivatives at different points rather than assuming they are constant and by including a separation-dependent weighting in order to reduce location-dependent path bias. Modern HDC methods also allow for spatially varying derivatives (Bergman and Engdahl 2001), and the appropriateness of separation-dependent weighting has been subject to debate. The Markov Chain Monte Carlo approach of Myers et al. (2007) takes a probabilistic approach that incorporates the probability of phase identification errors in the estimation of uncertainty.

Some examples of applications of WCC combined with relative location for microearthquakes at a local scale include Switzerland (Deichmann and Garcia-Fernandez 1992), Kilauea volcano, Hawaii (Got et al. 1994; Rubin et al. 1998), California (Dodge et al. 1995; Shearer 1997; Waldhauser et al. 1999; Waldhauser and Ellsworth 2000; Schaff et al. 2002), Iceland (Slunga et al. 1995), the Soultz geothermal field, France (Rowe et al. 2002b), Mount Pinatubo, the Philippines (Battaglia et al. 2004), New Zealand (Du et al. 2004b), and Soufriere Hills volcano, Montserrat (Rowe et al. 2004). At a regional scale, Schaff and Richards (2004) determined relative locations for an earthquake sequence in China using WCC on Lg waves. For teleseismic data, WCC/relative location methods have been applied to body waves for nuclear explosions at the Balapan test site (Phillips et al. 2001; Thurber et al. 2001), to the aftershock sequences of the great Sumatra earthquakes (Pesicek et al. 2010), and to global seismicity (Waldhauser and Schaff 2008).

Time Reversal and Migration Location

An idea adapted from the exploration industry by McMechan (1982), but long-established in the acoustics literature (Andersen et al. 2008), completely avoids arrival-time picking, instead relying on the time reversibility of the wave equation to “image” earthquake source locations. In a classic (but “reciprocal”) time-reversal experiment reported by Parvulescu and Clay (1965), a microphone was placed about 25 ft away from a loudspeaker in a room, and short impulsive signals were emitted by the loudspeaker and recorded on tape through the microphone. The tape was then played back in reverse through the loudspeaker, resulting in a simple “click” sound at the microphone due to the reverse-time focusing of the signal. A similar field demonstration was successfully carried out with two research ships at an oceanic trench (Parvulescu and Clay 1965). See the encyclopedia entry on time reversal in seismology by Larmat and Clay (2019). Time-reversal localization of microearthquakes in near-real time is now common practice in the enhanced petroleum recovery (fracking) domain, where it is generally termed migration location (Artman et al. 2010). An

accurate seismic velocity model is important for effective focusing at the hypocenter, and the quality of the results is dependent on the acquisition geometry, as well as signal-to-noise ratio. Pesicek et al. (2014) compared pick-based, double-difference location with cross-correlation to migration for an industry dataset. The two methods yielded comparable results, but the migration approach may be preferred for small aperture surface arrays due to its rapid processing capability.

Summary

Despite being one of the oldest quantitative problems in seismology, it is clear that there is still much to learn about earthquake location. The combination of WCC and high-precision relative location brought new excitement to this rather old field of research, and time-reversal methods are providing an approach that is very effective in near-real-time analysis situations. At the same time, the needs of nuclear explosion monitoring have motivated the careful study of absolute location capability and statistically consistent uncertainty estimation.

Cross-References

- Seismic Monitoring of Nuclear Explosions
- Seismic Tomography
- Time-Reversal in Seismology

Bibliography

- Allen RV (1978) Automatic earthquake recognition and timing from single traces. *Bull Seismol Soc Am* 68:1521
- Allen RV (1982) Automatic phase pickers: their present use and future prospects. *Bull Seismol Soc Am* 72:S225
- Anderson BE, Griffa M, Larmat C, Ulrich TJ, Johnson PA (2008) Time reversal. *Acoust Today* 4:5
- Artman B, Podladtchikov I, Witten B (2010) Source location using time-reverse imaging. *Geophys Prospect* 58:861
- Aster R, Borchers B, Thurber C (2019) Parameter estimation and inverse problems, 3rd edn. Elsevier/Academic Press, Amsterdam
- Battaglia J, Thurber C, Got J-L, Rowe C, White R (2004) Precise relocation of earthquakes following the June 15, 1991 explosion of mount Pinatubo (Philippines). *J Geophys Res* 109:B07302
- Bergman EA, Engdahl ER (2001) Validation and generation of reference events by cluster analysis. *Seism Res Lett* 72:3
- Billings SD (1994) Simulated annealing for earthquake location. *Geophys J Int* 118:680
- Billings SD, Sambridge MS, Kennett BLN (1994) Errors in hypocenter location: picking, model, and magnitude dependence. *Bull Seismol Soc Am* 84:1978
- Bolt BA (1960) The revision of earthquake epicentres, focal depths and origin times using a high-speed computer. *Geophys J R Astron Soc* 3:433
- Bolt BA (2006) Earthquakes. W.H. Freeman, New York
- Bondár I, McLaughlin K (2009) A new ground truth data set for seismic studies. *Seismol Res Lett* 80:465
- Bratt SR, Bache TC (1988) Locating events with a sparse network of regional arrays. *Bull Seismol Soc Am* 78:780–798
- Buland R (1976) The mechanics of locating earthquakes. *Bull Seismol Soc Am* 66:173
- Buland R (1986) Uniform reduction error analysis. *Bull Seismol Soc Am* 76:217
- Chang AC, Shumway RH, Blandford RR, Barker BW (1983) Two methods to improve location estimates – preliminary results. *Bull Seismol Soc Am* 73:281
- Cichowicz A (1993) An automatic S-phase picker. *Bull Seismol Soc Am* 83:180
- Deichmann N, Garcia-Fernandez M (1992) Rupture geometry from high precision relative hypocentre locations of microearthquake clusters. *Geophys J Int* 110:501–517
- Diehl T, Deichmann N, Kissling E, Husen S (2009) Automatic S-wave picker for local earthquake tomography. *Bull Seismol Soc Am* 99:1906
- Dodge DA, Beroza GC, Ellsworth WL (1995) Foreshock sequence of the 1992 Landers, California, earthquake and its implications for earthquake nucleation. *J Geophys Res* 100:9865
- Douglas A (1967) Joint epicenter determination. *Nature* 215:47
- Draper NR, Smith H (1981) Applied regression analysis. Wiley, New York
- Du W, Thurber CH, Eberhart-Phillips D (2004a) Earthquake relocation using cross-correlation time delay estimates verified with the bispectrum method. *Bull Seismol Soc Am* 94:856
- Du W, Thurber CH, Reyners M, Eberhart-Phillips D, Zhang H (2004b) New constraints on seismicity in the Wellington region, New Zealand, from relocated earthquake hypocenters. *Geophys J Int* 158:1088
- Earle P, Shearer PM (1994) Characterization of global seismograms using an automatic-picking algorithm. *Bull Seismol Soc Am* 84:366
- Engdahl ER (2006) Application of an improved algorithm to high precision relocation of ISC test events. *Phys Earth Planet Inter* 158:14
- Engdahl ER, van der Hilst R, Buland R (1998) Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bull Seismol Soc Am* 88:722
- Evernden JF (1969) Precision of epicenters obtained by small numbers of world-wide stations. *Bull Seismol Soc Am* 59:1365
- Flinn EA (1965) Confidence regions and error determinations for seismic event location. *Rev Geophys* 3:157
- Font Y, Honn K, Lallemand S, Liu C, Chiao L (2004) Hypocentre determination offshore of eastern Taiwan using the maximum intersection method. *Geophys J Int* 158:655
- Fremont M-J, Malone SD (1987) High-precision relative location of earthquakes at Mount St. Helens, Washington. *J Geophys Res* 92:10223
- Geiger L (1910) Herdbestimmung bei erdbeben ans den ankunftszeiten. *Königlichen Gesellschaft der Wissenschaften zu Göttingen* 4:331
- Geiger L (1912) Probability method for the determination of earthquake epicenters from the arrival time only. *Bull St Louis Univ* 8:60
- Gill PE, Murray W, Wright MH (1981) Practical optimization. Academic, New York
- Gomberg JS, Shedlock KM, Roecker SW (1990) The effect of S-wave arrival times on the accuracy of hypocenter estimation. *Bull Seismol Soc Am* 80:1605
- Got J-L, Frechet J, Klein F (1994) Deep fault plane geometry inferred from multiplet relative relocation beneath the south flank of Kilauea. *J Geophys Res* 99:15375
- Husen S, Hardebeck JL (2010) Earthquake location accuracy. *Community Online Res Stat Seismicity Anal*. <https://doi.org/10.5078/corssa-55815573>

- Jeffreys H (1939) *Theory of probability*. Oxford University Press, London
- Jordan TH, Sverdrup KA (1981) Teleseismic location techniques and their application to earthquake clusters in the south-Central Pacific. *Bull Seismol Soc Am* 71:1105
- Kennett BLN, Engdahl ER, Buland R (1995) Constraints on seismic velocities in the earth from travel times. *Geophys J Int* 122:108
- Klein FW (1978) Hypocenter location program HYPOINVERSE. U.S. Geological Survey Open-File Report 78–694, p 113
- Larmat C, Clay CS (2019) *Time reversal methods*. Springer, Encyclopedia of Solid Earth Geophysics
- Leonard M (2000) Comparison of manual and automatic onset time picking. *Bull Seismol Soc Am* 90:1384
- Leonard M, Kennett BLN (1999) Multi-component autoregressive techniques for the analysis of seismograms. *Phys Earth Planet Inter* 113:247
- Lomax A, Virieux J, Volant P, Berge C (2000) Probabilistic earthquake location in 3D and layered models: introduction of a Metropolis-Gibbs method and comparison with linear locations. In: Thurber CH, Rabinowitz N (eds) *Advances in seismic event location*. Kluwer, Amsterdam, pp 101–134
- Maeda N (1985) A method for reading and checking phase times in autoprocesing system of seismic wave data. *Zisin Jishin* 38:365
- McMechan GA (1982) Determination of source parameters by wavefield extrapolation. *Geophys J R Astron Soc* 71:613
- Milne J (1886) *Earthquakes and other earth movements*. Appleton, New York
- Myers SC, Johannesson G, Hanley W (2007) A Bayesian hierarchical method for multiple-event seismic location. *Geophys J Int* 171:1049
- Parvulescu A, Clay CS (1965) Reproducibility of signal transmission in the ocean. *Radio Electron Eng* 29:223
- Pavlis GL (1986) Appraising earthquake hypocenter location errors: a complete, practical approach for single event locations. *Bull Seismol Soc Am* 76:1699
- Pavlis GL (1992) Appraising relative earthquake location errors. *Bull Seismol Soc Am* 82:836
- Pesicek JD, Thurber CH, Zhang H, DeShon HR, Engdahl ER, Widiyantoro S (2010) Teleseismic double-difference relocation of earthquakes along the Sumatra-Andaman subduction zone with a three-dimensional model using a 3-D model. *J Geophys Res* 115: B10303. <https://doi.org/10.1029/2010JB007443>
- Pesicek JD, Child D, Artman B, Cieřlik K (2014) Picking versus stacking in a modern microearthquake location: comparison of results from a surface passive seismic monitoring array in Oklahoma. *Geophysics* 79:KS61–KS68
- Phillips WS, Hartse HE, Steck LK (2001) Precise relative location of 25 ton chemical explosions at Balapan using IMS stations. *Pure Appl Geophys* 158:173
- Poupinet G, Ellsworth WL, Fréchet J (1984) Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras fault, California. *J Geophys Res* 89:5719
- Pujol J (2000) Joint event location: the JHD technique and applications to data from local seismic networks. In: *Advances in seismic event location*. Kluwer Academic Publishers, Dordrecht, pp 71–99
- Rabinowitz N (1988) Microearthquake location by means of nonlinear simplex procedure. *Bull Seismol Soc Am* 78:380
- Rodi WL (2006) Grid-search event location with non-Gaussian error models. *Phys Earth Planet Inter* 158:55
- Rodi WL, Myers SC (2007) Modeling travel-time correlations based on sensitivity kernels and correlated velocity anomalies. In: *Proceedings of the 29th monitoring research review: ground-based nuclear explosion monitoring technologies*, Denver, 25–27 Sept
- Rowe CA, Aster RC, Borchers B, Young CJ (2002a) An automatic, adaptive algorithm for refining phase picks in large seismic data sets. *Bull Seismol Soc Am* 92:1660
- Rowe CA, Aster RC, Phillips WS, Jones RH, Borchers B, Fehler MC (2002b) Using automated, high-precision repicking to improve delineation of microseismic structures at the Soultz geothermal reservoir. *Pure Appl Geophys* 159:536
- Rowe CA, Thurber CH, White RA (2004) Dome growth behavior at Soufriere Hills volcano, Montserrat, revealed by relocation of volcanic event swarms, 1995–1996. *J Volcanol Geotherm Res* 134:199
- Rowlett H, Forsyth DW (1984) Microearthquakes and recent faulting at the intersection of the Vema fracture zone and the mid-Atlantic ridge. *J Geophys Res* 89:6079
- Rubin AM, Gillard D, Got J-L (1998) A reinterpretation of seismicity associated with the January 1983 dike intrusion at Kilauea volcano, Hawaii. *J Geophys Res* 103:10003
- Rubin AM, Gillard D, Got J (1999) Streaks of microearthquakes along creeping faults. *Nature* 400:635
- Ruud BO, Husebye ES (1992) A new three-component detector and automatic single station bulletin production. *Bull Seismol Soc Am* 82:221
- Sambridge M, Gallagher K (1993) Earthquake hypocenter location using genetic algorithms. *Bull Seismol Soc Am* 83:1467
- Sambridge MS, Kennett BLN (1986) A novel method for hypocentre location. *Geophys J R Astron Soc* 87:679
- Satriano C, Lomax A, Zollo A (2008) Real-time evolutionary earthquake location for seismic early warning. *Bull Seismol Soc Am* 98:1482
- Schaff DP, Richards PG (2004) Repeating seismic events in China. *Science* 303:1176
- Schaff DP, Bokelmann GHR, Beroza GC, Waldhauser F, Ellsworth WL (2002) High-resolution image of Calaveras fault seismicity. *J Geophys Res* 107:2186
- Schöffel H-J, Das S (1999) Fine details of the Wadati-Benioff zone under Indonesia and its geodynamic implications. *J Geophys Res* 104:13101
- Shearer PM (1997) Improving local earthquake locations using the L1 norm and waveform cross correlation: application to the Whittier narrows, California, aftershock sequence. *J Geophys Res* 102:8269
- Sleeman R, van Eck T (1999) Robust automatic P-phase picking: an on-line implementation in the analysis of broadband seismogram recordings. *Phys Earth Planet Inter* 113:265
- Slunga R, Rögnvaldsson ST, Böðvarsson R (1995) Absolute and relative locations of similar events with application to microearthquakes in southern Iceland. *Geophys J Int* 123:409
- Sumiejski L, Thurber C, DeShon H (2009) Relocation of earthquake families associated with the 2006 eruption of Augustine volcano using the equal differential time method. *Geophys J Int* 176:1017
- Thurber CH (1985) Nonlinear earthquake location: theory and examples. *Bull Seismol Soc Am* 75:779
- Thurber CH, Engdahl ER (2000) Advances in global seismic event location. In: Thurber C, Rabinowitz N (eds) *Advances in seismic event location*. Kluwer Academic Publishers, Dordrecht, pp 3–22
- Thurber C, Trabant C, Haslinger F, Hartog R (2001) Nuclear explosion locations at the Balapan, Kazakhstan, nuclear test site: the effects of high-precision arrival times and three-dimensional structure. *Phys Earth Planet Inter* 123:283
- Thurber C, Roecker S, Zhang H, Baher S, Ellsworth W (2004) Fine-scale structure of the San Andreas fault and location of the SAFOD target earthquakes. *Geophys Res Lett* 31:L12S02
- Vidale JE (1986) Complex polarization analysis of particle motion. *Bull Seismol Soc Am* 76:1393–1405
- Waldhauser F, Ellsworth WL (2000) A double-difference earthquake location algorithm: method and application to the northern Hayward fault, California. *Bull Seismol Soc Am* 90:1353
- Waldhauser F, Schaff DP (2008) Large-scale relocation of two decades of Northern California seismicity using cross-correlation and double-difference methods. *J Geophys Res* 113:B08311

- Waldhauser F, Ellsworth WL, Cole A (1999) Slip-parallel seismic lineations on the Hayward fault and the San Andreas fault near Parkfield, California. *Geophys Res Lett* 70:246
- Wang J, Teng T (1997) Identification and picking of S phase using an artificial neural network. *Bull Seismol Soc Am* 87:1140
- Wilcock WS, Toomey DR (1991) Estimating hypocentral uncertainties for marine microearthquake surveys: a comparison of the generalized inverse and grid search methods. *Mar Geophys Res* 13:161
- Wolfe CJ (2002) On the mathematics of using difference operators to relocate earthquakes. *Bull Seismol Soc Am* 92:2879
- Zhang H, Thurber C, Rowe C (2003) Automatic P-wave arrival detection and picking with multiscale wavelet analysis for single-component recordings. *Bull Seismol Soc Am* 93:1904
- Zhou H-W (1994) Rapid three-dimensional hypocentral determination using a master station method. *J Geophys Res* 99:15439
- Zoback MD, Hickman S, Ellsworth WL (2011) Scientific drilling into the San Andreas fault zone – an overview of SAFOD's first five years. *Sci Drill* 11:14

Earthquakes, PAGER

David J. Wald, Kishor S. Jaiswal, Kristin D. Marano and Michael Hearne
U.S. Geological Survey, Denver Federal Center, Lakewood, CO, USA

Definition and Introduction

PAGER, short for Prompt Assessment of Global Earthquakes for Response, is an automated system developed and run by the US Geological Survey (USGS) that produces information concerning the impact of significant earthquakes around the world within approximately 20 min of any magnitude 5.5 or larger event. PAGER rapidly assesses earthquake impacts by combining populations exposed to estimates of shaking intensity with models of economic and fatality losses based on past earthquakes in each country or region of the world. The main purpose of the PAGER system is to inform emergency responders, government and aid agencies, and the media regarding the scope of the potential disaster. Earthquake alerts – which were formerly sent based on event magnitude and location or population exposure to shaking – are generated based on the estimated range of fatalities and/or economic losses.

The USGS's National Earthquake Information Center (NEIC) in Golden, Colorado, locates over 30,000 earthquakes a year. Tragically, about 20–30 of these cause significant damage, injuries, or fatalities. The USGS often detects earthquakes well before eyewitness reports are available. The USGS must then decide rapidly whether Federal and international agencies should be alerted to a potentially damaging event. In the past, the USGS primarily relied on the experience and intuition of its on-duty seismologists to estimate the impact of an event. To quantify and improve the accuracy of

the assessment, the USGS has developed PAGER, an automated system to rapidly estimate the shaking distribution, the number of people and settlements exposed to severe shaking, and the range of possible fatalities and economic losses. The estimated losses in turn trigger the appropriate color-coded alert (green, yellow, orange, or red), which correspond to suggested levels of response: no response, local/regional, national, or international response, respectively. In addition to direct alert notifications, PAGER provides important supplementary information, including comments describing the dominant types of vulnerable buildings in the region, population exposure, fatality reports from a few previous nearby earthquakes, and a summary of regionally specific information concerning the potential for secondary hazards, such as earthquake-induced landslides, tsunamis, and liquefaction. A map provides quick visual overviews of shaking levels and population densities, and a table summarizes the predicted shaking intensity at nearby cities and towns. This impact summary information is available on an event-by-event basis on the USGS Earthquake Program web site [<http://earthquake.usgs.gov/>] and as a downloadable and printable one-page report with an accompanying description, such as that shown in Fig. 1.

Fundamental to such a system, the USGS NEIC operates the robust computational and communication infrastructure necessary for earthquake response. PAGER results are generally available within 20 min of a significant earthquake, shortly after the determination of its location and magnitude. However, information on the extent of shaking will be uncertain in the minutes and hours following an earthquake and typically improves as additional sensor data and reported intensities are acquired and incorporated into models of the earthquake's source. Users of PAGER should account for the inherent uncertainty in the initial PAGER estimates and be attentive to updates posted to the USGS web site.

The PAGER Process

In general, the shaking-related impact of an earthquake is controlled by the distribution and severity of shaking, the population exposed at each shaking intensity level, and how vulnerable that population is to building damage at each intensity level. Population vulnerability is dominated by the degree of seismic resistance of the local building stock. The PAGER system takes many such factors into account (Wald et al. 2010a, b). At the heart of PAGER are the timely and accurate earthquake location and magnitude determinations that the USGS has been producing for decades. PAGER uses these earthquake parameters to calculate estimates of ground shaking by using the methodology and software developed for ShakeMap (<http://earthquake.usgs.gov/shakemap/>). The ShakeMap system produces maps of regional ground shaking using a site-specific ground-motion amplification map, seismic wave attenuation equations, and



Earthquake Shaking **Orange Alert**



USAID
FROM THE AMERICAN PEOPLE

M 7.8, NEAR THE COAST OF ECUADOR

Origin Time: Sat 2016-04-16 23:58:37 UTC (18:58:37 local)

Location: 0.37°N 79.94°W Depth: 19 km

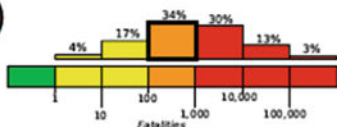
FOR TSUNAMI INFORMATION, SEE: tsunami.gov

Created: 3 hours, 26 minutes after earthquake

PAGER

Version 5

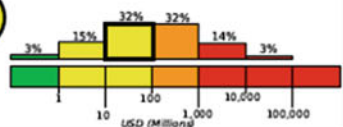
Estimated Fatalities



Orange alert level for shaking-related fatalities. Significant casualties are likely and the disaster is potentially widespread. Past events with this alert level have required a regional or national level response.

Yellow alert level for economic losses. Some damage is possible. Estimated economic losses are less than 1% of GDP of Ecuador.

Estimated Economic Losses

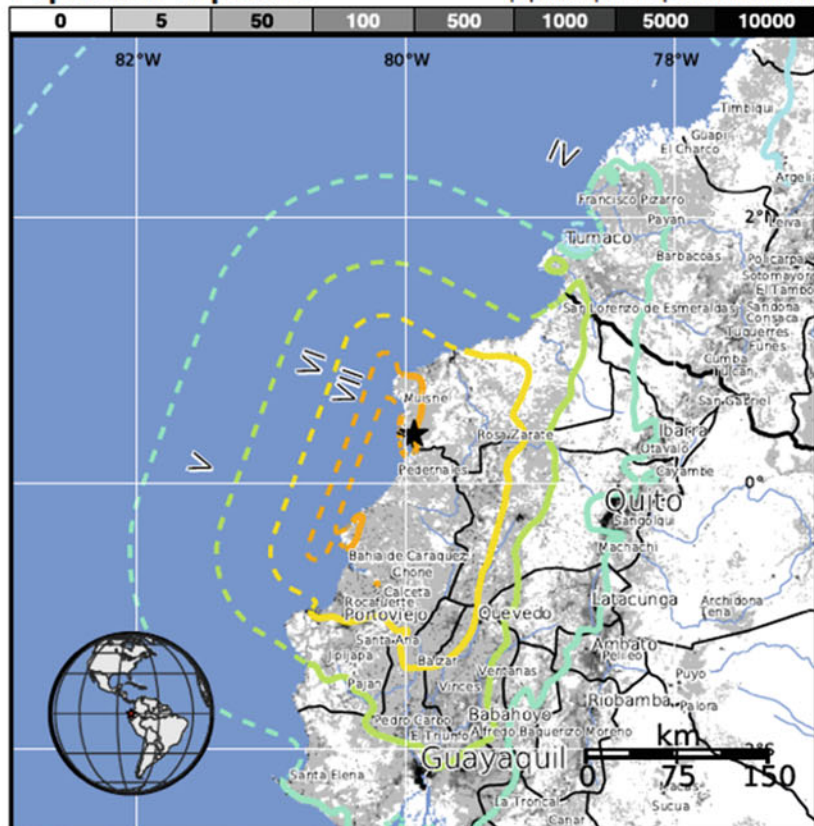


Estimated Population Exposed to Earthquake Shaking

ESTIMATED POPULATION EXPOSURE (k = x1000)	--*	115k*	5,238k*	5,971k	2,085k	1,767k	96k	0	0
ESTIMATED MODIFIED MERCALLI INTENSITY	I	II-III	IV	V	VI	VII	VIII	IX	X+
PERCEIVED SHAKING	Not felt	Weak	Light	Moderate	Strong	Very Strong	Severe	Violent	Extreme
POTENTIAL DAMAGE	Resistant Structures	none	none	none	V. Light	Light	Moderate	Moderate/Heavy	Heavy
	Vulnerable Structures	none	none	none	Light	Moderate	Moderate/Heavy	Heavy	V. Heavy

*Estimated exposure only includes population within the map area.

Population Exposure



Structures:

Overall, the population in this region resides in structures that are vulnerable to earthquake shaking, though some resistant structures exist. The predominant vulnerable building types are unreinforced brick masonry and mud wall construction.

Historical Earthquakes (with MMI levels):

Date (UTC)	Dist. (km)	Mag.	Max MMI(#)	Shaking Deaths
1979-03-01	31	5.7	VIII(4k)	0
1996-03-28	212	5.9	VIII(14k)	27
1987-03-06	245	7.1	IX(2k)	1k

Recent earthquakes in this area have caused secondary hazards such as tsunamis and landslides that might have contributed to losses.

Selected City Exposure

from GeoNames.org

MMI City	Population
VIII Muisne	13k
VIII Tosagua	15k
VIII Pedernales	6k
VII Bahia de Caraquez	37k
VII Rosa Zarate	42k
VII Chone	45k
VII Portoviejo	170k
V Guayaquil	1,952k
IV Pasto	382k
IV Quito	1,400k

bold cities appear on map

(k = x1000)

PAGER content is automatically generated, and only considers losses due to structural damage.

Limitations of input data, shaking estimates, and loss models may add uncertainty.

<http://earthquake.usgs.gov/earthquakes/eventpage/us20005j32>

Event ID: us20005j32

Earthquakes, PAGER, Fig. 1 Example of the PAGER one-page summary product showing the color-coded impact scale indicated alert levels (top), the total population exposed per color-coded intensity level (center, near top), a map of the population density and contoured

intensity level (lower left), summary statements on country-specific structures and historical earthquakes (center right), and selected cities with population and intensity level (lower right)

recorded ground motions as well as reported intensities (Wald et al. 2005). The number of people exposed to each shaking intensity level is then calculated by combining the maps of estimated ground shaking with a comprehensive worldwide population database (Landscan, from Oak Ridge National Lab).

Next, based on the population exposed to each intensity level of shaking, the PAGER system estimates total losses based on country-specific models developed from economic and casualty data collected from past earthquakes (Jaiswal et al. 2009). To calibrate the loss models, the USGS has generated an atlas of approximately 6,000 ShakeMaps for significant global earthquakes that have occurred during the last 40 years (Allen et al. 2009). The calibration of loss methodologies relies on this atlas and on fatality and damage data aggregated or collected by the NEIC. Finally, the alert levels, determined by estimated ranges of fatalities and economic loss, are produced, with the higher of the two levels setting the overall alert level. The overall alert level determines which users are actively notified, and, at the same time, all PAGER information and content is automatically distributed to the web on the USGS Earthquake web pages, as part of the earthquake summary information, for immediate consumption.

The PAGER Earthquake Impact Scale

PAGER employs an earthquake impact scale (EIS, Wald et al. 2010a) that is based on two complementary criteria. One, based on the estimated cost of damage, is most suitable for domestic events; the other, based on estimated ranges of fatalities, is generally more appropriate for global events, particularly in developing countries. Simple thresholds, derived from the systematic analysis of past earthquake impacts and associated response levels, turn out to be quite effective in communicating the predicted impact and response needed after an event, characterized by alerts of green (little or no impact), yellow (regional impact and response), orange (national-scale impact and response), and red (international response). Corresponding fatality thresholds for yellow, orange, and red alert levels are 1, 100, and 1,000, respectively. For damage impact, yellow, orange, and red thresholds are triggered by estimated losses reaching \$1 million, \$100 million, and \$1 billion, respectively. The rationale for a dual approach to earthquake alerting stems from the recognition that relatively high fatalities, injuries, and homelessness dominate in countries where local building practices typically lend themselves to high collapse and casualty rates, and it is these impacts that lend to prioritization for international response. In contrast, it is often financial and overall societal impacts that trigger the level of response in regions or countries where prevalent earthquake-resistant construction practices greatly reduce building collapse and resulting fatalities.

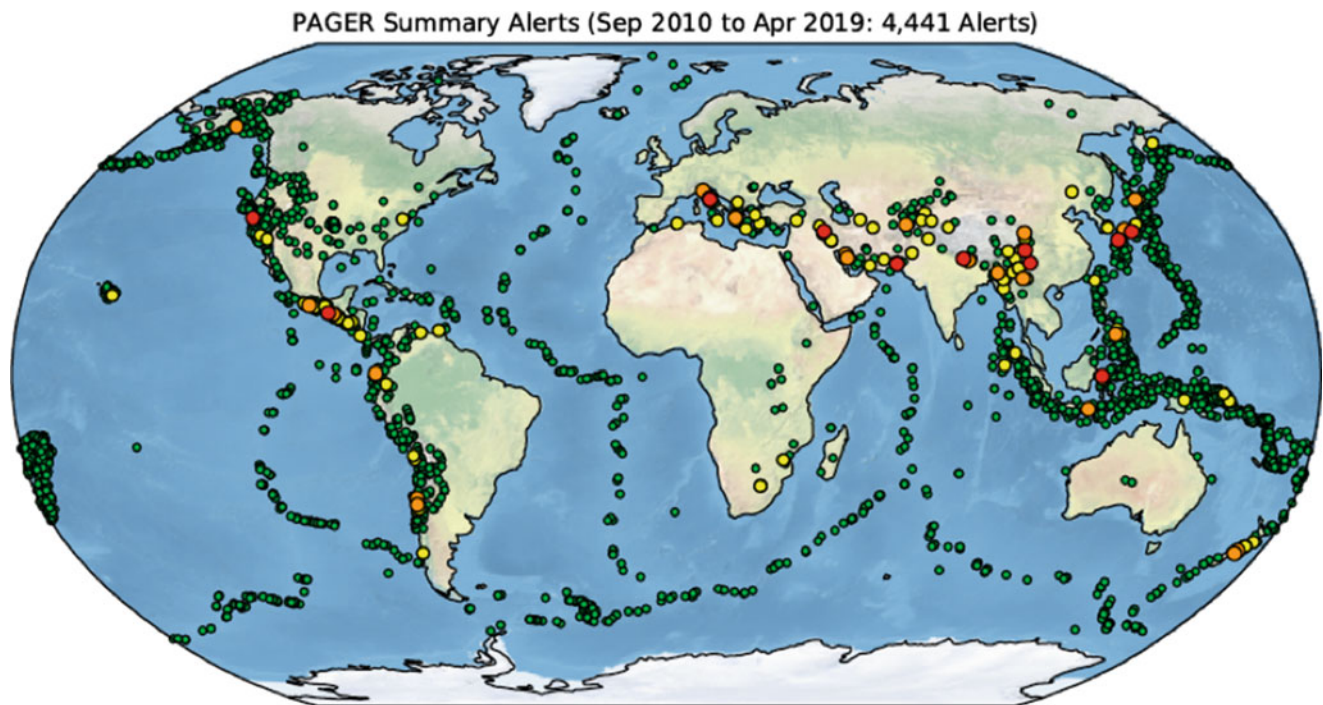
Since PAGER calculations are available well in advance of ground-truth observations or news accounts, PAGER information can play a primary alerting role for domestic as well as international earthquake disasters. An example of the PAGER summary product, or *onePAGER*, for a destructive magnitude 7.8 earthquake in 2016 near the coast of Ecuador that killed several hundred people is shown in Fig. 1. This earthquake reached an orange alert level based on projected fatalities and a yellow alert level for economic losses. The alert level is based on the median loss estimate; the uncertainty in the alert level can be gauged by the histogram showing the percent likelihood that adjacent alert levels (or loss/fatality ranges) occur. Accompanying text clarifies the nature of the alert based on experience from past earthquakes and provides context on the total economic losses in terms of the percent of the gross domestic product (GDP) of the country affected.

Though PAGER uses simple and intuitive color-coded alerting criteria, it preserves the necessary uncertainty measures by which one can gauge the likelihood of adjacent alert levels. Utilizing the earthquake impact scale, PAGER's rapid loss estimates can adequately recommend alert levels and suggest appropriate response protocols, despite their uncertainties; demanding or awaiting observations or loss estimates with a high level of accuracy may increase the losses. Fig. 2 shows a global map of PAGER alerts over the past 10 years, since its inception in 2010. Over this time period, there have been 5,136 events with PAGER reports; of these, the final summary alert levels were 4,903 green; 175 yellow; 40 orange; and 18 red. This frequency equates to approximately 20 yellow, 3–4 orange, and 1–2 red alerts per year.

Ongoing PAGER Developments

USGS is improving the PAGER system to include losses from recent earthquakes, more comprehensive loss-estimate methodologies that take into account detailed building inventories that account for sub-country-level regional variations, more complete population demographics (including time of day occupancy shifts), and better tools to compute building damage (e.g., Jaiswal et al. 2010). Such data sets are very difficult and time consuming to acquire and are effectively unavailable in many areas of the globe. In particular, more detailed within-country building inventories are key for describing the dominant vulnerable structures that contribute to casualties in the region. Knowledge of the main collapse “culprits” is vital for response, public safety, recovery, and long-term mitigation. The USGS PAGER team is working with the Global Earthquake Model (GEM) Foundation to refine and calibrate both the PAGER and GEM national loss models for improved real-time loss estimates.

Related USGS developments under the auspices of the PAGER project include rapid determination of fault



Earthquakes, PAGER, Fig. 2 Summary map showing PAGER alerts from its initial deployment in September 2010 through March 2019. Alerts shown are the maximum of the economic or fatality alert for each event

geometry, size, and rupture characteristics, refined global estimates of seismic soil-amplification patterns, ShakeMap enhancements, and ground-motion and loss-uncertainty analyses. Ultimately, the PAGER system will also directly incorporate impact estimates due to earthquake-induced landslide and liquefaction, integrating ground failure probability products that are maturing to the point of near-real-time operational use (Allstadt et al. 2016). Other strategies for improving rapid loss estimates include ongoing efforts to condition loss estimates on ground-truth loss data as they become available and the use of satellite imagery to refine damage locations and locations of ground failure and to improve upon modeled loss estimates themselves.

Summary

The Prompt Assessment of Global Earthquakes for Response (PAGER) system provides rapid (less than 20 min) post-earthquake information concerning the impact of significant earthquakes within approximately 20 min of any magnitude 5.5 or larger event around the world. PAGER has been operating automatically for nearly a decade, starting in 2010, and produces approximately 20 yellow, 3–4 orange, and 1–2 red alerts per year. These color-coded alerts, tied to the higher of the likely fatalities or economic losses (see Fig. 1), are widely used for response decision-making.

PAGER rapidly assesses earthquake impacts by comparing populations exposed to estimates of shaking intensity and models of economic and fatality losses based on past earthquakes in each country or region of the world. The main PAGER output – a one-page summary document – is sent to thousands of emergency responders, government and aid agencies, and the media to the scope of the potential disaster. In addition, PAGER loss estimates and information products are also fully integrated into the USGS Earthquake Information “event page” automatically generated for all earthquakes detected by the USGS. This impact summary information is available on the USGS earthquake web pages and as a printable, one-page report with accompanying descriptions such as those shown in Fig. 1.

Cross-References

- [Earthquakes, Intensity](#)
- [Earthquakes, ShakeMap](#)

Bibliography

- Allen TI, Wald DJ, Earle PS, Marano KD, Hotovec AJ, Lin K, Hearne MG (2009) An atlas of ShakeMaps and population exposure catalog for earthquake loss modeling. *Bull Earthq Eng* 7:701–718
- Allstadt KE, Thompson EM, Wald DJ, Hamburger MW, Godt JW, Knudsen KL, Jibson RW, Jessee MA, Zhu J, Hearne M, Baise LG,

- Tanyas H, Marano KD (2016) USGS approach to real-time estimation of earthquake-triggered ground failure – results of 2015 workshop. U.S. Geological Survey Open-File Report 2016–1044. <https://doi.org/10.3133/ofr20161044>
- Jaiswal KS, Wald DJ, Hearne M (2009) Estimating casualties for large earthquakes worldwide using an empirical approach. U.S Geological Survey Open-File Report 2009–1136, 78pp
- Jaiswal KS, Wald DJ, D'Ayala D (2010) Developing empirical collapse fragility functions for global building types. *Earthquake Spectra* 27(3):775–795
- Wald DJ, Worden BC, Lin K, Pankow K (2005) ShakeMap manual: technical manual, user's guide, and software guide. U.S. Geological Survey, Techniques and Methods 12-A1, 132pp
- Wald DJ, Jaiswal KS, Marano KD, Bausch D (2010a) An earthquake impact scale. *Nat Haz Rev* 12:125–139
- Wald DJ, Jaiswal KS, Marano KD, Bausch DB, Hearne MG (2010b) PAGER – rapid assessment of an earthquake's impact. U.S. Geological Survey Fact Sheet No. 2010–3036

Earthquakes, ShakeCast

Kuo-wan Lin, David J. Wald and Daniel Slosky
U.S. Geological Survey, Denver Federal Center, Lakewood,
CO, USA

Definition

ShakeCast[®] – short for *ShakeMap Broadcast* – is a fully automated software system for delivering specific ShakeMap products to critical users and for triggering established post-earthquake response protocols. ShakeCast is a freely available, postearthquake situational awareness software application that automatically retrieves earthquake shaking data from ShakeMap to compare ground shaking intensity measures against users' facilities (Lin and Wald 2008; Lin et al. 2018). ShakeCast then generates potential damage assessment and inspection priority notifications, maps, and web-based products for critical users, emergency managers, and those on a need-to-know basis.

Background

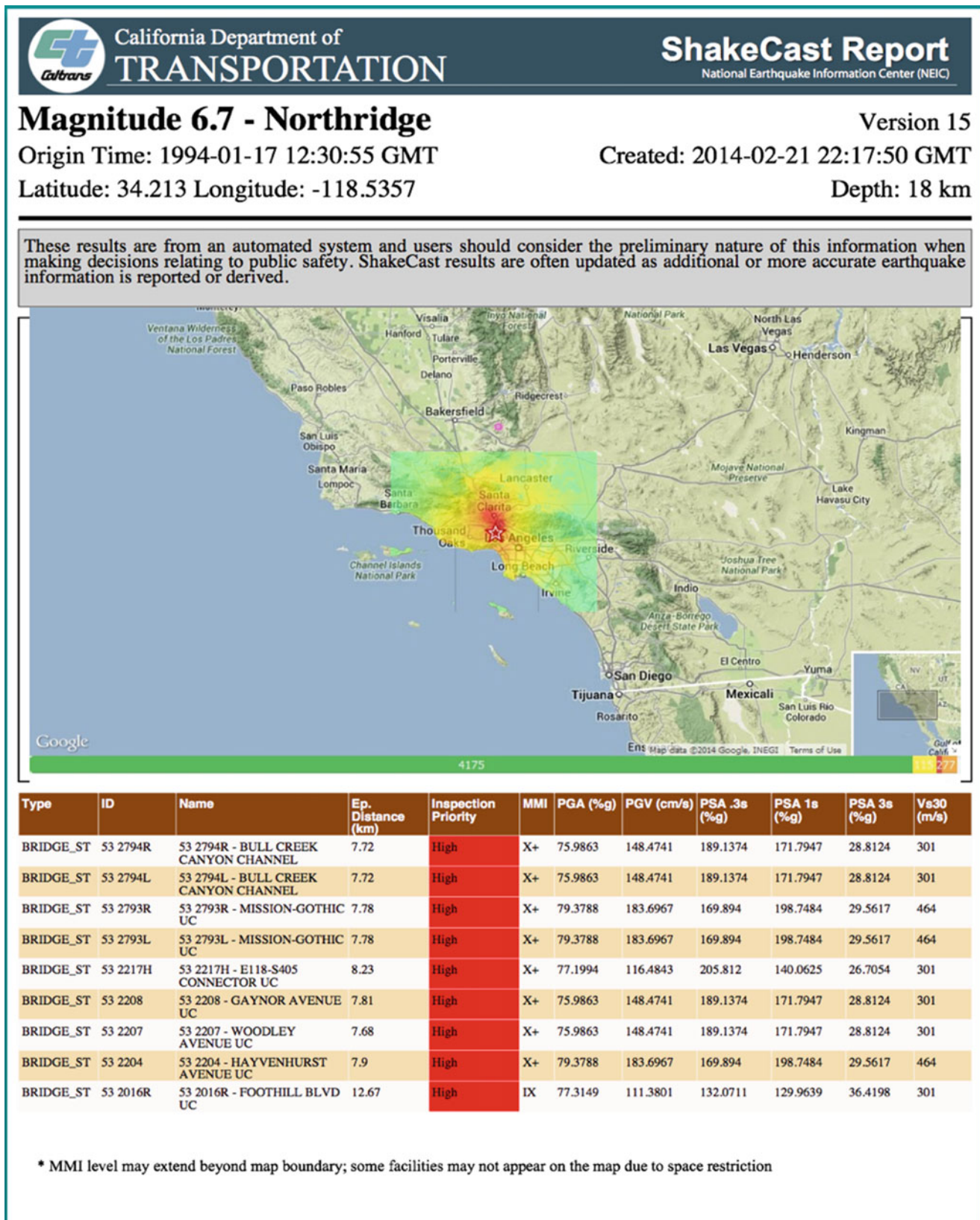
When a potentially damaging earthquake occurs, emergency actions are needed to ensure safety, restore system functionality, and minimize losses. Situational awareness in the immediate aftermath of a disastrous earthquake is of fundamental importance for an effective societal response. While overall disaster management is critical, a successful, organized response is dependent on the collective efforts of the community at large. Businesses, utilities and critical lifeline managers, emergency responders, and a wide variety of other parties have an urgent need for information so they can

make informed decisions and take quick actions. ShakeCast allows such groups to control and optimize the earthquake information they receive. With ShakeCast, they can automatically determine the shaking value at *their* facilities, set thresholds for notification of damage states for each facility, and automatically notify (via cell phone or email) specified operators, inspectors, etc., within their organizations who are responsible for those particular facilities so they can prioritize response.

ShakeMap is a well-established tool for portraying the extent of potentially damaging shaking following an earthquake (Wald et al. 1999; Worden and Wald 2016). ShakeMaps are automatically generated for both small and large earthquakes and can be found online at <http://earthquake.usgs.gov/shakemap>. The ShakeMap system was developed and is used for emergency response, loss estimation, and public information. However, for an informed response to a serious earthquake, critical users must go beyond just ShakeMap to understand the likely extent and severity of impact on the facilities for which they are responsible. In order to facilitate the use of ShakeMap for critical preearthquake planning and postearthquake response, USGS has developed ShakeCast.

ShakeCast is deployed by public and private emergency responders, lifeline operators, and facility engineers to automatically receive and process ShakeMap products for situational awareness, inspection priority, or damage assessment. For example, the California Department of Transportation (Caltrans) utilizes ShakeCast to identify bridges and buildings that require damage inspections following an earthquake, with more than 24,820 state and local bridges and facilities cataloged in their ShakeCast system. There are 568 groups and individuals within Caltrans who receive emails within 10 min of an earthquake of magnitude 4.0 and greater occurring in California. Caltrans can then respond quickly to potentially damaging incidents and focus inspection efforts where needed, better protecting the transportation network. Figure 1 is an example of the ShakeCast product that is automatically sent to users as a quick summary of the situation, including which facilities need to be inspected first.

Caltrans' ShakeCast system routinely sends notifications to Caltrans responders for all earthquakes greater than M4.0 that have resulted in significant ground shaking within California; over 300 events were evaluated in the past decade. Though widely felt in the region and reported through the media, the vast majority of these earthquakes did not produce ground shaking strong enough to cause damage to Caltrans structures. A few events, however, have generated significant shaking and triggered ShakeCast "Potential Impact" notifications: the 2008 M5.4 Chino Hills, 2010 M7.2 Calexico, and 2014 M6.0 Napa earthquakes. For each, ShakeCast generated messages to responders and damage to bridges was observed.



Earthquakes, ShakeCast, Fig. 1 Example of the ShakeCast report for a rerun of the 1994 magnitude 6.7 Northridge, California earthquake created for testing and calibration purposes for the California Department of Transportation (Caltrans). This is a standard ShakeCast product,

which gives a visual overview of the facility locations and ShakeMap shaking levels above a summary of the facility (bridge and overpass) locations, inspection priorities, and shaking levels

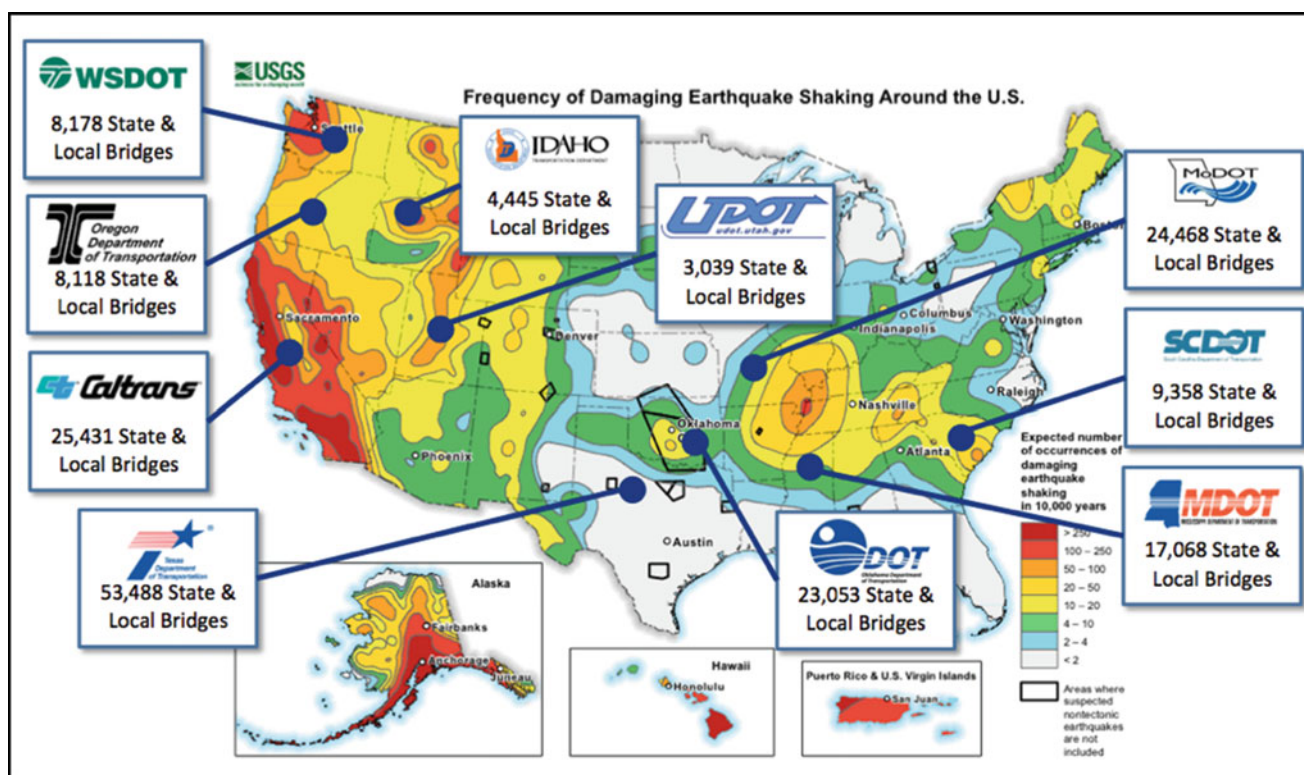
ShakeCast System Overview

The goal of the ShakeCast system is to provide simplified, near-real-time damage and inspection metrics in order to facilitate users' inspection priorities and response protocols (green, yellow, orange, and red priority ratings). Building or bridge fragilities can be evaluated for potential damage level and likelihood within ShakeCast using individual ShakeMap intensity measures (specifically, Modified Mercalli Intensity (MMI), peak horizontal ground acceleration (PGA), peak horizontal ground velocity (PGV), and 0.3, 1.0, and 3.0 s spectral accelerations) for each facility (Worden and Wald 2016).

The ShakeCast application is built upon open-source (Python) code, providing standard, freely available software for all users. All ShakeMap and ShakeCast files and products are nonproprietary to simplify interfacing with existing users' response tools and to encourage user-made enhancement to the software. ShakeCast software development utilizes the GitHub framework for both code distribution and documentation. Any GitHub user can contribute to the development of the ShakeCast application or submit feature requests in the form of "issues." Most users of ShakeCast operate the software in a cloud computing environment. During the development of ShakeCast, a conscious decision was made to provide

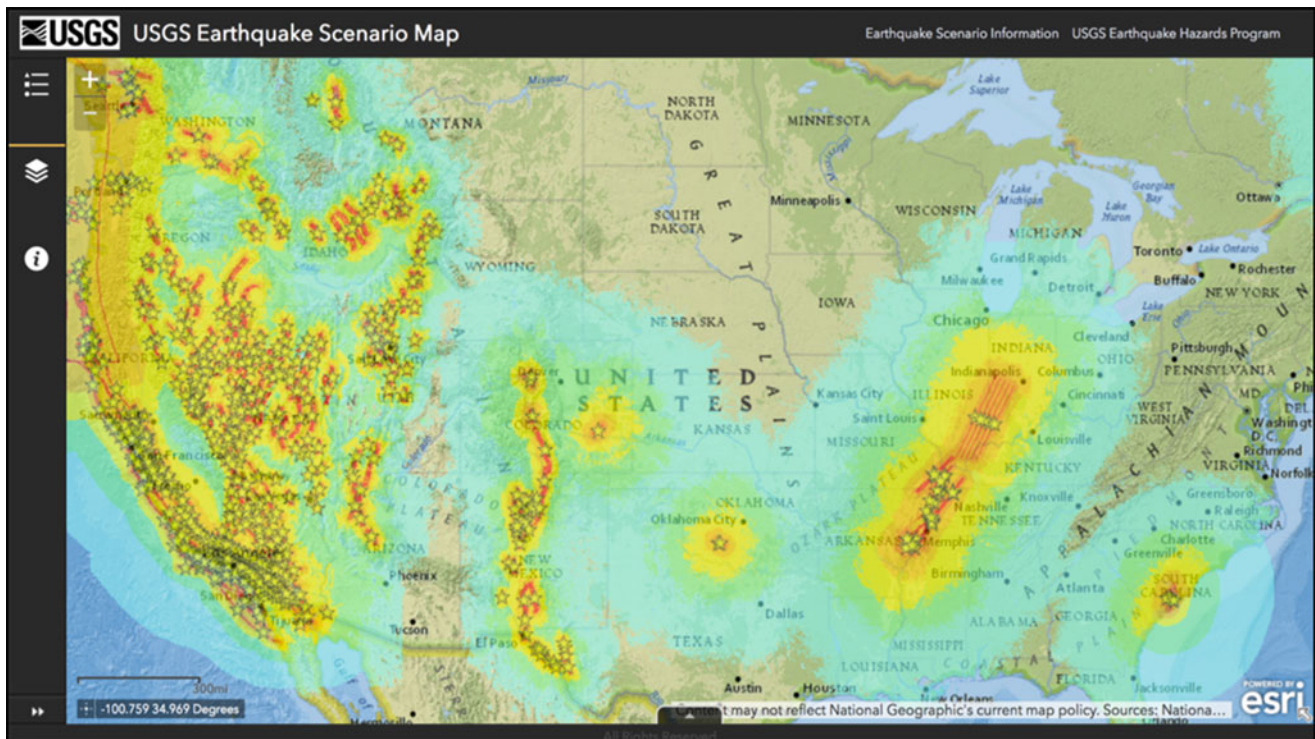
the software to critical users to run their own response operations in-house, rather than providing a centralized service (like ShakeMap), for a number of reasons. First, a user's inventory may be proprietary and thus analyses must be made in-house. Second, for a large and changing inventory, the user is more likely to keep an up-to-date database, and it would be unproductive for USGS to take on such a role. Finally, users can customize damage functions, modify recipient contact information, and customize who receives notifications, what they receive, and under what conditions, at their convenience.

Initial ShakeCast setup involves (1) populating a database of facility locations and types, (2) assigning fragilities using specific ShakeMap parameters (e.g., intensity, peak or spectral acceleration) and the corresponding likely "green," "yellow," and "red" damage states (such as "damage unlikely," "damage possible," and "damage likely" thresholds), (3) specifying who receives notifications by listing addresses of facility managers and response personnel (email, cell phone), and (4) selecting under which circumstances the alerts are sent (e.g., "damage possible" at specific facilities). Users can also customize the content of the summary report that is delivered internally, for example, adding organizational-specific links and images, or listing facilities based on their likely damage state.



Earthquakes, ShakeCast, Fig. 2 Distribution of shaking hazards from the USGS and State transportation department partners in the Transportation Pooled Fund. Ten state Departments of Transportation

(DOT) – CA, ID, MO, MS, OK, OR, SC, TX, UT, and WA – have partnered with the USGS ShakeCast team



Earthquakes, ShakeCast, Fig. 3 USGS ShakeMap earthquake scenario portal for events in the continental USA. Any such scenario can be imported into ShakeCast to evaluate effects on facilities for training exercises or evaluating capacity to respond

Whereas many ShakeCast users host their own ShakeCast software and operations (a ShakeCast “instance”), many life-line, government agencies, and other large groups – who could benefit from operating ShakeCast – have difficulty running and maintaining a 24×7 software application in-house. Thus, USGS converged on ShakeCast software variations with two main options, in-house and remote (USGS operated) ShakeCast software instances. *ShakeCast Remote* has become an option for hosting multiple instances of the software using a single server to take advantage of the dynamic scaling of server provision in a modern virtualization environment. For example, a USGS-hosted ShakeCast server for the Transportation Pooled Fund, a consortium of a dozen State Department of Transportation (Fig. 2; Turner et al. 2018), is an example of such a deployment, utilizing one to a dozen instances. Such ShakeCast instances are running in their own confined domain with the full features of ShakeCast, operated remotely by the USGS, sending alerts and notifications to the appropriate recipients dependent on the region of any earthquake.

Outlook

The ShakeCast development team coordinates with the ShakeMap team to bring upgrades and features to end users.

Several upgrades to the USGS ShakeMap system are noteworthy, particularly as they pertain to ShakeCast. Improvements have been made to (1) event-specific metadata, product archiving, and technical documentation; (2) additional gridded parameters (including interpolated rock-motion shaking estimates); and (3) ground motion interpolation, including spatial variability characterization and directivity functions (Worden and Wald 2016).

In addition to real-time notification, ShakeCast can also generate and deliver scenario earthquakes for facility response plans and earthquake drills or exercises. This includes routine testing of the system, earthquake scenario exercises, and evaluating performance and response under potential earthquake conditions. A new online catalog of more than 900 ShakeMap earthquake scenarios includes the deterministic ruptures from the 2014 version of the USGS national seismic hazard model. The systematic collections of ShakeMaps regarding event selection, ground motion prediction equations (GMPEs), and site response provide more opportunities for ShakeCast testing and evaluation (Fig. 3).

Cross-References

- [Earthquakes, Intensity](#)
- [Earthquakes, ShakeMap](#)

Bibliography

- Lin K, Wald DJ (2008) ShakeCast manual. U.S. Geological Survey Open File Report 2008-1158, p 90
- Lin K, Wald DJ, Kircher CA, Slosky D, Jaiswal KS, Luco N (2018) Advancements to and user applications of the USGS ShakeCast system. In: Proceedings of eleventh national conference on earthquake engineering, Earthquake Engineering Research Institute, Los Angeles
- Turner L, Wald DJ, Lin K, Chiou B, Slosky D (2018) State transportation agencies partner to deploy and enhance ShakeCast. In: Proceedings of eleventh world conference on earthquake engineering, Earthquake Engineering Research Institute, Los Angeles
- Wald DJ, Quitoriano V, Heaton TH, Kanamori H, Scrivner CW, Worden CB (1999) TriNet "ShakeMaps": rapid generation of peak ground motion and intensity maps for earthquakes in southern California. *Earthquake Spectra* 15(3):537–556
- Worden CB, Wald DJ (2016) ShakeMap manual online: technical manual, user's guide, and software guide. U.S. Geological Survey. <https://doi.org/10.5066/F7D21VPQ>, usgs.github.io/shakemap

Earthquakes, ShakeMap

David J. Wald, C. Bruce Worden, Eric M. Thompson and Michael Hearne
U.S. Geological Survey, Denver Federal Center, Lakewood, CO, USA

Definition and Introduction

ShakeMap[®] is an open-source software program employed to automatically produce a suite of maps and products that portray the geographical extent and severity of potentially damaging shaking following an earthquake. ShakeMap's primary purpose is to provide post-earthquake situational awareness for emergency management and response as well as damage and loss estimation. The availability of ShakeMaps immediately after a significant earthquake is critical for the identification of areas likely to be most damaged. Principal users include first responders, utility companies, response and aid agencies, scientists and engineers, and the media. Maps are made publicly available via the Internet within several minutes of an earthquake's occurrence. ShakeMap is widely deployed in seismically active, well-instrumented portions of the USA and internationally in numerous countries including Italy, Iceland, Greece, Costa Rica, and Switzerland, among others, and the US Geological Survey's (USGS) global system provides approximate maps for all earthquakes of magnitude 5.5 and larger worldwide.

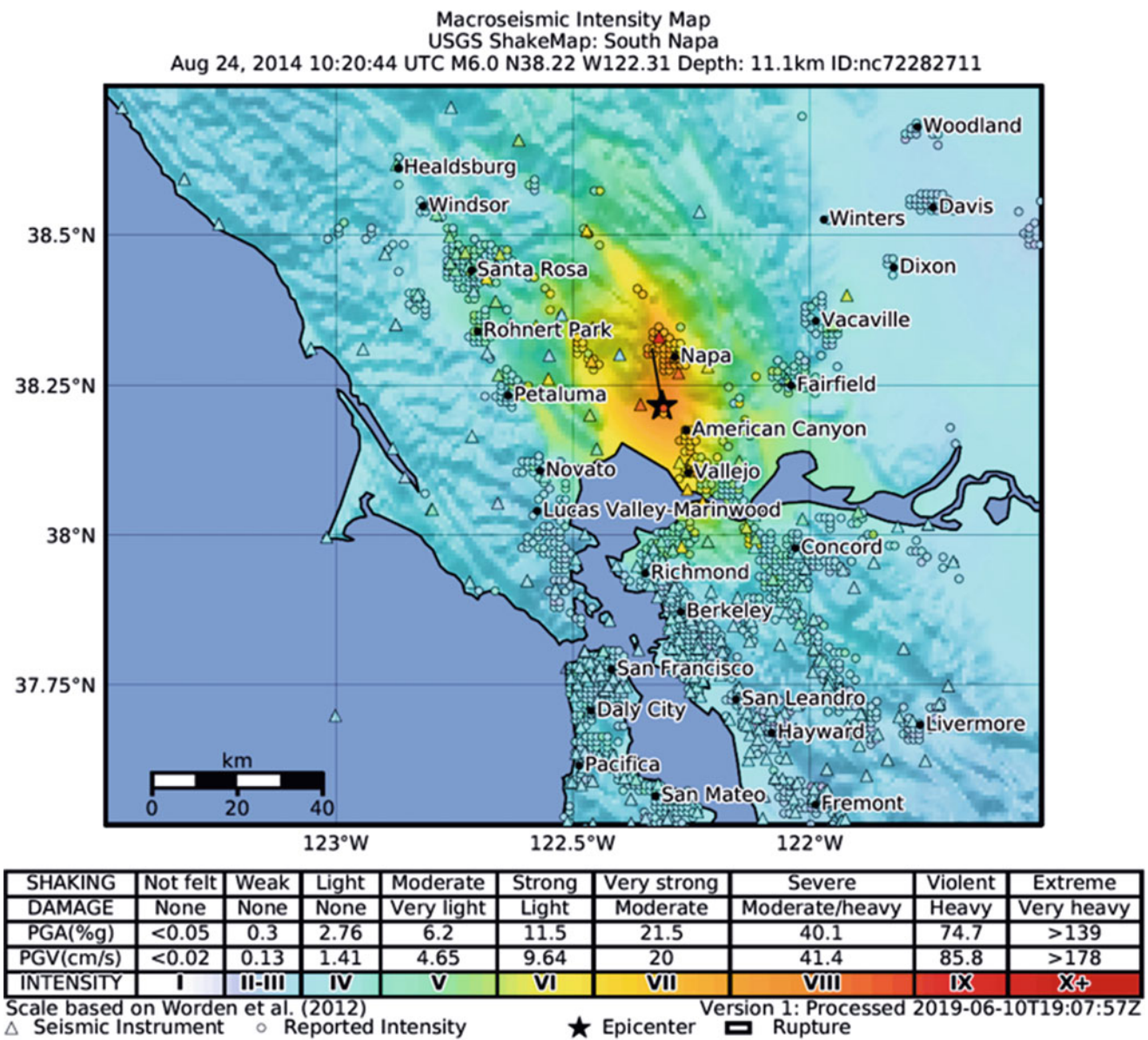
Making a ShakeMap entails deriving a detailed description of the shaking over a large region by integrating a variable distribution of seismic stations and macroseismic intensity data. This requires intelligently guided interpolation of measured and estimated ground motions unless the measured

recordings are extremely abundant (as in Japan, for instance). In the ShakeMap algorithms, empirically based ground-motion estimates, modified by simple geologically based, frequency- and amplitude-dependent site-correction factors, provide useful first-order shaking constraints to guide interpolation in areas that are not well instrumented. Thus, essential elements of a ShakeMap system include (1) near-real-time peak ground motion and/or intensity observations; (2) a ground-motion site-amplification map for the region; (3) the ShakeMap algorithms, which combine and guide interpolation of ground-motion observations with built-in shaking estimates (Worden et al. 2018) to produce the suite of shaking-hazard maps; and (4) mechanisms for automatically transferring ShakeMap products to users, including via interactive web pages and through Internet-based product feeds. For larger events (magnitude greater than 7.0), ShakeMap also uses a description of the fault that ruptured during the event, though this often takes hours to characterize so it is added to later revisions of the maps.

An example ShakeMap for the 2014 magnitude 6.0 South Napa earthquake just north of San Francisco, CA, that employs data from seismic stations from the California Integrated Seismic Network (CISN) is shown in Fig. 1. The ShakeMap shaking pattern is well-constrained at and near seismic stations (triangles in Fig. 1) and macroseismic observations (circles in Fig. 1), and the shaking pattern is less well-informed elsewhere. Station locations thus provide a sense of how well-constrained the shaking is in any region of a ShakeMap. The seismic station and macroseismic observation data in aggregate also help provide an overall fit to the data by informing the modeled shaking of overall trends. Whether showing the signature ShakeMap product – the image of shaking intensity (Fig. 1) – or contours of other shaking parameters like PGA (Fig. 2), the overall goal of ShakeMap is to provide a recognizable depiction of the areas of the strongest shaking for any earthquake.

Historical Context

The most common information available immediately following a damaging earthquake has traditionally been magnitude and hypocenter location. However, the damage pattern is not a simple function of these two parameters alone, and more detailed information is necessary to properly evaluate the situation. ShakeMap fills this role, and it has proven to be a useful, descriptive display for rapidly assessing the scope and extent of shaking and potential damage following an earthquake. Because large earthquakes are generated over a fault surface, not at a single point, the location of the earthquake (the epicenter) tells us only where the earthquake started, not necessarily where the shaking was the greatest. For a large earthquake, damage can occur hundreds of miles from the epicenter (e.g., see Fig. 3). Other factors – such as rupture



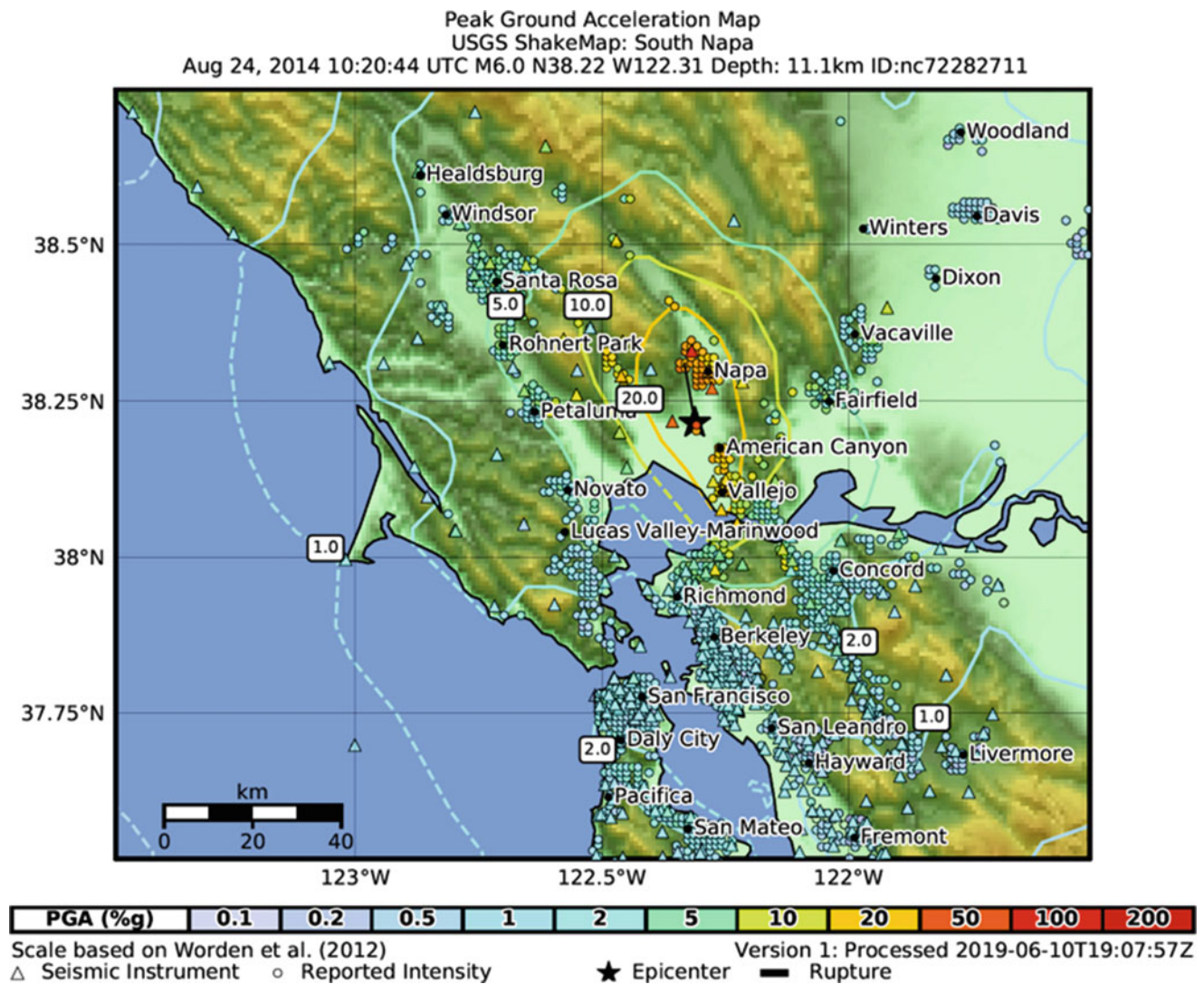
Earthquakes, ShakeMap, Fig. 1 Intensity ShakeMap for the magnitude 6.0 South Napa, California, earthquake of August 24, 2014. Triangles indicate locations of seismic stations, and circles show observed macroseismic observations contributing to the map, each color-coded to

the intensity scale shown below the map. The surface projection of the fault trace is shown with a black line connected to the epicenter (black star)

direction, underground earth structure, and shallow geology – influence the amount of shaking in a particular area. For example, although emergency responders identified many areas of heavy damage soon after both the 1994 Northridge and the 1989 Loma Prieta, California, earthquakes, additional regions of severe damage were only belatedly discovered. A ShakeMap displays the distribution of ground shaking within minutes after an earthquake so that emergency managers can properly deploy resources to those locations with the greatest potential damage.

Part of the strategy in designing rapid-response ground-motion maps was to determine the best format for reliable

presentation of the maps, given the diverse audiences. In an effort to simplify and maximize the flow of information to the public, ShakeMap was designed to produce not only peak-ground acceleration and velocity maps used by engineers but also an instrumentally derived, estimated Modified Mercalli Intensity map. The color-coded instrumental intensity map makes it easier to relate the recorded ground motions to the felt experience and damage distribution (e.g., Fig. 1). ShakeMap was first developed for earthquakes in Southern California as part of the TriNet Project, a joint effort by the USGS, the California Institute of Technology, and the California Geological Survey (Wald et al. 1999). Following the



Earthquakes, ShakeMap, Fig. 2 Peak ground acceleration (PGA) contour ShakeMap for the magnitude 6.0 South Napa, California, earthquake of August 24, 2014. Contours show PGA in percent g. Symbology is the same as in Fig. 1

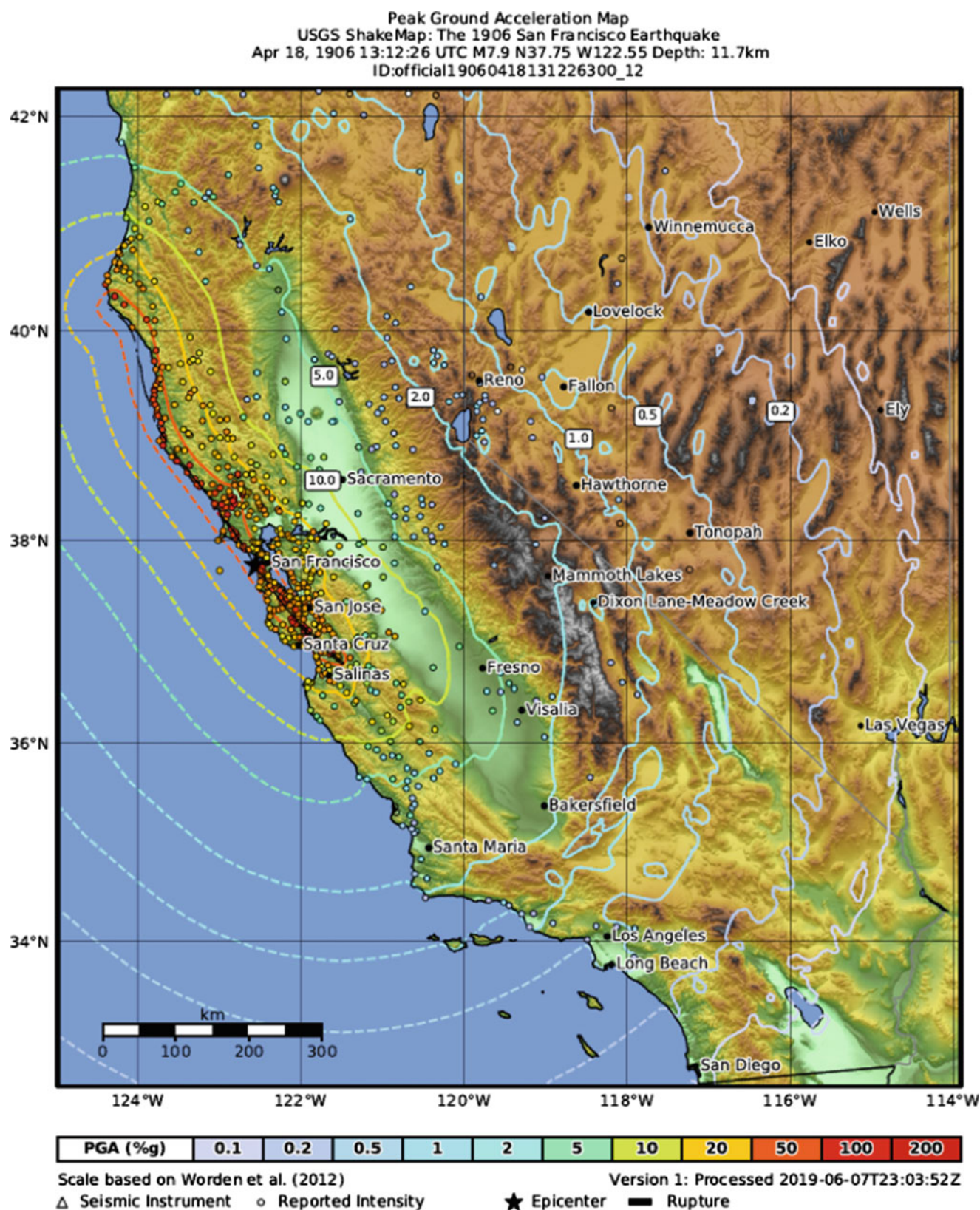
Northridge earthquake, older analog instruments were replaced with a modern seismic network employing near-real-time digital communications. A by-product of that improved network, ShakeMap was further enhanced by advances in telecommunications and computer-processing speed, as well as research that improved empirical ground-motion predictions and the relations among recorded ground motions and shaking intensities. A prototype ShakeMap system was first deployed in Southern California in 1997; refinements to the ShakeMap system continue today (e.g., Worden et al. 2018).

Related Systems and Uses

ShakeMap has facilitated the development and implementation of additional post-earthquake analyses and tools. Separate map layers are generated, each comprised of the spatial distribution of a specific peak ground-motion parameter

(specifically acceleration, velocity, or response-spectral acceleration) as well as a layer for instrumentally derived seismic intensities. In addition to maps, ShakeMap produces a grid containing the value of each parameter, as well as geographic information system (GIS) files and a host of other products intended for a variety of other uses and users. In particular, peak ground motion and spectral acceleration are fundamental parameters for earthquake-engineering and loss-estimation analyses; these gridded parameters of shaking can be combined with building inventory or lifeline databases to rapidly produce maps of estimated damage. All the ShakeMap products and formats can be automatically retrieved via GeoJSON feeds and other web services (Worden et al. 2017).

One commonly used loss-modeling approach, both for post-earthquake response and pre-earthquake mitigation, is the Federal Emergency Management Agency's (FEMA) Hazards US (HAZUS) software. FEMA routinely uses HAZUS to compute earthquake losses based on ShakeMap input within the United



Earthquakes, ShakeMap, Fig. 3 ShakeMap depiction of the shaking pattern experienced during the Great 1906 San Francisco earthquake. Since there were no digital seismic stations at the time, we rely on reported macroseismic observations (circles) that provide not only

constraints of the intensity at their locations but they also allow for estimates of the peak ground motions by employing the intensity to ground-motion conversion equations of et al. (2012). The 1906 rupture on the San Andreas Fault is shown as a black line

States. A secondary system for ShakeMap, ShakeCast[®] (short for ShakeMap Broadcast) is a fully automated system for delivering specific ShakeMap products to critical users and for triggering established post-earthquake response protocols (Lin and Wald 2008). ShakeCast allows utility and transportation agencies, businesses, and other large organizations to control and optimize the earthquake information they receive. With ShakeCast, users can automatically determine the shaking value at their facilities, set thresholds for notification of different degrees of damage for each facility, and then automatically notify (via text or email) specified operators, inspectors, or responders within their organizations who are responsible for those particular facilities so they can prioritize response. For example, via ShakeCast, the California Department of Transportation (Caltrans) can prioritize inspection of thousands of bridges and overpasses statewide, an otherwise potentially overwhelming task.

Another secondary system, also developed primarily as a by-product of the USGS global implementation of ShakeMap, is called PAGER (Prompt Assessment of Global Earthquakes for Response). PAGER combines shaking intensity with exposed populations to rapidly and automatically estimate economic and human impacts worldwide (Wald et al. 2010). Unlike regional ShakeMap systems that rely on dense instrumentation, the global ShakeMap system relies primarily on the predictive aspects of the ShakeMap system; it incorporates site amplifications based on approximate soil conditions derived from topographic slope and utilizes macroseismic intensity data, if available, from near-real-time, Internet-based data collection systems such as the USGS “Did You Feel It?” system (Wald et al. 2011).

ShakeMaps also now provides the shaking constraints needed to rapidly estimate the likelihood and distribution of landslides and liquefaction for earthquakes around the globe (Allstadt et al. 2016). These secondary earthquake-shaking-induced ground failures can contribute substantially to earthquake losses and complicate post-earthquake response, particularly when they occur along transportation corridors.

Though developed primarily for post-earthquake, near-real-time situational awareness, ShakeMap algorithms, which incorporate empirical ground-motion estimation capabilities, are widely used for other purposes. One side-benefit of the built-in ability to estimate ground motions for a specified earthquake is generating hypothetical earthquake ground-motion maps. Such “Scenario ShakeMaps” are useful for response planning and long-term earthquake mitigation, since portraying the overall pattern of shaking for potential future events is useful for earthquake drills and loss assessments. Hundreds of USGS ShakeMap scenarios are available on the USGS ShakeMap web site (<http://earthquake.usgs.gov/shakemap/>). Calibration of earthquake loss models is facilitated by creating ShakeMaps for historical earthquakes, which has been done for significant earthquakes around the globe from

1970 to recent (the ShakeMap Atlas; Allen et al. 2009; Garcia et al. 2012). Some of these historic ShakeMap events have constraints from ground motion, intensity, and fault extent; others are purely predictive. Nonetheless, such historic ShakeMaps can be used for loss calibration or for better scientific understanding of the nature of past earthquakes.

For example, though the magnitude 7.8 of 1906 San Francisco earthquake occurred in the pre-digital instrumentation era, the widespread shaking experienced and the damage observed were well-documented. Macroseismic intensities were assigned both at the time and more recently by revisiting historic accounts in newspapers and other archives of that period (Boatwright and Bundock 2008), providing ample data for producing an informative ShakeMap (Fig. 3).

Summary

Over the past two decades, rapid earthquake information has evolved from providing a magnitude and location to depicting the variations of the intensity of shaking over the region potentially affected by each earthquake. The standard approach to depicting shaking is now with the ShakeMap system, which allows the best estimates of shaking intensity and other peak ground-motion parameters (including peak acceleration, peak velocity, and spectral acceleration at various periods).

The shaking estimates in ShakeMap are computed by a logical, weighted combination of shaking estimates, recorded ground motions, and reported or assigned macroseismic intensities over the area of interest. Often the estimated motions are strongly influenced by the dimensions of the fault that rupture, so rupture finiteness is properly taken into account as soon as that information becomes available, typically a few hours or so after a major earthquake.

ShakeMap is used for a variety of purposes, primarily for post-earthquake situational awareness. As such, many ShakeMap formats are provided via the USGS Earthquake Program web pages and data feeds that include geospatial grids, station files, faults, GIS, KML, GeoJSON, and all the metadata needed to follow changes to ShakeMaps since they often evolve as more data become available. However, historic and scenario ShakeMaps are widely used for planning and mitigation purposes so that responders can plan ahead.

Cross-References

- [Earthquakes, Did You Feel It?](#)
- [Earthquakes, Intensity](#)
- [Earthquakes, PAGER](#)
- [Earthquakes, ShakeCast](#)
- [Earthquakes, Strong-Ground Motion](#)

Bibliography

- Allen TI, Wald DJ, Earle PS, Marano KD, Hotovec AJ, Lin K, Hearne MG (2009) An Atlas of ShakeMaps and population exposure catalog for earthquake loss modeling. *Bull Earthq Eng* 7:701–718
- Allstadt KE, Thompson EM, Wald DJ, Hamburger MW, Godt JW, Knudsen KL, Jibson RW, Jesse MA, Zhu J, Hearne M (2016) USGS approach to real-time estimation of earthquake triggered ground failure – results of 2015 workshop. U.S. Geological Survey Open-File Report 2016–1044, 13pp
- Boatwright J, Bundock H (2008) Modified Mercalli intensity maps for the 1868 Hayward earthquake plotted in ShakeMap format. U.S. Geological Survey Open-File Report 2008–1121. <https://pubs.usgs.gov/of/2008/1121/>
- Garcia D, Mah RT, Johnson KL, Hearne MG, Marano KD, Lin K-W, Wald DJ (2012) ShakeMap Atlas 2.0: an improved suite of recent historical earthquake ShakeMaps for global hazard analyses and loss model calibration. In: 15th World conference on earthquake engineering, 24–28 September 2012, Lisbon, 10pp
- Lin K, Wald DJ (2008) ShakeCast Manual. U.S. Geological Survey Open File Report 2008–1158, 90pp
- Wald DJ, Quitoriano V, Heaton TH, Kanamori H, Scrivner CW, Worden CB (1999) TriNet “ShakeMaps”: rapid generation of peak ground motion and intensity maps for earthquakes in southern California. *Earthquake Spectra* 15(3):537–556
- Wald DJ, Jaiswal KS, Marano KD, Bausch DB, Hearne MG (2010) PAGER – rapid assessment of an earthquake’s impact. U.S. Geological Survey Fact Sheet No. 2010–3036
- Wald DJ, Quitoriano V, Worden CB, Hopper M, Dewey JW (2011) USGS “Did You Feel It?” Internet-based macroseismic intensity maps. *Ann Geophys* 39:688–709
- Worden CB, Thompson EM, Hearne MG, Wald DJ (2017) ShakeMap v4 manual: technical manual user’s guide, and software guide. <http://usgs.github.io/shakemap>
- Worden CB, Thompson EM, Baker JW, Bradley BA, Luco N, Wald DJ (2018) Spatial and spectral interpolation of ground-motion intensity measure observations. *Bull Seismol Soc Am* 108(2):866–875. <https://doi.org/10.1785/0120170201>

Earthquakes, Strong-Ground Motion

Giuliano F. Panza^{1,2}, Cristina La Mura¹, Fabio Romanelli¹ and Franco Vaccari¹

¹Department of Geosciences, University of Trieste, Trieste, Italy

²Earth System Physics Section/Sand Group, The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy

Definition

Strong-ground motion An earthquake-induced ground motion capable of damaging man-made environment.

Introduction

Earthquake ground motion is a natural phenomenon associated with a sudden release of energy due to a fault rupture. Strong-motion seismology deals with seismic waves radiated by large earthquakes. More precisely, it is concerned with measurement, interpretation and prediction of strong shaking generated by damaging earthquakes.

Because of the engineering need for estimates of future strong-ground motion, the scope of strong-motion seismology encompasses seismic source theory and all other aspects of seismology that can provide insights helpful in making the best possible earthquake ground-shaking scenarios given the limited observational data.

Accumulated data have been providing very important knowledge about rupture processes of earthquakes, propagation paths and site effects on ground motion, and the relation between ground motion, damage, and other earthquake related phenomena (e.g., landslides earthquake induced).

The most important goals of strong-motion seismology are the improvement of the physical understanding of the processes controlling strong-ground motion and the reliable assessment of seismic hazard.

To assess the hazard for engineering design applications and due to the limited availability of strong-ground motion records, it has become increasingly common to compute broadband synthetic seismograms that allow us to perform realistic waveform modeling for different seismotectonic environments. The modeling has to take simultaneously into account the properties (e.g., dimensions, directivity, and near-field effects) of the radiating source, lateral heterogeneities along the path, and local site effects.

The joint use of reliable synthetic signals and observations can be fruitfully used for design purposes. In fact, even if recently strong-motion records in near-fault, soft soils, or basin conditions have been obtained, their number is still very limited to be statistically significant for seismic engineering applications.

The lack of a representative set of observations is due to the low frequency of large earthquakes and to the difficulty of providing a proper instrumental coverage of all the areas prone to strong shaking. Despite of the growing strong-motion registration networks, the existing database is still not complete with respect to the possible scenario earthquakes, and the lack of completeness will last for many years if not centuries. The variability of ground motion, due to different causes, e.g., spatial variability, source parameter variability, azimuthal variability (Strasser and Bommer 2009), show how the current available strong-motion data still represent only a small sample of all physically possible ground motion. The mathematical modeling, with different degrees of complexity, based on probabilistic concepts cannot

fill in the gap due to the lack of knowledge about the physical process behind an earthquake (Panza et al. 2004). Therefore, to resort to broadband synthetic seismogram is a fundamental step: where no records are available, synthetic signals can be used to estimate the ground motion without having to wait for a strong earthquake to occur.

Description of Strong-Ground Motion

A seismic record of ground acceleration, due to a damaging shaking, is usually defined a strong-motion datum. There is not a quantitative rule, but it is generally assumed that damaging earthquakes are those of magnitude 5 and above. The threshold to be exceeded in terms of peaks acceleration for ground motion to be considered strong motion is quite blurred. The strongest recorded earthquake motions have peaks between 1 and 3 g, but these are very rare events. Usually, damages are caused by ground motion, which exceeds the threshold of 0.1 g, but Bolt and Abrahamson (2003) lower this limit to 0.05 g.

Due to the importance of strong motion for earthquake engineering, a number of different parameters have come into use to represent various characteristics of strong motion. Actually, the most appropriate way to describe or specify the characteristics of ground motion for the purpose of seismic design is still an open problem. A large number of parameters have been proposed for measuring the capacity of earthquakes to damage structures. However, to account for the complex characteristics of earthquake-induced strong-ground motion in the engineering analysis and design, the need arose for a more inclusive definition of the existing parameters and for the introduction of new ones. In fact, the adoption of inadequate parameters can lead to the definition of unrealistic design earthquakes and consequently to the unreliable evaluation of the seismic risk.

The parameters fundamentally involved in the evaluation of the level of severity associated with strong motion are, for engineering purposes, the frequency content, the amplitude, and the effective duration.

Beginning about the 1960s, the most used ground-motion parameter for quantification of ground motion was the peak ground acceleration (PGA), as the inertia forces depend directly on acceleration. PGA is simply obtained from unprocessed accelerograms. However, PGA is a poor indicator of damage, since, as observed, time histories with the same PGA could be very different in their frequency content, strong-motion duration, and energy level. In fact, PGA may be associated with high frequency pulses, which do not produce significant damage to the buildings as most of the impulse is absorbed by the inertia of the structure with little deformation. On the other hand, a more moderate acceleration may be associated with a long-duration pulse of low

frequency (acceleration pulse), which gives rise to a significant deformation of the structure. The peak ground velocity (PGV) is another useful parameter for the characterization of ground-motion amplitude. The velocity is less sensitive to the higher-frequency components of the ground motion, so the PGV should be more useful than the PGA in characterizing the damaging potential of ground motion. For design purposes, the peak ground displacement (PGD), generally associated with the lower-frequency components of an earthquake ground motion, has been recognized as useful parameter, mostly for the design of seismically isolated structures.

The response spectrum, firstly proposed by Benioff (1934), defined as the maximum response of a damped harmonic system to input motion (Anderson 2003), is a parameter widely used for engineering purposes.

Another ground-motion parameter, very important for engineers, is the duration of shaking (Villaverde 2009). Duration can impact structures and response of soils: structures that can stand one or two large oscillations may collapse under prolonged motions, and while a second of shaking of sandy soil can be tolerated, repeated oscillations can induce liquefaction (Jennings 2003). But there is not a unique definition of duration, as clearly evidenced in Bommer and Martinez-Pereira (1998), where 30 definitions of strong-motion duration are reported. The available definitions can be classified in three different groups, or as a combination of them (Montejo and Kowalsky 2008): (1) “bracketed” durations (Bolt 1969), in which the duration is defined to be the time interval between the first and the last exceedances of a threshold of acceleration; (2) “significant” durations that define the duration of strong motion as the time interval during which a given percentage of the total energy of the record is released; and (3) frequency-dependent durations, in which the duration of strong-ground motion is analyzed separately in several narrow-frequency bands (Bolt 1969; Trifunac and Westermo 1977). The following two examples of duration measurement show the limits of this parameter. A way to estimate duration is to measure the interval of time within which acceleration exceeds some threshold, usually 0.05 g (bracketed duration, Bolt 1969); another measure is the interval of time in which 90% of the integral of the acceleration-squared is recorded. These two estimations yield opposite results as the distance from the source increases. Peak ground motions decrease, as a rule, with increasing distance from the source (for exceptions, see e.g., Panza and Suhadolc 1989; Fah and Panza 1994), but energy becomes dispersed, so the time that is necessary to wait for the arrival of the 90% of the total energy in the seismogram increases (Anderson 2003).

Several equations, commonly referred to as attenuation relations or ground-motion prediction equations (GMPE), have been proposed over the years from strong-ground motion observations and theoretical studies. These

mathematical expressions give ground-motion intensity measures as a function of earthquake magnitude and distance. They may be determined empirically, by a regression of recorded ground motion, or theoretically, using seismological models to generate synthetic ground motions that account for the source, site, and path effects. Theoretical attenuation relations between various source parameters are used to estimate source parameters for earthquakes that have not yet occurred, or for which parameters of interest are unknown (e.g., Geller 1976; Oth et al. 2007; Somerville et al. 1997). The other approach uses empirical relations between source parameters compiled for a huge number of earthquakes (e.g., Wells and Coppersmith 1994). In the past few years, the availability of additional strong-motion data has allowed the incorporation of additional parameters, such as type of fault, fault orientation; further recent relationships have increasingly become more refined, although, at the same time, more complicated. Attenuation relations are usually grouped into three sets (Panza et al. 2004): shallow crustal earthquakes in active tectonic regions, shallow crustal earthquakes in stable tectonic regions, and subduction zone earthquakes. These relations can differ in the assumed functional form, the data selection criteria, the number and definition of independent variables, and the statistical treatment of the data. A review can be found in Douglas (2003), where the complete procedure that needs to be followed to derive equations for the estimation of ground motion from recorded strong-motion data is outlined as well. As the coefficients of the attenuation relations are determined empirically by means of regression analysis, they are quite sensitive to the used data set, an example of the dependence of attenuation relations on the procedure followed in the data processing is given by Parvez et al. (2001) for the Himalayas. The Izmit, Turkey, earthquake of August 17, 1999 and the Chi Chi, Taiwan, September 20, 2000, did show the inadequacy of the existing database: as explained in Anderson (2007), the GMPEs available at that moment underestimated the peak accelerations for those events. The discrepancies could be due to a wide variety of factors, so these examples suggest that a large amount of data and a deep physical understanding of the processes involved in seismic events are needed to improve the quality of GMPEs (Anderson 2007). A common main source of error can be mathematical shortcomings that arise when attenuation relations are used in geologically complex areas (e.g., Peresan et al. 2010).

As stressed earlier by Aki: “a major goal of strong-motion seismology is to be able to synthesize strong-motion seismograms suitable for use in engineering analysis” (Anderson 2003, p. 942). The use of computational modeling has the aim to simulate the wave generation and propagation process in complex media. The formulation of physical-mathematical models aimed to represent the effective complexity of the phenomenon and their resolution is a

complicated affair. With the exception of a very limited number of simple situations, the physical-mathematical approach does not lead to closed form solutions, i.e., fully analytical, so it is necessary to employ numerical procedures, more or less complex, or approximated theories (Bolt 1999).

Earthquake Source Effects on Strong-Ground Motion: Near-Field Effects

Source effects on far-field seismograms are discussed by Aki and Richards (1980, Chap. 14). Seismic data used in earthquake engineering are occasionally collected in the near-field, but the question where the near-field ends and the far-field begins is gradational and will not be addressed here.

Near-fault ground motions often contain large long-period pulses (Somerville 2002). There are two causes for these pulses: one is constructive interference of the dynamic shaking due to rupture directivity effects, the other is due to the movement of the ground associated with the permanent offset of the ground. For keeping these two causes separated, the terms “directivity pulse” and “fling step” have been introduced (Bolt and Abrahamson 2003).

Large earthquakes are produced by rupture that starts at a point and travels with a velocity that is generally only slightly smaller than the shear-wave velocity in the medium. Therefore, when rupture is towards the station, the propagation causes most of the seismic energy to arrive in a single large pulse occurring at the beginning of the record. Forward-directivity effects occur when the rupture front propagates towards the recording site and the slip direction on the fault is aligned with the site. Backward-directivity effects, occurring when the rupture front propagates away from the recording site, generates opposite effects, i.e., longer-duration motions, lower amplitudes, and longer periods. This phenomenon has been first explained by H. Benioff in 1955, when analyzing the records of the 1952 Kern County, California, earthquake. The conditions for generating forward-rupture directivity effects are readily met in strike-slip faulting, where the rupture propagates horizontally along the strike either unilaterally or bilaterally, and the fault slip direction is oriented horizontally in the direction along the strike of the fault. As described by Bolt and Abrahamson (2003), for strike-slip earthquakes, the rupture directivity is observed on the fault normal component, and the static displacement fling step is observed on the fault parallel component. The conditions required for forward directivity are also met in dip-slip faulting, including both reverse and normal faults. The alignment of both the rupture direction and the slip direction up dip on the fault plane produces rupture directivity effects at sites located around the surface exposure of the fault (or its up dip projection if it does not break the surface). Consequently, it is generally the case that all sites located near the surface

exposure of a dip-slip-fault experience forward-rupture directivity when an earthquake occurs on that fault.

For dip-slip earthquakes, the separation of the effects (rupture directivity and fling step) on the fault components is more complicated: the rupture-directivity effect will be strongest on the fault normal component at a location direct up-dip from the hypocenter, and the fling-step will also be observed on the horizontal component perpendicular to the strike of the fault, i.e., directivity-pulse effects and fling-step effects occur on the same component for dip-slip faults. The horizontal records of the 1966 Parkfield and 1971 San Fernando, California, earthquakes were the first to be discussed (Bolt 1975) as examples of near-fault velocity pulses. The pulses from both directivity and fling-step may attenuate differently. Hisada and Bielak (2003) investigated the combined effects of fling-step and rupture directivity by paying special attention to the contribution of static and dynamic Green's functions. Fling effects are found to be stemmed mainly from the static Green's function. They are dominant in the vicinity of the surface fault and negligible for buried faults since the static Green's function decreases rapidly, as $(1/r^2)$, with the distance from the fault. When the observation point is above the buried fault, the fling step disappears because the slip dislocation of the fault cannot fling the ground, due to the presence of a continuous medium above the fault itself. Directivity effects derive mainly from the dynamic Green's function and attenuate as $(1/r)$ to $(1/\sqrt{r})$, more slowly than the fling. The combination of both the effects of rupture directivity and fling-step results in inclined directions, with respect to the fault plane, of maximum velocities, and displacements.

The destructive potential of near-fault ground motions was manifested in the 1994 Northridge and 1995 Kobe earthquakes. In each of these earthquakes, peak ground velocities as high as 175 cm/s were recorded. The period of the near-fault pulses recorded in both of these earthquakes lies in the range from 1 to 2 s, comparable with the natural periods of structures, such as bridges and buildings, many of which were indeed severely damaged.

Near-fault recordings from recent earthquakes indicate that this pulse is a narrow band pulse whose period increases with magnitude. This magnitude dependence of the pulse period causes the response spectrum to have a peak whose period increases with magnitude, such that the near-fault ground motions from moderate-magnitude earthquakes may exceed those of larger earthquakes at intermediate periods (around 1 s).

Wave-Propagation Effects

Wave propagation in heterogeneous media is addressed in another chapter herein (see ► [“Propagation of Elastic Waves: Fundamentals”](#)), so, here we limit the analysis to site effects due to their impact on strong-ground motion.

Ground motion can be highly amplified, or deamplified, by local effects. The essential aspects of site effects on strong-ground motion are widely addressed in literature (see e.g., Kawase 2003). The pioneering works where site effects are well-recognized date back to 1930s and are due to Sezawa and Ishimoto (Kawase 2003), but a first comment on the effect of near-surface conditions on shaking intensity during the 1906 San Francisco earthquake dates back to 1910: “Experience shows that the damage done by destructive earthquakes is much greater on alluvial soil than on solid rock” (Reid 1910). Almost all recent destructive earthquakes have shown the relevance of site effects in the amplification of the seismic ground motion. The classical example of such a situation is the September 19, 1985 earthquake that, in Mexico City, about 400 km far away from the epicenter, caused a large victims toll (about 8000), destroyed or heavily damaged hundreds of buildings, and caused few billion dollars damage (Fah et al. 1994; Anderson et al. 1986). The presence of lacustrine sediments is the cause of the large ground motion, up to 40 times larger than that of firm ground (Chávez-García and Salazar 2002). The modeling of wave propagation in basins, and more generally in complex 2D and 3D regions, using numerical methods, such as finite difference (FDM), finite element (FEM), spectral element method (SEM), boundary integral equation, and boundary element method (BIEM and BEM) is an active field of research (see e.g., Aagaard et al. 2001; Bielak et al. 1991; Bouchon and Sánchez-Sesma 2007; Graves 1996; Komatitsch and Tromp 1999, 2002a, b; Komatitsch and Vilotte 1998; Komatitsch et al. 2004; Oprea and Zahradnik 1999, 2002; Pitarka 1999; Yoshimura et al. 2003). These methods differ in accuracy with respect to different structural features of the complex heterogeneous models and considerably in the computational efficiency. For comparison of several modeling methods, we refer to the review paper by Takenaka et al. (1998), but the most recent review of the computational and numerical modeling methods can be found in the book edited by Wu and Maupin (2007). Very recently, a new analytical approach has been developed and validated performing simulations in the area of the Kanto basin, Japan (La Mura et al. 2010). It is generally observed that at stations throughout the basin larger amplitudes and longer durations of ground shaking are observed, but the pattern can be very complicated. The amplification patterns depend not only on the site conditions, but also on the characteristics of the incoming wavefield and, therefore, on the source and the wave propagation path. Clear examples are given in Field (2000).

Last but not least, we mention the liquefaction phenomenon, largely observed, for example, during the Niigata earthquake, Japan, on June the 16th, 1964 (Youd 2003; Bardet 2003). Liquefaction of water-saturated sands is believed to occur when the pore pressure, increased by the strong-ground motion, reaches the effective confining pressure. The

likelihood that an earthquake will liquefy a site depends on many site characteristics (e.g., mean grain size, percentage of fine-grained sediment, porosity and level of saturation) and on the regional geology, which influences the amplification and attenuation of strong-motion amplitude. Ground-motion modeling may help in defining a priori the liquefaction potential (e.g., Nunziata et al. 2008).

Strong-Ground Motion Modeling

Due to the destructive potential of earthquakes, a goal is to develop methodologies for modeling strong-ground motions from seismic sources related to active faults and subduction zones with high probability of earthquake occurrence. Confidence in such modeling relies on the theoretical part of strong-motion seismology and on the calibration of the techniques against the empirical data.

One of the most common methods to simulate broadband strong-ground motion is the use of records from small earthquakes occurred in the vicinity of the source area of a large earthquake. This kind of modeling of the strong motion is known as empirical Green's function method, and it has been proposed by Hartzell in 1978 and largely reviewed and applied through the years (see e.g., Irikura 1986; Dan et al. 1990). The empirical Green's function method has been applied in simulating strong-ground motions from a large event, even using observed records of just one or two small events. The records are used as the empirical Green's functions over the fault plane of the large event by correcting only attenuation effects due to difference in distance from each subfault to the site. The advantage of this approach relies in the fact that the seismograms contain all the complexities of wave propagation. On the other hand, the serious disadvantage is that the empirical Green's functions may not be available for the desired source-station path and may originate by an earthquake with a focal mechanism and/or depth different from the desired one.

Very recently, a procedure, called "recipe of strong-motion prediction," has been developed to the aim of obtaining ground-motion time histories from specific earthquakes (Irikura et al. 2004). Such a recipe is based on source characteristics obtained from the waveform inversion using strong-motion data. The idea is to follow a procedure, outlined as a recipe, for characterizing the source model estimating three kinds of source parameters: outer-, inner-, and extra-fault parameters. The outer are to outline the overall properties of the target earthquake, such as entire source area and seismic moment; the inner are parameters characterizing stress heterogeneity inside the fault area; the extra parameters are considered to complete the source model, such as the starting point and propagation pattern of the rupture.

The theoretical approach to the simulation of strong motion, based on the computation of synthetic seismograms,

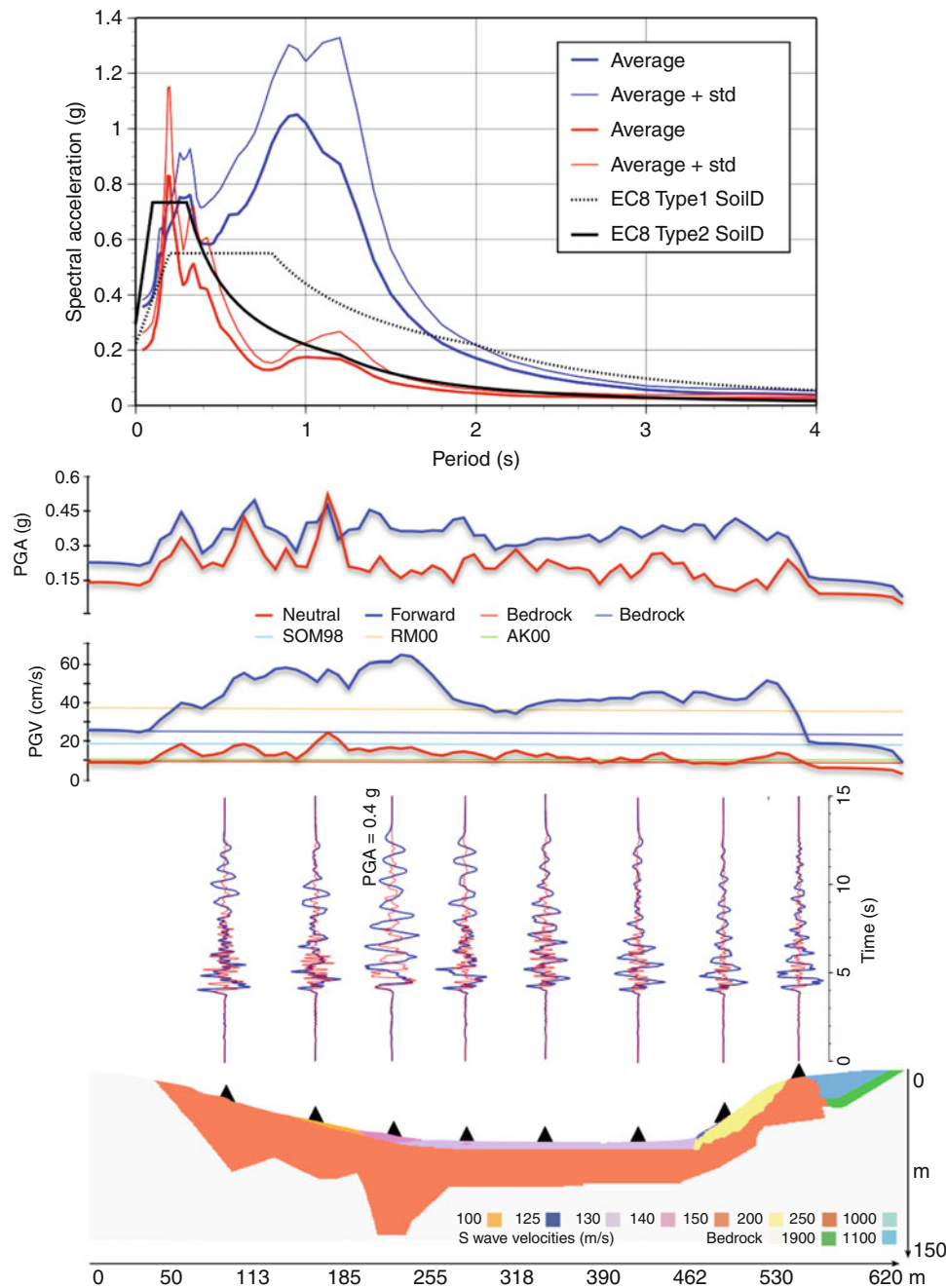
has experienced an impressive growth in the last decades. A description of both analytical and numerical methods can be found, e.g., in Anderson (2003), Bolt and Abrahamson (2003), Bielak et al. (2003). Hybrid techniques, obtained combining analytical and numerical methods, e.g., modal summation method and finite difference (Panza et al. 2001), wave number method and finite difference (Zahradnik and Moczo 1996), finite element and boundary integral method (Bielak et al. 1991), have been developed and a review of these methods dating back to 1980 can be found in Moczo et al. (1997). Numerical-modeling methods for anelastic wave propagation, that take into consideration the earthquake source, propagation path, and local site effects, have become available for the estimation of the site responses and, more generally, for addressing the problem of solving the wave equation in 3D models. Then simulations based on finite difference (see e.g., Sato et al. 1998) or, more recently, finite element (see e.g., Bielak et al. 2003) and Spectral Element Method (SEM) (see e.g., Komatitsch and Tromp 2002a, b; Komatitsch et al. 2004) have been largely used in order to simulate strong-ground motion in 3D structures. A very recent development (La Mura et al. 2010) extended the very efficient analytical modal summation (Panza et al. 2001) to 3D anelastic structures.

Strong-Motion Simulation: An Example

In this section, an example of simulation of strong ground motion is discussed. The determination of the seismic ground motion due to an earthquake with a given magnitude and epicentral distance from the site has been done following a theoretical approach. In order to perform an accurate and realistic estimate of ground shaking scenario, it is necessary to make a parametric study that takes into account the complex combination of the source and propagation parameters, in realistic geological structures.

The realistic modeling of ground motion requires the simultaneous knowledge of the geotechnical, lithological, geophysical parameters, and topography of the medium on one side, and tectonic, historical, paleoseismological, seismotectonic models on the other, for the best possible definition of the probable seismic source.

To deal both with realistic source and structural models, including topographical features, a hybrid method, that combines modal summation and the finite difference technique (e.g., Panza et al. 2001) and optimizes the use of the advantages of both methods, has been used. Wave propagation is treated by means of the modal-summation technique from the source to the vicinity of the local, heterogeneous structure that has to be modeled in detail. A laterally homogeneous, anelastic structural model is adopted that represents the average crustal properties of the region. The generated wavefield is



Earthquakes, Strong-Ground Motion, Fig. 1 From bottom to top: local geotechnical model schematizing the surface geology along the Warth bridge. Black triangles show the sites of the abutments and of the piers along the section. Transverse acceleration time series computed at the eight pier sites. The amplitude of the signals is normalized with respect to the maximum one (0.4 g). Blue and red curves correspond to the case of forward and neutral directivity, respectively. Peak Ground Velocities (PGV) calculated along the profile, for forward (blue curves) and neutral (red curves) directivity, for a unilateral rupture and for the bedrock case (BED), that can be compared with the regression analysis for PGV made using a database of near-fault, forward-directivity

motions, proposed by: Somerville (1998) -SOM98- azure curve, Alavi and Krawinkler (2000) -AK00- orange curve and Rodriguez-Marek (2000) -RM00- green curve, for a magnitude equal to 5.5. Peak Ground Accelerations (PGA) calculated along the profile, for forward (blue curves) and neutral (red curves) directivity, for a unilateral rupture. Average Response Spectra Accelerations (ARSA) and ARSA + one standard deviation computed for the eight sites; blue and red curves correspond to the case of forward and neutral directivity, respectively. The black lines correspond to the two (Type 1 and 2) Design Spectra suggested by EC-8 (EN 1998) for a 5.5 magnitude earthquake and for a soil of class D

then introduced in the grid that defines the heterogeneous area and it is propagated according to the finite differences scheme. With this approach, source, path, and site effects are all taken into account simultaneously.

An example of seismic-input computation is illustrated. The computations are performed for the Warth bridge, freeway A2, 63 km south of Vienna (Austria), a location where no seismic records are available (see Romanelli et al. 2003, 2004). The information about the possible seismic input is limited to an estimate of the macroseismic intensity, in the range from VI to VIII (MSK), the value of the magnitude, M , of the nearest largest recorded event, $M = 5.5$ and the most probable focal depth of strong earthquakes, in the range from 6 to 11 km.

The synthetic time signals are calculated for the three components of motion and the working magnitude is 5.5 (seismic moment equal to 1.8×10^{17} Nm). The study of possible directivity effects in the direction of the Warth bridge (see cross section in Fig. 1) has been performed with the method developed by Gusev (2010), based on the modeling of a Haskell-type seismic source (Haskell 1964, 1966), where a stochastic component allows to build a spectrum (amplitude and phase) of the source function that takes into account both the rupture process and directivity effects. The near-source effect is shown in Fig. 1, where the acceleration time series at the eight sites are plotted in the case of forward (blue curves) and neutral (red curves) directivity for a unilateral rupture. In the figure are also reported the Peak Ground Velocities (PGV) and Peak Ground Accelerations (PGA) along the profile, for neutral (N) and forward (F) directivity, calculated for a unilateral rupture, both for the bedrock case and taking into account the local site conditions. The curves can be compared with the regression analysis for PGV made using a database of near-fault, forward-directivity motions proposed by Somerville (1998), Alavi and Krawinkler (2000) and Rodriguez-Marek (2000). From the results it is evident the strong influence of the rupture and site effects on the ground-motion characteristics, in terms of amplitude and duration. The accelerograms exhibit the greatest peaks in the frequency range, from 1 to 6 Hz, and reach considerable peak values, around 0.4 g.

The analysis of the computed strong-ground motion can be carried out in the time domain (broadband ground-motion time series) and other domains (e.g., Fourier and response spectra). The results show that lateral heterogeneities and source effects can produce strong spatial variations (e.g., more than doubling the amplitudes) in the ground motion (also in terms of differential motion) even over distances of a few tens of meters. In other words, considering that an increment of one intensity degree in the MCS scale roughly corresponds to a doubling of the PGA (Cancani 1904), a general result of the modeling is that a difference greater

than one unit in the seismic intensity can be experienced at sites as close as a few tens of meters.

Conclusions

Solution to earthquake engineering problems, either design or estimation of seismic hazard, are today very different than they were before a strong interaction between seismologists and engineers. Today the differences between strong motion and weak motion have been softened, due to progress in instrumentation and the awareness that both kinds of motions are useful to the understanding of the nature of earthquakes. Today it is unthinkable to study seismic-rupture processes and seismic-wave generation of large earthquakes without near-source strong-motion data; however, while waiting for the enlargement of the strong-motion data set, a very useful approach is the development and use of modeling tools based, on one hand, on the theoretical knowledge of the physics of the seismic source and of wave propagation and, on the other hand, exploiting the rich database about the geotechnical, geological, tectonic, seismotectonic, and historical information already available. With these efforts, the prediction of strong motions through physical modeling is adequate and desirable.

Cross-References

- [Earthquakes, Intensity](#)
- [Earthquakes, PAGER](#)
- [Earthquakes, ShakeMap](#)
- [Propagation of Elastic Waves: Fundamentals](#)

Bibliography

- Aagaard BT, Hall JF, Heaton T (2001) Characterization of near source ground motions with earthquake simulations. *Earthquake Spectra* 17:177–207
- Aki K, Richards PG (1980) *Quantitative Seismology*. In: Theory and method. W.H. Freeman and Co, San Francisco
- Alavi B, Krawinkler H (2000) Consideration of near-fault ground motion effects in seismic design. In: *Proceedings of the 12th world conference on earthquake engineering*. The New Zealand Society for Earthquake, Auckland, p 8
- Anderson JG (2003) Strong-motion seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*. IASPEI, vol 81B. Academic, San Diego, pp 983–965
- Anderson JG (2007) Physical processes that control strong ground motion. In: Kanamori H (ed) *Treatise on geophysics*, volume 4, earthquake seismology, vol 4. Elsevier, Amsterdam, pp 513–565
- Anderson JG, Bodin P, Brune JN, Prince J, Singh SK, Quaaas R, Onate M (1986) Strong ground motion from the Michoacan, Mexico, earthquake. *Science* 233:1043–1049

- Bardet J (2003) Advances in analysis of soil liquefaction during earthquakes. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology. IASPEI, vol 81B. Academic, San Diego, pp 1175–1201
- Benioff H (1934) The physical evaluation of seismic destructiveness. *Bull Seismol Soc Am* 24:9–48
- Bielak J, MacCamy RC, McGhee DS, Barry A (1991) Unified symmetric BEM-FEM for site-effects on ground motion-SH waves. *J Eng Mech* 117:2265–2285
- Bielak J, Loukakis K, Hisada Y, Yoshimura C (2003) Domain reduction method for three-dimensional earthquake modeling in localized regions, part I: theory. *Bull Seismol Soc Am* 93:817–824
- Bolt BA (1969) Duration of strong motion. In: Proceedings 4th WCEE, Rome, pp 1304–1315
- Bolt BA (1975) The San Fernando earthquake, 1971. Magnitudes, aftershocks, and fault dynamics. In: Bulletin 196. California Division of Mines and Geology, Sacramento, Chap. 21
- Bolt BA (1999) Estimating seismic ground motion. *Earthquake Spectra* 15:187–197
- Bolt BA, Abrahamson NA (2003) Estimation of strong seismic ground motions. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology. IASPEI, vol 81B. Academic, San Diego, pp 983–1001
- Bommer JJ, Martinez-Pereira A (1998) The effective duration of earthquake strong motion. *J Earthq Eng* 3(2):127–172
- Bouchon M, Sánchez-Sesma FJ (2007) Boundary integral equations and boundary elements methods in elastodynamics. In: Wu RS, Maupin V, Dmowska R (eds) Advances in wave propagation in heterogeneous earth, advances in geophysics, vol 48. Elsevier–Academic, New York, pp 157–166
- Cancani A (1904) Sur l'emploi d'une double échelle sismique des intensités, empirique et absolue. *G Beitr* 2:281–283
- Chávez-García FJ, Salazar L (2002) Strong motion in Central Mexico: a model based on data analysis and simple modeling. *Bull Seismol Soc Am* 92:3087–3101
- Dan K, Watanabe T, Tanaka T, Sato R (1990) Stability of earthquake ground motion synthesized using different small-event records as empirical Green's functions. *Bull Seismol Soc Am* 80:1433–1455
- Douglas J (2003) Earthquake ground motion estimation using strong-motion. A review of equations for the estimation of peak ground acceleration response spectral ordinates. *Earth Sci Rev* 61:43–104
- EN (1998) General rules, seismic actions and rules for buildings. CEN European Committee for Standardization, Bruxelles
- Fah D, Panza GF (1994) Realistic modeling of observed seismic motion in complex sedimentary basins. *Ann Geofis* 37:1771–1797
- Fah D, Suhadolc P, Mueller S, Panza GF (1994) A hybrid method for the estimation for ground motion in sedimentary basins: quantitative modeling for Mexico City. *Bull Seismol Soc Am* 84:383–399
- Field EH, SCEC Phase III Working Group (2000) Accounting for site effects in probabilistic seismic hazard analyses of Southern California: overview of the SCEC phase III report. *Bull Seismol Soc Am* 90(6B):S1–S31
- Geller RJ (1976) Scaling relations for earthquake source parameters and magnitudes. *Bull Seismol Soc Am* 66:1501–1523
- Graves RW (1996) Simulating seismic wave propagation in 3D elastic media using staggered-grid finite differences. *Bull Seismol Soc Am* 86:1091–1106
- Gusev AA (2010) Broadband kinematic stochastic simulation of an earthquake source: a refined procedure for application in seismic hazard studies. *Pure Appl Geophys*
- Hartzell SH (1978) Earthquake aftershocks as Green's functions. *Geophys Res Lett* 5:1–4
- Haskell NA (1964) Total energy and energy spectra density of elastic wave radiation from propagating faults. *Bull Seismol Soc Am* 54:1811–1841
- Haskell NA (1966) Total energy and energy spectral density of elastic wave radiation from propagating faults. 2. A statistical source model. *Bull Seismol Soc Am* 56:125–140
- Hisada Y, Bielak J (2003) A theoretical method for computing near-fault ground motions in layered half-spaces considering static offset due to surface faulting, with a physical interpretation of fling step and rupture directivity. *Bull Seismol Soc Am* 93:1154–1168
- Irikura K (1986) Prediction of strong acceleration motion's using empirical Green's function. In: Proceedings of 7th Japan earthquake engineering symposium, Tokyo, pp 151–156
- Irikura K, Miyake H, Iwata T, Kamae K, Kawabe H, Dalgner LA (2004) Recipe for predicting strong ground motions from future large earthquakes. In: Proceedings 13th world conference of earthquake engineering, Paper no. 1371
- Jennings PC (2003) An introduction to the earthquake response of structures. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology. IASPEI, vol 81B. Academic, San Diego, pp 1097–1125
- Kawase H (2003) Site effects on strong ground motion. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology. IASPEI, vol 81B. Academic, San Diego, pp 1013–1030
- Komatitsch D, Tromp J (1999) Introduction to the spectral-element method for 3-D seismic wave propagation. *Geophys J Int* 139:806–822
- Komatitsch D, Tromp J (2002a) Spectral-element simulations of global seismic wave propagation, I. Validation. *Geophys J Int* 149:390–412
- Komatitsch D, Tromp J (2002b) Spectral-element simulations of global seismic wave propagation, II. 3-D models, oceans, rotation, and self-gravitation. *Geophys J Int* 150:303–318
- Komatitsch D, Vilotte JP (1998) The spectral-element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures. *Bull Seismol Soc Am* 88:368–392
- Komatitsch D, Liu Q, Tromp J, Peter Suss P, Stidham C, John H, Shaw JH (2004) Simulations of ground motion in the Los Angeles basin based upon the spectral-element method. *Bull Seismol Soc Am* 94:187–206
- La Mura C, Yanovskaya TB, Romanelli F, Panza GF (2010) Three-dimensional seismic wave propagation by modal summation: method and validation. *Pure Appl Geophys*. <https://doi.org/10.1007/s00024-010-0165-2>
- Moczo P, Bystricky J, Kristek J, Carcione JM, Bouchon M (1997) Hybrid modeling of P-SV seismic motion at inhomogeneous viscoelastic topographic structures. *Bull Seismol Soc Am* 87:1305–1323
- Montejo LA, Kowalsky MJ (2008) Estimation of frequency-dependent strong motion duration via wavelets and its influence on nonlinear seismic response. *Comput Aided Civ Inf Eng* 23:253–264
- Nunziata C, De Nisco G, Panza GF (2008) Evaluation of liquefaction potential for building code. In: Santini A, Moraci N (eds) Seismic engineering conference commemorating the 1908 Messina and Reggio Calabria earthquake. Melville, New York, pp 370–377
- Oprsal I, Zahradnik J (1999) Elastic finite-difference method for irregular grids. *Geophysics* 64:240–250
- Oprsal I, Zahradnik J (2002) Three-dimensional finite-difference method and hybrid modeling of earthquake ground motion. *J Geophys Res* 107(B8):2161. <https://doi.org/10.1029/2000JB000082>
- Oth A, Wenzel F, Wust-Bloch H, Gottschammer E, Ben-Avraham Z (2007) Parametrization of a composite attenuation relation for the Dead Sea area based on 3D modeling of elastic wave propagation. *Pure Appl Geophys* 164:23–37
- Panza GF, Suhadolc P (1989) Realistic simulation and prediction of strong ground motion. In: Carlomagno GM, Brebbia CA (eds) Computers and experiments in stress analysis. Springer, New York, pp 77–98
- Panza GF, Romanelli F, Vaccari F (2001) Seismic wave propagation in laterally heterogeneous anelastic media: theory and applications to seismic zonation. In: Dmowska R, Saltzman B (eds) Advances in geophysics, vol 43. Academic, San Diego, pp 1–95
- Panza GF, Romanelli F, Vaccari F, Decanini L, Molaioli F (2004) Seismic ground motion modeling and damage earthquake scenarios:

- a possible bridge between seismologists and seismic engineers. In: Chen YT, Panza GF, Wu ZL (eds) *Earthquake hazard, risk, and strong ground motion*. Seismological, Beijing, pp 323–349
- Parvez AI, Gusev AA, Panza GF, Petukhin AG (2001) Preliminary determination interdependence among strong motion amplitude, earthquake magnitude and hypocentral distance for the Himalayan region. *Geophys J Int* 144:577–596
- Peresan A, Zuccolo E, Vaccari F, Gorshkov A, Panza GF (2010) Neodeterministic seismic hazard and pattern recognition techniques: time-dependent scenarios for north-eastern Italy. *Pure Appl Geophys*. <https://doi.org/10.1007/s00024-010-0166-1>.
- Pitarka A (1999) 3D elastic finite-difference modeling of seismic motion using staggered grids with non-uniform spacing. *Bull Seismol Soc Am* 89:54–68
- Reid HF (1910) The California earthquake of April 18, 1906. Report of the state earthquake investigation commission, Vol. II, The Mechanics of the earthquake. Carnegie Institution of Washington, Publication no. 87
- Rodriguez-Marek A (2000) Near-fault seismic site response. PhD dissertation, Department of Civil Engineering, University of California, Berkeley, p 451
- Romanelli F, Vaccari F, Panza GF (2003) Realistic modelling of the seismic input: site effects and parametric studies. *J Seismol Earthq Eng* 5:27–39
- Romanelli F, Vaccari F, Panza GF (2004) Realistic modelling of the effects of asynchronous motion at the base of bridge piers. *J Seismol Earthq Eng* 6:17–26
- Sato T, Helmberger DV, Somerville PG, Graves RW, Saikia CK (1998) Estimates of regional and local strong motions during the great 1923 Kanto, Japan, earthquake (Ms 8.2). Part 1: source estimation of calibration event and modeling of wave propagation path. *Bull Seismol Soc Am* 88:183–205
- Somerville PG (1998) Development of an improved representation of near fault ground motions. In: *Proceedings SMIP98 seminar on utilization of strong motion data*. California Strong Motion Instrumentation Program, Oakland, pp 1–20
- Somerville PG (2002) Characterizing near fault ground motion for the design and evaluation of bridges. In: Nimis R, Bruneau M (eds) *Proceedings of the third national seismic conference and workshop on bridges and highways*. MCEER, Buffalo, pp 137–148
- Somerville PG, Smith NF, Graves RW, Abrahamson NA (1997) Modification of modification of empirical strong ground motion attenuation relations to include the amplitude and duration effects of rupture directivity. *Seismol Res Lett* 68:199–222
- Strasser FO, Bommer JJ (2009) Review: strong ground motions – have we seen the worst? *Bull Seismol Soc Am* 99:2613–2637
- Takenaka H, Furumura T, Fujiwara H (1998) Recent developments in numerical methods for ground motion simulation. In: Irikura K, Kudo K, Okada H, Sasatani T (eds) *The effects of surface geology on seismic motion*, vol 2. Balkema, Rotterdam, pp 91–101
- Trifunac MD, Westermo BD (1977) A note on the correlation of frequency dependent duration of strong earthquake ground motion with the modified Mercalli intensity and the geologic conditions at the recording site. *Bull Seismol Soc Am* 67:917–927
- Villaverde R (2009) *Fundamental concepts of earthquake engineering*. CRC, Boca Raton
- Wells DL, Coppersmith KJ (1994) New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bull Seismol Soc Am* 84:974–1002
- Wu RS, Maupin V (2007) Advances in wave propagation in heterogeneous earth. In: Dmowska R (ed) *Advances in geophysics*, vol 48. Elsevier–Academic, San Diego
- Yoshimura C, Bielak J, Hisada Y, Fernandez A (2003) Domain reduction method for three-dimensional earthquake modeling in localized regions part II: verification and applications. *Bull Seismol Soc Am* 93:825–840
- Youd TL (2003) Liquefaction mechanisms and induced ground failure. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*. IASPEI, vol 81B. Academic, San Diego, pp 1159–1173
- Zahradnik J, Moczo P (1996) Hybrid seismic modeling based on discrete-wave number and finite-difference methods. *Pure Appl Geophys* 148:21–38

Earthquakes, Volcanogenic

J. W. Neuberg

Institute of Geophysics and Tectonics, School of Earth and Environment, University of Leeds, Leeds, UK

Definition

- | | |
|------------------------------|---|
| Volcano-tectonic earthquakes | “Conventional” double-couple earthquakes of shallow tectonic origin, occurring in a volcanic setting, and comprising primary (<i>P</i>) and secondary (<i>S</i>) waves as well as surface waves. |
| Rock-fall events | High-frequency (above 2 Hz), “cigar-shaped” signals caused by the surface impact of solid, crystallized lava dome material, consisting mainly of surface waves. |
| Low-frequency earthquakes | Earthquakes with low-frequency content (0.5–2 Hz), generated at the interface between a magmatic fluid and solid rock, causing parts of the volcanic plumbing system such as conduits, dykes, and cracks to resonate. |
| Volcanic tremor | Either unstructured, random background noise on a volcano caused by magma movement, or merging swarms of low-frequency earthquakes that retain all characteristics of the single events. |
| Very-long-period earthquakes | Seismic transients with periods between tens of seconds and several minutes, caused by magma/gas movement in the shallow plumbing system. |

Volcano Seismology

Introduction

Volcano seismicity covers a large variety of seismic signals caused by the complex interaction of magmatic and hydrothermal fluids with solid rock within the volcanic edifice leading to stress changes, brittle failure of rock, pressure variations, magma fracturing, and fragmentation. Volcano seismology has the aim to identify and separate different seismic signals and attribute them to physical processes, a task that helps volcanologists to understand the dynamics of the magmatic systems.

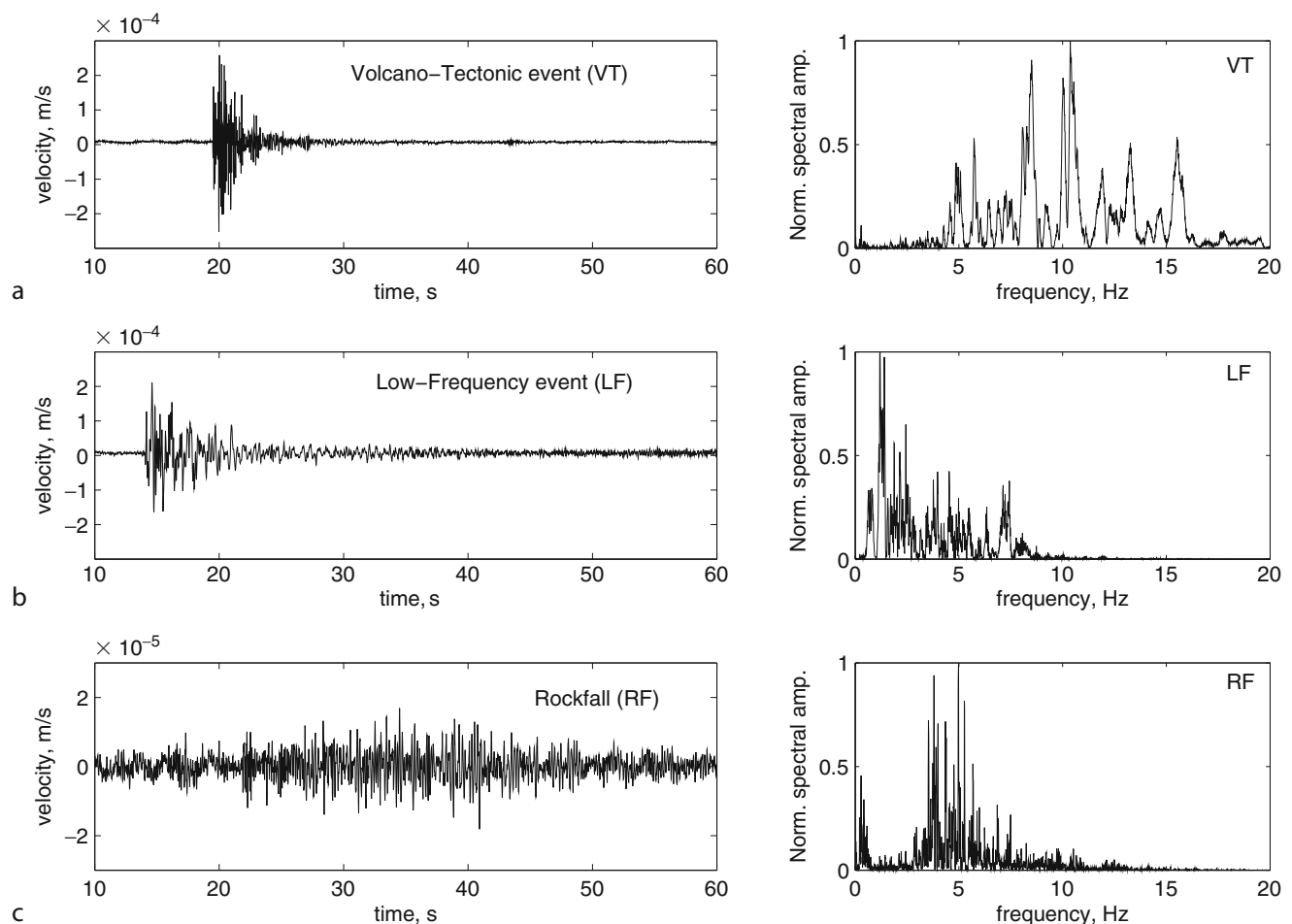
The interpretation of different volcanogenic seismic signals and the development of conceptual and numerical models are at the core of any attempt to forecast volcanic eruptions. While volcano seismology emerged from earthquake seismology, it has evolved into a separate discipline in its own right, dealing with a large complexity of seismic source processes in a volcanic environment.

Crucial for the interpretation of seismic signals is the examination of their frequency content by Fourier or frequency analysis by which the signal is decomposed into harmonic waves and contributions of single frequencies are compared with each other. This decomposition is called a frequency spectrum, and its repeated application in consecutive time windows is referred to as a spectrogram. It turns out that different volcanogenic signals can be easily identified and separated by their frequency content; hence, we will make frequent use of this methodology in the following sections where the characteristics and the possible origin and source mechanism of different volcanogenic signals are described. Single event types are classified according to their suspected

source mechanism rather than using a phenomenological classification relying on their waveform and frequency. A comparison of the three main types of volcanogenic signals and their respective frequency spectra are shown in Fig. 1, while volcanic tremor and very-long-period transients are depicted in Figs. 3, 4, and 5, respectively. An overview describing the development of seismic methods over the last decade using these volcanogenic signals can be found in Chouet and Matoza (2013).

Volcano-Tectonic Earthquakes

Volcanoes are often located in tectonically active areas such as subduction zones, where earthquakes are generated merely by changes in the regional stress field due to the continuous movement of the tectonic plates. When magmatic and hydrothermal processes occur at the same location and time, the local stress field is further modulated and impacted upon by these processes. It is this complex interaction between the regional stress field and local stress modulation by volcanic processes such as dyke intrusion and (de)pressurization of the



Earthquakes, Volcanogenic, Fig. 1 Different event types in time and frequency domain, recorded on Soufriere Hills volcano on Montserrat: (a) volcano-tectonic earthquake (VT), (b) low-frequency earthquake, and (c) rock-fall event

volcanic plumbing system that leads to the generation of volcano-tectonic earthquakes, referred to in the following as VTs for brevity. These stress changes can be created directly by magma emplacement, or indirectly by temperature changes or pressure changes as a result of magma–water interaction in a hydrothermal system.

Assuming that the pressure modulation through a magmatic body produces only a localized effect on the overall stress field impacting only on close distance of a few hundred meters from the magmatic source, the earthquake location in itself indicates the emplacement of fresh magma (or its withdrawal). Hence, the locations of VT earthquakes over a certain time period are an indicator of changes in the system rather than providing a picture of a so-called magma chamber. VTs are generated by *stress changes*, as a consequence of *magma movement*. In this way, Battaglia et al. (2005) traced the rising magma dyke through time at Piton de la Fournaise, La Reunion. White and McCausland (2016) used VTs to detect initial stress changes and the corresponding volumes of intruding magma and demonstrated the forecasting potential of VTs in several examples.

Being generated by brittle fracture of rocks along a fault plane, VTs have all the characteristics of local earthquakes: containing a clear primary (*P*) wave, followed by the secondary (*S*) wave, both in the frequency range of 1–20 Hz (see Fig. 1a). If the recordings of several seismic stations are available, the hypocenter of the event can be located using standard seismological tools, as well as other seismic source parameters, which can be obtained by the fault plane solution. Typical VTs range in magnitude *M_b* between 1 and 3.

Another major difference to usual tectonic earthquakes lies in the geometry of the fault plane. While tectonic events can be sufficiently modeled using piecewise straight fault planes, VTs might be generated on strongly curved fault planes in the volcanic plumbing system of dykes and conduits. This has significant implications on amplitude estimations and the determination of a so-called moment tensor that describes the interaction of forces in the source process. Some of the consequences can be an underestimation of amplitudes and a misinterpretation of the source time history (Contreras-Arratia and Neuberg 2019).

Rock-Fall and Lahar Events

This event type is produced by the impact of falling rock fragments on the surface of the volcanic edifice, and therefore, is typical for dome-building eruptions involving highly viscous magma. There is a continuum between small rock avalanches to highly energetic, so-called pyroclastic flows, which are a mixture of hot gas and rock fragments, ranging from huge boulders to fine volcanic ash particles. The impact that generates the seismic signal depends on rock size, impact angle, and the property of the surface.

The signal, a typical example of which is depicted in Fig. 1c, has a “cigar-shaped” waveform, increasing in amplitude as more and more rocks form a rock avalanche, gaining speed and momentum. Secondary phases can be identified in the waveform as the flow reaches a change in topography slope converting its kinetic energy into a seismic signal. The frequency content is very high, containing mostly surface waves. The end of the “cigar” is reached once the material involved in the flow has settled. Rock-fall signals occur sometimes as secondary events in connection with low-frequency events that are able to cause significant seismic shaking of a dome structure (see below) or volcanic explosions resulting in a shower of volcanic ejecta and collapsing eruption columns. Another signal with a similar waveform is caused by so-called *lahars*, high-density mudflows that can keep big boulders in suspension. These mudflows occur often after heavy rainfall or through melting icecaps that cover parts of the volcano. While rock-fall and lahar events do not offer a detailed insight into magmatic processes within the volcano, they can still be used as a proxy for magmatic extrusion rate, assuming that the lava disintegrates and falls off the dome at the same rate as it is extruded. A further use of lahar and rockfall events lies in the real-time monitoring and determination of the hazard and risk these phenomena might cause.

Low-Frequency Earthquakes

Low-frequency earthquakes, depicted in Fig. 1b and in the following referred to as LF, proved to be pivotal to modern volcano seismology. Several examples have been quoted where these events have been successfully used to forecast volcanic eruptions (e.g., Chouet 1996). Their occurrence alone has been used as an indicative tool even though a full understanding of their source mechanism had not been achieved. On Galeras volcano, Colombia, the occurrence of a few low-frequency earthquakes played a major role in the heated scientific and public debate regarding whether lives could have been saved had their occurrence been recognized as a warning sign for an imminent volcanic eruption (Gil Cruz and Chouet 1997).

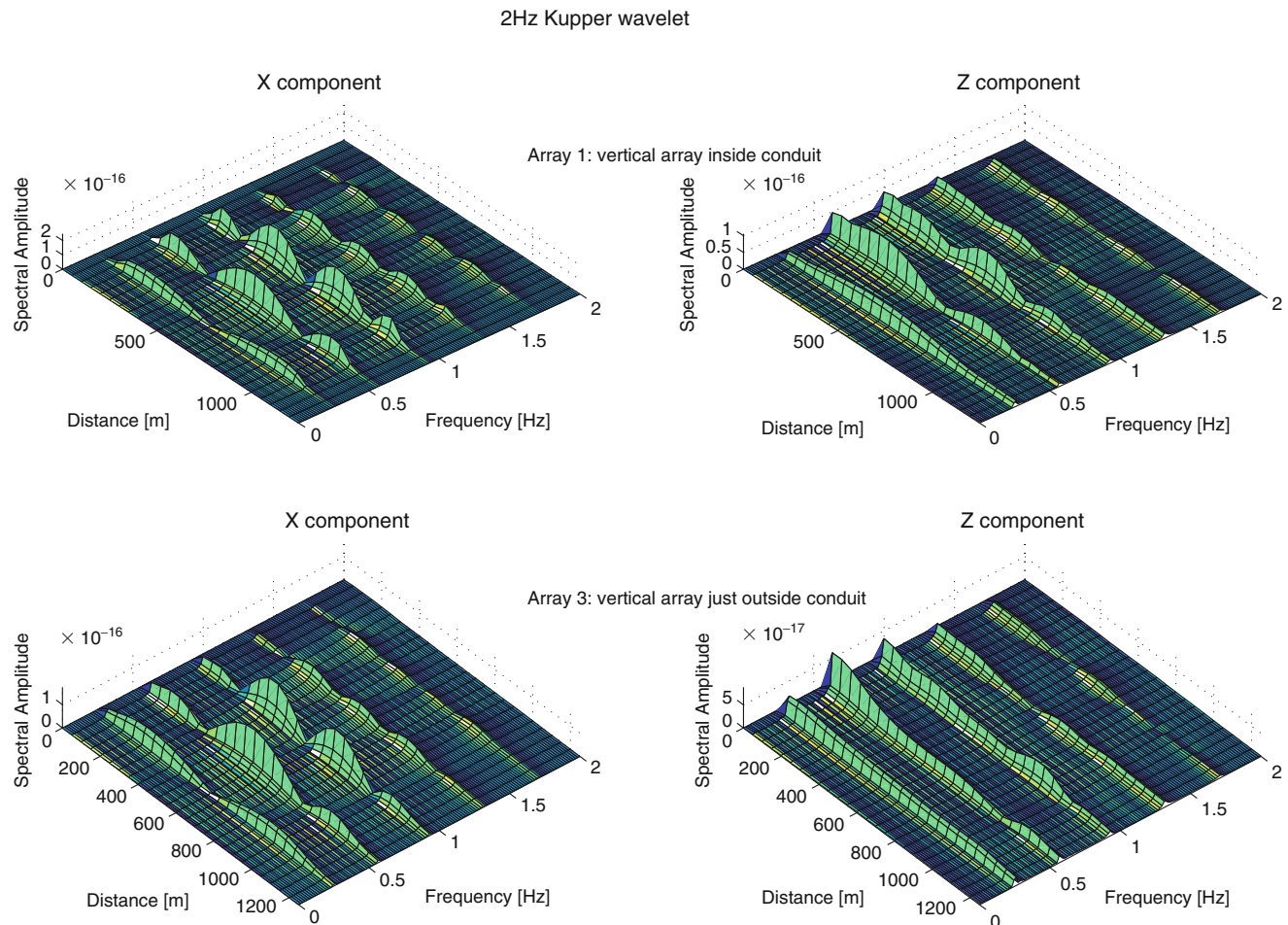
Low-frequency earthquakes occupy a spectral range of approximately 0.2–10 Hz and form a spectral continuum between two end-members of seismic categories: *long-period events* and so-called *hybrid events*, which are similar to long-period events but carry in addition a high-frequency onset. On Soufriere Hills volcano in Montserrat, it has been demonstrated that the amount of high-frequency energy observed in the seismic signal can vary from station to station, and also at one station from event to event. Therefore, long-period and hybrid earthquakes have been grouped together under a more generic term as *low-frequency earthquakes* (Neuberg et al. 2000). These events have been observed on many, mostly dome-building volcanoes, and have been empirically

associated with the pressurization of the volcanic system (Lahr et al. 1994; Neuberg et al. 1998).

Very different from tectonic events, low-frequency earthquakes are generated at a boundary between a fluid, such as gas or magma contained in conduits, dykes, or cracks, and the solid surrounding rock (e.g., Ferrazzini and Aki 1987; Chouet 1988; Neuberg et al. 2000). Most of the seismic energy is trapped in the fluid-filled plumbing system – conduit, dyke, or crack – and resonates within this system while only a small part can escape and propagate through the volcanic edifice and is recorded at the surface. Hence, low-frequency earthquakes are generated by resonating interface waves – hence the term *conduit resonance* – however, these interface waves must not be confused with acoustic waves in the fluid, often referred to as organ pipe modes. Acoustic and interface waves differ in one important aspect: Interface waves are dispersive; hence, the velocity varies with wavelength, and it is the slow moving part of the wave field that explains the low-frequency character (Ferrazzini and Aki 1987). Figure 2 shows the

horizontal and vertical seismic component of a synthetic, resonating interface wave for a two-dimensional dyke model. The single overtones, or modes, are separated by frequency in the depth–frequency plot. The low-frequency seismic wave can be considered as the superposition of these modes. The frequency content is controlled by the interface wave velocity which, in turn, is determined by the rigidity μ of the surrounding rock, the bulk modulus of the fluid κ , the wavelength of the interface wave, and the width of the fluid-filled container.

Low-frequency earthquakes are observed on many volcanoes, and most observations include one or several of the following characteristics: (1) LFs occur in swarms, (2) the waveforms are very similar, (3) their occurrence is highly regular, (4) with increasing occurrence, the LF swarms merge into continuous tremor, and most importantly, (5) major dome collapses or other eruptive phases are preceded by LF swarms (Miller et al. 1998; Aspinall et al. 1998; Hellweg 2003). The highly periodic occurrence of LFs in



Earthquakes, Volcanogenic, Fig. 2 Eigen modes of a low-frequency earthquake source: spectral amplitude distribution for horizontal (X) and vertical (Z) component inside (above) and outside (below) the conduit for single frequencies. The superposition of the frequency contributions

represents the source signal inside the conduit and outside. Note that the vertical component is affected by the interaction with the free surface, while the horizontal component is not

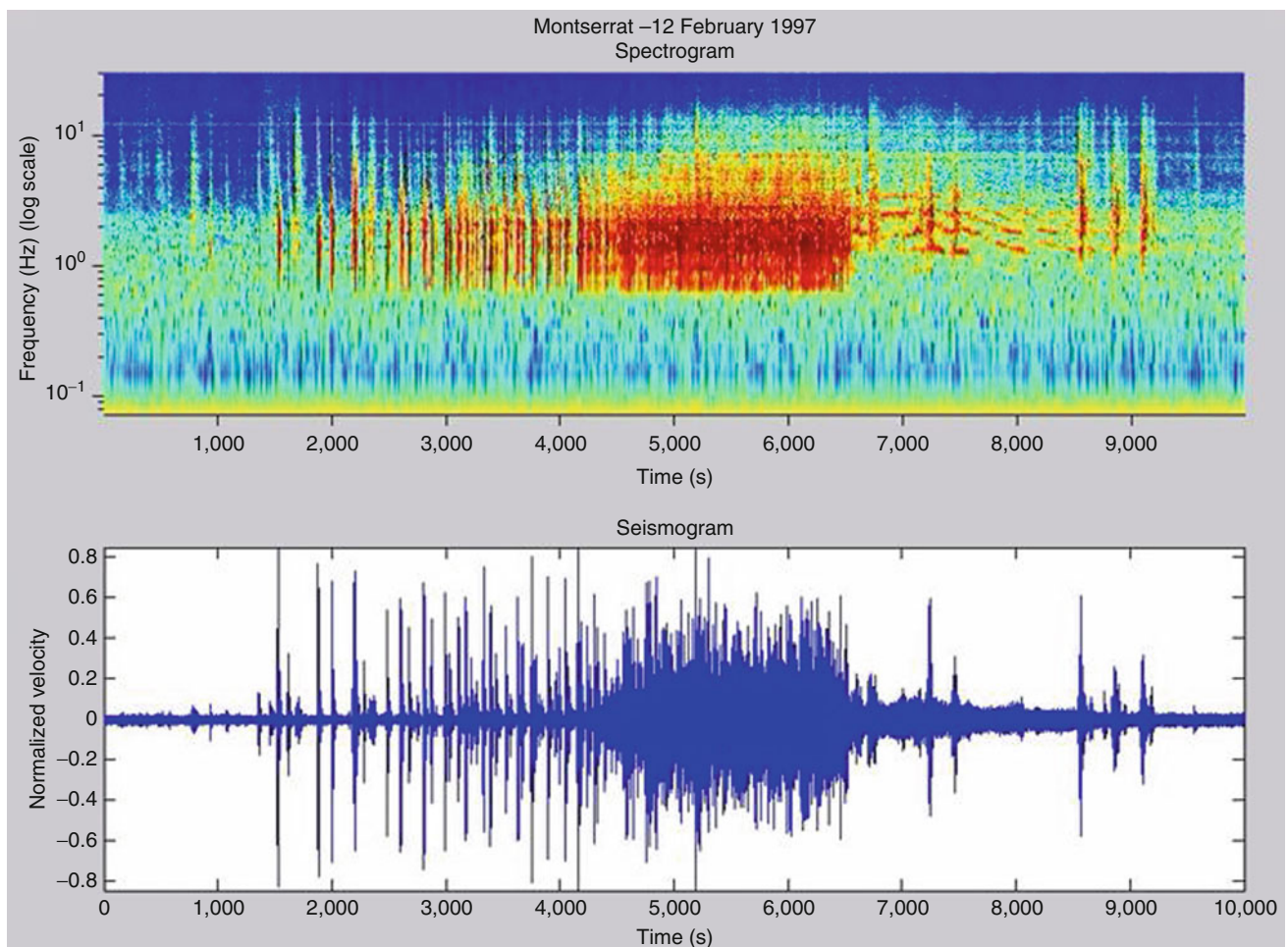
swarms of similar waveform can be seismologically interpreted as a stationary seismic source with a repetitive non-destructible source mechanism (Green and Neuberg 2006). The other exciting aspect is their potential as a forecasting tool for volcanic eruptions or dome collapses. Both Galeras volcano in Colombia and Soufriere Hills volcano in Montserrat are examples where LF swarms occurred before such events; however, the occurrence of the earthquakes does not guarantee a volcanic event. This demonstrates that the mere occurrence of LFs is not yet a reliable forecasting tool unless the actual trigger mechanism is understood and a direct link between the seismogram and source mechanism can be made.

Several trigger mechanisms for LFs have been proposed, including magma–water interaction (Zimanowski 1998), stick-slip motion of magma plugs (Goto 1999; Iverson et al. 2006), magma flow instabilities (Julian 1994), periodic release of gas–ash mixtures into open cracks (Molina et al.

2004), and repeated, brittle fracturing of magma in glass-transition (Neuberg et al. 2006).

Volcanic Tremor

As indicated in the previous section, one definition of volcanic tremor is the superposition of low-frequency earthquakes as they merge in time. Figure 3 depicts an example in time and frequency domain from Soufriere Hills volcano in Montserrat where single LF earthquakes merge into tremor. If the highly periodic occurrence, or trigger mechanism, is maintained, the tremor signal exhibits a spectral signature of integer harmonic spectral lines (see Fig. 3) if the time window from which the spectral slice is computed is long enough to cover several events. The distance between the spectral lines is inversely related to the time separation between LFs of the same waveform, origin, and time history. As the frequency of occurrence increases, the distance between spectral lines will increase as well. In this way, one can measure the inter-event time, or



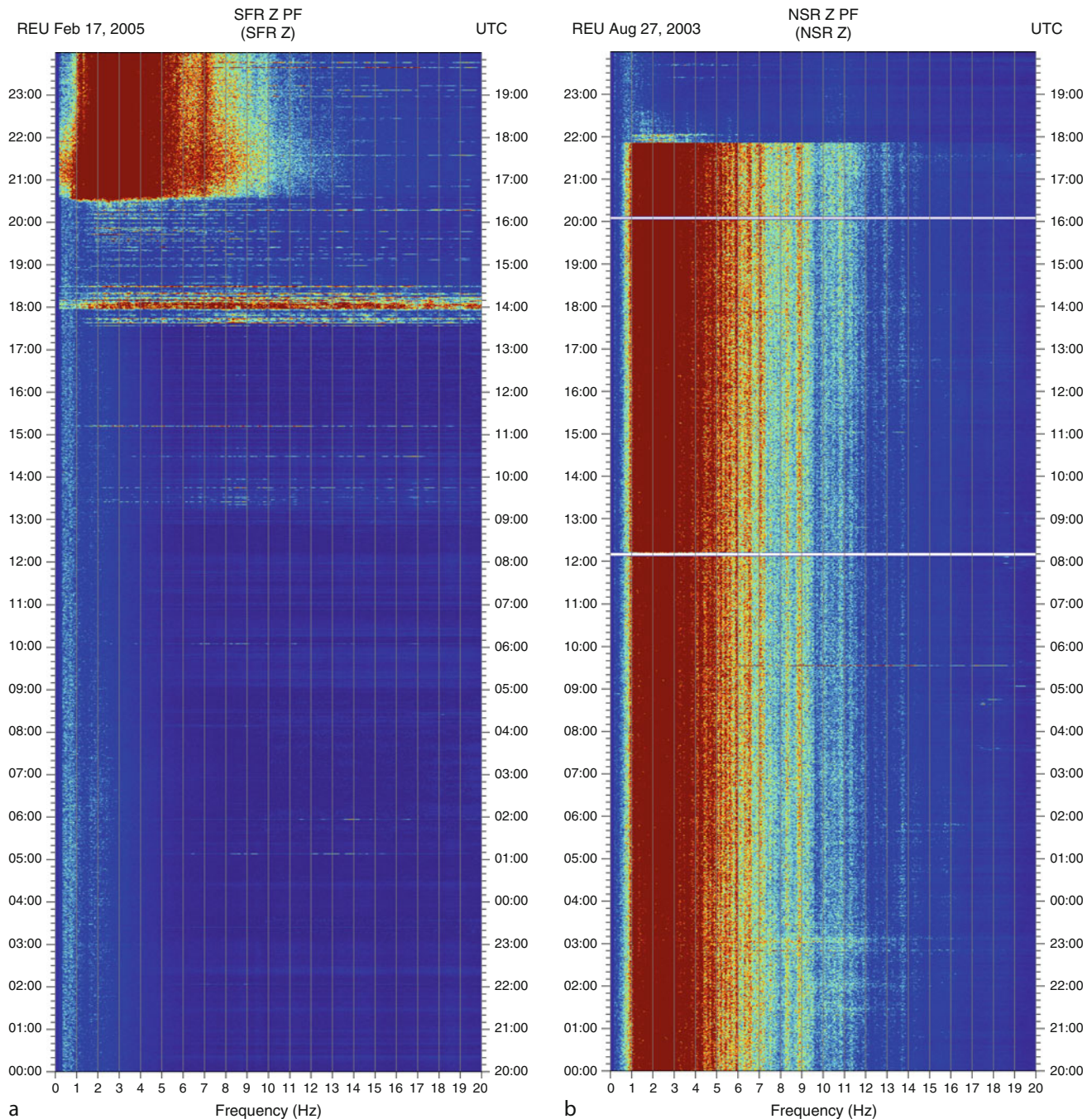
Earthquakes, Volcanogenic, Fig. 3 Seismic low-frequency earthquake swarms merging into tremor: spectrogram (upper part) and seismogram (lower part). Note the *lines* in the spectrogram toward the end of

the signal, representing a highly periodic repeated excitation of identical events. The spectral distance between the lines corresponds to the excitation frequency

event rate, by using the separation of spectral lines, even though the single events have merged in time into a continuous tremor signal. For this type of volcanic tremor, it has been shown that the spectral signature of the tremor signal is identical to the low-frequency earthquakes of which it is composed (Latter 1981).

Another definition of volcanic tremor is more general and refers to the background noise caused by magmatic processes.

Figure 4 shows an example from Piton de la Fournaise, La Reunion, where volcanic tremor is produced by lava extruding from fissures. The onset and end of the tremor signal are remarkably sharp and coincide with the start and finish of the eruption. The spectrum of the signal is “white”; hence, it is composed of a large number of frequencies and does not have any spectral structure. The origin of the tremor is due to pressure fluctuations during turbulent magma flow (Julian



Earthquakes, Volcanogenic, Fig. 4 Spectrograms of continuous tremor at Piton de la Fournaise, La Reunion. (a) Onset of tremor is preceded by a short swarm of VT events, (b) continuous tremor during

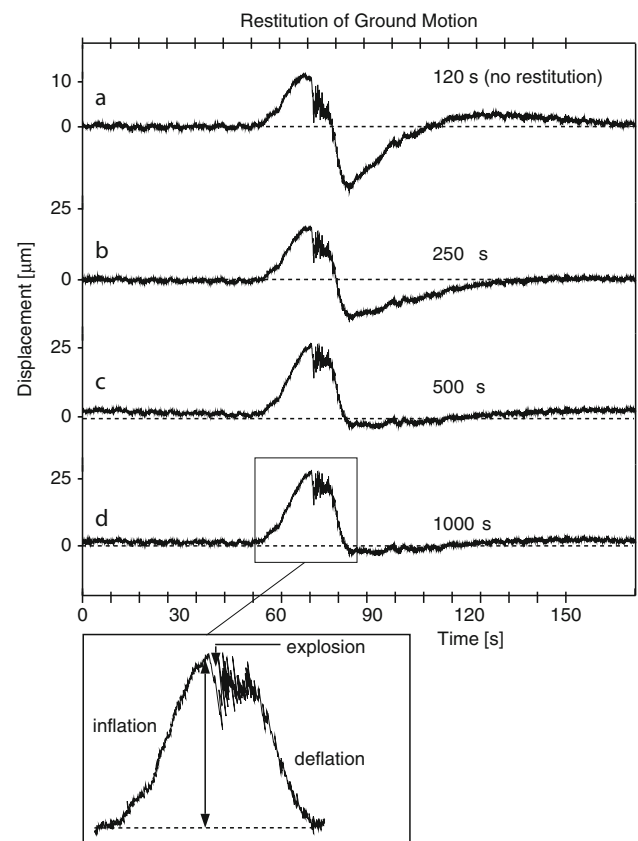
lava extrusion, note the extremely sharp end of the tremor signal when magma extrusion stops

1994). Another tremor source of this kind can be attributed to hydrothermal activity or degassing. While such a tremor signal does not have any distinct seismic phases (like *P* and *S* wave arrival) that could be used to determine the source location on the base of relative travel times, the relative amplitudes of the signal can be utilized to obtain an approximate location of the tremor source as demonstrated by Taisne et al. (2011).

Very-Long-Period Earthquakes

With the availability of broadband seismic sensors, seismology opened a new window to very-long-period signals caused by transients of pressure variations in the volcanic edifice with periods ranging from tens of seconds to several minutes (Neuberg et al. 1994; Wielandt and Forbriger 1999). Measurements in this frequency range close the gap between classic seismology and ground deformation as determined by strain-, tilt-meters, and GPS. However, while ground deformation can be measured directly as ground *displacement*, a seismometer detects the ground *velocity*, or the temporal change of the displacement, due to its electromagnetic measurement principle, reacting to the *change* of the magnetic flux by inducing an electric current. The advantage of such a system lies in the high sensitivity of the detection system; the disadvantage is that several data processing steps are necessary to retrieve the ground displacement, if this is the quantity required for modeling or further interpretation. For short-period signals, this does not pose any problem as the instrument response is constant in this narrow (high-) frequency band; hence, the seismogram has only to be integrated to convert ground velocity into displacement. However, for a broadband seismometer, the instrument response is highly frequency dependent and lower frequencies are suppressed by the seismometer. In order to retrieve the ground displacement from the measured ground velocity, the effect of the instrument response has to be taken into account and, if possible, to be removed. This is in addition to the integration step in the data processing sequence.

Modern broadband seismometers have a “flat” frequency response typically in the range between 50 and 0.008 Hz, expressed in periods, between 0.02 and 120 s. Figure 5 shows the effect of the removal of the instrument response in an example from Stromboli volcano. The seismic record of a Strombolian eruption shows an ultra-long-period transient in the vertical direction, caused by an ascending magma/gas plug with an eruptive, high-frequency seismic signature superimposed. In order to represent the long-period part of the signal properly, the original velocity seismogram has to be corrected for the instrument response as well as integrated. This processing sequence is shown in Fig. 5, where the velocity seismogram is converted to displacement by integration (Fig. 5a), and where finally, after the instrument response has been successfully removed, the seismogram (Fig. 5d)



Earthquakes, Volcanogenic, Fig. 5 Very-long-period signal from Stromboli volcano after removal of instrument response. (a) through (d) indicate the period limit to which the spectral energy has been included in the restitution. While the trace in (a) contains only spectral components up to 120 s period, the true ground displacement is reconstructed by including all spectral components to a period of 500 s (c) or even 1000 s (d). The final result in the lowermost trace shows the ground displacement associated with the rising gas slug, the Strombolian explosion as well as the deflation after the explosion

represents the true ground displacement at the location of the seismic sensor. The comparison between corrected and uncorrected displacement demonstrates the importance of this processing steps. An interpretation or model of the uncorrected trace (Fig. 5a) would comprise an upward movement of the volcanic edifice prior to the Strombolian eruption followed by a deflation and slow recovery of the edifice. The corrected trace offers a much simpler interpretation (Neuberg and Luckett 1996): The edifice inflates while the magma plug is ascending, and once the gas is released through the eruption the edifice deflates to its original level. This example demonstrates that each seismic sensor, and indeed each recording instrument, has to be considered a *filter* that affects and modifies the original physical signal in a certain way, often introducing a frequency-dependent change in amplitude and phase that results in an entirely different waveform. To correct for these effects is absolutely essential when very-long-period seismic signals are interpreted in terms of volcanic source processes.

Volcano Seismology in a Wider Volcanological Context

Volcanogenic earthquakes have been the first signals to be used to monitor and assess volcanic activity, and are today still at the core of any volcano monitoring program. However, if combined with other observations and integrated into wider conceptual and numerical volcanological models, their potential as a forecasting tool increases even further. In the following section, we give a few examples how volcanogenic earthquakes have been used to assess several aspects of volcanic activity.

Estimation of Magma Ascent Rate

A very exciting example from Montserrat has been reported by Voight et al. (1998), where a radial tilt signal recorded on Chances Peak, an old dome structure adjacent to the active lava dome of Soufriere Hills volcano, correlated perfectly with the cyclic occurrence of seismic swarms. Neuberg et al. (2006) reinvestigated this correlation and found that seismic activity, in this case, low-frequency earthquake swarms, commences when the tilt signal goes through an inflection point (Fig. 6). Furthermore, low-frequency seismicity ceases when the tilt signal goes through a second turning point. The tilt, which is the rotation of the volcano's flank, is caused by a combination of a pressure buildup in a shallow magma reservoir and the shear stress acting across the conduit wall by the motion of viscous magma at shallow depth (Green et al. 2006). Hence, the cyclic occurrence of LF swarms can be directly related to the magma ascent rate: As the pressure in the magma has reached a certain threshold, magma starts to move causing low-frequency seismicity by either a stick-slip motion (Iverson et al. 2006) or by magma rupture (Neuberg et al. 2006). This causes the turning point in the tilt behavior;

however, the ground deformation (tilt) still increases to its maximum where the magma ascent rate peaks. At this point, the magma ascent slows down as gas bubble diffusion cannot keep up with the pressure drop, caused by the magma movement. This can be clearly seen in both the tilt signal and the seismicity rate depicted in the inset of Fig. 6. At the second turning point, magma ascent rate has become so low or even stopped completely that no further LF seismicity is produced. This gives gas diffusion a chance to catch up, leading to a renewed pressurization of the system again, and a new cycle has started.

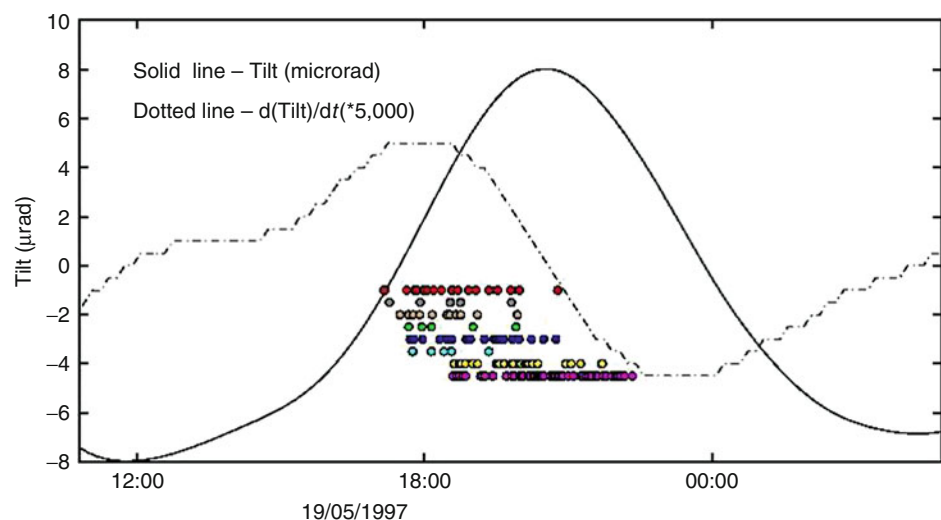
Tungurahua volcano in Ecuador shows a similar anti-correlated behavior of tilt and seismicity pointing to a mechanism where shear stress at the conduit wall is responsible for both the generation of LFs and tilt. Shear stress partitioning leads to the anticorrelation where tilt decreases when shear stress is released by slip through the generation of earthquakes (Neuberg et al. 2018).

These examples demonstrate how LF seismicity can be interpreted following a conceptual model that allows to estimate the changes in magma ascent rates on the base of seismicity, even if no tilt signal is available. Using such a model, Hammer and Neuberg (2009) found an accelerating seismicity rate in consecutive LF swarms on Montserrat – as a proxy for an accelerated magma ascent – which led to a dome collapse in 1997.

Combining Seismicity with Infrasonic Signals

A second example of a combined interpretation uses infrasonic acoustic signals with seismicity. The monitoring of infrasonic signals on active volcanoes has become routine in the last 10 years and turned out extremely helpful for open volcanic systems.

Earthquakes, Volcanogenic, Fig. 6 Correlation between tilt and seismicity on Soufriere Hills volcano, Montserrat. The low-frequency seismicity is sorted in different event “families.” The different colors indicate different families of similar seismic waveform which implies that each family has a common source mechanism and origin. It commences and ceases at the inflection points of the tilt signal (solid line), equivalent to the maximum and minimum of its time derivative (dotted line)



See Johnson (2003) for an overview on infrasonic monitoring techniques. Compared to seismic waves, infrasonic signals are less affected by the medium (steam and air) through which they propagate, and the relative timing between the two types of signals can be used to dissect a complex volcanic event such as a dome collapse, which can last for several hours and exhibit distinct phases. Green and Neuberg (2005) modeled a partial dome collapse at Montserrat in 2004 and used the relative timing between seismic and infrasonic pulses to separate several distinct explosions during the collapse. Furthermore, they identified a plume forming degassing event as the underlying cause of the collapse. A further application of a combined interpretation can be found in Ripepe et al. (2002) where relative timing between infrasound and seismic signal revealed the relative height of a magma column on Stromboli volcano from where an explosion occurred.

Detection of Lahars and Pyroclastic Flows

Pyroclastic flows, lava flows, and lahars often escape direct visual observation, but the fast determination of their location can be crucial in terms of risk mitigation. As pointed out above, the seismic signals produced by such events lack any clear seismic phase that could be used for a classic travel-time-based location procedure. Jolly et al. (2002) devised a method based on a relative seismic amplitude estimation across a seismic monitoring network, which allowed to follow the path of a flow in near real-time. The basic idea is to determine the location of the seismic surface source by assuming that the amplitudes of the surface waves caused by the flow decay by the inverse of the distance to a seismic station and by inelastic attenuation. They performed a so-called grid search inversion across the volcanic edifice to locate the flow. Through this technique, they were able to resolve fine details of the flow propagation such as a “piling-effect” that occurred when the front of the pyroclastic flow came to a halt while the rest of the flow piled up behind it, causing the seismic source traveling up hill!

On Piton de la Fournaise, a similar method is used (Taisne et al. 2011) to detect the source of basaltic lava flows, which are responsible for the generation of a tremor signal. In this way, the observatory staff is able to inform the authorities in near real-time about the source location of the lava flow.

Detecting Stress Changes in Volcanic Settings

Recent advances have made it possible to detect temporal variations in properties of seismic signals that are caused by stress changes in the medium that they penetrate. These changes can be seen in directional variations of the wave speed (also referred to as anisotropy), energy loss, and seismic source parameters (fault plane solution) of volcano-tectonic events.

Gerst and Savage (2004) detected a change in anisotropy at Mt Ruapehu, New Zealand, which coincided with the

eruptive activity in 1995 and 1996. They interpreted these changes to be caused by changes in the stress due to magmatic activity that altered the orientation and volume of cracks in the volcanic edifice. Using noise correlation techniques, Brenguier et al. (2008) observed at Piton de la Fournaise, La Reunion, similar changes in seismic velocity caused by the pressurization of the volcanic edifice due to ascending magma. Roman et al. (2006) found systematic variations in the orientation of the main pressure axis, which was derived from fault plane solutions of volcano-tectonic events on Soufriere Hills volcano on Montserrat. They identified distinct episodes in the eruptive history of Soufriere Hills volcano, where the regional stress field was superimposed by local stress perturbations that caused the pressure axis determined from VT fault plane solutions to swing around by 90°. Again, these stress changes were interpreted as caused by ascending magma, and indeed, these episodes preceded in some cases the restart of dome growth and magma extrusion. While the methods mentioned in this section do not necessarily utilize only volcanogenic earthquakes, they employ seismological tools to monitor the stress perturbations caused by volcanic activity, and therefore, could lead to new methods of volcanic forecasting.

Summary and Conclusions

Monitoring and interpretation of volcanogenic earthquakes are core tasks of any volcano observatory. The separation into different event types allows us to attribute seismic activity to certain volcanic processes. While some event types are directly caused by brittle failure during magma ascent, or by extrusion of magma at the surface, other seismic signals provide us with indirect evidence of changes caused by the volcanic-magmatic system. Volcano seismology emerged from classic earthquake seismology as a separate discipline focusing on the special character of seismo-volcanic source processes. This development is ongoing, and classic seismic techniques such as the determination of seismic moment tensors will be applied in the future to the wider variety of seismo-volcanic sources which can be significantly different from usual earthquake sources acting on a single fault plane. With the availability of seismic broadband sensors, volcano seismology overlaps with high-frequency acoustic emissions on one side of the seismic spectrum and with deformation signals on the low-frequency end. Multidisciplinary approaches combining seismology with other signals recorded on volcanoes will lead to more detailed insights into volcanic processes.

The use of *machine learning* and *artificial intelligence* will allow us to jointly interpret increasing amounts of data in volcano observatories leading to improved risk assessment and mitigation.

Cross-References

- [Earthquake Source Theory](#)
- [Earthquake, Focal Mechanism](#)
- [Earthquakes, Location Techniques](#)
- [GPS, Tectonic Geodesy](#)
- [Magnetic Anisotropy](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Instrumentation](#)
- [Seismological Networks](#)

Bibliography

- Aspinall W, Miller A, Lynch L, Latchman J, Stewart R, White R, Power JA (1998) Soufriere Hills eruption, Montserrat, 1995–1997: volcanic earthquake locations and fault plane solutions. *Geophys Res Lett* 25:3397–3400
- Battaglia J, Ferrazzini V, Staudacher T, Aki K, Cheminee J-L (2005) Pre-eruptive migration of earthquakes at the Piton de la Fournaise volcano (Reunion island). *Geophys J Int* 161:549–558
- Brenguier F, Shapiro NM, Campillo M, Ferrazzini V, Duputel Z, Coutant O, Nercissian A (2008) Towards forecasting volcanic eruptions using seismic noise. *Nat Geosci* 1:126–130
- Chouet B (1988) Resonance of a fluid-filled crack: radiation properties and implications for the source of long-period events and harmonic tremor. *J Geophys Res* 101:4375–4400
- Chouet B (1996) Long-period volcano seismicity: its source and use in eruption forecasting. *Nature* 380:309–316
- Chouet BA, Matoza RS (2013) A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption. *J Volcanol Geotherm Res* 252:108–175. <https://doi.org/10.1016/j.jvolgeores.2012.11.013>
- Contreras-Arratia R, Neuberg JW (2019) Complex seismic sources in volcanic environments: radiation modelling and moment tensor inversions. *J Volcanol Geotherm Res* 381:262–272
- Ferrazzini V, Aki K (1987) Slow waves trapped in a fluid-filled crack: implications for volcanic tremor. *J Geophys Res* 92:9215–9223
- Gerst A, Savage MK (2004) Seismic anisotropy beneath Ruapehu volcano: a possible eruption forecasting tool. *Science* 306:1543–1547
- Gil Cruz F, Chouet B (1997) Long-period events, the most characteristic seismicity accompanying the emplacement and extrusion of a lava dome in Galeras volcano, Columbia, in 1991. *J Volcanol Geotherm Res* 77:121–158
- Goto A (1999) A new model for volcanic earthquakes at Unzen volcano: melt rupture model. *Geophys Res Lett* 26:2541–2544
- Green DN, Neuberg J (2005) Seismic and infrasonic signals associated with an unusual collapse event at the Soufriere Hills volcano Montserrat. *Geophys Res Lett* 32:L07308
- Green D, Neuberg J (2006) Waveform classification of volcanic low-frequency swarms and its implication. *J Volcanol Geotherm Res* 153:51–63
- Green DN, Neuberg J, Cayol V (2006) Shear stress along the conduit wall as a plausible source of tilt at Soufriere Hills volcano Montserrat. *Geophys Res Lett* 33:L10306
- Hammer C, Neuberg J (2009) On the dynamical behaviour of low-frequency earthquake swarms prior to a dome collapse of Soufriere Hill volcano Montserrat. *Geophys Res Lett* 36:L06305
- Hellweg M (2003) The polarization of volcanic seismic signals: medium or source? *J Volcanol Geotherm Res* 128:159–176
- Iverson RM, Dzurisin D, Gardner CA, Gerlach TM, LaHusen RG, Lisowski M, Major JJ, Malone SD, Messerich JA, Moran SC, Pallister JS, Qamar AI, Schilling SP, Vallance JW (2006) Dynamics of seismogenic volcanic extrusion at Mount St Helens in 2004–05. *Nature* 444:439–443
- Johnson JB (2003) Generation and propagation of infrasonic air waves cc from volcanic explosions. *J Volcanol Geotherm Res* 121:1–14
- Jolly AD, Thompson G, Norton GE (2002) Locating pyroclastic flows on Soufriere Hills volcano, Montserrat, West Indies, using amplitude signals from high dynamic range instruments. *J Volcanol Geotherm Res* 118:299–317
- Julian B (1994) Volcanic tremor: nonlinear excitation by fluid flow. *J Geophys Res* 99:11859–11877
- Lahr JC, Chouet BA, Stephens CD, Power JA, Page RA (1994) Earthquake classification, location and error analysis in a volcanic environment. *J Volcanol Geotherm Res* 62:137–151
- Latter JH (1981) Volcanic earthquakes and their relationship to eruptions at Ruapehu and Ngauruhoe volcanoes. *J Volcanol Geotherm Res* 9:293–309
- Miller A, Stewart R, White R, Luckett R, Baptie B, Aspinall W, Latchman J, Lynch L, Voight B (1998) Seismicity associated with dome growth and collapse, at the Soufriere Hills volcano Montserrat. *Geophys Res Lett* 25:3401–3404
- Molina I, Kumagai H, Yepes H (2004) Resonances of a volcanic conduit triggered by repetitive injections of ash-laden gas. *Geophys Res Lett* 31:L03603
- Neuberg J, Luckett R (1996) Seismo-volcanic sources on Stromboli volcano. *Ann Geofis* 39:377–391
- Neuberg J, Luckett R, Ripepe M, Braun T (1994) Highlights from a seismic broadband array on Stromboli volcano. *Geophys Res Lett* 21:749–752
- Neuberg J, Baptie B, Luckett R, Stewart R (1998) Results from the broad-band seismic network on Montserrat. *Geophys Res Lett* 25:3661–3664
- Neuberg J, Luckett R, Baptie B, Olson KB (2000) Models for tremor and low-frequency events on Montserrat. *J Volcanol Geotherm Res* 101:83–104
- Neuberg JW, Tuffen H, Collier L, Green D, Powell T, Dingwell D (2006) The trigger mechanism of low-frequency earthquakes on Montserrat. *J Volcanol Geotherm Res* 153:37–50
- Neuberg JW, Collinson ASD, Mothes PA, Ruiz MC, Aguaiza S (2018) Understanding cyclic seismicity and ground deformation patterns at volcanoes: intriguing lessons from Tungurahua volcano, Ecuador. *Earth Planet Sci Lett* 482:193–200
- Ripepe M, Harris AJL, Carniel R (2002) Thermal, seismic and infrasonic evidences of variable degassing rates at Stromboli volcano. *J Volcanol Geotherm Res* 118:285–297
- Roman DC, Neuberg J, Luckett R (2006) Assessing the likelihood of volcanic eruption through analysis of volcano-tectonic earthquake fault-plane solutions. *Earth Planet Sci Lett* 248:229–237
- Taisne B, Brenguier F, Shapiro NM, Ferrazzini V (2011) Imaging the dynamics of magma propagation using radiated seismic intensity. *Geophys Res Lett* 38:2–6. <https://doi.org/10.1029/2010GL046068>
- Voight B, Hoblitt R, Clarke A, Lockhart A, Miller A, Lynch L, McMahon J (1998) Remarkable cyclic ground deformation monitored in real time on Montserrat, and its use in eruption forecasting. *Geophys Res Lett* 25:3405–3408
- White R, McCausland W (2016) Volcano-tectonic earthquakes: a new tool for estimating intrusive volumes and forecasting eruptions. *J Volcanol Geotherm Res* 309:139–155
- Wielandt E, Forbriger T (1999) Near-field seismic displacement and tilt associated with the explosive activity of Stromboli. *Ann Geofis* 42:407–416
- Zimanowski B (1998) Phreato-magmatic explosions. In: Freundt A, Rosi M (eds) From magma to tephra. Elsevier, Amsterdam, pp 25–53

Electrical Properties of Rocks

Takashi Yoshino

Institute for Planetary Materials, Okayama University,
Misasa, Japan

Definition

Electrical conduction	The movement of electron or electrically charged particles through a medium.
Electrolyte	Any substance containing free ions that make the substance electrically conductive. The most typical electrolyte is an ionic solution.
Fermi level	The highest energy of occupied states in a solid at 0 K. The Fermi level in conductors lies in the conduction band, in insulators it lies in the valence band, and in semiconductors it falls in the gap between the conduction band and the valence band.
Stern layer	One of the two electrically charged layers of electrolyte ions: the layer of ions immediately adjacent to the surface and the layer in the neighborhood of a negatively charged surface.

Introduction

The interpretation of electromagnetic data obtained from the electromagnetic (EM) induction method requires knowledge of electrical properties of rocks comprising the Earth's interior. The electrical properties of rock are now widely used in electrical conductivity (resistivity) and induced polarization to identify zones of ore deposits and to probe the thermal and compositional structure in the Earth's interior. For example, electrical resistivity measurements have been firstly used in 1960s to prospect the sulfide ore deposits. Electrical conductivity of the crust and mantle by the electromagnetic (EM) induction method represents an important contribution to our knowledge of processes in the Earth's interior independently of the seismological studies, because the electrical conductivity is sensitive to small amounts of accessory phases such as fluid, melt and metal in rocks, and small amounts of hydrogen and ferric iron in rock-forming minerals.

The bulk electrical properties of rocks contain complex combination of various electrical processes consisting of solids (silicates, oxides, and metal), pores filled with air or liquid, and interfaces between solid and liquid, because rocks are composites of wide variety of minerals occasionally containing fluid and melt. Since 1960s, many studies of electrical properties in rocks and rock-forming minerals have been performed as functions of frequency, temperature,

pressure, oxygen fugacity, water content, applied electric field and other variables. Recent development of analytical technique and an improvement of resolution for the electromagnetic observations has led to an increase of interest in the electrical properties of rocks at depth within the Earth. At such depths, pressures and temperatures are very high, and laboratory studies in which these pressures and temperatures can be realized have been needed to predict the electrical properties at high pressures corresponding to the depth of the Earth's interior. Within the last two decades, many efforts were concentrated on the measurement of high-pressure phases to obtain knowledge of temperature, composition, and water content in the Earth's interior. As a result, the maximum range of pressure where the electric conductance can be measured currently covers the pressure condition of the core-mantle boundary (2900 km depth).

Definition of Electrical Conductivity (Resistivity)

When an electric potential V is applied across a material, electric current (I) in a material is proportional to the voltage across it according to Ohm's law: $V = IR$, where R is the electrical resistance, which depends on the intrinsic resistivity ρ of the material and on the geometry (length l and area A through which the current passes).

$$R = \frac{\rho l}{A}$$

The electrical conductivity (σ) is the inverse of the resistivity. The current density (J) can be expressed as $J = E/\rho = E\sigma$. The conductivity is one of the properties of materials that varies most widely, from 10^7 (S/m) typical of metals to 10^{-20} (S/m) for good electrical insulators. Semiconductors have conductivities in the range 10^{-6} – 10^4 (S/m).

Electrical Conduction

An electrolytic process can be considered as a dominant mechanism for electron charge transport in shallower crustal rocks developing pores and cracks that filled with fluid electrolytes. At high saturation conditions, conducting currents in electrolyte primarily arise from the flow of ions in electric field. The conductivity of an electrolyte is controlled by solute concentration, ionic charge, and other physical parameters such as temperature and pressure. Fluid saturated rocks contain interface between mineral and fluid. Surface conductivity can be considered as an additional electrical conductivity of fluid in the vicinity of the charged surface. Concentration of ions becomes higher close to the charged surfaces because of attractive electrostatic forces induced by the surface

charge. This layer with high concentration is called as the electrochemical double layer (DL) composed of Stern layer in contact with solid and diffusive layer in contact with fluid. Higher concentration of ions in the Stern layer leads to higher conductivity. The significant influence of surface conductivity on bulk electrical properties appears for porous samples saturated with dilute electrolytes and low porous samples where the thickness of the DL is the same order as crack width. A mixed conduction process of electrolyte and surface conductivity is most likely at low saturation conditions.

As pressure and temperature increase, electrical conduction in rock-forming mineral is likely to control the bulk electrical conductivity of rocks by reduction of open space and interface effect in rocks. Solids can be classified in terms of their conductivity as conductors, semiconductors, or insulators. The classifications of solid materials can be visualized in terms of the band theory and are strongly correlated with the interatomic spacing in the solid. The electron energy levels in a solid are often expressed in relation to the Fermi energy. In metals, the electrons occupy states up to the Fermi level. Conduction occurs by promoting electrons into the conduction band that starts at the Fermi level. The promotion energy is negligibly small so that at any temperature electrons can be found in the conduction band. The number of electrons participating in electrical conduction is extremely small. In semiconductors and insulators, the valence band is filled, and no more electrons can be added, following Pauli's principle. In insulators, there is a large energy gap between the valence and conduction bands, so huge energy is needed to promote an electron to the conduction band. Semiconductors can be divided into intrinsic or extrinsic. Intrinsic conduction means that electrical conductivity does not depend on impurities, thus intrinsic means pure. In extrinsic semiconductors, the conductivity depends on the concentration of impurities. Electronic conduction occurs by electrons and electron holes. The electrical conductivity of semiconductors can be enhanced by doping to produce *n* (negative)-type or *p* (positive)-type semiconductors.

Silicates and oxides are usually insulator at ordinary temperature. Since the large energy gap of insulator could be overcome by heat, the silicate minerals can behave as a semiconductor at high temperatures corresponding to conditions of the Earth's interior. Mobility of lattice defects and a small amount of impurity such as hydrogen, ferric iron, and so on enhances electrical conduction of silicate minerals at high temperatures. The conductivity of ionic-bonded crystals is proportional to number and mobility of charge carriers. It is known from the Nernst-Einstein equation that electrical conductivity (σ) depends on the number (N) of electric charge carriers per unit volume;

$$\sigma = Nze\mu$$

where z is the charge number, e is the charge of electron, and μ is mobility.

The dependence of conductivity on temperature is like other thermally activated processes:

$$\sigma = A \exp(-\Delta H/2kT)$$

where A is a pre-exponential constant, ΔH is an activation enthalpy, k is the Boltzmann constant, and T is absolute temperature. Plotting log normal σ vs. reciprocal temperature ($1/T$) produces a straight line of slope $\Delta H/2k$ from which the band gap energy can be determined. Extrinsic semiconductors have, in addition to this dependence, dependence due to the thermal promotion of electrons from donor levels or holes from acceptor levels. The dependence on temperature is also exponential but it eventually saturates at high temperatures where all the donors are emptied or all the acceptors are filled. This means that at low temperatures, extrinsic semiconductors have larger conductivity than intrinsic semiconductors. At high temperatures, both the impurity levels and valence electrons are ionized, but since the impurities are very low in number and they are exhausted, eventually the behavior is dominated by the intrinsic type of conductivity. The activation enthalpy (ΔH) can be also defined as a function of pressure (P) as follows: $\Delta H = \Delta E + P\Delta V$, where ΔE is the activation energy and ΔV is the activation volume.

Ionic (intrinsic) conduction can occur at high temperatures through the creation of cation vacancies. In the ferromagnesian mantle minerals (silicates and oxides), the charge carriers are generally vacancies in magnesium site. The activation enthalpy of ionic conduction is usually higher than 2 eV. For this reason, extremely high temperatures are needed to observe the ionic conduction. Proton conduction is a kind of ion conduction. Because the proton is overwhelmingly small ion but massive compared with electron, the proton is the only ion which is expected to be highly mobile in solids. The energy barrier for the transportation of charge carriers for proton conduction is relatively low because of its small bonding energy and small ionic radius. The activation enthalpy of proton conduction is usually less than 1 eV and tends to decrease with increasing a number of hydrogen in minerals at very low concentrations. The proton conduction is a dominant conduction mechanism at low temperatures, whereas it has been known that the presence of hydrogen in crystal structure can occasionally enhance the ionic conduction.

The most dominant conduction mechanism for the ferromagnesian minerals such as olivine is hopping (small polaron) conduction. Hopping conduction occurs through charge transfer between neighboring ions of different valence. A ferrous ion (Fe^{2+}) generally substitutes an Mg ion in an Mg site. However, a certain proportion of ferric ion, which has an electron hole, exists dependent on oxygen fugacity, temperature, pressure, and crystal structure. Transfer of an electron hole between ferrous and ferric iron sites carries an electric charge, which is the hopping conduction. If only electron holes migrate in crystal, the energy barriers for the migration

would be low. However, presence of an electric hole significantly affects the local structure of the ionic crystal. A presence of ferric iron generates an extra positive charge, which repulses cations and attracts anions. This complex of the local strains is called “small polaron.” The migration of electron holes associates that of the small polaron. The relatively large activation enthalpy is needed to move a small polaron. The hopping conduction of usual ferromagnesian silicates has relatively large activation enthalpy (>1 eV). Activation energy for hopping conduction decreases with increasing iron concentration. The conductivity of ferromagnesian minerals is also controlled by the oxidation state of iron. A variation of the oxygen fugacity from an oxidizing to a reducing environment decreases the conductivity, due to the reduction of Fe^{3+} to Fe^{2+} . At the same oxygen buffer, the conductivity tends to increase with increasing iron content in minerals.

Electrical Conductivity of Liquid-Bearing Rocks

The effective conductivity of liquid-bearing rocks is very sensitive to the pore geometry in rocks. The enormous contrast in conductivity between silicates and liquids implies that electrical currents will flow almost totally through the pore liquid within rocks. Archie (1942) proposed an empirical equation that relates the bulk conductivity of σ_r to the conductivity of a pore saturating fluid σ_f . The rocks were assumed to exhibit interconnected porosity, and no disturbance due to the presence of reactive minerals was considered. The ratio σ_f/σ_r is known as the formation factor, F . Archie's law forms the basis for resistivity log interpretation:

$$F = \frac{\sigma_f}{\sigma_r} = C\phi^{-n},$$

where C and n are constants. Archie's law is not a geometrical model. There are various theoretical models for estimating the bulk conductivity of liquid-bearing rocks from the conductivity of liquid and solid phases. A frequently used model to predict the maximum effective bulk conductivity of an interconnected two-phase mixture is Hashin and Shtrikman (HS) upper bound (Hashin and Shtrikman 1963). This upper bound is representative for liquid phases distributed along grain boundaries and filling triple junctions of spherical grains. The opposite case representing the isolated spheres of liquids in solid grains is known as the HS lower bound.

Electrical Conductivity of Crust

The shallower zone of the upper crust consists of a sequence of sedimentary rocks along with fractured crystalline rocks. Because these rocks contain pores and cracks that are filled

with fluid electrolytes, they are moderately good conductors of electricity. As well as electrolyte conductivity, surface conductivity caused by electrochemical reactions between the fluid electrolyte and the mineral surface contribute additionally to the bulk conductivity. This zone has conductivities varying from about $10^{-0.5}$ S/m in recent sediments to 10^{-3} S/m or more in weathered crystalline rock. As pressure increases, the rock conductivity decreases due to the reduction of porosity. Remarkable changes in electrical properties often occur at mid-crustal depth, probably caused by the development of incipient fractures. If the fractures are filled with fluid electrolytes, electrical conductivity would largely enhance compared to the surrounding dry crystalline rocks. Secular changes in electrical conductivity might be used in predicting earthquakes.

The lower crustal rocks mainly consist of crystalline igneous or metamorphic rocks which are thought to be much more dense, having little pore space in which water may collect. Since the main rock-forming minerals are highly insulating at crustal temperatures, electrical conductivity in such rocks should be expected to be lower values than the upper crust. However, magnetotelluric (MT) and deep geomagnetic soundings (GDS) have revealed an unexpected high conductivity anomaly in the lower crust. A very small portion of well conducting material may cause enhanced conductivity of the whole rock mass provided that material fills interconnected pore space or covers grain surfaces. To generate the conductive layer, an indispensable interconnectivity of the well conducting phase must be established and maintained over large distances under deep crustal conditions.

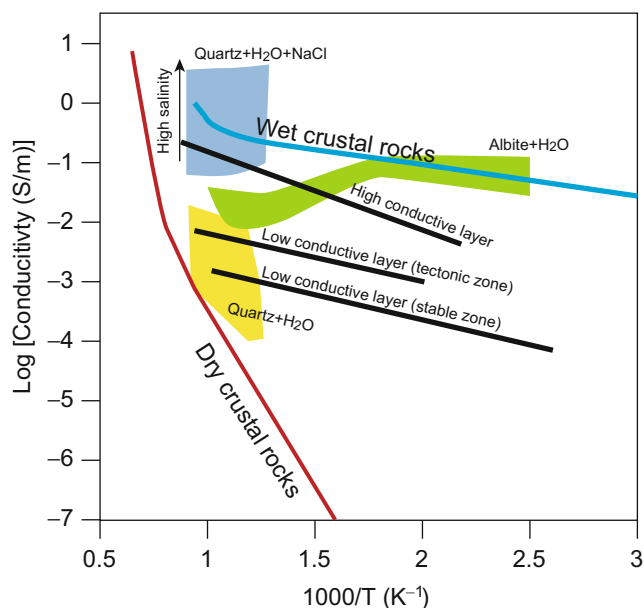
Geophysicists traditionally considered aqueous fluids as the likely origin (Shankland and Ander 1983), because in active tectonic environments, aqueous fluid can supply via metamorphic dehydration, igneous activity, and introduction of fluids derived from dehydration of subducting plate. Figure 1 shows conductivity-temperature curves of conductivity layers in crust compared with conductivity measurement results of fluid-bearing rocks. Laboratory conductivity measurements suggest that the electrical conductivity of the aqueous fluid-bearing quartzite was found to be more than a few orders of magnitude higher than that of dry crustal rocks, but the conductivity is definitely lower than the observed values of the high-conductivity anomalies. The conductivity of the brine-bearing quartzite increases with salinity and fluid fraction, but slightly increases with increasing temperature. Unlike brine-free system, the small decrease of the conductivity with temperature is quite consistent with the temperature dependence of the specific conductivity for NaCl solution. Crustal rocks with fluid of high salinity (>10 wt%) is needed to account for the high conductive regions with the conductivity of 10^{-1} S/m. There is a small negative dependence of bulk conductivity on aqueous fluid fraction. Because the mobility of dissolved ions has little temperature dependence, the bulk conductivity maintains high values even at low temperatures.

On the other hand, in stable cratons fluids would be gravitationally driven but would be able to gain access only to the upper 10 km or so of the crust. Below this depth, the crust would be “dry.” The discovery of grain-boundary films on minerals in high-grade rocks (Frost et al. 1989) has led to a very reasonable alternative to aqueous fluids as a cause of conductivity anomalies. Duba and Shankland (1982) used the upper bound formula to calculate the absolute amount of well-ordered graphite required to enhance the conductivity of dry rocks. They found that a volume fraction of carbon of only $5 \times 10^{-6}\%$ could cause a conductivity of 0.1 S/m if well interconnected. However, high wetting angles of CO_2 – quartz system cannot establish along grain boundaries to form a connecting graphite film upon reduction of CO_2 . In addition, graphite films on the grain boundary of silicate minerals are readily destroyed by high interfacial energy (Yoshino and Noritake 2011). The origin of high conductivity anomalies observed in the lower crust where aqueous fluid is absent is still an unsolved problem.

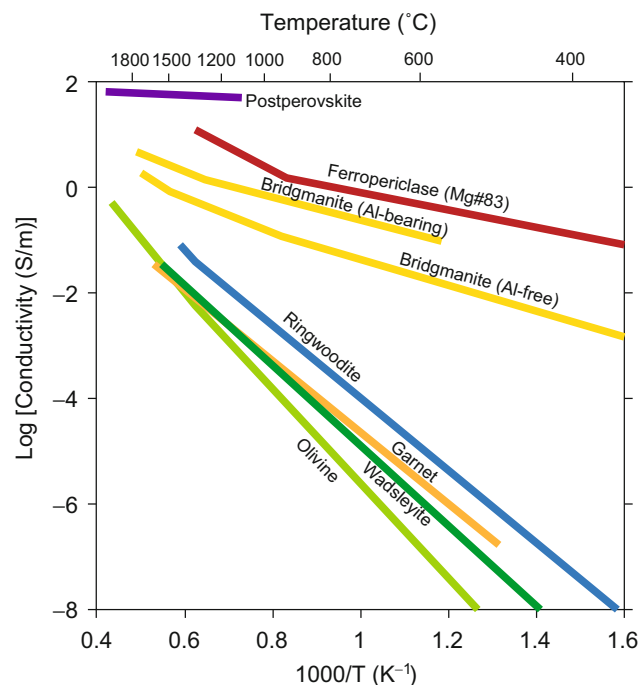
Electrical Conductivity of the Earth’s Mantle

Electrical conductivity of mantle rocks tend to be similar to that of the most dominant phase, unless small amounts of conductive phase establish interconnection. Conductivity generally increases with increasing depth to the top of the

lower mantle, from 10^{-3} – 10^{-2} to 10^0 S/m. This increase is considered to be the semiconductive behavior of the constituent minerals at high temperatures because mantle temperature adiabatically increases with depth, and by pressure-induced phase transformations of dominant mantle minerals. Figure 2 shows the electrical conductivity of the main constituent mantle minerals in the Arrhenius plot. The mantle minerals with higher-pressure stability field tend to have higher conductivity in order of olivine, wadsleyite, ringwoodite, bridgmanite, and post-perovskite. A phase transformation affects electrical properties due to a change in the atomic arrangement (crystal structure), as well as the nature of chemical bonding. The phase transitions would yield the conductivity jumps at some certain depths where seismic velocity suddenly changes. Thus, the electrical conductivity measurements of the main mantle minerals to depths of up to the core-mantle boundary indicate that the Earth’s mantle is zoned electrically. This tendency can explain the geophysically observed electrical conductivity structure in the mantle well. In the lower mantle, there is no phase transition down to the D” seismic discontinuity. Because activation enthalpy for small polaron and ionic conduction is smaller for mantle minerals with higher-pressure stability field, as pressure increases, temperature dependence becomes smaller. For this reason, a uniform electrical conductivity is expected to exist in the lower mantle. On the other hand, the D” layer composed of post-perovskite might be an abnormally high



Electrical Properties of Rocks, Fig. 1 Field conductivity-temperature curves of crust (solid lines) compared with laboratory measurement of conductivity in “dry” granite and “wet” granite with 1–2% water (modified after Shankland and Ander 1983). Also shown are laboratory measurements of conductivity in fluid-bearing crustal rocks (Shimojuku et al. 2014; Guo et al. 2015)



Electrical Properties of Rocks, Fig. 2 Electrical conductivity of mantle minerals with $\text{Mg\#} = 90$; olivine, wadsleyite, ringwoodite, bridgmanite, ferro-periclase ($\text{Mg\#} = 83$), and post-perovskite in Arrhenius plot modified after Yoshino (2010)

conductive layer. Thus there should be huge conductivity jump at the D'' discontinuity.

Origin of High Conductive Anomalies in the Mantle

High-conductivity anomalies in the mantle have been observed and predicted in various depths and would be caused by thermal anomaly, relatively high concentrations of water, and iron in mantle minerals or a presence of fluid or partial melt. The current understandings of their conductivity anomalies are summarized here.

Oceanic Asthenosphere

The high conductive layer with anisotropic signature at the top of the asthenosphere (60–100 km depth) was observed near the Eastern Pacific Rise and cannot be explained by the conductivity characteristics of ordinary anhydrous olivine (Evans et al. 2005). One possibility is that partial melting occurs in the region and the melt forms an anisotropic, interconnected conductive path in the partially molten peridotites. Another possibility is that olivine in the region contains considerable amounts of hydrogen inside crystal structures, and thus electrical conductivity is substantially elevated by proton conduction (Karato 1990). Wang et al. (2006) reported large effect of water on electrical conductivity of olivine, and concluded that 10^{-2} wt.% water is enough to explain high conductivity anomaly of oceanic asthenosphere. On the other hand, Yoshino et al. (2006) reported that the electrical conductivity of single crystal hydrous olivine is nearly isotropic and distinctly lower than the observed conductivity. The majority of subsequent studies did not support the contribution of hydrous olivine to the conductivity. Thus hydration of olivine cannot account for the geophysical observations, which instead may be explained by the presence of partial melt with basaltic composition elongated in the direction of plate motion. The experimental results demonstrate that alignment of 2 vol% of the melt in peridotite causes conductivity anisotropy by one order of magnitude higher in the direction parallel to the shear compared with normal to the shear on the shear plane (Zhang et al. 2014). The expected melt fraction in the regions where conductivity anisotropy is observed is only up to 0.2 vol%. Because electrical conductivities of hydrous and carbonate melts are more than a few orders of magnitude higher than those of hydrated olivine, high conductivities in the asthenosphere may be explained by the presence of very small amounts of hydrous carbonate melt in peridotite (Gaillard et al. 2008; Sifré et al. 2014). However, recent studies have pointed out that the effect of temperature dependence on the electrical conductivity of hydrous olivine becomes larger at higher temperatures because of a change of conduction mechanism in hydrous olivine. The debate on whether the origin of the conductivity anomaly is water or melt has continued.

Mantle Transition Zone

The Earth's mantle transition zone (410–660 km depth) has been considered as a large water storage area because wadsleyite and ringwoodite, which are high-pressure polymorphs of olivine, can enter significant amounts of water in their crystal structures. If these phases are dry, their conductivity values are too low to explain the conductivity-depth profiles in the transition zone obtained by the semi-global electromagnetic induction studies. Therefore, effect of water on electrical conductivity was investigated to determine the water content in this zone. Water content on electrical conductivity of their minerals was estimated to be 0.1 wt.% from the average conductivity profile in the mantle transition zone (Yoshino and Katsura 2013). On the other hand, the conductivity anomalies observed in the mantle wedge require high water content over the solubility limit. Such high conductivity values suggest the presence of supercritical fluids in this region, leading to a persistence of local water saturation and continuous release of supercritical fluids at the stagnant slab roof on geological time scales. This phenomenon can account for both the high-conductivity and seismic-velocity anomalies observed in the mantle wedge at the transition-zone depth.

The Base of the Lower Mantle

Perturbations in the Earth's rotation rate at decadal time-periods require the existence of dissipative coupling at the CMB, leading to the presence of a thin conductive layer with a conductance of 10^8 S or greater. Several models have been proposed to explain anomalously high electrical conductance in some regions above the CMB: (1) capillary rise of iron-rich melt into the mantle; (2) a suction mechanism for iron entrainment into the mantle; (3) penetration of molten iron into the mantle by morphological instability; (4) presence of post-perovskite or extremely Fe-rich (Mg,Fe)O with high electrical conductivity. The first three requires at least a 200-m thick metallic solid layer at the top of the core. The conductivity data of post-perovskite showed extremely high conductivity ($>10^2$ S/m) in comparison with that of bridgmanite, and little temperature dependence at the conditions of D'' layer (Ohta et al. 2008). A ~300-km-thick post-perovskite layer provides an alternative explanation.

Conclusions

The electrical conductivity of rocks within the Earth shows an ultimate variation that exceeds eight orders of magnitude. The electrical conductivity measurements of rocks as a function of temperature, pressure, oxygen fugacity, and impurity concentrations can constrain thermal and compositional structure in the solid Earth. At shallower depths, electrolytes filled with cracks or grain boundaries control the conductivity of rocks,

whereas the conductivity of the mantle rocks is mainly controlled by semiconductive behavior of the constituent minerals. In the last two decades, development of measurement techniques for electrical properties of rocks and minerals under high-pressure and high-temperature conditions has improved our understanding of the electrical structure of the Earth's deep interior up to the core-mantle boundary. Although some discrepancies of data among laboratories have remained due to the difficulty of the electrical conductivity measurement at high pressures, it is not doubtful that the Earth's mantle is electrically zoned, and the conductivity generally increases with depth. Further improvement of experimental techniques will provide us accurate reference conductivity-depth model up to depth of the core-mantle boundary and more detailed knowledge of chemical composition and thermal distribution in the present Earth's interior by comparison with the results of the electromagnetic surveys.

Cross-References

- [Earth's Structure, Lower Mantle](#)
- [Earth's Structure, Upper Mantle](#)
- [Magnetotelluric Interpretation](#)

Bibliography

- Archie GE (1942) The electrical resistivity log as an aid in determining some reservoir characteristics. *Trans Am Inst Min Metall Pet Eng* 146:54
- Cole KS, Cole RH (1941) Dispersion and absorption in dielectrics I. Alternating current field. *J Chem Phys* 1:341
- Duba AG, Shankland TJ (1982) Free carbon and electrical-conductivity in the Earth's mantle. *Geophys Res Lett* 9:1271
- Evans RL, Hirth G, Baba K, Forsyth D, Chave A, Mackie R (2005) Geophysical evidence from the MELT area for compositional controls on oceanic plates. *Nature* 437:249
- Frost BR, Fyfe WS, Tazaki K, Chan T (1989) Grain-boundary graphite in rocks and implications for high electrical conductivity in the lower crust. *Nature* 340:134
- Gaillard F, Malki M, Iacono-Marziano G, Pichavant M, Scailliet B (2008) Carbonatite melts and electrical conductivity in the asthenosphere. *Science* 322:1363
- Guo X, Yoshino T, Shimojuku A (2015) Electrical conductivity of albite-(quartz)-water and albite-water-NaCl systems and its implication to the high conductivity anomalies in the continental crust. *Earth Planet Sci Lett* 412:1
- Hashin Z, Shtrikman A (1963) A variational approach to the theory of elastic behaviour of multiphase materials. *J Mech Phys Solids* 11:127
- Karato S (1990) The role of hydrogen in the electrical conductivity of the upper mantle. *Nature* 347:272
- Ohta K, Onoda S, Hirose K, Shinmyo R, Shimizu K, Sata K, Ohishi Y, Yasuhara A (2008) The electrical conductivity of post-perovskite in earth's D layer. *Science* 320:89
- Shankland TJ, Ander ME (1983) Electrical conductivity, temperature, and fluids in the lower crust. *J Geophys Res* 88:9475
- Shimojuku A, Yoshino T, Yamazaki D (2014) Electrical conductivity of brine-bearing quartzite at 1 GPa: implications for fluid content and salinity of the crust. *Earth Planets Space* 66:2
- Sifré D, Gardés E, Massuyeau M, Hashim L, Hier-Majumder S, Gaillard F (2014) Electrical conductivity during incipient melting in the oceanic low-velocity zone. *Nature* 509:81
- Wang D, Mookherjee M, Xu Y, Karato S (2006) The effect of water on electrical conductivity of olivine. *Nature* 443:977
- Yoshino T (2010) Laboratory electrical conductivity measurement of mantle minerals. *Surv Geophys* 31:163
- Yoshino T, Katsura T (2013) Electrical conductivity of mantle minerals: role of water in conductivity anomalies. *Annu Rev Earth Planet Sci* 41:605
- Yoshino T, Noritake F (2011) Unstable graphite films on grain boundaries in crustal rocks. *Earth Planet Sci Lett* 306:186
- Yoshino T, Matsuzaki T, Yamashita S, Katsura T (2006) Hydrous olivine unable to account for conductivity anomaly at the top of the asthenosphere. *Nature* 443:973
- Zhang B, Yoshino T, Yamazaki D, Manthilake G, Katsura T (2014) Electrical conductivity anisotropy in partially molten peridotite under shear deformation. *Earth Planet Sci Lett* 405:98

Electrical Resistivity Surveys and Data Interpretation

- M. H. Loke¹, D. F. Rucker², J. E. Chambers³,
P. B. Wilkinson³ and O. Kuras³
¹Geotomo Software Sdn. Bhd., Penang, Gelugor, Malaysia
²hydroGEOPHYSICS, Inc., Tucson, AZ, USA
³Geophysical Tomography Team, British Geological Survey, Keyworth, Nottingham, UK

Synonyms

Ω.m, ohm.meter; 1-D, one-dimensional; 2-D, two-dimensional; 3-D, three-dimensional

Definition

Electrical survey – Mapping the subsurface resistivity by injecting an electric current into the ground

Electrode – Commonly a metal rod through which the current is injected into the ground or is used to measure the induced voltage on the ground surface

Least-squares resistivity inversion – Finding the subsurface resistivity model that minimizes the sum of squares of the differences between the measured and calculated apparent resistivity values

Introduction

Electrical resistivity surveys map the subsurface structures by making electrical measurements near the ground surface. An electric current is injected into the ground through two

electrodes, and the voltage difference is measured between two other electrodes (Fig. 1a). The true subsurface resistivity can be estimated by making the measurements of potential difference at different positions of the current and potential electrodes, converting these values into apparent resistivity and then inverting the data set. The ground resistivity is related to various geological parameters such as the mineral and fluid content, porosity, and degree of water saturation in the rock (*electrical properties of rocks*). Over the past 25 years, the resistivity method has undergone rapid developments in instrumentation, field survey techniques, data interpretation, and novel applications in the geosciences. It has now become one of the standard geophysical exploration techniques widely used in environmental, engineering, hydrogeological, and mining investigations (Loke et al. 2013).

The basic data from a resistivity survey are the positions of the current and potential electrodes, the current (I) injected into the ground, and the resulting voltage difference (ΔV) measured between the potential electrodes (Fig. 1a). The current and voltage measurements are then converted into apparent resistivity (*instrumentation, electrical resistivity*), ρ_a , by using the following formula:

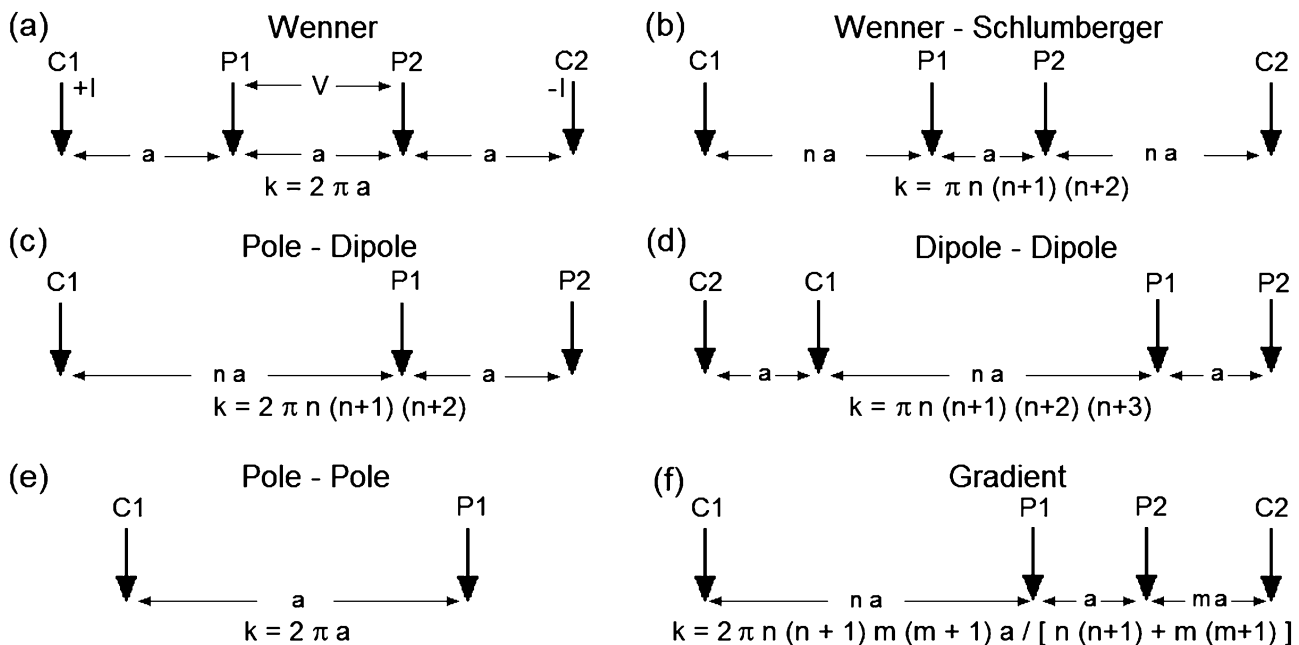
$$\rho_a = k \frac{\Delta V}{I}, \quad (1)$$

where k is the geometric factor that depends on the configuration of the current and potential electrodes (Koefoed 1979).

Over the years, various electrode configurations (or arrays) have been developed. Figure 1 shows the arrangements for some commonly used arrays. A discussion on the merits of the different arrays is given by Dahlin and Zhou (2004), Szalai and Szarka (2008), and Martorana et al. (2017). There has also been substantial work in nontraditional arrays that maximize resolution of the subsurface (Wilkinson et al. 2012) or maximize the use of unused channels on multi-channel resistivity meters (Cubbage et al. 2017). Determining the true subsurface resistivity from the apparent resistivity values is the data inversion problem.

Traditional Profiling and Sounding Surveys

The resistivity survey method has its origin in the 1910s due to the work of the Schlumberger brothers. Traditionally resistivity surveys were divided into profiling and sounding surveys. The distances between the electrodes are kept fixed in a profiling survey, and the four electrodes are moved along the survey line. The data interpretation for profiling surveys was mainly qualitative. The second type of survey is the vertical sounding method (usually with the Schlumberger array) where the center point of the electrode array remains fixed, but the spacing between the electrodes is increased to obtain deeper subsurface information. Apparent resistivity plotted as a function of the current electrode spacing gives information on subsurface resistivity variations (Fig. 2a).

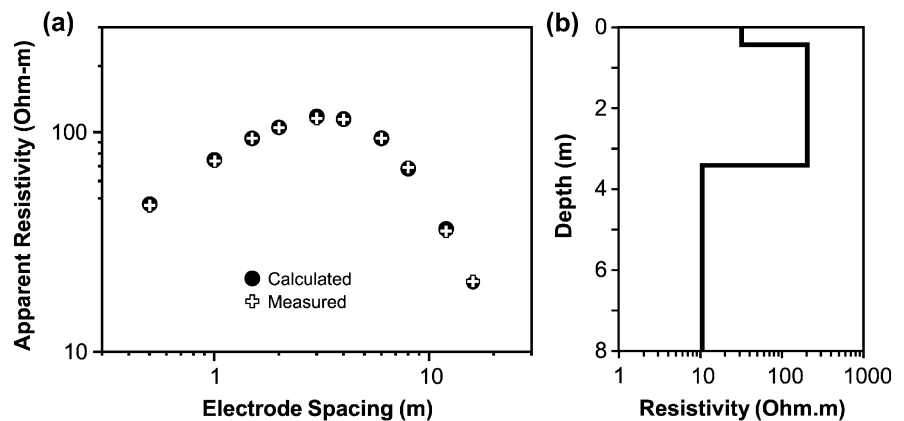


Electrical Resistivity Surveys and Data Interpretation, Fig. 1 Common arrays used in resistivity surveys and their geometric factors. The current is injected into the ground through the C1 and C2 electrodes, while the voltage difference is measured between the P1 and

P2 electrodes. For the arrays with fewer than four electrodes, the remaining electrodes are placed at sufficiently large distances so that they do not affect the measurements

Electrical Resistivity Surveys and Data Interpretation,

Fig. 2 Example of (a) 1-D sounding curve and (b) layered earth model



Quantitative data interpretation for sounding surveys uses a one-dimensional (1-D) earth model with a series of horizontal layers (Fig. 2b). One commonly used method for 1-D data inversion is the iterative damped least-squares method (*inverse theory, linear*) (Marquardt 1963; Inman 1975; Koefoed 1979) that is based on the following equation:

$$[\mathbf{J}_i^T \mathbf{J}_i + \lambda_i^2 \mathbf{I}] \Delta \mathbf{q}_i = \mathbf{J}_i^T \mathbf{g}_i, \quad (2)$$

where the data misfit vector $\Delta \mathbf{g}$ contains the difference between the logarithms of the measured and the calculated apparent resistivity values and $\Delta \mathbf{q}$ is a vector consisting of the deviation of the estimated model parameters from the true model. Here, the model parameters are the logarithms of the resistivity and thickness of the model layers. $\mathbf{J}$ is the Jacobian matrix of partial derivatives of apparent resistivity with respect to the model parameters. λ is a regularization factor that stabilizes the ill-conditioned Jacobian matrix usually encountered for geophysical problems (*inverse theory, singular value decomposition*). i represents the iteration number. Starting from an initial model (such as a homogeneous earth model), this method iteratively refines the model so as to reduce the data misfit (Fig. 2a) to a desired level. A detailed description on the use of the iterative damped least-squares method for the inversion of resistivity sounding data is given by Inman (1975).

The resistivity sounding method has been used for many years, particularly in groundwater exploration. The greatest limitation of this method is that it does not take into account lateral changes in the subsurface resistivity which can result in errors in the interpreted resistivity and thicknesses of the layers. A two-dimensional (2-D) survey and interpretation model is required for more accurate results.

Two-Dimensional Resistivity Imaging Surveys

Since the early 1990s, development of multielectrode resistivity meter system (*instrumentation, electrical resistivity*)

has made 2-D surveys a practical tool for mapping moderately complex geological environments. The apparent resistivity measurements from the survey are commonly plotted in the form of a pseudosection (Loke et al. 2013). The pseudosection (Fig. 3a) is a useful method to present the data in a pictorial form, but it gives a distorted picture of the subsurface because the shapes of the contours depend on the type of array used as well as the true subsurface resistivity.

A 2-D model that consists of a large number of cells is commonly used to interpret the data (Loke et al. 2013). The finite-difference (*numerical method, finite difference*) or finite-element methods (*numerical method, finite element*) are used to calculate the apparent resistivity values for the 2-D model. A nonlinear optimization method is then used to automatically change the resistivity of the model cells, so as to reduce the data misfit. Smoothness or other constraints are usually incorporated to stabilize the inversion procedure such that numerical artifacts are avoided. As an example, the following equation includes a model smoothness constraint to the least-squares optimization method:

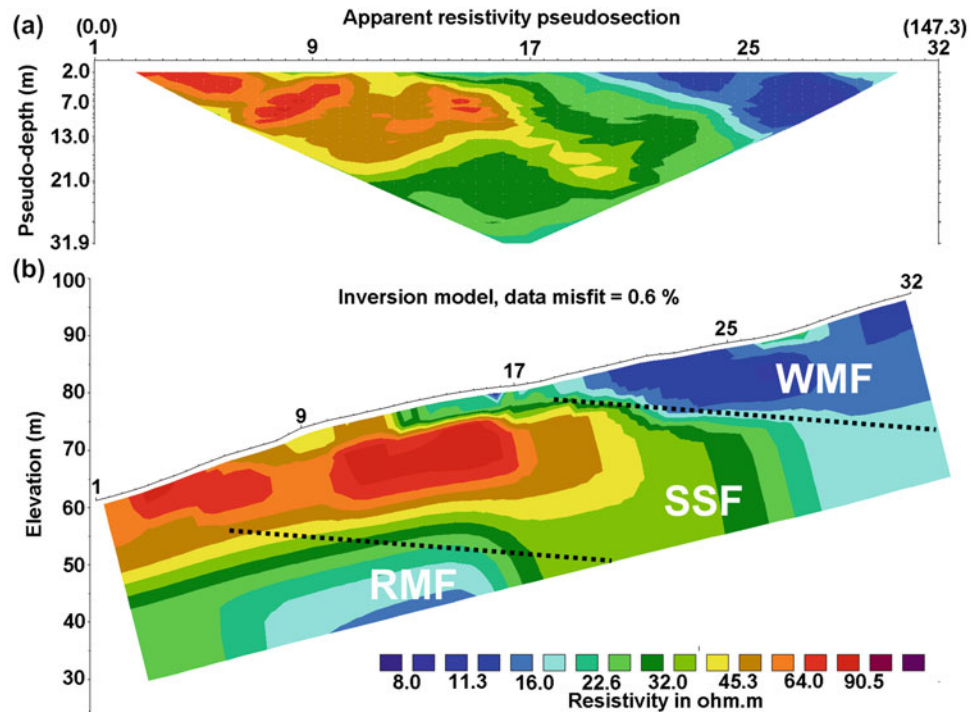
$$[\mathbf{J}_i^T \mathbf{R}_d \mathbf{J}_i + \lambda_i^2 \mathbf{W}^T \mathbf{R}_m \mathbf{W}] \Delta \mathbf{q}_i = \mathbf{J}_i^T \mathbf{R}_d \mathbf{g}_i - \lambda_i^2 \mathbf{W}^T \mathbf{R}_m \mathbf{W} \mathbf{q}_{i-1}. \quad (3)$$

$\mathbf{W}$ includes the roughness filters in the x-, y-, and z-directions (deGroot-Hedlin and Constable 1990). $\mathbf{R}_d$ and $\mathbf{R}_m$ are weighting matrices employed if the L1-norm inversion method is used (Farquharson and Oldenburg 1998). A number of computer software packages are now available that can automatically carry out the inversion of a 2-D survey data set in seconds.

Figure 3a shows an example of a measured apparent resistivity pseudosection from a survey using the dipole-dipole array. The resistivity model after inversion is shown in Fig. 3b. This survey was part of a long-term landslide monitoring program. From the base to the top of the slope, the formations are the Lias Group Redcar Mudstone Formation (RMF), Staithes Sandstone and Cleveland Ironstone

Electrical Resistivity Surveys and Data Interpretation,

Fig. 3 Example of an (a) apparent resistivity pseudosection and (b) 2-D inverse model



Formation (SSF), and Whitby Mudstone Formation (WMF). Slope failure occurs within the WMF in the upper parts of the slope (Uhlemann et al. 2017).

Three-Dimensional Resistivity Imaging Surveys

Since all geological structures tend to be three-dimensional (3-D) in nature, a 3-D resistivity survey and interpretation model (Fig. 4) should give the most accurate results. Although it has not reached the same level of usage as 2-D surveys, it is now more widely used in very complex areas such as in environmental and mineral exploration surveys (Loke et al. 2013). Many of the early 3-D surveys used the pole-pole array over relatively small grids (up to about 20 by 20 electrodes) with measurements in different directions (Loke and Barker 1996). The use of other arrays, such as the pole-dipole and Wenner-Schlumberger, is now becoming more common in 3-D surveys that involve thousands of electrode positions (Loke et al. 2013). Ideally, the electrodes should be arranged in a rectangular grid to get a full 3-D coverage. However, a more cost-effective strategy is usually followed wherein 3-D data sets are collated from independent 2-D survey lines. This strategy greatly reduces the cost of a 3-D survey and also allows for 2-D inversions of each individual line. The data interpretation techniques used for 2-D surveys can be extended to 3-D surveys (Loke and Barker 1996). Fast computer software that takes minutes to hours to

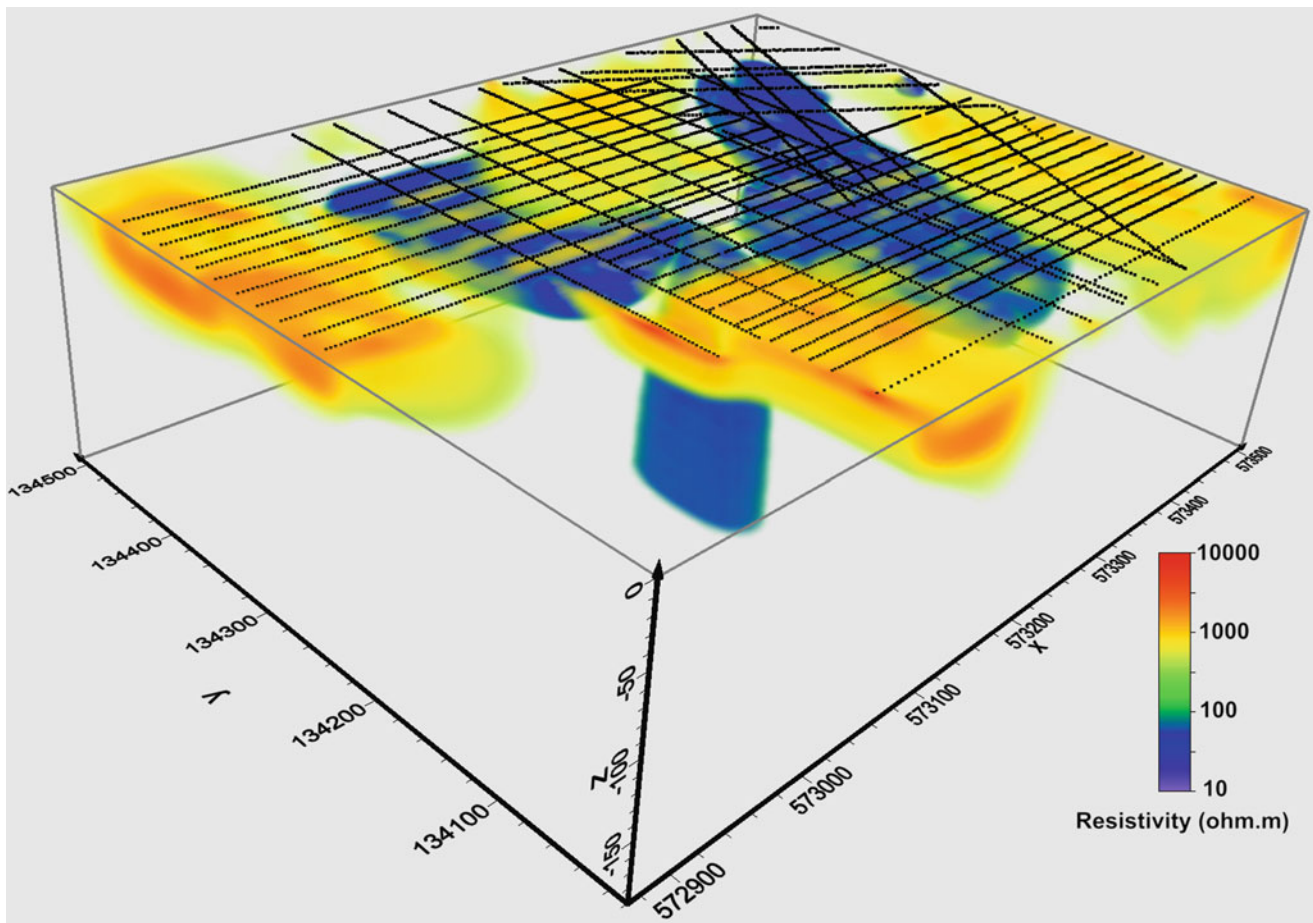
invert a 3-D data set (depending on size) on multicore PCs are now available. Other nonlinear optimization methods such as neural networks, simulated annealing, and conjugate gradient techniques have also been used for resistivity data inversion (Pellerin and Wannamaker 2005).

Figure 4 shows an example from the Hanford site in Washington state, USA, where liquid waste of high ionic strength (mostly comprised of sodium nitrate) was disposed in trenches and concrete cribs (Rucker et al. 2009). Different resistivity survey phases were carried out using 2-D lines in different directions and spacings. The linear high resistivity features near the surface are due to the trenches and concrete cribs. The leakage zones are marked by the prominent low resistivity zones in the deeper regions.

The field survey and data inversion methods have been adapted for other types of problems such induced polarization (IP) surveys (Fiandaca et al. 2013), measurements across boreholes (Wilkinson et al. 2008), aquatic surveys (Rucker and Noonan 2013; Dahlin and Loke 2018), cylindrical geometries (Chambers et al. 2004), measurements with long electrodes (Rucker et al. 2010), and mobile survey systems (*instrumentation, electrical resistivity*).

4-D Time-Lapse Surveys

In time-lapse surveys, measurements using the same array configurations are repeated, usually at regular intervals, to



Electrical Resistivity Surveys and Data Interpretation, Fig. 4 A 3-D survey (electrodes marked by black dots) and model. Blue regions have resistivity of less than 100 ohm.m, while light-green and orange regions have resistivity above 400 ohm.m

monitor temporal changes in the subsurface resistivity. The least-squares optimization method can be modified for time-lapse surveys, so as to minimize temporal changes in the model resistivity values (Kim et al. 2009; Loke et al. 2014), as shown by the following equation:

$$\begin{aligned} & [\mathbf{J}_i^T \mathbf{R}_d \mathbf{J}_i + (\lambda_i^2 \mathbf{W}^T \mathbf{R}_m \mathbf{W} + \alpha_i^2 \mathbf{M}^T \mathbf{R}_t \mathbf{M})] \Delta \mathbf{q}_i \\ &= \mathbf{J}_i^T \mathbf{R}_d \mathbf{g}_i - (\lambda_i^2 \mathbf{W}^T \mathbf{R}_m \mathbf{W} + \alpha_i^2 \mathbf{M}^T \mathbf{R}_t \mathbf{M}) \mathbf{q}_{i-1}. \end{aligned} \quad (4)$$

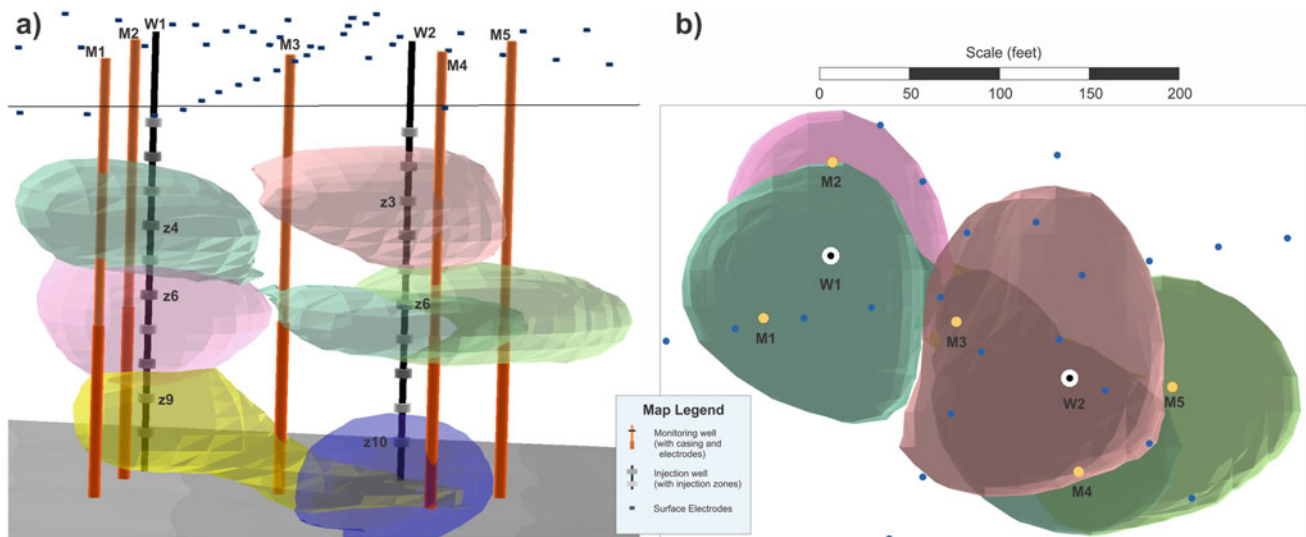
$\mathbf{M}$ is a difference matrix (with regularization factor α) applied across the time models to minimize the difference in the resistivity of each model cell and the corresponding cell for the next temporal model. The data misfit vector $\mathbf{g}$ now incorporates the measurements from all the time-series data sets, while the model vector $\mathbf{q}$ incorporates the resistivity values from all the temporal models (Kim et al. 2009).

Sophisticated integrated monitoring data acquisition systems have been developed in recent years (Uhlemann et al. 2017; Poje et al. 2018) that can automatically carry out the measurements at preset times and transmit the data over the Internet to a processing center where it is automatically

inverted. Figure 5 shows an example from a survey to monitor the flow of a leaching reagent through a rock ore heap as part of a secondary recovery method at a gold mine (Poje et al. 2018). Plots of the change in saturation (calculated from the change in the resistivity) show the regions within the rock heap penetrated by the leaching reagent. The volumes of the regions covered by the plumes provided an estimate of the value of the gold and silver extracted using the known concentration of the metals.

Summary

The electrical resistivity survey method has undergone tremendous changes over the past 25 years. While traditional resistivity profiling and sounding surveys are still used, 2-D imaging surveys are now the method of choice for most areas as they can accurately map moderately complex structures. The field equipment and computer interpretation software are widely available. 3-D surveys now play an increasingly important role in very complex areas. In many cases the 3-D data set is collated from a number of separate 2-D survey



Electrical Resistivity Surveys and Data Interpretation, Fig. 5 Stacked plumes of 5% saturation increase for different zones from injects at wells W1 and W2 (a) side view and (b) overhead view

lines. There have been rapid developments within the past decade in the instrumentation, data acquisition techniques, and inverse modeling techniques for 4-D time-lapse surveys to monitor temporal changes in the subsurface resistivity.

Cross-References

- [Electrical Properties of Rocks](#)
- [Instrumentation, Electrical Resistivity](#)
- [Inverse Theory, Linear](#)
- [Inverse Theory, Singular Value Decomposition](#)
- [Numerical Methods, Finite Difference](#)
- [Numerical Methods, Finite Element](#)

Acknowledgments Chambers, Wilkinson, and Kuras publish with the permission of the Executive Director of the British Geological Survey (UKRI-NERC).

Bibliography

- Chambers JE, Loke MH, Ogilvy RD, Meldrum PI (2004) Non-invasive monitoring of DNAPL migration through a saturated porous medium using electrical impedance tomography. *J Contam Hydrol* 68:1–22
- Cabbage B, Noonan GE, Rucker DF (2017) A modified Wenner array for efficient use of eight-channel resistivity meters. *Pure Appl Geophys* 174:2705–2718
- Dahlin T, Loke MH (2018) Underwater ERT surveying in water with resistivity layering with example of application to site investigation for a rock tunnel in central Stockholm. *Near Surface Geophysics* 16:230–237
- Dahlin T, Zhou B (2004) A numerical comparison of 2D resistivity imaging with ten electrode arrays. *Geophys Prospect* 52:379–398
- deGroot-Hedlin C, Constable S (1990) Occam's inversion to generate smooth, two-dimensional models from magnetotelluric data. *Geophysics* 55:1613–1624
- Farquharson CG, Oldenburg DW (1998) Nonlinear inversion using general measures of data misfit and model structure. *Geophys J Int* 134:213–227
- Fiandaca G, Ramm J, Binley A, Gazoty A, Christiansen AV, Auken E (2013) Resolving spectral information from time domain induced polarization data through 2-D inversion. *Geophys J Int* 192(2):631–646
- Inman JR (1975) Resistivity inversion with ridge regression. *Geophysics* 40:798–817
- Kim JH, Yi MJ, Park SG, Kim JG (2009) 4-D inversion of DC resistivity monitoring data acquired over a dynamically changing earth model. *J Appl Geophys* 68:522–532
- Koefoed O (1979) *Geosounding principles 1: resistivity sounding measurements*. Elsevier Science Publishing Company, Amsterdam
- Loke MH, Barker RD (1996) Practical techniques for 3D resistivity surveys and data inversion. *Geophys Prospect* 44:499–523
- Loke MH, Chambers JE, Rucker DF, Kuras O, Wilkinson PB (2013) Recent developments in the direct-current geoelectrical imaging method. *J Appl Geophys* 95:135–156
- Loke MH, Dahlin T, Rucker DF (2014) Smoothness-constrained time-lapse inversion of data from 3-D resistivity surveys. *Near Surface Geophysics* 12:5–24
- Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. *SIAM J Appl Math* 11:431–441
- Martorana R, Capizzi P, D'Alessandro A, Luzio D (2017) Comparison of different sets of array configurations for multichannel 2D ERT acquisition. *J Appl Geophys* 137:34–48
- Pellerin L, Wannamaker PE (2005) Multi-dimensional electromagnetic modeling and inversion with application to near-surface earth investigations. *Comput Electron Agric* 46:71–102
- Poje M, Salgado L, Rucker DF, Levitt M, Strohmeyer J, McNeill M, Rucker K, Berumen R, Galbraith H (2018) Optimizing gold leaching

- with the aid of electrical resistivity geophysics. *Fast Times* 23(2):62–68
- Rucker DF, Noonan GE (2013) Using marine resistivity to map geotechnical properties: a case study in support of dredging the Panama Canal. *Near Surface Geophysics* 11:625–637
- Rucker DF, Levitt MT, Greenwood WJ (2009) Three-dimensional electrical resistivity model of a nuclear waste disposal site. *J Appl Geophys* 69:150–164
- Rucker D, Loke MH, Levitt MT, Noonan GE (2010) Electrical resistivity characterization of an industrial site using long electrodes. *Geophysics* 75:WA95–WA104
- Szalai S, Szarka L (2008) On the classification of surface geoelectric arrays. *Geophys Prospect* 56:159–175
- Uhlemann S, Chambers J, Wilkinson P, Maurer H, Merritt A, Meldrum P, Kuras O, Gunn D, Smith A, Dijkstra T (2017) Four-dimensional imaging of moisture dynamics during landslide reactivation. *J Geophys Res Earth Surf* 122(1):398–418
- Wilkinson PB, Chambers JE, Lelliott M, Wealhall P, Ogilvy RD (2008) Extreme sensitivity of crosshole electrical resistivity tomography measurements to geometric errors. *Geophys J Int* 173:49–62
- Wilkinson PB, Loke MH, Meldrum PI, Chambers JE, Kuras O, Gunn DA, Ogilvy RD (2012) Practical aspects of applied optimised survey design for electrical resistivity tomography. *Geophys J Int* 189:428–440

Electromagnetic Methods, Imaging Magma Bodies

Kathryn A. Whaler

School of GeoSciences, University of Edinburgh, Edinburgh, UK

Definition: Essential Concepts

Electromagnetic (EM) methods probe the electrical resistivity, or its inverse, electrical conductivity, of the subsurface. The magnetotelluric (MT) method depends on time-dependent naturally occurring magnetic fields external to the Earth inducing eddy currents in the subsurface whose strength and geometry depend on the subsurface resistivity distribution. These decaying currents generate magnetic fields, and by monitoring changes in both the electric and magnetic fields at the surface, information about the subsurface resistivity can be inferred. More slowly varying magnetic fields penetrate deeper, so separating the signals into their frequency components allows depth resolution of the resistivity structure. Penetration depth also depends on the subsurface resistivity. It is characterized by the “skin depth,” $\delta \approx 500\sqrt{\rho T}$, which, in a uniform Earth of resistivity ρ (Ωm), is the depth (in m) at which a signal of period T (in s) is attenuated to $1/e$ of its surface amplitude. Lightning discharges (sferics) trapped in the ionospheric waveguide and interactions between the magnetosphere and solar wind provide the signal at short and long periods respectively. Between the two, centered on about 1 s period, lies the reduced power “dead band.”

The electric and magnetic fields are measured in (or resolved into) two orthogonal horizontal directions, and the MT data are often supplemented by measurements of the vertical magnetic field. Other than on ice-covered surfaces, electric fields are deduced from the potential difference between two (nonpolarizing) electrodes, typically separated by around 100 m. The magnetic field is measured by induction coils – many turns of fine wire around a metallic core – at shorter periods (up to 10,000 s or so), and fluxgate magnetometers for longer periods. The measured field variations are very small, so it is important to install equipment carefully and minimize noise. Electrode separation must be measured accurately, and the electrode lines, and induction coils or magnetometers must be orientated, and the induction coils or magnetometer components leveled, precisely. Equipment is buried to reduce wind noise, wherever possible away from trees. Moving people, equipment, and vehicles generate noise, and electrical infrastructure (e.g., power lines, generators, electric fences) is especially problematic, although notch filters at the appropriate frequency and its harmonics can help. Typically, practitioners aim to locate sites at least 500 m from “cultural” noise. This limits the locations at which MT data acquisition is possible. Depending on the noise levels, depth of interest and typical resistivities of the study area, data will be recorded for half a day up to several months, with 1–3 days being typical in magmatic regions where the crust and uppermost mantle are of primary interest.

Small-scale structure in the shallow subsurface generates galvanic (noninductive) effects that distort the MT signals. Controlled source transient electromagnetic (TEM), also known as time-domain EM (TDEM), methods can be used to infer the shallow structure and correct for this galvanic distortion if the resistivity structure is one- or two-dimensional. In three-dimensions, joint inversion with MT data is undertaken. TEM/TDEM involves injecting a current pulse into a wire loop laid onto the surface and measuring its decay and the induced magnetic field at the loop’s center.

Sea water effects distort – even dominate – data from sites near the coast. “Near” is indicated by the skin depth – longer periods can be affected even if the coastline is many tens of kilometers away. This can be a significant problem when attempting to use MT to study volcanic islands, although corrections can be calculated from bathymetry models.

Controlled source EM (CSEM) methods are also used, particularly in the marine environment. They overcome loss of power in the dead band, and the source power can be increased where necessary to improve signal-to-noise. However, the processing and modeling of CSEM data is more complicated – in passive MT, the signal is treated as a “harmonic plane wave from infinity,” whereas the source geometry and distance from the receivers must be accounted for in CSEM.

MT data are processed in the frequency domain using robust methods to estimate and then manipulate the

so-called impedance tensor or transfer function, a 2×2 tensor whose four elements are the ratio of the electric to magnetic field components in the two orthogonal horizontal directions. If data are recorded synchronously at a second, less noisy site, remote reference methods can be used to improve the impedance tensor. These four (complex) numbers can be expressed in terms of their amplitude and phase; with a suitable scaling, the amplitude has units of resistivity and is referred to as apparent resistivity, and the phase is the lag between the magnetic and electric fields. The structure of the impedance tensor indicates the dimensionality required to explain the data. Over a uniform medium, the diagonal elements of the impedance tensor vanish and the two off-diagonal elements are equal; the apparent resistivity is the actual resistivity of the medium, and the phase is 45° . The diagonal elements also vanish if the resistivity structure is one- or two-dimensional, in the latter case only so long as the data are rotated into the geoelectric strike direction (i.e., the direction along which structure does not vary). Rotation to find the optimum geoelectric strike direction can be coupled with accounting for galvanic distortion; this is normally applied simultaneously to all sites at all periods intended for two-dimensional modeling. Phase data are unaffected by galvanic distortion, so some practitioners prefer to interpret only the phase information, which can be described by a quantity known as the phase tensor. The data are normally viewed as $\log(\text{apparent resistivity})$ and phase (or various properties of the phase tensor components) as a function of $\log(\text{period})$ (or $\log(\text{frequency})$) on a site-by-site basis, known as data curves, or else for a number of sites as a pseudo-section, with $\log(\text{period})$ acting as the depth proxy. Both types of plot help visualize the main features of the subsurface resistivity structure. For example, apparent resistivity decreasing or phase increasing with period indicates resistivity decreasing with depth. Galvanic distortion results in a vertical shift of apparent resistivity curves, and uncorrected data will place structure at the wrong depth. When the vertical magnetic field is also measured, the ratio of vertical to each horizontal component of the magnetic field (also estimated robustly in the frequency domain) form a vector known as the tipper. In the absence of lateral resistivity variations, the vertical field vanishes; thus, lateral variations “tip” the magnetic field out of the horizontal plane. Tippers are normally displayed as induction vectors whose real components point away from (in the Wiese convention) subsurface conductors. However, they are also affected by steep topography often associated with volcanoes. Thus, dimensionality analysis (and geoelectrical strike determination, in the two-dimensional case) is based on the structure of the impedance tensor, accounting for noise and galvanic distortion, and the tipper values.

The relationship between MT data and subsurface resistivity is strongly nonlinear, meaning inversion is fundamentally non-unique, as well as suffering from the usual ambiguities

associated with insufficient, inconsistent, and noisy data. Most practical algorithms minimize a combination of misfit to the data and a measure of model roughness (e.g., Constable et al. 1987). This, and the inherent diffusive propagation of the signal, tends to smear structure, although MT data resolve the top surfaces of conductors well. Their lower surfaces are considerably harder to image, even without issues of model regularization and data resolution, as currents will not penetrate deep into a conductor. Sensitivity analysis, to ascertain which model features are required to fit the data, can help, for example, by testing whether the conductor can be confined to a certain depth range. These tests rely on identifying systematic mismatches between data and their model predictions as much as an overall root-mean-square misfit.

The highly nonlinear nature of the MT inverse problem means inversion is computationally intensive, especially as the need to match boundary conditions “at infinity” means the model area or volume must be much larger than that covered by the data, and include a number of air layers above the Earth’s surface. There are guidelines for determining the appropriate model area/volume and the dimensions of model cells within it. Both isotropic and anisotropic inversion codes are used; data obtained over a two-dimensional anisotropic subsurface can be distinguished from those obtained over a three-dimensional isotropic subsurface by site-to-site consistency in the phase (Simpson and Bahr 2005). It is possible to introduce extra parameters accounting for galvanic distortion into inversion, for example, if TEM data are unavailable (e.g., Kelbert et al. 2014). A number of implementations of well-established algorithms are available for community use, both for data processing and subsurface resistivity modeling. Useful reviews of all aspects of the MT method can be found in the books of Simpson and Bahr (2005) and Chave and Jones (2012). Its application to magmatic systems is reviewed by Magee et al. (2018).

Observations

Fluids, including magma and partial melt, are generally considerably less resistive than solid rock, making them ideal for detection by EM induction methods that are particularly sensitive to conductive material. The signal propagates diffusively and hence it is inherently larger scale (particularly at depth) than seismic methods involving wave propagation. On the other hand, fluids and higher temperatures lower seismic wavespeeds, attenuate the signals, and thus make melt regions harder to image seismically. Thus, EM induction methods, especially MT, have been used extensively to study magmatic systems, in both terrestrial and marine, and many different tectonic environments. Here, a few examples are mentioned to illustrate the type of information obtained and problems faced.

MT data have imaged several low resistivity features beneath the central Andes, interpreted as dykes and magma reservoirs associated with Volcán Uturuncu. A sensitivity test showed that the intermediate depth (2–12 km) dykes were required. Beneath Unzen volcano, Japan, it was inferred from the resistivity structure that a magma column ascended at 20 m/day until it interacted with a water saturated layer; this explained the temporal changes in eruptive styles of the volcano in 1990–1991. The top few kilometers of the resistivity structure beneath Merapi volcano, Indonesia, have been probed, using MT and TEM. Owing to cultural noise from the high population density, only short period MT data were reliable. Pronounced topography effects were separated from resistivity structure when modeling and interpreting induction vectors. Conductive structures were attributed to hot, saline fluids (i.e., a hydrothermal system) rather than partial melt, contributing to the debate as to whether there is a superficial magma reservoir beneath Merapi. Other studies of hydrothermal systems include Mount Fuji, and Meakandake volcanoes, Japan, using MT and Mount St Helens, USA, using CSEM. Mount St Helens has a localized, shallow conductor that extends downwards to join a deeper and more extensive mid-crustal conductor, interpreted as zones of partial melt of silicic magma that erupted in 1980. Beneath Tianchi volcano, on the Chinese-North Korean border, a mid-crustal conductor, inferred to be a magma chamber, was imaged at a time of precursors of a possible eruption (earthquake activity and an increase in spring and surface water temperature). MT is also frequently used to explore high enthalpy geothermal systems, for example, in Iceland and the East African rift valley.

Monitoring of magmatic systems has been undertaken by both time-lapse and continuous EM measurement. Usu volcano, Japan, has been active since the seventeenth century and erupted in 1977–1978. A later MT study found that an initially conductive zone corresponding to the magmatic intrusion associated with the eruption had become resistive, indicating cooling from 800 °C to 50 °C. Impedance tensor changes on timescales of months at Sakurajima volcano, Japan, were ascribed to degassing. Continuous monitoring at La Fournaise, Réunion Island, recorded apparent resistivity decreases associated with the large 1998 eruption and changes to the phase tensor that were attributed to the injection of a N-S striking dyke. At quiet sites, it is possible that hydrologically related seasonal geothermal changes could also be monitored with MT.

MT data from Mount Vesuvius illustrate problems that can arise with galvanic distortion. Data curves from two studies had similar shapes, but the apparent resistivities corrected for galvanic distortion using TDEM were two log units lower. Interpretation of the uncorrected data was of a deep (~8–10 km) magma reservoir, whereas the corrected data suggest that there is a shallow brine layer and no deep

reservoir. The longer period data in both cases were affected by pervasive noise from DC electrified railways.

Marine EM studies have focused on mid-ocean ridges. At the fast-spreading East Pacific Rise, a ~10 km wide sub-vertical conductor slightly displaced from the ridge axis, connected to a deep, broad conductive zone, was interpreted to be a channel efficiently transporting melt to the base of the crust. From MT and CSEM, melt supply to crustal magma chambers beneath the slow-spreading Reykjanes ridge in the Atlantic Ocean appears to be intermittent and rapid. However, the extension of the slow-spreading Red Sea rift on land in Afar, Ethiopia, is underlain by a conductor inferred to contain voluminous melt, implying it is a persistent feature.

Interpretation

How can magma volumes be determined from EM studies? Bulk electrical resistivity is sensitive to melt distribution as well as the presence of electrically conductive fluid phases such as water. For instance, well-connected melt gives a lower bulk resistivity than isolated melt pockets, for the same melt fraction and resistivity. Two-phase mixing laws (e.g., Schmeling 1986) predict bulk resistivity from melt resistivity and geometry in the rock matrix. Resistivities of basaltic and rhyolitic melts have been measured in laboratory experiments (e.g., Laumonier et al. 2015; Guo et al. 2016), but are significantly altered by relatively modest changes in temperature, pressure, silica, sodium, and water content. Even if eruptive products have been well characterized, they may not reflect the current subsurface magma/partial melt composition. Crystallization within a magma reservoir can alter melt composition. These uncertainties make direct inference of melt fraction difficult. Nonetheless, information from laboratory studies, petrology, and geochemistry is always an aid to interpretation; see the review by Pommier (2014).

Other geophysical observations are usefully combined with resistivity information in the study of magmatic systems. For example, seismic data have corroborated the association of low resistivity anomalies with melt. The inference of ~4% connected melt at 10 km depth beneath the Taupo Volcanic Zone, New Zealand, was supported by low Vs values from receiver function studies; this was just beneath the brittle-ductile transition as indicated by the base of the seismogenic zone. Earthquake locations, along with Vp values, were used to infer repeat magma injection along a fault intersection near St George's Hill volcanic center, Montserrat. MT indicating extensive melt beneath an active Quaternary magmatic segment in Afar, Ethiopia, was supported by low Pn and surface wave velocities, high Vp/Vs values, and low amplitude receiver functions. Crustal conductors along the northern flanks of the main Ethiopian rift coincide with locations where receiver functions either have amplitudes too low to

interpret or indicate high V_p/V_s values. MT data indicate that earthquakes occurring deeper than the brittle-ductile transition in the main Ethiopian rift, but away from the Quaternary magmatic segments, are magma induced. They are focused where the lower crust is being modified by magma emplacement driven by partial melting of the upper mantle. MT results also indicate lower crustal magmatic underplating supplied by upper mantle melt at the Great basin – Colorado plateau transition in the USA; the inference is supported by seismic anisotropy (shear wave splitting) information.

Electrical anisotropy can also be inferred directly from two-dimensional MT data – the amount by the phase difference between the off-diagonal impedance tensor components, and the direction by the geoelectrical strike. There is moderate agreement between the direction and amount of seismic and electrical anisotropy beneath Quaternary magmatic segments in Afar, Ethiopia, where oriented melt pockets were proposed as the cause of large seismic anisotropy in the lower crust. Schmeling (1985, 1986) predicts bulk seismic wave speeds and electrical resistivity as a function of melt arrangement (tubes, disks, etc.), orientation (e.g., horizontal or vertical), and amount; joint analysis might enable a more robust interpretation of magma distribution. Gravity models of profiles across the main Ethiopian rift and an inactive Quaternary magmatic segment in Afar, Ethiopia, suggest partial melt coincident with low resistivity anomalies.

Sometimes resistivity evidence is difficult to reconcile with other information. The depth to the Antiplano-Puna magmatic body (APMB) in Bolivia is 3–8 km greater from MT than seismic receiver function and ambient noise tomography methods, although the slowest velocities are at the conductor's surface. There is a resistive body beneath Aluto volcano, Ethiopia, where other data including deformation, degassing, and seismic indicated a magma reservoir – in fact, the magmatic region below the superficial hydro-geothermal system is significantly more resistive than the surrounding region. Here, V_p values are low, suggesting a gaseous melt that could therefore have poor connectivity and hence high resistivity. A geochemical study suggests that Aluto's magmatic system is normally dominated by crystalline mushes containing ephemeral, small ($< 1 \text{ km}^3$) segregations of melt, another plausible explanation of its low resistivity.

Sensitivity tests can also aid in resistivity model interpretation. For example, beneath a rifting region of Afar, Ethiopia, models show a large conductor extending into the upper mantle, where melt should be too buoyant to reside. Confining the conductor to the crust significantly increased the overall data misfit, systematically misfitting the longer period data, and made crustal resistivity unfeasibly low. It was inferred to have a conductance (vertically integrated conductivity) of at least 10 kS, a value reasonably robust to its thickness – a similar conductance is obtained by decreasing the resistivity and the thickness of the conductor while

maintaining the data fit. The APMB in Bolivia has a conductance exceeding 20 kS and is at least 6 km thick. To put these values into context, 1 km of seawater has a conductance of about 5 kS.

Uncertainties in inferred magma volumes associated with the inability of MT to resolve melt distribution have been investigated. Geochemical evidence suggests melt in the mantle-straddling conductor beneath Afar should be in vertical stacks of sills, not a single large body as in regularized inversion models. Constrained inversion in which the conductor was divided up into horizontal layers of alternating resistive (solid rock) and conductive (sills containing partial melt) material provided an equally good fit to the data. The conducting layers had a much lower resistivity than when the conductive material was in a single large body, but the total melt volume in the two models was broadly similar. Hence, it is possible that geologically unrealistic MT models can nevertheless be used to infer magma volumes.

Conclusions

EM methods are becoming a standard tool to investigate and monitor magmatic systems. Developments in data acquisition, processing, and modeling, alongside laboratory investigations of the resistivity of melt over a range of compositions and conditions, have greatly increased their utility over the last few decades. The most robust interpretations are obtained when multiple data sets are combined.

Cross-References

- [Electrical Properties of Rocks](#)
- [Electromagnetic Pulsations and Magnetic Storms](#)
- [Geoelectromagnetism](#)
- [Instrumentation, EM](#)
- [Magnetotelluric Data Processing](#)
- [Magnetotelluric Interpretation](#)
- [Magnetotellurics, Crustal Imaging](#)
- [Magnetovariation Studies](#)

Bibliography

- Chave AD, Jones AG (eds) (2012) *The magnetotelluric method: theory and practice*. Cambridge University Press, Cambridge, UK. <https://doi.org/10.1017/CB09781139020138>
- Constable SC, Parker RL, Constable CG (1987) Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics* 52:289–300
- Guo X, Zhang L, Behrens H, Ni H (2016) Probing the status of felsic magma reservoirs: constraints from the P – T – H_2O dependences of electrical conductivity of rhyolitic melt. *Earth Planet Sci Lett* 433:54–62

- Kelbert A, Meqbel N, Egbert GD, Tandon K (2014) ModEM: a modular system for inversion of electromagnetic geophysical data. *Comput Geosci* 66:40–53
- Laumonier M, Gaillard F, Sifre D (2015) The effect of pressure and water concentration on the electrical conductivity of dacitic melts: implication for magnetotelluric imaging in subduction areas. *Chem Geol* 418:66–76
- Magee C et al (17 authors) (2018) Magma plumbing systems: a geophysical perspective. *J Petrol* 59:1217–1251. <https://doi.org/10.1093/petrology/egy064>
- Pommier A (2014) Interpretation of magnetotelluric results using laboratory measurements. *Surv Geophys* 35:41–84
- Schmeling H (1985) Numerical models on the influence of partial melt on elastic, anelastic and electrical properties of rocks. Part I: elasticity and inelasticity. *Phys Earth Planet In* 41:34–57
- Schmeling H (1986) Numerical models on the influence of partial melt on elastic, anelastic and electrical properties of rocks. Part II: electrical conductivity. *Phys Earth Planet In* 43:123–136
- Simpson F, Bahr K (2005) *Practical magnetotellurics*. Cambridge University Press, Cambridge, UK

Electromagnetic Pulsations and Magnetic Storms

G. S. Lakhina¹ and B. T. Tsurutani²

¹Indian Institute of Geomagnetism, New Panvel (W), Navi Mumbai, India

²Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Definition

Electromagnetic Pulsations	Short period geomagnetic field fluctuations in the range of 0.2 to 600 s.
Magnetic Storms	Large scale global disturbances in the geomagnetic field that last for about half a day to several days.
Magnetospheric substorms	Magnetospheric disturbances, resulting from an enhanced coupling between the solar wind and the geomagnetic field, that follow a set pattern lasting for a period ~ half an hour to a few hours.

Electromagnetic Pulsations

Variations in the Earth's magnetic field occur from hundreds of thousands of years to a fraction of a second. It is believed that the long period variations, from hundreds of thousands to tens of years, are caused by the geodynamo processes. On the other hand, the solar output is thought to drive, either directly or indirectly, the geomagnetic field variations at shorter periods varying from about 22 years to a fraction of a second. The geomagnetic field fluctuations at the shortest periods,

ranging from about ~0.2 to 600 s, are called electromagnetic pulsations or simply magnetic pulsations. Therefore, magnetic pulsations have frequencies that fall in the range of ultra-low-frequency (ULF: ~0 to 30 Hz) electromagnetic waves. The amplitudes of the electromagnetic pulsations are found to increase as their time periods increase (or their frequencies decrease). The amplitudes of the magnetic pulsations vary over a large range, from several hundreds of nT at the longest time periods (~600 s) to fractions of a nT at the shortest time periods (~0.2 s). On the ground, the magnetic pulsations are detected by magnetometers whereas in the magnetosphere, they are recorded by magnetic and electric field sensors onboard spacecraft.

Classification of Magnetic Pulsations

Depending on the waveforms and periods, the magnetic pulsations observed on ground can be classified into two broad types. The International Association of Geomagnetism and Aeronomy (IAGA) has formally done such a classification (Jacobs et al. 1964). Magnetic pulsations with quasi-sinusoidal waveforms are called pulsation continuous (Pc) whereas those with irregular waveforms are called pulsation irregular (Pi). Each main class is further subdivided into period bands that roughly characterize a particular type of pulsation as shown in Table 1.

It should be noted that during periods of disturbed geomagnetic activity, such as substorms and magnetic storms, giant magnetic pulsations, called Pc6 or Ps6, are observed with periods ranging from ~600–900 s. Recently, the seasonal and solar cycle modulation of the occurrence and characteristics of Pc3 pulsations at low latitudes has been discussed by Villante et al. (2019).

Generation Mechanisms of Magnetic Pulsations

Several plasma processes occurring in the solar wind and in the Earth's magnetosphere are responsible for the magnetic pulsations that are observed on the ground or in near-Earth space (McPherron 2005).

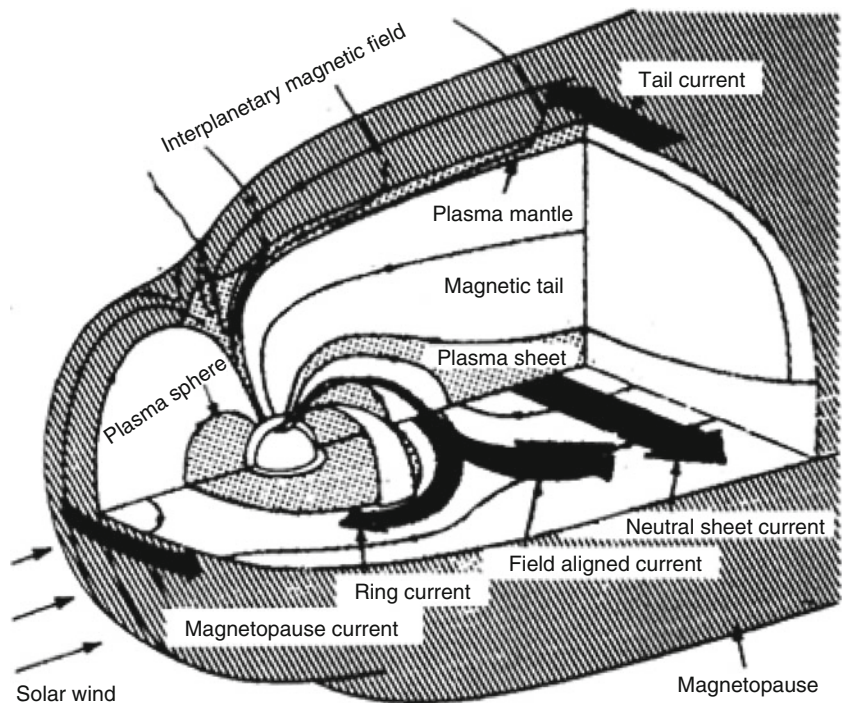
The Earth's magnetosphere is formed by the interaction of solar wind plasma coming out from the Sun and the geomagnetic field. This interaction leads to the geomagnetic field getting confined in a region of space around the Earth, which is known as the magnetosphere. Figure 1 gives a schematic of the three-dimensional view of the magnetosphere. It shows various important plasma regions as well as the current systems existing in the magnetosphere and at its boundary. Several plasma waves and instabilities are driven by the fluctuations in these magnetospheric currents as well as by the nonthermal distributions existing in the magnetosphere. Hence, variations in solar wind parameters can

Electromagnetic Pulsations and Magnetic Storms, Table 1 Classification of Magnetic Pulsations

	Continuous pulsations					Irregular pulsations	
	Pc 1	Pc 2	Pc 3	Pc 4	Pc 5	Pi 1	Pi 2
Period range (seconds)	0.2–5	5–10	10–45	45–150	150–600	1–40	40–150
Frequency range (mHz)	200–5000	100–200	22–100	7–22	2–7	25–1000	7–25

Electromagnetic Pulsations and Magnetic Storms,

Fig. 1 Schematic three-dimensional view of the earth’s magnetosphere formed by the interaction of solar wind from the Sun (from the left) with the geomagnetic field. Small arrows indicate the direction of the magnetic field lines. Thick arrows show the direction of electric currents. Various current systems present in the magnetosphere are shown



produce dramatic effects on the type of waves observed at a particular location at Earth.

Observations show that continuous pulsations having the same frequency range may have different characteristics, for example, their polarization, harmonic structure, or spatial location may not be the same. This suggests that there can be different generation mechanisms for Pc pulsations. In the literature, there are two popular mechanisms for the long-period (Pc3 to Pc5) magnetic pulsations. The first mechanism is based on field-line resonance theory where a monochromatic surface wave is excited by some plasma instability (e.g., Kelvin-Helmholtz (Kivelson and Pu 1984), drift mirror mode (Hasegawa 1969), etc.) at the magnetopause. This instability resonantly couples with a shear Alfvén wave associated with local field line oscillations (Chen and Hasegawa 1974). The second mechanism is based on cavity modes where global fast modes are excited throughout the entire magnetospheric cavity in response to sudden impulses in the solar wind (Kivelson and Southwood 1985). The cavity modes couple to field line resonances that drive currents in the ionosphere producing magnetic pulsations. However, some daytime Pc3 to Pc5 pulsations are believed to be driven directly by the solar

wind pressure fluctuations (Kepko et al. 2002). On the other hand, short-period continuous pulsations (Pc1 and 2) are thought to be excited by an electromagnetic ion cyclotron instability in the equatorial region of the magnetosphere. This instability is driven by the temperature anisotropy of the energetic protons either from the ring current or the plasma sheet region (Cornwall 1965). The irregular magnetic pulsations (Pis) are believed to be generated mainly by the coupling of global cavity modes to field-line resonances (Cheng et al. 1998). Nighttime Pi2 pulsations are most likely generated by earthward propagating fast mode waves launched at substorm onset by large-scale magnetic reconfiguration associated with cross-tail current disruptions (Baumjohann and Glassmeier 1984; Keiling and Takahashi 2011).

Substorms and Magnetic Storms

Substorms and magnetic storms are geomagnetic activity phenomena. A substorm (Akasofu 1964) consists of a set pattern of evolution of ionospheric auroral forms and magnetospheric disturbances lasting typically for about ~30 min to

E

an hour or so. Generally, many substorms occur in a day. During substorms, energy stored in the magnetotail is released explosively directed towards the inner magnetosphere in the midnight sector. This leads to the excitation of discrete auroras which become intense, widespread, and much more agitated than normal. Further, the field-aligned currents and the auroral (ionospheric) electrojets are intensified which produce disturbances in the geomagnetic field. The solar wind is the ultimate source of energy for substorms.

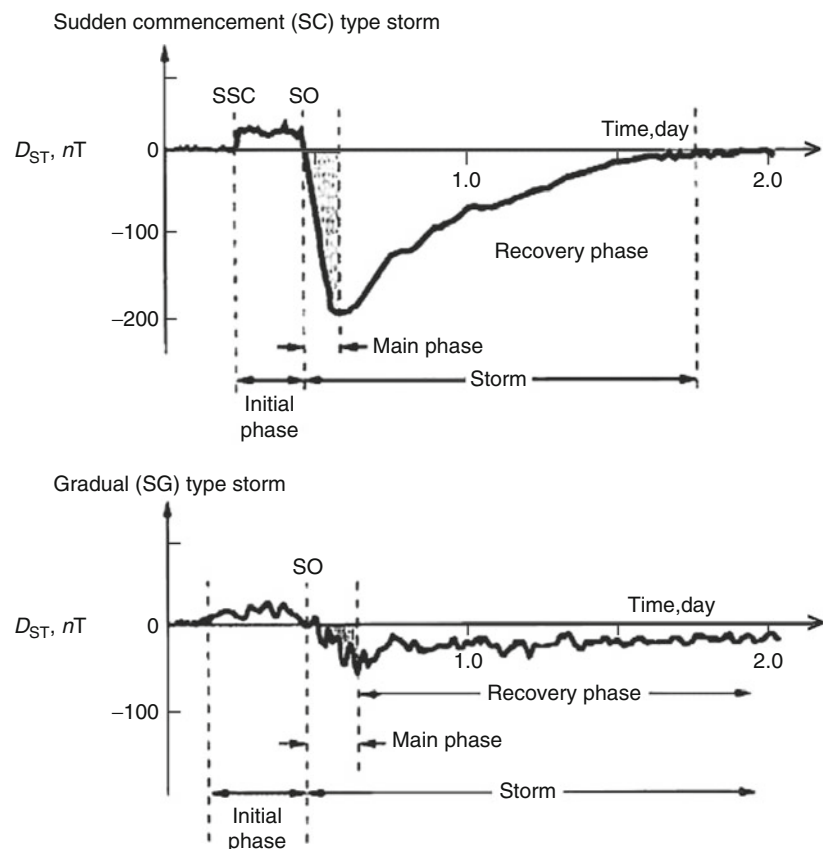
Magnetic storms (Gonzalez et al. 1994) are driven directly by solar drivers like interplanetary coronal mass ejections (ICMEs) and flares, and Corotating Interaction Regions (CIRs) associated with solar wind high speed streams (HSSs) emanating from coronal holes (Tsurutani et al. 1995, 2006). Magnetic storms occur during long intervals of intense southward interplanetary magnetic fields which considerably enhance the magnetic reconnection process (Dungey 1961). This leads to strong plasma injection from the magnetotail towards the inner magnetosphere causing intense global auroras at high and mid-latitudes, and at the same time intensifying the ring current which produces a diamagnetic decrease in the Earth's magnetic field measured at near-equatorial magnetic stations. This latter effect is the *main phase* of the magnetic storm.

Some magnetic storms are characterized by a sudden increase in the ground horizontal magnetic field intensity

shortly before the main phase [see top panel of Fig. 2]. These are known as the *sudden commencement* (SC) storms or *sudden impulse* (SI^+) prior to the storms. This sudden increase in the horizontal component of geomagnetic field is due to the compression of the magnetosphere by the interplanetary shock and downstream compressed plasma. The interval between the sudden commencement and the storm main phase used to be known as the storm *initial phase* but now is not considered to be part of the magnetic storm (Joselyn and Tsurutani 1990). However, all magnetic storms do not have an initial phase. On the other hand, a magnetic storm not accompanied by a SI^+ is called a *gradual geomagnetic storm* (SG) type [bottom panel of Fig. 2]. The *main phase* of a magnetic storm is caused by intensification of the ring current energetic ions injected into the inner magnetosphere ($L \sim 2-7$). During the main phase, the horizontal components of the low-latitude ground magnetic fields are significantly depressed over a time span of one to a few hours. After the ring current particle injection ends or starts to subside, and the ring current particle losses dominate input, the storm *recovery phase* begins. The recovery phase may extend for ~ 10 h or more. The intensity of a geomagnetic storm is measured in terms of the disturbance storm time index (Dst) or SYM-H (SYM-H is basically the same as Dst except the higher time resolution) index. The amount of magnetic field decrease is a measure of the total kinetic energy of the ring current particles (Dessler and Parker 1959).

Electromagnetic Pulsations and Magnetic Storms,

Fig. 2 Schematics of a magnetic storm sudden commencement (SC) driven by an ICME (top) and gradual (SG) type caused by a CIR (bottom). All storms may not have initial phases. (Slightly modified from Tsurutani et al. 2006)



Relationship Between Magnetic Pulsations and Substorms and Magnetic Storms

Drastic changes occur in magnetospheric magnetic field configuration and plasma populations during substorms and magnetic storms. Therefore, magnetic pulsations that are generated internally, e.g., Pc1 and 2 by electromagnetic ion cyclotron instability, Pc3 and 4 by drift mirror instability, and Pi1 and 2 associated with the formation of substorm current wedges, are directly affected during substorms and magnetic storms. Pc6 magnetic pulsations are also generated during substorms and magnetic storms (Moiseev et al. 2016). Planetary distribution of Pc5 pulsations also undergo changes during magnetic storms (Kleimenova and Kozyreva 2014; Marin et al. 2014). Injection of energetic oxygen ions into the magnetosphere during magnetic storms has a tendency to reduce the frequency of long period-pulsations, especially Pc5s. Further, the ULF wave power increases substantially during magnetic storms. Pc5 pulsations are considered as a factor of accelerating energetic electrons to relativistic energies by radial diffusion during storm main phases (Hudson et al. 1999). It is also possible that such effects occur during storm recovery phases as well (Pilipenko et al. 2010).

Magnetic Storms and Society

Intense ($Dst < -100$ nT) and super-intense ($Dst < -500$ nT) geomagnetic storms create hostile space weather conditions that can harm Earth-orbiting spacecraft and technological systems at ground. High fluxes of relativistic “killer” $\sim$ MeV electrons during such magnetic storms can cause malfunctioning and even failure of sensitive instruments and satellite command systems. Some other deleterious effects are failure of satellite communication and navigational systems, satellite data loss and loss of low earth-orbiting satellites caused by an expanded upper atmosphere. The elevated radiation level can impose a threat to astronauts and jetliner passengers. Geomagnetically induced currents (GICs) can cause damage to power line transformers and corrosion of long pipelines and cables.

Future Outlook

Observations by several spacecraft, such as the Time History of Events and Macroscale Interactions during Substorms (THEMIS), Van Allen Probes (VAP), Cluster, Magnetospheric MultiScale (MMS) mission and others, have led to a new understanding of the properties of ULF waves and their effects on the dynamics of energetic particles in the radiation belt (Sigsbee et al. 2016; Pokhotelov et al. 2016; Paulson et al.

2017; Ren et al. 2019). Korotova et al. (2015) attribute the generation of localized Pc 4 pulsations, with an amplitude of ~ 8 nT in the pre-midnight inner magnetosphere observed by VAP-B on 1 May 2013, due to drift-bounce resonance of hot protons. By analyzing the data from VAP on 8 November 2012, Takahashi et al. (2015) found that plasmaspheric ULF waves showed strong spectral power in the 5–40 mHz band consisting of multiharmonic toroidal waves visible up to the eleventh harmonic. They suggest that most likely this wave activity was driven by the foreshock broadband ULF waves. In another study involving data from Arase spacecraft and ground based stations, Takahashi et al. (2018) proposed that the existence of ULF waves plays an important role in accelerating electrons up to relativistic energies (Murphy et al. 2018). Using a 3-D MHD model for ULF waves, Degeling et al. (2018) have shown that accessibility of dayside ULF wave power to radiation belt electrons is controlled by the convection of plasma density. Rae et al. (2018) have proposed a new mechanism for the loss of energetic radiation belt electrons based on direct modulation of the loss cone by localized compressional ULF waves.

Shi et al. (2018) have analyzed long-lasting (2–3 days) poloidal ULF waves between 5 and 10 mHz observed by multiple satellite and three high-latitude Super Dual Auroral Radar Network (SuperDARN) radars during the recovery phase of a geomagnetic storm on 24–27 January 2016. They suggest the bump-on-tail ion distributions at 1–3 keV to be the source of wave excitation. By using fortuitous multipoint THEMIS spacecraft observations during a rare isolated fast plasma jet impinging on the magnetopause boundary, Archer et al. (2019) have provided an unambiguous direct observation that the resulting magnetopause motion and magnetospheric ULF waves at well-defined frequencies can only be explained in terms of the magnetopause surface eigenmode mechanism (Chen and Hasegawa 1974).

Zhang et al. (2018) have done a statistical study of the Pc5–6 ULF waves in the magnetotail using 8 years of THEMIS data and examined the influence of solar wind parameters and the geomagnetic activity level on the features of these ULF waves. From the analysis of dayside ULF event observed by THEMIS-A on 11 January 2010, Shen et al. (2017) have concluded that the ULF wave was excited by the cavity mode mechanism (Kivelson and Southwood 1985). On the other hand, from a multi-spacecraft study, Wang et al. (2017) have demonstrated that quiet time Pc5 ULF waves are driven by the combined effects of Kelvin-Helmholtz instability and ion foreshock perturbations. Based on in situ observations of chorus and Pc4-Pc5 ULF waves by VAP and ground-based observations of pulsating auroras, Jaynes et al. (2015) have proposed that substorm-driven Pc4–5 ULF pulsations modulate chorus waves, which in turn produce pulsating electron precipitation into the Earth’s atmosphere via cyclotron resonance process.

To conclude, magnetic pulsations play an important role in the transport of the solar wind energy to the magnetosphere, and significantly affect the dynamics of the inner magnetosphere and outer radiation belt during geomagnetic storms. Electromagnetic pulsations and magnetic storms form an important component of space weather. The ground-based studies of magnetic pulsations and magnetic storms offer a unique and simple way of monitoring the conditions in the magnetosphere and solar wind. Measurements of magnetic pulsations and magnetic storms can be utilized for geophysical surveys to probe the subsurface conductivity structure of the Earth.

Cross-References

- [Geoelectromagnetism](#)
- [Geomagnetic Field, Measurement Techniques](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Geomagnetically Induced Currents](#)
- [Magnetic Data Enhancements and Depth Estimation](#)
- [Magnetic Methods, Airborne](#)
- [Magnetometers](#)
- [Magnetovariation Studies](#)

Acknowledgments GSL thanks Indian National Science Academy, New Delhi, for support under the INSA-Honorary Scientist Scheme.

Bibliography

- Akasofu SI (1964) The development of the auroral substorm. *Planet Space Sci* 12:273–282
- Archer MO, Hietala H, Hartinger MD, Plaschke F, Angelopoulos V (2019) Direct observations of a surface eigenmode of the dayside magnetopause. *Nat Commun* 10:615. <https://doi.org/10.1038/s41467-018-08134-5>
- Baumjohann W, Glassmeier KH (1984) The transient response mechanism and Pi2 pulsations at substorm onset: review and outlook. *Planet Space Sci* 32:1361–1370
- Chen L, Hasegawa A (1974) A theory of long-period magnetic pulsations, I, steady state excitation of field line resonances. *J Geophys Res* 79:1024–1032
- Cheng CC, Chao JK, Hsu TS (1998) Evidence of the coupling of a fast magnetospheric cavity mode to field line resonances. *Earth Planets Space* 50:683–697
- Cornwall JM (1965) Cyclotron instabilities and electromagnetic emission in the ultra low frequency and very low frequency ranges. *J Geophys Res* 70(1):61–69. <https://doi.org/10.1029/JZ070i001p00061>
- Degeling AW, Rae IJ, Watt CEJ, Shi QQ, Rankin R, Zong Q-G (2018) Control of ULF wave accessibility to the inner magnetosphere by the convection of plasma density. *J Geophys Res Space Phys* 123:1086–1099. <https://doi.org/10.1002/2017JA024874>
- Dessler AJ, Parker EN (1959) Hydromagnetic theory of magnetic storms. *J Geophys Res* 64:2239–2252
- Dungey JW (1961) Interplanetary magnetic field and the auroral zones. *Phys Res Lett* 6:47–48
- Gonzalez WD, Joselyn JA, Kamide Y, Kroehl HW, Rostoker G, Tsurutani BT, Vasyliunas VM (1994) What is a geomagnetic storm? *J Geophys Res* 99(A4):5771–5792
- Hasegawa A (1969) Drift mirror instability in the magnetosphere. *Phys Fluids* 12:2642–2650
- Hudson MK, Elkington SR, Lyon JG, Goodrich CC, Rosenberg TJ (1999) Simulations of radiation belt dynamics driven by solar wind variations. In: Burch J, Carovillano RL, Antiochos SK (eds) *Sun-Earth plasma connections*. AGU, Washington, DC, p 171
- Jacobs JA, Kato Y, Matsushita S, Troitskaya VA (1964) Classification of geomagnetic micropulsations. *J Geophys Res* 69:180–181
- Jaynes AN, Lessard MR, Takahashi K, Ali AF, Malaspina DM, Michell RG, Spanswick EL, Baker DN, Blake JB, Cully C, Donovan EF, Kletzing CA, Reeves GD, Samara M, Spence HE, Wygant JR (2015) Correlated Pc4–5 ULF waves, whistler-mode chorus and pulsating aurora observed by the Van Allen probes and ground-based systems. *J Geophys Res Space Phys* 120:8749–8761. <https://doi.org/10.1002/2015JA021380>
- Joselyn JA, Tsurutani BT (1990) Geomagnetic sudden impulses and storm sudden commencements: a note on terminology. *Eos* 71:1808–1809
- Keiling A, Takahashi K (2011) Review of Pi2 models. *Space Sci Rev* 161:63–148. <https://doi.org/10.1007/s11214-011-9818-4>
- Kepko L, Spence HE, Singer HJ (2002) ULF waves in the solar wind as direct drivers of magnetospheric pulsations. *Geophys Res Lett* 29(8):1197. <https://doi.org/10.1029/2001GL014405>
- Kivelson MG, Pu ZY (1984) The Kelvin–Helmholtz instability on the magnetopause. *Planet Space Sci* 32:1335. [https://doi.org/10.1016/0032-0633\(84\)90077-1](https://doi.org/10.1016/0032-0633(84)90077-1)
- Kivelson MG, Southwood DJ (1985) Resonant ULF waves: a new interpretation. *Geophys Res Lett* 12:49–52
- Kleimenova NG, Kozyreva OV (2014) Planetary distribution of geomagnetic pulsations during a geomagnetic storm at solar minimum. *Izvestiya Phys Solid Earth* 50:102–111. <https://doi.org/10.1134/S1069351313060062>
- Korotova GI, Sibeck DG, Takahashi K, Dai L, Spence HE, Kletzing CA, Wygant JR, Manweiler JW, Moya PS, Hwang K-J, Redmon RJ (2015) Van Allen probe observations of drift-bounce resonances with Pc 4 pulsations and wave–particle interactions in the pre-midnight inner magnetosphere. *Ann Geophys* 33:955–964. <https://doi.org/10.5194/angeo-33-955-2015>
- Marin J, Pilipenko V, Kozyreva O, Stepanova M, Engebretson M, Vega P, Zesta E (2014) Global Pc5 pulsations during strong magnetic storms: excitation mechanisms and equatorward expansion. *Ann Geophys* 32:319–331. <https://doi.org/10.5194/angeo-32-319-2014>
- McPherron RL (2005) Magnetic pulsations: their sources and relation to solar wind and geomagnetic activity. *Surv Geophys* 26:545–592
- Moiseev AV, Baishev DG, Mullayarov VA, Samsonov SN, Uozumi T, Yoshikawa A, Koga K, Matsumoto H (2016) The development of compression long-period pulsations on the recovery phase of the magnetic storm on May 23, 2007. *Cosm Res* 54:31–39. <https://doi.org/10.1134/S0010952516010123>
- Murphy KR, Inglis AR, Sibeck DG, Rae IJ, Watt CEJ, Silveira M, Plaschke F, Claudepierre SG, Nakamura R (2018) Determining the mode, frequency, and azimuthal wave number of ULF waves during a HSS and moderate geomagnetic storm. *J Geophys Res Space Phys* 123:6457–6477. <https://doi.org/10.1029/2017JA024877>
- Paulson KW, Smith CW, Lessard MR, Torbert RB, Kletzing CA, Wygant JR (2017) In situ statistical observations of Pc1 pearl pulsations and unstructured EMIC waves by the Van Allen probes. *J Geophys Res Space Phys* 122:105–119. <https://doi.org/10.1002/2016JA023160>
- Pilipenko V, Kozyreva O, Belakhovsky V, Engebretson MJ, Samsonov S (2010) Generation of magnetic and particle Pc5 pulsations during the recovery phase of strong magnetic storms. *Proc R Soc A* 466:3363–3390. <https://doi.org/10.1098/rspa.2010.0079>

- Pokhotelov D, Rae IJ, Murphy KR, Mann IR, Ozeke L (2016) Effects of ULF wave power on relativistic radiation belt electrons: 8–9 October 2012 geomagnetic storm. *J Geophys Res Space Phys* 121:11,766–11,779. <https://doi.org/10.1002/2016JA023130>
- Rae IJ, Murphy KR, Watt CEJ, Halford AJ, Mann IR, Ozeke LG, Sibeck DG, Clilverd MA, Rodger CJ, Degeling AW, Singer HJ (2018) The role of localized compressional ultra-low frequency waves in energetic electron precipitation. *J Geophys Res Space Phys* 123:1900–1914. <https://doi.org/10.1002/2017JA024674>
- Ren J, Zong QG, Zhou XZ, Spence HE, Funsten HO, Wygant JR, Rankin R (2019) Cold plasmaspheric electrons affected by ULF waves in the inner magnetosphere: a Van Allen probes statistical study. *J Geophys Res Space Phys* 124:7954–7965. <https://doi.org/10.1029/2019JA027009>
- Shen XC, Shi QQ, Zong QG, Tian AM, Nowada M, Sun WJ, Zhao HY, Hudson MK, Wang HZ, Fu SY, Pu ZY (2017) Dayside magnetospheric ULF wave frequency modulated by a solar wind dynamic pressure negative impulse. *J Geophys Res Space Phys* 122:1658–1669. <https://doi.org/10.1002/2016JA023351>
- Shi X, Baker JBH, Ruohoniemi JM, Hartinger MD, Murphy KR, Rodriguez JV, Nishimura Y, McWilliams KA, Angelopoulos V (2018) Long-lasting poloidal ULF waves observed by multiple satellites and high-latitude super DARN radars. *J Geophys Res Space Phys* 123:8422–8438. <https://doi.org/10.1029/2018JA026003>
- Sigsbee K, Kletzing CA, Smith CW, MacDowall R, Spence H, Reeves G, Blake JB, Baker DN, Green JC, Singer HJ, Carr C, Santolík O (2016) Van Allen probes, THEMIS, GOES, and cluster observations of EMIC waves, ULF pulsations, and an electron flux dropout. *J Geophys Res Space Phys* 121:1990–2008. <https://doi.org/10.1002/2014JA020877>
- Takahashi K, Denton RE, Kurth W, Kletzing C, Wygant J, Bonnell J, Dai L, Min K, Smith CW, MacDowall R (2015) Externally driven plasmaspheric ULF waves observed by the Van Allen probes. *J Geophys Res Space Phys* 120:526–552. <https://doi.org/10.1002/2014JA020373>
- Takahashi N, Seki K, Teramoto M, Fok M-C, Zheng Y, Matsuoka A, Higashio N, Shiokawa K, Baishev D, Yoshikawa A, Nagatsuma T (2018) Global distribution of ULF waves during magnetic storms: comparison of Arase, ground observations, and BATSRUS + CRCM simulation. *Geophys Res Lett* 45. <https://doi.org/10.1029/2018GL078857>
- Tsurutani BT, Gonzalez WD, Gonzalez ALC, Tang F, Arballo JK, Okada M (1995) Interplanetary origin of geomagnetic activity in the declining phase of the solar cycle. *J Geophys Res* 100:21717–21733
- Tsurutani BT, Gonzalez WD, Gonzalez ALC, Guarnieri FL, Gopalswamy N, Grande M, Kamide Y, Kasahara Y, Lu G, McPherron RL, Soraas F, Vasyliunas V (2006) Corotating solar wind streams and recurrent geomagnetic activity: a review. *J Geophys Res* 111:A07S01. <https://doi.org/10.1029/2005JA011273>
- Villante U, Tiberi P, Pezzopane M (2019) On the seasonal and solar cycle variation of the ULF fluctuations at low latitudes: a comparison with the ionospheric parameters. *J Atmos Solar–Terres Phys* 190:96–107. <https://doi.org/10.1016/j.jastp.2019.05.005>
- Wang C-P, Thorne R, Liu TZ, Hartinger MD, Nagai T, Angelopoulos V, Wygant JR, Breneman A, Kletzing C, Reeves GD, Claudepierre SG, Spence HE (2017) A multi-spacecraft event study of Pc5 ultralow-frequency waves in the magnetosphere and their external drivers. *J Geophys Res Space Phys* 122:5132–5147. <https://doi.org/10.1002/2016JA023610>
- Zhang S, Tian A, Shi Q, Li H, Degeling AW, Rae IJ, Forsyth C, Wang M, Shen X, Sun W, Bai S, Guo R, Wang H, Fazakerley A, Fu S, Pu Z (2018) Statistical study of ULF waves in the magnetotail by THEMIS observations. *Ann Geophys* 36:1335–1346. <https://doi.org/10.5194/angeo-36-1335-2018>

Electronic Geophysical Year

William K. Peterson¹, Daniel N. Baker¹, C. E. Barton², Peter Fox³, Mark A. Parsons³ and Emily A. CoBabe-Ammann⁴

¹Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO, USA

²Australian National University, Canberra, ACT, Australia

³Rensselaer Polytechnic Institute, Troy, NY, USA

⁴Research and Innovations Office, University of Colorado, Boulder, CO, USA

Definition

eGY. Electronic Geophysical Year. An international science year (2007–2008) celebrating the IGY + 50 anniversary that focused on increasing access to scientific data and information for all people.

Introduction and Concept

Sharing scientific data and information for the common good has long been at the heart of good scientific practice. For the geosciences, the outstandingly successful International Geophysical Year, 1957–1958 (IGY), was a watershed moment in international data sharing. During IGY, in addition to initiating much new research, scientists and member countries worked together to emplace numerous geophysical observatories around the globe and establish a network of World Data Centers for archiving data and for making data available without restrictions.

The Electronic Geophysical Year, 2007–2008 (*eGY*), was conceived in the conviction that the benefits of this “Information Commons” are even greater today than they were 50 years ago. Modern information and communications technologies endow us with unprecedented capabilities for sharing data and information, creating new opportunities for Earth and Space Science Research, and facilitating international coordination to focus effort worldwide, avoid replication, and harmonize standards.

The *eGY* VISION. We can achieve a major step forward in Earth and space science capability, knowledge, and usage throughout the world for the benefit of humanity by accelerating the adoption of modern, visionary practices for managing and sharing data and information.

History

In 1999, the International Union of Geodesy and Geophysics (IUGG) called on its Scientific Associations to propose

activities to mark the 50-year anniversary of IGY. The International Association of Geomagnetism and Aeronomy (IAGA) responded through a resolution passed at the IUGG General Assembly in Sapporo in 2003 to lead an Electronic Geophysical Year.

eGY began on July 1, 2007, and ended on December 31, 2008, exactly 50 years after the start and end of IGY. Support for eGY came from IAGA, IUGG, NASA, the United States National Science Foundation, United States Geological Survey, and the Laboratory for Atmospheric and Space Physics (LASP) of the University of Colorado. In-kind contributions came from the American Geophysical Union (AGU), the National Centre for Atmospheric Research in Boulder, and the volunteer labor of eGY participants.

eGY formally ended on December 31, 2008, exactly 50 years after the IGY ended. eGY achieved its vision primarily through implementation and adoption of virtual observatories. As with the IGY, the mission and goals of eGY are being carried forward through national and international organizations.

Scope

eGY focused on the international science community to achieve a step increase in making past, present, and future geoscientific data (including information and services) rapidly, conveniently, and openly available. The themes of eGY were electronic data location and access, data release and permission, data preservation and rescue, data integration and knowledge discovery, capacity building in developing countries (mainly improving Internet connectivity), and education and outreach. Promoting the development of virtual observatories (VOs) and similar user-community systems for providing open access to data and services was a central feature of eGY.

Operations

Working groups were established to focus effort on each of the eGY theme areas. National eGY groups were particularly active in Canada, China, India, Japan, Russia, and the UK. A Web site served not only as a means of networking among eGY participants but was developed as an information resource for data and information practice and policy.

Presentations, theme sessions, town hall meetings, workshops, and demonstrations of VO technology and working systems were organized at international conferences run by AGU, the Asia Oceania Geophysics Society (AOGS), ICSU's Committees on Data for Science and Technology (CODATA), the Committee on Space Research (COSPAR), the European Geosciences Union (EGU), IAGA, and IUGG. A Virtual Observatories in Geosciences (VOiG) conference was held in Denver in June 2007, with selected papers published in the

April 2008 issue of Earth Science Informatics (<https://link.springer.com/journal/12145/1/1>).

In 2008, an Earth & Space Science Informatics Summit was held in Rome to establish a better understanding and cooperation among various geoinformatics and space science informatics initiatives and programs that are growing rapidly worldwide. A report is available at http://laspl.colorado.edu/media/projects/egy/files/ESI_Summit_Report_final.pdf.

One of the first actions of eGY was to establish a declaration for an Earth and Space Science Information Commons. It states

The Electronic Geophysical Year (eGY) joins with the International Council for Science, the World Summit on the Information Society, and many other bodies in recognizing that knowledge is the common wealth of humanity. We have a shared responsibility to create and implement strategies to realize the full potential of digital information for present and future generations. In the 21st century and beyond, access to digital information and new technologies for information integration and knowledge discovery will influence the free and productive development of societies around the world. Providing ready and open access to the vast and growing collections of cross-disciplinary digital information is the key to understanding and responding to complex Earth system phenomena that influence human survival.

and includes eight articles addressing the major aspects of an Earth and Space Science Information Commons (mirroring the eGY themes).

Achievements

Principal legacies of eGY are (1) stronger awareness of the role that informatics plays in modern research, (2) adoption of virtual observatories and similar systems for accessing data, information, and services, and (3) an expanding infrastructure at the international and national levels. As with IGY, the mission of eGY is being carried forward through existing and newly formed national and international organizations.

Advances in which the eGY effort played a role include:

- Establishment of the Earth & Space Science Informatics Division in EGU
- Establishment of COSPAR's Panel for Informatics
- Establishment of an IUGG Union Commission for Data & Information
- Strengthening of AGU's Earth & Space Science Informatics Special Focus Group to where it is now a formal section within AGU
- IUGG readmission as a member of CODATA and support for a CODATA delegate
- Drafting implementation guidelines for the Global Earth Observing System of Systems (GEOSS) data and information policy
- Formation of an International Polar Year (IPY) Data Policy and Management Subcommittee under ICSU, the World

Meteorological Organization, and CODATA as a CODATA Task Group

- Contributions to development of the IPY Data and Information Service and the Polar Information Commons.
- Formation of a CODATA Task Group: eGY Earth & Space Science Data Interoperability
- ICSU/SCID and SCCID: eGY membership of SCID to review data policy and representation on ICSU's Strategic Coordinating Committee for Information and Data
- eGY-Africa as a contribution to IUGG's Geoscience in Africa initiative

Summary

The Electronic Geophysical Year (eGY) was formed to honor the 50th anniversary of the International Geophysical Year (IGY). It focused on increasing access to scientific data and information, by accelerating the adoption of modern, visionary practices for managing and sharing data and information.

Cross-References

- [International Geophysical Year](#)
- [International Polar Year 2007–2008](#)
- [International Year of Planet Earth](#)

Bibliography

- Baker DN (2008) A 21st-century vision for geophysical data management. *Phys Today* 61(9):54. <https://doi.org/10.1063/1.2982123>
- Baker DN, Barton CE (2009) EGY: progress in global earth and space science informatics. *Data Sci J* 8:226. <https://doi.org/10.2481/dsj.008-008>
- Baker DN, Barton C, Rodger AS, Fraser B, Thompson BV, Papatitashvili VO (2004) Moving beyond the IGY: the electronic geophysical year (eGY) concept. *Eos* 85(11):105–112. <https://doi.org/10.1029/2004EO110001>
- Baker DN, Barton CE, Peterson WK, Fox P (2008) Informatics and the 2007–2008 electronic geophysical year. *Eos* 89(48):485–500. <https://doi.org/10.1029/2008EO480001>
- Barton C, Smith R, Weaver R (2010) Data practices, policy, and rewards in the information era demand a new paradigm. *Data Sci J* 9:IGY95–IGY99. https://doi.org/10.2481/dsj.SS_IGY-003
- CoBabe-Ammann E, Peterson WK, Baker D, Fox P, Barton C (2007) The electronic geophysical year (2007, 2008): eScience for the 21st century. *Lead Edge* 26(10):1294–1295
- Fox P (2008) Virtual observatories in geosciences. *Earth Sci Inf* 1:3. <https://doi.org/10.1007/s12145-008-0001-8>
- Parsons MA, Godøy Ø, LeDrew E, de Bruin TF, Danis B, Tomlinson S, Carlson D (2011) A conceptual framework for managing very diverse data for complex interdisciplinary science. *J Inf Sci* 37(6):555–569. <https://doi.org/10.1177/0165551511412705>
- Peterson WK (2009) Open access to digital information: opportunities and challenges identified during the electronic geophysical year. *Data Sci J* 8:S108–S112. <https://doi.org/10.2481/dsj.IGY-002>

Energy Budget of the Earth

Jean-Claude Mareschal¹ and Claude Jaupart²

¹Centre GEOTOP-UQAM, University of Québec, Montréal, QC, Canada

²Université de Paris Institut de Physique du Globe, Paris, France

Definition

The Earth is a physical system with an energy budget where the outgoing energy must be balanced against all the sources of energy. The planet is not in equilibrium: the difference between the energy flowing out of the Earth and the internal heat production is accounted for by the secular cooling of the mantle and core.

Total Heat Loss of the Earth

The main constraint on the total heat loss of the Earth comes from the many heat flux measurements that have been made both on continents and oceans. Because the continents and the oceans have different thermal structures and because the heat flux data are affected by different biases in each region, it is necessary to analyze separately the continental and the oceanic heat losses (Jaupart et al. 2015).

Continental Heat Flow

There are more than 18,000 reported heat flux measurements on the continents and their margins, not counting measurements made out of equilibrium in oil wells. This is double the number of data in the compilation by Pollack et al. (1993), with many additions by individual researchers. Heat flow data are very unevenly distributed geographically with the majority of them coming from Eurasia and North America. Two large continents, Antarctica and Greenland, are almost unsampled. The raw average of all continental heat flux values is 79.7 mW m^{-2} . This value is biased because the data set includes many data from the USA that were collected for geothermal exploration in anomalously hot regions. The bias is made evident when data from the USA and the rest of the world are analyzed separately: the average is 112.4 mW m^{-2} for the USA vs 65.7 for the rest of the world. One way to remove the bias is by area-weighting: i.e., estimating the average over sufficiently large windows ($1^\circ \times 1^\circ$), and then taking the average of all the windows. It yields an average value for all the continents of 63 mW m^{-2} (Table 1). Multiplying by the continental surface area of $210 \times 10^6 \text{ km}^2$ gives 13.5 TW. An alternative method to calculate continental heat loss is to bin the data by “age”, determine the average value within each age group, integrate

Energy Budget of the Earth, Table 1 Continental heat flux statistics^a

	$\mu(Q)$ mW m ⁻²	$\sigma(Q)$ mW m ⁻²	$N(Q)$
World			
All values	79.7	162	14,123
Averages 1° × 1°	65.3	82.4	3024
Averages 2° × 2°	64.0	57.5	1562
Averages 3° × 3°	63.3	35.2	979
USA			
All values	112.4	288	4243
Averages 1° × 1°	84	183	532
Averages 2° × 2°	78.3	131.0	221
Averages 3° × 3°	73.5	51.7	128
Without USA			
All values	65.7	40.4	9880
Averages 1° × 1°	61.1	30.6	2516
Averages 2° × 2°	61.6	31.6	1359
Averages 3° × 3°	61.3	31.3	889

^a μ is the mean, σ is the standard deviation, and N is the number of values

the product of heat flux times the areal age distribution for continental crust. It gives about the same value as the area weighted average.

Oceanic Heat Flow

There are more than 20,000 heat flux measurements reported in the oceans but these data are not directly used to calculate the total heat loss through the oceans. The oceanic heat loss is calculated from a model of cooling of the sea-floor. The measurements are used to determine the parameters of this model. The cooling model of the sea-floor is a direct consequence of the sea-floor spreading hypothesis. It assumes that sea floor forms at mid-oceanic ridges and cools as it moves away from the ridge. The surface heat flux is calculated by solving the heat conduction equation. The solution shows that heat flux q decreases with age τ as:

$$q(\tau) = C_Q \tau^{-1/2}$$

where C_Q depends on the temperature of the magma ascending at the sea floor and on the thermal properties (thermal diffusivity and conductivity) of the cooling lithosphere.

In order to compare the oceanic heat flux measurements with the model, data have been binned by age. Figure 1 shows that the observed heat flux is systematically less than predicted for sea floor ages <60 My and that heat flux is approximately constant and higher than predicted for ages >80 My.

Young Sea Floor

Only the heat that is conducted through the sea floor can be measured. Near the mid-oceanic ridges, where there are large open fractures in the shallow crust, much heat is transported by the movement of hydrothermal fluids. This convective

component of the heat flux accounts for the difference between the observations and the cooling model. At depth, fractures are sealed by the confining pressure and cooling is by conduction only. In regions of the sea floor that are well sealed by sediments, the heat flux observations do fit the cooling model and can be used to determine the constant C_Q . Estimates of C_Q vary between 470 and 510 mW m⁻² My^{1/2}.

Old Sea Floor

After 80My, the heat flux levels off and is higher than calculated for a half space cooling model. The common interpretation is that cooling of the mantle triggers the onset of small scale convection which maintains a constant temperature or heat flux at fixed depth below the sea floor. Experiments designated to measure the heat flux on sea floor older than 100My have shown that it is almost constant and equal to 48 mW m⁻².

Bathymetry

One consequence of the cooling of the sea floor is that the average density of a rock column increases because of thermal contraction:

$$\Delta\rho_m = -\alpha\rho_m\Delta T$$

where α is the thermal expansion coefficient, ρ_m is mantle density, and ΔT is the temperature difference at time τ . The change in bathymetry can be calculated from the isostatic balance condition:

$$\begin{aligned} \Delta h &= \frac{\Delta\rho_m d}{(\rho_m - \rho_w)} = \frac{-\alpha}{C_p(\rho_m - \rho_w)} \int_0^d \Delta T(z, \tau) dz \\ &= \frac{\alpha}{C_p(\rho_m - \rho_w)} \int_0^\tau [q(0, \tau) - q(d, \tau)] d\tau \end{aligned}$$

where C_p is the specific heat of the lithospheric rocks and ρ_w density of sea water. For a half space, $d \rightarrow \infty$ and $q(d, \tau) \rightarrow 0$, we have:

$$\Delta h = \frac{\alpha}{C_p(\rho_m - \rho_w)} \int_0^\tau q(\tau) d\tau$$

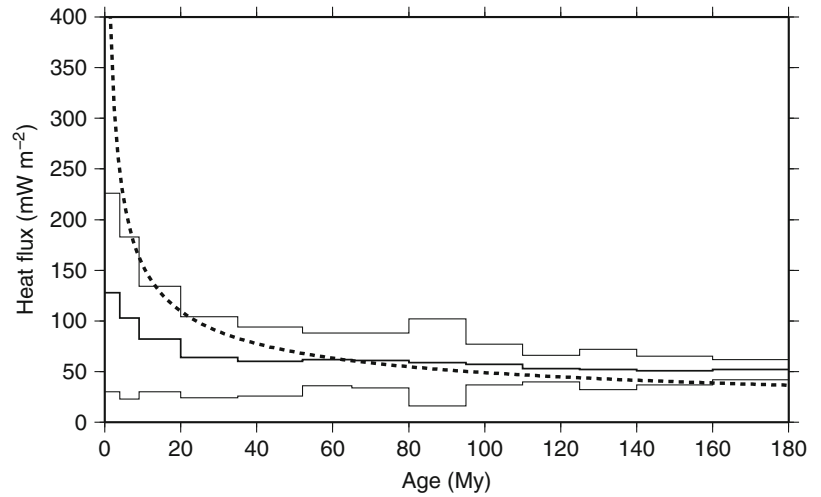
It gives:

$$h = h_0 + \frac{2\alpha\rho_m}{C_p(\rho_m - \rho_w)} C_Q \tau^{1/2} = h_0 + C_H \tau^{1/2}$$

The bathymetry of the sea floor fits very well the cooling model for ages less than 80My. For greater ages, the bathymetry becomes flat, confirming that heat supplied at the base balances the heat loss at the surface of the plate (Fig. 2). The

Energy Budget of the Earth,

Fig. 1 Observed and predicted oceanic heat flux as a function of sea floor age. The dotted line is the predicted heat flux for the half space cooling model. The average observed heat flux ($\pm$ one standard deviation) within different age groups is shown by continuous (thin) lines. (Data from Stein and Stein (1994))



constant C_H is related to C_Q and, with standard values for the physical properties of the mantle ($\alpha = 3.1 \times 10^{-5} \text{ K}^{-1}$, $C_p = 1170 \text{ J kg}^{-1} \text{ K}^{-1}$, $\rho_m = 3330 \text{ kg m}^{-3}$, and $\rho_w = 1000 \text{ kg m}^{-3}$), we obtain $C_H/C_Q = 704 \text{ m}^3 \text{ W}^{-1} \text{ My}^{-1}$. Measurements of the depth to the basement on old sea floor have been used to determine a value of $345 \text{ m My}^{-1/2}$ for C_H which translates as $480 \text{ mW m}^{-2} \text{ My}^{1/2}$ for C_Q .

Age Distribution of the Sea Floor

In order to determine the total oceanic heat loss, we have to integrate the heat flux times the areal distribution of sea floor ages. The distribution of sea floor ages has been very well determined from studies of the marine magnetic anomalies (Müller et al. 2008). There are few ages higher than 180 My and the areal distribution appears to decrease linearly with age τ , such that the areal distribution can approximated by:

$$C_A(1 - \tau/180) \quad (1)$$

where τ is age in My (Fig. 3). In order to account for the total area covered by the oceans including the marginal basins ($300 \times 10^6 \text{ km}^2$) the accretion rate $C_A = 3.4 \text{ km}^2 \text{ year}^{-1}$.

Integrating separately sea floor younger and older than 80 My gives:

$$\begin{aligned} Q_{80-} &= \int_0^{80} C_Q \tau^{-1/2} C_A (1 - \tau/180) d\tau = 24.3 \text{ TW} \\ Q_{80+} &= q_{80} \int_{80}^{180} C_A (1 - \tau/180) d\tau = 4.4 \text{ TW} \\ Q_{\text{oceans}} &= 29 \pm 1 \text{ TW} \end{aligned} \quad (2)$$

The cumulative heat loss as a function of age depends strongly on the age distribution (Fig. 4); with a lower accretion rate and a rectangular age distribution, the oceanic heat loss could be as low as 22 TW.

Hot Spots

Hotspots bring additional heat to the oceanic plates. This heat is not accounted for by the plate cooling model and should be added to the heat loss. The extra heat flux is barely detectable, but the heat that has been put in the plate can be estimated from the volume of the swell of the sea floor: it amounts to 2–4 TW.

Energy Budget of the Mantle

The total heat loss at the Earth surface includes continental heat losses 14 TW, cooling of the sea floor, 29 TW, and heat from the hot spots, 3 TW. The resulting value for the total energy loss at the Earth's surface, 46 TW, is not much different from the first estimates by Williams and von Herzen (1974) (Table 2).

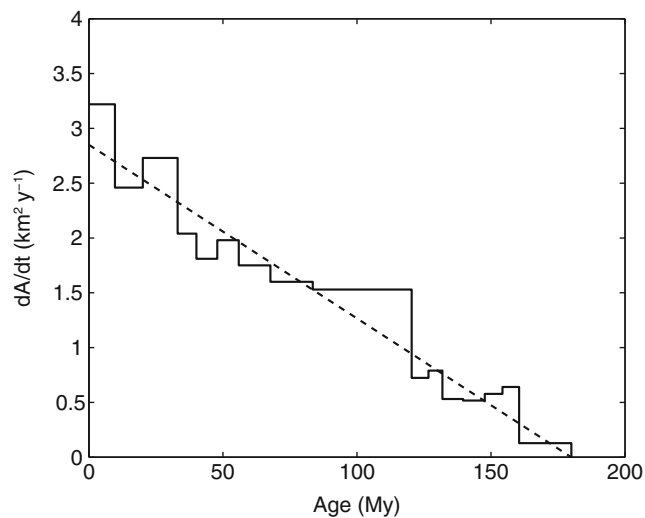
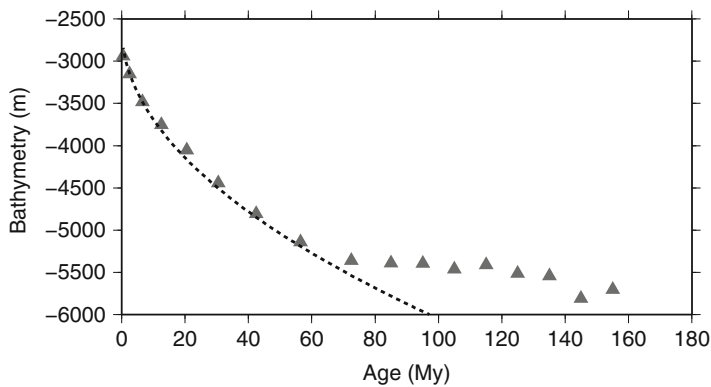
Part of this heat is produced by the decay of radio-active elements (U, Th, and K) in the continental crust which has been estimated by multiple and extensive sampling (Rudnick and Fountain 1995; Rudnick and Gao 2014; Jaupart and Mareschal 2014). It is estimated to be 7–8 TW. This leaves 38–39 TW for the mantle heat loss, either by cooling of the sea floor or by conduction through the continental lithosphere. Three main components balance the mantle energy loss: radiogenic heat production in the mantle, heat flux from the core, and secular cooling of the mantle. Other sources (tidal dissipation, gravitational settling, changes in gravitational potential energy due to thermal contraction) contribute less than 1 TW.

Radiogenic Heat Production

Different approaches have been used to estimate the composition of the Earth's mantle and its radiogenic heat production. Earth was formed by accretion of planetesimals in the Solar

Energy Budget of the Earth,

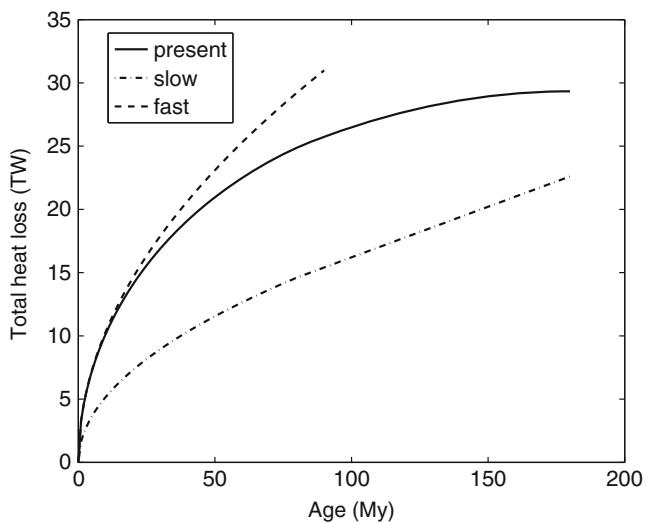
Fig. 2 Worldwide average sea floor bathymetry as a function of age (solid triangle), and predicted bathymetry for a cooling half space (dotted line). Bathymetry data from Crosby and McKenzie (2009)



Energy Budget of the Earth, Fig. 3 Age distribution of the sea floor, not including the marginal basins: distribution from isochrons (solid line); best fit to the distribution (dotted line). (Data from Cogné and Humler (2004))

nebula that have the same origin as the chondritic meteorites. However, the accretion process and the giant impact that formed the Moon have modified Earth’s composition relative to the meteorites. Estimates of mantle composition rely on comparing elemental ratios on Earth and meteorites, to infer the composition of bulk silicate Earth (BSE), i.e., the mantle and crust. Another method assumes that compositional trends in meteorites and mantle rocks record processes in the solar nebula and the mantle, and that they will intersect at the composition of their common source. Both methods yield a heat production of ≈ 20 TW for BSE. Accounting for the radio-elements that have been stored in the continental lithosphere leaves 13 TW in the convecting mantle. It is useful to compare the mantle heat production to its heat loss: this ratio is referred to as the *Urey ratio* of the mantle. It is estimated to be in the range 0.12–0.49 with a preferred value of 0.29.

A lower bound for mantle heat production can be obtained from the composition of the rocks that are extracted from the



Energy Budget of the Earth, Fig. 4 Cumulative total heat loss of the sea floor for three different distributions of the sea floor age: the present triangular distribution, triangular distribution with twice the present spreading rate and oldest sea floor 90My, rectangular distribution with oldest sea floor 180My and spreading rate half of present. The total oceanic area is 300×10^6 km², and $C_Q = 490$ mW m⁻²My^{-1/2}; for ages >80My, heat flux is 48 mW m⁻²

Energy Budget of the Earth, Table 2 Estimates of the continental and oceanic heat flux and global heat loss

	continental mW m ⁻²	Oceanic mW m ⁻²	Total TW
Williams and von Herzen (1974)	61	93	43
Davies (1980)	55	95	41
Slater et al. (1980)	57	99	42
Pollack et al. (1993)	65	101	44
Davies and Davies (2010)	71	105	47
Jaupart et al. (2015) ^a	65	94	46

^aThe average oceanic heat flux does not include the contribution of hotspots. The total heat loss estimate does include 3 TW from oceanic hotspots

mantle to form the sea floor. This only gives a lower bound because it does not take into account a possible hidden reservoir that has not been tapped by the extraction of the MORBs. This estimate is 11 TW for the mantle; accounting for the heat production in the continental lithosphere gives a total radiogenic heat production higher than 18 TW for BSE. These various estimates are summarized in Table 3.

Heat Flux From the Core

Some authors have assumed that the heat flux from the core is equal to the total heat transported by the hotspots. There is no reason that it should be so. The heat carried away by hotspots originating at the core mantle boundary must be less than the core's output. This gives a lower bound of 4 TW. A different bound can be obtained by calculating the minimum heat flux that would be conducted along an adiabat in the core. The core is thought to have a high thermal conductivity ($\approx 80 \text{ W m}^{-1} \text{ K}^{-1}$) and the heat flux conducted along the adiabat is 40 mW m^{-2} , with a total output of 5 TW. More recent measurements and calculations of the core thermal conductivity (Pozzo et al. 2012; Gomi et al. 2013) have led to much higher values of thermal conductivity than previously thought, in the range $130\text{--}200 \text{ W m}^{-1} \text{ K}^{-1}$. Such values

imply a total core heat loss between 9 and 17 TW. How much energy is required to power the geodynamo has been much debated. Present estimates vary within a wide range (9–17 TW) with an average value of 11–12 TW (Nimmo 2015; Buffett 2015).

Secular Cooling

Table 4 compares different ways of breaking down the energy budget of the mantle. Stacey and Davis (2008) add the radio-activity in the continental crust to that of BSE and thus overestimate the total heat production. Both Stacey and Davis (2008) and Davies (1999) assumed that all the heat from the core is carried away by the hotspots, although the hotspots provide only a weak lower bound on core heat flow.

The difference between the total mantle energy loss and the inputs from the core, radio-activity, and other sources, $38\text{--}11\text{--}11 = 16 \text{ TW}$, must be accounted for by the secular cooling of the mantle. Because of the large uncertainties on the core heat loss and on mantle heat production, the range for mantle cooling is implausibly wide (Table 4). The present mean value of 16 TW corresponds to a mantle cooling rate of $\approx 100 \text{ K Gy}^{-1}$. Mantle temperatures can be calculated from the composition of the Mid oceanic ridge basalts (MORB).

Energy Budget of the Earth, Table 3 Radio-element concentration and heat production in meteorites, in the Bulk Silicate Earth, in Earth mantle and crust

	U (ppm)	Th (ppm)	K (ppm)	A* (pW kg ⁻¹)
CI Chondrites				
Palme and O'Neill (2003)	0.008	0.030	544	3.5
McDonough and Sun (1995)	0.007	0.029	550	3.4
EH Chondrites				
Wasson and Kalleyman (1988)	0.009	0.030	/	
Bulk Silicate Earth				
From CI Chondrites				
Javoy (1999)	0.020	0.069	270	4.6
From EH Chondrites				
Kaminski and Javoy (2013) ^a	0.018 ± 0.001	0.060 ± 0.003	217 ± 11^b	4.0 ± 0.2
From Chondrites and Lherzolites trends				
Hart and Zindler (1986)	0.021	0.079	264	4.9
From Elemental Ratios and Refractory Lithophile Elements abundances				
McDonough and Sun (1995)	$0.020 \pm 20\%$	$0.079 \pm 15\%$	$240 \pm 20\%$	4.8 ± 0.8
Palme and O'Neill (2003)	$0.022 \pm 15\%$	$0.083 \pm 15\%$	$261 \pm 15\%$	5.1 ± 0.8
Lyubetskaya and Korenaga (2007)	0.017 ± 0.003	0.063 ± 0.011	190 ± 40	3.9 ± 0.7
Jackson and Jellinek (2013)	0.014 ± 0.003	0.055 ± 0.011	166 ± 30	3.4 ± 0.5
Depleted MORB source				
Workman and Hart (2005)	0.0032	0.0079	25	0.59
Average MORB mantle source				
Su (2000); Langmuir et al. (2005)	0.013	0.040	160	2.8
Peridotites	0.006	0.02	100	1.5
Continental crust				
Rudnick and Gao (2014)	1.3	5.6	$1.5 \cdot 10^4$	330
Jaupart and Mareschal (2014)	/	/	/	293–352

^aU and Th values deduced from the Ca concentration and the chondritic U/Ca and Th/Ca ratios

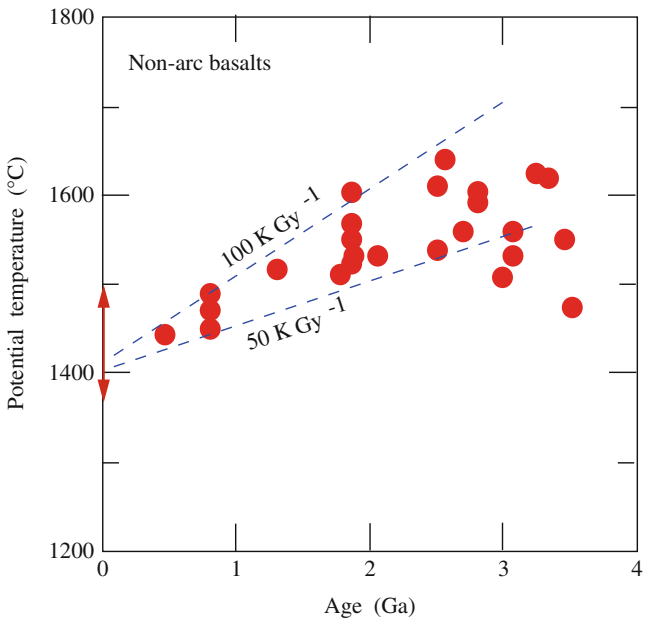
^bCalculated from the U value and $K/U = 12,000$

Energy Budget of the Earth, Table 4 Various estimates of the breakdown of the global budget in TW. Stacey and Davis (2008) have erroneously added heat production in the continental crust to that of bulk silicate earth. They also assume that gravitational potential energy

released by thermal contraction produces heat rather than strain energy. Both Stacey and Davis (2008) and Davies (1999) assume that hotspots carry all the heat from the core

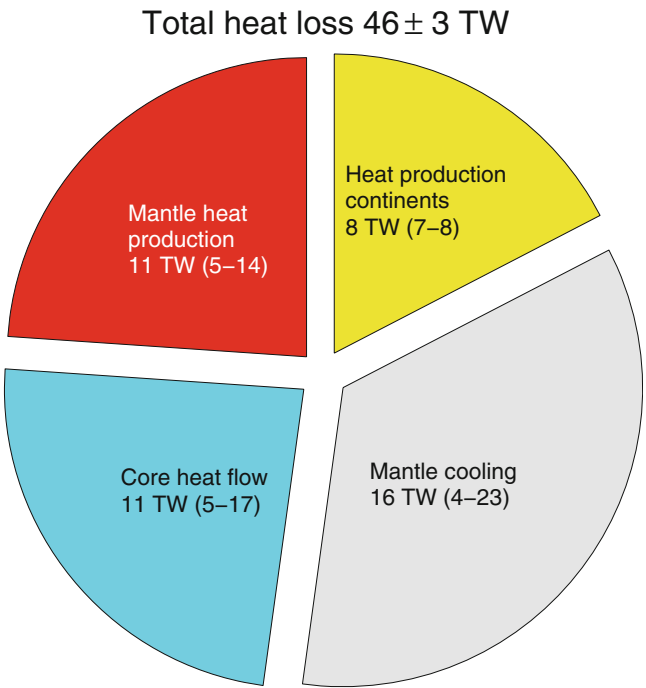
	Stacey and Davis (2008)	Davies (1999)	Jaupart et al. (2015) range (preferred value)
Total heat loss	42	41	46
Continental heat production	8	5	7–8 (8)
Upper mantle		1.3	
Lower mantle		11–27	
Mantle heat production	19	12–28 ^b	9–17 (11)
Latent heat			
(Core differentiation)	1.2	<1	
Mantle differentiation	0.6	0.3	0.3
Gravitational			
(thermal contraction)	2.1		
Tidal dissipation		0.1	0.1
Core heat loss	3	5	5–17 (11)
Mantle cooling	10	9 ^a	4–24 (16) ^c

^aMantle cooling is fixed
^bLower mantle heat production is variable and calculated to fit the mantle cooling rate
^cMantle cooling is adjusted to fit the other terms in the budget



Energy Budget of the Earth, Fig. 5 Temperature of the source of nonarc basaltic rocks as a function of age. The vertical arrow at zero age shows the range of mantle potential temperature estimates. Most of the data fall between the 50 K Gy⁻¹ and 100 K Gy⁻¹ lines (Herzberg et al. 2010). The convergence of the trend to the present range of mantle temperatures gives confidence in these results

Petrological studies on nonarc basaltic rocks (Herzberg et al. 2010) suggest a cooling rate in the range 50–100 K Gy⁻¹ (Fig. 5). The present cooling rate falls near the high end of the range obtained in the petrology studies. There is a scatter in the estimated mantle temperatures, but there is a trend



Energy Budget of the Earth, Fig. 6 Proposed breakdown of the present energy budget of the Earth. With heat production in the continental lithosphere, mantle heat production, and core heat loss constrained, the mantle cooling rate is adjusted to fit the total energy loss determined from surface heat flux, allowing an implausible wide range

indicating lower cooling rate in the past and possibly heating of the mantle before 3 Ga. These differences between the present and secular rates of cooling raise interesting questions about long term changes in the energy budget (Fig. 6).

Present and Long Term Cooling Rates

Is the present rate of energy loss equal to the long term rate? The energy lost through the sea floor depends on the age distribution of the sea floor. It is likely that the energy lost through the sea floor fluctuates with the super continent cycle and that the present energy loss is different from the long term mean. The second point is that radioactive heat production in the mantle was higher in the past than now for two reasons: one reason is the rundown of radio-elements, the other is the extraction of continental crust that removed radio-elements from the mantle. Finally, the heat loss from the mantle depends on the total continental area as much more heat escapes through the sea floor than through the continental lithosphere. The input from the core may also oscillate and have been different in the past, when the frequency of geo-magnetic reversals was lower than in recent times.

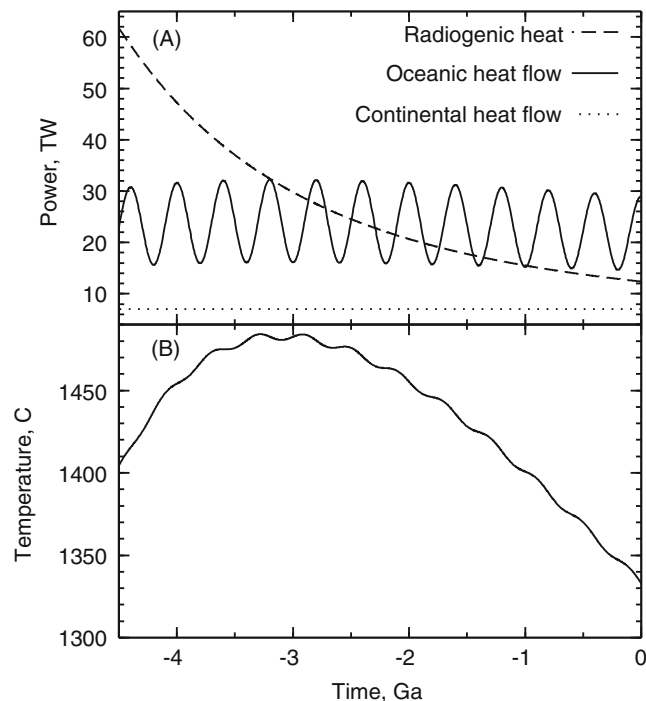
The relative importance of these effects remains difficult to assess because we do not precisely know how the volume of continental crust changed with time, but Fig. 7 presents one plausible thermal history of the mantle with the added

assumption that the mantle heat loss oscillates by ± 7 TW over a period of 400My corresponding to the supercontinent cycle.

Perspectives

The past years have seen the development of underground neutrino observatories that can detect geo-neutrinos, i.e., neutrinos from the decay of U and Th in the Earth, which appeared to be a new avenue for estimating the radio-genic power of the mantle (Dye 2012). First results from 2 land observatories (Kamioka, in Japan, and the small Borexino detector, in Italy) have yielded an estimated 18 TW total heat production for the crust and mantle but still leave a large uncertainty due to measurements errors and the poorly estimated crustal contribution to the measured neutrino activity (Kamland collaboration, (Gando, A. and 45 collaborators) 2013; Borexino collaboration group 2013).

New estimates of the core thermal conductivity imply a much higher rate of cooling (>10 TW) for the core than previously thought. These values of thermal conductivity will need further confirmation, but they narrow down the estimated core heat flux and consequently imply tighter constraints on the mantle energy budget.



Energy Budget of the Earth, Fig. 7 Effect of fluctuations of oceanic heat loss over a supercontinent time-scale of $\tau_w = 400$ My on the secular thermal evolution of the mantle, from Labrosse and Jaupart (2007). The model calculations rely on a simple parametrization of sea floor spreading that leads to a weak secular variation of oceanic heat loss. Heat loss fluctuations due to changes of sea floor spreading are superimposed on the secular trend. These changes occur on a time-scale that is short compared to that of the secular trend and leave almost no trace on mantle temperatures. Top panel (a): the various contributions to the Earth's energy budget that are used in the calculation. Bottom panel (b): predicted evolution of the average mantle temperature through time

Summary

The Earth loses energy as heat flows out through its surface. The total energy loss of the Earth has been estimated at 46 ± 2 TW, of which 14 TW come through the continents while 32 TW come from the cooling of the sea floor. After removing the heat production of radio-active elements in the continental lithosphere (8 TW), the energy loss from the mantle is 38 TW. Three main components balance the mantle energy budget: Heat flow from the core (9–17 TW), radiogenic heat production in the mantle (9–17 TW), and the secular cooling of the mantle (4–29 TW). The uncertainty on the exact contribution of each component is much larger than that on the total heat loss. The best estimate of the present cooling rate of the mantle (110 K Gy^{-1}) is the upper range of estimates from petrology of the secular cooling rate for the past 3Gy. The present cooling rate may be different from its long term average because of variations in sea floor spreading rates and higher mantle heat production in the past.

Cross-References

- [Core Dynamo](#)
- [Heat Flow Determinations, Continental](#)
- [Heat Flow, Seafloor: Methods and Observations](#)
- [Lithosphere, Continental: Thermal Structure](#)
- [Lithosphere, Oceanic: Thermal Structure](#)
- [Radiogenic Heat Production in the Continental Crust](#)
- [Radiogenic Heat Production of Rocks](#)

Bibliography

- Borexino collaboration group (2013) Measurement of geo-neutrinos from 1353 days of Borexino. *Physics Letters B* 722:295–300
- Buffett BA (2015) Core-mantle interactions. In: Olson P (ed) *Treatise of geophysics*, vol 8, 2nd edn. Elsevier, New York, NY, pp 213–233, Chapter 8
- Cogné J-P, Humler E (2004) Temporal variation of oceanic spreading and crustal production rates during the last 180 My. *Earth Planet Sci Lett* 227:427–439
- Crosby AG, McKenzie D (2009) An analysis of young ocean depth, gravity and global residual topography. *Geophys J Int* 178:1198–1219
- Davies GF (1980) Review of oceanic and global heat flow estimates. *Rev Geophys Space Phys* 18:718–722
- Davies G (1999) *Dynamic Earth: plates, plumes, and mantle convection*. Cambridge University Press, New York
- Davies JH, Davies DR (2010) Earth's surface heat flux. *Solid Earth* 1:5–24
- Dye ST (2012) Geoneutrinos and the radioactive power of the Earth. *Rev Geophys* 50:3007
- Gomi H, Ohta K, Hirose K, Labrosse S, Caracas R, Verstraete MJ, Hernlund JW (Nov. 2013) The high conductivity of iron and thermal evolution of the Earth's core. *Phys Earth Planet In* 224:88–103
- Hart SR, Zindler A (1986) In search of a bulk-Earth composition. *Chem Geol* 57:247–267
- Herzberg C, Condie K, Korenaga J (2010) Thermal history of the Earth and its petrological expression. *Earth Planet Sci Lett* 292:79–88
- Jackson MG, Jellinek AM (2013) Major and trace element composition of the high $^3\text{He}/^4\text{He}$ mantle: implications for the composition of a nonchondritic Earth. *Geochem Geophys Geosyst* 14(8):2954–2976
- Jaupart C, Mareschal JC (2014) Constraints on crustal heat production from heat flow data. In: Rudnick RL (ed) *Treatise on geochemistry*, The crust, vol 4, 2nd edn. Elsevier, New York, pp 65–84
- Jaupart C, Labrosse S, Lucazeau F, Mareschal JC (2015) Temperatures, heat and energy in the mantle of the Earth. In: Bercovici D (ed) *Treatise on geophysics*, the mantle, vol 7. Elsevier, New York, pp 223–270
- Javoy M (1999) Chemical Earth models. *C R Acad Sci Paris* 329:537–555
- Kaminski E, Javoy M (2013) A two-stage scenario for the formation of the Earth's mantle and core. *Earth Planet Sci Lett* 365:97–107
- KamLAND collaboration, (Gando, A. and 45 collaborators) (2013) Reactor onoff antineutrino measurement with KamLAND. *Phys Rev D* 88:033001
- Labrosse S, Jaupart C (2007) Thermal evolution of the Earth: secular changes and fluctuations of plate characteristics. *Earth Planet Sci Lett* 260:465–481
- Langmuir CH, Goldstein SL, Donnelly K, Su YK (2005) Origins of enriched and depleted mantle reservoirs. In: Suppl FM (ed) *Eos trans*, vol 86. American Geophysical Union, Washington, DC, Abstract V23D-02, p 1
- Lyubetskaya T, Korenaga J (2007) Chemical composition of Earth's primitive mantle and its variance: 1. Method and results. *J Geophys Res* 112:3211
- McDonough WF, Sun SS (1995) The composition of the Earth. *Chem Geol* 120:223–253
- Müller RD, Sdrolias M, Gaina C, Roest WR (2008) Age, spreading rates, and spreading asymmetry of the world's ocean crust. *Geochem Geophys Geosyst* 9:4006–+
- Nimmo F (2015) Energetics of the core. In: Olson P (ed) *Treatise of geophysics*, vol 8, 2nd edn. Elsevier, New York, NY, pp 27–55, Chapter 2
- Palme H, O'Neill HSC (2003) Cosmochemical estimates of mantle composition. In: Carlson RW (ed) *Treatise on geochemistry*, vol 2, Mantle and core. Elsevier, New York, NY, pp 1–38, Chapter 1
- Pollack HN, Hurter SJ, Johnston JR (1993) Heat flow from the Earth's interior: analysis of the global data set. *Rev Geophys* 31:267–280
- Pozzo M, Davies C, Gubbins D, Alfè D (May 2012) Thermal and electrical conductivity of iron at Earth's core conditions. *Nature* 485(7398):355–358
- Rudnick RL, Fountain DM (1995) Nature and composition of the continental crust: a lower crustal perspective. *Rev Geophys* 33:267–310
- Rudnick RL, Gao S (2014) Composition of the continental crust. In: Rudnick RL (ed) *Treatise on geochemistry*, The crust, vol 4, 2nd edn. Elsevier, New York, pp 1–51
- Sclater JG, Jaupart C, Galson D (1980) The heat flow through oceanic and continental crust and the heat loss from the Earth. *Rev Geophys* 18:269–311
- Stacey FD, Davis PM (2008) *Physics of the Earth*, 4th edn. Cambridge University Press, Cambridge, UK.
- Stein CA, Stein S (1994) Constraints on hydrothermal heat flux through the oceanic lithosphere from global heat flow. *J Geophys Res* 99:3081–3095
- Su YJ (2000) Mid-ocean ridge basalt trace element systematics: constraints from database management, ICPMS analysis, global data compilation and petrologic modeling. Unpub. PhD thesis, Columbia University, New York 1, 569 pp
- Wasson JT, Kalleyman GW (1988) Composition of chondrites. *Phil Trans R Soc Lond, ser A* 325:535–544
- Williams DL, von Herzen RP (1974) Heat loss from the Earth: new estimate. *Geology* 2:327–330
- Workman RK, Hart SR (2005) Major and trace element composition of the depleted MORB mantle (DMM). *Earth Planet Sci Lett* 231:53–72

Energy Partitioning of Seismic Waves

Kalachand Sain

Wadia Institute of Himalayan Geology, Dehradun, India

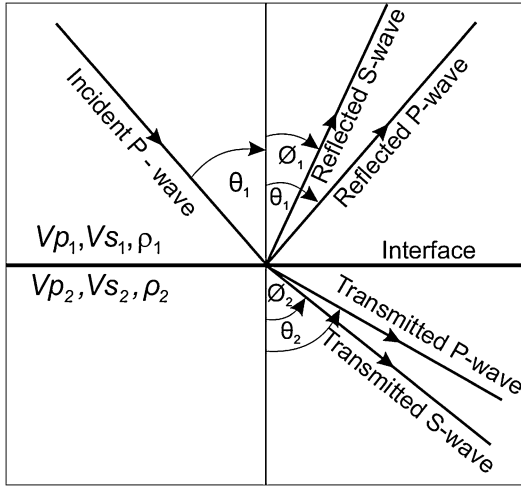
Synonyms

Partitioning – conversion

Definition

When an incoming P-wave strikes an interface between two isotropic homogeneous elastic media at an angle other than the vertical, a portion of the P-wave energy is converted into S-wave energy, which gets reflected and transmitted in the same way as the P-wave does. This is known as the energy partitioning of seismic wave.

Snell's law: The angles of incidence, reflection, and transmission, as shown in Fig. 1, are related by the Snell's law, which says that the ratio of sines of angles of incidence, reflection, and transmission to velocities of respective media across the interface is constant. In seismology, this constant is defined as the ray parameter, R_p as:



Energy Partitioning of Seismic Waves, Fig. 1 Ray diagram for the partition of a plane wave energy at a horizontal interface

$$\frac{\sin \theta_1}{V_{P1}} = \frac{\sin \theta_2}{V_{P2}} = \frac{\sin \phi_1}{V_{S1}} = \frac{\sin \phi_2}{V_{S2}} = R_P \quad (1)$$

where V_P , V_S , and ρ are the P-wave velocity, S-wave velocity, and density, respectively, and subscripts 1 and 2 represent the first and second layers, respectively.

Reflection and transmission coefficients: A reflection coefficient describes either the amplitude or the energy of a reflected wave relative to an incident wave; similarly a transmission coefficient describes either the amplitude or the energy of a transmitted wave relative to an incident wave. When a plane wave strikes a horizontal interface vertically (i.e., at normal incidence), the reflection coefficient, R_C and the transmission coefficient, T_C are defined as:

$$R_C = \frac{Z_2 - Z_1}{Z_2 + Z_1} \quad (2)$$

$$T_C = \frac{2Z_1}{Z_2 + Z_1} \quad (3)$$

where Z_1 and Z_2 are the acoustic impedance (P-wave velocity multiplied by density) of the first and second media, respectively.

Zoeppritz equation: When the seismic ray obliquely strikes the interface (Fig. 1), separated by three elastic parameters (V_P , V_S , and ρ) on either side, the R_C and T_C as expressed by equations (2) and (3) are not valid. The reflection and transmission coefficients or amplitudes of both P-P and mode converted P-S-waves as a function of angle can be derived using the Zoeppritz equations (Zoeppritz 1919) that can be easily expressed through a matrix (Waters 1981) as:

$$Q = P^{-1}R \quad (4)$$

where P is a 4×4 square matrix and, R and Q are 4×1 column matrices whose elements are shown as:

$$P = \begin{bmatrix} \sin \theta_1 & \cos \phi_1 & -\sin \theta_2 & \cos \phi_2 \\ -\cos \theta_1 & \sin \phi_1 & -\cos \theta_2 & -\sin \phi_2 \\ \sin 2\theta_1 & \frac{V_{P1}}{V_{S1}} \cos 2\phi_1 & \frac{\rho_2 V_{S2}^2}{\rho_1 V_{S1}^2} 2 \sin \theta_1 \cos \theta_2 & -\frac{\rho_2 V_{P1} V_{S2}}{\rho_1 V_{S1}^2} \cos 2\phi_2 \\ \cos 2\phi_1 & -\frac{V_{S1}}{V_{P1}} \sin 2\phi_1 & -\frac{\rho_2 V_{P2}}{\rho_1 V_{P1}} \cos 2\phi_2 & -\frac{\rho_2 V_{S2}}{\rho_1 V_{P1}} \sin 2\phi_2 \end{bmatrix} \quad (5)$$

$$Q = [R_{PPR} \ R_{PSVR} \ R_{PPT} \ R_{PSVT}]^T \text{ and} \quad (6)$$

$$R = [-\sin q_1 \ -\cos q_1 \ \sin 2q_1 \ -\cos 2f_1]^T$$

where R_{PPR} and R_{PSVR} are the reflection coefficients and, R_{PPT} and R_{PSVT} are the transmission coefficients of P-P and mode converted P-S waves, respectively.

Figure 2(a) displays the coefficients of reflections and transmissions for P-P and converted P-S waves as a function of incident angles (θ) as per equation (4) for the model shown in the figure itself. At zero angle or offset, the R_C for P-P wave is mainly governed by the P-wave velocity. However, at nonzero angle or offset, its value depends on both the P and S wave velocities. The nature and amplitudes of curves vary with the elastic parameters across the interface (Fig. 2b). These elastic parameters mainly depend on the mineralogical constituents, fluids, pressure, temperature, etc.

Approximation to the Zoeppritz equation led to an elegant method of reducing these curves based on just two parameters. According to Shuey (1985), the P-wave reflection coefficient, $R_C(\theta)$ for an incidence angle (θ) up to $\sim 30^\circ$ can be expressed as:

$$R_C(\theta) \approx A + B \sin^2 \theta \quad (7)$$

where A is the intercept or normal incidence reflection coefficient, and B is the gradient.

$$A = \frac{1}{2} \left(\frac{\Delta V_P}{V_P} + \frac{\Delta \rho}{\rho} \right), B = A A_0 + \frac{\Delta \sigma}{(1 - \sigma)^2} \quad (8)$$

$$A_0 = B_0 - 2(1 + B_0) \frac{1 - 2\sigma}{1 - \sigma}, B_0 = \frac{\Delta V_P / V_P}{\Delta V_P / V_P + \Delta \rho / \rho},$$

$$V_P = (V_{P2} + V_{P1})/2, V_S = (V_{S2} + V_{S1})/2, \rho = (\rho_2 + \rho_1)/2,$$

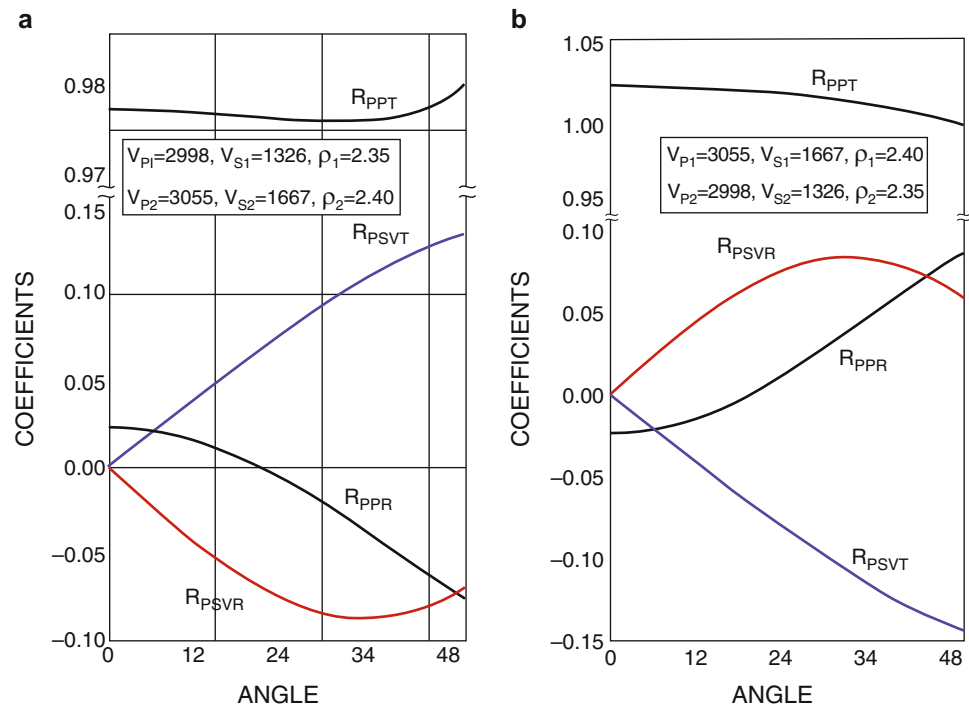
$$\Delta V_P = (V_{P2} - V_{P1}), \Delta V_S = (V_{S2} - V_{S1}), \Delta \rho = (\rho_2 - \rho_1),$$

$$\Delta \sigma = (\sigma_2 - \sigma_1), \sigma = (\sigma_2 + \sigma_1)/2, \sigma_1 = (5V_{P1}^2 - V_{S1}^2) / (V_{P1}^2 - V_{S1}^2)$$

Equations (7) and (8) are valid up to high A for a negative impedance contrast across an interface, and to about $A = 0.2$ for a positive impedance contrast. Since 30° angle corresponds to the maximum angle that we generally encounter

Energy Partitioning of Seismic Waves, Fig. 2

(a) Reflection and transmission coefficients as a function of incidence angles for the model shown inside. (b) Reflection and transmission coefficients as a function of incidence angles for the model shown inside in which the layer parameters for model in (a) are interchanged



in normal seismic data acquisition, this approximation is widely used to determine the amplitude versus offset (AVO) attributes for the interpretation of seismic data.

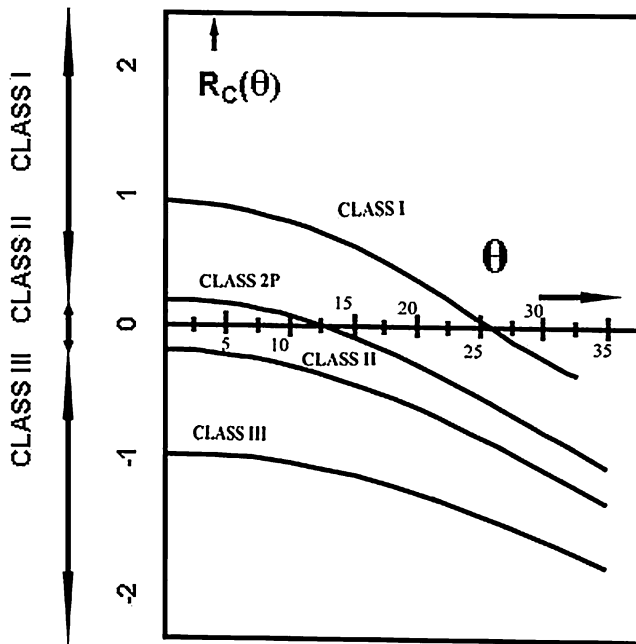
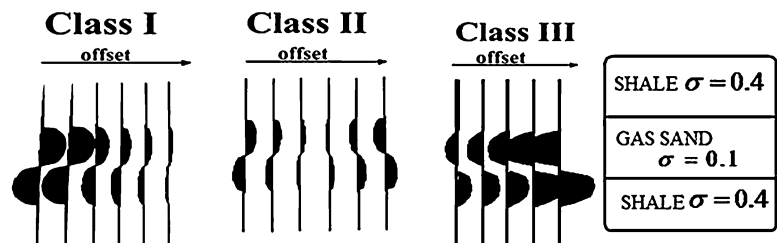
AVO response for a realistic earth model: The reflections from various subsurface interfaces exhibit a wide range of AVO characteristics that depend upon the rock types and fluid contents. From the field data, the amplitudes and offsets of reflected phases can be converted into reflection coefficients and angles (Warner 1990). By modeling using the Zoeppritz equation, we can derive various elastic parameters (Ojha and Sain 2007, 2008; Arun et al. 2018, 2020) or Poisson ratio (σ) that can be used for the exploration of gas hydrates or hydrocarbons. Any overpressure in the trapped gas can also be known from the AVO anomaly (Tinivella 2002; Dash et al. 2004) that helps in avoiding drilling hazard. In oil industries, the AVO attributes have gained considerable popularity for hydrocarbon detection, lithology identification, and reservoir characterizations or fluid parameter analysis (Dahl and Ursin 1992; Castagna and Backus 1993; Castagna et al. 1998; Dahroug et al. 2018) due to the fact that seismic amplitudes at interfaces are affected by the variations of physical properties just above and below the interfaces. Ojha et al. (2010) have demonstrated through a field example that a plot of A versus B coupled with rock physics modeling can be used as a rapid tool for detecting free-gas below a bottom simulating reflector or BSR, and quantifying the amount of gas-hydrate and free-gas along a seismic reflection profile.

The gas-sand detection is the most promising application of AVO analysis. The characteristically low V_P/V_S ratio or σ of gas sand is expected to be differentiated from other low-impedance layers, such as coals and porous brine or oil sands. Rutherford and Williams (1989) have defined three distinct classes of gas-sand AVO anomalies (Figs. 3 and 4) mainly based on the intercept, A and gradient, B attributes.

The Class 1 is a high impedance sand anomaly and occurs when the normal-incidence P-wave reflection coefficient is strongly positive. This shows a strong amplitude decrease with offset and a possible phase change at far offset. The Class II anomaly occurs when the gas sand has similar impedance to the neighboring shale. For the gas sand scenario, the normal incidence reflection coefficient is close to zero and becomes negative at larger offsets. The Class III is the classic Gulf of Mexico AVO anomaly where gas sand has low impedance compared to shale. The normal incidence reflection coefficient is negative and becomes more negative as offset increases. The “bright spot” is associated with the Class III AVO anomaly and can be used as a direct indicator for hydrocarbons.

If the rock overlying the reservoir has a velocity appreciably lower than that of the reservoir itself (e.g., carbonate reservoir capped by shale), the effect of hydrocarbon is to decrease the impedance contrast and reduce the reflection coefficient producing a “dim spot.” Class I and Class II AVO anomalies can produce the reverse effect. When the rock overlying the reservoir has a velocity slightly smaller

Energy Partitioning of Seismic Waves, Fig. 3 Seismogram showing AVO response for three classes of gas-sand anomalies and a typical reservoir showing the Poisson ratio



Energy Partitioning of Seismic Waves, Fig. 4 AVO curves for three classes of gas-sand anomalies.

than that of the reservoir rock, lowering the impedance of reservoir rock by hydrocarbons may invert the sign of reflection, producing a polarity reversal. The sand types and presence of gas can be identified from the classical A-B crossplot. Recently, Hussein et al. (2020) illustrated this by categorizing different gas zones based on crossplot between A and B seismic attributes using a data set in the El-Mansoura field, onshore Nile Delta in Egypt.

Summary

Because of energy partitioning, we can record both P- and S-waves using multicomponent sensors, and derive both the P- and S-wave seismic velocities and their ratio (V_p/V_s), which can provide reliable and more accurate information than an individual wave can do. The AVO tools have been very efficient in oil industry for the identification of lithology and delineation of fluids content. The AVO response can also be utilized in determining the overpressure, which is prerequisite to avoid any drilling hazard.

Acknowledgments I am grateful to the Director, Wadia Institute of Himalayan Geology for permission. The Ministry of Earth Sciences, Govt. of India is thankfully acknowledged for sponsoring the gas-hydrate research at CSIR-NGRI, Hyderabad, India.

Bibliography

- Arun KP, Sain K, Kumar J (2018) Elastic parameters from constrained AVO inversion: application to a BSR in the Mahanadi offshore, India. *J Nat Gas Sci Eng* 50:90–100
- Arun KP, Sain K, Kumar J (2020) Application of constrained AVO inversion: 2-D modeling of gas hydrates and free gas in Mahanadi basin, India. *J Nat Gas Sci Eng* 78(103287):1–19
- Castagna JP, Backus MM (1993) Offset dependent reflectivity: theory and practice of AVO analysis. *Soc Expl Geophys, SEG, Tulsa*, 8, p 355
- Castagna JP, Swan HW, Foster DJ (1998) Framework for AVO gradient and intercept interpretation. *Geophysics* 63:948–956
- Dahl T, Ursin B (1992) Non-linear AVO inversion for a stack of anelastic layers. *Geophys Prospect* 40(2):243–265
- Dahroug AM, Sharafeldin SM, Mabrouk WM, Noaman MH (2018) Contribution of integrating seismic coherence and AVO attributes in delineating sand bars reservoirs, offshore Nile Delta, Egypt, a case study. *Egypt J Pet* 27:595–603
- Dash RK, Sain K, Thakur NK (2004) Overpressure detection from seismic AVO response: an application to gas hydrates. *Curr Sci* 86:985–990
- Hussein M, Abu El-Ata A, El-Beihy M (2020) AVO analysis aids in differentiation between false and true amplitude response: a case study of El Mansoura field, onshore Nile Delta, Egypt. *J Pet Explor Prod Technol* 10:969–989
- Ojha M, Sain K (2007) Seismic velocities and quantification of gas hydrates from AVA modeling in the western continental margin of India. *Mar Geophys Res* 28:101–107
- Ojha M, Sain K (2008) Appraisal of gas hydrates/free-gas from VP/VS ratio in the Makran accretionary prism. *Mar Petrol Geol* 25:637–644
- Ojha M, Sain K, Minshull TA (2010) Assessment of gas hydrates saturation in the Makran accretionary prism using the offset dependence of seismic amplitudes. *Geophysics* 75(2):C1–C6
- Rutherford SR, Williams RH (1989) Amplitude-versus-offset variations in gas sands. *Geophysics* 54:680–688
- Shuey RT (1985) A simplification of Zoeppritz equations. *Geophysics* 50:609–614
- Tinivella U (2002) The seismic response to overpressure versus gas hydrate and free gas concentration. *J Seismic Expl* 11:283–305
- Warner M (1990) Absolute reflection coefficients from deep seismic reflections. *Tectonophysics* 173(1–4):15–23
- Waters KH (ed) (1981) *Reflection seismology*. A Wiley Interscience Publication, California 1, pp 6–42
- Zoeppritz K (1919) On the reflection and penetration of seismic waves through unstable layers. In: Goettinger N (ed) *Erdbebenwellen VIII B*. pp 66–84

Equatorial Electrojet

Archana Bhattacharyya

Indian Institute of Geomagnetism, Navi Mumbai, India

Definition

The equatorial electrojet (EEJ) is a narrow band of intense electric current that flows along the dip equator in the day side ionospheric E region with peak current density at an altitude of around 105 km.

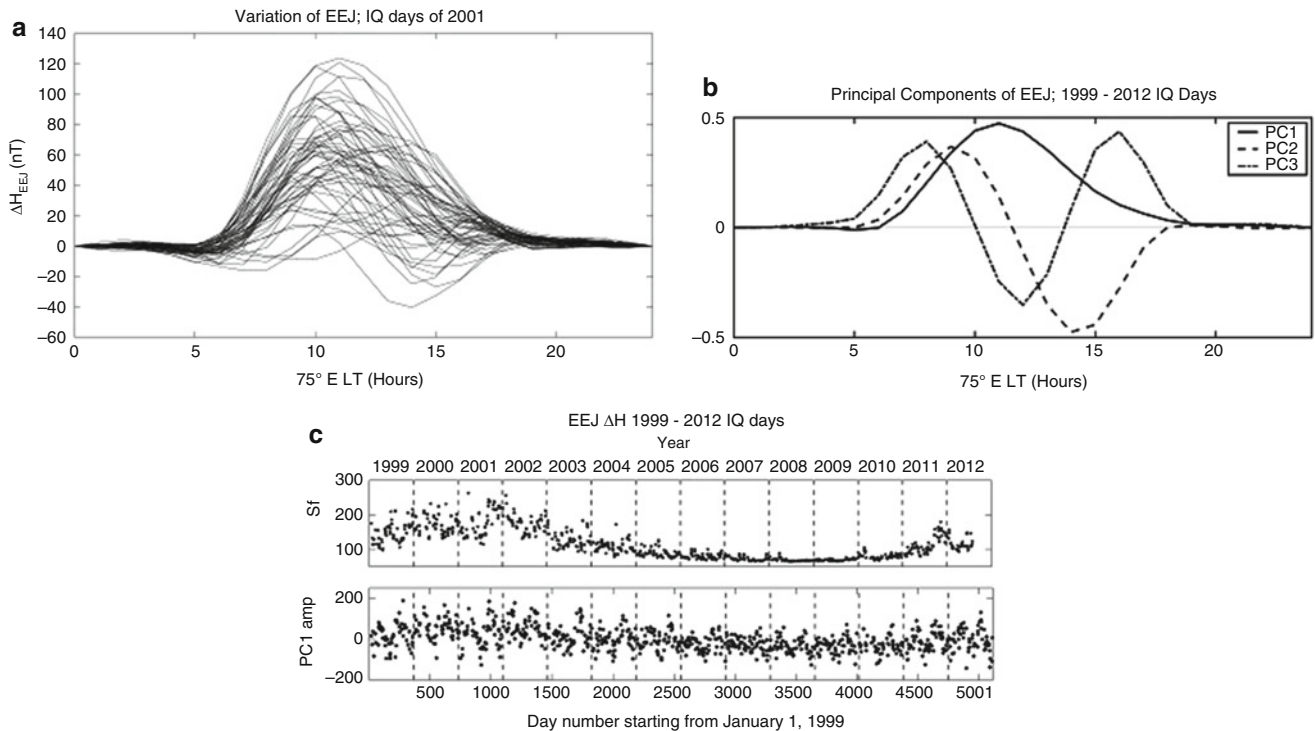
Cause

Electrical conductivity of the partially but significantly ionized upper atmosphere of Earth is anisotropic due to presence of Earth's magnetic field. In the ionospheric E region, movement of electrons in response to an electric field is influenced by the geomagnetic field, whereas for the heavier ions, collisions with neutral particles have much greater influence. This difference in the behavior of electrons and ions gives rise to a large Hall conductivity in the E region. Atmospheric tides, driven mainly by solar heating and to a lesser extent by the gravitational pull of moon, cause the neutral atmosphere to undergo global-scale oscillations, with harmonic periods of a day. These oscillations, through ion-neutral collisions, set the electrically conducting ionospheric plasma into motion, thereby generating an electric field in E region due to dynamo action in the presence of the geomagnetic field. At the dip equator, where the geomagnetic field is horizontal and northward directed, an eastward electric field in magnetic east-west direction drives a downward Hall current, which is restricted due to limited vertical extent of the E region. This causes charges to accumulate producing an upward polarization electric field much larger than the original eastward electric field in a region where Hall conductivity is much larger than Pedersen conductivity. The upward electric field now drives an eastward Hall current that greatly exceeds the original eastward Pedersen current. Together these eastward currents constitute the normal equatorial electrojet (EEJ), which depends on an effective Cowling conductivity (Forbes 1981; Yamazaki and Maute 2017).

Characteristics of the EEJ

The EEJ has been observed using magnetometers on the ground as well as on board rockets and low Earth orbiting

satellites. Geomagnetic field variations recorded at observatories near the dip equator and at low-latitude observatories outside the influence of the EEJ but in the same longitude zone show a typical enhancement of about 40–100 nT in the horizontal component of the geomagnetic field during daytime at the equatorial location, due to the EEJ. Observations by satellites such as Ørsted, SAC-C, and CHAMP have yielded information about spatial features of the EEJ (Olsen and Stolle 2012). Assuming the EEJ to flow in a thin sheet, average current density estimated from CHAMP data is 0.19 A/m, while latitudinal extent of eastward currents of the EEJ at noontime is estimated to be about $\pm 2^\circ$ around the dip equator, where the current peaks. Smaller westward return currents are found to peak at approximately $\pm 5^\circ$ from the dip equator. It is yet to be established how return currents close the EEJ. CHAMP data has shown the EEJ to be correlated over a short longitude range of approximately $\pm 15^\circ$. Figure 1a shows the daily variation of EEJ strength, ΔH_{EEJ} , computed from measurements at a dip equatorial location and an off-equatorial location outside the influence of the EEJ, in the Indian longitude zone, on International Quiet (IQ) days of the year 2001. The EEJ shows considerable day-to-day variations, even for magnetically quiet days. On some days, the magnetic signatures recorded at observatories or by satellites indicate a daytime current flowing in the westward direction. This is called the counter electrojet, a phenomenon not yet understood (Rastogi 1989). The EEJ also varies on seasonal and solar cycle time scales. Role of solar EUV flux in altering ionospheric electrical conductivity and contribution of variability of atmospheric tides in changing the dynamo electric field can be seen in a principal component analysis of the daily variations of EEJ in the Indian region (Bhattacharyya and Okpala 2015). First three principal components (PCs) for IQ days of 14 years, 1999–2012, are shown in Fig. 1b. These three PCs together account for 96% of the variance in ΔH_{EEJ} during this period. Observed daily 10.7 cm solar flux, which is used as a proxy for solar EUV radiation responsible for ionizing Earth's upper atmosphere, is plotted in top panel of Fig. 1c for the days considered. Corresponding daily amplitudes for PC1, which contributes the most to day-to-day variations of quiet time EEJ around 11 LT, are plotted in bottom panel of Fig. 1c. PC1 amplitudes show a semiannual variation with equinoctial maxima, mostly due to upward propagating tides generated in the lower atmosphere, modulated by solar EUV flux. Geomagnetic storms and substorms can also alter the equatorial ionospheric electric field. The EEJ thus reflects changes in equatorial ionospheric electric field, due to both forcing from lower atmosphere and forcing from magnetosphere above due to transient events on the sun.



Equatorial Electrojet, Fig. 1 (a) Daily variation of EEJ strength, ΔH_{EEJ} , in Indian region on IQ days of the year 2001, as a function of local time (LT) at 75°E longitude. (b) First three principal components, PC1, PC2, and PC3, computed using ΔH_{EEJ} for 794 IQ days from the

years 1999–2012. (c) Top – Observed daily 10.7 cm solar flux, Sf (in units of $10^{-22} \text{ W m}^{-2} \text{ Hz}^{-1}$) for the IQ days considered; bottom – Corresponding daily amplitudes for PC1 (in nT)

Uses of EEJ Observations

Electric field in the equatorial ionospheric E region plays a major role in determining spatial distribution of ionospheric plasma throughout the low-latitude ionosphere. This can be seen from global maps of vertical total electron content (TEC) of the ionosphere derived from records of radio signals transmitted from the Global Navigation Satellite Systems (GNSS). Ionospheric plasma distribution affects the performance of satellite-based communication and navigation systems such as GNSS. The EEJ is a useful indicator of the state of the low-latitude ionosphere during daytime. In processing of data collected in magnetic surveys as a part of geophysical exploration for natural resources, it is necessary to remove the diurnal variation in the geomagnetic field due to the EEJ. The EEJ is also a natural source for electromagnetic induction studies. However, its latitudinal extent sets a limit on the depth within the Earth to which electrical conductivity can be determined using temporal variations of the EEJ.

Cross-References

- [Electromagnetic Pulsations and Magnetic Storms](#)
- [Total Electron Content in Seismo-ionospheric Studies](#)

Bibliography

- Bhattacharyya A, Okpala KC (2015) Principal components of quiet time temporal variability of equatorial and low-latitude geomagnetic fields. *J Geophys Res Space Physics* 120:8799–8809. <https://doi.org/10.1002/2015JA021673>
- Forbes JM (1981) The equatorial electrojet. *Rev Geophys Space Phys* 19:469–504
- Olsen N, Stolle C (2012) Satellite geomagnetism. *Annu Rev Earth Planet Sci* 40:441–465. <https://doi.org/10.1146/annurev-earth-042711-105540>
- Rastogi RG (1989) The equatorial electrojet. In: Jacobs J (ed) *Geomagnetism*, vol 3. Academic, San Diego, pp 461–525
- Yamazaki Y, Maute A (2017) Sq and EEJ – a review on the daily variation of the geomagnetic field caused by ionospheric dynamo currents. *Space Sci Rev* 206:299–405. <https://doi.org/10.1007/s11214-016-0282-z>

F

Fault Zone Guided Waves

Peter Eric Malin
Earth and Ocean Sciences, Duke University, Durham,
NC, USA

Synonyms

Fault zone normal and leaky modes; Fault Zone Trapped Waves (FZTW); F_g , F_L , F_R , and F_ϕ

Definition

Fault Zone Guided Waves (FZGW) are elastic waves largely confined to zones of low seismic velocities associated with faults. On seismograms, they are designated by the phase name F_g . FZGW can be viewed as resulting from the constructive interference of P and/or S body waves, either totally or partially reflected from edges of the fault damage zone where seismic velocities increase relative the velocities of the fault core. They can also be modeled as the normal and leaky modes of this same structure.

F_g waves on large faults are analogous to seismic surface and in-seam waves (Aki and Richards 1980; Dresen and Rüter 1994). For those familiar with the former, the Earth structure giving rise to F_g waves can be visualized with the diagram shown in Fig. 1. Like surface waves, F_g waves can be divided into Love and Rayleigh type modes, F_L and F_R , depending on their motion with respect to the plane of the fault. The Love type motion is parallel to this plane, whereas the Rayleigh types include motions perpendicular to the plane. Leaky mode F_g waves have also been observed, the first being PSV F_ϕ waves, also with plane-perpendicular motions (Ellsworth and Malin 2011).

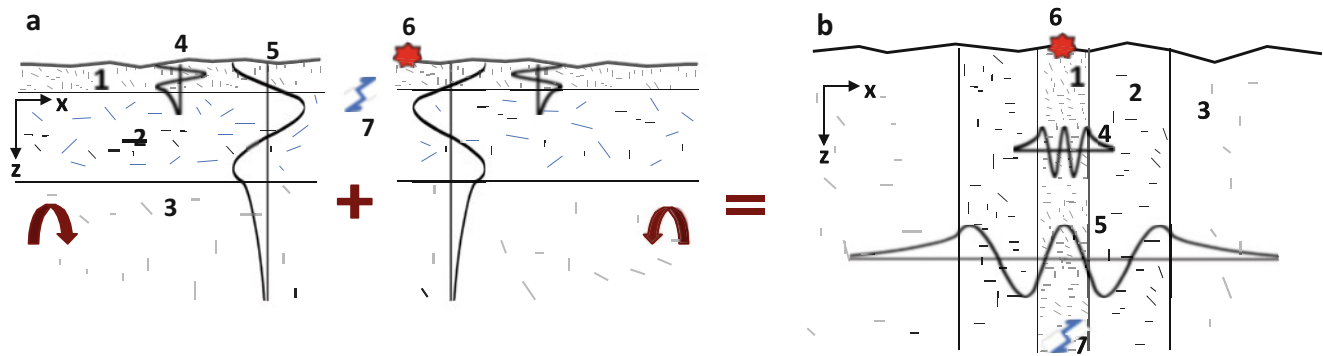
The frequency contents and propagation velocities of F_g waves are functions of the geometry of a fault zone's P and

S wave velocity structure. Modeling of F_L , F_R , and F_ϕ waves in major faults shows these zones to be on the order of 10–200 m wide, with velocity reductions between 5% and 20% of the surrounding rock, with some instances of both much narrower and wider damage zones and up to 50% velocity reductions (e.g., Li et al. 2016, 2019; Catchings et al. 2016). These characteristics have been used to help understand the nature of earthquake rupture. The low-velocity channel causes F_g waves to exhibit frequency-versus-velocity dispersion, with phase velocities of low frequencies being higher than those of higher frequencies. F_g -group velocities can thus show Airy Phase cutoffs.

The geometric spreading of FZGW energy is cylindrical, as opposed to spherical. However, F_ϕ waves attenuate more rapidly due to their leakage of P-S converted waves. Fault-zone discontinuities, heterogeneities, and anisotropies can distort or completely disperse or attenuate F_g waves. As a result, fault segmentations and other kinds of offsets can disrupt F_g wave propagation, a fact that has been used to map fault-zone lateral and vertical extents. However, the ultimate depths to which F_g -supporting low-velocity zones exist continue to be a source of investigation. Current observational limits ranging between a small fraction of a km and 12 km, as in the cases of the San Andreas and San Jacinto faults in California.

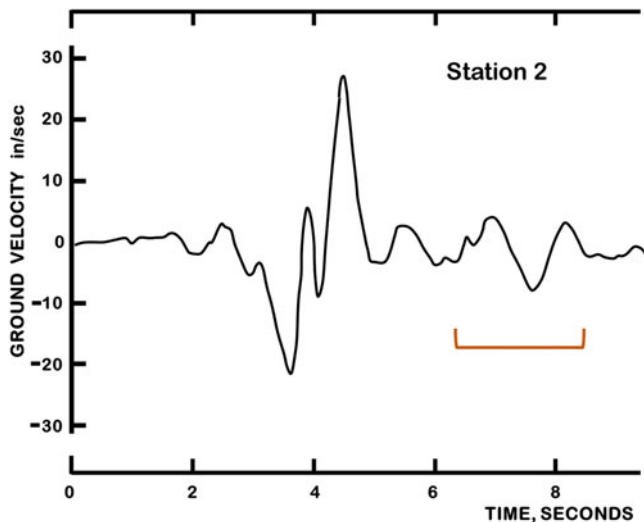
A Short and Recent History of Fault Zone Guided Wave Seismology and Controversies

Two years after the 28 September 1966 M 6.0 Parkfield, California, earthquake on the San Andreas Fault, Keiiti Aki presented a model of the earthquake's strong motion close to the fault. (Aki and Paul Richards wrote the 1980 *Quantitative Seismology* text most currently active seismologists used in their university educations.) Recorded a few tens of meters off the SAF in Chalome Valley the waveform motion included a



Fault Zone Guided Waves, Fig. 1 A way of visualizing fully trapped a Fault Zone Guided Wave by analogy to the generation and propagation of ordinary seismic surface waves. (a) Mirror image, half space, cross sections showing (1) a low-velocity weathering layer over (2) a faster layer, which in turn overlies (3) an even faster basement. The z axis points down into the earth and the x axis along the ground. Surface waves (4 and 5), fully trapped in the low-velocity layers (1 and 2), can be described as interfering plane waves that propagate in the x direction in the lower velocity layers and exponentially die in the z direction in the higher velocity layers (2 and 3). (e.g., see <https://www.acs.psu.edu/drussell/Demos/EvanescenceWaves/EvanescenceWaves.html>.) These waves can be excited by near-surface sources (such as explosions) and earthquakes at appropriate depths. They can also be described as normal modes of the velocity structure. (b) Rotating the cross sections in (a) as

indicated and joining them at their former land surfaces creates two vertical low-velocity channels sandwiched between two higher velocity quarter spaces. The former near-surface weathering layer (1) now represents the core of a fault embedded in a broad damage zone (2). In this case, the interfering plane waves propagate along the z direction and exponential decay is along the x direction. As shown in Fig. 3, especially in the event that the P-wave velocities in the core approach the S-wave velocities in the damage zone, a propagating plane wave may leak out of this boundary as well. A similar situation can occur between the damage zone and the high-velocity quarter spaces. These signals form the leaky modes of this velocity structure. Inseam guided waves can also be visualized by turning diagram (b) 90° clockwise and relocating the ground surface on top of the upper high-velocity layer



Fault Zone Guided Waves, Fig. 2 Strong motion record from the 1966 Parkfield Earthquake (Aki 1968). Shown is the horizontal transverse component perpendicular to the San Andreas Fault rupture plane in Chalome Valley. The underscore indicates a normally dispersed low-frequency pulse that arrived much later than Aki's models of seismic waves in faults described. Aki was aware of this – and bothered by it – and eventually, with the aid of borehole stations directly on the San Andreas Fault at Middle Mountain, was able to resolve its origin as Fault Zone Guided Waves (Li et al. 1990). (Figure modified from Aki (1982)

late, unaccounted-for, wave train (Aki 1968; Fig. 2). Aki was aware of it, and in his reflective way, simply commented that the model did not match the data, but the Earth was not at fault (K. Aki, personal comment, 1989).

Around this same time, Karen McNally, working in Tom McEvilly's seismology group at University of California Berkeley, and then later as Professor at UC Santa Cruz, also noted near-fault, late-arriving phases. Often these signals had larger amplitudes and lower frequencies than any of the direct body waves (K. McNally, personal communication, 1990). Generated by deep microearthquakes, these seismic waves were decidedly not ordinary surface waves trapped in near-surface, low-velocity layers.

Relatively low frequency and apparently normally dispersed, the puzzle of these near-SAF fault signals began to unravel when numerous examples were found in near-fault borehole recordings from Middle Mountain, north of Parkfield. The first published observational report on these examples was by Yong-Gang Li, with associated theoretical analysis by Yehuda Ben-Zion, both of whom were working in Aki's lab at USC (Li et al. 1990; Ben-Zion 1990). (For a reference to borehole seismic observations at Parkfield and elsewhere see: Malin et al. 1989; https://seismo.berkeley.edu/bdsn/hrsn_overview.html; Ellsworth and Malin 2011; Bohnhoff and Malin, in this volume. Earlier work with models and observations suggestive of the existence of FZGW were presented by Cormier and Spudich 1984).

Li and Ben-Zion each went on to publish numerous fault guided wave observational and theoretical studies, references to which can be found variously in Li (2018) and Ben-Zion (1998). Li and Ben-Zion also landed on opposite sides of an amicable debate on the depth at which F_g waves could

originate, with the implication being the depth to which velocity contrasts of the sort shown in Fig. 1 extend. This question is reviewed on both sides by Yang (2015).

Ben-Zion et al. (2007), for example, argue that F_g -supporting velocity contrasts do not extend much deeper than 5 km at most, with the most definitive supporting data coming from the San Jacinto fault in southern California (Qin et al. 2018). The apparent discrepancy between F_g -generating events located deeper than 5 km being resolved by scattering or other types of coupling in shallower low-velocity fault zones.

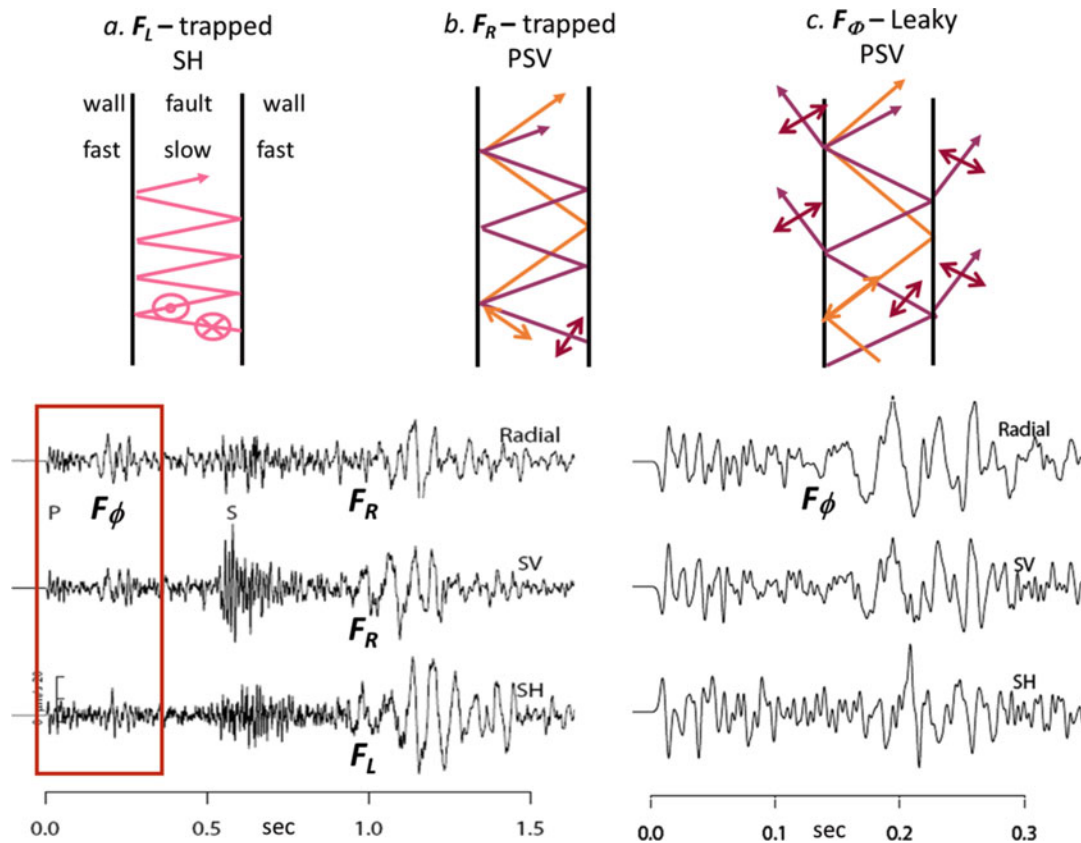
Alternatively, claims of F_g waves originating directly from low-velocity zones extending to 7.2, 10.6, 11.7, and 12.6 km depths along the San Andreas are presented by Ellsworth and Malin (2011), Li and Malin (2008), and Wu et al. (2010).

Perhaps these differences originate in the location and maturity of the fault supporting the guided wave as opposed to complicated wave propagation and scattering effects.

Types of Fault Zone Guided Waves and Their Particle Motions

When fully trapped by total reflection from either the fault zone's internal velocity structure or the damage zone's edges, F_g waves can be separated into two types of signals (Malin et al. 1996; Fig. 3a, b). Analogous to the case of ordinary surface waves and their PSV and SH relation to the ground surface, this separation is based on wave motions with respect to the plane of the fault zone. Fault guided wave motions parallel to the fault plane consist of constructively interfering S-waves and are known as Love types, which are designated by the phase name F_L . Motions with components perpendicular to the damage plane are known as Rayleigh types and are designated as F_R waves. Both F_L and F_R phase velocities range between the highest and lowest S-wave velocities of the fault zone (e.g., Ben-Zion 1998).

Fault guided waves can also become only partially trapped in a damage zone, leaking part of their energy out of this zone. As a result, these signals attenuate more rapidly (Fig. 3c).



Fault Zone Guided Waves, Fig. 3 Upper: (a) Love type motion (F_L) in a low-velocity fault zone between half spaces. The circles with a dot and cross represent back and forth motion parallel to the plane of the fault zone. (b) Motion for the Rayleigh type (F_R) guided wave, the double ended arrows signifying its constructively interfering P and S components. (c) Motion for the leaky mode (F_ϕ) guided wave.

Lower: Fault Zone Guided Waves observed in the San Andreas Fault Observatory at Depth (Ellsworth and Malin 2011). Each wave type is identified by its direction of motion with respect to the plane of this fault. The box on the left is expanded on the right to better reveal the normal dispersion of the F_ϕ wave train. (Figure modified from Ellsworth and Malin 2011)

Currently, due probably to this attenuation, only P-to-S type leaky F_g waves have been recognized in seismograms. In fact, P-to-S type leaky F_g waves have only been recorded in deep boreholes drilled into major faults like the San Andreas Fault (Ellsworth and Malin 2011) and in high-resolution, dense, multichannel fault observatories (Qin et al. 2018).

This phase, known as an F_ϕ wave, is most strongly excited when the S-wave velocity outside either the fault zone core or the damage zone is near the P-wave velocity inside it. Its particle motion still has fault-zone perpendicular components, but the F_ϕ wave propagates faster than the wall-rock-velocity of the direct S-wave. Other types of leaky F_g waves, originating from the many organ pipe modes of a given fault zone's velocity structure, have yet to be documented. Such rapidly attenuating signals are likely to be observed only in boreholes drilled close to their fault zone sources and on dense, multisensor, arrays covering the fault core, damage zone, and surrounding country rock.

The Effects of Fault Geometry, Velocity Structure, Anisotropy, and Heterogeneity

Bracketing the current endmembers of FZGW observation-modeling in 1-D and 3-D velocity structures are the studies of Igel et al. (2002), Jahnke et al. (2002), and Gulley et al. (2017a, b). The former two studies deal with sources and receivers on and off layered fault zones, restrictions along faults, low-velocity zone offsets, and 3-D heterogeneity. Their results investigating on- and off-fault sources were expanded upon by Fohrmann et al. (2004). Some of the fundamental findings include:

- Lateral and vertical fault-zone discontinuities more significantly alter F_g waveforms
- Moderate velocity gradients, small heterogeneities, and changes in width have little effect
- Sources outside and below shallow fault zones (<5 km) can generate F_g signals in faults
- Alternatively, the shallow part of a deep fault zone allows for the trapping of F_g signals

The two papers by Gulley et al. discuss modeling methods and effects of faults with cross-fault velocity gradients rather than layers, fault plane parallel isotropy and anisotropy. They find that:

- Fault models with exponentially decaying velocities toward their cores differ from three-layer ones with the Airy phases being absent
- For small wall velocity contrasts, F_g in transverse-isotropy faults depend on fault parallel S-velocities

- Fault-aligned structures enhance the amplitude and length of F_R
- Moving sources or receivers away from the fault removes higher frequencies F_R first

All these facets of F_g -waves impact the use of observational data to infer fault-zone properties and their relationship to earthquake rupture processes (e.g., Kuwahara and Ito 2002).

Conclusions

Whereas observations and modeling of basic F_g waves F_L , F_R , and F_ϕ in relatively complex fault structures is now well established, other types of modes have yet to be discovered and described. It is likely that other types of leaky mode F_g waves will attenuate rapidly away from their earthquake sources. This makes it likely that leaky mode F_g waves will be most readily observed in deep boreholes drilled close to event hypocenters or then on dense arrays of seismographs on shallow faults. Currently, forward modeling appears to be the preferred method for extracting fault zone properties from F_g observations, due primarily to an insufficient amount of lateral and vertical observational coverage for FZGW tomography along major faults. Compared to the century-long history of mapping of the Earth's complex velocity structure with seismic waves, mapping of fault-zone properties, widths, depths, segmentation, heterogeneity, anisotropy, and healing after earthquakes using F_g -waves is just beginning.

Cross-References

- Body Waves
- Borehole Seismic Networks and Arrays
- Continental Crustal Structure
- Earthquake, Focal Mechanism
- Earthquake, Magnitude
- Earthquakes and Crustal Deformation
- Earthquakes, Strong-Ground Motion
- Energy Partitioning of Seismic Waves
- Magnetic Anisotropy
- Numerical Methods, Finite Difference
- Propagation of Elastic Waves: Fundamentals
- Seismic Diffraction
- Seismic Imaging, Overview
- Seismic Instrumentation
- Seismic Phase Nomenclature: The IASPEI Standard
- Seismic Signals in Well Water Observations
- Seismic Wave Propagation in Real Media: Numerical Modeling Approaches
- Seismic Waves, Scattering
- Seismic, Reflectivity Method

- Seismogram Interpretation
- Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting
- Single and Multichannel Seismics
- Slow Earthquake
- Surface Waves

Bibliography

- Aki K (1968) Seismic displacement near a fault. *J Geophys Res* 73:5359. <https://doi.org/10.1029/JB073i016p05359>
- Aki K (1982) Strong motion prediction using math model techniques. *Bull Seismol Soc Am* 72(6):29–41
- Aki K, Richards PG (1980) Quantitative seismology: theory and methods. WH Freeman and Company, New York
- Ben-Zion Y (1990) The response of two half spaces to point dislocations at the material interface. *Geophys J Int* 101(3):507–528. <https://doi.org/10.1111/j.1365-246X.1990.tb05567.x>
- Ben-Zion Y (1998) Properties of seismic fault zone waves and their utility for imaging low-velocity structures. *J Geophys Res Sol EA* 103:12567–12585. <https://doi.org/10.1029/98JB00768>
- Ben-Zion Y, Peng Z, Lewis M, McGuire J (2007) High-resolution imaging of fault zone structures with seismic fault zone waves. *Sci Drill.* <https://doi.org/10.2204/iodp.sd.s01.23.2007>
- Catchings RD, Goldman MR, Li Y-G, Chan JH (2016) Continuity of the West Napa–Franklin fault zone inferred from guided waves generated by earthquakes following the 24 august 2014 M_w 6.0 South Napa earthquake. *Bull Seismol Soc Am* 106(6):2721–2746. <https://doi.org/10.1785/0120160154>
- Cormier VF, Spudich P (1984) Amplification of ground motion and waveform complexity in fault zones: examples from the San Andreas and Calaveras Faults. *Geophys J Int* 79(1):135–152. <https://doi.org/10.1111/j.1365-246X.1984.tb02846.x>
- Dresen L, Ruter H (1994) Seismic coal exploration: in-seam seismics. Pergamon Technol Eng. 443pp
- Ellsworth WL, Malin PE (2011) Deep rock damage in the San Andreas Fault revealed by P- and S-type fault-zone-guided waves. *Geol Soc London Spec Publ* 359:39–53. <https://doi.org/10.1144/SP359.3>
- Fohrmann M, Igel H, Jahnke G, Ben-Zion Y (2004) Guided waves from sources outside faults: an indication shallow fault zone structure? *Pure Appl Geophys* 161:2125. <https://doi.org/10.1007/s00024-004-2553-y>
- Gulley AK, Kaipio JP, Eccles JD, Malin PE (2017a) A numerical approach for modelling fault-zone trapped waves. *Geophys J Int* 210(2). <https://doi.org/10.1093/gji/ggx199>
- Gulley AK, Eccles JD, Kaipio JP, Malin PE (2017b) The effect of gradational velocities and anisotropy on fault-zone trapped waves. *Geophys J Int* 210(2):964. <https://doi.org/10.1093/gji/ggx200>
- Igel H, Jahnke G, Ben-Zion Y (2002) Numerical simulation of fault zone trapped waves: accuracy and 3-D effects. *Pure Appl Geophys* 159:2067–2083. <https://doi.org/10.1007/s00024-002-8724-9>
- Jahnke G, Igel H, Ben-Zion Y (2002) Three-dimensional calculations of seismic fault zone waves in various irregular structures. *Geophys J Int* 151:416. <https://doi.org/10.1046/j.1365-246X.2002.01784.x>
- Kuwahara Y, Ito H (2002) Fault low velocity zones deduced by trapped waves and their relation to earthquake rupture processes. *Earth Planets Space* 54:1045–1048. <https://doi.org/10.1186/BF03353299>
- Li Y (2016) Seismic wave propagation in anisotropic rocks with applications to defining fractures in earth crust, Chapter 1. In: Rock anisotropy, fracture and earthquake assessment. China High Education Press with De Gruyter, Beijing/Boston, pp 11–141
- Li Y-G (2017) Fault-zone guided wave, ground motion, landslide and earthquake forecast. De Gruyter. <https://doi.org/10.1515/9783110543551-011>
- Li Y-G, Malin PE (2008) San Andreas Fault damage at SAFOD viewed with fault-guided waves. *Geophys Res Lett* 35:L08304. <https://doi.org/10.1029/2007GL032924>
- Li Y-G, Leary P, Aki K, Malin PE (1990) Seismic trapped modes in the Oroville and San Andreas Fault zones. *Science (New York, NY)* 249:763–766. <https://doi.org/10.1126/science.249.4970.763>
- Malin PE, Blakeslee SN, Alvarez MG, Martin AJ (1989) Microearthquake imaging of the Parkfield asperity. *Science* 244(4904):557–559. <https://doi.org/10.1126/science.244.4904.557>
- Malin PE, Lou M, Rial J (1996) F R waves: a second fault-guided mode with implications for fault property studies. *Geophys Res Lett* 23:3547–3550. <https://doi.org/10.1029/96GL03424>
- Qin L, Ben-Zion Y, Qiu H, Share P-E, Ross ZE, Vernon FL (2018) Internal structure of the San Jacinto Fault zone in the trifurcation area southeast of Anza, California, from data of dense seismic arrays. *Geophys J Int* 213(1):98–114. <https://doi.org/10.1093/gji/ggx540>
- Wu J, Hole JA, Snoke JA (2010) Fault zone structure at depth from differential dispersion of seismic guided waves: evidence for a deep waveguide on the San Andreas Fault. *Geophys J Int* 182(1):343–354. <https://doi.org/10.1111/j.1365-246X.2010.04612.x>
- Yang H (2015) Recent advances in imaging crustal fault zones: a review. *Earthq Sci* 28(2):151–162. <https://doi.org/10.1007/s11589-015-0114-3>

Fiber Optic Distributed Strain Sensing for Seismic Applications

Thomas Reinsch¹, Philippe Jousset¹ and Charlotte M. Krawczyk^{2,1}

¹GFZ German Research Centre for Geosciences, Potsdam, Germany

²Institute for Applied Geosciences, Technical University Berlin, Berlin, Germany

Definition

Fiber optic distributed strain sensing relies on the interrogation of an optical fiber using an appropriate light source, typically a laser. When light travels through an optical fiber, it interacts with the fiber glass structure. The interaction can be elastic (Rayleigh) or inelastic (Raman or Brillouin) scattering phenomena. Among other, the scattered light can carry information about the current elongation of the optical fiber. Comparing the elongation from successive measurements, strain changes can be detected with a high temporal and spatial resolution. If a fiber optic cable is coupled to the ground, these strain changes are a measure of the ground motion and can hence be used for seismic applications.

Introduction

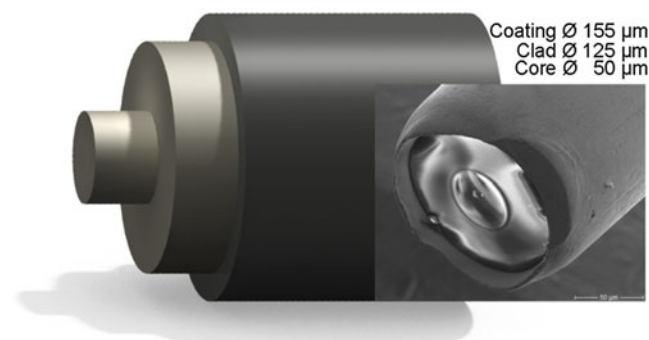
To determine crustal properties distribution, seismic source processes, and wave propagation mechanisms, the acquisition

of dense seismic and ground motion datasets is required (e.g., Jousset et al. 2018). Seismic and ground-motion datasets are typically acquired measuring acceleration, velocity (e.g., geophones and broadband sensors), or position (e.g., GNSS sensors) with individual sensors located in a favorable geometry. A high-resolution image of the subsurface, thereby, requires a dense spatial coverage of sensors.

In recent years, fiber optic sensor technologies have been increasingly used for seismic applications. Different acquisition strategies have been applied to measure the subsurface deformation: measurements at a single point using specially engineered fiber optic point sensors, quasi-distributed measurements along an optical fiber with an array of point sensors, or fully distributed, where deformation is detected using interaction of light with the molecular structure of the fiber itself. We will consider fully distributed technologies in the following, only.

Optical Fibers

When light is coupled into an optical fiber, it is typically guided within the core of the fiber. Optical fibers are mainly made of doped silica glass to engineer a desired refractive index. The core of a fiber is surrounded by a clad, also made of silica but with a lower refractive index. For a given angle of incidence, light propagating within the core towards this interface fulfils the condition for total reflection, is reflected back towards the core and hence confined within the fiber. If only a single mode of propagation is possible for a given wavelength, these types of fibers are called “single-mode fibers,” in contrast to “multi-mode fibers” where multiple propagation modes are possible. Single-mode fibers that are mostly used for distributed strain sensing have typical diameters for the core/clad of 8–10/125 μm , multi-mode fibers often 50 or 62.5/125 μm , respectively (Refi 1998). In order to protect the optical fiber from external forces and to preserve the fiber’s mechanical strength, a protective coating is applied (Refi 1998) with typical diameters of, e.g., 155 μm (Fig. 1).



Fiber Optic Distributed Strain Sensing for Seismic Applications, Fig. 1 Schematic of a multimode optical fiber. Inset: Scanning electron microscopy image of a multimode fiber

Depending on the application, mostly polymer (e.g., acrylate) coatings materials are used. For high-temperature applications, also metal coatings (e.g., aluminium, copper, or gold) are manufactured.

Cable Coupling

To assess the coupling between the optical fiber and the subsurface, a detailed evaluation of the cable structure and the cable installation procedure is required. Typically, optical fibers are embedded in a multilayer cable structure. Within such cables, the optical fibers are embedded in a tube surrounded by a gel-type filling compound (Refi 1998). Depending on the application, the tube material is either some hard plastic or metal. In addition, to the tube with the optical fiber, at least some strength members (e.g., steel wires or Kevlar fibers) are used to give the cable the required strength. For most cables, a polymer jacket is extruded on the cable. For long wavelength and small amplitude signals, typical for seismic applications, the cable structure can be assumed to behave fully elastic. In addition, friction at the surface of the cable is generally sufficient for a good strain transfer between optical fiber and the ground for seismic frequencies (Reinsch et al. 2017). However, if short wavelength or high amplitudes are of interest, a tight cable design is required and the transfer function for the strain transfer to the optical fiber should be studied in detail.

Determination of Strain Changes Along Optical Fiber

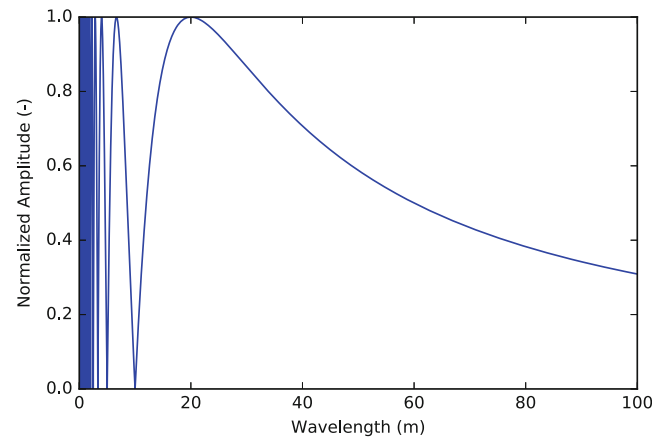
When light travels through an optical fiber, part of the light is scattered at random inhomogeneities within the optical fiber. While the dominant scattering phenomenon is elastic Rayleigh scattering, also inelastic Raman and Brillouin scattering phenomena occur. When light is scattered within the optical fiber, a fraction of the light is captured and travels back into the direction of the incident light. For an incident light pulse and an appropriate receiver set-up, the two-way travel-time of the light can be used to locate the scattering position along the fiber. This method is called optical time domain reflectometry (OTDR) (Barnoski and Jensen 1976; Personick 1977) and was introduced as a measure for the signal attenuation along optical fibers and fiber optic installation ▶ “Fiber-Optic Sensing in Geophysics, Temperature Measurements”. Based on the OTDR technology, several measurement principles capable of quantifying distributed dynamic strain changes along the entire optical fiber have been developed (e.g., Masoudi and Newson 2016, and references therein). The technology of distributed strain sensing, interrogated at seismic frequencies, is typically based on elastic Rayleigh

scattering from a coherent laser pulse and referred to as distributed vibration sensing (DVS) or distributed acoustic sensing (DAS) (Hartog et al. 2013). We will use the term DAS in the following. If a coherent laser pulse is launched into the fiber, not only the amplitude but also the phase of the back-scattered photons can be analyzed (Φ -OTDR). The phase of the backscattered photons is sensitive to both, temperature and strain changes along the optical fiber. As the temporal scale of temperature changes is typically long compared to strain changes induced by seismic waves, the measured signal can be attributed to the detection of ground motion. To detect dynamic phase changes along an optical fiber, for example, an imbalanced Mach-Zehnder interferometer has been used (Posey et al. 2000). For any section of the fiber, the phase-difference $\Delta\Phi$ of photons scattered at both ends of that sections is linearly related to the length of this section (Masoudi and Newson 2017). When the section of the sensing fiber is unperturbed, the length and, consequently, the phase difference $\Delta\Phi$ remains unchanged. Any perturbation inducing a strain ϵ on the fiber will change that difference. Measurements, capable to resolve perturbations, in the nc-range are reported (Masoudi and Newson 2017). A fiber optic cable can hence be considered a system consisting of a large number of one component relative strain gauges. The length of these sections is typically referred to as gauge length.

DAS Versus Conventional Seismic Sensing

Conventional seismological sensors probe ground velocity or acceleration at specific locations, e.g., point sensors. DAS sensors measure average strain along the fiber over the gauge-length (typically 10 m). At a single location, The measurement of deformation is insensitive to the direction of the incoming signal. Whether a seismic wave arrives from one side or the other towards a given interval in the optical fiber will both lead to the same deformation – contraction or extension (Frignet and Hartog 2014; Papp et al. 2017). Only by comparing neighboring traces, the spatial propagation of the incident wave can be compared and the polarity of the signal can be analyzed.

DAS is sensitive to strain changes in direction of the optical fiber, only. To overcome the limitations of a sensor array with one component channels, various cable designs have been proposed and tested to increase broadside sensitivity. One design, the “helically wound cable”, where the optical fiber is wrapped under a specific angle around a central cable member, gained special interest in literature. Kuvshinov (2015) theoretically describes such a helically wound cable that was extensively tested in the field by Hornman (2016). The author was able to show that the helically wound fiber is capable of detecting p-waves independent of the angle of incidence.



Fiber Optic Distributed Strain Sensing for Seismic Applications, Fig. 2 Normalized maximum amplitude with respect to wavelength for a gauge length of 10 m

The acquisition of strain or strain changes over a specified gauge length introduces a spatial filter (e.g., Dean et al. 2017). Figure 2 shows the effect of the geometric filter caused by an acquisition system with a gauge length of 10 m and an incident wave in axial direction. As the optical fiber is interrogated with a laser pulse of a given duration (i.e., pulse width = length in the optical fiber), typically a few meters of fiber are illuminated at the same time. As light is back-scattered from a larger interval, the pulse width also effects the amplitude response of the DAS system (Dean et al. 2017). Whereas algorithms in seismology typically rely on processing in the frequency domain, distributed strain sensing data allows for and requires processing in both, time and space or frequency and wavenumber.

Localization of the Sensor Position

Typically, conventional seismic point sensors are localized using accurate GPS measurements. For fully distributed fiber optic sensors, this concept has to be modified. Distributed fiber optic data is recorded as a function of the optical fiber length. Although the position of the fiber optic cable is known with a high accuracy, the position of individual traces along the optical fiber can differ from a simple correlation of the distance between the readout unit and individual positions along the fiber optic cable. Aspects to consider are:

- Fiber optic cables are never installed without an excess length at the termination at both ends. For splicing fibers, sufficient excess cable length is required to pull the cable to a good working environment. This is especially true for connections in subsurface installations, where the cable is pulled into a tent or a splicing van. In order to have a sufficient safety margin to allow for splicing connectors

to the optical fibers, at least a couple of meters of optical fibers are placed in so-called splice closures. Such is also true if there is a splice somewhere along the fiber optic cable to be measured (Refi 1998).

- If the connector at the end of the optical fiber is different from the socket, a the readout-unit, an additional patch cord or cross-over cable of a few meters has to be used.
- The distance of individual traces along an optical cable is determined by the two-way travel time of light from a readout unit multiplied by the speed of light in the optical fiber. The speed of light can be expressed in terms of the refractive index of the glass. A slight inaccuracy in the refractive index determination can easily add up to a positioning error of several meters for a measured length of a few kilometers.
- Typical fiber optic cables are produced with an excess length of optical fiber. The excess length is designed such that even when stretching the cable under load, the fiber is not stretched more than 0.33% (Refi 1998). Depending on the application, this design condition typically results in an excess length of a few per-mill.

Due to the different elements detailed above, the measured length between individual position along the optical fiber, determined by the time of flight for the light, does not necessarily correspond to a geographical distance in a given reference coordinate system. In order to associate the geographical position of any measured trace along the cable, signal sources with a known position and a known characteristic can be used. For surface application, hammer shots can be used. In down-hole applications, tapping on the wellhead can generate a sufficient signal. Such a procedure is suitable to locate individual traces with a meters accuracy.

Selected Case Studies

Natural hazard assessment and efficient crust exploration require dense seismic observations both in time and space. Seismological techniques provide ground-motion data, whose accuracy depends on sensor characteristics and spatial distribution. Dynamic distributed strain sensing allows for a high spatial resolution that can be achieved along fiber optic installations. For downhole installations, several studies were performed to evaluate the potential of Vertical Seismic Profiling, by Rector and Mangriotis surveys. Generally, it has been shown that although fiber optic data has a lower signal to noise ratio compared to conventional borehole seismic sensors, the higher spatial resolution allows to acquire high resolution exploration results (e.g., Mestayer et al. 2011; Cox et al. 2012; Daley et al. 2013; Götz et al. 2018).

The densification of ground motion sensors, offered by the fiber optic DAS technology, not only allows for a spatially

un-aliased data but also for the detection of microseismic events, typically undetected by conventional seismic networks. Li and Zhan (2018) used template matching methods to analyze DAS data from the Brady Hot Springs geothermal field, Nevada, USA. The authors found a significant number of microseismic events below the noise level in the DAS data and could extend the local earthquake catalogue to study the correlation of microseismicity and subsurface pressure changes associated to the operation of a nearby geothermal power plant. On the same array, Wang et al. (2018) compared DAS data with ground motion data from colocated geophones. The authors compared the signal to noise ratio (SNR) from both recordings and found that the SNR of a single DAS channel is typically lower by a factor of two when compared at frequencies of a few Hertz but increases for lower frequencies. In recent years, it has been shown that dynamic strain determination is also possible with conventional fiber-optic cables deployed for telecommunication. Lindsey et al. (2017) showed that it is possible to monitor ground motion from regional seismic and teleseismic events using telecommunication fiber-optic cables. Dou et al. (2017) and Ajo-Franklin et al. (2019) recorded ambient noise using DAS and showed the potential of DAS data for time-lapse imaging of shear wave velocity structures. Dou et al. (2017) were able to achieve a repeatability of 2% in the model domain for the uppermost 20 m. Jousset et al. (2018) acquired high resolution spatially un-aliased broadband strain data associated to seismic signals from natural and man-made sources on Reykjanes Peninsula, Iceland. Structural features, such as normal faults and volcanic dykes were identified in the Reykjanes Oblique Rift with unprecedented resolution (4-m data spacing) along a 15 km fiber-optic cable layout. Conventional seismometer recordings, acquired simultaneously, validated the spectral amplitude of the DAS response between 0.1 and 100 Hz bandwidth. Jousset et al. (2018) suggested that the networks of fiber-optic telecommunication lines worldwide could be used as seismometers opening a new window for seismic hazard assessment and exploration.

More and more applications in various fields of Earth Sciences are emerging, e.g., measurements in marine environment, taking advantage of the network of optical cables (e.g., Lindsey et al. 2019; Marra et al. 2018; Sladen et al. 2019; Williams et al. 2019; Jousset 2019). First measurements were also performed on active volcanoes and cities build on the flanks of active volcanoes (Jousset et al. 2019) using infrastructure of temporary or permanent monitoring networks (Contrafatto et al. 2019).

Summary and Conclusions

Fiber optic distributed strain sensing has shown to be capable to complement conventional ground motion sensors, such as,

e.g., geophones or broad band stations. Individual fiber optic sensor positions have a lower signal to noise ratio compared to conventional sensors. Furthermore, fiber optic sensors are typically only sensitive in the direction of an optical fiber. The high densification of distributed sensor positions, however, offers the opportunity for unprecedented subsurface imaging. Due to the continuous improvement of broadband internet access, fiber optic infrastructure is pervasive in urban environments. Also, numerous marine cables crossing all major oceans and connecting almost all parts of the world are already deployed. In addition to the installation of dedicated sensing fibers, utilizing such an infrastructure for distributed sensing opens up an opportunity for a plethora of novel applications. Dense data acquisition will help to further unravel earth's structure and dynamics.

Acknowledgments The authors would like to acknowledge the ZELMI at Berlin University of Technology for the SEM image. This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 654497 (GeoWell) and 691728 (DESTRESS). This work was also supported by the GFZ German Research Centre for Geosciences.

Bibliography

- Ajo-Franklin JB, Dou S, Lindsey NJ, Monga I, Tracy C, Robertson M, Tribaldos VR, Ulrich C, Freifeld B, Daley T, Li X (2019) Distributed acoustic sensing using dark fiber for near-surface characterization and broadband seismic event detection. *Sci Rep* 9(1). <https://doi.org/10.1038/s41598-018-36675-8>
- Barnoski MK, Jensen SM (1976) Fiber waveguides: a novel technique for investigating attenuation characteristics. *Appl Opt* 15(9):2112–2115
- Contrafatto D, Currenti G, Jousset P, Larocca G, Messina A, Pellegrino D, Pulvirenti M, Rapisarda S, Zuccarello L, Chalari A, Krawczyk CM (2019) Monitoring volcanic activity with fiber-optic distributed acoustic sensing – first experiments at the Etna volcano. *Geophysical Research Abstracts*. 21(EGU2019-13088-1). EGU General Assembly 2019
- Cox B, Wills P, Kiyashchenko D, Mestayer J, Lopez J, Bourne S, Lupton R, Solano G, Henderson N, Hill D, Roy J (2012) Distributed acoustic sensing for geophysical measurement, monitoring and verification. *CSEG Rec* 37(2):7–13
- Daley TM, Freifeld BM, Ajo-Franklin J, Dou S, Pevzner R, Shulakova V, Kashikar S, Miller DE, Goetz J, Henningses J, Lueth S (2013) Field testing of fiber-optic distributed acoustic sensing (DAS) for subsurface seismic monitoring. *Lead Edge* 32(6):699–706
- Dean T, Cuny T, Hartog AH (2017) The effect of gauge length on axially incident p-waves measured using fiber optic distributed vibration sensing. *Geophys Prospect* 65(1):184–193
- Dou S, Lindsey N, Wagner AM, Daley TM, Freifeld B, Robertson M, Peterson J, Ulrich C, Martin ER, Ajo-Franklin JB (2017) Distributed acoustic sensing for seismic monitoring of the near surface: a traffic-noise interferometry case study. *Sci Rep* 7(1). <https://doi.org/10.1038/s41598-017-11986-4>
- Frignet BG, Hartog A (2014) Optical vertical seismic profile on wireline cable. In: SPWLA 55th annual logging symposium
- Götz J, Lüth S, Henningses J, Reinsch T (2018) Vertical seismic profiling using a daisy-chained deployment of fiber-optic cables in four wells simultaneously – case study at the ketzin carbon dioxide storage site. *Geophys Prospect* 66(6):1201–1214
- Hartog AH, Kotov O, Liokumovich L (2013) The optics of distributed vibration sensing. In: Second Eage workshop on permanent reservoir monitoring 2013 – current and future trends. EAGE. <https://doi.org/10.3997/2214-4609.20131201>
- Hornman J (2016) Field trial of seismic recording using distributed acoustic sensing with broadside sensitive fiber-optic cables. *Geophys Prospect* 65(1):35–46
- Jousset P (2019) Illuminating Earth's faults. *Science*. 366(6469):1076–1077. <https://doi.org/10.1126/science.aaz7750>
- Jousset P, Reinsch T, Ryberg T, Blanck H, Clarke A, Aghayev R, Hersir GP, Henningses J, Weber M, Krawczyk CM (2018) Dynamic strain determination using fiber-optic cables allows imaging of seismological and structural features. *Nat Commun* 9(1). <https://doi.org/10.1038/s41467-018-04860-y>
- Jousset P, Currenti G, Tilmann F, Zuccarello L, Chalari A, Reinsch T, Krawczyk C (2019) Towards seismic and volcanic hazard assessment with distributed acoustic sensing in fiber optic cable. *Geophys Res Abstr* 21(EGU2019-17281). EGU General Assembly 2019
- Kuvshinov B (2015) Interaction of helically wound fiber-optic cables with plane seismic waves. *Geophys Prospect* 64(3):671–688
- Li Z, Zhan Z (2018) Pushing the limit of earthquake detection with distributed acoustic sensing and template matching: a case study at the Brady geothermal field. *Geophys J Int* 215(3):1583–1593
- Lindsey NJ, Martin ER, Dreger DS, Freifeld B, Cole S, James SR, Biondi BL, Ajo-Franklin JB (2017) Fiber-optic network observations of earthquake wavefields. *Geophys Res Lett* 44(23):11,792–11,799
- Lindsey NJ, Dawe TC, Ajo-Franklin JB (2019) Illuminating seafloor faults and ocean dynamics with dark fiber distributed acoustic sensing. *Science* 366(6469):1103–1107
- Marra G, Clivati C, Luckett R, Tampellini A, Kronjäger J, Wright L, Mura A, Levi F, Robinson S, Xuereb A, Baptie B, Calónico D (2018) Ultrastable laser interferometry for earthquake detection with terrestrial and submarine cables. *Science*:eaat4458. <https://doi.org/10.1126/science.aat4458>
- Masoudi A, Newson TP (2016) Contributed review: distributed optical fiber dynamic strain sensing. *Rev Sci Instrum* 87(011501):9
- Masoudi A, Newson TP (2017) High spatial resolution distributed optical fiber dynamic strain sensor with enhanced frequency and strain resolution. *Opt Lett* 42(2):290–293
- Mestayer J, Cox B, Wills P, Kiyashchenko D, Lopez J, Costello M, Bourne S, Ugueto G, Lupton R, Solano G, Hill D, Lewis A (2011) Field trials of distributed acoustic sensing for geophysical monitoring. *SEG Tech Prog Exp Abstr* 30(1):4253–4257
- Papp B, Donno D, Martin JE, Hartog AH (2017) A study of the geophysical response of distributed fiber optic acoustic sensors through laboratory-scale experiments. *Geophys Prospect* 65(5):1186–1204
- Personick S (1977) Photon probe-an optical-fiber time-domain reflectometer. *Bell Syst Tech J* 56(3):355–366
- Posey R, Johnson JG, Vohra S (2000) Strain sensing based on coherent rayleigh scattering in an optical fiber. *Electron Lett* 36(20):1688–1689
- Reif JJ (1998) Fiber optic cable. abc Tele Training, Inc, Geneva
- Reinsch T, Thurler T, Jousset P (2017) On the mechanical coupling of a fiber optic cable used for distributed acoustic/vibration sensing applications – a theoretical consideration. *Meas Sci Technol* 28(12):127003. <https://doi.org/10.1088/1361-6501/aa8ba4>
- Sladen A, Rivet D, Ampuero JP, De Barros L, Hello Y, Calbris G, Lamare P (2019) Distributed sensing of earthquakes and ocean-solid Earth interactions on seafloor telecom cables. *Nat Commun*. <https://doi.org/10.1038/s41467-019-13793-z>
- Wang HF, Zeng X, Miller DE, Fratta D, Feigl KL, Thurber CH, Mellors RJ (2018) Ground motion response to an ML 4.3 earthquake using co-located distributed acoustic sensing and seismometer arrays. *Geophys J Int* 213(3):2020–2036
- Williams EF, Fernández-Ruiz MR, Magalhaes R et al (2019) Distributed sensing of microseisms and teleseisms with submarine dark fibers. *Nat Commun* 10:5778. <https://doi.org/10.1038/s41467-019-13262-7>

Fiber-Optic Sensing in Geophysics, Temperature Measurements

Jan Henninges¹ and Ali Masoudi²

¹Geoenergy, GFZ German Research Centre for Geosciences, Potsdam, Germany

²ORC, University of Southampton, Southampton, UK

Definition

Temperature measurements for geophysical applications using fiber-optic methods

Temperature: formation temperature, borehole temperature

Introduction

Temperature measurement has a very long history in geophysics where it has been used for many different applications employing various sensing methods. With most relevance for solid earth geophysics, the direct measurement of temperature in boreholes is considered here.

Within the past decades, continuous developments in fiber-optic sensing have resulted in new sensor types, with certain advantages over conventional methods for temperature measurements. For measurements in boreholes, mainly electronic sensors have previously been used, either incorporated in downhole tools for wireline logging, or deployed as thermistor chains.

There has been a rapid evolution of fiber-optic sensing methods since the 1990s. Apart from temperature, several other physical quantities like pressure or strain and chemical parameters can be measured.

Distributed temperature sensing (DTS) is probably by far the most frequently applied fiber-optic technology. The initial development was driven by the energy sector, for monitoring of power cables and transformers in the electricity transmission networks. In earth sciences, Hurtig et al. (1993) carried out early tests of applying DTS for temperature measurements in boreholes. Wisian et al. (1998) compare DTS to other conventional temperature logging methods. Since the year 2000, there is a strong increase in number of scientific publications, mainly in the oil and gas sector, with a steady growth until now.

Within this entry, we first explain the design of the most common types of optical fibers. We then give an overview of different sensor types and configurations, followed by a description of the most important methods for temperature sensing. Particular emphasis is on Raman DTS, which is the method most frequently applied. Next, we describe how optical fibers are deployed in boreholes. The final section then

gives an overview of the applications of fiber-optic temperature measurements in solid Earth geophysics and related disciplines.

Optical Fibers

Optical fibers for downhole use usually have cylindrical cross-section with several co-axial layers. The light mainly propagates within the core of the fiber, which is surrounded by an outer layer referred to as the “cladding.” Both core and cladding are primarily made of silica glass (SiO_2). In order for a fiber to act as a waveguide, the refractive index of the core is set to be greater than that of the cladding. This is typically achieved by adding small amounts of other elements referred to as “dopants” either to the core or the cladding region of the glass to adjust the refractive index accordingly.

Within the fiber core, the light can follow different paths at different angles with respect to the fiber axis, which are referred to as propagation “modes.” Standard single-mode telecom fibers have a very small core diameter of typically 8–10 μm , and only support one propagation mode. Multi-mode fibers have larger cores with typical diameters of 50 μm . Due to their different path lengths, different modes travel at different velocities along the fiber, which results in a spreading of the light pulse, which is also known as modal dispersion. In a graded-index fiber, the differences in mode velocities are reduced by gradually changing the refractive index from the core to the cladding, resulting in much reduced dispersion.

The outer layer, or coating, of an optical fiber serves to improve the mechanical characteristics of the fiber and protects it from undesirable environmental influences. The coating material is selected according to the temperature and chemical substances the fiber will be exposed to, and can be made of a combination of materials, including plastic (e.g., acrylate, silicone, polyimide) or metals (e.g., aluminum, copper, gold). The coating material primarily determines the temperature tolerance of optical fibers.

Sensor Configurations, Sensing Methods

The field of optical fiber sensors falls into two broad categories, namely point sensors and distributed sensors. A point temperature sensor can quantify the temperature at a discrete, predetermined point. The role of the optical fiber in such a sensor is to transmit the probe light to the point of interest where the sensing node is installed and carry the altered light back to the interrogator for analysis.

The architecture of some of the point sensors allows them to be daisy chained, or multiplexed, to form an array. Such a

sensor array, which is also known as a “quasi-distributed” sensor, can be interrogated by a single device, hence significantly reducing the per-node cost of such sensing systems.

Point Sensors, Sensor Arrays

A wide range of optical phenomena have been employed to realize single point optical fiber temperature sensors. Temperature sensors based on optical microfibers, fiber Bragg gratings (FBG), and optical interferometers are among the optical probes that have been used for thermometry (Yin et al. 2008). Despite their high sensitivity and wide sensing range, not all of these probes can be ruggedized for field deployment purposes. For geophysical applications, only ruggedized FBG-based sensors have been adopted in practice. Besides

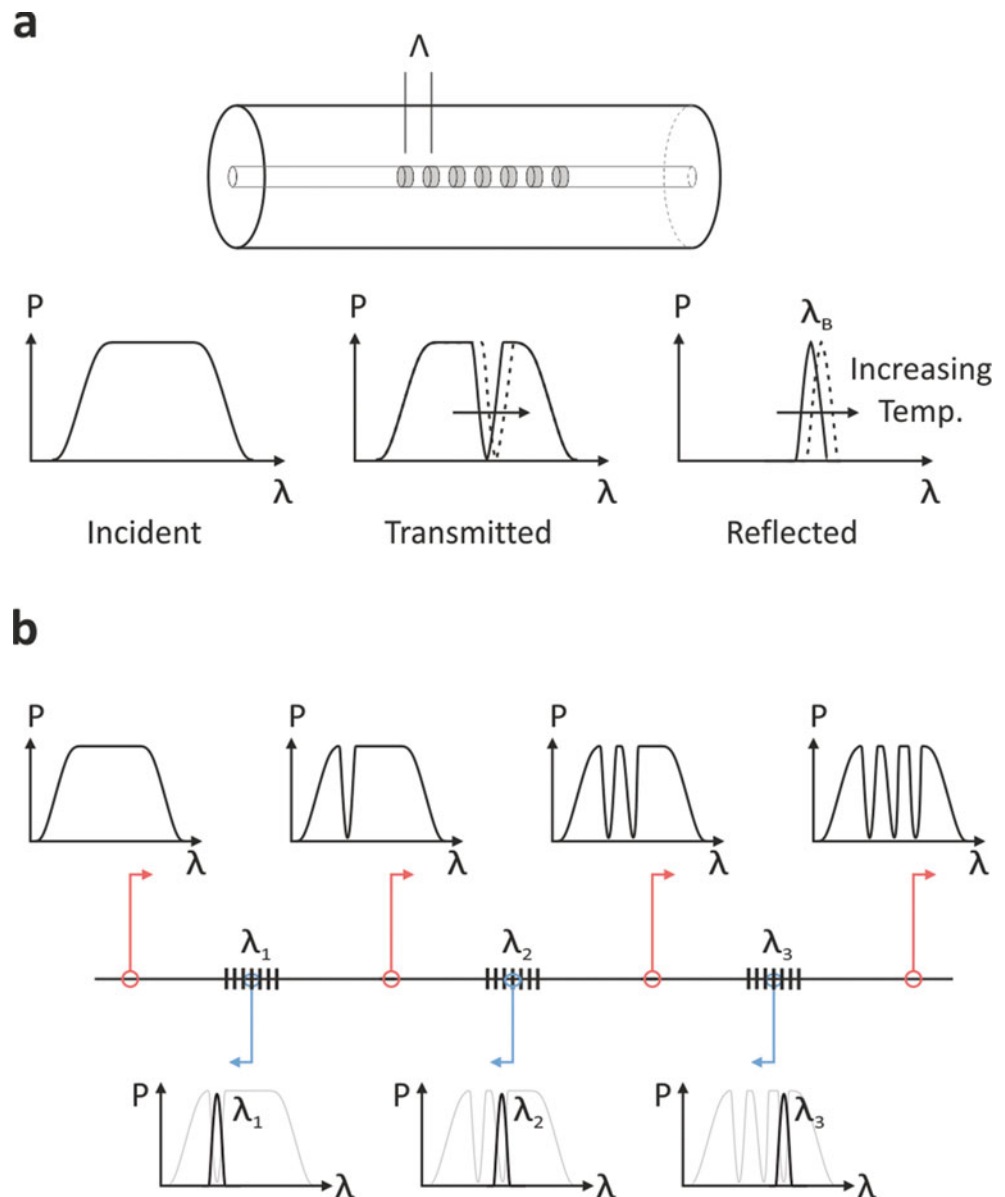
temperature sensing, FBG sensors are also used for strain and pressure measurement.

An FBG is a reflective structure in an optical fiber which is formed by a periodic variation in the refractive index (RI) of the fiber core (Fig. 1a). The pitch of this periodic structure dictates the wavelength at which the FBG reflects light. The relationship between the pitch and the reflected wavelength is given by

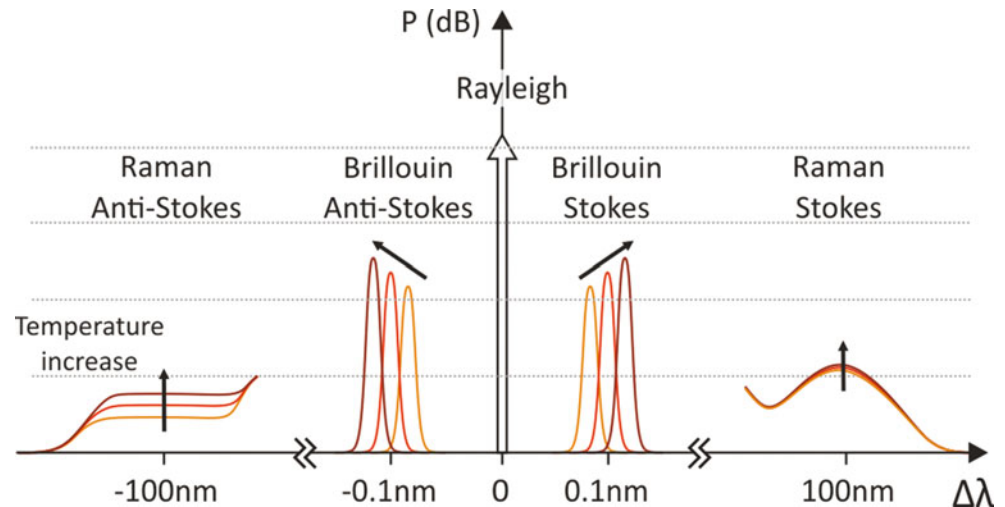
$$\lambda_B = 2n_{eff}\Lambda \quad (1)$$

where λ_B is the Bragg reflection wavelength, n_{eff} is the effective RI of the fiber, and Λ is the pitch of the periodic structure. Any increase in temperature of the fiber results in

Fiber-Optic Sensing in Geophysics, Temperature Measurements, Fig. 1 Fiber Bragg grating (FBG) temperature sensors. (a) FBG structure and its response to broadband incident light. (b) An FBG sensing array with three sensing nodes that reflect the light at three different wavelengths



Fiber-Optic Sensing in Geophysics, Temperature Measurements, Fig. 2 Light scattering mechanisms and their spectra in silica fiber for 1550 nm incident light. Different curves show the evolution of the backscattered light spectrum as a function of temperature



a thermal expansion that changes the pitch length and the RI. The temperature sensitivity of the Bragg wavelength is given by

$$\Delta\lambda_B = \lambda_B(\alpha + \beta) \Delta T \quad (2)$$

where α and β are the thermal expansion and thermo-optic coefficients of the fiber. This equation links the shift in the Bragg wavelength to a change in temperature based on two fixed coefficients. Commercial downhole FBG-based gauges with temperature accuracy around ± 0.1 K over a temperature range of 25–150 °C are available.

FBGs with different Bragg wavelengths can be fabricated and spliced together to form a sensing array also known as quasi-distributed sensor (Fig. 1b). A single broadband source can interrogate several FBGs with different Bragg wavelengths (James et al. 1996). The sensing procedure of each individual FBG in this array follows the same principle elucidated earlier. Most advanced commercial FBG interrogators can analyze up to 150 sensing points along 1 km long optical fiber.

GaAs-based optical fiber sensors have also been developed for temperature measurement in harsh environmental conditions. But their use, so far, has been limited to process monitoring, medical applications, and high radiation environments. The main advantage of GaAs-based optical fiber thermometers over their FBG counterparts is their insensitivity to external strains.

Distributed Sensing

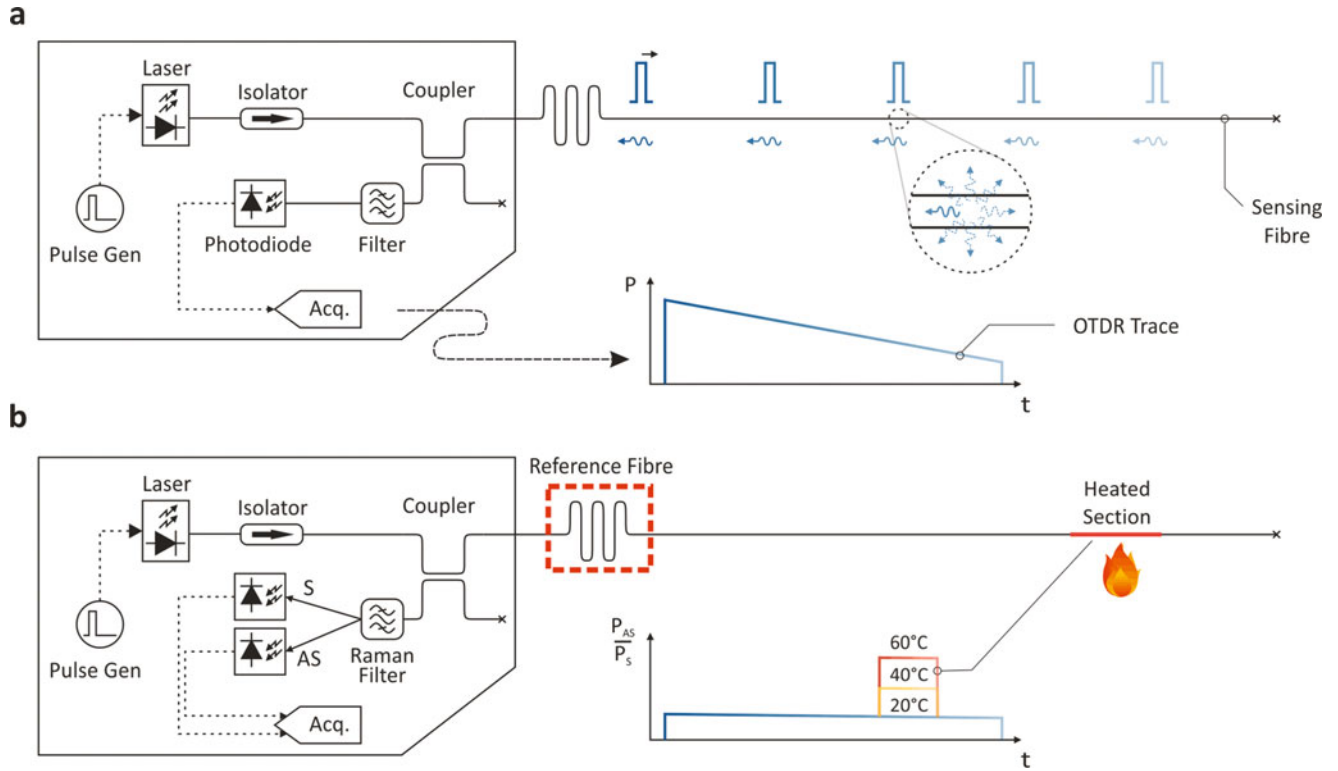
The operation of distributed optical fiber temperature sensors is markedly different from that of single point sensors. In a *distributed temperature sensor* (DTS), the optical fiber is used both as a medium to carry the light and as a sensing element. A DTS system relies on the light scattering inside the fiber to

spatially resolve the temperature along tens of kilometers of optical fiber.

There are three light scattering mechanisms in an optical fiber: Rayleigh, Brillouin, and Raman (Fig. 2). Although all three scattering processes are temperature dependent, only Brillouin and Raman scattering have been employed for DTS. The effect of an increase in temperature on the spectrum of Brillouin and Raman scattering is shown in Fig. 2. DTS systems map the temperature distribution along an optical fiber by measuring and analyzing the spectrum of the backscattered light at each point on the fiber.

Interrogation Techniques

To spatially resolve the temperature distribution along an optical fiber, the backscattered light from each individual section of that fiber needs to be captured first. The most commonly used interrogation technique employed for analysis of light scattering in an optical fiber is known as *optical time domain reflectometry* (OTDR). In this technique, a laser source is modulated to form an optical pulse (Fig. 3a). As the probe pulse propagates down the fiber, a small fraction of the light is scattered in all directions through the three mechanisms mentioned earlier. The portion of the scattered light which is recaptured and guided back toward the front-end of the fiber is called the backscattered light. The time it takes for the backscattered light from a point on the fiber to reach the front-end of the fiber depends on the distance between the two points. Therefore, the backscattered light from the sensing fiber appears as a time-domain trace known as OTDR trace where each point on the trace corresponds to the backscattered light from an interval on the sensing fiber. Since the backscattered light from the fiber has all three scattering components (i.e., Rayleigh, Brillouin, and Raman), an optical filter is added before the photodiode to transmit the desired component of the backscattered light and block the rest.



Fiber-Optic Sensing in Geophysics, Temperature Measurements, Fig. 3 (a) Schematic diagram of optical time domain reflectometry (OTDR) system and its typical OTDR trace, (b) building blocks of a

distributed temperature sensor (DTS) based on OTDR interrogation technique. The diagram shows the effect of temperature changes on the OTDR trace

Other interrogation techniques such as *optical frequency domain reflectometry* (OFDR) have also been used to devise DTS systems (Hartog 2017).

DTS Based on Raman Scattering

Raman-based OTDR systems are the most widely used and well-established technique for distributed temperature sensing. The simplicity of an OTDR setup combined with inherent temperature sensitivity of Raman scattering makes Raman-based OTDR systems the best candidate for distributed temperature measurement. The most common technique used to measure the temperature is by analyzing the intensity of Raman Stokes and anti-Stokes signals. As illustrated in Fig. 2, for a fixed temperature change, the variation in the intensity of Raman anti-Stokes is much more significant compared to that of Raman Stokes. Therefore, the temperature of the fiber can be determined from the ratio R between the intensity of the Raman anti-Stokes, I_{AS} , and the Raman Stokes, I_S , components (Hartog 2017):

$$R(T) = \frac{I_{AS}(T)}{I_S(T)} = \frac{K_{AS}}{K_S} \left(\frac{\lambda_S}{\lambda_{AS}} \right)^4 \exp \left(-\frac{S_{LE}}{T} \right) \quad (3)$$

where K is the scattering coefficient, λ the wavelength, S and AS refers to Raman Stokes and anti-Stokes, respectively, S_{LE}

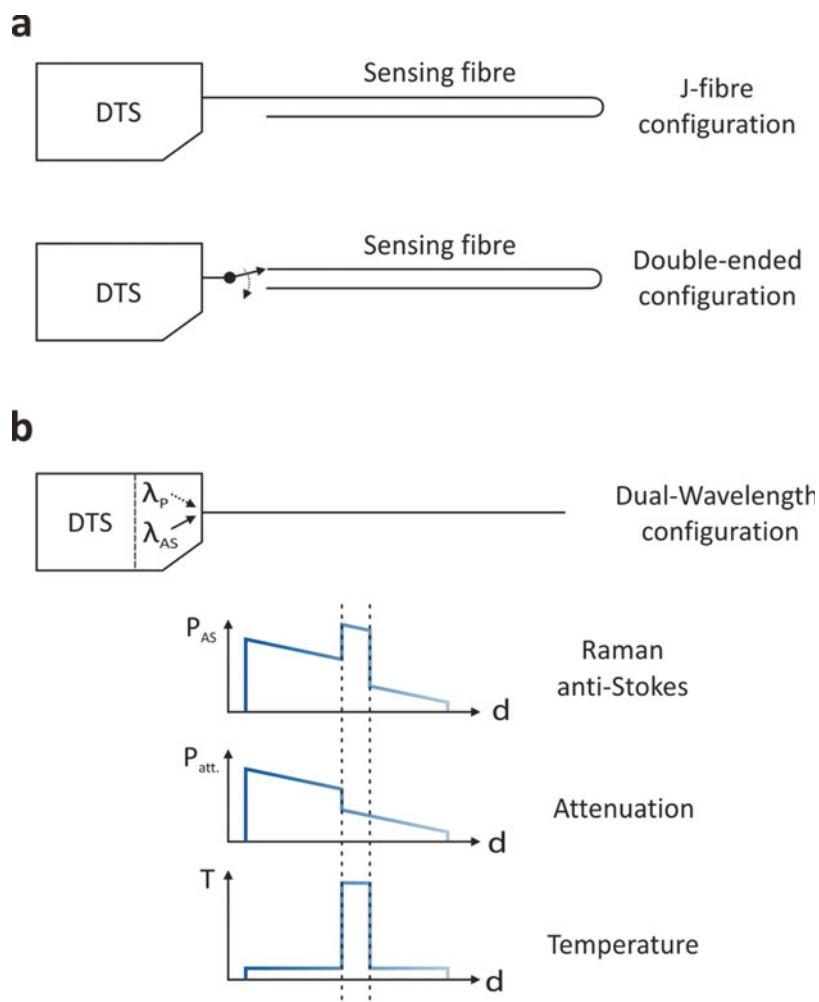
coefficient defines the temperature sensitivity of the system, and T is the temperature in Kelvin. The S_{LE} coefficient can be determined by calibration, where the Raman ratio is measured at known temperatures.

Figure 3b shows a typical block diagram of a Raman-based DTS system based on OTDR interrogation technique. In this DTS embodiment, an optical filter is used to separate the Raman Stokes and anti-Stokes backscattered light to form two OTDR traces, one for Stokes Raman and the other for anti-Stokes Raman. By dividing the two OTDR traces and compensating for wavelength-dependent, or differential, attenuation, the temperature distribution along the fiber can be determined with the help of Eq. (3). DTS systems use a stretch of fiber at a known temperature, which is referred to as the reference fiber or reference coil, in order to eliminate the influence of variable parameters within the interrogator.

As mentioned earlier, Brillouin scattering can also be used for temperature measurement. However, unlike Raman scattering, which is only sensitive to temperature, Brillouin scattering is sensitive to both temperature and strain. Therefore, Raman-based DTSs have an advantage in applications where only the temperature needs to be mapped. However, in studies where both the strain and temperature need to be measured, distributed sensing systems based on Brillouin scattering should be employed.

Fiber-Optic Sensing in Geophysics, Temperature Measurements,

Fig. 4 Distributed temperature measurement with differential attenuation compensation based on (a) special fiber configuration, and (b) dual-wavelength measurement



Implementation

As discussed in the previous section, the temperature at any point on the fiber is proportional to the ratio between the intensity of the anti-Stokes and Stokes Raman backscattered light. The intensity of the Raman backscattered light depends on both the Raman scattering coefficients as well as the fiber attenuation at the Raman anti-Stokes (λ_{AS}) and Stokes (λ_S) bands. In an ideal optical fiber and under ideal environmental conditions, the differential attenuation between the two Raman bands is fixed, which makes it easy to eliminate the effect of fiber attenuation on the Raman backscattered light and, subsequently, the temperature measurement. In reality, the differential attenuation along the fiber can be variable. Under harsh environmental conditions such as high temperature and excessive pressure, the fiber attenuation can be altered through a process known as hydrogen ingress or hydrogen darkening (see section below). A distributed temperature measurement along an optical fiber in which the effect of differential attenuation has not been compensated for can result in a significant disparity between the actual and measured temperature.

Two techniques have been devised to eliminate the effect of differential attenuation on temperature measurements, one based on special fiber configuration and the other based on multi-wavelength measurement (Hartog 2017).

One approach to eliminate the effect of optical fiber attenuation is to arrange the sensing fiber in a loop configuration as shown in Fig. 4a. In both configurations, the optical fiber is arranged in such a way to provide two separate Raman scattering read-outs at any given section. It is shown that the two read-outs at a given section can be combined to compensate for any signal attenuation due to fiber loss and to provide an accurate temperature measurement.

The main disadvantage of this technique is the feasibility of such fiber arrangements. In many cases, the access to both ends of the fiber is not feasible. The other issue with this compensation approach is its vulnerability to fiber breakage. The double-ended method ceases to work if the fiber is damaged at any point along the loop.

The other compensation method relies on measuring the fiber attenuation at the anti-Stokes Raman band to adjust the intensity of the Raman backscattered trace. This multi-

wavelength measurement procedure is illustrated in Fig. 4b. At the first step, a probe pulse (λ_P) is launched into the fiber to obtain the OTDR trace at the anti-Stokes Raman band. To assess the fiber attenuation, a second probe pulse at the anti-Stokes Raman band (λ_{AS}) is launched into the fiber and the Rayleigh backscattered light is collected. By analyzing the intensity of the Raman backscattered light and comparing it with the fiber loss measurement, an accurate temperature measurement can be achieved.

Temperature Accuracy and Resolution

The measurement accuracy, that is, the difference between measured and true absolute temperature, or systematic error, of a Raman DTS interrogator depends on the calibration and stability of the system, as well as how well differential attenuation has been compensated for. For example, Henninges et al. (2005) have reported an accuracy of ± 0.3 °C for a calibrated Raman DTS system using double-ended processing, both under field and laboratory conditions, while Voigt et al. (2011) observed a temporal drift of ± 2 °C over a period of 5 days attributed to unstable optical connectors using single-ended processing.

The temperature resolution, that is, the repeatability of the measurement, or random error, depends on the following parameters:

- Integration time, that is, length of time interval over which a measurement is averaged
- Length and attenuation of the sensing fiber
- Noise figure of the components used in the interrogator

The temperature resolution is proportional to the inverse of the square root of integration time, and example numbers achievable with a current Raman DTS instrument are listed in Table 1. While the fiber length and signal attenuation are features characteristic to an individual DTS installation, the integration time can be adjusted to the conditions and requirements of an individual measurement campaign.

Hydrogen Darkening

Under influence of hydrogen, rapid degradation of the properties of optical fibers has been observed (Smithpeter et al. 1999). Additional optical loss due to absorption is caused by diffusion of hydrogen into interstitial sites in the silica lattice,

and chemical reaction with the glass constituents resulting in hydroxyl (OH) formation, which is amplified by dopants such as germanium or phosphorus (Williams et al. 2000). Appropriate shielding is especially difficult at elevated temperatures, as the diffusion of hydrogen through steel increases at temperatures >150 °C. The interrogation approaches described above for Raman-based DTS are only effective at low levels of hydrogen ingress. Other measures to avoid hydrogen darkening include modification of the fiber dopants, use of hermetic coatings with a carbon layer, design of sensor cables including hydrogen barriers, and operation of the interrogator at wavelengths that are less affected by hydroxyl absorption.

Sensor Cables, Deployment Methods

In order for the fiber to maintain its optical transmission properties during deployment, that is, during installation and operation, it needs to be protected against external influences like mechanical stress or ingress of hydrogen. The protective jacketing or armor must be able to tolerate the harsh environmental conditions which can prevail in boreholes, like high temperature and pressure, corrosive fluids, as well as the occurring mechanical loads.

For temperature sensing applications, the fibers are usually inserted into a thin stainless steel tube with about 2 mm diameter, which is typically filled with a water-blocking gel. This is a standard technology for optical fiber cables which are also used in other application fields, for example, in the telecom sector. Depending on the individual deployment conditions, an appropriate sensor cable design has to be employed. Sometimes, a hybrid design is used, that is, with optical fibers and electrical conductors, for simultaneous operation of different fiber-optic and electronic sensors in parallel. For boreholes, basically two different deployment methods described in the following are used in practice.

Permanent Installation

Here ruggedized cables, often with a second steel tube, and an outer plastic jacketing are used. Such cables can be strapped to production tubing using special protector clamps at the tubing joints, which is standard industry practice in oil and gas wells. Cables can also be installed behind casing and cemented (e.g., Henninges et al. 2005). This has advantages for the measurement due to the proximity of the sensor cable to the formation, but it is also more challenging to avoid damage during installation or perforation, or potential leakage pathways in the cement.

Another method is the use of fibers which are pumped into hydraulic control lines (e.g., Brown et al. 2000), which are typically 0.64 cm in diameter, and can be up to several

Fiber-Optic Sensing in Geophysics, Temperature Measurements, Table 1 Examples for temperature resolution of a Raman DTS measurement, as a function of fiber length and integration time. (After Voigt et al. 2011)

Fiber length (m)	RMS error (°C), 60 s integration time	RMS error (°C), 10 min integration time
1000	0.05	0.02
4000	0.28	0.04

kilometers in length. As a great advantage, the fiber can be replaced in case of fiber degradation.

Temporary/Retrievable Installation

A fiber-optic sensor cable can also be lowered down inside a well, similar to a conventional wireline operation. This method can be applied in already existing wells, or in cases where the well design makes a permanent installation to the desired target depth difficult. When using distributed sensing methods, the entire profile along the installation interval can be monitored simultaneously once the sensor cable is in place. This is a great advantage over conventional wireline techniques, which require that the logging tool is moved along the section of interest with a defined speed.

For this purpose, optical fibers can be deployed using a slickline cable with a design similar to the one for permanent installations described above, as well as using coiled tubing, or inside a fiberglass rod. Here nevertheless only data acquired on the optical fibers, eventually together with data from the logging winch like cable tension and cable speed during running into the hole, can be recorded.

Henninges et al. (2011) developed a hybrid borehole measurement system, with an armored wireline cable containing both electrical conductors and optical fibers. This allows for parallel deployment of fiber-optic sensors and electric down-hole tools. In addition, data like gamma-ray or casing-collar locator logs can be simultaneously recorded, allowing for proper depth correlation. With the advent of distributed acoustic sensing (DAS), there is increasing interest for development of such hybrid wireline cables (e.g., Hartog et al. 2014).

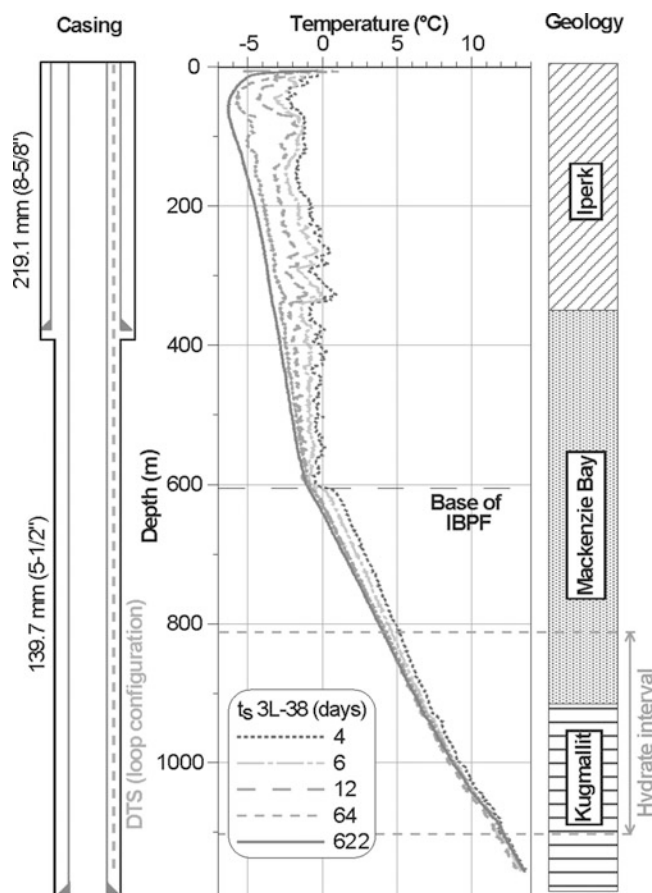
Applications, Methods of Evaluation

Passive Monitoring

Formation Temperatures, Heat Flow Studies

One of the classical applications of borehole temperature measurements is the determination of formation temperatures, for example, for heat flow studies. The prerequisite for this is that the well is in thermal equilibrium with the surrounding formation. It is important to consider that this may not always be the case, and that the well temperature can be influenced from the movement of fluids within the borehole, for example, due to fluid circulation during drilling.

Henninges et al. (2005) performed repeated DTS measurements in three wells to a depth of 1200 m over several years and monitored the decay of the thermal disturbance after drilling in order to determine the undisturbed formation temperatures (Fig. 5). From the geothermal gradients, profiles of the in situ thermal conductivities were derived. The use of permanently installed sensor cables enabled to perform



Fiber-Optic Sensing in Geophysics, Temperature Measurements, Fig. 5 DTS temperature profiles for successive times after completion (t_s) of the Mallik 3L-38 well. Left: well sketch showing completion and position of DTS sensor cable. Right: geological sequence boundaries. IBPF: base of ice-bearing permafrost. (After Henninges et al. (2005))

measurements even after abandonment, when the wells were not accessible anymore for wireline equipment.

Flow Profiling

The temperature profile of wells can also be used to derive estimates of the flow profile during production or injection. Conventionally, production logging is performed using downhole tools equipped with impeller flowmeters and differential manometers which is operationally expensive. Evaluation methods using temperature are based on models for the heat transfer inside and around the wellbore, for which analytical approximations have been developed by Ramey (1962). The temperature of the produced fluids depends on the geothermal temperature and the thermal properties as well as the flowing velocities of the borehole fluids. In addition, thermal effects caused by the pressure drop between the formation and the well (Joule-Thomson effect) have to be accounted for.

The method has been greatly facilitated by availability of DTS because of the simplicity of the measurement process.

Brown et al. (2000) derived flow profiles from DTS temperature data in extended-reach deviated oil wells, with completions where conventional production logging would not have been possible. Evaluation was performed in conjunction with other production data using a wellbore simulator, as solutions are not always unique using temperature alone.

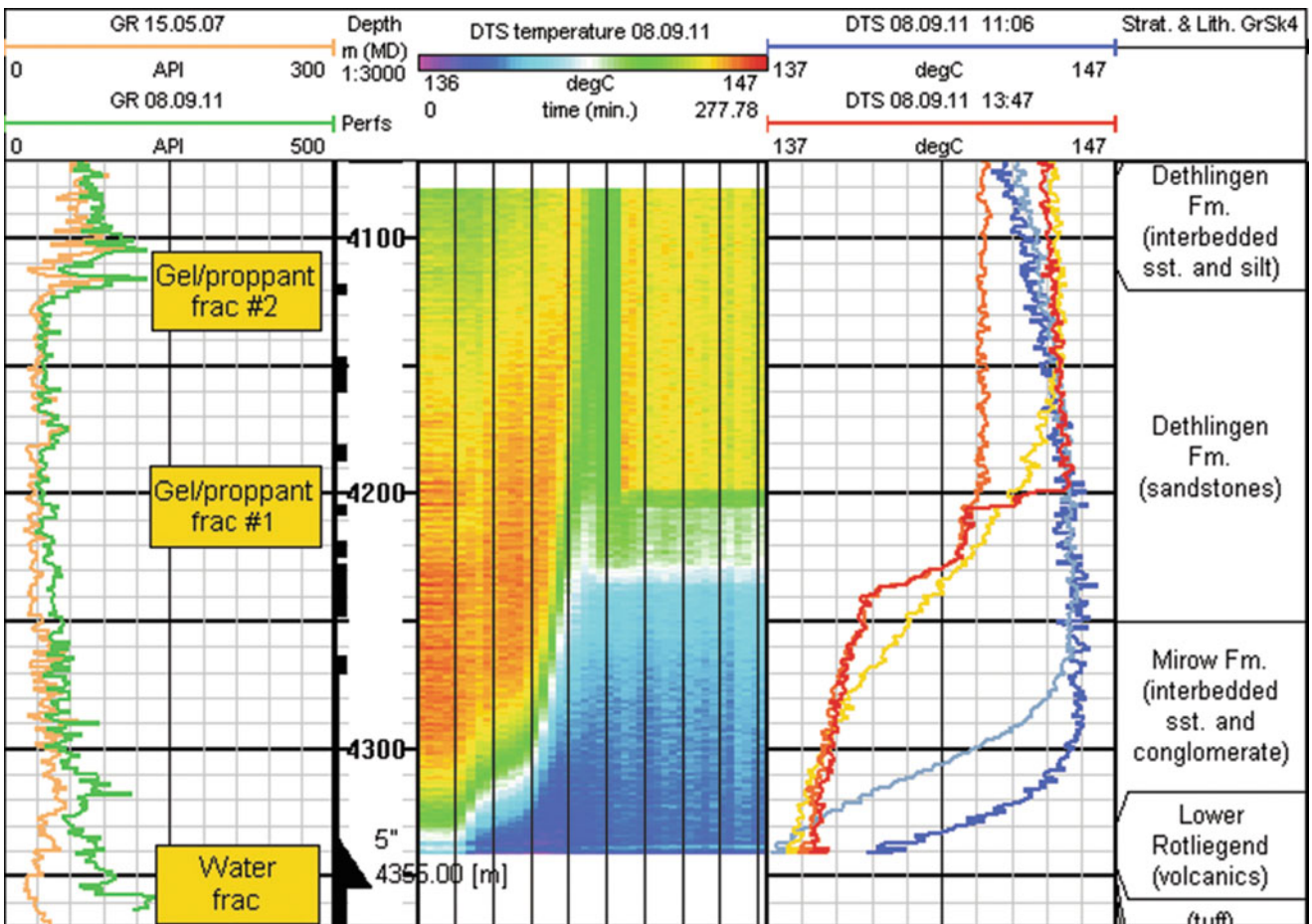
Flow profiling using temperature is also of interest for geothermal wells. Henninges et al. (2012) performed DTS measurements during a production test in a 4300 m deep geothermal well using a hybrid wireline cable (Fig. 6). Using this data, they were able to derive information on temporal variations of flow from individual reservoir zones characterizing the hydraulic behavior of this enhanced geothermal system (Blöcher et al. 2016).

A different method is based on observing the temperature recovery after shut-in of injection into a well (warm-back). Using this technique, permeable regions that have accepted larger fluid volumes can be identified through differences in the cooling effect and the duration of the recovery process (Witterholt and Tixier 1972).

Well Cementing and Well Treatments

The DTS method has also been successfully applied for on-line monitoring of well cementing and well treatments like hydraulic fracturing or thermal stimulation. Hartog (2017) contains an extensive review of different applications and related publications, primarily from the oil and gas sector. More recently, the DTS method has also increasingly been used in combination with DAS, which enables to derive additional information on flow inside the well from noise generated at injection or production zones, or changes of the propagation of acoustic waves along the well.

Another important application for DTS is the monitoring of steam injection during heavy oil production. Here process temperatures of 220–270 °C are being used, and DTS is applied to understand where the formation has been heated and in order to maintain the desired temperatures, as well as to avoid steam breakthrough at the producer well. The high temperatures create the risk of fiber degradation by hydrogen darkening, which requires appropriate mitigation measures to be applied (see section “Distributed Sensing”).



Fiber-Optic Sensing in Geophysics, Temperature Measurements, Fig. 6 DTS temperature data and gamma ray logs (GR) during a production test in the GrSk4 well. Acquisition times of DTS profiles

not displayed in log header: light blue 12:29, gold: 13:02, orange: 13:28. (After Henninges et al. (2012))

In practice, special interrogation techniques to compensate for additional loss have been applied, or the fibers are installed in hydraulic control lines, which enable to replace them after degradation.

High-Temperature Applications

In order to avoid the adverse effects of hydrogen ingress in high temperature and high pressure environments, either fibers with metal coating or fibers with hermetic coatings including a carbon layer have been proposed. Until now, metal-coated fibers exhibit high attenuation, especially at low temperatures (Reinsch and Henningses 2010), which limits their use to short sensing ranges. Hermetic coating fibers have been tested in geothermal wells in the United States, where they maintained their optical properties over 7 months at 171–182 °C, but rapidly failed in another well when exposed to temperatures up to 243 °C (Smithpeter et al. 1999). During a field test in a high-temperature geothermal well in Iceland, Reinsch et al. (2013) have successfully deployed fibers with hermetic coating at temperatures up to 230 °C over a period of 14 days, but reported different signs of degradation occurring after exposure to temperatures above 300 °C.

Another approach is the use of pure-silica core (PSC) fibers without dopants in the fiber core, which makes them less susceptible to hydrogen darkening. Kaura and Sierra (2008) successfully tested single mode PSC fibers for DTS monitoring during a 20-month trial at temperatures up to 250 °C in a steam injection well. Nevertheless, single-mode DTS systems have lower spatial and temperature resolution due to lower signal intensity compared to multimode systems. Step-index PSC fibers are not suitable for multimode DTS systems due to pulse dispersion, which degrades spatial resolution.

A third approach has been used by Ikeda (2003), who performed DTS measurements in a geothermal well in Japan at temperatures of 275 °C for 43 h of continuous operation, while flushing the tube through which fiber was deployed with nitrogen, in order to avoid degradation due to hydrogen ingress.

Active Methods

Fluid Logging, Thermal Tracer/Dilution Techniques

Flow profiling using temperature requires a measurable difference between the temperatures of the borehole and the surrounding formation. In situations with small temperature differences (e.g., low geothermal gradient along the borehole, closely spaced flow zones, horizontal wells, low flow rates and fluid velocities) the temperature difference between the fluids can be enhanced by induction of a temperature change within the well. Information about the flow along the well is derived from mixing models and by tracking the movement of

temperature features, which is similar to earlier established methods using solute tracers or fluid electrical conductivity logging.

The desired change in temperature inside the well can be induced by local fluid injection through tubing (Hurtig et al. 1994; Leaf et al. 2012) or a downhole electrical heat source (Read et al. 2015). Such methods have mainly been applied in shallow groundwater wells at depths up to 100 m until now, primarily within the context of detecting flow in fractured rock masses.

A special challenge is that fast measurements with high precision are needed. The required resolution might in some cases be difficult to achieve with the typical performance of DTS instruments.

Heat-Pulse Method

This method is used to determine in situ thermal conductivities of the formation and to derive information about heat advection and fluid flow along boreholes. It is based on the well-known theory of heat conduction along a constantly heated line source (Carslaw and Jaeger 1959), and has originally been developed in hydraulic engineering for leak detection along surface sealings and dams. The borehole is heated along its axis using electrical conductors as a controlled heat source and temperature changes are monitored with DTS.

Freifeld et al. (2008) performed an experiment in a 400 m deep borehole, which was heated for 64 h using a heating power of about 20 W/m. Thermal conductivities were determined with 1 m resolution by numerical inversion of the temperature data recorded during the 58-h cooling phase after heating. Prevedel et al. (2014) performed similar measurements along a well to a depth of 690 m. They further extended the method to investigate the properties of both the completion of the well and the surrounding rock formation by evaluating early- and late-time data of the heating and cooling phases.

The method is also used in shallow geothermal wells, where it is referred to as enhanced geothermal response test (EGRT). Lehr and Sass (2014) determined profiles of thermal conductivity along a 400 m borehole heat exchanger, and identified zones with heat convection caused by fluid flow inside the formation around the well.

Another application of this method has been for distributed measurements of soil water content at meter to kilometer scales (Sayde et al. 2010).

Other Applications in Related Disciplines

The DTS method has also increasingly been used for measurements at surface, especially in surface water bodies. Selker et al. (2006) reviewed the applications of temperature mapping in hydrological sciences, where, for example, dynamic processes in lakes and rivers can be investigated using DTS measurements. As one example, they

measured the temperature distribution along a 4 km telecom cable at the bottom of Lake Geneva between France and Switzerland.

Another application is the detection of heat discharges associated with fluid emissions at the sea floor for volcano monitoring. Carlino et al. (2016) installed a 2400 m long cable at the Campi Flegrei caldera in Southern Italy, extending from the shore into the sea to a maximum water depth of about 80 m. As for the previous example, the measurements were performed using the BOTDA method (stimulated Brillouin scattering) under the assumption that the fiber within the sensor cable was not subjected to strain and the measured effects could be related to changes in temperature only.

Surface measurements using optical fibers of telecommunication and data transmission cables have great potential for further applications in science and engineering since large networks of such cables already exist worldwide both in on- and offshore regions (You 2010).

Summary

Fiber-optic sensing has opened up new possibilities for geophysical measurements with a broad range of applications.

The optical fiber transmits the signal and can also act as the sensing element. There are no electronic parts required down-hole. Some of the resulting advantages of fiber-optic sensors compared to conventional electronic sensors are:

- Ruggedness
- Flexibility
- High temperature tolerance
- Ease of deployment
- Immunity to electromagnetic interference

The core of an optical fiber guides the optical waves and is made of silica glass. The coating layer protects the fiber from harmful influences and is made of plastic materials or metals, which primarily determine the temperature tolerance of the optical fiber. Depending on the sensing method, different types of fibers are used.

Fiber Bragg grating (FBG) sensors, which are also used for measurement of strain and pressure, have mainly been adopted as point sensors. Several 10s of these sensors can be interrogated on a single fiber through wavelength multiplexing interrogation technique.

Distributed temperature sensing (DTS) methods enable sampling with high spatial and temporal resolution over long distances of up to tens of kilometers in length. They rely on different light scattering mechanisms within the optical fiber. Most systems are based on the optical time domain reflectometry (OTDR) technique. Here the backscattered light

is captured as a function of time, and the position along the fiber is determined from the travel time of the signal.

DTS systems based on Raman scattering are used most often. Temperature is determined from the ratio of the back-scattered light at two distinct wavelengths. Other wavelength-dependent effects (differential attenuation) need to be avoided or adequately compensated. One of such effects is hydrogen darkening, which is caused by diffusion of hydrogen into the fiber at elevated temperatures. The accuracy of the measured temperature depends on the calibration and stability of the DTS system. The achievable resolution depends on the signal attenuation, and scales with the inverse of the square root of integration time.

Sensor cables used for deployment under harsh conditions of a borehole environment have to be designed to protect the optical fibers against mechanical stress and hydrogen ingress. The optical fiber is usually inserted into a thin stainless steel tube and further elements for reinforcement and protection are added, according to the deployment conditions. Hybrid cable designs with additional electrical conductors also exist.

For permanent installations, ruggedized cables are used which are deployed along tubing, or behind casing. Another method is to pump the fiber into a capillary tubing. This enables to replace the fiber in case it is damaged or degraded. Temporary or retrievable installations are either performed similar to a slickline operation, using a coiled tubing, or hybrid wireline logging cables.

There are many applications for DTS in borehole temperature measurements, especially if transient processes involving temporal changes of temperature along extended sections of a borehole need to be studied. These can be classified in passive methods, where the (naturally) occurring temperature changes are recorded, and active methods, where an additional source of heat is introduced inside the borehole:

1. Passive monitoring

- Determination of formation temperatures, observation of the decay of drilling disturbance, for example, for heat flow studies
- Flow profiling during injection or production, warm-back monitoring, leakage detection
- Well cementing and well treatments (hydraulic fracturing, thermal stimulation)
- High temperature applications (geothermal wells, steam injection)

2. Active methods

- Fluid logging, thermal tracer/dilution techniques: determination of the flow profile along well from mixing models, or by tracking the movement of temperature features
- Heat-pulse method: determination of the in situ thermal conductivity, or zones of fluid flow and/or heat advection

Other applications in related disciplines include surface measurements, for example, in hydrology or volcano monitoring, using optical fibers of telecom cables, of which large networks are already in place worldwide.

Cross-References

- [Geophysical Well Logging](#)
- [Heat and Ground Water Flow](#)
- [Heat Flow Determinations, Continental](#)

Bibliography

- Blöcher G, Reinsch T, Henningses J, Milsch H, Regensburg S, Kummerow J, Francke H, Kranz S, Saadat A, Zimmermann G, Huenges E (2016) Hydraulic history and current state of the deep geothermal reservoir Groß Schönebeck. *Geothermics* 63:27–43. <https://doi.org/10.1016/j.geothermics.2015.07.008>
- Brown GA, Kennedy B, Meling T (2000) Using fibre-optic distributed temperature measurements to provide real-time reservoir surveillance data on Wytch Farm field horizontal extended-reach wells, SPE 62952. In: SPE annual technical conference and exhibition, Dallas, 1–4 Oct 2000
- Carlino S, Mirabile M, Troise C, Sacchi M, Zeni L, Minardo A, Caccavale M, Daranyi V, De Natale G (2016) Distributed-temperature-sensing using optical methods: a first application in the offshore area of Campi Flegrei Caldera (Southern Italy) for Volcano monitoring. *Remote Sens* 8(8). <https://doi.org/10.3390/rs8080674>
- Carslaw HS, Jaeger JC (1959) *Conduction of heat in solids*, 2nd edn. Clarendon Press, Oxford, UK
- Freifeld BM, Finsterle S, Onstott TC, Toole P, Pratt LM (2008) Ground surface temperature reconstructions: using in situ estimates for thermal conductivity acquired with a fiber-optic distributed thermal perturbation sensor. *Geophys Res Lett* 35:L14309
- Hartog A (2017) An introduction to distributed optical fibre sensors, Series in Fiber optic sensors. CRC Press, Boca Raton
- Hartog A, Frignet B, Mackie D, Clark M (2014) Vertical seismic optical profiling on wireline logging cable. *Geophys Prospect* 62(4):693–701. <https://doi.org/10.1111/1365-2478.12141>
- Henningses J, Huenges E, Burkhardt H (2005) In situ thermal conductivity of gas-hydrate-bearing sediments of the Mallik 5L-38 well. *J Geophys Res* 110:B11206. <https://doi.org/10.1029/2005JB003734>
- Henningses J, Baumann G, Brandt W, Cunow C, Poser M, Schrötter J, Huenges E (2011) A novel hybrid wireline logging system for downhole monitoring of fluid injection and production in deep reservoirs. In: Conference proceedings, 73rd EAGE conference & exhibition, Vienna (CD-ROM), C043
- Henningses J, Brandt W, Erbas K, Moeck I, Saadat A, Reinsch T, Zimmermann G (2012) Downhole monitoring during hydraulic experiments at the In-situ Geothermal Lab Groß Schönebeck. In: Proceedings, 37th workshop on geothermal reservoir engineering, Stanford (CD-ROM), pp 51–56
- Hurtig E, Schrötter J, Grosswig S, Kühn K, Harjes B, Wieferig W, Orrell RP (1993) Borehole temperature measurements using distributed fibre optic sensing. *Sci Drill* 3(6):283–286
- Hurtig E, Grosswig S, Jobmann M, Kühn K, Marschall P (1994) Fibre-optic temperature measurements in shallow boreholes: experimental application for fluid logging. *Geothermics* 23(4):355–364
- Ikeda N (2003) Fracture reservoir characterization by fiber-optic distributed temperature log. In: Twenty-eighth workshop on geothermal reservoir engineering, Stanford University, Stanford, 27–29 Jan 2003
- James SW, Dockney ML, Tatam RP (1996) Simultaneous independent temperature and strain measurement using in-fibre Bragg grating sensors. *Electron Lett* 32(12):1133–1134. <https://doi.org/10.1049/el:19960732>
- Kaura J, Sierra J (2008) High-temperature fibers provide continuous DTS data in a harsh SAGD environment. *World Oil* 229:47–54
- Leaf AT, Hart DJ, Bahr JM (2012) Active thermal tracer tests for improved hydrostratigraphic characterization. *Ground Water* 50(5):726–735. <https://doi.org/10.1111/j.1745-6584.2012.00913.x>
- Lehr C, Sass I (2014) Thermo-optical parameter acquisition and characterization of geologic properties: a 400-m deep BHE in a karstic alpine marble aquifer. *Environ Earth Sci* 72(5):1403–1419. <https://doi.org/10.1007/s12665-014-3310-x>
- Prevedel B, Martens S, Norden B, Henningses J, Freifeld BM (2014) Drilling and abandonment preparation of CO₂ storage wells – experience from the Ketzin pilot site. *Energy Procedia* 63:6067–6078. <https://doi.org/10.1016/j.egypro.2014.11.639>
- Ramey HJ (1962) Wellbore heat transmission. *J Pet Technol* 14(4):427–435
- Read T, Bense VF, Hochreutener R, Bour O, Le Borgne T, Lavenant N, Selker JS (2015) Thermal-plume fibre optic tracking (T-POT) test for flow velocity measurement in groundwater boreholes. *Geosci Instrum Method Data Syst* 4(2):197–202. <https://doi.org/10.5194/gi-4-197-2015>
- Reinsch T, Henningses J (2010) Temperature-dependent characterization of optical fibres for distributed temperature sensing in hot geothermal wells. *Meas Sci Technol* 21(9):094022. <https://doi.org/10.1088/0957-0233/21/9/094022>
- Reinsch T, Henningses J, Ásmundsson R (2013) Thermal, mechanical and chemical influences on the performance of optical fibres for distributed temperature sensing in a hot geothermal well. *Environ Earth Sci* 70:3465–3480. <https://doi.org/10.1007/s12665-013-2248-8>
- Sayde C, Gregory C, Gil-Rodriguez M, Tuffillaro N, Tyler S, Nvd G, English M, Cuenca R, Selker JS (2010) Feasibility of soil moisture monitoring with heated fiber optics. *Water Resour Res* 46(6). <https://doi.org/10.1029/2009WR007846>
- Selker JS, Thévenaz L, Huwald H, Mallet A, Luxemburg W, Van De Giesen N, Stejskal M, Zeman J, Westhoff M, Parlange MB (2006) Distributed fiber-optic temperature sensing for hydrologic systems. *Water Resour Res* 42(12). <https://doi.org/10.1029/2006WR005326>
- Smithpeter C, Normann R, Krumhansl J, Benoit D, Thompson S (1999) Evaluation of a distributed fiber-optic temperature sensor for logging wellbore temperature at the Beowawe and Dixie Valley geothermal fields. In: Twenty-fourth workshop on geothermal reservoir engineering, Stanford University, Stanford, 25–27 Jan 1999
- Voigt D, van Geel J, Kerkhof O (2011) Spatio-temporal noise and drift in fiber optic distributed temperature sensing. *Meas Sci Technol* 22(8). <https://doi.org/10.1088/0957-0233/22/8/085203>
- Williams GR, Brown G, Hawthorne W, Hartog AH, Waite PC (2000) Distributed temperature sensing (DTS) to characterize the performance of producing oil wells. In: Wang A, Udd E (eds) Proceedings of SPIE, Boston, vol 4202, pp 39–54. <https://doi.org/10.1117/12.411726>
- Wisian KW, Blackwell DD, Bellani S, Henfling JA, Normann RA, Lysne PC, Förster A, Schrötter J (1998) Field comparison of conventional and new technology temperature logging systems. *Geothermics* 27(2):131–141
- Witterholt EJ, Tixier MR (1972) Temperature logging in injection wells. Paper presented at the fall meeting of the Society of Petroleum Engineers of AIME, San Antonio, 1 Jan 1972
- Yin S, Ruffin PB, Yu FTS (2008) *Fiber optic sensors, Optical science and engineering*, vol 132, 2nd edn. CRC Press, Boca Raton
- You Y (2010) Harnessing telecoms cables for science. *Nature* 466(7307):690–691. <https://doi.org/10.1038/466690a>

Floating Seismographs (MERMAIDS)

Yann Hello and Guust Nolet

IRD Geoazur and Université Côte d'Azur, Sophia Antipolis, France

Definition

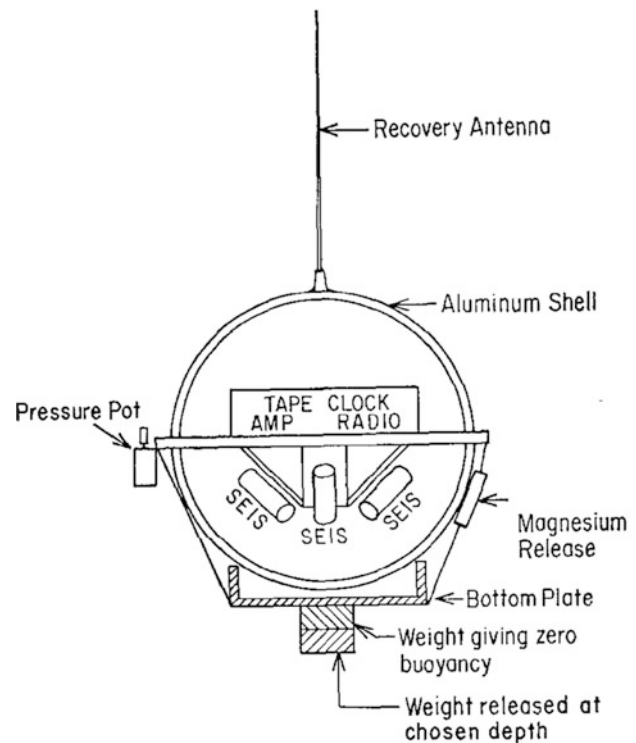
MERMAIDS are Lagrangian floats, drifting passively with the deep ocean current, detecting seismic waves, and recording seismograms (and possibly other data) which they transmit by satellite after coming to the surface. The name is an acronym for Mobile Earthquake Recording in Marine Areas by Independent Divers (Simons et al. 2009).

History

The need to observe the abyssal current in the oceans led John Swallow in the UK and Henry Stommel in the USA to develop submarine floats in the mid-1950s (see Riser et al. 2018, for a review). Though the early floats were equipped with a hydrophone, this was only used to track the float position by recording intermittently fired explosives from a surface vessel.

The first seismological application was by Bradner et al. (1970), who equipped a spherical aluminum float (Fig. 1) with a 3-component seismometer to record microseismic noise between 0.02 and 5 Hz at a depth of 1200 m. D'Spain and Hodgkiss (1991) used 17-inch diameter Benthos glass shell as a float to measure the background noise level in the 1–20 Hz frequency band. Nolet (1991) analyzed data from this instrument and discovered a weak signal of an earthquake of magnitude 6.8 at a depth of 26 km in the Andreanoff Islands (distance 47°) showing that floating seismographs could in principle be used to record teleseisms.

At the time, the technology was not yet advanced enough to have floats operate independently in mid-ocean. This changed with the development of autonomous sensors for oceanography such as RAFOS and SOLO floats (Riser et al. 2018 – SOLO stands for Sounding Oceanographic Lagrangian Observer, the acronym RAFOS is 'SOFAR', for SOUNd Fixing And Ranging, spelled backwards.), which can transmit a small volume of data by satellite (Argos) and be located by GPS when surfacing, thus obviating the need for tracking by a surface vessel. A modern float has a bladder that allows it to vary its volume and seek the depth where it is neutrally buoyant, since the density of seawater increases slowly with depth. A nominal cruising depth of 1500 m is a good compromise to get a good signal to noise ratio during monitoring and time needed to rise to surface in order to reduce the power consumption. Simons et al. (2009) successfully tested a



Floating Seismographs (MERMAIDS), Fig. 1 The spherical float used by Bradner et al. (1970) to record microseismic noise

hydrophone attached to a SOLO float to record a number of earthquake signals.

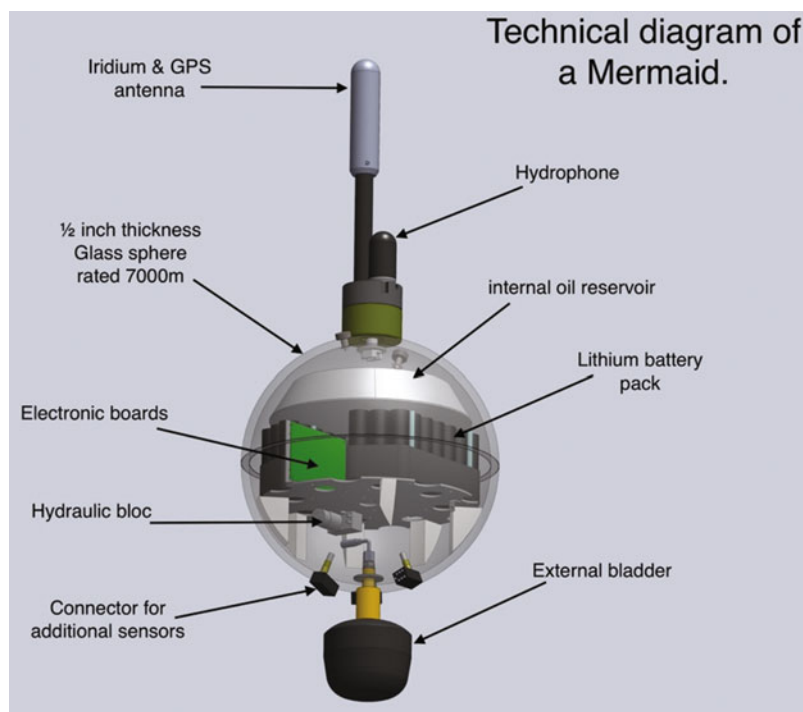
Adapting a RAFOS float, Hello et al. (2011) subsequently developed the first MERMAID capable of autonomously detecting earthquakes and transmitting these by satellite. The current version of the MERMAID uses a 17-inch glass sphere and was developed by OSEAN-SAS and the University of the Côte d'Azur, both located in the south France.

Technical Description

In the MERMAID, all hydraulic and electronic components are encapsulated in the glass sphere except for a bladder which adapts the float volume and controls its buoyancy in the water column (Fig. 2). Oil is pumped from an inner reservoir to the bladder outside. The piston can pump oil against a pressure up to 550 bar, equivalent to a depth of 5500 m, which forces the float to ascent to the surface from a 4000 m maximum cruising depth. To descend, the oil transfer from the bladder to the inside is controlled by valve or bypass. The float measures the pressure during its descent. It calculates its speed and activates the pump or opens the valve to maintain a speed that varies according to the distance to the desired depth. For the ascent, the float will leave its cruising depth by gradually increasing its vertical speed to a maximum to reach the surface.

Floating Seismographs

(MERMAIDS), Fig. 2 The spherical glass float used by OSEAN-SAS and the University of the Côte d'Azur in France to detect, record and transmit earthquakes by satellite communication. (Credit: OSEAN-SAS)



The MERMAID contains two circuit boards. A very low power (3,3 V/7,6 mA), acquisition board interacts with a pilot board that operates the pump. Both are powered up by lithium disposable batteries, like the rest of the system (motor, pump, Iridium modem, GPS receiver). For a battery with capacity 5 KWh, the MERMAID has an expected life time of 6 years of continuous monitoring the hydrophone signal at a depth of 1500 m, if a selection of 150 Kbytes of triggered seismogram data is transmitted every week.

Signal Discrimination

After the float descends to its drift depth, it oscillates around the level of natural buoyancy, typically 1 cycle/hour, but the pressure variations introduced by this are too little and too slow to affect seismic signals. To record acoustic signals, the MERMAID is equipped with a hydrophone that monitors the pressure in the acoustic frequency band (0.1–10 Hz). This is a useful passband for seismic P waves. The MERMAID float depth is too close to the surface to record low frequency waves (surface waves, normal modes) for which the pressure perturbation increases only slowly with depth from zero at the surface. The pressure signal is continuously analyzed by comparing energy averages over a short time window with that over a long-time window. When the STA/LTA or signal-to-noise (SNR) ratio triggers an alert, a wavelet decomposition is computed. Seismic P waves are efficiently converted to acoustic waves at the sea bottom and have a distinctive

signature characterized by the power in different wavelet scales, but converted S waves have only been observed at intermediate distance ($<20^\circ$). A measure of wavelet power, combined with the SNR, may trigger a return to the surface for transmission of the seismogram per satellite (Iridium). For specific details of the algorithm, see Sukhovich et al. (2011). If the detected signal is characterized as a P wave but rated too weak, it is recorded in a buffer to be transmitted later by Iridium. On surfacing, the major events are always transmitted first.

This signal discrimination is necessary to optimize the cost of data transmission and limit power consumption. However, continuous sampled data are stored in a circular buffer with a capacity of 6 Gbytes. Any signals missed by the algorithm can be retrieved using two-way communication by satellite, and continuous signals can be read from the buffer when the float is recovered.

A Typical Mission

A typical recording cycle comprises several stages. The first stage is the descent. The acoustic detection algorithm is activated during the second stage when it drifts at the assigned depth for a specified duration. This second stage may be cut short if the acquisition board detects a seismic signal of significant amplitude. The hydrophone data acquisition is stopped before the ascent, to avoid interference of pump noise with earthquake detection algorithm. Once at the

Floating Seismographs (MERMAIDS), Fig. 3

A Mermaid being deployed from a research vessel



surface, the float fully fills its oil bladder to gain maximum buoyancy for optimal satellite communication even in high seas.

The float uses the GPS signal to get its geographical coordinates and to calculate the internal clock drift prior to synchronizing the clock on the MERMAID acquisition card against GPS time. GPS and Iridium communication system use the same helical special antenna to receive and transmit. Two-way communication allows for a command file containing float parameters such as depth and trigger level for the next dive to be downloaded. Seismograms are transferred to the SD card of the pilot board which performs the Iridium data upload. The float drift at depth is typically 4 km/day and rarely exceeds 10 km/day, whereas drift at the surface is usually between 1 and 2 km/hr; drift measurements by repeated GPS localizations allow for an accurate extrapolation of the float position during recording at depth (Nolet et al. 2019). Voltage level, internal vacuum, and external pressure are measured and a last GPS fix for MERMAID's clock synchronization is obtained before diving again.

If a major earthquake occurs in an area covered by one or more MERMAIDS, the floats have the option to revert to "lander" mode. This means that they can be commanded go down to the ocean bottom and stay anchored, much as an ocean bottom hydrophone (OBH). They can transmit recorded aftershock data by surfacing at programmed intervals until commanded to revert to their original mission at the end of the swarm. Similarly, landers can be rapidly deployed

for this purpose as an OBH network to monitor aftershocks of a large earthquake. Continuous recordings may be recovered at the end of a rapid response campaign while transmitting identified events in quasi-real time at every surfacing.

Deploying Mermaid can be done by simply pushing it overboard from a small vessel, or using a crane or U-Frame from a research vessel (see Fig. 3). To recover a Mermaid, it can be commanded to stay at the surface and send its geographical coordinates. It can be lifted aboard by hand from a small vessel or with a snap-hook and a crane on a larger vessel.

Other Sensors

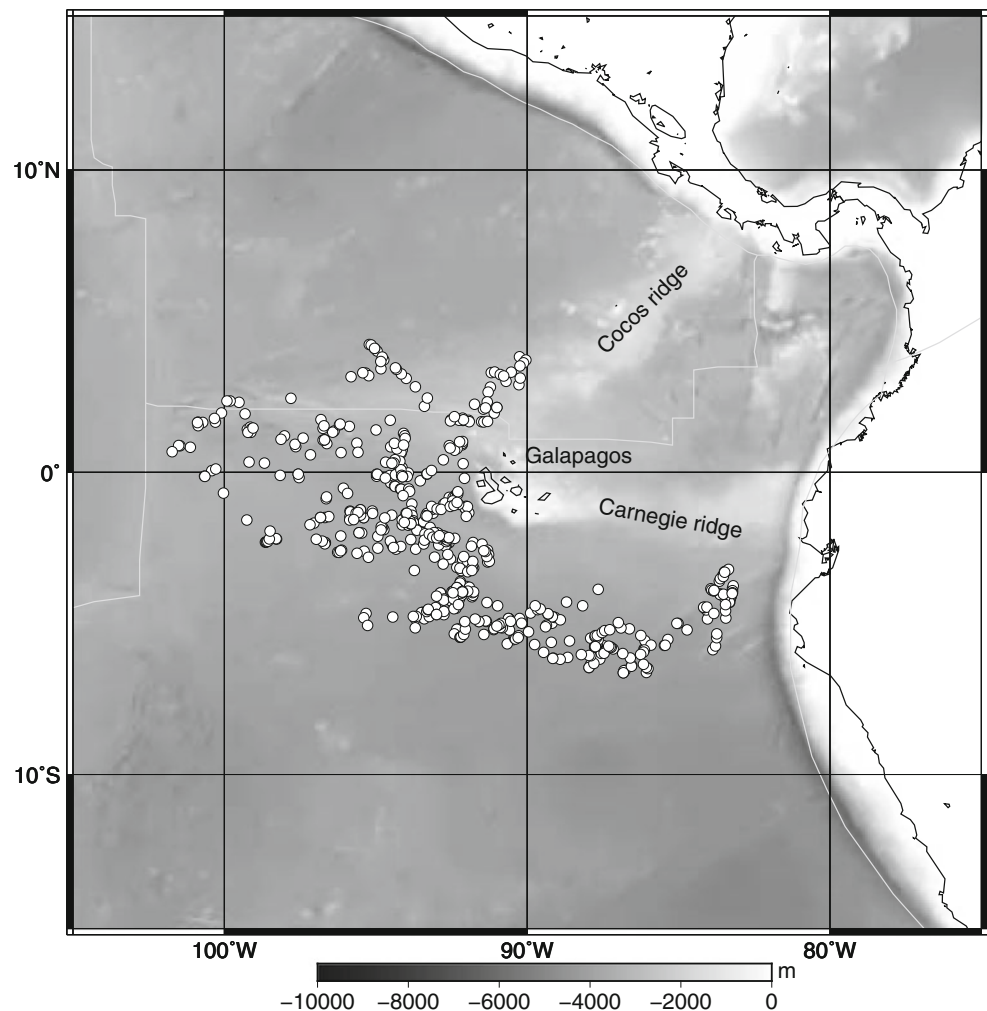
The MERMAID is a very recent addition to seismic instrumentation, but in its current version integrates functionality for oceanography, notably the option to accommodate a Sea-bird SBE 41 Conductivity Temperature Depth (CTD) sensor to measure salinity and temperature down to 4000 m depth. These data are compatible with the Argo program to monitor temperature, salinity, and abyssal currents in the oceans (Riser et al. 2018). The next development is a fully multidisciplinary Multi-MERMAID hosting multiple sensors.

Multi-MERMAID

The MERMAID board is equipped with eight channels for external sensors. For example, a high frequency hydrophone allows for the monitoring of rainfall and storms over the oceans; cracking and sliding of ice sheets and icebergs;

Floating Seismographs

(MERMAIDS), Fig. 4 The 9 MERMAIDS drift freely with the ocean current. Circles show the 580 locations where one of them successfully recorded a P wave, between May 2014 and October 2016. The grayscale denotes bathymetry in m. (Modified after Nolet et al. 2019)



tracking of whales and ecological stressors (ship and airgun noise). Sensors for bathymetric, magnetic, optical, geochemical, or biological measurements can also be added. Flexibility is obtained by allowing the data management by the MERMAID board to be programmed by the scientists. A Domain Specific Language is scheduled to be operational in 2020. The software can be updated using two-way communication by satellite.

Recent Results

Sukhovich et al. (2015) showed that a large fraction of earthquakes with magnitude >6.5 are detected by MERMAIDS (35–63%, depending on the signal to noise ratio being higher than the programmed threshold, typically 2–3). Although they are a very recent addition to seismic instrumentation, Mermaids have already contributed scientific results of significance. Sukhovich et al. (2015) recorded more than 200 events of an earthquake swarm near the Indian Ocean

triple junction that were not observed by land stations. His analysis showed the ability of Mermaids to detect nearby earthquakes with Richter scale magnitudes below 3. Nolet et al. (2019) used an array of 9 first generation MERMAIDS in the Pacific to image the plume under the Galapagos islands. This study resulted in a significant increase in image resolution in the upper mantle, due to the fact that the sensors change their position continuously, thus avoiding redundant ray paths (Fig. 4). Analysis of the images showed that the Galapagos plume carries a heat flux that is a factor of 6–7 higher than was assumed earlier on the basis of the bathymetry. Since 2019, a large network of 49 Mermaids is operating in the South Pacific in an effort to image the “superplume” beneath Tahiti (Simon 2020).

Summary

Floating seismographs have been made possible by recent advances in satellite communication, GPS miniaturization,

low power high performance electronic components, and increased battery capacities. Though limited to the recording of acoustic (P) waves in the water columns, they can cover large areas and offer a low-cost complement to OBS deployments. They bring detailed resolution to seismic tomography and the option of a fast response to monitor aftershocks or seismic swarms.

Cross-References

- [Earth's Structure, Global](#)
- [Earthquake, Magnitude](#)
- [Ocean Bottom Seismics](#)
- [Seismic Instrumentation](#)
- [Seismic Noise](#)
- [Seismic Tomography](#)
- [Seismicity, Intraplate](#)
- [Seismicity, Subduction Zone](#)
- [Seismogram Interpretation](#)
- [Seismological Networks](#)

Bibliography

- Bradner H, de Jerphanion LG, Langlois R (1970) Ocean microseism measurements with a neutral buoyancy free-floating midwater seismometer. *Bull Seismol Soc Am* 60:1139–1150
- D'Spain GL, Hodgkiss WS (1991) The simultaneous measurement of infrasonic acoustic particle velocity and acoustic pressure in the ocean by freely drifting swallow floats. *IEEE J Ocean Eng* 16:195–207
- Hello Y, Ogé A, Sukhovich A, Nolet G (2011) Modern MERMAIDS: new floats image the deep earth. *EOS Trans Am Geophys Un* 92:337–338
- Nolet G (1991) Imaging the deep earth: technical possibilities and theoretical limitations. In: A. Roca (ed) *Proceedings of the XXIIth Assembly ESC, Barcelona*, pp 107–115
- Nolet G, Hello Y, van der Lee S, Bonnieux S, Ruiz MC, Pazmino NA, Deschamps A, Regnier MM, Font Y, Chen YJ, Simons FJ (2019) Seismic tomography with floating seismometers: a first application to Galapagos. *Sci Rep* 9:1326
- Riser SC, Swift D, Drucker R (2018) Profiling floats in SOCCOM: technical capabilities for studying the southern ocean. *J Geophys Res Oceans* 123:4055–4073
- Simon JD (2020, June) Recording earthquakes in the oceans for global seismic tomography by freely-drifting robots. Ph.D. thesis, Princeton University
- Simons FJ, Nolet G, Georgief P, Babcock JM, Regier LA, Davis RE (2009) On the potential of recording earthquakes for global seismic tomography by low-cost autonomous instruments in the oceans. *J Geophys Res Solid Earth* 114:B05307
- Sukhovich A, Irissou J-O, Simons FJ, Ogé A, Hello Y, Deschamps A, Nolet G (2011) Automatic discrimination of underwater acoustic signals generated by teleseismic P-waves: a probabilistic approach. *Geophys Res Lett* 38:L18605
- Sukhovich A, Bonnieux S, Hello Y, Irissou J-O, Simons FJ, Nolet G (2015) Seismic monitoring in the oceans by autonomous floats. *Nat Commun* 6:8027

Fractal Scaling of Earthquakes

Simanchal Padhy¹ and Vijay P. Dimri²

¹Seismological Observatory, CSIR-National Geophysical Research Institute, Hyderabad, India

²CSIR-National Geophysical Research Institute, Hyderabad, India

Synonyms

Earthquake statistics; Power law

Definition

Fractals define a statistical distribution that is self-similar across different scales (scale-free/scale-invariance) characterized by a power-law with no characteristic length scale. Mathematically, they represent a class of non-linear system of equations with a self-similar solution, which can be modelled by using a recursive algorithm.

Scaling leads to a power-law relationship between the scale and the corresponding measure. It can be expressed as $T[f(x)] = f(\lambda x) = \lambda^n f(x)$, where T is scaling transform, x the scale variable, $f(x)$ the function of x , λ the scale factor, n the scaling coefficient.

A 1-D profile $z(x)$ is self-affine if it is statistically similar under scaling transformation $(x, z) \rightarrow (\lambda x, \lambda^H z)$, where λ is scale factor and H the Hurst exponent or roughness index ($0 < H < 1$). A self-affine profile is self-similar for $H = 1.0$, meaning that it is statistically identical with transformation $(x, z) \rightarrow (\lambda x, \lambda z)$. Therefore, a scaled portion of a self-similar profile appears statistically identical to the whole if the same scale factor is applied in both x and z directions, which for a self-affine case requires different scale factors depending on direction. Self-affine fractals/scaling are band-limited with smooth cut-offs.

Introduction

Heterogeneity exists in earth on scales from centimeters to hundreds of kilometers characterized by the scaling laws for different scale ranges with the crossovers separating them. Models explaining them are scale dependent; i.e., a single model cannot characterize the heterogeneity over the whole range, which requires going beyond self-similarity, i.e., multi-fractality. Scale-affine, band-limited fractal scaling, a special case of multi-fractality, can properly explain observations with local transient features characterized by crossovers.

In this chapter, we define fundamental notions of both the self-affine and self-similar fractal scalings, and their application in seismology. For the application, we present statistical treatment of distributed heterogeneities and their influence on rupture and seismic wave propagation and scattering, and finally suggest ways to reliably interpret observed scaling relations.

Fractal scaling implies invariance with respect to the two fundamental scaling symmetries of nature such as dilation (change in unit of length $r \rightarrow \lambda r$; λ the scale factor), and translation (change in the origin of coordinate system $r \rightarrow r + b$, b the constant). They satisfy, both in time or frequency domain, the power-law form of any frequency-size distribution $F(s) \propto s^{-\alpha}$ (s is scale parameter, and α scaling exponent) that is valid for a large, at least for one order of the s values.

In a self-similar set S of N non-overlapping subsets, each subset of S is congruent to λS ($0 < \lambda < 1$), meaning that the whole set is composed of N copies of itself after scaling down by λ (i.e., identical under translation and rotation). The fractal or similarity dimension D of the set S then follows the relation $N\lambda^D = 1$. For self-affinity, however, each of the N subsets is scaled down from the whole by a different λ_n , satisfying $\sum_{n=1}^N \lambda_n^D = 1$.

A self-affine series $h(x)$ defined along a profile of length L can be formulated with Fourier sine series as $h(x) = \sum_{n=1}^{\infty} A_n \sin(2\pi nx/L)$, where the coefficient A_n is related to wave length $\lambda_n = L/n$, $n = 1, 2, 3, \dots$ through a power-law $A_n = C\lambda_n^{\frac{2-D}{2}}$, where C is constant and D the fractal dimension (Turcotte 1991); this relation was obtained by counting the number of rectangular boxes of various sizes to cover the self-affine profile in contrast to the square boxes of a given size used to cover the self-similar profile.

For a 1-D profile of slip roughness, self-affinity means invariance of roughness of a profile observed at direction-dependent different scales and is characterized by a power-law dependence of power spectrum on wave-number k as $p(k) \propto k^{-\beta}$, with $\beta = 2H + 1$. Cross-sections of topography and bathymetry are examples of self-affine function (Brownian motion or random walk) where the height scales as $|h(x+l) - h(x)| \approx l^H$; they can be reproduced by integration of white Gaussian noise ($\beta = 0$).

Scaling of Earthquake Sources

Earthquake scaling, generally referred to as self-similar unless otherwise stated, implies identical physics of earthquake sources under a change of scale. This means that similar physics holds good for all earthquakes regardless of

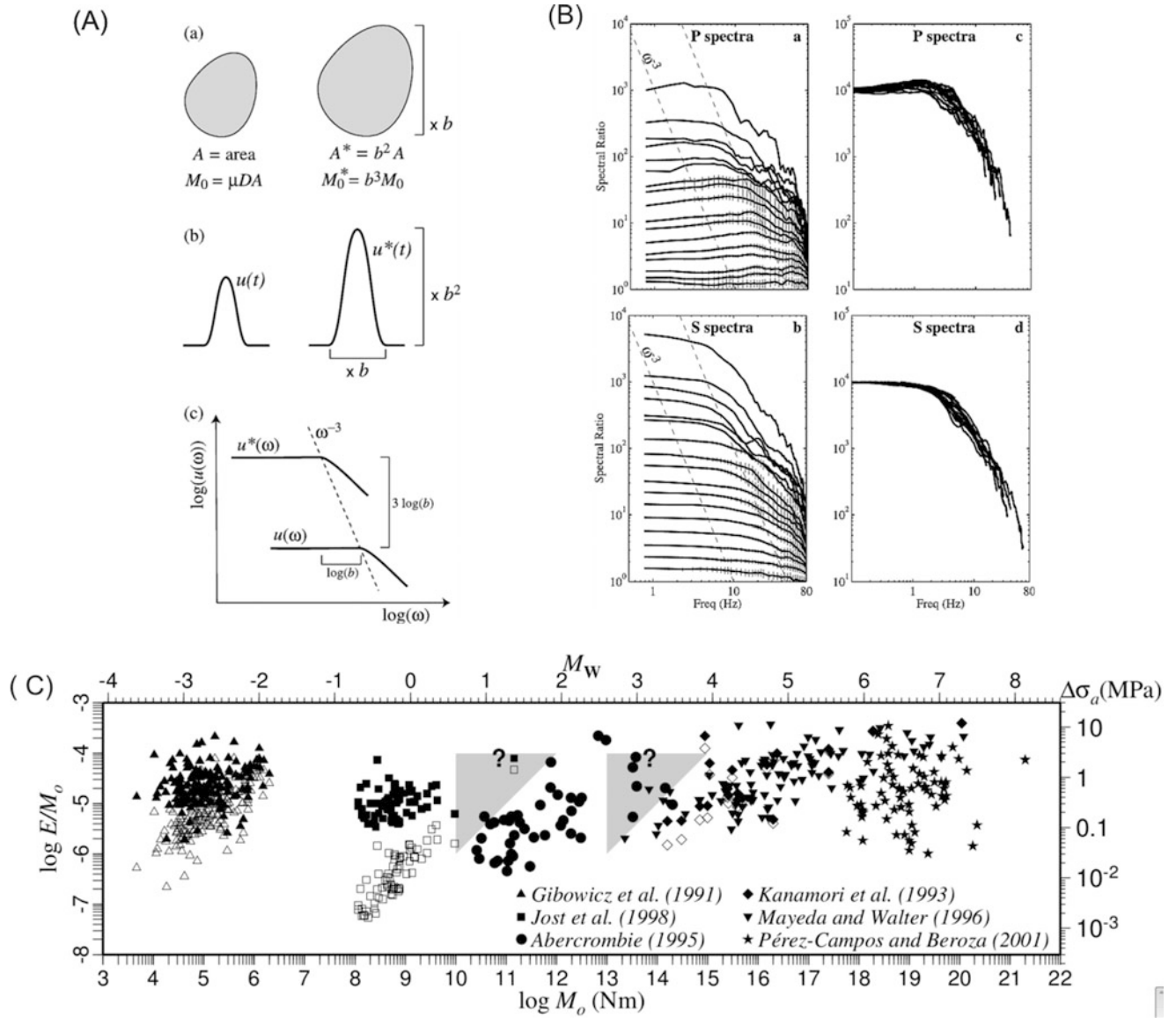
size, depth, tectonics, and geology. Figure 1A demonstrates the concept of earthquake scaling. Increase in earthquake size by a factor of b scales (increases) the displacement, fault area, and seismic moment by b , b^2 , and b^3 , respectively. Self-similarity predicts, for an identical source-receiver geometry in a loss-less media with constant rupture velocity between earthquakes, that the far-field displacement pulse will increase in duration by b , and in amplitude by b^2 ; the displacement spectra of different earthquakes scaled proportionally have identical shapes but shifted along the ω^{-3} line, as found from the use of similarity theorem in frequency domain (Prieto et al. 2004).

There is a long-standing debate on the existence scaling in seismology: whether stress drop and apparent stress (ratio of seismic energy to seismic moment) are constant, independent of seismic moment (or magnitude). One way to test this hypothesis is stacking of the log spectra of P - and S -waves, after the events are corrected for site and path-related effects. The identical shapes of the log spectra of events of different moment bins that are offset along the ω^{-3} line (Fig. 1B) (Prieto et al. 2004) suggest self-similarity of earthquake spectra. Observations such as the (i) $f_c \propto M_0^{-1/3}$ relation between the corner frequency and seismic moment, and (ii) almost constant apparent stress with moment, suggest a self-similar constant stress-drop scaling for the events (M_L 1.8–3.4) in southern California.

Limited bandwidth of the recorded events underestimates the seismic energy, with the energy more than the corner frequency lacking from the data (missing energy). The events, corrected for such finite bandwidth effects, show almost constant values of apparent stress (3×10^{-5}) and stress drop (1 MPa) over 17 orders of seismic moment (Ide and Beroza 2001) (Fig. 1C). They report that any departure from self-similar scaling could be a manifestation of inadequate correction of data during pre-processing for the site- and path-related effects, finite-frequency effects, etc. (for reviews see Ide and Beroza 2001; Prieto et al. 2004; Padhy and Subhadra 2016 and references therein).

Scaling of Distribution in Slip Roughness

Roughness in fault/fracture surface exhibits fractal geometry, but it is not yet clear whether slip distribution of real earthquakes is self-similar or self-affine. Spatial correlations of slip maps compiled from large earthquakes show self-affinity behavior with an average $H = 0.71 \pm 0.23$, regardless of size and source dimension (Mai and Beroza 2002). Self-affine roughness for the H values different from 1.0 implies a larger roughness variability compared to self-similarity ($H = 1.0$). A generalized self-affine fractal, in the assumption of anisotropic fractal roughness, properly explains both constant stress-drop (static) and complex rupture propagation (dynamic) models of earthquakes. Heterogeneity in slip can



Fractal Scaling of Earthquakes, Fig. 1 Concepts of self-similar earthquake scaling. (A) Effects of self-similarity on (a) rupture area, and the seismic moment and energy when an earthquake is scaled (increased) in size by a factor of b , (b) far-field displacement source functions, and (c) far-field displacement spectra. (B) Displacement spectra of events shown for different moment bins (left) and those obtained after shifting along the ω^{-3} line (right) for P- and S-waves. (After Prieto

et al. 2004; © 2004 AGU) Note the common shape of spectra within errors suggesting the self-similar earthquake scaling. (C) Variation of apparent stress and stress drop with seismic moment for different data sets. The open and closed symbols, respectively, represent the original and estimates corrected for the limited bandwidth effect. The gray triangles are the locations of events with missing energy. (After Ide and Beroza 2001; © 2001 AGU)

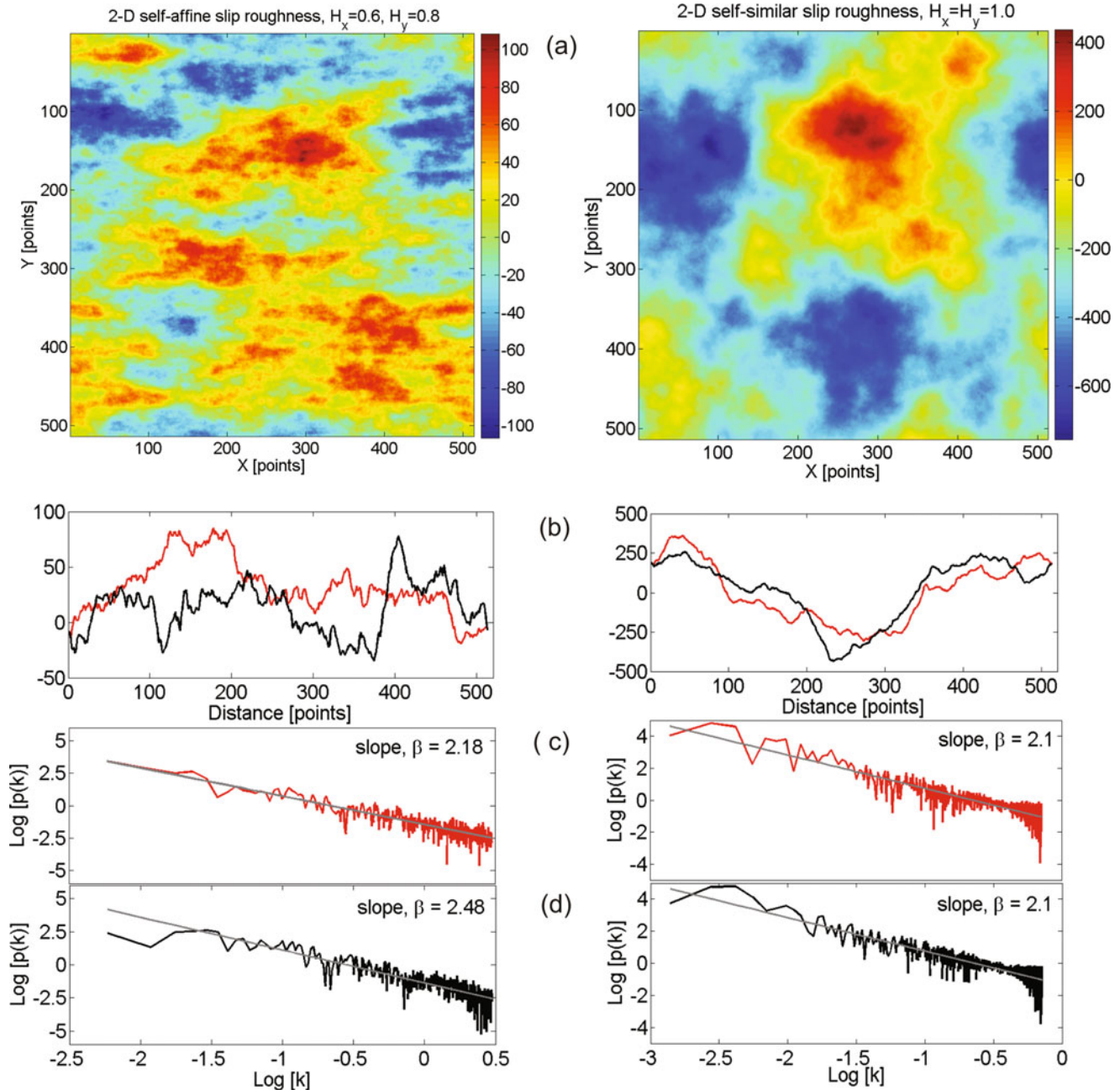
stochastically be described by spatially filtered random functions with a spectral decay of $k^{-\alpha}$, where k is wave number and the exponent α , a measure of smoothness in spatial slip distribution. This slip model helps in (i) interpreting the scaling of average slip with rupture length ($\alpha \sim 1.5$ suggests constant slip-length ratio and thus constant stress drop), and (ii) correlating with frequency-size event statistics ($\alpha < 1.5$ implies more small events than large ones) (Liu-Zeng et al. 2005).

Stress and strain changes corresponding to the slip-distribution of the type mentioned above have a spectral decay of $k^{-\alpha+1}$, $\alpha = 1.0$ represents a Gaussian and $\alpha < 1.0$ representing stress changes that are deficient in long-wavelength variations compared to random functions. In 2-D, the amplitude spectrum of the stress field can be expressed as $\tau(k) \propto k_c^{-(1+H)} = \text{const}$ for $k \leq k_c$, and $\tau(k) \propto k^{-(1+H)}$ for $k > k_c$ (Mai and Beroza 2002), where $k = \sqrt{k_x^2 + k_z^2}$ in x-z plane, suggesting that the spectral decay

is governed by the Hurst exponent H above the corner wave number k_c . The random stress field is stochastically realized by the exponent H and the correlation length $a_c = \frac{2\pi}{k_c}$. The exponent H governs the high wave-number fall-off of the stress heterogeneity spectrum that is related to fractal dimension D by $D = d + 1 - H$, d the Euclidean dimension. The stress spectrum takes the form of $|\tau(k > k_c)| \propto k^{-1}, k^{-1.5}, k^{-2}$

for $H = 0, 0.5$, and 1.0 , where $|\tau(k)| \propto k^{-1}$ corresponds to the self-similar scaling (stress drop independent of rupture size).

Figure 2a shows the map of slip roughness of size (512 points $\times$ 512 points) generated from one random realization with the H values along the X - and Y -axes: ($H_x = 0.6$, $H_y = 0.8$ for self-affinity (**left panel**) and $H_x = H_y = 1.0$ for self-similarity (**right panel**). The slopes of the power spectra



Fractal Scaling of Earthquakes, Fig. 2 Scaling of roughness of fault surface for self-affine (left panel) and self-similar (right panel) fractal media. (**Left panel**) (a) Map of fault roughness generated with one of the random realization of the stochastic process ($H_x = 0.6$, $H_y = 0.8$). (b) Roughness along the two selected profiles parallel to X - and Y -axes. (c–d)

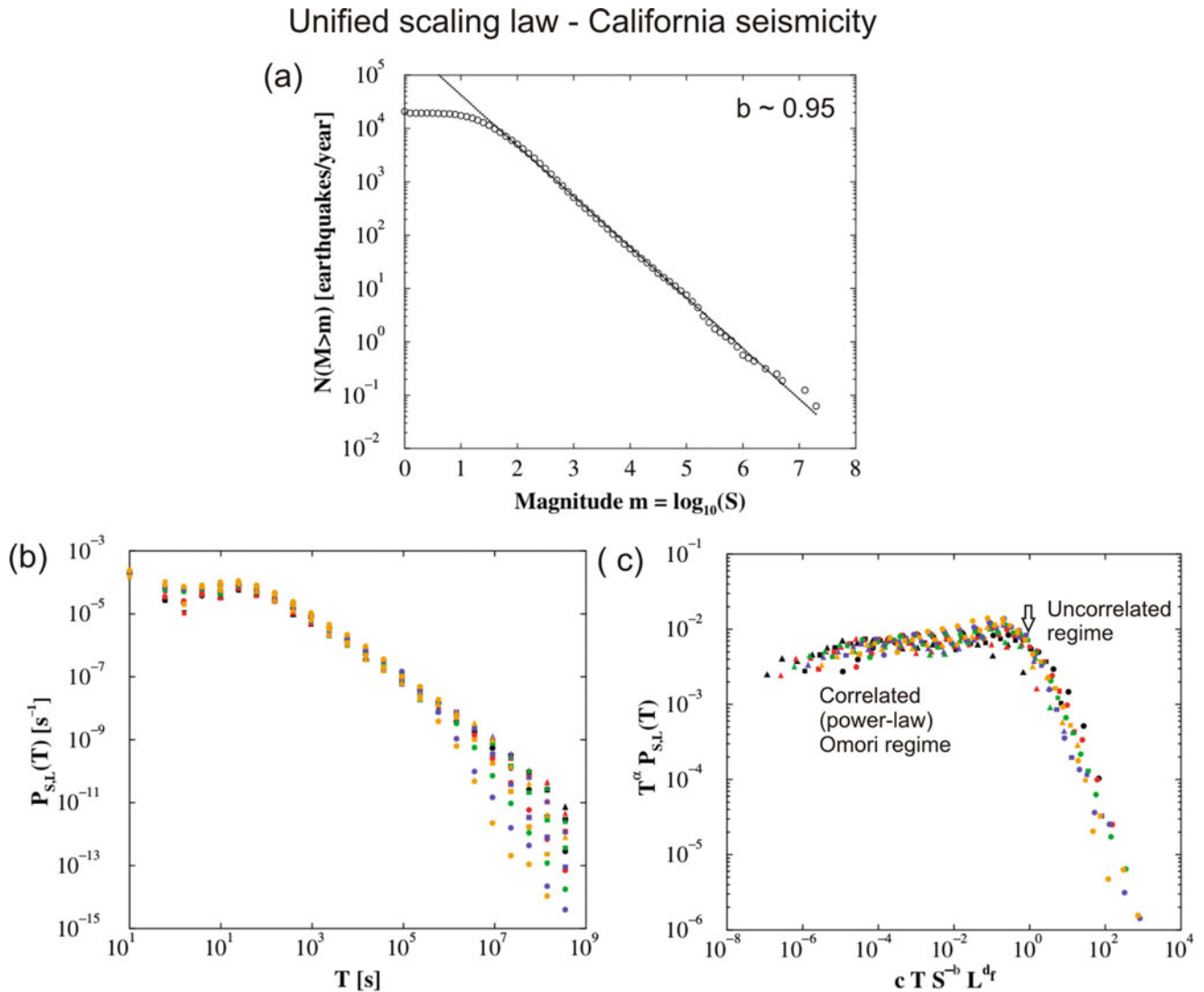
power spectra of the profiles shown in (b). The spectral exponents estimated from the slopes to the spectral fits (solid lines) are shown. (**Right panel**) Same as the left panel, but for a self-similar distribution ($H_x = H_y = 1.0$)

of the two profiles selected along X - and Y -axes of the map are compared for both the scaling types (Fig. 2b–d); their spectral dimensions are nearly the same for the self-similar surface, while they differ for self-affinity, as expected. The roughness variability for the self-similar surface is relatively smoother than that for the self-affine one, because of the variation with direction of roughness for the latter.

Unified Scaling Law

There is a widespread evidence of earthquake scale-invariance from observations such as the magnitude (m)/energy (s)

distribution in time $N(M > m) = 10^a - b m$, $N(S > s) \propto s^{-b}$, $m = \log s$, $b \sim 1$ (Gutenberg-Richter, G-R law) (Fig. 3a), the temporal correlation T^{-p} , $p \sim 1$ (Omori law), $m \propto \log(cM_0)$ ($c \sim 1.5$), $M_0 \propto A^{3/2}$, $N(S > s) \propto L^{-D}$ (M_0 : seismic moment, A : rupture area, L : rupture length, and D : fractal dimension) and space-time clustering of events with a fractal distribution (Dimri 2005). Unified scaling law provides a framework explaining all these features together, where a scale of $S^{-b}L^D$ is suggested combining S and L (Bak et al. 2002). The distribution of time interval between two successive earthquakes (waiting time) T occurring within an area of size L with magnitudes more than m shows a band-limited power-law regime with a cut-off (upper limit) waiting time (Fig. 3b). It shows the



Fractal Scaling of Earthquakes, Fig. 3 Unified scaling law and SOC. (a) Magnitude (≥ 2) distribution of earthquakes (circles) in California. The G-R frequency-size distribution (solid line) with $b \sim 0.95$ is shown. (b) Distribution of waiting time between earthquakes shown for different values of cut-off magnitude S (or m) and areal extent of size L : $m = 2$ (circles), 3 (squares), and 4 (triangles); and $L = 0.25^\circ$ (black), 0.5° (red),

2° (blue), and 4° (orange). (c) Presentation of rescaled data shown in (b) for $T > 38$ s, re-plotted with $T^\alpha P_{SL}(T)$ as a function of $x = c T S^{-b} L^D$, $c = 10^{-4}$. The inverted arrow marks the region of correlated power-law regime from uncorrelated one. The values $\alpha = 1$, $b = 1$, $d_f = 1.2$ allow all data to fall onto a single curve. (Bak et al. 2002; © 2002 APS)

Omori law regime increases with increasing S (or m), and decreases with increasing L . Rescaling the axes [$x = TS^{-b}L^D$ and $y = T^p P_{S,L}(T)$] of Fig. 3b shifts the curves, causing them to fall onto a single curve $y = f(x)$, implying a unified scaling law for earthquakes (Fig. 3c) (Bak et al. 2002). Temporal correlation of events is thus decided by the product $x = TS^{-b}L^D$ but not on T , S , and L individually. The curve $f(x)$ is almost constant over eight orders of magnitude for $x < 1$, corresponding to a correlated Omori law regime, while it decays fast for $x > 1$, related to uncorrelated events (Bak et al. 2002). Any departure from power-law scaling can significantly be reflected with this analysis.

Statistical Treatment of Fractal Heterogeneities in Earth

Heterogeneities in the earth are fractal in a wide scale range. The growth of fractal structures is attributed to several geological processes. It is of considerable interest to study those processes quantitatively through fractal analyses of the distributed heterogeneities in geological structures. The fluctuations in seismic velocity can be realized with a random function $f(\mathbf{x})$ of zero mean $\langle f \rangle = 0$ and finite covariance defined by $\langle f(\mathbf{x})f(\mathbf{x} + \mathbf{r}) \rangle = \sigma^2 N(\mathbf{r})$, where σ is the standard deviation and $N(\mathbf{r})$ the ACF with $N(\mathbf{0}) = 1$. The energy spectrum can be obtained from the Fourier spectrum of the ACF as $E(\mathbf{k}) = \sigma^2 \int N(\mathbf{r}) e^{-i\mathbf{k} \cdot \mathbf{r}} d\mathbf{r}$, where $\mathbf{k}$ is the wave vector. The self-affine fractal properties of the medium inhomogeneities can be described by the vonKarman ACF as $N(\mathbf{r}) = \frac{2^{1-H}}{\Gamma(H)} \left(\frac{r}{a}\right)^H K_H\left(\frac{r}{a}\right)$, and the corresponding energy

$$\text{spectrum as } E^d(\mathbf{k}) = \sigma^2 \left| \frac{\Gamma(H+\frac{d}{2})}{\Gamma(H)} \right| \pi^{d/2} \frac{(2a)^d}{(1+a^2 k^2)^{H+\frac{d}{2}}} \quad (\text{Lord 1954}),$$

where a is correlation length, H the Hurst exponent characterizing the roughness of the medium, and d is Euclidean dimension. The energy spectrum of the distributed heterogeneities is used to predict wave propagation and scattering of seismic waves in 3-D fractal media.

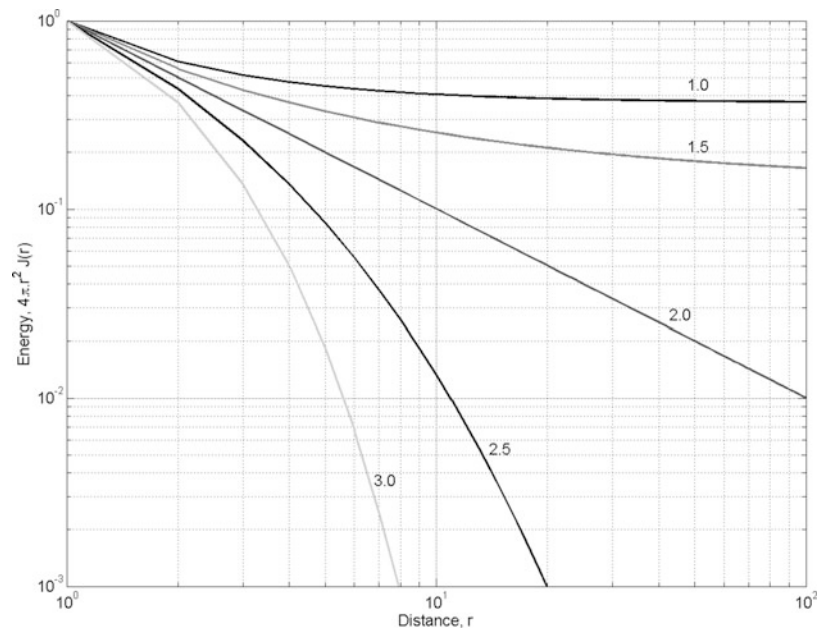
Seismic Attenuation in Fractal Media

Seismic waves at high frequencies (say, 1–25 Hz) undergo scattering from heterogeneities in velocity with length scale in the range of a few hundred meter to a few kilometer. Other heterogeneities on the similar scale length could be in the density, geometry or fluid content of cracks/fractures associated with shear faulting. A fractal medium refers to scale-invariance in fractures/faults in the earth's crust, suggesting that (a) all the cracks have same aspect ratio, (b) crack-lengths are power-law distributed, and (c) fracture surfaces have the same roughness (Main et al. 1990). There is an upper and lower bound to the length scale defining the power-law (band-limited self-affine law). This band-limited fractal geometry satisfies the physical processes at a finite range of frequencies in order to satisfy the Kramers-Kronig relation for explaining the frequency dependence of attenuation in the lithosphere, characterized by a dimensionless factor $Q(f)$ (Main et al. 1990). In 3-D, the power spectrum of the heterogeneities in velocity produced by distributed cracks takes the form $S(k) \propto k^{-\beta}$ that explains the power-law frequency dependence of Q^{-1} as $Q^{-1} \propto \omega^3 - \nu$.

For a fractally homogeneous spatial distribution of scatterers with number density $n(r)$ [$n(r) \propto r^{D-3}$], where r is

Fractal Scaling of Earthquakes,

Fig. 4 Dependence of seismic energy with distance for different values of fractal dimension in a fractal medium of fractally distributed heterogeneities



distance and D is fractal dimension, energy decay with distance of seismic waves can be expressed as $J(r) = \frac{1}{4\pi r^2} e^{\left[-\frac{G}{D-2}r^{D-2} + \text{const}\right]}$ for $D \neq 2$ and $J(r) = \frac{1}{4\pi r^2} \frac{\text{const}}{r^G}$ for $D = 2$, where G is a constant, a measure of scatterer cross-section (Sato 1988). Exponential decay of energy with distance is described as a special case for $D = 3$ (Sato 1988). Figure 4 shows the energy decay with distance for different values of fractal dimension obtained from the above expression with $G = 1$, and $\text{const} = 1$, illustrating the dependence of the shape of the decay with fractal dimension.

Summary and Outlook

Scaling, a powerful technique in science, allows one to analyze earthquake process on any other scale than it is observed. Earthquake phenomena described by fractal distributions are said to scale, meaning that same physics must work at different scales analyzed. Scaling analyses can detect regions of power laws and its types in earthquake data. Observation of self-similar power-law scaling in seismology is widespread. Presence of more than one distinct scaling region in 1-D signal suggests to go beyond self-similarity (multi-fractality and hence self-affinity as its special case), however. In 2-D, self-affine, rather than self-similar map of slip roughness can tell us larger and anisotropic roughness variability. Such detailed source slip models are important for high-frequency ground-motion modelling. There is a great debate on the existence of self-similar scaling in earthquake studies, which needs proper treatment of high-quality data (sufficient data length, wide magnitude range, etc.) before interpretation. Inadequate corrections of the data for the path- and site-related and finite-frequency effects (lack of energy at frequency more than the corner-frequency) may yield biased estimates of scaling relations. Also, the use of correct theory in deriving the scaling laws is important. The unified scaling functions that can rescale the multidimensional (space, time, and magnitude) scaling relationships to a function of single variable confidently detects the region of power-law regime in the distribution. The analogy among the scale invariance of earthquake processes, observation of self-similar earthquake statistics and self-similar structures at different scales is still poorly understood, which probably needs detailed numerical studies to explain observed scaling relations.

Cross-References

► Fractals and Chaos

Acknowledgments We are grateful to Professor Harsh K. Gupta (the Editor) for inviting us to contribute this chapter. Director, CSIR-NGRI is thanked for granting permission for its publication [NGRI/Lib/2020/Pub-32].

Bibliography

- Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. *Phys Rev Lett* 88:178501-1–178501-4
- Dimri VP (2005) Fractals in geophysics and seismology: an introduction. In: Dimri VP (ed) *Fractal behavior of the earth system*. Springer, New York, pp 1–22
- Ide S, Beroza GC (2001) Does apparent stress vary with earthquake size? *Geophys Res Lett* 28:3349–3352
- Liu-Zeng J, Heaton T, DiCaprio C (2005) The effect of slip variability on earthquake slip-length scaling. *Geophys J Int* 162:841–849
- Lord RD (1954) The use of the Hankel transform in statistics I. General theory and examples. *Biometrika* 41:44–55
- Mai PM, Beroza GC (2002) A spatial random field model to characterize complexity in earthquake slip. *J Geophys Res* 107(B11):2308
- Main IG, Peacock S, Meredith PG (1990) Scattering attenuation and the fractal geometry of fracture systems. *Pure Appl Geophys* 133:283–304
- Mandelbrot BB (1983) *The fractal geometry of nature*. W. H. Freeman, New York
- Padhy S, Subhadra N (2016) Spectral scaling and seismic efficiency for earthquakes in northeast India. *Bull Seismol Soc Am* 106(4):1613–1627
- Prieto GA, Shearer PM, Vernon FL, Kilb D (2004) Earthquake source scaling and self-similarity estimation from stacking of P and S spectra. *J Geophys Res* 109. <https://doi.org/10.1029/2004JB003084>
- Sato H (1988) Fractal interpretation of the linear relation between logarithms of maximum amplitude and hypocentral distance. *Geophys Res Lett* 15:373–375
- Turcotte DL (1991) Fractals in geology: what are they and what are they good for? *GSA Today* 1(1):2–4

Fractals and Chaos

Ravi P. Srivastava¹, Nimisha Vedanti² and Vijay P. Dimri²

¹Equinor ASA, Bergen, Norway

²CSIR-National Geophysical Research Institute, Hyderabad, India

Synonyms

Power law/Scaling/Long-range dependence and Disorder/Randomness; Self-similar/Self-affine

Definition

Fractals – Often refers to the geometry of the objects. Objects having a fractional dimension are termed as a fractal. Also, such objects are self-similar in nature; thus, their shape is independent of the scale and remains the same under dilatation or contraction. The concept is extended to the time series analysis, and the time series in which the power spectrum follows power law is termed as “fractal time series.”

Chaos – Often refers to the dynamic state of a system. The behavior of chaotic systems is highly sensitive to initial conditions. A small change in initial conditions, even due to rounding-off errors in computation, can yield widely diverging outcomes for chaotic systems, rendering long-term prediction impossible. The chaos can be quantified by Lyapunov exponent if it is positive.

Introduction

The origin of fractals dates back to the eighteenth century in the study of nondifferentiable functions by Karl Weierstrass, Georg Cantor, and Felix Hausdorff; however, the term fractal was coined by Mandelbrot (1967, 1983) in a paper published in science “How long is the coast of Britain?” and later in his book “Geometry of Nature.” Fractals and Chaos are relatively new mathematical innovations, which find their applications in several branches of pure and applied sciences. Applications of fractals and chaos have been widely explored in solid earth geophysics, and its successful applications are demonstrated in many areas of geophysics viz: gravity, magnetic modeling (Turcotte 1997; Dimri 2005; Maus and Dimri 1995), seismic inversion, earthquake frequency magnitude relationship (Ouchi and Uekawa 1986; Dimri et al. 2005), analysis of reflectivity sequence (Painter et al. 1995; Toverud et al. 2001), and fractal representation of soil structures to characterize soil water properties in terms of porosity and particle size distribution (Rieu and Sposito 1991). Another important study is the simulation of fractures using a fractal method (Borodich 2000; Tanaka et al. 2001; Dimri and Srivastava 2019). Chaos has been used in studies related to the atmosphere, climate, and earthquake studies (Grassberger and Procaccia 1983; Srivastava 2005).

Basic Concepts

Power Law (Scaling Behavior)

Spectral analysis of fractal time series shows that power (P) is proportional to frequency (f) raised to power a real exponent (β), termed as scaling exponent (Fig. 1a, e). In mathematical form it is represented as:

$$P(f) \propto f^{-\beta} \quad (1)$$

This behavior is also known as scaling law.

The fractal analysis of time series is fundamentally based on the concept of fractional Brownian motion and fractional Gaussian noise. The fractal time series, also known as time series having long memory are characterized by fractional Gaussian noise (fGn). A time series is said to follow fGn characteristic (i) if its histogram is fat-tailed Gaussian,

(ii) Covariance, semivariogram, and power spectrum follows power-law behavior, and (iii) Hurst coefficient lies within [0.5, 1.0] as shown in Fig. 1. A popular method of Hurst coefficient estimation is Re-Scaled range analysis (R/S analysis). Below we discuss the definition and basic concepts of fractional Brownian motion and fractional Gaussian noise.

Fractional Brownian Motion (fBm)

Fractional Brownian motion is a generalization of classical Brownian motion (cBm). The fractal nature of fBm can be understood from the following mathematical definition. Assume that $f(t)$ is a single-valued continuous function of time/space (t). Let the increments (or difference of adjacent values) $\delta f = f(t_2) - f(t_1)$ follow Gaussian distribution (Molz and Boman 1993) with the following properties:

$$\left. \begin{aligned} E < f(t_2) - f(t_1) > &= 0, \text{ and} \\ E < [f(t_2) - f(t_1)]^2 > &= \alpha |t_2 - t_1|^{2H}, \end{aligned} \right\} \quad (2)$$

where $E<X>$ is the mean or expected value of X , α is constant of proportionality and H is an exponent known as Hurst coefficient (Mandelbrot and Van Ness 1968; Voss 1985a, b) lying between 0 and 1 ($0 \leq H \leq 1$). The expected values are obtained by averaging over many increments of size $|t_2 - t_1|$. The equation (2) holds for any value of time increment $\delta t = |t_2 - t_1|$. It is obvious from equation (2) that for $H = 0.5$, $E < [f(t_2) - f(t_1)]^2 > = \alpha |t_2 - t_1|$, so the fractional Brownian motion reduces to classical Brownian motion (cBm).

Fractional Gaussian Noise (fGn)

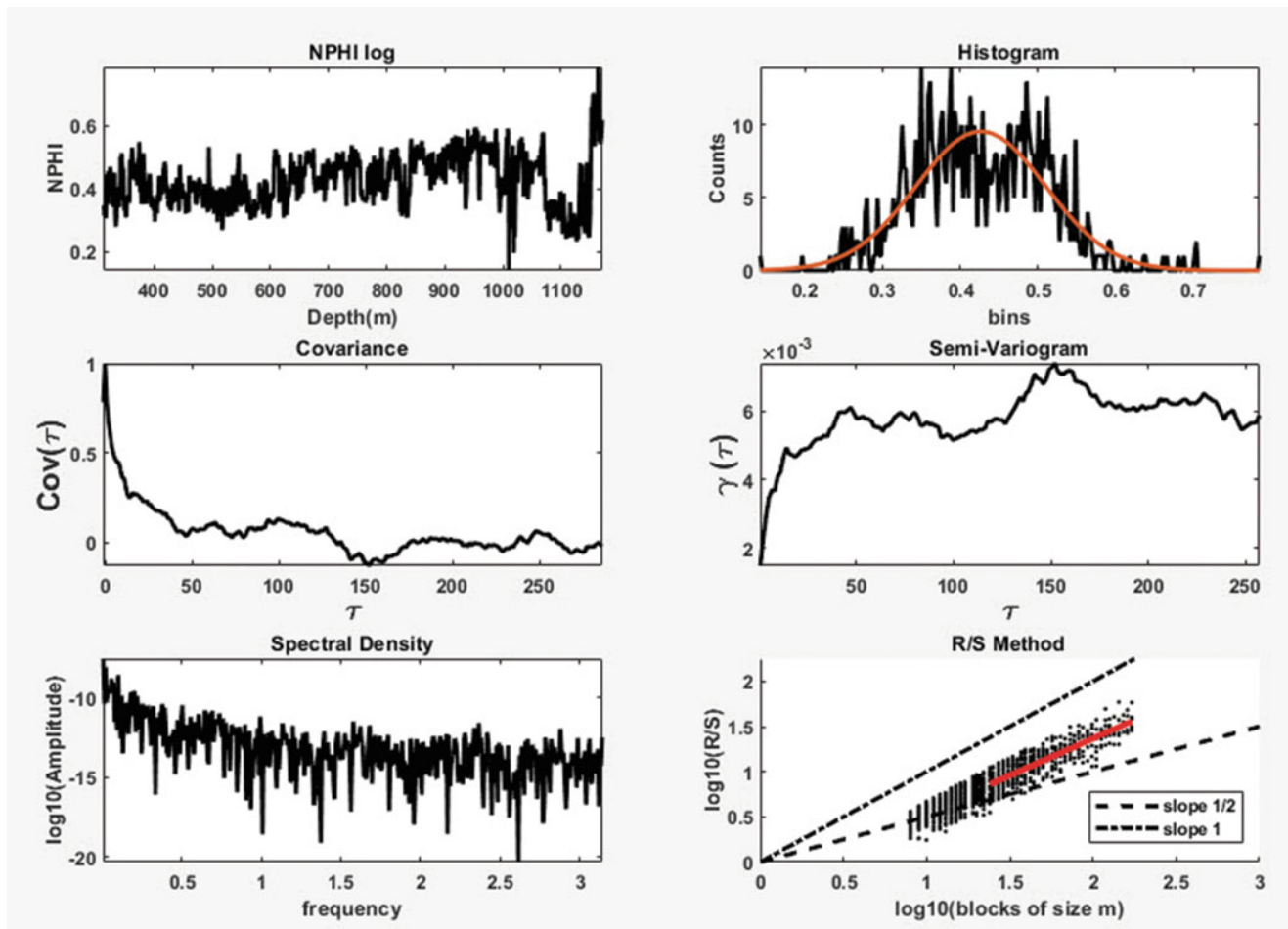
The concept of fractional Gaussian noise is intertwined with the fractional Brownian motion. The detailed discussion and criteria to distinguish between these two are explained in the paper by Hewett (1986) and Hardy and Beier (1994). Here we describe in brief the relationship between fBm and fGn.

Let, $x_1, x_2, x_3, \dots, x_n$ be the sequence of numbers separated by time interval δt that represents a fBm process, then one can generate corresponding fGn or time derivative process by computing the sequence $[(x_2 - x_1), (x_3 - x_2), (x_4 - x_3), \dots, (x_n - x_{n-1})]/\delta t$. Conversely, if one starts with a fGn sequence of numbers given by $y_1, y_2, y_3, \dots, y_n$, then numerical integration will yield the corresponding fBm sequence given by:

$[y_1, (y_1 + y_2), (y_1 + y_2 + y_3), \dots, (y_1 + y_2 + y_3 + \dots + y_n)]$. A time/space series with Hurst coefficient $H > 0.5$ is characterized as fGn.

Fractal Signal Analysis

Geophysical signals are often in the form of time/space series, which can be characterized by a combination of stochastic



Fractals and Chaos, Fig. 1 Test of fractal behavior of the time series (a) given neutron porosity (NPHI) well log data (b) Histogram of the NPHI well log data (c) Covariance of the NPHI well log data (d) Semivariogram of the NPHI well log data (e) Log power spectrum of

the NPHI well log data (f) Hurst coefficient ($H = 0.80$) of the NPHI well log data showing fractional Gaussian noise (fGn) characteristic of the well log

component, trend component, and periodic component (Malamud and Turcotte 1999). To quantify the stochastic component of the time series, it is necessary to specify the statistical distribution of values and persistence. There are a variety of techniques to quantify the strength of persistence, but the most commonly used is the spectral analysis, where the Fourier spectrum of time series is plotted against frequency (or wavenumber in case of space series) and the value of the slope, known as scaling exponent (say β), gives an estimation of persistence. Persistence study is useful in understanding the nature of the phenomenon. Depending upon the value of scaling exponent, persistence can be characterized as “weak” or “strong.” Scaling exponent controls the balance of high and low frequencies and determines the degree of smoothness or correlation of the time series.

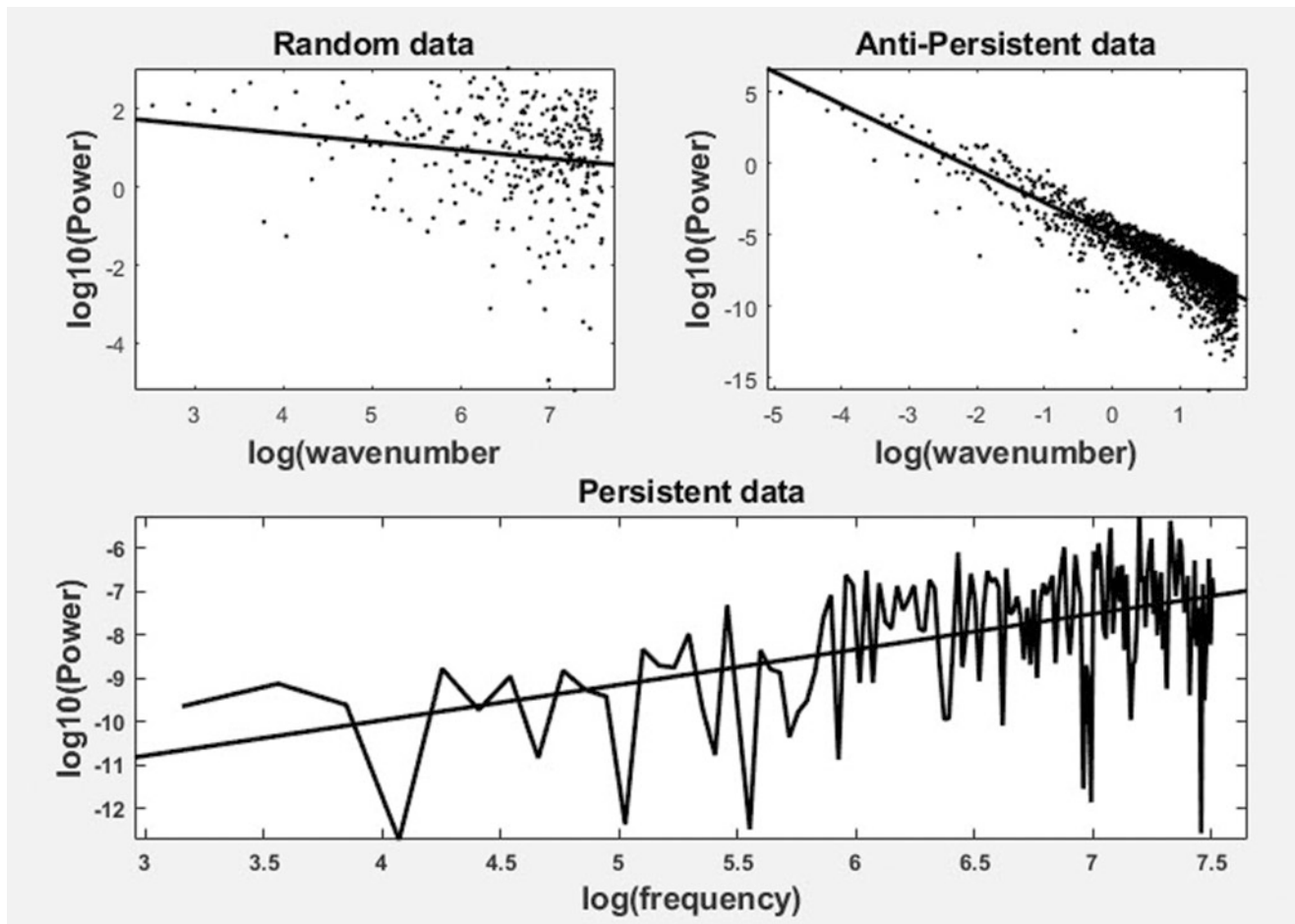
Analyzing Geophysical Time Series

Persistence of the time series is a measure of the correlation between adjacent values of the terms in time series. The

uncorrelated or random data with zero persistence is the white noise. The case $\beta = 0$ (Fig. 2a) has a power spectrum, independent of frequency, and is white noise. The values of such a time series are uncorrelated and completely independent of one another. The time series is antipersistent if adjacent values are anticorrelated. When $\beta < 0$ (Fig. 2b), the series is anticorrelated and each successive value tends to have the opposite sign. Noise with $\beta = -1$ is sometimes called flicker noise, while that with $\beta = -2$ is the Brownian noise. The time series is persistent if adjacent values are positively correlated to each other. For $\beta > 0$ (Fig. 2c), the series is positively correlated, i.e., persistent. Most of the well-logs exhibit antipersistent character, whereas reflectivity shows persistent character.

Fractal Behavior of Various Physical Properties

The Earth’s inherent complexity makes it challenging to infer the location, distribution, and structure of rock types, grain size distribution, material strength, porosity, permeability,



Fractals and Chaos, Fig. 2 Power spectrum of (a) random data, (b) antipersistent data, and (c) persistent data. (Modified from Dimri 2005)

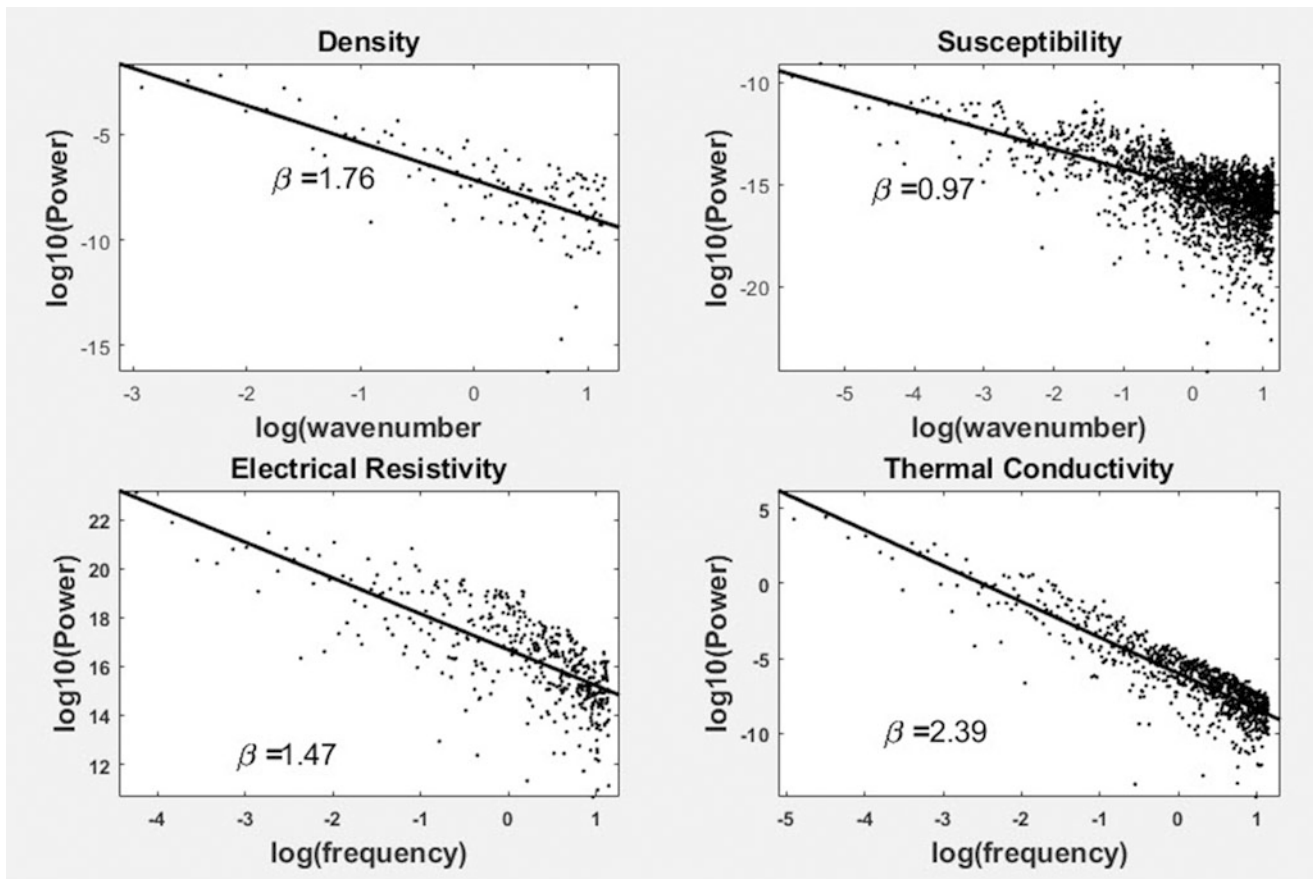
etc. from the direct observations; these characteristics are often inferred from the distribution of fundamental physical properties such as density, electrical conductivity, acoustic impedance, and many others properties. For a geophysical study, the physical response of the Earth can be approximated by the convolution model (Dimri 1992). The extraction of useful information from the observations made over the surface needs advanced processing and interpretation techniques. Physical properties of the earth system follow fractal (scaling) distribution. Examples showing the scaling power spectrum of well-log data from deep KTB pilot borehole are shown in Fig. 3. Observed scaling distribution of physical properties is contrary to the commonly assumed random distribution of properties, which offers simple mathematical formulation. An attempt has been made to reformulate interpretation techniques of geophysical exploration based on scaling (fractal) distribution of physical properties (Dimri 2005; Dimri et al. 2012).

Fractals in Seismology

Fractal concepts can be efficiently used to characterize the seismicity pattern of a region. The fractal nature of the spatial

distribution of earthquakes was first demonstrated by Kagan and Knopoff (1980), Hirata and Imoto (1991), and Hirabayashi et al. (1992). The hypocenter distribution data suggest that the changes in fractal dimension could be a useful precursor parameter for earthquakes as it is a measure of the degree of clustering of seismic events. A change in fractal dimension corresponds to the dynamic evolution of the states of the system. Generally, the values of fractal dimension cluster around 1 when the system is relatively stable and decreases to lower values around 0.3 prior to the failure. The decrease in the value of the fractal dimension before a big earthquake is observed by several authors (Ouchi and Uekawa 1986; De Rubies et al. 1993). Hirata et al. (1987) showed that the change in fractal dimension with time is mainly dependent on the crustal conditions of the study region. On similar lines, fractal characteristics of the spatio-temporal sequence were studied to understand the seismic activity in one of the most active seismic zones, the “Circum-Pacific Seismic Belt” and its vicinity (Yin et al. 2019).

A recent study (Chen et al. 2020) reveals that the occurrence of strong (Magnitude >6.0) shallow earthquakes appears to follow a fractal pattern called Devil’s Staircase



Fractals and Chaos, Fig. 3 Power spectrum of (a) Density (b) Susceptibility (c) Electrical resistivity, and (d) Thermal conductivity. (Data from KTB borehole, Germany)

(also known as Cantor function), where clusters of earthquake events are separated by long but irregular intervals of no earthquakes. This finding differs from the previous pattern prediction by classical earthquake modeling that suggests the periodic or quasiperiodic occurrence of earthquakes based on cycles of build-up and release of tectonic stress. In nature, Devil's Staircase pattern is found in sedimentation sequences. Changes in uplift and erosion rates and reversals in Earth's magnetic field, among many other examples.

Fractal Porous Media

Discussion on the fractal theory for the solid Earth geophysics would be incomplete without a comprehensive discussion on the porous media. Studies have shown that porous media exhibit fractal characteristics, and since then, several attempts have been made to describe the fractal characters of a porous media. The properties like porosity and permeability have steadily attracted much attention in the past decades to describe the flow in fractal media (Vadapalli et al. 2014; Chattopadhyay and Vedanti 2016). The fractal characteristics are being used to model the fluid flow in complex fractured media. An analytical model of dimensionless permeability developed on the basis of the fractal theory is proposed by

Miao et al. (2019) for a porous media. In the latest development, it was demonstrated that the surface conductivity also obeys a power-law dependence on the specific surface area, called as surface Archie's law, and the power exponent is theoretically related to the porosity exponent entering Archie's law and to the fractal dimension of the pore network. Thus, a new permeability model is derived by combining the new surface Archie's law and the Kozeny–Carman model (Wang and Revil 2020).

Chaos

And now a bubble bursts, and now a world
- An essay on Man, Alexander Pope

Chaos theory is a branch of mathematics that deals with nonlinear dynamical systems. Nonlinear means that due to feedback or multiplicative effects between the components, the result becomes something more significant than just adding up the individual parts. Dynamical means the system changes over time based on its current state (Boeing 2016). The concept of chaos was proposed by Lorenz, who advocated that a small perturbation in the coupled system can

amplify to a great deal in some other part of the system. The chaotic systems are known to have a finite fractal dimension (Moharir 2000). The chaotic behavior of the system can be assessed by the quantitative estimate of Lyapunov exponent. The Lyapunov exponent measures the sensitivity of a dynamical system to small changes in the initial conditions and used for identifying chaos. For instance, if two orbits start close to each other in a chaotic system, then they will tend to move apart from each other exponentially. This process can be mathematically expressed as (Taylor 2005):

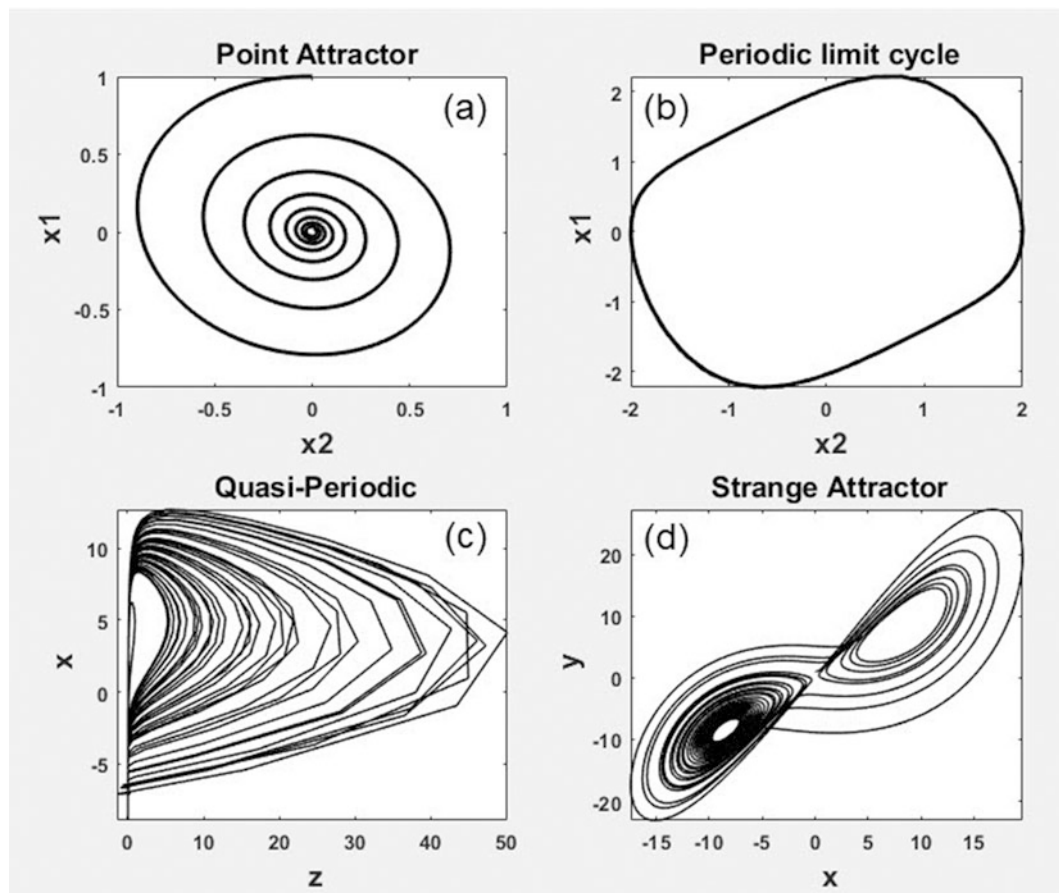
$$\frac{d}{d_0} = e^{\gamma(t-t_0)} \quad (3)$$

where d_0 is the initial displacement between a starting point and a nearby neighbor at initial time t_0 . The variable d is the displacement at time $t > t_0$. The γ is the Lyapunov exponent. It is evident from equation (3) that if $\gamma > 0$, then d/d_0 grows exponentially with time; if $\gamma < 0$, then d/d_0 tends to zero, and if $\gamma = 0$, then the change in displacement with time is non-exponential. The above equation (3) can be used to calculate

Lyapunov exponent, and a system is said to be chaotic if the Lyapunov exponent is positive (> 0).

Various evolutionary fluctuations exhibit the features of a dissipative dynamical system with possible chaos in phase space (Tiwari 2000). If all the trajectories of evolutionary processes converge toward a subset of phase space irrespective of their initial conditions, then the subset is called an “attractor” (Tsonis and Elsner 1988). There are mainly two types of attractors, nonstrange and strange attractors. There are three known types of nonstrange attractors, (a) fixed point attractor, (b) limit cycle attractor, and (c) torus attractor. These are nonfractal attractors with predictable behavior. Each of the three nonstrange attractors is associated with a specific type of orbit behavior (trajectory). The *fixed point* is associated with attractors whose orbits approach an equilibrium state; the *limit cycle* corresponds to periodic motion, and the *torus* corresponds to quasiperiodic motion orbit (Fig. 4).

Attractor provides a vital link between Chaos theory and Fractal geometry. The nature of the attractor is characterized by its fractal dimension. A “strange attractor” will have a noninteger (fractal) dimension. The term strange is used to



Fractals and Chaos, Fig. 4 Trajectory of the types of attractors: (a) point attractor obtained from a solution of a 2D differential equation (b) periodic limit cycle attractor obtained from Vande pol oscillator

(c) quasiperiodic attractor is a solution of Lorentz ODE using Rossler approach (d) Chaotic attractor is a solution of Lorentz ODE

those attractors showing extreme sensitivity to the initial condition, often they are chaotic, but it is not essential (they can be nonchaotic). In a chaotic regime, orbits on an attractor are nonperiodic; thus, the orbit behavior (or trajectory) of the chaotic attractors is a strange geometrical structure that is fractal in nature. The orbit behavior of a strange attractor is shown in Fig. 4d, which is nonperiodic and nonpredictable. The fractal dimension of a chaotic system indicates the number of ordinary differential equations needed to model the process, or in other words, it is indicative of the number of variables required for an almost complete description of the dynamic system (Tsonis and Elsner 1988).

Summary

Geophysical interpretation techniques have been reformulated after the findings from the analysis of physical properties recorded in different deep bore wells worldwide, which shows that the physical properties of the earth system follow fractal instead of random distribution (Dimri 2005). Research in solid earth geophysics employing fractals and chaos includes modeling of earth's magnetic and gravity fields, and it has been shown that the rocks have "scaling" stratification (Lovejoy et al. 2008; Lovejoy and Schertzer 2007). Another significant research in this field includes simulation of fGn series, which helps to improve the inversion results by way of providing a suitable initial model (Srivastava and Sen 2009, 2010). In atmospheric sciences, interesting research is done to model the atmospheric dynamics, which repeats scales after scales and behaves like a cascade (Lovejoy et al. 2009). There are several interesting applications of chaos in climate and rainfall modeling (Hosking 1984; Sivakumar 2000).

Bibliography

- Boeing G (2016) Visual analysis of nonlinear dynamical systems: chaos, fractals, self-similarity, and the limits of prediction. *Systems* 4(4):37. <https://doi.org/10.3390/systems4040037>
- Borodich FM (2000) Self-similar models and size effect of multiple fractures. *Fractals* 9(1):17–30
- Chattopadhyay PB, Vedanti N (2016) Fractal characters of porous media and flow analysis. In: Dimri VP (ed) *Fractal solutions for understanding complex systems in earth sciences*. Springer, Cham
- Chen Y, Liu M, Luo G (2020) Complex temporal patterns of large earthquakes: devil's staircases. *Bull Seismol Soc Am*. <https://doi.org/10.1785/0120190148>
- De Rubies V, Dimitriu P, Papa Dimtriu E, Tosi P (1993) Recurrent patterns in the spatial behavior of Italian seismicity revealed by the fractal approach. *Geophys Res Lett* 20:1911–1914
- Dimri V (1992) *Deconvolution and inverse theory*. Elsevier Science Publishers, Amsterdam/London/New York/Tokyo
- Dimri VP (2005) Fractals in geophysics and seismology: an introduction. In: Dimri V (ed) *Fractal behaviour of the earth system*. Springer, Berlin, pp 1–18
- Dimri VP, Srivastava RP (2019) Is geology scaling? *Earth Space Sci* 6. <https://doi.org/10.1029/2018EA000523>
- Dimri VP, Nimisha V, Chattopadhyay S (2005) Fractal analysis of aftershock sequence of Bhuj earthquake – a wavelet based approach. *Curr Sci* 88(10):1617–1620
- Dimri VP, Srivastava RP, Vedanti N (2012) *Fractal models in exploration geophysics: applications to hydrocarbon reservoirs*. Elsevier Science, London
- Grassberger P, Procaccia I (1983) Characteristics of strange attractors. *Phys Rev Lett* 50:346–349
- Hardy HH, Beier RA (1994) *Fractals in reservoir engineering*. World Scientific Publishing, Singapore
- Hewett TA (1986) Fractal distribution of reservoir heterogeneity and their influence on fluid transport. In: 61st annual technical conference, SPE, paper 15386
- Hirabayashi T, Ito K, Yoshi T (1992) Multifractal analysis of earthquakes. *Pure Appl Geophys* 138:591–610
- Hirata T, Imoto M (1991) Multifractal analysis of spatial distribution of micro earthquakes in the Kanto region. *Geophys J Int* 107:155–162
- Hirata T, Sato T, Ito K (1987) Fractal structure of spatial distribution of microfracturing in rock. *Geophys J R A Soc* 90:369–374
- Hosking JRM (1984) Modeling persistence in hydrological time series using fractional differencing. *Water Resour Res* 20(12):1898–1190
- Kagan YY, Knopoff L (1980) Spatial distribution of earthquakes, the two point correlation function. *Geophys J R A Soc* 62:303–320
- Lovejoy S, Schertzer D (2007) Scaling and multifractal fields in the solid earth and topography. *Nonlinear Process Geophys* 14:465–502
- Lovejoy S, Gaonac'h H, Schertzer D (2008) Anisotropic scaling models of rock density and the earth's surface gravity field. *Math Geosci* 40:533–573. <https://doi.org/10.1007/s11004-008-9171-7>
- Lovejoy S, Schertzer D, Allaire V, Bourgeois T, King S, Pinel J, Stolle J (2009) Atmospheric complexity or scale by scale simplicity? *Geophys Res Lett* 36:L01801. <https://doi.org/10.1029/2008GL035863>
- Malamud BD, Turcotte DL (1999) Self affine time series I: generation and analysis. In: Dmowska R, Saltzman B (eds) *Advances in geophysics: long-range persistence in geophysical time series*, vol 40. Academic, San Diego, pp 1–87
- Mandelbrot BB (1967) How long is the coastline of Britain? Statistical self-similarity and fractional dimension. *Science* 156:636–638
- Mandelbrot BB (1983) *The fractal geometry of nature*. W. H. Freeman, New York
- Mandelbrot BB, Van Ness JW (1968) Fractional Brownian motions, fractional noises, and applications. *SIAM Rev* 10:422–437
- Maus S, Dimri VP (1995) Potential field power spectrum inversion for scaling geology. *J Geophys Res* 100:12605–12616
- Miao T, Chen A, Xu Y, Cheng S, YU B (2019) A fractal permeability model for porous-fracture media with the transfer of fluids from porous matrix to fracture. *Fractals* 27(6):1950121
- Moharir PS (2000) Concepts similar to self-similarity in science. In: Dimri VP (ed) *Application of fractals in earth sciences*. A. A. Balkema, Rotterdam, pp 33–43
- Molz FJ, Roman GK (1993) A fractal-based stochastic interpolation scheme in subsurface hydrology. *Water Resour Res* 29(11):3769–3774
- Ouchi T, Uekawa T (1986) Statistical analysis of the spatial distribution of earthquakes before and after large earthquakes. *Phys Earth Planet Int* 44:211–225
- Painter S, Beresford G, Paterson L (1995) On the distribution of seismic reflection coefficients and seismic amplitudes. *Geophysics* 60(4):1187–1194
- Rieu M, Sposito G (1991) Fractal fragmentation, soil porosity, and soil water properties: I. theory. *Soil Sci Soc Am J* 55:1231–1238
- Sivakumar B (2000) Fractal analysis of rainfall observed in two different climatic regions. *Hydrol Sci J/des Sciences Hydrologiques* 45(5):727–738

- Srivastava HN (2005) Earth system modeling through chaos. In: Dimri VP (ed) *Fractal behaviour of the earth system*. Springer, Berlin, pp 179–203
- Srivastava RP, Sen MK (2009) Fractal based stochastic inversion of poststack seismic data using very fast simulated annealing. *J Geophys Eng* 6:412–425
- Srivastava RP, Sen MK (2010) Stochastic inversion of prestack seismic data using fractal based prior. *Geophysics* 75(3):R47–R59. <https://doi.org/10.1190/1.3379322>
- Tanaka M, Kato R, Kayama A (2001) Growth and linkage of cracks and formation of creep fracture pattern simulated by a multi-crack growth model. *Fractals* 9(2):223–230
- Taylor RL (2005) Attractors: nonstrange to chaotic. *Soc Ind Appl Math Undergrad Res Online* 4:72–80
- Tiwari RK (2000) Detecting chaos from geophysical time series. In: Dimri VP (ed) *Application of fractals in earth sciences*. A. A. Balkema, Rotterdam, pp 195–213
- Toverud T, Dimri VP, Ursin B (2001) Comparison of deconvolution methods for scaling reflectivity. *J Geophys* 22:117–123
- Tsonis AA, Elsner JB (1988) The weather attractor over very short time scale. *Nature* 333:545–547
- Turcotte DL (1997) *Fractals and chaos in geology and geophysics*. Cambridge University Press, Cambridge
- Vadapalli U, Srivastava RP, Vedanti N, Dimri VP (2014) Estimation of permeability of a sandstone reservoir by a fractal and Monte Carlo simulation approach: a case study. *Nonlinear Process Geophys* 21:9–18
- Voss RF (1985a) Random fractals: characterization and measurement. Paper presented at NATO advanced study institute on scaling properties of disordered media, Geilo, Norway
- Voss RF (1985b) Random fractal forgeries. In: *Fundamental algorithms for computer graphics*. Ilkley, Berlin: Springer-Verlag, pp. 805–835
- Wang H, Revil A (2020) Surface conduction model for fractal porous media. *Geophys Res Lett* 47:e2020GL087553. <https://doi.org/10.1029/2020GL087553>
- Yin L, Li X, Zheng W, Yin Z, Song L, Ge L, Zeng Q (2019) Fractal dimension analysis for seismicity spatial and temporal distribution in the circum-Pacific seismic belt. *J Earth Syst Sci* 128(1):22

Free Oscillations of the Earth

Sarva Jit Singh¹ and Sunita Rani²

¹Indian National Science Academy, New Delhi, India

²Department of Mathematics, Guru Jambheshwar University of Science and Technology, Hisar, India

Definition

Free oscillations of the Earth are standing waves along the surface of the Earth. These oscillations are typically observed after a large earthquake or an explosion because the Earth starts to vibrate as a whole as does a ringing bell when it is struck. The study of the Earth's free oscillations is based on the theory of vibrations of an elastic sphere. The problem increases in complexity as we proceed from an isotropic homogeneous sphere to models of the Earth which take into

account its anelasticity, anisotropy, radial and lateral heterogeneity, rotation, gravity, and prestress.

Introduction

The seismological data can be analyzed and interpreted in terms of three basic concepts. The first is based on the geometrical ray theory of optics. Most of our knowledge about the interior of the Earth has come from the application of the ray-theoretic methods to seismological data. Ray theory is useful for periods shorter than a minute. Travel times of seismic phases constitute an important component of the ray theory concept.

The second concept is that of surface waves. The study of the dispersion of surface waves helps us in determining the regional structure, e.g., low-velocity zones, oil-bearing structures, etc. The surface wave theory is useful in the period range 10–300 s.

The third concept is that of the normal-mode theory or free oscillations of the Earth. In here, the field is viewed as a sum of standing waves instead of travelling waves. The standing waves can be seen as peaks in the amplitude spectrum of a seismogram. Theoretical seismograms can be computed for spherically symmetric, nonrotating, elastic, isotropic (SNREI) Earth models by mode summation techniques at low frequencies. At high frequencies, the number of modes required to represent the field becomes so large that the mode summation technique is not practicable.

The problem of the oscillations of an elastic sphere is an old one. It attracted many classical physicists of the nineteenth century, including S.D. Poisson, Lord Kelvin, G.H. Darwin, and H. Lamb. In a classic paper, Lamb (1882) discussed the various modes of oscillations of a uniform elastic sphere and calculated the important roots of the frequency equation. He classified the oscillations of a sphere as of the “first class” and the “second class.” In an oscillation of the first class, the dilatation and the radial component of the displacement vanish everywhere so that there is no change of shape. In the case of the oscillations of the second class, the radial component of the vorticity vanishes. The oscillations of the first class are now known as toroidal or torsional oscillations and the oscillations of the second class as spheroidal or poloidal oscillations.

The Earth differs from a homogeneous elastic sphere in two important respects: self-gravitation and radial heterogeneity. Gravity does not affect the toroidal oscillations because these are divergence-free and have no radial component of the displacement, so that the density distribution remains unchanged. But gravity plays an important role in the case of the spheroidal oscillations. Bromwich (1898) investigated the effect of self-gravitation, simplifying his work by dealing only with an incompressible sphere. Love (1911)

showed that a self-gravitating sphere of the size and mass of the Earth and of rigidity that of steel has a free period of oscillation of almost exactly 1 h for the mode ${}_0S_2$ (the fundamental mode of the spheroidal oscillation of the second order). Obviously, if this mode were to be observed, instruments would be needed that had sensitivity at periods very much longer than the periods of seismic motion then being routinely recorded.

Shida (1925) designed an electromagnetic seismograph with pendulum period 180s and galvanometer period 1200s for observing Earth's free oscillations. However, the technology of that time was incapable of achieving the precision and sensitivity that were demanded. It was Hugo Benioff who first constructed an electromagnetic strainmeter which was capable of recording free vibrations for as long as 1 h. However, he had to wait for a big earthquake to make an appropriate use of the strainmeter developed by him, since the Earth will produce its deeper modes only when a large amount of kinetic energy is released by an earthquake or an explosion.

Benioff was able to record a ground motion of 57-min period for the Kamchatka earthquake of 4 November 1952 on his strainmeter. He (Benioff 1958) attributed it to the mode ${}_0S_2$ of the oscillations of the Earth. This was the first time that the natural oscillations of the whole Earth had been observed. It was but natural that several seismologists started investigating the problem of the free oscillations of the Earth theoretically and calculating the periods of the normal modes of real Earth models. Jobert (1956) applied Rayleigh's principle in calculating the period for the mode ${}_0T_2$ (the fundamental mode of the toroidal oscillation of the second order) as 43.54 min for the Bullen B model of the Earth. In the following year, Jobert (1957) applied the same method in obtaining the period for the mode ${}_0S_2$ as about 53 min. Pekeris and Jarosch (1958) applied variational methods in calculating the period of the mode ${}_0S_2$ as 53 min. Soon after, Alterman et al. (1959) integrated the equations of motion numerically and calculated the periods of various normal modes of the Earth. In particular, they showed that the period of the mode ${}_0S_2$ should be 53.7 min for Bullen B model of the Earth. A wide range of normal modes of the whole Earth was observed by several seismological groups in the USA for the great Chilean earthquake of 22 May 1960. The agreement between the observed periods and the periods calculated by Alterman et al. (1959) was excellent. This was the beginning of a new branch of seismology – the low-frequency seismology. Impetus to research in this new field was provided by several factors: (i) availability of instruments capable of recording long-period oscillations of the Earth; (ii) availability of high-speed electronic computers for Fourier analysis of the recorded data; and (iii) availability of computer codes for numerical solution of differential equations efficiently.

A network of low-noise accelerometers deployed all over the globe in the 1970s permitted routine detection of normal-mode spectral peaks following earthquakes of relatively modest magnitude.

Observations of the 1960 Chilean earthquake indicated that the mode ${}_0S_2$ did not have a single spectral peak but, instead, was composed of at least two lines with periods 53.1 and 54.7 min. Indeed, Backus and Gilbert (1961) and Pekeris et al. (1961) showed theoretically that for the mode ${}_0S_2$, five spectral lines exist if the rotation of the Earth is taken.

The work of Alterman et al. (1959) and their contemporaries did not include the effects of lateral heterogeneity, ellipticity, prestress, anisotropy, and anelasticity. Considerable progress has since been made in the direction of taking one or more of these features into account. Madariaga (1972) studied toroidal oscillations of a laterally heterogeneous Earth. Kanamori and Anderson (1977) and Woodhouse (1980) discussed the effect of the departure from perfect elasticity. Woodhouse and Dahlen (1978) investigated the effect of a general aspherical perturbation on the free oscillations of the Earth. Excellent reviews on the free oscillations of the Earth have been given by Stoneley (1961), Alterman et al. (1974), Lapwood and Usami (1981), Buland (1981), Dahlen and Tromp (1998), and others.

Toroidal Oscillations of a Uniform Elastic Sphere

On applying the method of separation of variables, the equation of motion for the toroidal oscillations of a homogeneous isotropic (uniform) elastic sphere of density ρ and Lamé's parameters λ and μ leads us to the following equation for the radial eigenfunction $W(r)$

$$\frac{d^2W}{dr^2} + \frac{2}{r} \frac{dW}{dr} + \left[\frac{\rho\omega^2}{\mu} - \frac{l(l+1)}{r^2} \right] W = 0, \quad (1)$$

where ω is the angular frequency and $l = 0, 1, 2, 3, \dots$ is a separation variable. Equation (1) can be solved in terms of the spherical Bessel functions. Solution of Eq. (1), which remains bounded as $r \rightarrow 0$, is of the form

$$W = A_0 j_l(\omega r/\beta), \quad (2)$$

where $j_l(\chi)$ is the spherical Bessel function of the first kind and $\beta = (\mu/\rho)^{1/2}$ is the shear wave velocity. The boundary condition that the surface $r = a$ of the sphere is traction-free yields the frequency equation:

$$\chi j_{l+1}(\chi) = (l-1)j_l(\chi), \quad (3)$$

where $\chi = \omega a/\beta$.

Since the frequency equation (3) is a transcendental equation, for each l , it possesses an infinite number of roots in χ . Let these roots be numbered in ascending order of magnitude and denoted by ${}_n\chi_l$ ($n = 0, 1, 2, \dots$). The corresponding frequencies and periods can be found through the relations.

$${}_n\omega_l = (\beta/a){}_n\chi_l, {}_nT_l = 2\pi/{}_n\omega_l. \quad (4)$$

Then ${}_n\omega_l$ are the eigenfrequencies, and ${}_nT_l$ are the eigenperiods. Further, $n = 0$ is associated with the fundamental mode, and $n = 1, 2$, etc. are associated with the first overtone, the second overtone, etc. For a given $l > 0$, there exists a family of n concentric nodal surfaces, i.e., interior surfaces at which the toroidal eigenfunction W vanishes. These nodal surfaces are given by

$$j_l\left(\frac{r}{a}{}_n\chi_l\right) = 0. \quad (5)$$

There are no toroidal oscillations associated with $l = 0$, since the corresponding displacement is identically zero. The value $l = 1$ yields a rigid rotation. The accompanying toroidal oscillations are known as rotatory vibrations. For $l = 1$, the frequency equation (3) becomes $j_2(\chi) = 0$, which may be expressed in the form

$$\frac{\tan \chi}{\chi} = \frac{3}{3 - \chi^2}. \quad (6)$$

The lowest roots of Eq. (6) are given, approximately, by

$$\chi = 5.763, 9.095, 12.322, 15.514 \dots \quad (7)$$

Equation (7) lists the first four overtones. For $l = 1$, the fundamental mode does not exist. For this mode ($l = 1, n = 0$), every part of the sphere shows rotation around some axis in one direction. Such a motion could not be excited without the application of an external torque and so would not be a free oscillation. Consequently, the frequency equation has no root corresponding to this mode (Lapwood and Usami 1981, p. 35).

For rotatory vibrations, the nodal spheres are given by

$$j_1\left(\frac{r}{a}{}_n\chi_l = 0\right)$$

which yields

$$\tan\left(\frac{r}{a}{}_n\chi_l\right) = \frac{r}{a}{}_n\chi_l. \quad (8)$$

Table 1 lists the dimensionless frequencies, periods, and nodal spheres for rotatory vibrations of a homogeneous sphere for the first four overtones.

From Eq. (3), the frequency equation for the toroidal oscillations of the second order ($l = 2$) is given by

$$\chi j_3(\chi) = j_2(\chi),$$

or, equivalently, by

$$\frac{\tan \chi}{\chi} = \frac{\chi^2 - 12}{5\chi^2 - 12}. \quad (9)$$

The corresponding dimensionless frequencies, periods, and nodal surfaces for the fundamental mode and the first four overtones are listed in Table 2.

Free Oscillations of the Earth, Table 1 Dimensionless frequency (χ), period (T in seconds), and nodal spheres (r_n) for the rotatory vibrations ($l = 1$) of a homogeneous sphere

n	χ	T^a	r_1/a	r_2/a	r_3/a	r_4/a
1	5.763	1113	0.780			
2	9.095	705	0.494	0.849		
3	12.322	521	0.365	0.627	0.885	
4	15.514	414	0.290	0.498	0.703	0.907

^aAssuming $a = 6371$ km, $\beta = 6.24$ km/s

Free Oscillations of the Earth, Table 2 Dimensionless frequency (χ), period (T in seconds), and nodal spheres (r_n) for the toroidal oscillations of the second order ($l = 2$) of a homogeneous sphere

n	χ	T^a	r_1/a	r_2/a	r_3/a	r_4/a
0	2.501	2565				
1	7.136	899	0.808			
2	10.514	610	0.548	0.865		
3	13.771	466	0.419	0.660	0.895	
4	16.983	378	0.339	0.535	0.725	0.914

^aAssuming $a = 6371$ km, $\beta = 6.24$ km/s

Spheroidal Oscillations of a Uniform Elastic Sphere

Spheroidal modes corresponding to $l = 0$ are known as radial modes because the associated particle motion is purely radial. The frequency equation for the radial vibrations is (Ben-Menahem and Singh 1981, p. 346)

$$\zeta \cot \zeta = 1 - \frac{1}{4} \left(\frac{\lambda + 2\mu}{\mu} \right) \zeta^2, \quad (10)$$

where

$$\zeta = \omega a / \alpha, \quad \alpha = [(\lambda + 2\mu)/\rho]^{1/2}. \quad (11)$$

It may be noted that the frequency equation for the spheroidal oscillations is a function of the ratio λ/μ . For the Poissonian case ($\lambda = \mu$), the lowest roots of Eq. (10) are given, approximately, by

$$\zeta/\pi = 0.816, 1.929, 2.936, 3.966, \dots \quad (12)$$

Assuming $a = 6371$ km, $\alpha = 9$ km/s, the corresponding periods (in seconds) are found to be

$$T = 1735, 734, 482, 357, \dots \quad (13)$$

The fundamental radial mode ${}_0S_0$ corresponds to an expansion and contraction of the whole sphere without its form changing. For the n th overtone ${}_nS_0$, there are n concentric nodal surfaces inside the sphere at which the displacement vanishes identically.

For $l = 1$, the fundamental mode ${}_0S_1$ does not exist, since it represents a rigid-body translation. Even though the geometric center of the sphere is displaced, its center of mass does not move (Lapwood and Usami 1981, p. 37). Thus the movement of the center does not contradict the physical requirement that the center of mass stays unmoved. The roots of the frequency equation representing the first four overtones ($l = 1, n = 1, 2, 3, 4$) are given, approximately, by

$$\chi = 3.424, 6.771, 7.444, 10.695, \quad (14)$$

assuming $\lambda = \mu$.

In the case of the spheroidal oscillations of the second order ($l = 2$), the oscillating sphere assumes alternatively the forms of a prolate and an oblate spheroid. Therefore, this mode of oscillation is known as the football mode.

Oscillations of an SNREI Earth Model

In the last two sections, we have discussed oscillations of a homogeneous, isotropic, elastic, non-gravitating, spherical

Earth model. This relieves the free oscillation problem of its mathematical intricacies while retaining its important features.

The equations of motion of a spherically symmetric, non-rotating, elastic, isotropic (SNREI) Earth model may be written in the form (Aki and Richards 1980; Ben-Menahem and Singh 1981)

$$\begin{aligned} \operatorname{div} \vec{T} + \rho [\operatorname{grad}(\Psi - gu_r) + \mathbf{e}_r g \operatorname{div} \mathbf{u}] \\ = \rho \partial^2 \mathbf{u} / \partial t^2, \end{aligned} \quad (15)$$

$$\nabla^2 \Psi = 4\pi G \operatorname{div}(\rho \mathbf{u}), \quad (16)$$

where

$\mathbf{u}$ = displacement vector,

$\vec{T} = \lambda \vec{I} \operatorname{div} \mathbf{u} + \mu(\nabla \mathbf{u} + \mathbf{u} \nabla)$ = stress tensor,

g = acceleration due to gravity,

G = gravitational constant,

Ψ = perturbation in the gravitational potential.

Equations (15) and (16) are to be solved subject to suitable boundary conditions. Thus, $\mathbf{u}$, $\mathbf{e}_r \cdot \vec{T}$, Ψ and $\partial \Psi / \partial r - 4\pi G \rho u_r$ must be continuous everywhere. However, at a boundary between a solid and a fluid or between two fluids, u_r rather than $\mathbf{u}$ must be continuous.

In the absence of external body force, the motion of the Earth consists of a superposition of free simple harmonic oscillations of the form

$$\mathbf{u}(\mathbf{r}, t) = \mathbf{u}(\mathbf{r}) \exp(i\omega t).$$

Therefore, the equation which governs the free oscillations of the Earth is obtained from Eq. (15) by making the substitution $\partial/\partial t \rightarrow i\omega$. This yields

$$\mathcal{L}\mathbf{u} + \rho\omega^2\mathbf{u} = 0, \quad (17)$$

where $\mathcal{L}$ is a self-adjoint linear operator. Equations (16) and (17) together with boundary conditions define an eigenvalue problem. Every model of the Earth will possess an infinite number of eigenfrequencies ω_i and a corresponding infinite number of eigenfunctions $\mathbf{u}_i(\mathbf{r})$.

As mentioned in the Introduction, the oscillations of the Earth are of two types – toroidal and spheroidal. At high frequencies, the toroidal oscillations correspond to the Love surface waves and SH body waves. Toroidal eigenfunctions are of the form

$$\mathbf{u}(\mathbf{r}) = W(r)[- \mathbf{e}_r \times \operatorname{grad} Y_{ml}], \quad (18)$$

where

$$Y_{ml} = P_l^m(\cos \theta) \exp(im\varphi)$$

is a surface spherical harmonic. For every value of l except zero, there is a fundamental toroidal mode ${}_0T_l$ and an infinite number of overtones ${}_nT_l$ ($n = 1, 2, \dots$). From Eq. (18) it is seen that for toroidal oscillations, $u_r = \text{div } \mathbf{u} = 0$.

At high frequencies, the spheroidal oscillations correspond to the Rayleigh surface waves and P - SV body waves. The spheroidal eigenfunctions are of the form

$$\mathbf{u}(r) = \mathbf{e}_r U(r) Y_{ml} + V(r) \text{grad } Y_{ml}. \quad (19)$$

For every value of l , there is a fundamental spheroidal mode ${}_0S_l$ and an infinite number of overtones ${}_nS_l$ ($n = 1, 2, \dots$). From Eq. (19), it can be verified that $(\text{curl } \mathbf{u})_r = 0$. On putting $\Psi(r) = P(r) Y_{ml}$, and using Eqs. (18) and (19), Eqs. (16) and (17) can be transformed into a set of four ordinary differential equations of the second order in U , V , W , and P . One of these is an equation in W alone and corresponds to the toroidal oscillations. The remaining three equations are coupled in U , V , and P and correspond to the spheroidal oscillations. These equations are awkward for numerical integration, because one needs to evaluate the derivatives of empirically determined quantities $\lambda(r)$, $\mu(r)$, and $\rho(r)$ in order to obtain the coefficients. This can be formally obviated by converting these equations into an equivalent set of eight linear differential equations of the first order which are free from the derivatives of λ , μ , and ρ . Out of these eight equations, two correspond to the toroidal oscillations and the remaining six to the spheroidal oscillations. We can thus express the equations governing the free oscillations of an SNREI Earth model in the form (Singh 1992)

$$\frac{dY}{dr} = A(r)Y, \quad (20)$$

where Y is an $(N \times 1)$ column matrix and $A(r)$ is an $(N \times N)$ square matrix. The number N depends upon the kind of oscillations and the nature of the medium (Table 3).

From Eqs. (17) and (18), we have the following radial equation for the toroidal oscillations

$$\mu \left(\frac{d^2 W}{dr^2} + \frac{2}{r} \frac{dW}{dr} \right) + \frac{d\mu}{dr} \left(\frac{dW}{dr} - \frac{W}{r} \right) + \left\{ \rho \omega^2 - \frac{l(l+1)}{r^2} \mu \right\} W(r) = 0. \quad (21)$$

This equation involves $d\mu/dr$. Numerical evaluation of the derivative of the empirically determined $\mu(r)$ can be avoided by defining $y_1 = W$, $y_2 = \mu \left(\frac{d}{dr} - \frac{1}{r} \right) y_1$.

The differential equation of the second order (21) is equivalent to two differential equations of the first order:

$$\begin{aligned} \frac{dy_1}{dr} &= \frac{y_1}{r} + \frac{y_2}{\mu}, \\ \frac{dy_2}{dr} &= \left[(l^2 + l - 2) \frac{\mu}{r^2} - \rho \omega^2 \right] y_1 - \frac{3}{r} y_2. \end{aligned} \quad (22)$$

For spheroidal oscillations, the set of six linear differential equations of the first order is obtained by defining

$$\begin{aligned} y_1 &= U, \quad y_2 = (\lambda + 2\mu) \frac{dU}{dr} + \frac{2\lambda}{r} U - \lambda \frac{l(l+1)}{r} V, \\ y_3 &= V, \quad y_4 = \mu \left[\frac{1}{r} (U - V) + \frac{dV}{dr} \right], \\ y_5 &= P, \quad y_6 = \frac{dP}{dr} - 4\pi G \rho U. \end{aligned} \quad (23)$$

In the case of the liquid core, $\mu = y_4 = 0$ and

$$y_3 = \frac{1}{r\omega^2} \left(g y_1 - \frac{y_2}{\rho} - y_5 \right). \quad (24)$$

Explicit expressions for the elements of the matrix $A(r)$ have been given by Singh (1992).

Equation (20) subject to an appropriate set of homogeneous boundary conditions can be integrated by the Runge-Kutta method. For a given SNREI Earth mode $\lambda(r)$, $\mu(r)$, and $\rho(r)$ and given l , there is a discrete set of values of ω for which the boundary conditions are satisfied. We denote this set by ω_i (eigenfrequencies) and the corresponding displacements by $\mathbf{u}_i$. The index i signifies the normal mode. Each i stands for a triplet (l, m, n) of numbers; l being the colatitudinal mode number, also known as the angular order ($l \geq 0$), m the azimuthal mode number ($-l \leq m \leq l$), and n the

Free Oscillations of the Earth, Table 3 Value of the number N indicating the order of the differential system

Oscillations	Medium	Gravity	N
Toroidal	Solid	No effect	2
Spheroidal	Solid	Yes	6
Spheroidal	Liquid	Yes	4
Spheroidal	Solid	No	4
Spheroidal	Liquid	No	2
Radial	Solid	Yes	2

radial mode number ($n \geq 0$). The numbers l, m, n describe the manner in which the displacement field depends upon the colatitude, the azimuth, and the radial distance, respectively. However, because of the spherical symmetry, the eigenfrequencies are degenerate in m (i.e., ω_l are independent of m). This degeneracy is removed when the rotation of the Earth, its ellipticity, or lateral heterogeneity is taken into account.

The spherical harmonic

$$P_l^m(\cos \theta)(\cos m\phi, \sin m\phi)$$

from which the eigenfunctions are derived has $(l - |m|)$ nodal latitude lines and $2|m|$ nodal longitude lines on the surface of the Earth. The value of m depends upon the source exciting the oscillations.

Table 4 compares the eigenfrequencies computed for PREM (Preliminary reference Earth model (Dziewonski and Anderson 1981)) with mean observed frequencies. The computed eigenfrequencies are obtained from the website (<http://stuplots.geophys.uni-stuttgart.de/~plots/Modes/modes.html>). In this website, the eigenfrequencies and eigenfunctions can be computed for PREM and three other Earth models for $0 \leq n < 50$, $0 \leq l < 200$, using the FORTRAN program MINOS written by Drs. Freeman Gilbert and Guy Masters. The mean observed frequencies are taken from the website (<http://igppweb.ucsd.edu/~gabi/rem.dir/surface/rem.surf.html>). The agreement between the computed and observed values is excellent.

The eigenfrequencies of the toroidal oscillations computed for PREM are shown in Fig. 1 as functions of the angular order l for the fundamental mode $n = 0$ and the first five

overtones $n = 1-5$. The corresponding eigenfrequencies of the spheroidal oscillations are shown in Fig. 2.

Effect of the Rotation of the Earth

As mentioned earlier, the eigenfrequencies of a spherically symmetric, nonrotating, elastic, isotropic Earth model are degenerate in the azimuthal mode number m . The introduction of the rotation removes this degeneracy. The effect of the diurnal rotation of the Earth can be calculated by carrying out a first-order perturbation calculation. Let Ω denote the uniform angular velocity of the Earth about its center. Assume that the observer is referred to a non-inertial frame, which, for all times, maintains a state of uniform rotation with angular velocity Ω . Let (x_1, x_2, x_3) be a Cartesian coordinate system in this uniformly rotating frame of reference, let the origin of the system coincide with the center of the Earth, and let e_3 be aligned along the axis of rotation, so that $\Omega = \Omega e_3$.

The rotation of the Earth introduces two body forces. One of them is the centrifugal force $r \sin\theta \Omega^2$ perpendicular to and away from the axis of rotation. The other force introduced by the rotation is the Coriolis force, which, per unit mass, is equal to $2\Omega(\partial \mathbf{u}/\partial t) \times e_3$.

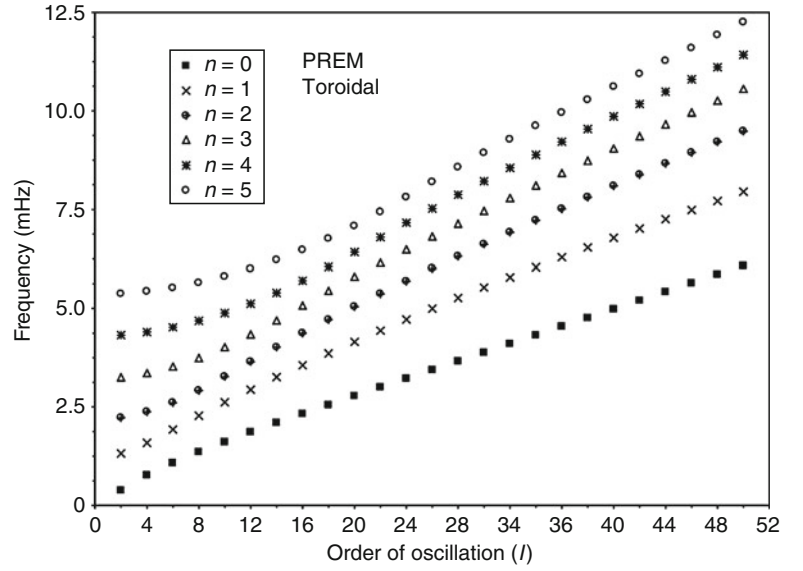
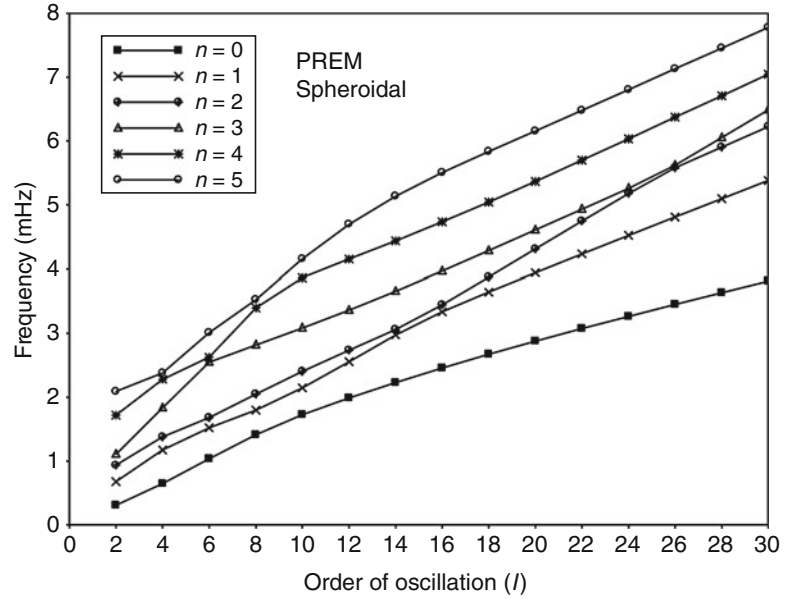
Assuming $\Omega/\omega_l \ll 1$ (the highest value of $\Omega/\omega_l \simeq 1/27$) and carrying out a first-order perturbation calculation, it can be shown that (Ben-Menahem and Singh 1981, p. 390)

$${}_n\omega_l^m = {}_n\omega_l + m\tau\Omega, \quad -l \leq m \leq l, \quad (25)$$

where

Free Oscillations of the Earth, Table 4 Fundamental mode ($n = 0$) eigenfrequencies ($f = 1/T = \omega/2\pi$), in mHz (10^{-3} cycles per second), for PREM (Preliminary reference Earth model). The corresponding mean observed frequencies are also shown

l	Toroidal		Spheroidal	
	Computed frequencies	Observed frequencies	Computed frequencies	Observed frequencies
2	0.37917	0.37970	0.30928	0.30945
3	0.58616	0.58630	0.46856	0.46850
4	0.76566	0.76593	0.64707	0.64682
5	0.92824	0.92845	0.84042	0.84002
6	1.07883	1.07913	1.03822	1.03755
7	1.22070	1.22115	1.23179	1.23105
8	1.35611	1.35660	1.41351	1.41285
9	1.48661	1.48707	1.57828	1.57755
10	1.61326	1.61367	1.72647	1.72560
15	2.21034	2.21116	2.34638	2.34495
20	2.77667	2.77759	2.87837	2.87795
30	3.88175	3.88290	3.81552	3.81480
40	4.97942	4.98070	4.71010	4.70840
50	6.07699	6.07900	5.61073	5.61030
100	11.56366	11.58500	10.24285	10.28900
150	17.01277	17.03500	14.99560	15.11000

Free Oscillations of the Earth,**Fig. 1** Eigenfrequency $f = 1/T = \omega/2\pi$ of the toroidal oscillations for PREM**Free Oscillations of the Earth,****Fig. 2** Eigenfrequency of the spheroidal oscillations for PREM. Only points corresponding to even values of $l = 2, 4, \dots, 30$ have been plotted. For the sake of clarity, points representing the eigenfrequencies for even values of l have been joined by continuous curves. However, the intervening curves do not represent any physical reality

${}_n\omega_l$ = eigenfrequencies of a nonrotating Earth model,
 ${}_n\omega_l^m$ = eigenfrequencies of the corresponding rotating Earth model.

The splitting parameter

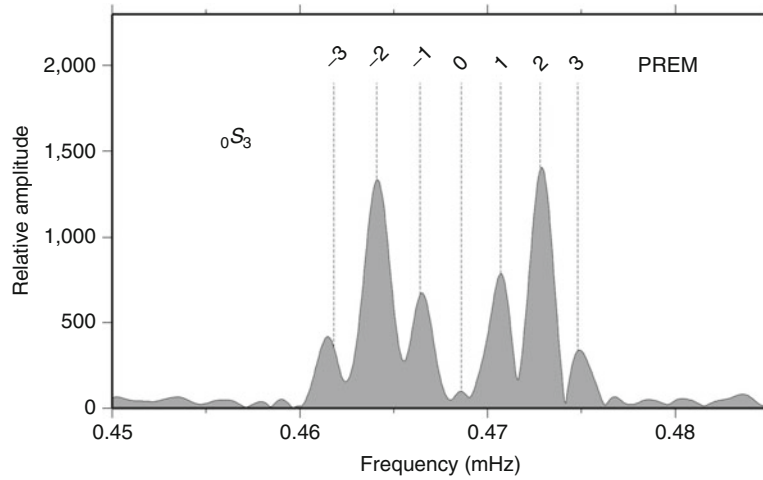
$$\tau = \frac{1}{l(l+1)} \quad (26)$$

for the toroidal oscillations, and

$$\tau = \frac{\int_0^a (2U + V)V\rho r^2 dr}{\int_0^a [U^2 + l(l+1)V^2]\rho r^2 dr} \quad (27)$$

for the spheroidal oscillations, where a is the radius of the Earth; U and V are defined in Eq. (19).

Therefore, for both toroidal and spheroidal oscillations, the degenerate eigenfrequency ${}_n\omega_l$ is resolved by a slow rotation of the Earth into $(2l+1)$ frequencies ${}_n\omega_l^m$ ($-l \leq m \leq l$). The set of $(2l+1)$ spectral lines for a given (l, n) is called a mode multiplet, and each member of this set is called a singlet. For the toroidal oscillations, the splitting parameter $[l(l+1)]^{-1}$ does not depend upon the Earth model. However, in the case of the spheroidal oscillations, τ depends upon the Earth model. The splitting of the terrestrial spectral lines is the elastodynamic analogy of the splitting of atomic spectral lines by a magnetic field discovered by Zeeman in 1896.



Free Oscillations of the Earth, Fig. 3 Fourier spectrum of the 27 February 2010 Maule (Chile) earthquake ($M_w = 8.8$) record provided by the superconducting gravimeter at the Black Forest Observatory (BFO). The rotation removes the degeneracy and splits the ${}_0S_3$ multiplet into seven singlets. The vertical dashed lines indicate the theoretically predicted

frequencies of the singlets for the spherically symmetric Earth model PREM, including the effect of the rotation of the Earth and hydrostatic flattening. (Figure made available by Dr. Rudolf Widmer-Schmidrig of BFO who holds the copyright)

Figure 3 shows the Zeeman splitting of the fundamental spheroidal mode of the third-order ${}_0S_3$ observed in the Fourier spectrum of the gravimeter record of the 27 February 2010 Maule (Chile) earthquake. The time series is 400 h long, and the effect of local barometric fluctuations has been removed. The rotation of the Earth removes the degeneracy of the multiplet; six out of seven singlets can be clearly identified. The singlet corresponding to the azimuthal mode number $m = 0$ is missing because either the source or the receiver is located on a nodal line of the mode.

The ellipticity of the Earth also removes the degeneracy giving (Dahlen 1968)

$${}_n\omega_l^m = {}_n\omega_l [1 + e(b + m^2c)] \quad (28)$$

to first order in ellipticity e , where $b = b(n, l)$, $c = c(n, l)$. Thus the ellipticity splits the degenerate eigenfrequency ${}_n\omega_l$ into $(l + 1)$ lines ${}_n\omega_l^m$ ($0 \leq m \leq l$).

Other departures from symmetry, e.g., lateral inhomogeneity, anisotropy, and initial stress, have the same effect of removing the degeneracy.

Jeans' Formula

The complete seismic field induced by a point source in an SNREI Earth model can be expressed as an infinite sum of normal modes. However, we know from seismogram analysis that most of the recorded Earth motion can be explained in terms of propagating waves. Therefore, there must exist a relationship between these two seemingly different aspects of seismic wave motion.

Consider a general term in the normal-mode solution expressed by Eqs. (18) and (19). The factor of this term which depends upon the time and the colatitude is derivable from

$$P_l^m(\cos \theta) \exp(i\omega t), \quad (29)$$

where $\omega = {}_n\omega_l$ is an eigenfrequency. Replacing the Legendre function by its asymptotic approximation for large values of l , this factor becomes

$$(-l)^m \left(\frac{1}{2\pi l \sin \theta} \right)^{1/2} \left[\exp \left\{ i \left[\omega t - (l + 1/2)\theta + \frac{\pi}{4} - \frac{m\pi}{2} \right] \right\} + \exp \left\{ i \left[\omega t + (l + 1/2)\theta - \frac{\pi}{4} + \frac{m\pi}{2} \right] \right\} \right]. \quad (30)$$

The first term describes a wave motion of frequency ω diverging from the pole $\theta = 0$, whereas the second term describes a wave motion diverging from the antipode $\theta = \pi$.

In cylindrical coordinates, a wave diverging from the axis may be expressed in the form

$$\exp[i(\omega t - k\Delta)], \quad (31)$$

where k is the wave number. Putting $\Delta = a\theta$ and comparing Eqs. (30) and (31), we get Jeans' formula (Jeans 1923):

$$l + \frac{1}{2} = ka \quad (l \gg 1).$$

This formula tells us that, if l is large, every mode of oscillation can be interpreted as a travelling wave whose parameters are functions of l and n and are given by

$$T = \frac{2\pi}{\omega}, c = \frac{a\omega}{l + \frac{1}{2}}, \Lambda = \frac{2\pi a}{l + \frac{1}{2}}, \quad (32)$$

where T is the period of the oscillation, c is the phase velocity, and Λ is the associated wavelength. These relations are found to yield a good approximation for $l \geq 7$. The connection between the free oscillations and travelling waves (mode-ray duality) has been studied by Jeans (1923), Ben-Menahem (1964), Brune (1964), and others.

Jeans formula can be used to determine the eigenfrequencies from the observed dispersion data, and then the normal-mode formalism can be used to find the elastic properties of the Earth.

Continuous Free Oscillations of the Earth

Free oscillations of the Earth are typically observed after a large earthquake. Continuous free oscillations of the Earth (primarily fundamental spheroidal modes between 2 and 7 mHz) were discovered in 1998 (Suda et al. 1998; Tanimoto 2001). Observational evidence points to the atmosphere (e.g., Kobayashi and Nishida 1998) and oceans (e.g., Rhie and Romanowicz 2004) as possible sources of excitation of these oscillations. Stochastic excitation of normal modes by surface pressure variations, caused by atmospheric and/or oceanic turbulence, may explain the overall amplitudes of their observed modes. Since the discovery of the continuous free oscillations of the Earth, also known as seismic hum, in 1998, a large number of geophysicists got interested in their study (see, e.g., Nishida et al. 2000, 2009; Fukao et al. 2002; Tanimoto 2005; Webb 2007). In a recent survey, Nishida (2013) discussed at length the observation of the continuous free oscillations of the Earth and their source of excitation. According to Nishida (2013), Earth's background free oscillations from 2 to 20 mHz are observed at globally distributed stations, excited modes of these oscillations are almost exclusively fundamental spheroidal and toroidal, and amplitudes of the toroidal modes are larger than those of the spheroidal ones.

Conclusions

We have described how to compute the eigenperiods of an SNREI Earth model. Modifications due to lateral inhomogeneity, anisotropy, and initial stress can be found by perturbation methods (Woodhouse and Dahlen 1978). The eigenperiods of different modes of vibration are increased by the anelasticity of the Earth. In addition, anelasticity results

in the decay of the amplitudes of the free oscillations with time. Necessary allowance for anelasticity can be made while constructing Earth models using normal-mode theory (Dziewonski and Anderson 1981).

The study of the free oscillations of the Earth has played a leading role in the development of geophysical inverse theory in the late 1960s (Backus and Gilbert 1967, 1968, 1970) and 1970s. This theory enables us optimal extraction of information regarding the internal constitution of the Earth from a finite set of gross Earth data (Wiggins 1972). Comprehensive exposition of the inverse theory has been given, among others, by Tarantola (1987) and Parker (1994). Dahlen and Tromp (1998) gave a comprehensive treatment of the problem of computing synthetic seismograms for a realistic three-dimensional model of the Earth, using normal-mode summation technique. In their recent review, Woodhouse and Deuss (2015) gave a detailed account of the theory and observations of the free oscillations of the Earth.

Major discoveries based on the observations of the Earth's free oscillations are (in chronological order):

1. Finite rigidity of the inner core (Dziewonski and Gilbert 1971)
2. 1D density distribution inside the Earth (Gilbert and Dziewonski 1975; Masters and Gubbins 2003)
3. Construction of PREM (Preliminary reference Earth model) by Dziewonski and Anderson (1981)
4. Construction of the first global 3D upper mantle velocity model (Masters et al. 1982)
5. Construction of 1D attenuation model (e.g., Masters and Gilbert 1983)
6. Tight upper bound on the maximum differential super rotation of the inner core (e.g., Laske and Masters 1999)
7. Regional variation of inner core anisotropy (e.g., Deuss et al. 2010)

Cross-References

- [Body Waves](#)
- [Inverse Theory, Monte Carlo Method](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Wave Propagation in Real Media: Numerical Modeling Approaches](#)
- [Seismic, Waveform Modeling and Tomography](#)
- [Surface Waves](#)

Acknowledgments The authors are thankful to the reviewer, Dr. Rudolf Widmer-Schmidrig of the Black Forest Observatory, for his useful suggestions and comments and for providing the Fourier spectrum of the 2010 Maule (Chile) earthquake gravimeter record shown in Fig. 3.

Bibliography

- Aki K, Richards PG (1980) Quantitative seismology, vol 1. W.-H. Freeman, San Francisco
- Alterman Z, Jarosch H, Pekeris CL (1959) Oscillations of the Earth. *Proc R Soc Lond A Math Phys Eng Sci* 252(1268):80–95
- Alterman Z, Eyal Y, Merzer AM (1974) On free oscillations of the Earth. *Surv Geophys* 1(4):409–428
- Backus GE, Gilbert F (1961) The rotational splitting of the free oscillations of the Earth. *Proc Natl Acad Sci U S A* 47:362–371
- Backus GE, Gilbert F (1967) Numerical applications of a formalism for geophysical inverse problems. *Geophys J R Astron Soc* 13(1–3):247–276
- Backus GE, Gilbert F (1968) The resolving power of gross Earth data. *Geophys J R Astron Soc* 16(2):169–205
- Backus GE, Gilbert F (1970) Uniqueness in the inversion of gross Earth data. *Philos Trans R Soc Lond A Math Phys Eng Sci* 266(1173):123–192
- Benioff H (1958) Long waves observed in the Kamchatka earthquake of November 4, 1952. *J Geophys Res* 63(3):589–593
- Ben-Menahem A (1964) Mode-ray duality. *Bull Seismol Soc Am* 54:1315–1321
- Ben-Menahem A, Singh SJ (1981) Seismic waves and sources. Springer, New York. Reprinted in 2000 by Dover Publications, New York
- Bromwich TJ (1898) On the influence of gravity on elastic waves, and, in particular, on the vibrations of an elastic globe. *Proc Lond Math Soc* 30(1):98–120
- Brune JN (1964) Travel times, body waves and normal modes of the Earth. *Bull Seismol Soc Am* 54(6):2099–2128
- Buland R (1981) Free oscillations of the Earth. *Annu Rev Earth Planet Sci* 9:385–413
- Dahlen FA (1968) The normal modes of a rotating, elliptical Earth. *Geophys J R Astron Soc* 16:329–367
- Dahlen FA, Tromp J (1998) Theoretical global seismology. Princeton University Press, Princeton
- Deuss A, Irving JCE, Woodhouse JH (2010) Regional variation of inner core anisotropy from seismic mode observations. *Science* 328:1018–1020
- Dziewonski AM, Anderson DL (1981) Preliminary reference Earth model. *Phys Earth Planet Inter* 25:297–356
- Dziewonski AM, Gilbert F (1971) Solidity of the inner core of the Earth inferred from normal mode observations. *Nature* 234:465–466
- Fukao Y, Nishida K, Suda N, Nawa K, Kobayashi N (2002) A theory of the Earth's background free oscillations. *J Geophys Res* 107(B9):11–11–10
- Gilbert F, Dziewonski AM (1975) An application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra. *Philos Trans R Soc Lond A Math Phys Eng Sci* 278:187–269
- Jeans JH (1923) The propagation of earthquake waves. *Proc R Soc Lond A Math Phys Eng Sci* 102(718):554–574
- Jobert N (1956) Evaluation de la période d'oscillation d'une sphère élastique hétérogène, par application du principe de Rayleigh. *C R Math* 243:1230–1232
- Jobert N (1957) Sur la période propre des oscillations sphéroïdales de la terre. *C R Math* 245:921–922
- Kanamori H, Anderson DL (1977) Importance of physical dispersion in surface waves and free oscillation problems-review. *Rev Geophys Space Phys* 15(1):105–112
- Kobayashi N, Nishida K (1998) Continuous excitation of planetary free oscillations by atmospheric disturbances. *Nature* 395(6700):357–360
- Lamb H (1882) On the vibration of an elastic sphere. *Proc Lond Math Soc* 13(1):189–212
- Lapwood ER, Usami T (1981) Free oscillations of the Earth. Cambridge University Press, Cambridge
- Laske G, Masters G (1999) Limits on differential rotation of the inner core from an analysis of the Earth's free oscillations. *Nature* 402(6757):66–69
- Love AEH (1911) Some problems of geodynamics. Cambridge University Press, Cambridge. Reprinted in 1967 by Dover Publications, New York
- Madariaga R (1972) Toroidal free oscillations of the laterally heterogeneous Earth. *Geophys J R Astron Soc* 27:81–100
- Masters G, Gilbert F (1983) Attenuation in the Earth at low frequencies. *Philos Trans R Soc Lond A Math Phys Eng Sci* 308:479–522
- Masters G, Gubbins D (2003) On the resolution of density within the Earth. *Phys Earth Planet Inter* 140:159–167
- Masters G, Jordan TH, Silver PG, Gilbert F (1982) Aspherical Earth structure from fundamental spheroidal mode data. *Nature* 298:609–613
- Nishida K (2013) Earth's background free oscillations. *Annu Rev Earth Planet Sci* 41(1):719–740
- Nishida K, Kobayashi N, Fukao Y (2000) Resonant oscillations between the solid Earth and the atmosphere. *Science* 287(5461):2244–2246
- Nishida K, Montagner JP, Kawakatsu H (2009) Global surface wave tomography using seismic hum. *Science* 326(5949):112
- Parker RL (1994) Geophysical inverse theory. Princeton University Press, Princeton
- Pekeris CL, Jarosch H (1958) The free oscillations of the Earth. In: Contributions in geophysics in honor of Beno Gutenberg. Pergamon, New York, pp 171–192
- Pekeris CL, Alterman Z, Jarosch H (1961) Rotational multiplets in the spectrum of the Earth. *Phys Rev* 122:1692–1700
- Rhie J, Romanowicz B (2004) Excitation of Earth's continuous free oscillations by atmosphere-ocean-seafloor coupling. *Nature* 431:552–556
- Shida T (1925) On the possibility of observing the free vibrations of the Earth. *Nagaoka Anniversary vol* 109–120
- Singh SJ (1992) Free oscillations of the Earth. *Curr Sci* 62:227–232
- Stoneley R (1961) The oscillations of the Earth. In: Physics and chemistry of the Earth, vol 4. Pergamon, New York, pp 239–250
- Suda N, Nawa K, Fukao Y (1998) Earth's background free oscillations. *Science* 279(5359):2089–2091
- Tanimoto T (2001) Continuous free oscillations: atmosphere-solid Earth coupling. *Annu Rev Earth Planet Sci* 29:563–584
- Tanimoto T (2005) The oceanic excitation hypothesis for the continuous oscillations of the Earth. *Geophys J Int* 160(1):276–288
- Tarantola A (1987) Inverse problem theory. Elsevier Sciences, Amsterdam
- Webb SC (2007) The Earth's 'hum' is driven by ocean waves over the continental shelves. *Nature* 445(7129):754–756
- Wiggins RA (1972) The general linear inverse problem: implications of surface waves and free oscillations for Earth structure. *Rev Geophys Space Phys* 10(1):251–285
- Woodhouse JH (1980) The coupling and attenuation of nearly resonant multiplets in the Earth's free oscillation spectrum. *Geophys J R Astron Soc* 61:261–283
- Woodhouse JH, Dahlen FA (1978) The effect of a general aspherical perturbation on the free oscillations of the Earth. *Geophys J R Astron Soc* 53(2):335–354
- Woodhouse JH, Deuss A (2015) Theory and observations – Earth's free oscillations. In: Treatise on geophysics, vol 1. Elsevier, Oxford, pp 79–115

Geodesy, Figure of the Earth

Kusumita Arora

National Geophysical Research Institute, Council of Scientific and Industrial Research (CSIR), Hyderabad, India

Definition

Geodesy is the discipline that deals with the measurement and representation of the Earth, its gravity field, and geodynamic phenomena (polar motion, Earth tides, and crustal motion) in three-dimensional time varying space.

“Figure of the Earth” in geodesy deals with the representation of the external geometry of the whole planet often also called physical geodesy, without going into the causal mechanisms. On the other hand, engineering geodesy is concerned with measuring specific parts or regions of Earth, which contribute to understanding of geophysical measurements and in turn to geodynamic processes. This article provides a brief description of the former.

Background

The shape of the Earth has always been speculated upon – historical accounts of sailors record that they feared falling off the edge of the Earth if they sailed too far. Early ideas about the figure of the Earth held the Earth to be flat and the heavens a physical dome spanning over it. Two early arguments for a spherical Earth were that lunar eclipses were seen as circular shadows which could only be caused by a spherical Earth, and that the pole star is seen lower in the sky as one travels south.

The flat disc advocated by Homer (in the ninth century B.C.) later gave way to the spherical body postulated by Pythagoras (in the sixth century B.C.) – an idea supported one hundred years later by Aristotle (384–322 B.C.). According to old texts, the first geodetic experiment took place around

250 B.C. by Eratosthenes (about 284–192 B.C.) in Egypt. Eratosthenes noted that the Sun on the summer solstice day at Syene (close today to Assouan) was at the zenith while its zenithal distance at Alexandria (approximately on the same meridian) was recorded as $7^{\circ}12'$. By knowing the geometrical distance between Syene and Alexandria thanks to a cadastre performed by Ptolemy Soter (about 360–283 B.C.), he deduced, the length of the terrestrial circumference as 40,500 km. Later Hipparchus (in the second century B.C.) recommended the use of latitudes and longitudes for cartography (in agreement with its Greek etymology). The great Indian mathematician Aryabhatta (476–550 A.D.) described the Earth as being spherical and rotating on its axis. He estimated the circumference of the Earth, accurate to 1% of today’s measurement. The medieval Persian geodesist Abu Rayhan Biruni (973–1048 A.D.) is sometimes regarded as the “father of geodesy” for his significant contributions to the field. He solved complex geodesic equations based on spherical trigonometry in order to accurately compute the Earth’s circumference and radius, which are very close to modern values.

The shape of the Earth is the combined result of the forces due to the various motions, rotation being the strongest under the influence of gravity including its spatiotemporal variations as well as horizontal tectonic forces (more in ► “[Geodesy, Physical](#)” and ► “[Gravity, Global Models](#)” in this volume).

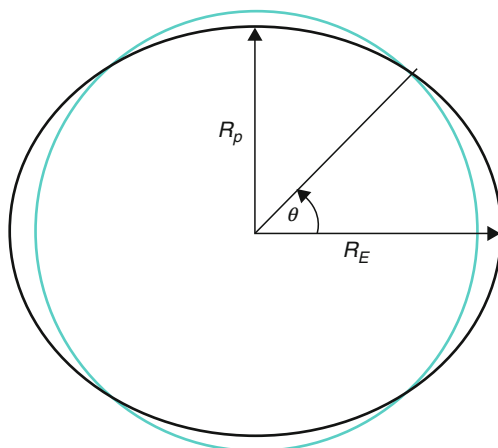
Mathematical Figures of the Earth

The actual shape of the Earth consisting of its topographic surface with its variety of land forms and water areas is not suitable for mathematical representations because the formulas which would be required to take these features into account would necessitate a prohibitive amount of computations. The Pythagorean concept of a spherical Earth offers a simple surface, which is mathematically easy to deal with. Many astronomical and navigational computations use it as a surface representing the Earth. While the sphere is a close

approximation of the true figure of the Earth and satisfactory for many purposes, to the geodesists interested in the measurement of long distances – spanning continents and oceans – a more exact figure is necessary. In the mid- to late- twentieth century, research across the geosciences contributed to drastic improvements in the accuracy in the description of the Figure of the Earth.

The concept of a perfectly spherical Earth was eventually refined by Sir Isaac Newton based on his theory of gravity, deriving that the Earth would not be perfectly spherical but slightly squashed at the poles and bulging at the Equator due to the forces acting upon it as it spins. Mathematically this shape is described as an oblate spheroid or as an ellipsoid of revolution (Fig. 1), an oval that revolves about its shortest dimension. This shape is most commonly used for measurements over long distances, across continents or oceans. Closer approximations utilize spherical harmonics for whole Earth shape or local approximations in terms of local reference ellipsoids. An equipotential ellipsoid is an ellipsoid that is defined to be an equipotential surface of normal gravity potential, determined by means of the ellipsoidal surface, the enclosed mass, and the angular velocity (independently of the internal density distribution). The equipotential ellipsoid furnishes a simple, consistent, and uniform reference system for all purposes of geodesy and does not include an atmosphere (more in ► “Geodesy, Physical”, ► “Gravity, Global Models”, and ► “Gravity Method, Satellite” in this volume).

Another method estimates the Earth’s surface using mean sea level of the ocean with all continents removed – this surface is equivalent to an equipotential surface and is called the Geoid. Due to variations in the Earth’s mass distribution (oceans and land), the Geoid has an irregular shape that is described as “undulating.” The ellipsoid of revolution mentioned above is a mathematical approximation of the Geoid. The Geoid, unlike the ellipsoid, is too complicated so far to



Geodesy, Figure of the Earth, Fig. 1 An oblate spheroid (black) superimposed on sphere (cyan), R_p is the polar radius 6357 km, and R_E is the equatorial radius 6378 km. θ is the latitudinal angle

serve as the computational surface on which to solve geometrical problems like point position. The separations between the two, referred to as Geoid undulations, Geoid heights, or Geoid separations, are irregular. The Geoid undulations over a spherical Earth are shown in Fig. 2, where the red colored areas indicate positive deviations of the Geoids from the ellipsoid and the blues indicate negative deviations of the Geoids from the ellipsoid. Many of these deviations have obvious spatial correlations with surface expressions of substantial topographic features, which in turn indicate significant geodynamic activity. However, there are also undulations, which do not bear such relations; their causal mechanisms remain to be established (more in ► “Geoid Undulation, Interpretation” in this volume). With the availability of decades of highly accurate, consistent, uniform data of gravitational measurements from satellites along with the data from an increasing number of GNSS satellites, the accuracy of geoid computations has reached new heights (more in ► “GPS, Tectonic Geodesy” in this volume).

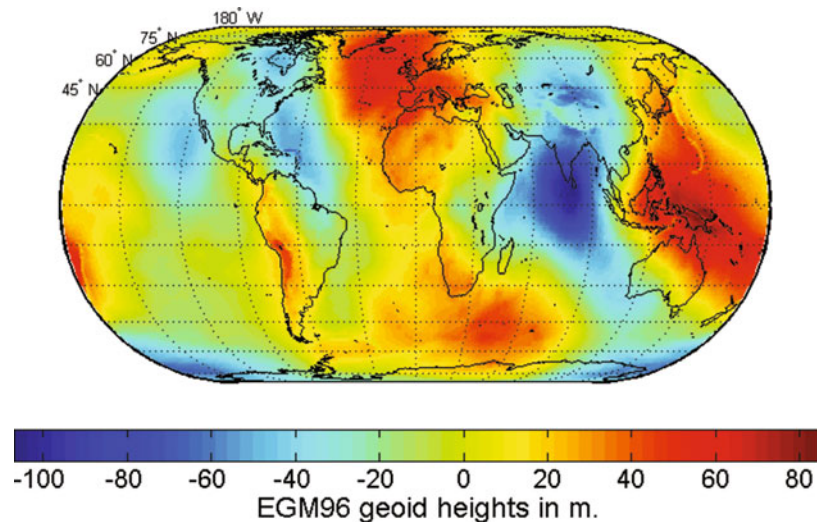
Since the last several decades, geodetic satellites using precise orbit computation determine the global geoid. Changes of geoid undulations and gravity anomalies on the Earth’s surface are determined from the moments of satellite orbits. EIGEN-1S, based on GPS satellite-to-satellite tracking and accelerometer data from CHAMP satellite produces a geoid with an approximation error of about 20 cm in terms of 5×5 degree block mean values, which is made possible because of CHAMP’s tailored orbit characteristics and dedicated instrumentation, providing continuous tracking and direct on-orbit measurements of nongravitational satellite accelerations (Reigber et al. 2002). The AUSGeoid2020 is a combined gravimetric-geometric model. The gravimetric component is a $1'$ by $1'$ grid based on data from gravity satellite missions (e.g., GRACE, GOCE), localized airborne and land gravity data. The geometric component is a $1'$ by $1'$ grid of quasigeoid-height separation values and is developed using a dataset of collocated GNSS ellipsoidal height and Australian Height Datum heights (<http://www.ga.gov.au/ausgeoid/>).

Apart from rotation, there are other motional forces, which affect the shape of the equipotential surface. Precession is the effect of external torque exerted on the Geoid causing the spin axis to describe a circular cone. This motion is caused by the gravitational attraction of other celestial masses, particularly the sun, moon, and planets. Precession arises because of the ellipticity of the Geoid, with torque exerted on the equatorial bulge, in combination with the obliquity of the axis of rotation in relation to the plane of revolution (the ecliptic and fixed, heliocentric space). The period of precession is presently about 25,780 years.

Polar Motion is the change in the position of the axis of rotation in relation to the surface of the Geoids, which was observed by five stations near latitude 39.1333° N, which began latitude observations in 1899. The Conventional

Geodesy, Figure of the Earth,

Fig. 2 Geoid undulations derived from observed gravitational satellite orbit perturbations (from <https://cddis.nasa.gov/926/egm96/egm96.html>). The greatest departure of the geoid from ellipsoidal form is a 105-m low south of India. This is very small compared with undulations of the land surfaces or the 5-km mean difference in level between continents and ocean floors



International Origin is defined as the mean position of the instantaneous pole during the period 1900–1905. Secular Polar Motion, quantified by astrometric and geodetic measurements as well as those of GNSS satellites is the drift in the pole position due to a combination of several forces: glacial isostatic adjustment, Earth's surface mass transport, seismogenic mass exchange, and changes in Earth's inertia tensor by mantle convection (Adhikari et al. 2018). It is a significant component of the Earth's orientation change, which undergoes variations on time scales from subdaily to decadal and longer.

Determining the exact figure of the Earth involves both surface geometry as well as geophysics. Without any idea of the Earth's interior, we can state a “constant density” of 5.515 g/cm^3 , and according to theoretical arguments such a body rotating like the Earth would have a flattening of 1:230. In fact, the measured flattening is 1:298.25, which is more similar to a sphere and a strong argument in favor of very compact Earth's core. Therefore, the density must be a function of the depth, reaching from about 2.7 g/cm^3 at the surface to approximately 15 g/cm^3 within the inner core. With increasing accuracy of the measured global gravity field from dedicated satellite missions, it is observed that the true vertical generally does not correspond to the theoretical vertical (deflection ranges from $2''$ to $50''$) because topography and all geological masses disturb the gravitational field. Therefore, the gross structure of the Earth's crust and mantle can be determined by geodetic-geophysical models of the subsurface (more in ► “Geodesy, Physical”, ► “Geoid”, ► “Gravity, Global Models”, and ► “Gravity Method, Satellite” in this volume).

Reference Coordinate System

Description of the shape and size of the Earth as well as measurement of distances on its surfaces and definitions of

position locations of any particular point inevitably depends upon the use of a Reference Coordinate System. Reference ellipsoids are customarily expressed by the size of the semi-major axis (a) and flattening (f). The quantity f is a purely mathematical one. The mechanical ellipticity of the Earth (dynamical flattening) is determined by observation and differs from the geometrical because the Earth is not of uniform density. The 1967 Geodetic Reference System posited a 6,378,160 m semimajor axis and 1:298.247 flattening. The 1980 Geodetic Reference System posited a 6,378,137 m semimajor axis and 1:298.257 flattening. This system was adopted at the XVII General Assembly of the International Union of Geodesy and Geophysics (IUGG).

The World Geodetic System 84 (WGS-84) is a standard for use in cartography, geodesy, and navigation. It comprises a standard coordinate frame for the Earth, a standard spheroidal reference surface (the *datum* or *reference ellipsoid*) for raw altitude data, and a gravitational equipotential surface (the *Geoid*) that defines the *nominal sea level*. Presently, WGS 84 uses the 1996 Earth Gravitational Model (EGM96) Geoid, revised in 2004. This Geoid defines the nominal sea level surface by means of a spherical harmonics series of degree 360 (which provides about 100 km horizontal resolution). The deviations of the EGM96 Geoid from the WGS 84 reference ellipsoid range from about -105 m to about $+85 \text{ m}$.

International Terrestrial Reference System (ITRS) was defined by the IUGG in 1991 as a geocentric nonrotating system identical to the International Astronomical Union (IAU) defined Geocentric Reference System with Geocentric Coordinate Time (TCG). Conventional Terrestrial System (CT) is geocentric and references north to the 1900–1905 mean pole and 0° longitude to Greenwich Observatory. Terrestrial coordinate systems are Earth-fixed and spin and revolve with the Earth. Terrestrial Reference System (TRS) is a spatial reference system co-rotating with the Earth in space, with positions of points attached to the solid surface of the Earth

having coordinates which undergo only small variations with time, due to geophysical, tectonic, or tidal deformations.

Nowadays, the realizations of the fundamental reference system supported by the International Association of Geodesy, the ITRF (International Terrestrial Reference Frame with the year of realization, at present the ITRF 2008) are very strongly recommended to be only used in precise geodesy for geosciences applications and for navigation satellite systems. Fortunately, the choice of fundamental parameters of WGS 84 was to be made in the beginning (in the 80s of last century) as close as possible to the ITRF (at this time defined by the BIH – the Bureau International de l’Heure). As a result, within 1 m there is no difference between what is defined as WGS 84 and the ITRF 2008, the latest international realization with a centimeter level of accuracy (more in ► “Geodesy, Ground Positioning, and Levelling” and ► “Geodesy, Networks, and Reference Systems” in this volume).

Datum

A datum is a framework that enables us to define coordinate systems. Definition of datum has improved considerably with advancements in measurement techniques through technology.

Traditional Geodetic Datum: To describe positions on the Earth accurately, traditionally two datum were required; a horizontal datum and a vertical datum.

Horizontal Datum: A horizontal datum is used to fix a position in the X and Y directions. A horizontal datum was traditionally defined using an ellipsoid and an origin. The ellipsoid was generally chosen for a best fit of the Geoid locally.

Vertical Datum: A Vertical Datum is used to fix a position in the vertical direction, up and down the Z axis. A vertical datum is a line, value, or set of values from which heights are measured.

The International Terrestrial Reference Frame (ITRF) is defined by the International Terrestrial Reference System and the Datum consistent with it and is based on a number of very accurate observations all over the Earth, employing Very Long Baseline Interferometry (VLBI), Satellite Laser Ranging (SLR), Lunar Laser Ranging (LLR), and the Global Positioning System (GPS) techniques. The measurements relate to a 3D Earth-centered Cartesian coordinate system. In changing to this Geocentric Datum, the Global Positioning System [GPS] can be most effectively used (more in ► “Geodesy, Networks, and Reference Systems” in this volume).

Summary

Geodesy, Figure of the Earth outlines the basic parameters by which the shape of the Earth is defined. The Earth

being a system in dynamic equilibrium is under the influence of many forces, all of which contribute to its final structure and thereby its detailed shape. The subject of Geodesy has evolved sufficiently technologically and mathematically to take into account the details of the Earth’s shape as well as the physics of the causal mechanisms.

Cross-References

- [Geodesy, Ground Positioning, and Levelling](#)
- [Geodesy, Networks, and Reference Systems](#)
- [Geodesy, Physical](#)
- [Geoid](#)
- [Geoid Undulation, Interpretation](#)
- [GPS, Tectonic Geodesy](#)
- [Gravity Method, Satellite](#)
- [Gravity, Global Models](#)

Bibliography

- Adhikari S, Caron L, Steinberger B, Reager JT, Kjeldsen KK, Marzeion B, Larour E, Ivins ER (2018) What drives 20th century polar motion? *Earth Planet Sci Lett* 502:126–132. <https://doi.org/10.1016/j.epsl.2018.08.059>
- Airy GB (1845) *Figure of the Earth*. In: *Encyclopaedia metropolitana*, B. Fellowes, London, vol 5. pp 165–240
- Bomford G (1954) Determination of the European geoid by means of vertical deflections. Rpt of Comm. 14, IUGG 10th Gen. Ass., Rome
- Bomford G (1962) *Geodesy*. Oxford Publications, London
- Brown NJ, McCubbine JC, Featherstone WE et al (2018) AUSGeoid2020 combined gravimetric–geometric model: location-specific uncertainties and baseline-length-dependent error decorrelation. *J Geod* 92:1457–1465. <https://doi.org/10.1007/s00190-018-1202-7>
- Fischer IK (1989) Geodesy: historical introduction. In: *Geophysics. Encyclopedia of earth science*. Springer, Boston
- Hooijberg M (1997) *Practical geodesy*. Springer, Berlin. ISBN 3-540-61826-0
- Moritz H, Hofmann B (2005) *Physical geodesy*. Springer, Wien/New York
- Needham J (1954/1959) *Science and civilization in China*. Cambridge University Press, London
- Reigber C, Balmino G, Schwintzer P, Biancale R, Bode A, Lemoine JM, König R, Loyer S, Neumayer H, Marty JC, Barthelmes F, Perosanz F, Zhu SY (2002) A high-quality global gravity field model from CHAMP GPS tracking data and accelerometry (EIGEN-1S). *Geophys Res Lett*. <https://doi.org/10.1029/2002GL015064>
- Smith JR (1986) From plane to spheroid. Determining the figure of the earth from 3000 BC to the 18th century Lapland and Peruvian survey expeditions. Landmark Enterprises, Rancho Cordova. ISBN 0-910845-29-8
- Smith JR (1996) *Introduction to geodesy. The history and concepts of modern geodesy*. Wiley, New York. ISBN 0-471-16660-X
- Todhunter I (1873) *A history of the mathematical theories of attraction and the figure of the earth: from the time of Newton to that of Laplace*, vol 2, Macmillan & Co., London

Geodesy, Ground Positioning, and Levelling

Stelios P. Mertikas

Geodesy and Geomatics Engineering Laboratory, School of Mineral Resources Engineering, Technical University of Crete, Chania, Greece

Definition

Geodesy	Science for the determination and representation of the figure, shape, and size of the earth and of other celestial bodies along with their external gravity field as a function of time
Positioning	Determination of coordinates for points on earth, sea, or space with respect to a reference system
Ground positioning	Determination of terrestrial coordinates
Levelling	The determination of heights of points with respect to a reference surface

Introduction

Geodesy, along with astronomy, is considered the oldest earth science and dates back to the Sumerians of 5000 years ago. The word “geodesy” is composite and is derived from the Greek noun “Γη = earth” and the verb “δαίω = divide.” It is, literally, the science used to “divide, distribute, and measure the earth.” The roots of geodesy can be found in ancient Egypt and Babylon where “the geometry of the earth” was used for the recovery of the land property boundaries, after, for example, the flooding of crops by the Nile River. The word “geodesy” appeared, for the first time, in the fourth century B.C. in Aristotle’s book *Metaphysika*. Before Aristotle, “geodesy” was called the “geometry of the earth.” Ancient Greeks and Romans were the first to invent geodetic or *surveying* instruments for constructing roads, buildings, and tunnels, laying water pipes and culverts, land surveying, etc. Great scientists such as Pythagoras, Aristotle, Eratosthenes, Poseidonius, W. Snellius, J. Picard, Newton, Gauss, Cassini, and many others contributed to geodesy, surveying, high-precision levelling, and mapping.

Today, *geodesy* is the science of accurately measuring and understanding (1) the geometric size, shape, and orientation of the earth, (2) its gravitational field, and (3) the variations of these quantities over time (Torge and Müller 2014; Herring 2009). The development of geodesy goes together with spectacular achievements in the fields of astronomy, mathematics,

probability and statistics, physics, map projections, fast computing, etc. As a first approximation, “geodesy” can be considered as the science of measuring distances or heights between terrestrial objects or measuring angles between lines and determining directions of lines for the establishment of coordinates of points on or near the earth’s surface. This is also called *ground positioning* or *surveying*.

One of the main tasks of geodesy is *positioning*. This is the accurate determination of coordinates of points on the earth’s surface, at sea, or in space with respect to a reference system (Petit and Lusum 2010). Positioning can be divided into *absolute positioning*, where coordinates are determined with respect to an inertial reference system, and into *relative positioning* when coordinates are derived with respect to other reference points on earth. Positioning constitutes essential information for mapping, cadastral surveys, highway and dam construction, earthquake and landslide monitoring, geodetic control networks, crustal and structural deformation, etc.

Ground positioning has traditionally been carried out using measurements of distances and angles (and or directions) between terrestrial lines with theodolites and measuring tapes (Anderson and Mikhail 1997; Kavanagh and Mastin 2013). Later, distances have been measured by other equipment using the electromagnetic radiation. Over the last two decades, the methods and instruments for measuring, acquiring, storing, calculating, reducing, and portraying geodetic ground positioning data have changed dramatically. New types of geodetic measurements, along with the traditional ones, are now being made with total electronic geodetic stations, terrestrial laser systems, inertial systems, ground-based radar interferometers, terrestrial wireless positioning systems, precise underwater positioning systems, and finally the space-based systems such as the global navigation satellite systems (Kaplan and Hegarty 2017).

Geodetic levelling is the precise determination of height difference between two or more points on the earth’s surface. It ultimately aims at the determination of the height above a reference surface, usually the mean sea surface. Again in levelling, the ancient Egyptians and the Babylonians were the first to employ it to connect the Nile River with the Red Sea and to apply levelling for irrigation. In the following sections, the geodetic ground positioning and levelling will be briefly elaborated.

Ground Positioning

The primary purpose of ground positioning is to provide coordinates of points on the earth’s surface based on either (1) the astronomic determination of latitude, longitude, and azimuth of a terrestrial line or (2) the observation of horizontal angles, distances, zenith angles, and height differences.

A *horizontal angle* is defined as that angle measured in the horizontal plane of the observer between two vertical planes. The *horizontal plane* is the tangent plane to the equipotential surface of the gravity field at the point of observation. In early geodetic surveys, painstaking efforts were employed to accurately measure angles and distances. Angles were – and are still to this day – measured using theodolites, and the distances were measured using tapes made from metal having a low coefficient of thermal expansion. Later, electronic instruments were used for measuring angles and distances.

Electromagnetic distance measuring devices have been used to determine the distance between geodetic stations on the ground. Electromagnetic pulses are transmitted to a retro-reflector (or even to the terrain target), and the emitted signal finally returns to the instrument. Intervisibility requirements and attenuation of the signal, mainly due to the atmosphere, allow measurements to be made at ranges up to about 50 km; but the best results are obtained at shorter distances with an accuracy even higher than 0.1 to 0.3 mm over 100 meters.

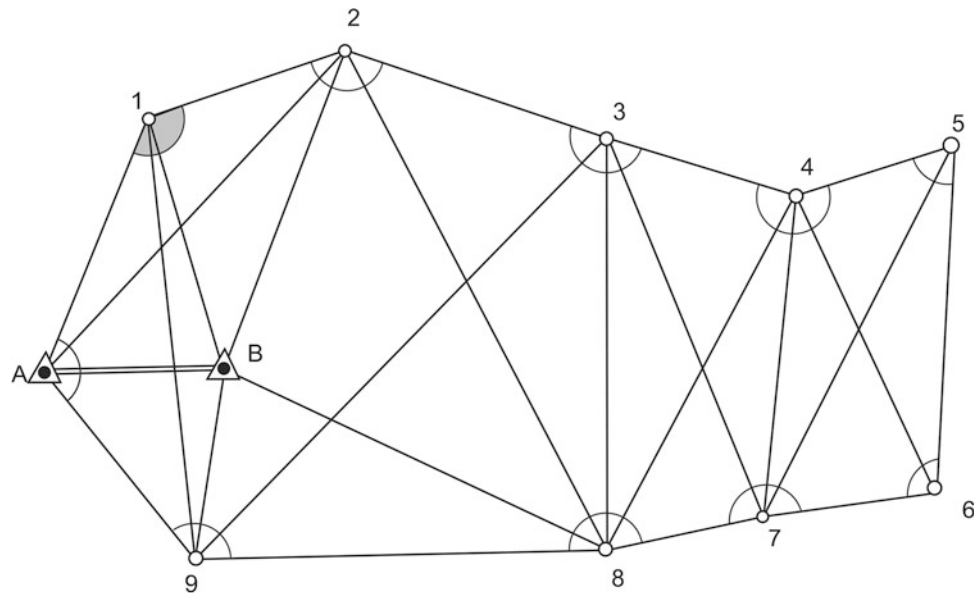
Point coordinates have thus been established by a process of triangulation, trilateration, or traverse surveys. *Triangulation* is the process of determining the relative positions of a

network of *control points* (Kavanagh and Mastin 2013) on the ground by theodolite measurements of directions or angles between intervisible stations (Fig. 1). Angles have been more easily measured, as compared to distances, particularly when large distances over mountainous and rugged terrain were involved. As the name implies, triangulation employs horizontal angles and a limited number of lengths – called *geodetic baselines* – of connected triangles to determine trigonometrically the positions of stations (vertices of triangles). Triangles have been the simplest geometric figure which permit rigorous checks and adjustments of field observational errors. To compensate for the errors in the measurements, triangulation networks are adjusted mainly using the method of least squares. This method not only yields the final and adjusted coordinates of the vertices of the triangles but also gives their coordinate accuracies (Ghilani 2017).

Trilateration is similar to triangulation process and again aims at the relative determination of positions using triangles but now with their lengths measured by electromagnetic radiation with electronic instruments (Fig. 2). Finally, *traverse surveys* consist of measurements of both horizontal (and vertical) angles at the traverse stations and distances

Geodesy, Ground Positioning, and Levelling, Fig. 1

An example of a triangulation network. The baseline AB is either known or measured, while all angles in the formed triangles are observed



Known parameters:

1. Geodetic baseline AB (distance is known),
2. Latitude and Longitude of stations A και B,
3. Azimuth of length AB.

Measurements:

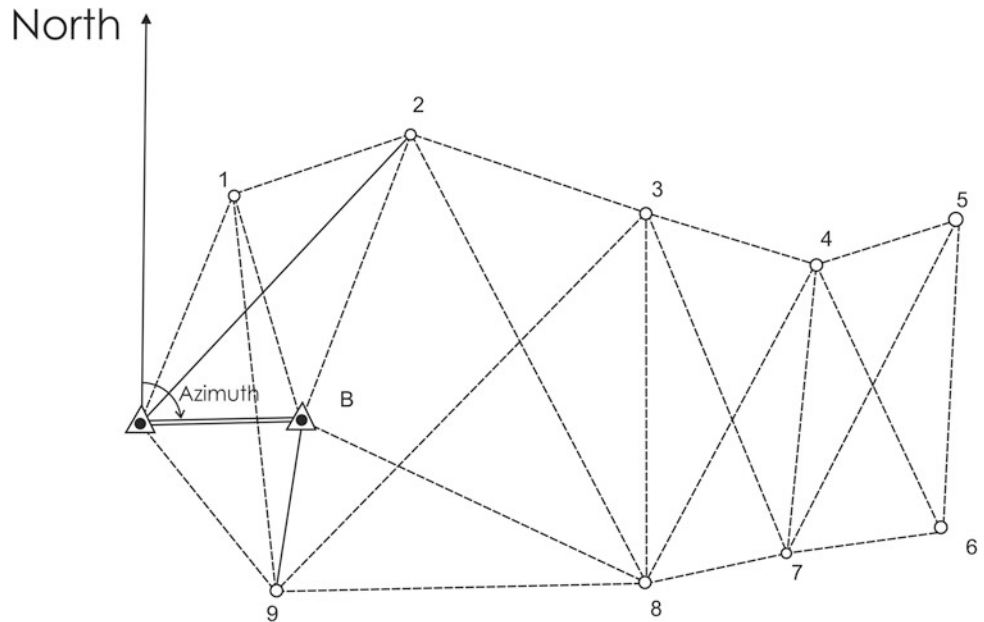
All horizontal angles to visible stations.

What is finally computed:

1. Distances to all control stations (triangle vertices),
2. Latitude and longitude of all stations,
3. Azimuth of all lines.

Geodesy, Ground Positioning, and Levelling, Fig. 2

A trilateration network where the azimuth of the baseline AB is determined to orient the network with respect to the north and all sides of the formed triangles are observed

**Known parameters:**

1. Geodetic baseline AB,
2. Latitude and longitude of stations A and B,
3. Azimuth of line AB.

Measurements:

All distances to all inter-visible sites (stations).

What is finally computed:

1. Angles and distances for all formed triangles,
2. Latitude and longitude for all stations,
3. Azimuth of all lines.

along a series of connected lines on the ground. Unlike triangulation and trilateration, in which stations are commonly widely separated (e.g., 20–60 km), traverse surveys generally follow routes of highways, railways, etc., with stations located close together (i.e., a few hundred of meters). The coordinates of these traverse stations are computed using basic trigonometric expressions as shown in Fig. 3.

Astronomic azimuths are used to determine the orientation of the network with respect to the ellipsoid. The *astronomic azimuth* is the angle in the observer's local horizontal plane between the astronomic meridian and the vertical plane passing through the target point. Two coordinates, the geodetic latitude and longitude of each station of the network, are then computed on this ellipsoid. The third coordinate of *height* above the reference ellipsoid is determined either from levelling and gravity data or from vertical angles.

All these geodetic measurements as well as other types (e.g., distance differences) are being made, today, by total geodetic stations, terrestrial laser systems, inertial systems, ground-based radar interferometry, wireless positioning, etc.

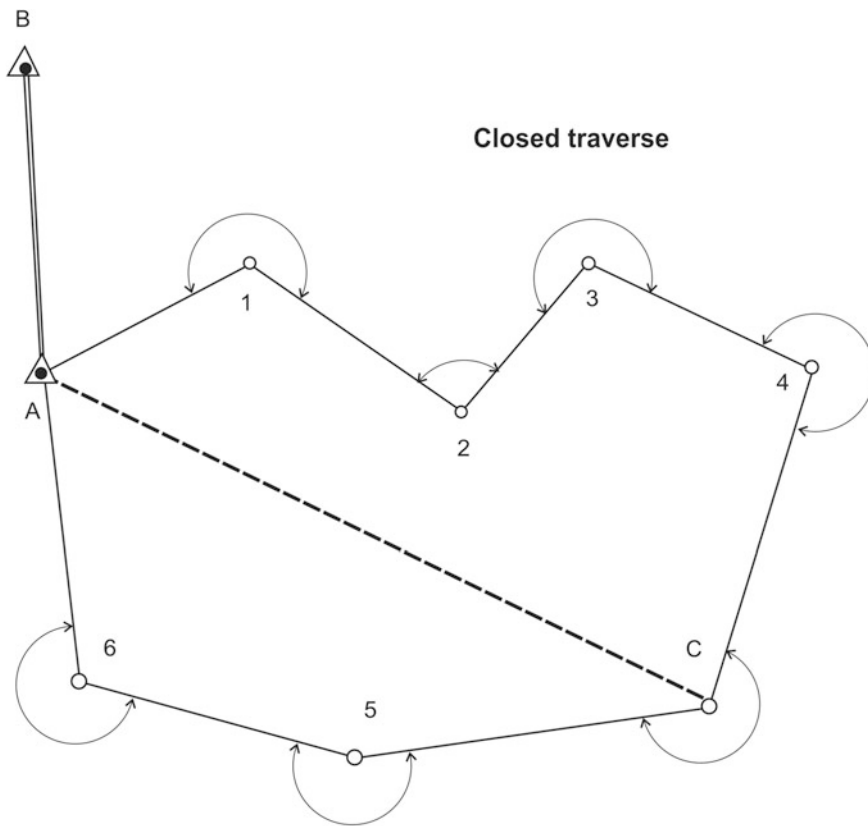
New kinds of theodolite measurement systems do not require a target and/or reflector to be monitored. They are

able to measure point coordinates without targeting the object to be observed (Reiterer et al. 2007). Such systems use the texture on the terrestrial surface or of an object to locate “points of interest” to make angle and distance measurements. However, these new instruments require well-trained personnel to operate them.

Terrestrial laser scanners (Parker 2019) can also provide detailed and highly accurate three-dimensional coordinates of terrestrial points rapidly and efficiently. Some scanners are suitable for indoor use and medium ranges (up to 100 m), while others are better for outdoor applications with longer ranges. Most systems use a laser beam, emitted by the instrument, which is directed into different directions, through a mechanical deflection mirror, to a number of reflective surfaces on the ground. The deflected laser light is detected, and the instrument measures the time of flight for long ranges and the phase measurement of the reflected signal for medium ranges. For various scanning rates and/or a vertical scan angles up to 90°, the deflection mirror rotates continuously at adjustable speed and records target *range*, *angle*, *signal amplitude*, and *time*. Some terrestrial laser scanners incorporate a camera to also capture a picture of an object being

Geodesy, Ground Positioning, and Levelling, Fig. 3

An example of a closed traverse. The geographic coordinates as well as the azimuth of the initial point A, with respect to point B, are known. Then the exterior (or interior) angles at each traverse station are observed along with the interstation distances. The Cartesian coordinates for all stations of the traverse and the azimuths for all lines are determined using trigonometry



Known parameters:

1. Latitude and longitude of point A ,
2. Azimuth of line AB.

Measurements:

1. The distances between all traverse stations
2. Interior angles at each traverse station.

What is finally determined:

1. Cartesian coordinates for all stations,
2. Azimuth for all lines in the traverse.

scanned. The accuracy of distance measurements is of the order of ± 1 cm (± 1 mm over 10 m), but this primarily depends on the intensity of the reflected laser light and on the reflectivity of the target surface. Terrestrial laser scanners have been used to characterize the kinematics of the landslide ground surface, by evaluating the involved displacement and deformation patterns. They have also been applied in surveying, mining, construction surveying, architecture, archaeology applications, deformation monitoring, etc.

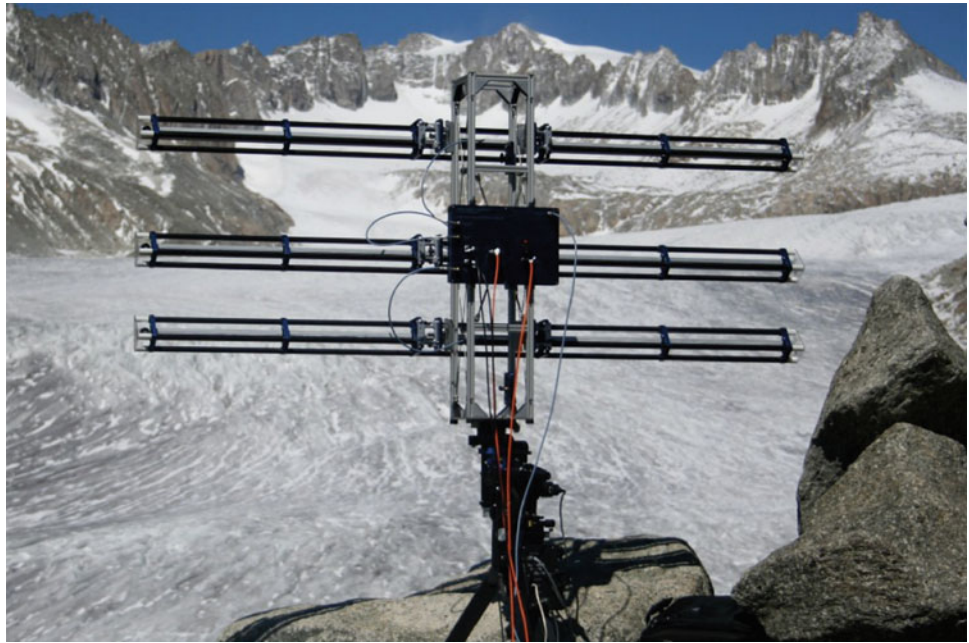
The standard wireless (i.e., Wi-Fi and the ultra-wide band) networks employ a number of transmitters on the ground and not only provide a way for finding the optimum communication relay routes in an ad hoc network of transmitters but are also capable for ground positioning. The locations of the communication nodes (i.e., transmitters) of these networks are known. Positioning is achieved by measuring the *time of arrival* and the *time difference of arrival* for the communication signals. Thus, the positional accuracy of a network user

depends of the ranging errors, the connectivity, and the available number of mobile nodes. The developed ultra-wide band systems provide a dynamic three-dimensional accuracy of better than 5 mm (Zuin et al. 2018). They are, at present, used for indoor wireless positioning, including robot tracking and monitoring, but future uses to geophysical applications are soon to emerge.

Geodetic positioning techniques have also been extended to underwater applications using acoustic ranging to measure crustal deformation with cm level accuracies at several thousand of meters depths (Gagnon et al. 2005; DeSanto et al. 2016).

Another promising technique for geodetic positioning, height mapping, and deformation monitoring, such as tectonic seismic and volcanic deformation as well as ice and rock-glacier motion and ground subsidence, is provided by the *ground-based interferometric radar systems* (Werner et al. 2008; de Macedo et al. 2017). These instruments use real

Geodesy, Ground Positioning, and Levelling, Fig. 4 A ground-based interferometric radar system to monitor deformation



aperture radar antennas to illuminate a target area on the ground (Fig. 4). The antennas are mounted in parallel on a rotational scanner. A radar image of the target area is formed, line by line, by rotating the antennas in azimuth and in the vertical axis. The range and the azimuth resolution of these radars are determined by the antenna beamwidth and the slant range. *Phase differences* (Cumming and Wong 2005) between successive images, acquired from the same viewpoint, are used to determine line-of-sight displacements better than 1 mm. These instruments are used for monitoring rapid moving landslides, and unstable mine slopes, as well as for terrain mapping. They can be used for short ranges from sensor to target (up to 1000 m) to medium-range applications (up to 4 km). The resulted accuracies are of the order of ± 1 mm, and the entire observed area can reach up to thousands of square meters, while the radar system measures the displacements of all points in the area illuminated by the antenna.

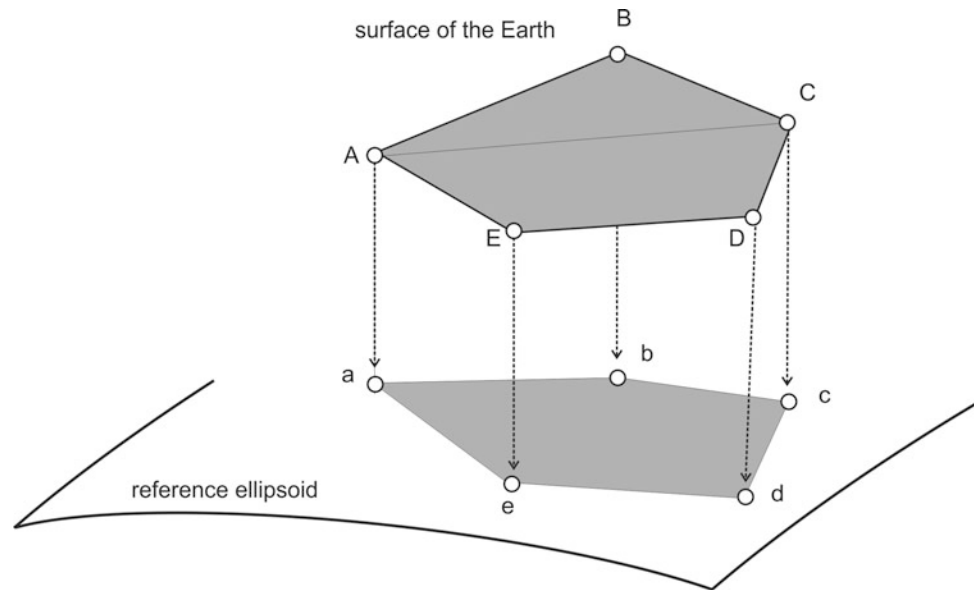
Finally, space geodetic techniques are based on processing signals received from earth-orbiting satellites to provide precise ground positioning and levelling (Herring 2009; Teunissen and Montenbruck 2017). These systems, called global navigation satellite systems (GNSS), include the US Global Positioning System (GPS), the Russian Global Navigation Satellite System (GLONASS), the European Galileo, the Chinese BeiDou, the Indian Gagan, the Japanese Quasi-Zenith Satellite System, etc. Although originally developed for military and navigational purposes, new measuring and processing techniques (Hofmann-Wellenhof et al. 2007; Leick 2003; Teunissen and Montenbruck 2017) were

developed to improve the originally designed positioning accuracy of several meters. Today, these systems provide coordinates of ground locations at the mm level in the horizontal as well as in the vertical sense (Blewitt 2010) over larger distances of several tens or hundreds of kilometers. In addition, they allow for three-dimensional positioning and do not require interstation visibility, their accuracy does not degrade with distance significantly, and they provide continuous (e.g., every 1 or 20 Hz) positioning with respect to a global reference frame. In this vein, they have been applied for mapping, precise ground positioning to monitor plate tectonics, seismic and volcano ground deformation, etc.

The computational process, for all the above geodetic positioning techniques, requires their reduction on a reference surface, usually taken as a flat horizontal surface or the *reference ellipsoid* (Petit and Lusum 2010) (Fig. 5). In the flat horizontal surface, the directions of the plumb line (and thus gravity vector) is considered parallel. The ellipsoid is the curved surface approximating the size and shape of the earth, and the Cartesian coordinate system is usually an earth-centered reference system for three-dimensional computations. Calculations involve solving equations derived from solid geometry, calculus, and statistics (Torge and Müller 2014; Meyer 2010).

The measured angles (directions) and distances are projected along normals onto the reference ellipsoid. The *geodetic datum* (Petit and Lusum 2010) is also defined that enables accurate location of features on earth through a mathematical model that describes the origin, orientation, and scale of coordinate systems.

Geodesy, Ground Positioning, and Levelling, Fig. 5 The observations of angles, distances, etc., measured on the earth's surface are reduced upon a reference ellipsoid and/or a flat horizontal plane



Levelling

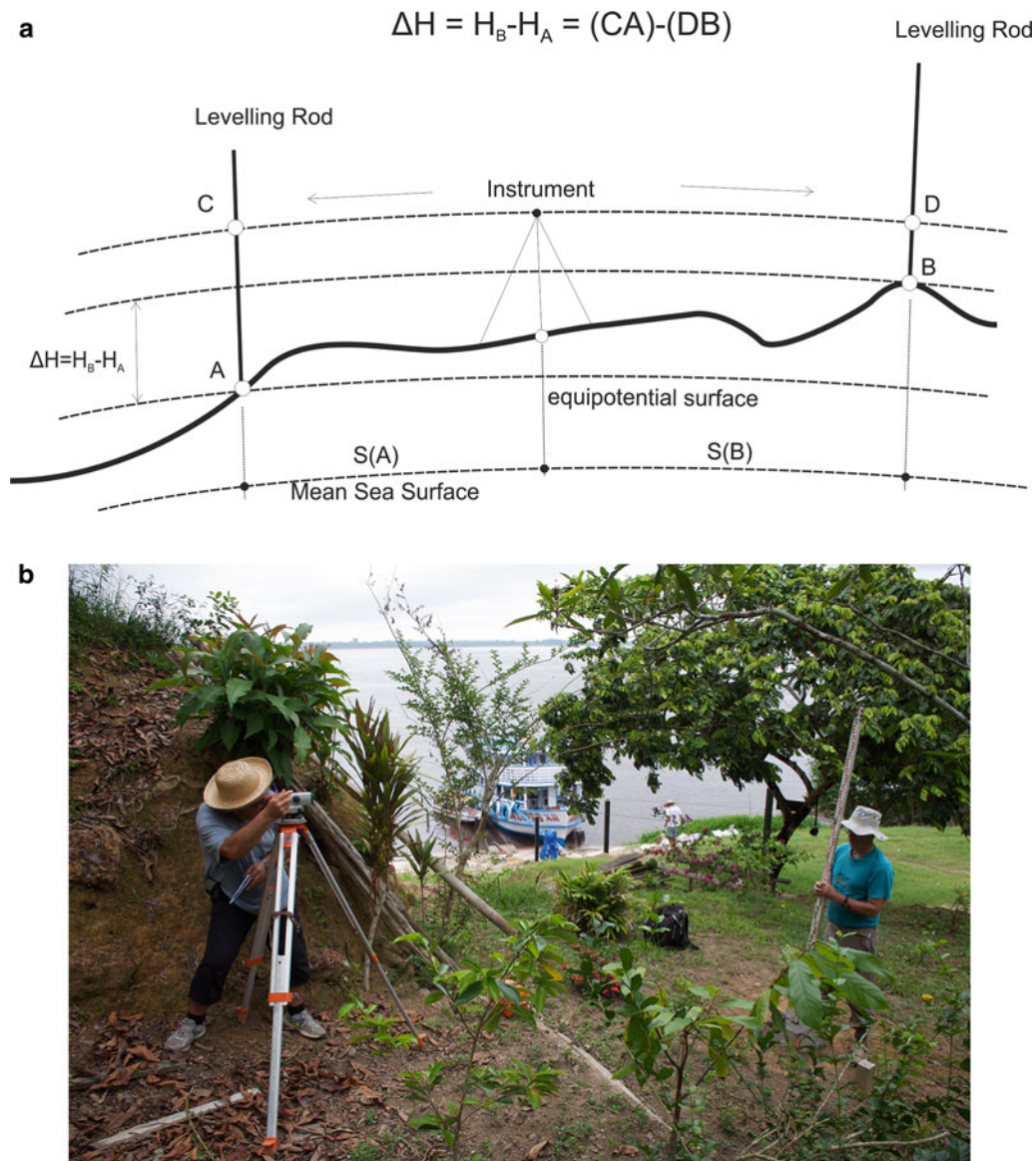
Levelling is the general term applied to indicate the process by which elevations of points or differences in elevation between two or more points on the surface of the earth are determined. Differences in elevation are determined by taping, differential levelling, and barometric levelling and indirectly by trigonometric levelling. All height determinations require the establishment of a vertical datum (Sanchez and Sideris 2017), usually taken as the mean sea level. This is realized by relative permanent objects on the ground, called *benchmarks*, whose elevation above an adopted geodetic datum is known.

Differential levelling is accomplished by making measurements to two or more points through a telescope. The instrument's telescope, with suitable magnification, is regularly set horizontally with its line of sight being parallel to the equipotential surface of the local gravity field. A horizontal line of sight is established within the telescope by a level vial or automatic compensator. The basic procedure is illustrated in Fig. 6. A vertical measurement is made to the backsight of a graduated rod, held on fixed points, and is added to its height. Thus the height of the levelling instrument is determined. A second measurement is made to determine the height of the foresight rod. The instruments can then be moved forward, and the levelling process is repeated until the elevation difference between the initial and the final point is established. Levelling procedures are limited by light of sight and to the distance between the level instrument and the rod. The readings on the rod can be taken either optically or electronically; the rod must be close enough to the level (about 40 m) to resolve the rod's readings in order to make

accurate readings. Today, digital levels measure electronically to bar-coded rods, but can also measure optically to conventional rods. These are easy to use, and because heights and distances are recorded electronically, reading and writing errors are avoided. They can be used for precise levelling of flat and sloping surfaces and for establishing the vertical component of grade and ground profiles and for subsidence monitoring and control. The accuracy in height determination is of the order of less than ± 1 mm per km of distance in double levelling.

Differences in elevation are proportional to differences in the atmospheric pressure. Thus barometers, which are instruments to measure atmospheric pressure, can be used to determine relative elevations of points on the earth's surface. Elevations determined by barometric levelling probably are several meters in error even after they are corrected for the effects of temperature and humidity. Barometric measurements are made at a benchmark, while a second roving barometer is carried from point to point in the field.

The difference in elevation between two points can also be established by measuring the vertical angle of the line from one point to the other and then computing the difference in elevation from either the slope distance between the two points. Trigonometric levelling is necessary when the elevations of inaccessible points must be determined. From two points of known position and elevation, the horizontal position of the unknown point is found by triangulation, and the vertical angles from the known points are measured. The technique is influenced by the atmosphere refraction (cloudy weather with good visibility with consistent temperature over measurement) and the earth curvature.



Geodesy, Ground Positioning, and Levelling, Fig. 6 Differential levelling. Two rods are placed vertically on points A and B, where the height difference between them is determined. The instrument is set up in the midway of A and B. A vertical measurement is made to the backsight

of the graduated rod placed at A. A second measurement is also made to the foresight rod at B. The height difference is determined by subtracting the foresight from the backsight reading

Summary

At present, the reference systems upon which all these geodetic positioning and levelling techniques are based have been well defined with unprecedented accuracy and reliability (Petit and Lusum 2010; Sanchez and Sideris 2017). The theory of errors for geodetic positioning and levelling has been well developed and applied in the measurements made. New data collection and processing technologies have emerged and matured. So, better results and reliable results are obtained, while the burden of adjustment computation

and efficiency has further improved with fast computing and improved measurements and processing algorithms. All in all, geodesy, ground positioning, and levelling have seen tremendous developments in recent years. New measurements (from optical to electronic) and new techniques (from angle measurement to phase and time and time-difference measurement of electromagnetic signals) and new instrumentation with increased accuracy (from meters to cm and mm at long distances) and capabilities (integration with Geographic Information Systems) have been emerged and matured for geophysical applications.

Cross-References

- [Geodesy, Figure of the Earth](#)
- [Geodesy, Networks, and Reference Systems](#)
- [Geoid](#)
- [GPS, Data Acquisition, and Analysis](#)
- [GPS, Tectonic Geodesy](#)
- [Gravity, Global Models](#)
- [Height Systems, Vertical Datums and their Unification](#)

Bibliography

- Anderson MJ, Mikhail EM (1997) *Surveying: theory and practice*. McGraw-Hill, Boston
- Blewitt G (2010) GPS and space-based geodetic methods. In Herring T (ed.), *Geodesy: Treatise on Geophysics*. Amsterdam, The Netherlands: Elsevier. Chap. 11.
- Cumming GI, Wong HF (2005) *Digital processing of synthetic aperture radar data: algorithms and implementation*. Artech House Publishers, New York
- DeSanto JB, Sandwell DT, Chadwell CD (2016) Seafloor geodesy from repeated side scan sonar surveys. *J Geophys Res-Solid Earth* 121:4800–4813. <https://doi.org/10.1002/2016JB013025>
- Gagnon K, Chadwell CD, Norabuena E (2005) Measuring the onset of locking in the Peru–Chile trench with GPS and acoustic measurements. *Nature* 434:205
- Ghilani DC (2017) *Adjustment computations: spatial data analysis*, 6th edn. Wiley, New York
- Herring T (ed) (2009) *Geodesy: treatise on geophysics*. Elsevier, Amsterdam
- Hofmann-Wellenhof B, Lichtenegger H, Wasle E (2007) *GNSS global navigation satellite systems: GPS, GLONASS, Galileo, and more*. Springer, Berlin
- Kaplan DE, Hegarty C (eds) (2017) *Understanding GPS/GNSS: principles and applications*. Artech House Publishers, New York
- Kavanagh B, Mastin T (2013) *Surveying: principles and applications*, 9th edn. Pearson, Boston
- Kuhn M, Zhang C, Mahfouz M, Fathy AE (2009) Real-time UWB indoor positioning system with millimeter 3-D dynamic accuracy, antennas and propagation society international symposium, APSURSI'09. IEEE, Charleston
- Leick A (2003) *GPS satellite surveying*, 2nd edn. Wiley, New York
- de Macedo CKA, Ramos FLG, Gaboardi C, Moreira JR, Vissirini F, Silva da Costa M (2017) A compact ground-based interferometric radar for landslide monitoring: the Xerém experiment. *IEEE J Sel Top Appl Earth Obs Remote Sens* 10(3):975–986. <https://doi.org/10.1109/JSTARS.2016.2640316>
- Meyer HT (2010) *Introduction to geometrical and physical geodesy: foundations of geomatics*. ESRI Press, New York
- Parker PM (2019) *The 2020–2025 world outlook for terrestrial laser scanning*. ICON Group International, Inc, Las Vegas, NV, USA
- Petit G, Lusum B (eds) (2010) *IERS conventions (2010), international earth rotation and reference systems service, IERS technical note no. 36*. Verlag des Bundesamts fuer Kartographie und Geodäsie, Frankfurt am Main
- Reiterer A, Lehmann M, Fabiankowski J, Kahmen H (2007) Quality control for building industry by means of a new optical 3D measurement and analysis system. In: *Proceedings of the 2007 IAG General Assembly*, Perugia, 2–13 July 2007
- Rueger JM (2003) *Electronic surveying instruments: a review of principles, problems and procedures*, monograph no. 18, School of Surveying and Spatial Information Systems. The University of New South Wales, Sydney, Australia
- Sanchez L, Sideris M (2017) G 2017. Vertical datum unification for the international height reference system. *Geophys J Int* 209:570–586
- Teunissen P, Montenbruck O (eds) (2017) *Springer handbook of global navigation satellite systems*. Springer Handbooks, Berlin
- Torge W, Müller J (2014) *Geodesy*, 4th edn. De Gruyter, Berlin
- Werner C, Strozzi T, Wiesmann A, Wegmüller U (2008) *A real-aperture radar for ground-based differential interferometry*. IGARSS, Boston, p 2008
- Yu K, Sharp I, Guo YJ (2009) *Ground-based wireless positioning*. John Wiley & Sons Ltd, West Sussex
- Zuin S, Calzavara M, Sgarbossa F, Persona A (2018) Ultra wide band indoor positioning system: analysis and testing of an IPS technology. *IFAC PapersOnLine* 51(11):1488–1492

Geodesy, Networks, and Reference Systems

Hayo Hase

Geodätisches Observatorium Wettzell, Bundesamt für Kartographie und Geodäsie, Bad Kötzing, Germany

Definition

Geodesy	Scientific discipline dealing with the measuring and representing of Earth in the domains of time, space, and the gravitational field.
Networks	Set of materialized geodetic reference points as carriers of a specific reference frame tied to each other by geodetic measurements.
Reference system	A reference system is based on one definition of the reference. A geodetic reference system is a definition of a specific coordinate system.
Reference frame	A reference frame is one among many possible realizations of a reference system. It is characterized by a set of reference points from geodetic networks and a reference epoch.

Introduction

The historical definition of geodesy was given by F. R. Helmert (1880): “Geodesy is the science of measuring and mapping of Earth’s surface.” According to this definition, the task of geodesy is expressed in Torge (2003): “The task of geodesy is to determine the figure and the external gravity field of the Earth and other objects in space as functions of time by surface- or spacebased observations.”

Therefore, geodesy deals with measurements and representation of Earth’s geometry and its gravitational field and geodynamic phenomena (Earth rotation, Earth tides, and

crustal motion) in three-dimensional and time-varying space. Metrology defines the physical units, which are needed in geodesy. The measurements are related to various geodetic instruments and different methods. Earth-bound instruments are mostly operated in global networks, while a space-borne instrument senses the entire Earth while in orbit. Earth models contain a mathematical description approximating the physical reality in which observation data are modeled. The description of Earth's figure and its external gravity field requires the definition of reference systems, which are constituted by measurements.

As it is very difficult to fix a static and very precise global coordinate system to the continuously changing Earth, a permanent monitoring of reference positions in geodetic networks is necessary. These permanent observing networks, in which fundamental stations for geodesy have an indispensable role, are the bases for global geodetic reference frames. Such reference frames are the backbone for a better understanding of the system Earth, for any kind of geoinformation systems, and for space exploration (Hase 1999).

Geodetic Reference Systems

Theory of Reference Systems

In geodesy, the terms *reference system* and *reference frame* have different meanings: A reference system is based on physical theories and reflects an ideal idea. A reference frame is one among many possible realizations of a reference system and based on geodetic observations and hence available to users.

Example: A global reference system can be defined in such a way that the origin of the coordinate system coincides with the center of mass. The center of mass is an abstract physical model and not accessible as reference point. However, geodetic observations of satellite orbits are useful to realize a corresponding reference frame because the center of mass of Earth coincides with one of the two foci of the elliptical orbit of the satellite. The used data set of observations delivers one particular reference frame with respect to the definition of the reference system.

Reference systems are standardized by the introduction of conventional reference systems which are based on the decisions of international committees. These conventional reference systems become available to users by conventional reference frames (Schneider 1999). Reference frames are always based on observation data, which are processed with introduced hypotheses of the used reduction models. Due to the arbitrary settings in the processing of observation data, the defined ideal of a reference system can only be approximated with different analysis strategies. This circumstance and the need for comparison call for a standardization of reduction models in order to compare consistently different reference

frames. The International Earth Rotation and Reference Systems Service (IERS) provides the IERS Conventions for the standardization of physical constants and models with respect to global reference frames (IERS 2010)

Two reference systems are of importance in global geodesy: (a) the barycentric space-bound celestial reference system (CRS) and (b) the geocentric Earth-bound terrestrial reference system (TRS). While the first system is related to distant objects in space, the latter one is related to terrestrial reference points or to reference points in the near-Earth space. Both systems are tied by the Earth rotation parameters (EOP).

The figure of Earth is approximated by Earth models. The geometric models are spherical or ellipsoidal, while the geophysical models are based on expansion of the spherical harmonic functions in order to model the geoid as an equipotential reference surface. In an ideal situation, models of both categories coincide with their origins and have parallel main axes.

Metrology

Geodetic measurements in the domains of space, time, and gravity field refer to physical units of the SI (Système International). In general the modeling of space geodetic measurements in Newtonian space-time shows deficiencies. The underlying physical model for the definition of physical units must be based on the theory of general relativity (Damour et al. 1991). This requires that the physical units are defined in infinitesimal small spatial areas with only local and not global validity. The question is whether physical units defined in local laboratory space can be used on the global scale for space geodesy.

The units for time scale, spatial scale, and gravitational potential cannot be defined independently from the object to be measured. The metrological units are only defined as precisely as the measurements of the geodynamic phenomena interacting with the laboratory and vice versa (Hase 1999).

The introduction of the constant of the speed of light, $c = 299,792,458$ m/s, enables the introduction of the physical units *meter* and *second* and a method to measure space-time intervals if the synchronization of two local clocks fulfills the Einstein synchronization condition. Generally this condition is not fulfilled on Earth. Therefore, the spatial distance of the clocks must be defined in relation to the desired resolution of the measurement in order to neglect the spatial curvature effects.

The global time scale is ideally related to the center of the mass of Earth as the effect of external gravitational bodies disappears. For practical reasons timekeeping is not possible at the center of Earth's mass. Therefore the time scale definition of "Temps Atomique International" (TAI) refers to a geocentric reference system with a definition of the second related to the equipotential surface of the rotating geoid. The world time is the *Coordinated Universal Time* (UTC) which is

TAI corrected by an integer number of leap seconds realized by local clocks. The unit of the time second is realized by cesium frequency standards. Its SI definition is: The second is the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium atom ^{133}Cs (BIPM 1998).

The metric scale is a secondary unit derived from the definition of the time second and the constant of the speed of light. The meter is defined as the distance traveled by light in vacuum in $1/299,792,458$ s. Depending on the definitions for speed of light and time second, the metric scale is only valid in the vicinity of the time second-defining clock. That is why large distances should be seen as space-time intervals which are characterized by two-way measurements or one-way measurements with clock transportation.

The mass of Earth could be seen as constant if there were no exchange of radiation with the ambient universe. With respect to the terrestrial reference frames, the solar system with its gravitating bodies is seen as sufficiently isolated. The universal gravitational constant, $G = 6.674,30 \cdot 10^{-11} \text{ m}^3/\text{kg}/\text{s}^2 \pm 0.000,15 \cdot 10^{-11} \text{ m}^3/\text{kg}/\text{s}^2$, is known with its relative standard uncertainty of $2.2 \cdot 10^{-5}$ (NIST Reference 2018). G is the only link between the SI unit for mass and the gravitational phenomena. Given that only in the magnitude 10^{-17} the tidal potential at the height of the geoid has an effect on the definition of the time second, the value of G may be applied in geocentric as well as in barycentric reference systems. The product of the universal gravitational constant with the mass of Earth is called the geocentric gravitational constant, $GM = 3.986,004,418 \cdot 10^{14} \text{ m}^3/\text{s}^2 \pm 8 \cdot 10^5 \text{ m}^3/\text{s}^2$, and can be determined with higher accuracy ($2 \cdot 10^{-9}$) than the individual factors (10^{-4}). It is used for geodetic Earth models.

The IERS Conventions and IERS Standards mirror the actual status of parameters and models to be used in order to obtain a consistent TRF approximating the underlying physical theories with the best available models and constants (IERS 2010).

Global Geodetic Reference Systems

With the beginning of the satellite age, the need for an accurate global geodetic reference system became evident. Also spatial and geographical information datasets extending over national boundaries are calling for a common global reference frame for collection, storage, visualization, and exchanging of geoinformation. On February 26, 2015, the General Assembly of the United Nations adopted Resolution 69/266 A global geodetic reference frame for sustainable development (UN69/266) urging member states to contribute.

Global geodetic reference systems have always been based on definitions and measurements with respect to the available scientific knowledge and technology. In former times solid Earth models had been sufficient for global reference frames,

scientific geodetic studies, and for mapping applications. Modern geodetic space techniques such as very-long-baseline interferometry (VLBI), satellite laser ranging (SLR), global navigation satellite systems (GNSS), and modern instrumentation of local sensors such as absolute gravity meter (AG) and superconducting gravity meter (SG), providing much higher resolutions in the measurements, showed the variability of the observables as a function of time. Consequently, the solid Earth models had to be replaced by viscoelastic Earth models.

For the purpose of reference frames, the viscoelastic Earth models must also include models of plate-tectonic motion because reference points at different continental plates expose different directions and velocities of proper motion over time. The exhaustion of the kinematic motion of individual reference points is achieved in the coordinate estimation under the condition that the squared sum of the individual point motion is made to zero. This condition is commonly expressed as *no-net-rotation* (NNR).

As it appears to be an impossible task to fix a static and very precise global coordinate system on the continuously changing Earth, a permanent monitoring of reference positions in geodetic networks became necessary. These permanent observing networks, in which fundamental stations for geodesy have an indispensable role, are the bases for global geodetic reference frames.

Currently two global geodetic reference systems have a major importance: ITRS and WGS84.

International Terrestrial Reference System (ITRS)

The most important conventional terrestrial reference system (CTRS) is the International Terrestrial Reference System (ITRS). The ITRS is defined as a Cartesian coordinate system in which (a) the origin coincides with the center of Earth's mass (including the masses of the hydrosphere and the atmosphere), (b) the metric scale is defined by the fixed velocity for the propagation of light in a gravitating environment and the physical definition for the unit of time according to SI, (c) the orientation is obtained by observation of Earth rotation relative to space-fixed objects of the celestial reference frame (ICRF), and (d) the system has no global residual rotation with respect to horizontal motions at the Earth's surface (NNR) (IUGG 1991). The definition of origin, scale, and orientation is the *geodetic datum* of a reference system. If a geographic coordinate system (latitude, longitude, ellipsoidal height) is chosen, the corresponding parameters of the desired reference ellipsoid must also be defined and become in general part of that datum. The GRS80 ellipsoid (Moritz 2000) is often used for the ITRS as best fitting the global figure of Earth's surface.

The database of the IERS Terrestrial Reference Frame (ITRF) is provided by International Services for VLBI,

SLR, GNSS, DORIS, and the collection of local surveys at collocation sites with multiple techniques.

The ITRF is published with respect to certain epochs, i.e., the ITRF2014 is based on observation data of more than 1,499 stations located in 975 sites, where about 10% of them are co-located with 2, 3, or 4 distinct geodetic instruments (IERS 2017). The ITRF coordinates and velocities of the reference points are estimated in the processing of observational data. Each reference frame refers to a specific epoch and shall allow linear extrapolation of the coordinates to any desired epoch. However, long-term series and certain geographical regions partly expose nonlinear evolution of point positions and require frequently updated regional reference frames or nonlinear motion models (Altamimi et al. 2016).

The updating of the ITRF is necessary not only due to the continuously acting geodynamic forces on the global network and the need to update reference points but also due to the introduction of improved models in the data reduction and new stations. For the generation of each new frame, the NNR condition is applied. The IERS Conventions contain the adopted constants, parameters, and models which shall be used for data processing (IERS 2010).

World Geodetic System 1984 (WGS84)

WGS84's geometrical base is comprised of (a) a global coordinate frame for the Earth and (b) a standard spheroidal reference surface (the *datum* or *reference ellipsoid*) for raw altitude data. In addition, the geophysical base is a gravitational equipotential surface model (the *geoid*) that defines the nominal sea level. The geophysical base was periodically updated as the underlying Earth Gravitational Model (EGM96, EGM2008) improved by its ground resolution from 200 km to 10 km.

The geometrical definition of WGS84 identifies Earth's center of mass as the origin of the coordinate system. The *zero longitude* is identified with the IERS Reference Meridian, which lies 5.31 arcsec east of the Greenwich Prime Meridian for historical reasons related to the NNSS Doppler System.

The WGS84 is defined and maintained by the US Department of Defense. The GPS control segment consists of only a few globally distributed reference stations, whose coordinates together with the broadcast ephemerides are expressed in this datum. The control segment network is also called *zero-order network* as it fixes the reference frame.

Consequently any GPS user who wants to determine a position in WGS84 is dependent on how accurately known the satellite orbits are with respect to the control segment reference stations. Due to orbit perturbations, the broadcast ephemerides are only approximations and need frequent updates. In absolute positioning mode, results are said to be in the WGS84 datum. In relative positioning model, a baseline vector between two stations can be determined with much

higher accuracy. However, the datum is almost completely defined by the base station, which has a much higher uncertainty than the baseline vector. It is therefore desirable to tie GPS baseline vectors to ITRF reference stations using the precise orbit product of the International GNSS Service (which is obtained from hundreds of network stations) and express sets of estimated coordinates related to the most recent ITRF. The *absolute accuracy* is defined by the true relationship of a position to the coordinate datum. The ITRF is multi-technique based and with a higher number of network stations, for which reason the ITRF provides a higher absolute accuracy than the WGS84 in its current configuration.

As it can be recognized by the defining parameters (Tables 1 and 2), the ITRS and WGS84 coincide within a few cm.

Continental Reference Systems

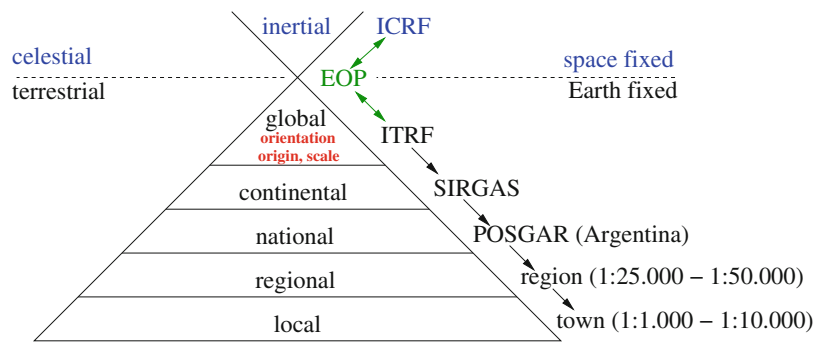
Like the surveying practice “working from the whole to the part,” continental and national reference frames are established using subsets of reference points defining the global terrestrial reference frame. These can be ordered hierarchically beginning with the global terrestrial reference frame (TRF) in which the origin, orientation, and scale are defined. The continental reference network contains the corresponding subset of the global reference points providing

Geodesy, Networks, and Reference Systems, Table 1 Defining geometrical and geophysical parameters of the Geodetic Reference System 1980 (GRS80). The dynamical flattening expressed by the dynamical form factor J_2 and determined by satellite orbit perturbations was given preference in the definition. The geometrical flattening $f = 1/298.257,222,101$ became a derived parameter (Moritz 1980)

Defining parameters of GRS80	Symbol	Value
Semimajor axis or equatorial radius	a	6,378,137.0 m
Dynamical form factor	J_2	0.001,082,63
Geocentric gravitational constant (incl. atmospheric mass)	GM	$398,600.5 \cdot 10^{+9} \text{ m}^3/\text{s}^2$
Angular velocity of Earth rotation	ω	$0.729,211,5 \cdot 10^{-4} \text{ rad/s}$

Geodesy, Networks, and Reference Systems, Table 2 Defining geometrical and geophysical parameters of the World Geodetic System 1984. As opposed to the GRS80, the geometrical flattening, f , is the defining quantity and the dynamical flattening given by the dynamical form factor $J_2 = 0.001,081,874$ which is a derived quantity

Defining parameters of WGS84	Symbol	Value
Semimajor axis or equatorial radius	a	6,378,137.0 m
Flattening $(a-c)/a$	f	$1/298.257,223,560$
Geocentric gravitational constant (incl. atmospheric mass)	GM	$398,600.441,8 \cdot 10^{+9} \text{ m}^3/\text{s}^2$
Angular velocity of Earth rotation	ω	$0.729,211,5 \cdot 10^{-4} \text{ rad/s}$



Geodesy, Networks, and Reference Systems, Fig. 1 Hierarchical relationship between different reference frames. The International Terrestrial Reference Frame (ITRF) is tied to the space-fixed International Celestial Reference Frame (ICRF) via the Earth orientation parameters (EOP). The Earth-fixed ITRF depends on the observational data from fundamental stations to define its origin, orientation, and scale. Subsequent frames on the continental, national, regional, or local level are consistent with each other if they are including reference points from the superior level (shown here is the example of Argentina in South

America). Thus the definition of the global reference frame is introduced into the continental SIRGAS and into the national POSGAR reference frame until to the level of cadastre in towns. The shape of the triangle symbolizes the increasing number of reference points. On the top level, only a few global reference points (fundamental stations) for the definition of scale, origin, and orientation are necessary. On the bottom level, the maximum number of reference points is needed for large-scale mapping and cadastre purposes

the consistent frame and is densified by additional reference points. Consistency to the global reference frame is achieved by the network adjustment with the introduction of the global definition of origin, scale, and orientation represented by the subset of global reference points. National reference frames are established by composing the network of the corresponding global and continental reference points on national territory and additional densifying of national reference points. Thus the densification of geodetic networks can be done in a consistent way to the most accurate global terrestrial reference frame, the ITRF. One example for consistency between hierarchical reference frames is the geocentric reference frame of the Americas, SIRGAS, and subsequent national reference frames in South-American countries (s. Fig. 1). The GRS80 is mostly adopted as the reference ellipsoid for the continental reference frames for consistency reasons to the ITRF (Table 3).

Continental vertical reference systems have to overcome the national height system definitions and to be made consistent at the borders. Usually in continent-wide measuring campaigns at tide gauges and at national first-order reference points with leveling, GNSS and gravimetry are performed to resolve this task. An example is the European Vertical Reference System (EVRS) in its realization of the European Vertical Reference Frame 2007 (Ihde et al. 2008).

National Geodetic Reference Systems

GNSS techniques turned the former particular national reference system into being part of a superior continental or global reference system. National geodetic networks are needed to provide a reliable common coordinate system for spatial datasets and geographical information data. At present most national geodetic network infrastructures consist of

Geodesy, Networks, and Reference Systems, Table 3 Continental reference frames

Continental reference frames	Acronym	Meaning	Earth model
Africa	AFREF	African Geodetic Reference Frame	WGS84
Asia-Pacific	APRGP	Asia and Pacific Regional Geodetic Project	GRS80
Europe	EUREF	European Reference Frame	GRS80
North and South America	SIRGAS	Geodetic System for the Americas	GRS80

continuously operating reference stations (CORS), which are linked to the ITRF. A GNSS CORS comprises a GNSS receiver with its antenna mounted on a stable monument at a safe and secure location equipped with a reliable power supply and a connection to the communication networks for data transmission. Different services make this infrastructure available to the users: (a) streaming real-time (raw) data over the Internet (NTRIP); (b) streaming real-time kinematic data (RTK) and differential GNSS data (DGNSS) for applications such as RTK survey, GIS data acquisition, and GNSS navigation; and (c) transfer of hourly or daily data files to an FTP server.

The national coverage with CORS depends on the resources available and on the tectonic situation. A good example is the Japanese GPS Earth Observation Network (GEONET) with more than 1,300 CORS. This network with 30–50 km distance between each CORS is used for land survey but also for the monitoring of seismic and volcanic activities.

For historical reasons national geodetic systems may still refer to a best-fitting ellipsoid as a reference surface to the

given national latitudes. With the introduction of the hierarchy of reference systems, the global Earth models are becoming more important especially for national reference systems due to consistency.

Vertical national reference systems are usually related to a tide gauge. The *zero height* can be related to an average level of the sea water height (i.e., mean sea level). Some countries transferred the vertical reference from the tide gauge by spirit leveling to a distant benchmark at a secure location. Vertical reference systems are preferably related to the sea level as the accessible part of the geoid. Vertical reference networks are measured not only by spirit leveling but also by GNSS combined with gravimetric measurements.

Historical Reference Systems

In the past, horizontal national reference systems were based originally on triangulation networks, when directions could be measured more precisely with theodolites than distances by rods. Triangulation networks had been oriented by astronomical azimuth measurements, setting the horizontal geodetic datum at the fundamental point. The scale was introduced by baseline enlargement networks.

The error in triangulation networks could reach 1–5 ppm with respect to the distance to the datum reference point. When electronics entered the geodetic instrumentation during the 1960s, electromagnetic signals (microwave, laser) allowed the measurement of distances more precisely than directions. Trilateration networks complemented or replaced the triangulation networks.

The introduction of space geodetic methods in the 1980s, i.e., Navy Navigation Satellite System based on the Doppler effect and the follow-up GPS system, replaced most of the historical national reference systems based on classical triangulation or trilateration methods. The advantages are the global availability and the superior accuracy of the GNSS. The disadvantages are the dependency on the control segment of the GNSS and the availability of energy and communication in order to be able to use it.

The separation of horizontal and vertical reference networks is the consequence of the historical development of the geodetic methods. Modern geodetic space techniques VLBI, SLR, and GNSS are three-dimensional methods and can be used to overcome the classical separation. This becomes possible if the geometrical and geophysical properties of the geodetic observations are modeled correctly in the corresponding reference frame.

Global Geodesy

International Services

The task of global geodesy requires a global infrastructure with geodetic observatories. Institutions in many countries

collaborate voluntarily in the global task to provide terrestrial reference frames by sharing their human, technical and financial resources, observational data, and geodetic products. This work has been initiated under the umbrella of the International Association of Geodesy (IAG). Even today, no country has been capable of reaching the same results on its own. This fact makes the nonprofit-oriented work of international services indispensable to achieve the ITRF and the ICRF. With the UN Resolution 69/266, the former scientific-driven global geodetic networks are complemented by an administrative support through the Committee of Experts on Global Geospatial Information Management (UN-GGIM) in order to enhance the global geodetic networks and improve the unequally available geoinformation infrastructure.

The international services can be classified according to their domains. In the space domain, the following services are the International VLBI Service (IVS), International Laser Ranging Service (ILRS), International Global Navigation Satellite System Service (IGS), and International DORIS Service (IDS). In the time domain, the Bureau International des Poids et Mesures (BIPM) in Paris hosts the office, which is generating the Universal Time (UT) scale as the time base. In the gravity field domain, the International Gravity Field Service (IGFS) with its Technical Support Center (IGFS-TSC) forms the umbrella for the Bureau Gravimétrique International (BGI), the International Service for the Geoid (ISG), the International Geodynamics and Earth Tide Service (IGETS), the International Center for Global Earth Models (ICGEM), the Combination Service for Time-variable Gravity Models (COST-G), and the International Digital Elevation Model Service (IDEMS). The International Earth Rotation and Reference Systems Service (IERS) is primarily using the products of the IVS, ILRS, IGS, and IDS for the computation of the ICRF, ITRF, and EOP. All of the services mentioned above are part of the International Association of Geodesy (IAG).

With the development of the geodetic space techniques, it became obvious that an international coordination is beneficial for the users of geodetic products in navigation, land surveying, mapping, and scientific research. Therefore most of the international services were created during the 1990s to continue previously existing project-related collaborations.

The IAG introduced a project called Global Geodetic Observing System (GGOS) as its contribution to the Global Earth Observing System of Systems (GEOSS). GEOSS is an effort of the Group of Earth Observation (GEO), which unifies more than 80 countries and more than 50 international organizations to improve the monitoring and hence the better understanding of global changes. The international services form the backbone for the realization of GGOS. The goal of GGOS is to provide a TRF with a global precision of 1 mm in position and 0.1 mm/yr. velocity determination (Plag and Pearlman 2009). This ambitious goal can only be reached if

more geodetic observatories with the characteristics of a fundamental station for geodesy will be implemented in a more homogeneous global distribution. The international services themselves try to improve their products so that they match the high precision demands of GGOS. This challenges the design of observing networks, improved observation schedules, new instruments and targets, improvements of the data acquisition and processing, and the homogenization and integration of the models of geodetic space techniques.

If the postulated precision can be achieved in the reference frames, trends in global changes will be discovered and resolved in shorter time periods. This compensates partly the lack of long-term monitoring data at many locations on Earth.

Fundamental Station for Geodesy

At fundamental reference points for global reference frames, one encounters the necessary instrumentation for observing in the space domain, in the time domain, and in the gravitational field domain of Earth. These domains correspond to the underlying physical model on which global geodetic reference systems need to be defined.

Each geodetic space method leads to a specific TRF, e.g., VLBI to a VLBI-TRF, SLR to a SLR-TRF, or GNSS to one of the GNSS-TRF. Each one of the specific TRFs contains site positions and velocities of the specific instruments. Each one of the different techniques has some advantages but also weaknesses regarding the realization of reference systems. Geodesists prefer to use all the techniques in a synergetic way in order to minimize the effect of otherwise undetectable systematic errors. Thus, the collocation of different geodetic space technique instruments at some sites is the key to achieve the best TRF. The space geodesy instruments also need to be complemented by local sensors in order to obtain a complete set of observables for the models used in data reduction.

A geodetic observatory in which all relevant space geodetic observation methods are operated is called a fundamental station if it complies with the following characteristics: (a) *permanency* of the operation by consideration of the time scales of geodynamic phenomena and therefore long-term assured production of geodetic data, (b) *complementariness* of the operated geodetic space techniques for the best possible coverage of the system Earth affecting phenomena as to obtain the best possible realization of a precise global reference system, (c) *redundancy* within the selected instruments for quality assurance of the observational data by independently obtained results of the same observables, and (d) *linking* of the technique-specific reference points by their corresponding space vectors, usually determined by a local survey within a local geodetic network, which is also used for the monitoring of the local site stability.

A fundamental station for geodesy enables the link and the transformation from one technique-specific reference frame

(e.g., VLBI) to another (e.g., GNSS) by coordinate transformation. Fundamental stations are constitutive for the realization of the ITRF which is based on observational data produced by VLBI, SLR, GNSS, and DORIS (Hase 1999).

Instrumentation of a Fundamental Station

The concept of a fundamental station was first realized at the Geodetic Observatory in Wettzell, Germany, in 1988. Since then the concept has been applied to other observatory sites, i.e., Ny-Alesund, Yebes, GGAO, Yarragadee, Hartebeesthoek, and AGGO. These observatories operate radio telescopes for geodetic/astrometric VLBI, optical telescopes with a laser system for SLR, and some GNSS receivers at the site plus an additional regional permanent network for monitoring the regional stability in the space domain. At these sites *local geodetic networks* are installed which allow to tie the telescope reference points to the GNSS antenna reference points and hence the individual technique-specific reference frames.

As the geodetic space techniques basically measure time intervals, a fundamental station needs to be equipped with frequency standards defining the time unit: the second. The instrument for the time domain is a time and frequency laboratory, which operates at least three atomic clocks based on frequency standards. Redundancy is an issue in timekeeping instrumentation. Otherwise, drifts in time cannot be related to one specific clock. That is, geodetic observatories may host three cesium normals for long-term time scale generation and up to three hydrogen maser for short-term time interval measurements.

Concerning the gravity field domain, a fundamental station must measure the gravity periodically with an absolute gravity meter and permanently with a superconducting gravity meter in order to determine the Earth tides. The daily gravitational deformations of Earth's crust reach several decimeters which is therefore the largest time variable spatial parameter in intercontinental baseline measurements and needs to be considered in the analysis of geodetic space techniques.

Fundamental stations for geodesy produce observational data in the space, time, and gravitational field domain.

In order to perform the observations correctly or to interpret the measurements afterward, additional local sensors are necessary. Meteorological sensors deliver input data for the application of refraction corrections during observation with the telescopes and compute signal travel times through the atmosphere. Hydrological sensors are requested to model seasonal variations in the gravity data. Laser gyroscopes deliver local Earth rotation parameters (as opposed to global determined EOP by geodetic space techniques), Seismometers deliver epochs of seismic events to model the nonlinearity of continental drift respective to the site motion (Fig. 2).

Geodesy, Networks, and Reference Systems, Fig. 2

The Argentinean-German Geodetic Observatory (AGGO) hosts as a fundamental station for geodesy a radio telescope for VLBI (in the back), a GNSS antenna (on the right), and a SLR telescope (in the front). The invariant points of the telescopes and the phase center of the GNSS antenna are object of a local survey in order to determine the spatial vectors between the reference points of each technique. Only the co-location of complementary instruments and its local survey enable the global reference frames with the quality of the ITRF. AGGO also hosts a time and frequency laboratory for the time reference and a gravity laboratory with absolute and superconducting gravity meter for the geopotential reference (not in this picture)



Summary

The task of geodesy is to establish global geodetic reference systems. These reference systems are defined theoretically in context of the knowledge of physics. Global reference systems are realized by geodetic observatories and measuring reference stations which mark reference points in the space and time domain and in the gravitational field of Earth. Metrology provides the corresponding physical constants and units to be used in the measurements. The cooperative use of the same geodetic method at various measuring platforms creates a geodetic network. A set of observational data from a geodetic network defines a specific geodetic reference frame which relates usually to a specific epoch. Continuous changes of Earth's surface require a permanent monitoring. Different terrestrial reference frames can be tied together by collocating geodetic space techniques.

Fundamental stations for geodesy establish reference points in technique-specific TRF which are tied together by the knowledge of the local spatial vectors between the technique-specific reference points. The ITRF is based on observational data of various space geodetic methods tied in a synergetic manner at collocated and fundamental stations. The WGS84 is based only on GPS observations and coincides with the multi-technique-based ITRF within 1–2 cm. Global change studies drive the global precision of TRF to 1 mm in position and 0.1 mm in point velocity. The International Association of Geodesy undertakes an effort with the Global Geodetic Observing System to set

up the necessary infrastructure and improve the geodetic products.

The use of geodetic satellite methods in positioning requires a precise knowledge of orbital parameters. Two out of six parameters are dependent on the variable Earth's rotation. Monitoring Earth's rotation requires the creation of a Celestial Reference System. The ICRF are based on quasar positions measured by global VLBI networks. The transformation parameters from the space-fixed celestial reference frame to the Earth-bound terrestrial reference frame are given by the time variable Earth orientation parameters (EOP).

Cross-References

- [Earth Rotation](#)
- [Earth Tides](#)
- [Geodesy, Figure of the Earth](#)
- [Geodesy, Ground Positioning, and Levelling](#)
- [Geodesy, Physical](#)
- [Geoid Determination, Computational Methods](#)
- [Geoid Determination, Theory and Principles](#)
- [GPS, Data Acquisition, and Analysis](#)
- [GPS, Tectonic Geodesy](#)
- [Gravity Field of the Earth](#)
- [Gravity Method, Satellite](#)
- [Gravity, Global Models](#)
- [Very Long Baseline Interferometry](#)

Bibliography

- Altamimi Z, Rebischung P, Métivier L, Collilieux X (2016) ITRF2014: a new release of the international terrestrial reference frame modeling nonlinear station motions. JGR. <https://doi.org/10.1002/2016JB013098>
- BIPM (1998) The international system of units (SI), 7th edn. Bureau International des Poids et Mesures, Paris
- Bursa M (1992) Parameters of common relevance of astronomy, geodesy and geodynamics. Bull Geod 66(2):193–197
- Damour T, Soffel MH, Xu C (1991) General relativistic celestial mechanics: I. method and definition of reference systems. Phys Rev D 43:3273–3307
- Hase H (1999) Theorie und Praxis globaler Bezugssysteme. Mitteilungen des BKG 13:40
- Helmert FR (1880) Die mathematischen und physikalischen Theorien der Höheren Geodäsie, Band I. Teubner Verlag, Leipzig
- IERS (2010) IERS Conventions 2010. In: Petit G, Luzum B (eds) IERS technical note 36. Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt am Main. 179 pp. ISBN 3-89888-989-6
- IERS (2017) Altamimi Z, Rebischung P, Métivier L, Collilieux X Analysis and results of ITRF2014. (IERS Technical Note; 38). Frankfurt am Main: Verlag des Bundesamts für Kartographie und Geodäsie, 2017. 76 pp. ISBN 978-3-86482-088-5 (print version)
- Ihde J, Sacher M, Mäkinen J (2008) Conventions for the Definition and Realization of a European Vertical Reference System (EVRS) – EVRS Conventions 2007, BKG Frankfurt am Main
- IUGG (1991) Resolution on the conventional terrestrial reference system
- Kovalevsky J, Mueller II (1981) Comments on conventional terrestrial and quasi-inertial reference systems. In Gaposchkin EM, Kolaczek, B Kluwer (eds) Proceedings of the 56th Colloquium of the IAU, Sept 8–12, 1980. Academic Publication, Dordrecht
- Moritz H (1980) Geodetic Reference System 1980
- Moritz H (2000) Geodetic Reference System 1980. J Geodesy 74(1):128–162. <https://doi.org/10.1007/s001900050278>
- NIST Reference (2018) 2018 CODATA recommended values. <https://www.physics.nist.gov/cgi-bin/cuu/Value?bg>
- Plag H-P, Pearlman M (2009) Global geodetic observing system. Springer-Verlag Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-02687-4>
- Schneider M (1992) Himmelsmechanik – Band I, Grundlagen, Determinierung. BI Wissenschaftsverlag, Mannheim/Leipzig/Wien/Zürich, p 250
- Schneider M (1999) Himmelsmechanik – Band IV, Theorie der Satellitenbewegung, Bahnbestimmung. Spektrum Akademischer Verlag, Heidelberg/Berlin/Oxford, Kap. 67.2
- SIRGAS. <http://www.sirgas.org/index.php?id=55>
- Torge W (2003) Geodäsie. Walter de Gruyter, Berlin
- UN69/266 (2015) A global geodetic reference frame for sustainable development. http://ggim.un.org/documents/A_RES_69_266_E.pdf

Geodesy, Physical

V. Chakravarthi
Centre for Earth, Ocean and Atmospheric Sciences,
University of Hyderabad, Hyderabad, India

Definition

Geodesy is a branch of Earth sciences that deals with the measurement and representation of the Earth's shape and its

gravitational field. Geodesy may be broadly divided into the areas of global (satellite) geodesy, geodetic surveys (national and supranational), and plane surveying (topographic surveying, cadastral surveying). Global geodesy includes the determination of the shape of the Earth, its orientation in space, and its external gravity field. A geodetic survey is for the determination of the Earth's surface and gravity field over a region that typically spans a country or a group of countries. In plane surveying, the details of the Earth's surface are determined on a local scale, and thus the curvature and gravity effects are generally ignored. A close relation exists between the above three areas of geodesy in the sense that geodetic surveys are linked to reference frames established by global geodesy, whereas plane surveys are generally referenced to control points established by geodetic surveys. Physical geodesy utilizes measurements and characteristics of the Earth's gravity field as well as theories relating it to deduce the shape of the geoid (See ► [“Geoid Determination, Theory and Principles”](#); ► [“Geoid Determination, Computational Methods”](#)) and thus the figure of the Earth (Haffman and Moritz 2005). With sufficient information regarding the Earth's gravity field (See ► [“Gravity Field of the Earth”](#)), it is possible to determine geoid undulations (See ► [“Geoid Undulation, Interpretation”](#)), gravimetric deflections, the Earth's flattening and other broad characteristics of the Earth, such as equatorial ellipticity – the C22, S22 terms of the spherical expansion of the gravity field, the C21, S21 terms related to the position of the rotation axis, etc. with a view to their application in geodesy. Recent developments in physical geodesy includes satellite gravity missions (See ► [“Gravity Method, Satellite”](#)), which enables one to map many parts of the world with 1 cm detailed geoid model and determination of global gravity field to 1 mGal accuracy (Sjöberg 2015).

Historical Development of Geodesy

Geodesy is one among the oldest sciences dealing with the planet Earth. Superseding the use of sphere as a model to describe the shape of the Earth, the oblate rotational ellipsoid became widely accepted as the model of choice in the early eighteenth century. The significance of the gravity field was recognized in the nineteenth century leading to the introduction of the geoid (See ► [“Geoid”](#)). Later in the second half of the twentieth century, satellite techniques (See ► [“Gravity Field, Temporal Variations from Space Techniques”](#)) have brought out commendable advancements for realization of three-dimensional concept of geodesy. A drastic increase in the accuracy of geodetic observations required that time variations be taken into account, which in turn leads the development of four-dimensional geodesy. Smith (1986) had given a detailed description on the historical development of geodesy.

Gravity Field of the Earth

A body rotating with the Earth experiences the gravitational force of the masses of the Earth, other celestial bodies as well as the centrifugal force due to the Earth's rotation (See ► “Earth Rotation”). The resultant is the force of gravity (See ► “Gravity Field of the Earth”). If the effects of centrifugal force and that of celestial bodies are removed from gravity, the left-out component is gravitation. Newton's law of gravitation states that the force of attraction between two-point masses m_1 and m_2 is directly proportional to the product of their masses and inversely to the square of the distance separating the centers of respective masses and is given as

$$F = \frac{-Gm_1m_2}{r^2} \frac{k}{r}, \quad (1)$$

where, F is the force on m_2 , r is the distance between m_1 and m_2 , and G is universal gravitational constant. The minus sign represents that the force is always attractive. By setting the mass at the attracted point to unity, the above equation transforms into the gravitational acceleration

$$g = \frac{-Gm}{r^2} \frac{k}{r}. \quad (2)$$

The vector k may be expressed by the position vectors r and r' in the Cartesian coordinate system as

$$k = r - r', r^T = (X, Y, Z), r'^T = (X', Y', Z'),$$

with the magnitude of

$$k = \sqrt{(X - X')^2 + (Y - Y')^2 + (Z - Z')^2}.$$

However, the Earth is composed of an infinite number of differential mass elements, dm . The gravitation on the unit mass results from the integral over the individual contributions. The equation for gravitational acceleration then takes the form

$$g = -G \iiint \frac{r - r'}{|r - r'|^3} dm. \quad (3)$$

The mass element dm can be expressed as

$$dm = \rho du, \quad (4)$$

where, ρ is density of volume element du .

The representation of gravitational acceleration and related computations are simplified if gravity potential (scalar quantity) is used. Because the gravitational field is invariant to rotations

$$\text{curl} g = 0 \quad (5)$$

The vector g may be represented as the gradient of a potential U (Sigl 1985),

$$g = \text{grad} U. \quad (6)$$

For a point mass m , one can write

$$U = \frac{GM}{r}, \text{ with } \lim_{r \rightarrow \infty} U = 0. \quad (7)$$

For the Earth, the potential can be expressed as

$$U = G \iiint \frac{dm}{r} = G \iiint \frac{\rho}{r} du, \lim_{r \rightarrow \infty} U = 0. \quad (8)$$

The potential indicates that the work that must be done by gravitation in order to move the unit mass from infinity to the point under consideration.

From the above relations, it is clear that once the density function ρ is known for the Earth, the gravitation can be calculated as a function of position (See ► “Gravity Modeling, Theory and Computation”). In reality, detailed density information is available for the upper layers of the Earth only (See ► “Earth, Density Distribution”), while global models merely consider radial density changes. Consequently, gravity field observations have to be used in order to model the exterior gravity field (Torge 2001).

Properties of the Gravitational Potential

Gravity potential U satisfies Laplace's equation (Murthy 1998) i.e.,

$$\nabla^2 U = 0, \quad (9)$$

at all points unoccupied by matter. All the space derivatives of this potential also satisfy Laplace's equation. However, at points occupied by matter, gravity potential satisfies Poisson's equation

$$\nabla^2 U = -4\pi Gd, \quad (10)$$

where d is the density at the point under consideration. Green's theorem connecting volume and surface integrations of potential fields has many important applications. If U and U' are two potential functions, whose first derivatives are finite, continuous and single valued throughout a volume v bounded by a closed surface, s , then

$$\int_v (U \nabla^2 U' - U' \nabla^2 U) dv = \int_s \left[U \frac{\partial U'}{\partial n} - U' \frac{\partial U}{\partial n} \right] ds, \quad (11)$$

where n is the direction of the outward drawn normal to the surface. The theorem is independent of the size or shape of the surface or the nature of distribution of masses causing U and U' . If U' is assumed to be constant everywhere and U is the Newtonian potential due to masses distributed both inside and outside the closed surface, s , then

$$\nabla^2 U' = 0, \frac{\partial U'}{\partial n} = 0, \quad (12)$$

reducing equation to the form

$$\int_v \nabla^2 U dv = \int_s \frac{\partial U}{\partial n} ds. \quad (13)$$

If U_{in} and U_{out} are respectively the potentials due to masses inside and outside the surface, s , then U_{out} obeys Laplace's equation while U_{in} satisfies Poisson's equation, i.e.,

$$\int_s \frac{\partial U}{\partial n} ds = -4\pi GM. \quad (14)$$

This is called Gauss' law, which states that the total normal gravitational flux over a closed surface is $-4\pi G$ times the enclosed mass. This law is valid irrespective of the type of mass distribution within the closed surface, s . Basic relationships can therefore be established between the observations in the gravity field and the parameters describing the surface.

Disturbing Potential

The difference between the actual (measured) gravity potential, W , and the normal gravity potential (constructed by matching the known reference ellipsoid with an equipotential surface), U , is called the anomalous or disturbing potential, T (Hoffman and Moritz 2005). The disturbing potential T is numerically smaller than U or W , and captures the detailed, complex variations of the true gravity field of the Earth from point-to-point, as distinguished from the overall global trend captured by the smooth mathematical ellipsoid of the normal potential. The relation between the geoid undulation, N , disturbing potential, T , and the force of gravity (F) computed from the normal gravity potential can be expressed by Bruns' formula as

$$N = \frac{T}{F}. \quad (15)$$

The Stokes' formula which enables one in determining the geoid from gravity data in physical geodesy can be expressed as

$$N = \frac{R}{4\pi F} \iint_{\sigma} \Delta g S(\Psi) d\sigma. \quad (16)$$

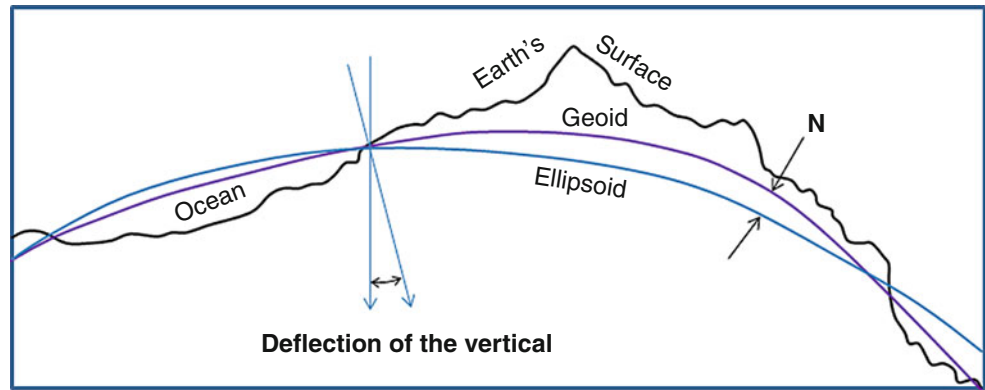
Here, Δg stands for gravity anomalies, and $S(\Psi)$ is the Stokes function derived by Stokes in a closed analytical form.

Geoid and Reference Ellipsoid

The equipotential surface of the Earth's gravity field that coincides with the Mean Sea Level (MSL) in the absence of currents, air pressure variation, etc. is called the geoid (See ► [“Geoid”](#)). Most of the efforts in physical geodesy are concentrated on the geoid determination with steadily increasing accuracy (Heiskanen and Moritz 1967, See ► [“Geoid Determination, Theory and Principles”](#)). The geoid is important for understanding more about ocean currents and is also used as a reference for traditional height systems and monitoring sea-level change. However, as in the case of any other equipotential surface of the Earth's gravity field the geoid surface is also an irregular surface and too complicated to serve as the computational surface. The mathematical surface that best fits the geoid is the reference ellipsoid. The relation between the geoid and the ellipsoid is shown in Fig. 1. The ellipsoid lies below the geoid surface in elevated regions and above the mean sea level (MSL) over the oceans. The geometrical separation between the geoid and the reference ellipsoid is called the geoid undulation, N (See ► [“Geoid Undulation, Interpretation”](#)), which generally varies globally between ± 110 m.

Gravity measurements are used in physical geodesy to determine the figure of the Earth (See ► [“Geodesy, Figure of the Earth”](#)). The justification in equating the Earth's shape to a model other than that of a sphere can be found from two observations, namely, (i) for a rotating sphere having uniform radius R and angular velocity ω the gravitational attraction at the equator and pole should differ by $R\omega^2$. For the values of R and ω relevant to the geoid, this figure works out to be 3.4 gals as against the observed value of 5.17 gals in gravity values at the equator and the pole. This discrepancy suggests an appreciable departure of the mean shape of the geoid from that of a sphere, also (ii) arc measurements at the pole and equator showed that the meridian degree at the pole is longer than that at the equator. The centrifugal acceleration causes the Earth's surface to flatten at the poles and to bulge at the equator. These effects would adequately explain the observed discrepancy in the difference of gravity at the equator and pole. It is thus reasonable to equate the Earth's shape to that of an oblate ellipsoid. The Earth's gravity field on the ellipsoidal surface varies only with the geographic latitude, φ , and is defined by the International Gravity Formula (Wilcox 1989, Also see ► [“International Gravity Formula”](#))

$$g(\varphi) = 978.0327(1 + 0.0053024 \sin^2 \varphi - 0.0000058 \sin^2 2\varphi) \text{ cm/s}^2.$$

Geodesy, Physical,**Fig. 1** Relation between Earth's topography, geoid and reference ellipsoid**Methods of In Situ Measurements**

Two distinctly different types of gravity measurements are made: absolute gravity measurements and relative gravity measurements. If the value of acceleration of gravity can be determined at the point of measurement directly from the data observed at that point, the gravity measurement is absolute (See ► [“Gravity Measurements, Absolute”](#)). If only the differences in the value of the acceleration of gravity are measured between two or more points, the measurements are relative (See ► [“Gravity Method, Surface”](#)). The most important gravity field quantities measured on the Earth's surface are gravity potential differences determined by leveling in combination with gravity measurement; the modulus of the gravity vector determined by absolute and relative gravity measurements; the direction of the gravity vector determined by geodetic-astronomical observations; and the components of the second-order gravity gradient tensor determined by torsion balance measurements.

Satellite Geodesy

Satellite geodesy utilizes artificial satellites and the moon as extraterrestrial targets or sensors. For a point mass Earth model, the orbital motion of a satellite is described by Kepler's laws. The actual gravitational field and nongravitational forces create orbital perturbations. Satellites used for geodetic applications differ in design, equipment, and orbit according to the respective observation technique and the mission purpose. Classical measurement methods based on orbital perturbations, introduced and employed from 1960s to 1980s demonstrated the efficacy of satellite observations for establishing large region geodetic control networks and gravitational field determination (See ► [“Gravity Method, Satellite”](#); ► [“Gravity Field, Temporal Variations from Space Techniques”](#)). Now the Global Positioning System (GPS) has become the most important method for a

fast and precise determination of geodetic positions (See ► [“GPS, Data Acquisition, and Analysis”](#); ► [“Geodesy, Ground Positioning, and Levelling”](#)). Today, the GPS governs three-dimensional positioning at all scales while laser distance measurements primarily contribute to global reference networks (See ► [“Geodesy, Networks, and Reference Systems”](#)). By monitoring the ocean surface, satellite altimetry contributes to gravity field modeling, and high-resolution global gravity recovery is now possible from satellite-to-satellite tracking and gravity gradiometry missions (e.g., GRACE and GOCE missions). The theory of satellite orbits and measurement methods was described in detail by Seeber (1993) and Kaula (1966).

Determining the Geoid

An intimate relationship exists between the height of the geoid, the Earth's gravitational attraction, and the distribution of mass within the Earth. The shape of the geoid does place important constraints on the mass distribution. Moreover, since the local vertical is defined as the direction of the gravitational attraction, which is always perpendicular to the geoid, geoid anomalies cause deflections of the vertical, which are of vital practical importance for surveyors. Determining the geoid is therefore an important objective in geodesy and geophysics.

In the presatellite era, the geoid was determined from the gravity field, which entailed measuring gravitational attraction or deflections of the vertical (or both) at a large number of locations around the Earth. Accuracy and resolution were limited by the number and distribution of observations that were feasible. Since the advent of artificial satellites, geoid determinations have become easier, more accurate, and of even higher resolution (See ► [“Geoid Determination, Computational Methods”](#); ► [“Geoid Determination, Theory and Principles”](#); ► [“Gravity Method, Satellite”](#)). Since the early 1980s, the mean sea level (or oceanic geoid) has been measured directly

by a radar-altimeter mounted on a satellite. It is important to remove the effects of waves, storms, ocean currents, and other similar effects, but ultimately this method gives a precision of a few centimeters in geoid height. Although they cannot be used over land, such measurements have yielded unprecedented high-resolution coverage of the geoid over the sea (Rapp 1982a, b; Sandwell and Schubert 1982; Cazenave and Monnereau 1987; Wunsch and Stammer 1998; Fu and Cazenave 2001). Besides, satellite gravity data is found to have tremendous applications, particularly in structural mapping, and modelling of large-scale geologic features in both onshore and offshore regions. However, it should be realized that the gravity data obtained from satellite altimetry is more reliable at longer wavelengths than at shorter wavelengths (Green et al. 2019). During 1980s and 1990s, Geosat and ERS-1 had been used to obtain high resolution marine geoid; and since 2010, the Cryosat-2 mission also has gained importance for measuring high resolution sea surface topography (Sandwell et al. 2014). In recent past, under the NASA Earth System Science Pathfinder (ESSP) Program the GRACE mission was developed in collaboration with the German Space Agency to accurately map the temporal variations in the Earth's gravity field. The mission launched in 2002 has provided the scientists an efficient and cost-effective way to map the Earth's gravity fields with high accuracy. The results from this mission has yielded crucial information about the distribution and flow of mass within the Earth and its surroundings. The gravity variations measured by GRACE has yielded important information on ice mass change over ice sheets and glaciers, ocean mass change due to land ice melt, and changes in terrestrial water storage over the continents (more details on the mission can be found at <http://www.csr.utexas.edu/grace>).

Further, satellite gradiometer mission (See ► “Gravity, Gradiometry”) named as Gravity field and steady-state Ocean Circulation Explorer (GOCE) was designed and launched by the European Space Agency (ESA) to determine gravity field anomalies with an accuracy of 1 mGal, and geoid with an accuracy of 1–2 cm. GOCE has provided gravity field data with unparalleled accuracy and spatial resolution during its life time between 2009 and 2013. The final gravity map and model of the geoid will provide users worldwide with a well-defined data product that will lead to a better understanding of the physics of the Earth's interior to gain new insights into the geodynamics associated with the lithosphere, mantle composition and rheology, uplift and subduction processes, a better understanding of the ocean currents and heat transport, a global height-reference system, to serve as a reference surface to study topographic processes and sea-level change, better estimates of the thickness of polar ice-sheets and their movement. Further details of the mission can be found at http://www.esa.int/esaLP/ESAYEK1VMOC_LPgoce_0.html.

Gravity Reduction

Gravity measurements provide values for the acceleration of gravity at points located on the physical surface of the Earth. Before these measurements can be used for most geodetic purposes, they must be converted into gravity anomalies. Gravity measured on the physical surface of the Earth must be distinguished from normal gravity referring to the surface of the reference ellipsoid. The measured gravity values show point to point variation on the Earth's surface because the measured field depends on latitude and elevation of the station and also due to the effects of topographic masses above sea level (Murthy 1998). The important corrections to the measured gravity data include latitude or normal correction, free-air correction, Bouguer correction, topographic correction, and isostatic correction (See ► “Gravity, Data to Anomalies”). The latitude correction calculates the normal gravity on the ellipsoidal surface, the free-air correction extrapolates it to the level of observation, the topographic and Bouguer corrections account for the gravity effect of the material between the geoid and the ground surface. Isostatic correction removes long-wavelength variations in the gravity field inversely related to topography. The magnitudes of these corrections more often equal to the measured field itself and therefore should be evaluated carefully and precisely.

Gravity Models

Two important models are in vogue (i) effect models and (ii) source models (Sunkel 1989). Effect models deals with the calculation of four parameters to define the best fitting ellipsoid to the shape of the Earth. These parameters are the product of the gravitational constant and the total mass of the Earth, a dimensionless dynamical form parameter in terms of a second degree spectral coefficient of the Earth's gravity field which is closely related to the flattening of the ellipsoid of revolution, the equatorial radius, and the angular velocity. However, the surface defined by the above parameters describes the global feature of the shape of the Earth and its gravity field (See ► “Geoid Determination, Computational Methods”; ► “Gravity, Global Models”). For more elaborate models, spherical harmonic coefficients have to be derived from surface and satellite data (Rapp and Cruz 1986). Ditmar et al. (2003) discussed the regularization issues of the computation of spherical harmonic coefficients of the gravity gradiometric data acquired by the GOCE satellite. On the other hand, source models presume a density model having radically changing but laterally constant mass density. Such a density model can be described by a few parameters and can therefore be compared with an ellipsoidal model.

Conclusions

The gravity field of the Earth, which forms the backbone for the studies relating to geometrical geodesy and physical geodesy, can be described in terms of potential, its derivatives both vertical and horizontal in several scales involving global, regional, and/or local information. The gradiometer missions will probably replace, even though not fully but to a large extent, the surface observations thereby pave the way for more detailed information on geoid with unprecedented accuracy.

Cross-References

- [Earth Rotation](#)
- [Earth, Density Distribution](#)
- [Geodesy, Figure of the Earth](#)
- [Geodesy, Ground Positioning, and Levelling](#)
- [Geodesy, Networks, and Reference Systems](#)
- [Geoid](#)
- [Geoid Determination, Computational Methods](#)
- [Geoid Determination, Theory and Principles](#)
- [Geoid Undulation, Interpretation](#)
- [Gravity Field of the Earth](#)
- [Gravity Field, Temporal Variations from Space Techniques](#)
- [Gravity Measurements, Absolute](#)
- [Gravity Method, Satellite](#)
- [Gravity Method, Surface](#)
- [Gravity, Data to Anomalies](#)
- [Gravity, Global Models](#)
- [Gravity, Gradiometry](#)
- [International Gravity Formula](#)

Bibliography

- Cazenave A, Monnereau M (1987) Seasat gravity undulations in the central Indian Ocean. *Phys Earth Planet Inter* 48:130–141
- Ditmar P, Kusche J, Klees R (2003) Computation of spherical harmonic coefficients from gravity gradiometry data to be acquired by the GOCE satellite: regularization issues. *J Geod* 77:465–477
- Fu LL, Cazenave A (2001) Satellite altimetry and earth sciences, a handbook of techniques and applications. Academic Press, London
- Green CM, Fletcher KM, Cheyney S, Dawson GJ, Campbell SJ (2019) Satellite gravity – enhancements from new satellites and new altimeter technology. *Geophys Prospect* 67:1611–1619
- Haffman BW, Moritz H (2005) Physical geodesy. Springer, Wien/New York
- Heiskanen WA, Moritz H (1967) Physical geodesy. W. H. Freeman, New York
- Kaula WM (1966) Theory of satellite geodesy. Blaisdell, Waltham, MA/Toronto/London
- Murthy IVR (1998) Gravity and magnetic interpretation in exploration geophysics. Geological Society of India, Bangalore

- Rapp RH (1982a) A global atlas of sea surface heights based on the adjusted Seasat altimeter data. Report 333. Ohio State University Department of Geodetic Science and Surveying, Columbus
- Rapp RH (1982b) A summary of the results from the OSU analysis of Seasat altimeter data. Report 335. Ohio State University Department of Geodetic Science and Surveying, Columbus
- Rapp RH, Cruz YJ (1986) Spherical harmonic expansions of the Earth's gravitational potential to degree 360 using 30' mean anomalies. Report 376. Ohio State University Department of Geodetic Science and Surveying, Columbus
- Sandwell DT, Schubert G (1982) Geoid height-age relation from SEASAT altimeter profiles across the Mendocino fracture zone. *J Geophys Res* 87:3949–3958
- Sandwell DT, Mueller RD, Smith WHF, Garcia E, Francis R (2014) New global marine gravity model from Cryosat-2 and Jason-1 reveals buried tectonic structures. *Science* 346(6205):65–67. <https://doi.org/10.1126/science.1258213>
- Seeber G (1993) Satellite geodesy. De Gruyter, Berlin/New York
- Sigl R (1985) Introduction to potential theory. Abacus Press, Cambridge
- Sjöberg LE (2015) The development of physical geodesy during 1984–2014 – a personal review. *J Geod Sci* 5:1–8
- Smith JR (1986) From plane to spheroid. Landmark Enterprises, Rancho Cordova
- Sunkel H (1989) Geodesy, physical. Encyclopedia of earth sciences series. Springer, Berlin
- Torge W (2001) Geodesy. Walter de Gruyter GmbH & Co, Berlin
- Wilcox LE (1989) Gravity anomalies: Interpretation. In: Geophysics. Encyclopedia of Earth Science. Springer, Boston, MA
- Wunsch C, Stammer D (1998) Satellite altimetry, the marine geoid, and the oceanic general circulation. *Ann Rev Earth Planet Sci* 26:219–253

Geodetic Pendulums, Horizontal Ultra Broad Band

Carla Braitenberg

Department of Mathematics and Geosciences, University of Trieste, Trieste, Italy

Synonyms

Ultra broad band long base tiltmeter

Definition

The ultra broad band horizontal geodetic pendulum is a device designed to measure the earth crustal movement in the range of periods typical of secular movements to seismic deformations. It measures the variation of the plumb line (vertical) with respect to the direction of two earth-fixed reference points (Zadro and Braitenberg 1999).

Instrument Description

The vertical pendulum is made of a mass suspended by a rod or wire that can oscillate in the vertical plane about a horizontal axis. For small amplitudes the oscillation period T depends on the moment of inertia K , the distance s of the center of mass from the rotation axis, the mass m , and the gravity acceleration g following the equation $T^2 = \frac{4\pi^2 K}{mgs}$. If the pendulum is allowed to oscillate about an inclined axis, the angle of the rotation axis with the vertical being ϕ , the period of oscillation is defined by $T^2 = \frac{4\pi^2 K}{\sin \phi mgs}$. For small values of the angle ϕ , the rotation axis is near to vertical, and the pendulum is called horizontal pendulum. Compared to the vertical pendulum of equal dimensions, the horizontal pendulum has an increased oscillation period. The first realizations of a horizontal pendulum were done in 1830 (Lorenz Hengler) and developed further by Zöllner (Zöllner 1871, 1872). The construction of Zöllner is made of a horizontal rod on which at one end a mass is attached. At the other end and at a small distance from it, the rod is connected by two wires to a housing structure (see Fig. 1). The two wires are near to vertical above each other. Supposing the upper fixed point is shifted with respect to the lower fixed point along the Meridian, so that the line connecting the two fixed points makes an angle ϕ with the vertical, then the rod has its equilibrium position in the Meridian plane. If the direction of the vertical changes due to the influence of the gravitational attraction of the moon or sun toward east by a small angle α , the rod will

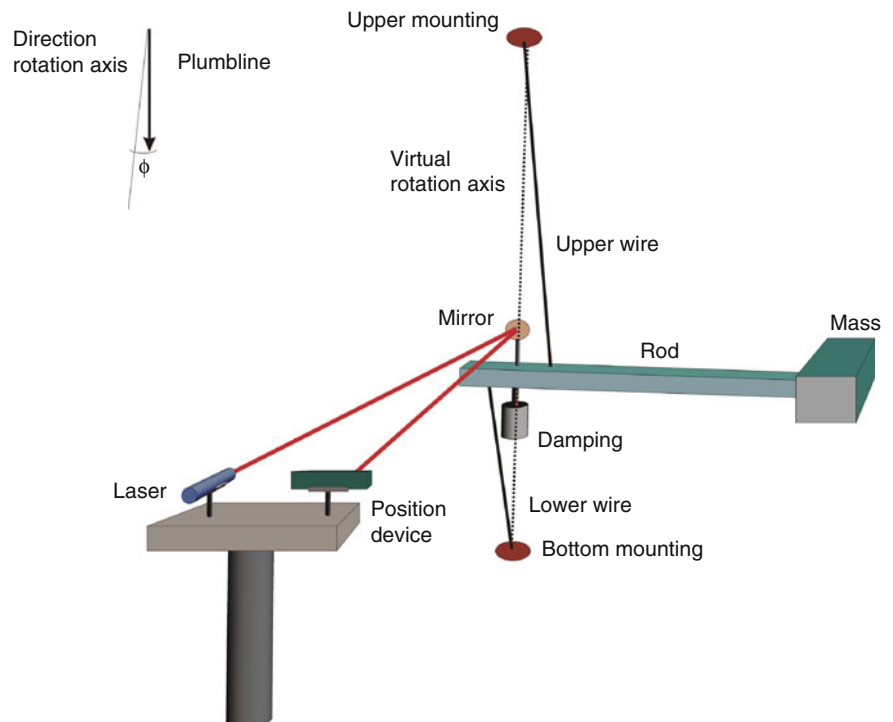
rotate toward east by the angle $\frac{\alpha}{\sin \phi}$. A similar rotation of the rod will be produced by the inclination of the housing structure or by a horizontal eastward movement of the housing structure as could be provoked by the passage of a seismic wave. In order to have an instrument sensitive in both NS and EW directions, usually a couple of orthogonally mounted pendulums is installed at one station. The movement of the rod is recorded by a space-sensitive device, as a magnetic induction coil mounted at the extreme of the rod or by an optical device that records the light ray reflected by a mirror mounted on the rod (e.g., Braitenberg et al. 2006).

Grotta Gigante Horizontal Pendulums

The largest prototype of a horizontal pendulum was built in 1959 by Antonio Marussi (Marussi 1959) in a natural cave (Grotta Gigante) situated in the Italian Karst at 275 m above sea level near the town of Trieste, Italy. Trieste is a port on the Adriatic Sea, 150 km East of Venice. The upper and lower wires of the horizontal pendulum-rod are fixed directly into the rock, at a vertical distance of 95 m. The oscillation period of the pendulum is near to 6 min, and damping is critical. The original recording of the pendulum was on a photographic paper, which was changed to a digital system in the year 2002. The instrument is very stable, and records multi-decennial crustal movements. The digital system has a sampling rate of 25 Hz, which makes it possible to observe fast movements of

Geodetic Pendulums, Horizontal Ultra Broad Band, Fig. 1

Cartoon of a horizontal geodetic pendulum. The rod with the mass rotates in the horizontal plane about the virtual rotation axis. The angle ϕ of the virtual rotation axis with the plumb line is essential for the amplification factor of the pendulum



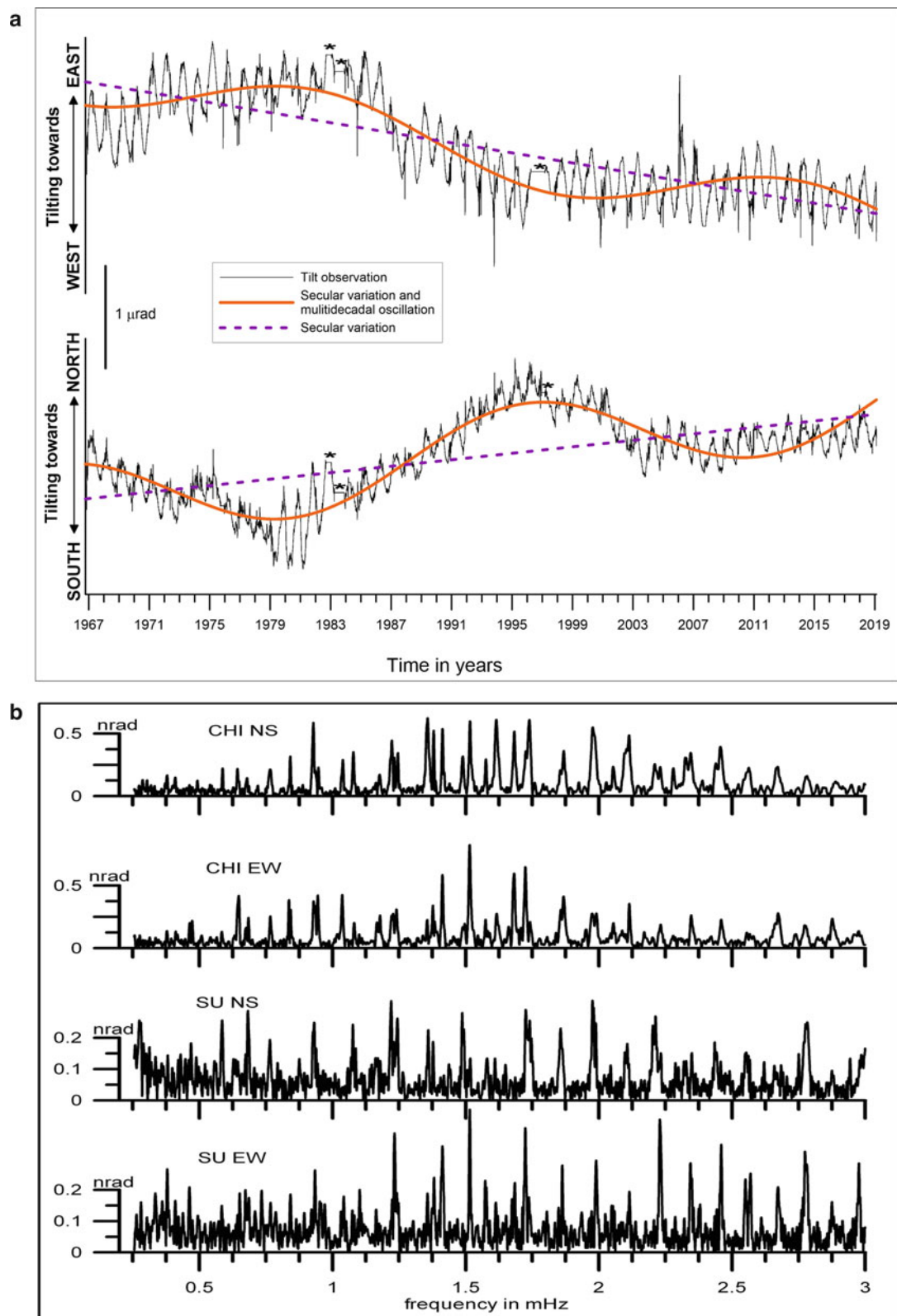
the pendulum caused by the passage of seismic waves. The smallest resolved tilt is 0.009 nrad (1 nrad = 10^{-9} rad).

The continuous record of tilt extends from 1966 to the present and demonstrates that the cave deforms continuously due to very different causes that act independently from each other. Each generates a characteristic signal contributing to the complex deformation of the cave. The graph in Fig. 2a displaying the entire tilt sequence shows a secular drift that adds to an oscillation with a period of 33 years. The long-term movement is a tilt of the cave toward NW, parallel to the coastline, and is probably tied to a deformation of the Adria plate that is bent below the load of the sediments in front of the South Alpine mountain chain. The cause of the multi-annual oscillation is presently unknown and could be tied to large-scale deformations imposed by the active deformation due to the collision of the Adria and the Eurasian plate, causing the seismicity at the contact between the sedimentary plain and the South Alpine mountains. The tilting of the cave with an annual period (mean period = 365.6 days and mean amplitude = 300 nrad for years 1966–2008; oscillation amplitude defined as half the peak-to-peak signal) has a characteristic orientation with extreme values in mid September and mid March. The annual tilting is due to the combined effect of the deformations due to the annual surface temperature variations (sinusoidal with an amplitude of 9.2 °C) and to the variable loading of the Adriatic Sea.

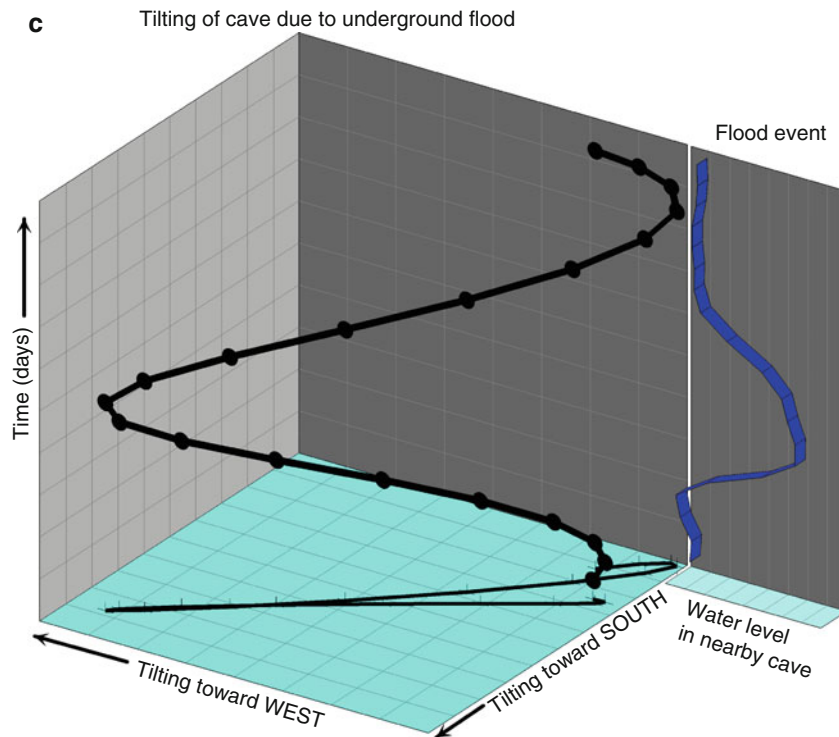
The earth tides generate a tilting that is due to the regular deformation of the earth in response to the gravitational attraction of moon and sun, with predominantly diurnal and semidiurnal periods. The amplitude of the earth tide tilting is 39 nrad, and the gravitational attraction and the load of the ocean tide contribute to an additional tilt signal at the same tidal frequencies but phase shifted with respect to the astronomical earth tide, leading to a total tidal tilting signal with a maximum amplitude of 90 nrad. The fault rupture of an earthquake leads to the generation of the well-known seismic waves that propagate along the earth surface and through earth's interior. Another movement caused by earthquakes is the oscillation of the earth at its resonance frequencies, called the free oscillations. The crustal rupture causes the earth to pulsate in the spherical and torsional oscillation modes. The gravest spheroidal mode 0S2 where the earth deforms with an ellipsoidal pattern has a 54-min period, the more complicated modes having shorter periods. The gravest torsional mode 0T2 has a period of 43.9 min, one earth hemisphere rotating with respect to the other around a central axis (Bolt and Marussi 1962). The geodetic pendulums of the Grotta Gigante station are the only instruments in the world that recorded the free oscillations generated by the greatest ($M = 9.5$, Chile, 1960; NEIC 2010), third greatest ($M = 9.1$, Sumatra Andaman Islands, 2004; NEIC 2010), and fifth greatest earthquake ($M = 8.8$; Chile, 2010; NEIC 2010) ever recorded in

instrumental seismology. The observations showed that the amplitude of the free oscillations of the Chile 1960 earthquake were up to four and a half times greater than those generated by the Sumatra-Andaman 2004 event (Braitenberg and Zadro 2007), the greatest enhancement being found for the torsional modes (see spectrum in Fig. 2b). The cave deforms not only due to periodic signals, similar to the ones we described before, but also due to transients, as pressure variations and underground water flow. The underground water flow is particularly interesting in karstic areas: The water from rainfall penetrates into the rock through pores and cracks until it finds an impermeable layer along which it flows downstream toward the sea. The Italian Karst is particular in this respect, as an entire surface river with an average flow of 9 m³/s reaching peaks of 100 m³/s during strong rainfalls, with extreme values up to 390 m³/s, disappears after its passage through a natural cave (Skocjanske Jame 2010). The underground river remains concealed for 33 km, until it emerges as a spring from the foot of the Karst and flows into the Adriatic Sea. The path of the flow is unknown, but its presence is identified by the geodetic pendulums. The cave responds in a characteristic direction to floods, with a signal that lasts in total between 4 and 8 days. The length of the deformation is indicative of the time the underground river takes to accommodate for the increased influx of water through the rainfall. The hydrologic-induced tilt signal is up to 300 nrad according to the existing records and is accompanied by uplift, as has been observed by GPS. The deformation is generated by an overpressure in the water-filled karst channel system (Braitenberg et al. 2019). Table 1 summarizes the different causes that generate a tilting of the cave together with the expected amplitude.

Besides the horizontal geodetic pendulums, other types of instruments exist with which tilt is observed. In the horizontal water tube, the differential water level at each end of the tube is measured by laser interferometry (e.g., Ferreira et al. 2006). The length of the tube extends over several tens of meters. Presently active instruments are found, e.g., at the Finnish Geodetic Institute, Metsähovi mine (Ruotsalainen 2008, 2018), and in California (Piñon Flat Observatory, University of California, San Diego). Vertical pendulums are mounted in boreholes for measuring tilt and the displacement of the pendulum is detected by a capacitive transducer with resolution of 1 nrad; examples are the tiltmeters of the Geodynamic Observatory Moxa, Thuringia, Germany. The application of a tilt array of this type to monitor pore pressure variations and minor earthquakes is given by Jahr et al. (2008). Portable short-base tiltmeters are used for volcanic monitoring, where the expected tilts are larger than for crustal deformation studies and therefore require less resolution and stability in time. The electronic tiltmeters record the movement of an air bubble in a conducting fluid. These instruments are used for volcanic monitoring of the USGS.



Geodetic Pendulums, Horizontal Ultra Broad Band, Fig. 2 (continued)



Geodetic Pendulums, Horizontal Ultra Broad Band, Fig. 2 The graph shows the continuous time series of tilt recorded by the horizontal geodetic pendulum of Grotta Gigante. (a) *Black continuous lines* are the NS and EW records with daily sampling. The regular oscillation is the annual tilting. The multi-decadal variation is the sum of an oscillation of 31 years (*thick line*) and a linear variation (*dashed line*). The spikes are due to tilting caused by floods of the underground karstic water flow. Asterisks mark data interruptions. (b) Spectral amplitudes in the

frequency range of the free oscillations following the earthquakes of Chile 1960 (CHI-NS, CHI-EW) and Sumatra-Andaman 2004 (SU-NS, SU-EW). (c) The tilting of the cave due to the underground flooding: The *thick line* shows tilting of the cave during the flood and the z-axis is the time. The *thinner line* shows the tilting during the flooding in the EW and NS direction. Furthermore, the water level in a nearby cave during a flood is shown

Geodetic Pendulums, Horizontal Ultra Broad Band, Table 1 Identified causes that deform the Grotta Gigante cave with characteristic periods and amplitudes

Cause of deformation	Characteristic period	Tilt maximum observed amplitude	Notes
Plate tectonics secular movement	Secular movement, near to static	31 nrad/year	
Plate tectonics multi-annual oscillation ^a	31 years	480 nrad	
Annual variation ^a	365.6 days	300 nrad	Orientation toward N60°E in mid March and toward N240°E in mid September
Earth tides ^a	Daily, half daily	39 nrad	
Ocean loading ^a	Daily, half daily	60 nrad	
Free oscillations ^a of the earth	54 min to a few minutes	Up to 0.7 nrad	
Underground water flow	A few days	Up to 300 nrad	

^aOscillation amplitude defined as half of the peak-to-peak excursion

Summary

The ultra broad band horizontal geodetic pendulum measures the inclination of a reference axis with respect to the vertical. The prototype installed in the Giant Cave in Northeastern Italy has shown that the deformations that affect the upper

crust are due to tectonic effects (seismic waves, free oscillations of the earth, coseismic deformation, tectonic plate deformation) and due to environmental effects (thermoelastic deformation, hydrologic water flow, ocean loading). The existing record of the greatest seismic event ever measured, the Chile 1960 earthquake, allows to make an absolute

magnitude comparison with recent events due to the fact that amplitude factor, the instrumental setup, and the location are perfectly known and controllable today.

Cross-References

- Earth Tides
- Free Oscillations of the Earth
- Great Earthquakes
- International Geophysical Year
- Seismic Instrumentation

Bibliography

- Bolt BA, Marussi A (1962) Eigenvibrations of the earth observed at Trieste. *Geophys J R Astron Soc* 6:299–311
- Braitenberg C, Romeo G, Taccetti Q, Nagy I (2006) The very-broad-band long-base tiltmeters of Grotta Gigante (Trieste, Italy): secular term tilting and the great Sumatra-Andaman islands earthquake of December 26, 2004. *J Geodyn* 41:164–174
- Braitenberg C, Zadro M (2007) Comparative analysis of the free oscillations generated by the Sumatra-Andaman islands 2004 and the Chile 1960 earthquakes. *Bull Seismol Soc Am* 97:S6–S17. <https://doi.org/10.1785/0120050624>
- Braitenberg C, Pivetta T, Barbolla DF, Gabrovsek F, Devoti R, Nagy I (2019) Terrain uplift due to natural hydrologic overpressure in karstic conduits. *Sci Rep* 9(3934):1–10. <https://doi.org/10.1038/s41598-019-38814-1>
- Ferreira AMG, d'Oreye NF, Woodhouse JH, Zürn W (2006) Comparison of fluid tiltmeter data with long-period seismograms: surface waves and Earth's free oscillations. *J Geophys Res* 111(B11):B11307/17
- Jahr T, Jentzsch G, Gebauer A, Lau T (2008) Deformation, seismicity, and fluids: results of the 2004/2005 water injection experiment at the KTB/Germany. *J Geophys Res* 113:B11410. <https://doi.org/10.1029/2008JB005610>
- Marussi A (1959) The University of Trieste station for the study of the tides of the vertical in the Grotta Gigante. In: *Proceedings of the III international symposium on earth tides*, Trieste, 1960, pp 45–52
- NEIC (2010) National Earthquake Information Center. <http://earthquake.usgs.gov/>
- Ruotsalainen HE (2008) Recording deformations of the earth by using an interferometric water level tilt meter. In: Korja T, Arhe K, Kaikkonen P, Korja A, Lahtinen R, Lunkka JP (eds) *Fifth symposium on the structure, composition, and evolution of the lithosphere in Finland*, Institute of Seismology, University of Helsinki, report S-53, pp 103–106
- Ruotsalainen HE (2018) Interferometric water level tilt meter development in Finland and comparison with combined earth tide and ocean loading models. *Pure Appl Geophys* 2018:175. <https://doi.org/10.1007/s00024-018-1875-0>
- Skocjanske Jame (2010). <http://www.park-skocjanske-jame.si/>
- Zadro M, Braitenberg C (1999) Measurements and interpretations of tilt-strain gauges in seismically active areas. *Earth-Sci Rev* 47:151–187
- Zöllner KF (1871) Über einen neuen Apparat zur Messung anziehender und abstoßender Kräfte, Kgl. sächs. Gesellsch. der Wissensch. zu Leipzig, 27. Nov. 1869 und 1. Juli 1871
- Zöllner KF (1872) Zur Geschichte des Horizontalpendels, Kgl. sächs. Gesellsch. der Wissensch. zu Leipzig, math.-phys. Klasse, November 1872

Links to Geodynamic Observatories with Active Tiltmeters

- Finnish Geodetic Institute.: http://www.fgi.fi/tutkimus/tiedot_aiheesta_eng.php?projekti=19
- Geodynamic Observatory Moxa, Germany.: http://www.geophysik.uni-jena.de/en/Institutions+_Labs/Observatory+Moxa.html
- LSBB – Low Noise Underground Laboratory, Rustrel, France.: <http://lsbb.eu/recherche/onde-environnement-radiatif/interferometrie-laser-inclinometrie/>
- Plate Boundary Project, Earthscope, USA., <https://www.unavco.org/projects/past-projects/pbo/pbo.html>
- St. Croix long-base tiltmeters, France.: <http://igets.u-strasbg.fr/stecroix.php>
- University California San Diego, USA.: <http://pfotrains.ucsd.edu>
- USGS volcano monitoring with tiltmeters, USA.: <http://volcanoes.usgs.gov/activity/methods/deformation/tilt/index.php>

Geodynamics

Alessandro M. Forte

Department of Geological Sciences, University of Florida, Gainesville, FL, USA

Definition

Geodynamics (noun, functioning in the singular). A broad, multidisciplinary branch of geophysics that investigates the causes and consequences of the forces acting on and inside the solid Earth and the rock movements and deformations they produce.

Geodynamic, Geodynamical (adjective). Relating or pertaining to the field of geodynamics.

The term geodynamics originates from two root words: “geo,” from the Greek *gē* meaning “earth,” and “dynamics,” from the Greek *dynamis* meaning “power.” Dynamics is a branch of physics, based on Newton’s second law of motion, which involves the study of forces acting on bodies and the resulting changes in their motion. As the name therefore suggests, the focus of geodynamics is on the forces that act on the solid Earth and the motions and deformation they produce.

Historically, in the late nineteenth and early twentieth century, when geodynamics was emerging as a distinct field of study, it also included the earliest systematic investigations of earthquakes and propagating seismic waves. These studies later evolved into a separate branch of geophysics called seismology. Modern geodynamics investigates forces and movements over time scales that exceed the characteristic periods of propagating earthquake waves and corresponding Earth oscillations (i.e., approximately ranging from a second to an hour). On longer time scales, for geodynamic processes ranging from Earth tides (hours to years) to glacial isostatic

adjustment (thousands of years) and ultimately plate tectonics (millions of years), the changes in applied forces and the resulting deformations are sufficiently slow that acceleration is entirely negligible (i.e., the deformation of the solid Earth is quasistatic).

The forces studied in geodynamics mainly originate from lateral changes in density $\Delta\rho$ on or within the Earth, thereby producing buoyancy or Archimedes forces $\Delta\rho g$ per unit volume, where g is Earth's gravitational field. For example, lateral density variations in anomalously hot or cold regions in the convecting mantle give rise to buoyancy forces $\Delta\rho g = -\alpha\rho\Delta Tg$, where α is the coefficient of thermal expansion and ΔT is the temperature anomaly relative to the ambient mantle. These thermal buoyancy forces drive the convective circulation in the mantle and the corresponding surface motions of the tectonic plates. Similarly, the topographic depressions Δh at the Earth's surface left behind by the rapid melting of the massive ice sheets that once covered Canada and Scandinavia about 20,000 years ago give rise to buoyancy forces $\Delta h\rho g$ per unit area that act on the surface (here ρ is the average density of the crust). The subsequent glacial isostatic response to this surface force involves slow viscous creep of the mantle and the resulting uplift of the crust in formerly ice-covered regions.

There are also external forces $\rho\nabla\varphi$ where, for example, $\nabla\varphi$ is the gradient of the gravitational potential generated by the attraction of the Sun and Moon, or it may represent the centrifugal force due to Earth's diurnal rotation. For the former case, these forces produce the tidal deformations in the Earth, or, for the latter case, they produce the oblate ellipsoidal shape (i.e., pole-to-equator "flattening") of the Earth. These tidal and rotational forces change over an enormous range of time scales: from 12 h to 18.6 years for tidal forcing and from thousands to hundreds of millions of years for rotational forcing. Tidal and rotational forces produce globally distributed deformations throughout the Earth's interior and at the surface.

A major area of geodynamics research is the measurement and quantitative modeling of forces and corresponding deformations in the crust, lithosphere, and deep mantle. An understanding of the relationship between force and deformation requires detailed knowledge of the rheology of the solid Earth that depends on the ambient temperature, pressure, and the time scales over which the forces are applied. For example, on time scales corresponding to the tides, the solid Earth behaves as an imperfectly elastic body. However, on the much longer time scales of mantle convection and plate tectonics, the solid mantle behaves as a fluid characterized by an effective viscosity that is extremely high.

The effective viscosity η of the mantle is one of the most important parameters in geodynamics because it controls the speed at which the mantle may flow and deform in response to buoyancy forces. The characteristic flow speed is

$v = \Delta\rho gLD/\eta$ where L represents the lateral dimension (horizontal distance) of the density anomalies and D is their vertical dimension (depth extent). Flow velocities in the mantle are of the order of several centimeters per year, similar to those of the moving surface plates. This flow velocity determines the efficiency of convective heat transport across the mantle and hence governs the thermal (and chemical) evolution of the Earth. For this reason, a great deal of work in the second half of the twentieth century, and up to the present day, is focused on constraining the value and spatial variation of mantle viscosity from the analysis of various geodynamic data sets (e.g., from postglacial uplift rates, geoid anomalies, long-term sea level variations, plate motions) as well as from theoretical and experimental studies of the creep properties of mantle minerals (e.g., Karato 2012).

The range of geophysical phenomena studied in geodynamics has increased greatly over the course of the twentieth century. For example, an early treatise on geodynamics by Love (1911) provided a detailed mathematical description of a range of problems that included the isostatic support of continents, the height of mountain chains and the forces required to maintain them, the Earth tides, the propagation of earthquake waves, and the oscillations of the entire Earth produced by earthquakes. With the advent of plate tectonics and the acceptance that some form of thermal convection in the mantle is necessary to drive the plate motions, the scope of geodynamical investigations has been greatly enlarged.

Modern geodynamics research deals with a multitude of problems relating to plate tectonics, heat and mass transfer and chemical mixing in the solid Earth, the rheology and dynamical coupling of the mantle and crust-lithosphere systems, and the impact of internal dynamics on surface geophysical and geological processes. The wide range of problems now included in contemporary geodynamics is well illustrated by Turcotte and Schubert (2014). Depending on the specific problem, it is difficult, if not impossible, to devise analogue laboratory experiments that can properly incorporate the complex spatial variations of the physical properties of the solid Earth. This difficulty is further compounded when it is necessary to model geodynamic processes over extremely large spatial and temporal scales, as in the case of mantle convection and plate tectonics.

To address these major challenges, the rapidly growing field of *computational geodynamics* has emerged, in which complex problems in solid Earth dynamics are now routinely solved on high-performance computers using numerical methods (e.g., Ismail-Zadeh and Tackley 2010; Gerya 2019). The need to develop robust, reliable, and openly accessible implementations of advanced computational algorithms to solve outstanding problems in geodynamics has motivated the growth of community-driven consortia of researchers, such as the "Computational Infrastructure of

Geodynamics” (<https://geodynamics.org/>) that maintains and disseminates open-source software packages for a wide range of applications in crystal dynamics, tectonics, mantle convection, and magnetohydrodynamics in the core.

Cross-References

- [Continental Drift](#)
- [Earth Tides](#)
- [Earthquakes and Crustal Deformation](#)
- [Geoid](#)
- [Gravity Field, Temporal Variations from Space Techniques](#)
- [Isostasy](#)
- [Lithosphere, Continental](#)
- [Lithosphere, Continental: Thermal Structure](#)
- [Lithosphere, Mechanical Properties](#)
- [Lithosphere, Oceanic](#)
- [Lithosphere, Oceanic: Thermal Structure](#)
- [Mantle Convection](#)
- [Mantle Viscosity](#)
- [Plate-Driving Forces](#)
- [Seismology, Global Earthquake Model](#)
- [Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity](#)

Bibliography

- Gerya T (2019) Introduction to numerical geodynamic modelling. Cambridge University Press, Cambridge
- Ismail-Zadeh A, Tackley P (2010) Computational methods for geodynamics. Cambridge University Press, Cambridge
- Karato S-i (2012) Deformation of earth materials: an introduction to the rheology of solid earth. Cambridge University Press, Cambridge
- Love AEH (1911) Some problems of geodynamics. Cambridge University Press, Cambridge
- Turcotte DL, Schubert G (2014) Geodynamics, 3rd edn. Cambridge University Press, Cambridge

Geoelectromagnetism

Antal Adam and Laszló Szarka
Geodetic and Geophysical Research, Institute of the
Hungarian Academy of Sciences, Sopron, Hungary

Definitions and Scope

Geoelectromagnetism, in the broad sense, includes the application of classical electrodynamics to various interrelated regions of space (Sun, interplanetary field, magnetosphere, ionosphere, atmosphere), on and below the Earth surface

(near-surface regions, crust, upper mantle [or lithosphere and asthenosphere using alternative nomenclature], lower mantle, and Earth’s core). Over the Earth surface the dominant phenomena of geoelectromagnetism are electromagnetic waves, in a wide spectrum. The electromagnetic waves do not penetrate deep into the Earth. Instead, the propagation of the electromagnetic field inside the Earth is a diffusion process.

In a narrower sense, geoelectromagnetism means the application of time variations of natural electromagnetic fields to the study of the electric resistivity distribution in the Earth’s interior, which gives indispensable information about the structure, composition, and processes of the subsurface.

The term electromagnetic geophysics is also used, but electromagnetic geophysics in general includes all exploration techniques, which may employ not only natural sources, but various man-made sources as well. In order to investigate very deep structures, extremely large-size sources of natural electromagnetic field variations should be used. In the following, we restrict ourselves to those geoelectromagnetic methods that apply natural electromagnetic field variations as the source to study deeper (lithosphere, asthenosphere, and upper mantle) geological structures.

Introduction

Of three electromagnetic parameters (magnetic permeability, electrical resistivity, and permittivity) associated with the subsurface geology, only the magnetic permeability and the electrical resistivity are relevant because of the diffusive nature of the electromagnetic field propagation. The magnetic permeability is usually assumed to be constant, which – apart from some exceptional cases – has been seen to be a good assumption. The electrical conductivity s of various earth materials varies over an extremely wide range; therefore, its spatial distribution reflects most structural/physical/petrological changes in the Earth. In practice, instead of the conductivity, the electrical resistivity ρ , where $\rho = 1/\sigma$, is often used. The ohm-meter (or Ωm) is the unit of measurement.

These methods are classified on the basis of the measured geoelectromagnetic field components. When only the electric field variations are measured, we speak about the *telluric* or *Earth current* technique, or *tellurics* (TT) for short. If only the geomagnetic field components are measured (usually as a function of the period of the time variations of the electromagnetic field), the method is known as “*Geomagnetic Deep Sounding*” (*GDS* for short), or by its other name: “*MagnetoVariation Sounding*” (MVS). Electrical conductivity anomalies (i.e., the zones of concentration of electric current) can be located by using the so-called “*MagnetoVariation Profiling*” (MVP), where the so-called induction vectors point to the current concentrations. A more sophisticated technique is the “*Magnetotelluric Sounding*” (MTS) or

shortly “*magnetotellurics*” (MT), where both the electric and magnetic field variations are measured in a broad period range (see ► “[Magnetic Data Enhancements and Depth Estimation](#)”).

Although the Geomagnetic Deep Sounding (the MagnetoVariation Sounding) is more or less coeval with observatory records (see, e.g., Lamb 1883), the history of almost all these variants dates back to about six decades. The MVP and its induction vectors were introduced in the second half of 1950s in studying the anomalous behavior of geomagnetic variations in North Germany (e.g., Schmucker 1959; Wiese 1965). The theoretical basis of the magnetotelluric (MT) method was laid by Tikhonov (1950) and Cagniard (1953). The first MT field measurements were carried out in the early 1960s. Application of the tellurics (i.e., the simplest technique) to investigate large sedimentary basins started in the middle of the twentieth century in Hungary (Kántás 1954) and elsewhere.

In the following, the principles of these geoelectromagnetic methods will be presented. We summarize the governing equations of MVS, MVP, and MT techniques and illustrate their field application. The telluric measurements being the special case of the magnetotelluric method used in the so-called S-interval (see Fig. 4) where the phase approximates zero, is not discussed here in detail.

Magnetovariational Sounding (MVS) and Magnetovariational Profiling (MVP)

By simultaneously measuring the time variations of the horizontal and vertical geomagnetic field components at a site, it is possible to determine the subsurface electrical resistivity. Carrying out the measurements in a broad period interval, the depth-dependence of the electrical resistivity can be estimated (see ► “[Magnetic Data Enhancements and Depth Estimation](#)”).

For a spherical Earth, this technique is able to provide information about the Earth’s conductivity down to the depths of 1,200–1,800 km in the mantle (Bailey 1973; Rokityansky 1982). For shallower studies (i.e., down to depths of less than approximately 100 km) the Earth’s surface may be considered as plane.

The magnetovariational profiling (MVP) aims at locating large lateral inhomogeneities of the Earth’s crust and upper mantle. In the recent book by Berdichevsky and Dmitriev (2008), this theme is discussed in detail under the title “*Magnetovariation Studies*.”

Deep Mantle MVS Studies

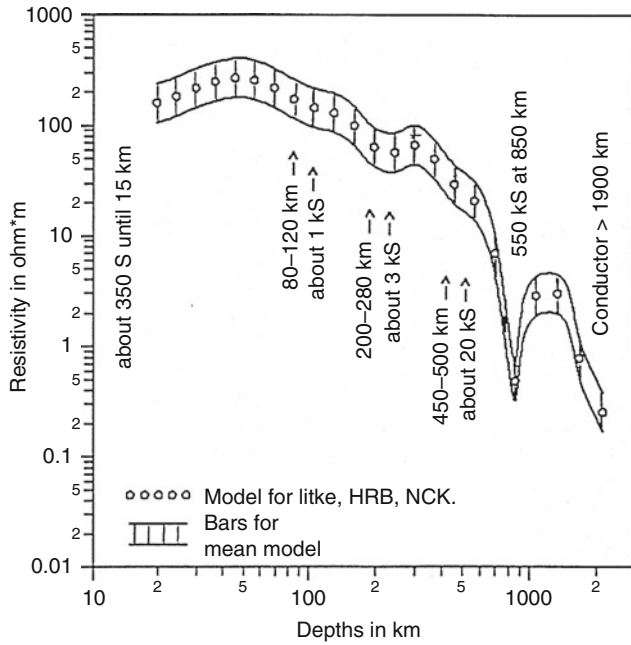
We follow Rokityansky (1982) nomenclature concerning the deep mantle studies. The field source of so-called Dst geomagnetic variations (used recently by Yu Semenov et al. 1997; Yu Semenov et al. 2008) is the ring current at the

equator. When the vertical magnetic field of the ring current source is approximated by the first term in its spherical harmonic expansion, that is, $H_z \tilde{P}_1^0(\cos \theta)$ where $\tilde{P}_1^0$ is the associated Legendre function in the first term of the expansion, the ratio of vertical (H_z) to horizontal ($\sqrt{H_x^2 + H_y^2}$) field component can be converted into the complex apparent resistivity (Rokityansky 1982). (In standard practice, x and y are the horizontal components, with the x -axis pointing northward, and the positive z -axis directed vertically downwards into the ground.) If the magnetic permeability (μ), the radius of the Earth (R), the geomagnetic colatitude of an observatory (Θ), and the angular frequency (ω) in the assumed time-dependence $e^{i\omega t}$ are known, then the apparent resistivity ρ is given as

$$\rho(\omega) = \frac{i\omega\mu}{4} R^2 \tan^2 \Theta \left| \frac{H_z(\omega)}{\sqrt{H_x^2 + H_y^2}} \right|^2.$$

Yu Semenov et al. (1997) used the daily mean values of geomagnetic observatories. Their first step in the analysis was a rotation of the observatory reference systems to determine the maximum Dst field (Schultz and Larsen 1983). For the agreement between the horizontal Dst field and the first spherical harmonic model of the source, a coherence test was carried out. The period ranges, which were in contradiction with the source model, were excluded from further analysis. The transfer functions between the vertical (H_z) and horizontal $\sqrt{H_x^2 + H_y^2}$ components were computed only if the coherence values were greater than 0.7 between them. The transfer functions were converted into apparent resistivity. Yu Semenov et al. (1997) combined the magnetovariational data with the results of magnetotelluric sounding in observatories and estimated, by using a robust technique, the resistivity model below the observatories. The mean geoelectric structure of the mantle below the Pannonian Basin is shown in Fig. 1 on the basis of data from Nagycenk, Litke (Hungary), and Hurbanovo (Slovakia). A more detailed combined MV + MT study in East-Central Europe pointed out a significant difference in the Precambrian and Phanerozoic mantle structures, separated by the Tornquist-Teisseyre Zone (TESZ line; Yu Semenov et al. 2008).

Schmucker (2003) proposal for the generalized horizontal spatial gradient method (gSHG) is promising for a reliable combination with the MT method as stated recently by Vozar and Semenov (2010) modeling. Kelbert et al. (2009) inverted long period geomagnetic response functions to derive a global-scale 3D model of electrical conductivity variations in the Earth’s mantle. Their results – one order magnitude conductivity increase in the subduction zones – support the presence of at least some water in these transition zones.



Geoelectromagnetism, Fig. 1 The mean geoelectrical structure of the mantle in the Pannonian Basin using the magnetic records of Nagycenk, Litke (Hungary), and Hurbanovo (Slovakia) for magnetovariational soundings (Yu Semenov et al. 1997)

Magnetovariational Profiling

It is evident from the Biot–Savart law that on the surface, just over and near the center of a horizontal lineament of high electrical conductivity in the subsurface, in which the telluric currents are gathered, the measured vertical magnetic component H_z is zero. More precisely, H_z is changing sign along a profile, crossing the lineament. At the same time, the horizontal component $\sqrt{H_x^2 + H_y^2}$ has a maximum over the lineament, with a moderate decrease on both sides. Therefore, the $H_z/\sqrt{H_x^2 + H_y^2}$ ratio can be drawn as a vector, that is, a quantity having magnitude and direction. It is called “induction vector” or “induction arrow.” Along a measuring profile crossing the lineament, the induction arrow will change sign. Induction arrow profiles at various periods provide useful depth information. The origin of the method goes back to the resolution of a paradox, namely when the vertical component of substorms in two German geomagnetic observatories (Wingst and Niemegk) proved to be of opposite sign. This paradox was attributed later to the North German anomaly (Schmucker 1959).

The source of the method is the same as that of all other geoelectromagnetic methods and a similar time series analysis is needed to separate the variations of different periods (frequencies) for single station and array measurements too. Array and single site transfer functions are applied.

Array Transfer Function

The normal field is that which would exist in the presence of one-dimensional layered structure in which the inhomogeneity is embedded. The anomalous field is then most simply defined as the total measured field minus the normal field.

Due to the linearity of the Maxwell field equations, there exists a linear relationship between the anomalous and normal magnetic field components (Schmucker 1970).

$$H_{xa} = T_{xx}H_{xn} + T_{xy}H_{yn} + T_{xz}H_{zn} + \epsilon_x$$

$$H_{ya} = T_{yx}H_{xn} + T_{yy}H_{yn} + T_{yz}H_{zn} + \epsilon_y$$

$$H_{za} = T_{zx}H_{xn} + T_{zy}H_{yn} + T_{zz}H_{zn} + \epsilon_z$$

where H_x , H_y , and H_z are the field components and ϵ is the error of the relation. All these T quantities are period dependent and complex. Subscript “a” and “n” refer to anomalous and normal parts, respectively. The assumption that the source field is uniform, allows the normal part of the H_z component to be taken as zero. Accordingly, the third term on the right-hand side of equations can be neglected.

Single site transfer function With the following condition

$$H_{zn} \ll H_{za}, H_{xa} \ll H_{xn}, H_{ya} \ll H_{yn}$$

the relation between the vertical and horizontal components is

$$H_{za} = T_{zx}H_x + T_{zy}H_y,$$

where T_{zx} and T_{zy} are complex values.

Induction vectors (arrows) are computed separately from the real and imaginary parts of the T values. In practice two variants are used (Wiese 1965; Parkinson 1959). The vector lengths in both cases are the same:

$$L_r = \sqrt{(\text{Re } T_{zx})^2 + (\text{Re } T_{zy})^2},$$

$$L_i = \sqrt{(\text{Im } T_{zx})^2 + (\text{Im } T_{zy})^2}.$$

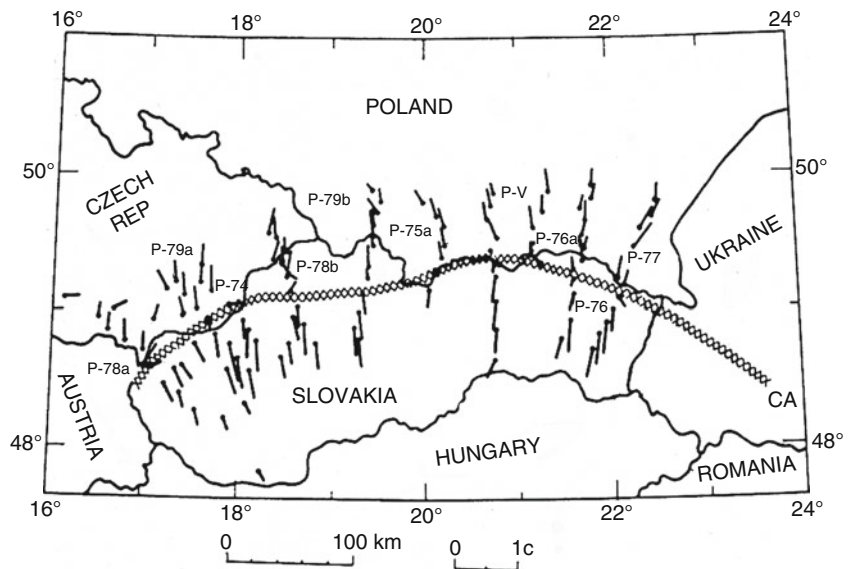
The azimuth of the Wiese vector (arrow)

$$\phi_w = \tan^{-1} \left(\frac{T_{zy}}{T_{zx}} \right)$$

gives the direction of the maximum correlation between the upward vertical and horizontal field component. This definition of the azimuth implies that the x-axis is directed northwards. The magnitude (length) represents the ratio of vertical field amplitude to the amplitude of the horizontal component with which it correlates (Arora 1997). The Wiese vectors point away from areas of higher conductivity. Conversely, the azimuth of the Parkinson arrow points towards the high electrical conductivity region. The

Geoelectromagnetism,

Fig. 2 Wiese induction vectors measured along the Carpathian conductivity anomaly. (Modified after Jankowski et al. 2005)



induction arrows when displayed on the map clearly identify the location and trend of subsurface conductivity anomalies (Arora 1997).

The MVP (induction vector) has an extended literature. Rokityansky (1982) provided a series of model studies. Arora (1997) published several alternative approximations, which might better determine the structure of the conductive bodies, for example, “perturbation arrows,” “hypothetical event analysis,” Z/H “pseudosection,” and 2D modeling.

Among the field studies conducted in East-Central Europe in which induction vectors were plotted, it is the Carpathian conductivity anomaly (indicated at first by Wiese 1965) that has received the greatest attention, attracting scientists from all neighboring countries. The distribution of the induction vectors is shown in Fig. 2 (after Jankowski et al. 2005) and one of the best models is shown in Fig. 3 (Zhdanov et al. 1986). For the assumed conductive region between 10 and 20 km there are two alternative hypotheses: (a) porous rocks saturated by mineralized water (ionic conduction), (b) conducting rock complex connected to graphite.

At present the induction vectors form an integral part of routine MT interpretation software (e.g., WinGlink) and the vertical magnetic components are also incorporated in MT inversion.

Magnetotelluric Soundings

Basic Formulas

Magnetotellurics (MT) means a joint measurement of the natural electric and magnetic field variations on the surface. Its energy source is partly the same as that of MVS and MVP. Namely, at frequencies $f < 5$ Hz the electromagnetic

(EM) source field is generated by ionospheric as well as magnetospheric phenomena (micropulsations, substorms, etc.) and at frequencies $f > 5$ Hz the EM field is due to worldwide lightning. The source field penetrates into the Earth, where it is reflected from all resistivity interfaces. The measured field is the sum of the primary (source) and secondary (subsurface-dependent) fields. The signal propagation in the subsurface is diffusive. *Skin depth* δ is the depth at which a uniform field is attenuated to $1/e$ of its surface amplitude in a conductor of uniform conductivity:

$$\delta = \sqrt{2/\mu\sigma\omega},$$

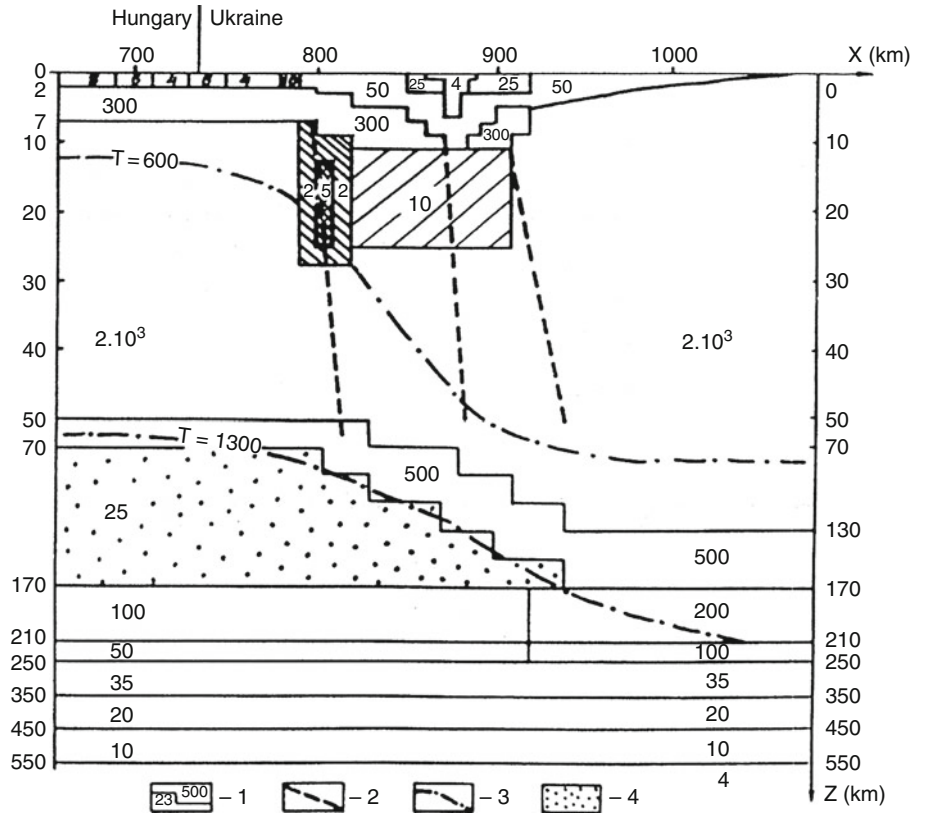
where σ is the conductivity (Sm^{-1}), μ is the magnetic permeability (Hm^{-1}), and $\omega \equiv 2\pi f$ is the angular frequency with f denoting the frequency (Hz). The free-space value of magnetic permeability ($\mu = \mu_0 = 4\pi \times 10^{-7} \text{Hm}^{-1}$) is generally assumed. As shown in the above equation, the lower the frequency is, the larger is the magnetotelluric depth of investigation.

The time variations of the two horizontal (E_x and E_y) components of the electric (or telluric) field are measured by using non-polarizable electrodes in a broad period range. The H_x and H_y horizontal geomagnetic field components are measured either with induction coils or fluxgates. From the time series of the horizontal electric and magnetic field components their frequency spectra are determined. In the simplest case, from $E_x(\omega)$ and $H_y(\omega)$ a complex impedance $Z(\omega)$ is computed as follows:

$$Z_{xy}(\omega) = \frac{E_x(\omega)}{H_y(\omega)}.$$

Goelectromagnetism,

Fig. 3 Goelectric and geothermal model of the Carpathian conductivity anomaly and the deepening asthenosphere toward the East European Platform along a NE–SW profile (so-called KAPG geotraverse II) crossing the Transcarpathian region in Ukraine. (After Zhdanov et al. 1986). 1, goelectric boundaries and resistivity values of blocks in Ωm ; 2, deep faults; 3, isotherms; 4, conducting asthenosphere



From Z the apparent resistivity ρ (i.e., the electrical resistivity of a fictitious homogeneous half-space) is defined as

$$\rho = \frac{1}{\omega\mu} |Z_{xy}|^2.$$

The phase difference between the electric and geomagnetic components is given as

$$\varphi = \arg Z_{xy} = \tan^{-1} \frac{\text{Im} Z_{xy}}{\text{Re} Z_{xy}}.$$

In the general case, the impedance is a period-dependent tensor:

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{pmatrix} \begin{pmatrix} H_x \\ H_y \end{pmatrix}$$

where Z_{xy} and Z_{yx} are the antidiagonal or main impedance elements and Z_{xx} and Z_{yy} are the diagonal or additional elements.

For a 1D model (where the goelectric structure varies exclusively with depth), the impedance elements are, theoretically, in the following relationships:

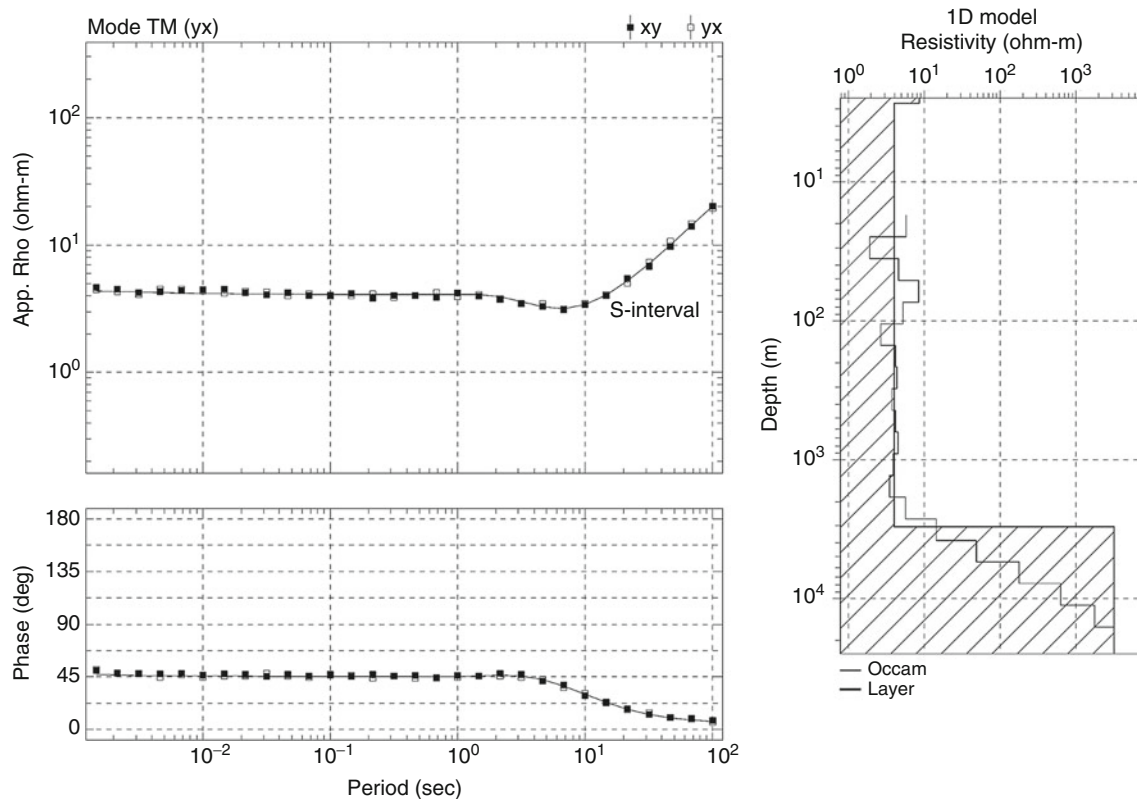
$$Z_{xy} = -Z_{yx} \text{ and } Z_{xx} = Z_{yy} = 0.$$

For a 2D model (where the resistivity section does not change in one of the horizontal direction named by strike direction), the apparent resistivity calculated from E_x and H_y components will differ from that derived from E_y and H_x , that is,

$$Z_{xy} \neq Z_{yx}$$

Furthermore, $Z_{xx} = Z_{yy} = 0$ is valid exclusively when the horizontal axes are aligned along and perpendicular to the strike direction.

One of the basic modes is called as *TE mode* (or *E-polarization*), where the electric field exists only in *strike direction*. The other basic mode is called as *TM mode* (or *H- or B-polarization*), where it is the magnetic field, which has one single component parallel to the strike direction. In the TE mode, there is also a measurable vertical magnetic component (H_z), which is usually also measured; in the TM mode, there is a measurable vertical electric component (which is not measured). The TE and TM mode apparent resistivities represent the goelectric structure in two different ways. The TE mode is more sensitive to deep



Geoelectromagnetism, Fig. 4 Magnetotelluric sounding curves for a layer (ID) model with $R_{h1} = 4 \Omega\text{m}$, h_1 (thickness) = 3 km, $R_{h2} = 2,000 \Omega\text{m}$. According to the Occam and few-layer inversions

the “S-interval” starts at a period of 10 s. At longer periods the telluric (Earth currents) method can be used to determine the thickness of the low resistivity sedimentary layer on a more resistive basement

conductive anomalies, while the TM mode – due to the charge accumulation at the resistivity boundaries – is more sensitive to the inhomogeneity boundaries.

When the model is 3D, the complete impedance tensor has four nonzero complex elements.

MT sounding curves are obtained by plotting apparent resistivity and impedance phase as functions of period (or square root of the period) over a wide period range. In Fig. 4, a 2 layer 1D resistivity model is shown. Note that the resistivity curve is fully logarithmic while the phase curve is semi-logarithmic. $Z_{xy}(\omega)$ and $Z_{yx}(\omega)$ serve as the basis of the TE and TM mode $\rho(\omega)$ and $\varphi(\omega)$ sounding curves. Computing the impedance elements in rotated coordinate systems, one obtains so-called polar diagrams, which characterize the anisotropy of the resistivity distribution.

Interpretation

Complex impedance elements recorded over a range of periods and at many sites along a profile or over an area, can be used to estimate the true resistivity distribution of the subsurface. This procedure is called inversion. Before carrying out inversion, the impedance elements should be corrected from distortions of both technical and structural origin. The “remote reference” method – using the data of simultaneous measuring sites in the

processing – reduces, for example, the effect of man-made electromagnetic noises (Gamble et al. 1979). Different *decomposition* techniques (e.g., (Larsen 1977; Groom and Bailey 1989; Bahr 1991); corrected by Prácser and Szarka 1999; Caldwell et al. 2004) enable suppression of the effect of a near-surface small 3D inhomogeneity causing the “static shift” of galvanic origin. At the same time they inform how well the impedance data correspond to a 2D resistivity model, and give an estimate for the strike direction, too. The papers by Bibby et al. (2005) and by Weaver et al. (2006) also explore the role of the phase tensor in determining dimensionality and strike direction. Due to the fact that the inversion is inherently nonunique, additional geophysical and geological information is needed in order to get a realistic model of the subsurface. The most frequently used 2D inversion algorithms in magnetotellurics are those by Rodi and Mackie (2001) and Siripunvaraporn and Egbert (2000) (see Fig. 7). The fitness of the modeled and measured data is characterized by the root-mean-square (rms) values. Various EM imaging procedures, exploiting the whole information content of the impedance tensor, may also be useful, as proposed by Szarka et al. (2005) and Weaver et al. (2000). For this purpose, Szarka and Menvielle (1997) recommended the system of independent rotational invariants of the impedance tensor. The usefulness

of this approach was demonstrated in detail by Novák (2009) using field data. In recent years, many papers have been published on 3D modeling and inversion and some have been presented at the biennial International Association of Geomagnetism and Aeronomy (IAGA) Workshops on Electromagnetic Induction in the Earth, including reviews of these topics, e.g., Avdeev (2005).

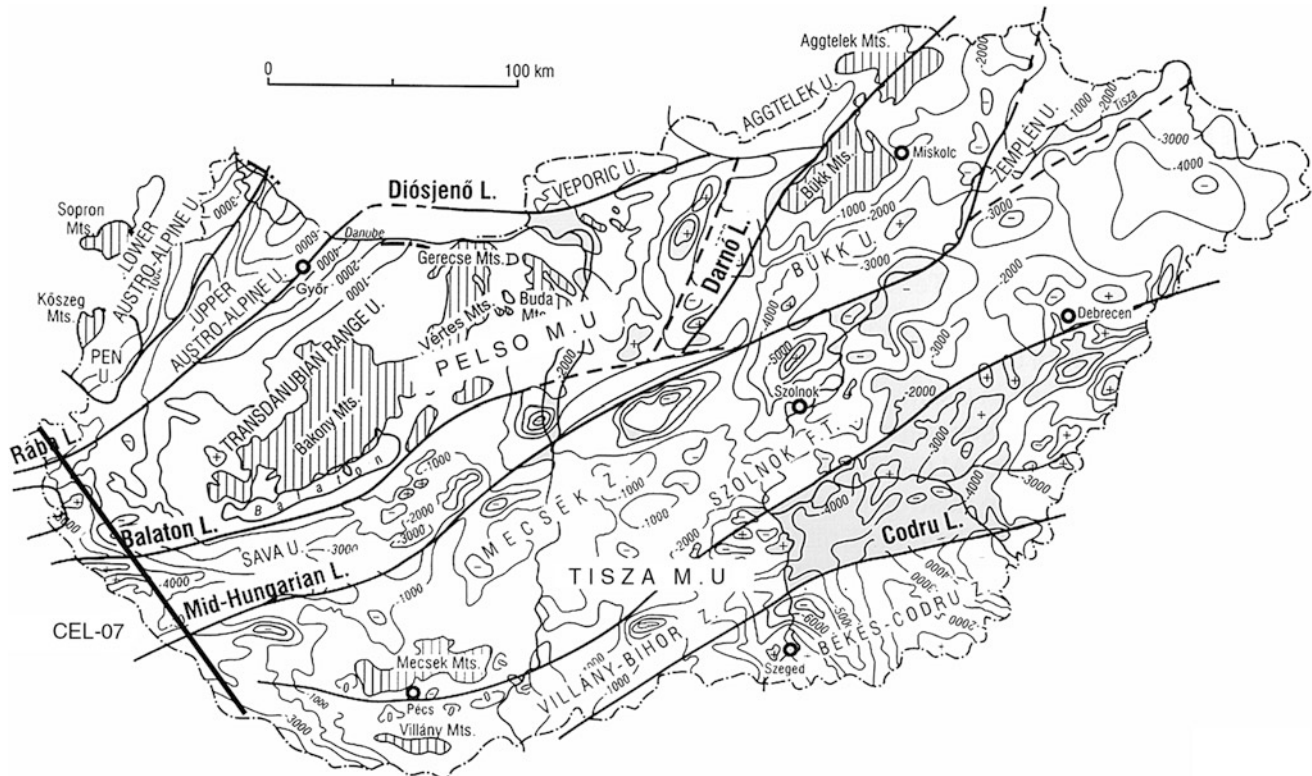
Application to Geosciences and Future Directions

MVS, MVP, MT, and to a lesser extent tellurics have received a very wide application in studying the structure and physical processes of the Earth at various local, regional, and global scales. The development of field applications has been disseminated at the biennial IAGA electromagnetic induction workshops since 1972 (see MTNet website, http://www.dias.ie/mtnet/workinggroup/wg_wsreviews.html) and at assemblies of IAGA. For example, at the last (11th) IAGA Assembly (Sopron, 23–30 August 2009, www.iaga2009sopron.hu), a wide spectrum of geoelectromagnetic studies, from the surface to the middle of mantle, was presented: theoretical and practical aspects of near-surface and environmental problems, oceanic geophysical exploration, formation and deformation of the continental lithosphere, crustal

tectonic processes, deep mantle structure, and data interpretation methods.

In order to illustrate the wealth of geoelectromagnetic applications, the authors present a selection from their own results from the Pannonian Basin, which is a very complex geological-geophysical unit of the Carpatho-Pannonian region. Such a summary cannot be done without mentioning other geoelectromagnetic investigations of the most well-known global geodynamic phenomena, such as the active plate tectonics in the Himalayas, Andes, San Andreas Fault, Afar region, etc. Earth being a “dynamic planet,” great efforts have also been devoted to mitigate the geohazards: earthquakes, volcanoes (as accompanying phenomena of the plate tectonics), and landslides (Bedrosian 2007; Korja 2007).

Before discussing the electrical conductivity anomalies/structures in the Pannonian Basin, we give a short description about its geology from those aspects that constrain the spatial distribution of electrical resistivity in the subsurface. The Pannonian Basin has a fairly thick and relatively uniform low electrical resistivity sedimentary cover of Late Tertiary origin, which was investigated in detail by telluric/magnetotelluric soundings. According to Haas (2001), the Pre-Neogene basement “shows a mosaic pattern of heterogeneous structural elements, a collage of allochthonous terrains derived from different parts of the Tethyan realm” (Fig. 5).



Goelectromagnetism, Fig. 5 Pre-Tertiary structural units of Hungary and depth of the pre-tertiary basement (Haas 2001)

In the basement there are two large Mega-units of composite terrains, which are separated by the ENE-WSW trending Mid-Hungarian Lineament. The South-Pannonian Mega-unit is the so-called TISIA (TISZA) Terrain and the North-Pannonian one is the so-called ALCAPA Mega-unit, which was formed from different smaller terrains. One of them is the PELSO in its southern part.

The evolution of the Pannonian Basin is significantly determined by the Miocene synrift, accompanied with a strong thermal heating, and leading to appraisal of crustal and mantle horizons and an extension of the basin by 10–30%, pre-dominantly in the NE–SW direction (Horváth and Berckhemer 1982). These features are extremely dominant in the deep extensional sedimentary sub-basins (e.g., in the Békés basin).

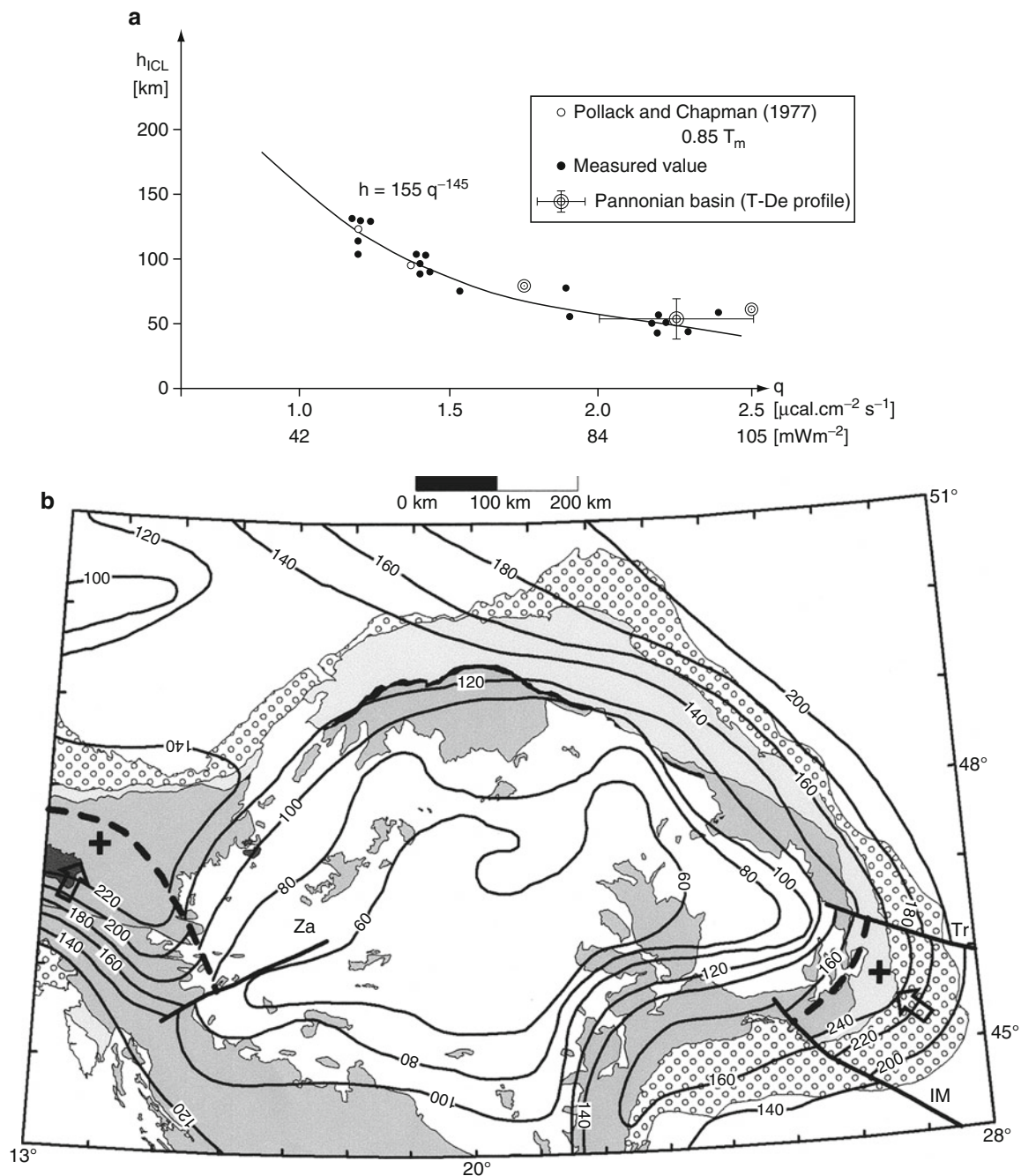
The spatial distribution of electrical resistivity/conductivity is influenced in the Pannonian Basin by the position, boundaries, and mosaic structure of the Megaunits and tectonic and thermal effects in the lithosphere, caused by synrift extension that has only been slowly compensated by the presently ongoing compression of the basin. The conductivity structures conserve the history of basin. Due to upheating of the Pannonian Basin – indicated by the high heat flow values at the surface of about 100 mW/m² – two geoelectric conductors are in an elevated position, in comparison to the surrounding mainly outer Carpathian: Paleozoic and Proterozoic regions.

Application of geoelectromagnetic methods is illustrated in five snapshots, representing various depth ranges: asthenosphere, mega-blocks, middle crust, upper crust, and near-surface.

1. *Asthenosphere.* The depth of the high-conductivity asthenosphere – in which the temperature approximates the solidus of the mantle peridotite, so that a partial melting can start – is in the Pannonian Basin at about 60 km depth, indicated by Ádám (1965). In platforms and shield areas the asthenosphere is at much deeper position – if it exists at all. Its estimated value is about 250 km as concluded by extended projects such as BEAR in Finland (Korja and BEAR Working Group 2003) and CEMES (Yu Semenov et al. 2008) in East-Central Europe (See also Fig. 3). The MT indications of the conductive asthenosphere at an elevated position within Pannonian Basin were confirmed with reflection seismics (Posgay 1975) and also with a thermobarographic diagram of well-equilibrated peridotite xenolith from the PELSO (Embey-Isztin et al. 2001). The elevated position is in close correlation with enhanced heat flow values (Fig. 6a, Ádám 1978). Its smallest depth of about 50 km was observed in the Békés extensional sub-basin (Ádám et al. 1996). The map of asthenospheric depth in and around the Pannonian Basin is shown by Fig. 6b as the result of seismic, seismological, and magnetotelluric measurements (See in Horvath et al. 2006, based partly on the MT data of Ádám and Wessztergom 2001).
2. *Middle crust.* In the middle crust the transition between the brittle and ductile zone is also temperature dependent, which might be due to dehydration as proposed by Hyndman and Shearer (1989). Ádám (1978, 1987) found a relation between the mid-crustal conductivity increase and the regional heat flow values. In the Pannonian Basin there is only a weak indication of this anomaly at a depth of about 15–17 km, but this is supported by the lower limit of seismic focal depths and by strength decrease. The electrical conductivity of the lower crust was reviewed by Jones (1992).
3. *Boundaries of mega-units.* The boundaries of the Mega-units and their mosaics in the Pannonian Basin were studied among others along the deep seismic CEL-07 profile in SW Hungary with 72 deep MTS. The distance between neighboring magnetotelluric stations was in average only 2 km. The CEL-07 profile is shown together with the complicated basement tectonic structure in Fig. 5.

The resistivity profile of the TE mode inversion along CEL-07 is shown in Fig. 7 (the TM mode inversion profile is strongly distorted by charge accumulation at the boundary of electrical inhomogeneities falsifying the values of the resistivity distribution). The TE mode inversion clearly expresses the conductivity anomalies at the boundaries of the Mega-units: the Balaton + Balatonfő double lineaments at the southern border of the PELSO M.u. (mosaic of the ALCAPA M.u.) and the Mid-Hungarian Line at the northern border of TISIA M.u. The horizontal electric conductance of the lineaments is about 2,500–3,000 S, starting at a depth of about 7 km. The source of the conductivity anomaly is certainly hot saline fluid containing also SiO₂ solution. Deep borehole data refer to accumulation of CO₂, increase of He concentration, etc. in the deep fractures. A high value of convective heat transfer, especially in volcanic eruption centers, also increases the electrical conductivity in these tectonic zones, which (e.g., the Balaton Line) is also indicated by the zero value of the induction vectors (Ádám et al. 2005).

4. *Crustal conductivity anomalies.* On the northern part of the CEL-07 resistivity profile there appears another conductivity anomaly that is assumed to be the continuation or ending of one of the extreme high fracture zones related to the Transdanubian Conductivity Anomaly (TCA) studied by 300 deep MT soundings (Ádám 2001). The TCA consists of a conductive layer at about 5 km depth and of three significant fracture zones in the NE–SW direction. Their horizontal electric conductance value can reach 10,000 S. Such high conductance values make it probable that the sources of the anomaly are graphite/graphitic shists as well as fluid, which are partly within, and partly under the thick Mesozoic limestone sequence of the Bakony Mts. In this

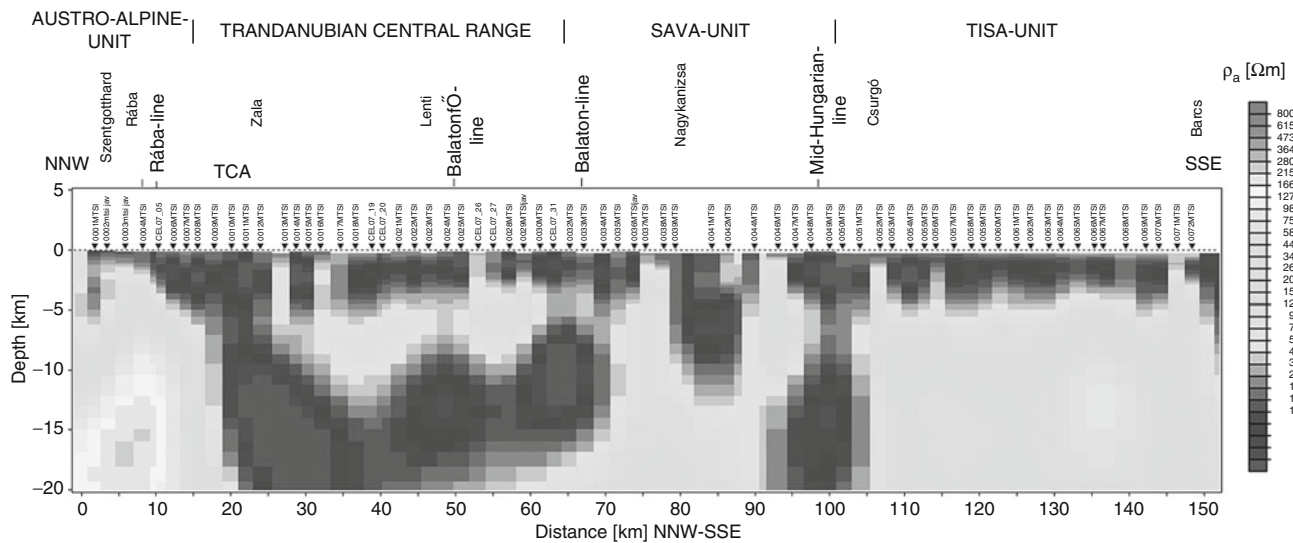


Geoelectromagnetism, Fig. 6 (a) Relation between the depth of the conductive asthenosphere and the regional heat flow (Ádám 1978), (b) The map of the asthenospheric depth in and around the Pannonian Basin

as results of seismic/seismological/magnetotelluric results (Horvath et al. 2006; Ádám and Wetztergom 2001)

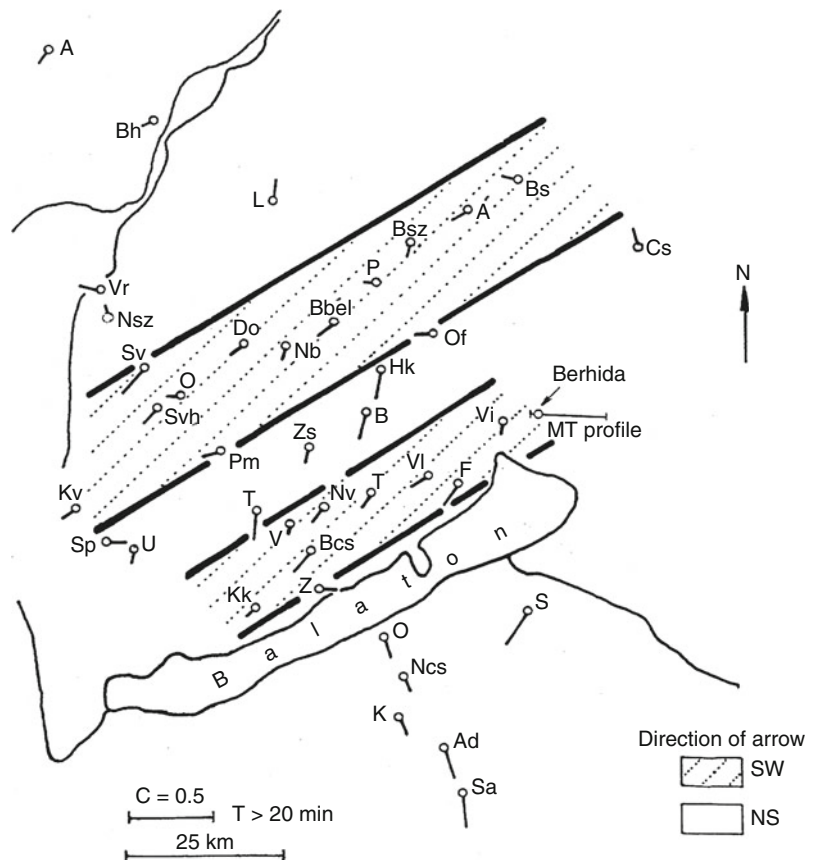
area there is significant seismic attenuation (Zsíros 1985), and this must be due to an electrical conductor (Glover and Ádám 2008). The large fractures as shown by the strongly reduced induction vectors (Fig. 8) are connected to the hypocenters of the earthquakes (e.g., Mór in, 1809, Berhida in 1985, rarely exceeding a magnitude of 6 on Richter's scale). Glover and Ádám (2008) found a correlation between the depth to the zone of this high electrical

conductivity anomaly and earthquake focal depths. According to Ogawa (2007) the earthquakes occur at the boundaries of the electric conductors, where only aseismic slip or elastic deformation can happen, rather than within the conductors themselves. In this way the energy can be transported towards the brittle zone, causing devastating earthquakes. In the presence of the hypothetical enhancement of magnetic permeability at the Curie depth (Kiss



Geoelectromagnetism, Fig. 7 The resistivity profile of the TE+ Hz nonlinear conjugate gradient inversion of the magnetotelluric measurements along the CEL-07 profile (RMS = 1.344132). See the conductors corresponding to the deep boundary fractures (Ádám et al. 2005)

Geoelectromagnetism, Fig. 8 Wiese induction vectors along the deep fractures of the TCA (Wallner 1977)



et al. 2005), MT interpretation of upper and middle crustal inhomogeneities should be carried out with a special care.

5. *Near-surface*. In order to study the near-surface depth range, either higher frequencies or local sources should be applied. Various high-frequency variants have been developed: instead of the classical MVP the so-called VLF; instead of the classical magnetotellurics: audiomagnetotellurics (AMT), radiomagnetotellurics (RMT), or VLF-EM. If the source is finite, we quickly arrive at various near-surface electromagnetic induction and direct current methods.

As a result of the various EM induction methods becoming increasingly important, the number of geoelectromagnetic papers published in international journals has shown an upward trend. Recently, Bedrosian (2007) reviewed MT measurements, which were carried out in the Himalayas (Lemonnier et al. 1999), in Southern Tibet (Li et al. 2003), in the Andean margin (Brasse et al. 2002), etc. In each of these dynamic areas the electrical conductivity anomalies appear at their most active segments, which has been also indicated by thermal activation (in forms of high heat flow), earthquakes, volcanoes, recent crustal movements, regional metamorphism (e.g., graphitization), partial melting, etc. Korja (2007) gives in his review a comprehensive description about the conductive anomalies in Europe. As Bedrosian (2007) stated, for a correct interpretation of MT (EM) data, due to the nonuniqueness of the geophysical inverse problems, independent geological and geophysical information is necessary. This conclusion is highly confirmed by our experiences with geoelectromagnetic results from the Pannonian Basin.

For further reading on various topics in geoelectromagnetism covered in this entry, both practical and theoretical, the reader is referred to the specialized monographs by Wait (1962), Berdichevsky and Zhdanov (1984), Weaver (1994), Simpson and Bahr (2005), and Zhdanov (2009) among others in sequence of the publication year.

Cross-References

- [Electrical Properties of Rocks](#)
- [Instrumentation, EM](#)
- [Magnetotelluric Data Processing](#)
- [Magnetotelluric Interpretation](#)

Bibliography

- Ádám A (1965) Einige Hypothesen über den Aufbau des oberen Erdmantels in Ungarn. *Gerlands Beiträge zur Geophysik* 74(1):20–40
- Ádám A (1978) Geothermal effects in the formation of electrically conducting zones and temperature distribution in the Earth. *Phys Earth Planet Inter* 17(2):P21–P28
- Ádám A (1987) Are there two types of conductivity anomaly (CA) caused by fluid in the crust? *Phys Earth Planet Inter* 45:209–215
- Ádám A (2001) Relation of the graphite and fluid bearing conducting dikes to the tectonics and seismicity (Review on the Transdanubian crustal conductivity anomaly). *Earth Planet Space* 53:903–918
- Ádám A, Wesztergom V (2001) An attempt to map the depth of the electrical asthenosphere by deep magnetotelluric measurements in the Pannonian Basin (Hungary). *Acta Geol Hung* 44:167–192
- Ádám A, Szarka L, Prácsér E, Varga G (1996) Mantle plumes or EM distortions in the Pannonian Basin? (Inversion of the deep magnetotelluric (MT) soundings along the Pannonian Geotransverse). *Geophys Trans* 40:45–78
- Ádám A, Novák A, Szarka L (2005) Tectonic weak zones determined by magnetotellurics along the CEL-7 deep seismic profile. *Acta Geodaetica et Geophysica Hungarica* 40(3–4):413–430
- Arora BR (1997) Chap 4. Geomagnetic deep sounding. In: Arora BR, Sri Niwas (eds) *Natural source electromagnetic induction in the earth*. New Age International, New Delhi, pp 80–128
- Avdeev PB (2005) Three-dimensional electromagnetic modelling and inversion from theory to application. *Surv Geophys* 26:767–799
- Bahr K (1991) Geological noise in magnetotelluric data: a classification of distortion types. *Phys Earth Planet Inter* 66:24–38
- Bailey RC (1973) Global geomagnetic sounding – methods and results. *Phys Earth Planet Inter* 7:234–244
- Bedrosian PA (2007) MT+, Integrating magnetotellurics to determine earth structure, physical state and processes. *Surv Geophys* 28:121–167
- Berdichevsky MN, Dmitriev MN (2008) *Models and methods of magnetotellurics*. Springer, Berlin, 563 pp
- Berdichevsky MN, Zhdanov MS (1984) *Advanced theory of the deep geomagnetic sounding*. Elsevier, Amsterdam, 408 pp
- Bibby HM, Caldwell TG, Brown C (2005) Determinable and non-determinable parameters of galvanic distortions in magnetotelluric. *Geophys J Int* 163:915–930
- Brasse H, Laezaeta P, Rath V, Schwabenberg K, Soyer V, Haak V (2002) The Bolivian Altiplano conductivity anomaly. *J Geophys Res* 107. <https://doi.org/10.1029/2001/J.B.000391>
- Cagniard L (1953) Basic theory of the magnetotelluric method of geophysical prospecting. *Geophysics* 18:605–635
- Caldwell TG, Bibby HM, Brown C (2004) The magnetotelluric phase tensor. *Geophys J Int* 158:457–469
- Embey-Isztin A, Downes H, Dobosi G (2001) Geochemical characterization of the Pannonian Basin mantle lithosphere and asthenosphere: an overview. *Acta Geol Hung* 44:259–280
- Gamble TB, Goubau WM, Clarke J (1979) Magnetotellurics with a remote reference. *Geophysics* 44:53–68
- Glover PWJ, Ádám A (2008) Correlation between crustal high conductivity zones and seismic activity and the role of carbon during shear deformation. *J Geophys Res* 113(B12210):1–8. <https://doi.org/10.1029/2008JB005804>
- Groom RW, Bailey RC (1989) Decomposition of magnetotelluric impedance tensors in the presence of local three-dimensional galvanic distortion. *J Geophys Res* 94:1913–1925
- Haas J (ed) (2001) *Geology of Hungary*. Eötvös University Press, Budapest, 317 pp
- Horváth F, Berckhemer H (1982) Mediterranean backarc basin. In: Berckhemer H, Ksu H (eds) *Alpine-mediterranean geodynamics, Geodynamics series, vol 7*. American Geophysical Union/Geological Society of America, Washington, DC/Boulder, pp 141–173
- Horváth F, Bada G, Szafián P, Tari G, Ádám A, Cloetingh S (2006) Formation and deformation of the Pannonian Basin constraints from observational data. In: Gee DG, Stephenson RA (eds) *European lithosphere dynamics, Memoires no. 32*. Geological Society, London, pp 191–206
- Hyndman RD, Shearer PM (1989) Water in the lower continental crust modelling magnetotelluric and seismic reflection results. *Geophys J Int* 98:343–365
- Jankowski J, Praus O, Józwiak V (2005) Carpathian anomaly of electrical conductivity: history of its discovery and interpretation. *Publ Inst Geophys Pol Acad Sci C-95(386):17–27*
- Jones AG (1992) Electrical conductivity of the continental lower crust. In: Fountain DM, Arculus RJ, Kay RW (eds) *Continental lower crust*. Elsevier, Amsterdam, pp 81–143
- Kántás K (1954) The results and perspectives of the Hungarian telluric research (in Hungarian). *Bányászati Lapok* 87:17
- Kelbert A, Schultz A, Egbert G (2009) Global electromagnetic constraints on transition-zone water content variation. *Nature* 460:1003–1006

- Kiss J, Szarka L, Prácser E (2005) Second order magnetic phase transition in the Earth. *Geophys Res Lett* 32:L24310. <https://doi.org/10.1029/2005GL024199>.
- Korja T (2007) How is the European lithosphere imaged by magnetotellurics. *Surv Geophys* 28:239–272
- Korja T, BEAR Working Group (2003) Is there an electrical asthenosphere beneath the Fennoscandinavian Shield. In: EGU-AGU joint meeting, Nice, 2003–2004, Abstract EAE03-A-14804
- Lamb H (1883) On electrical motions in a spherical conductor. *Philos Trans R Soc Lond* 174:519–549
- Larsen J (1977) Removal of local surface conductivity effects from low frequency mantle response curves. *Acta Geodaetica Geophysica et Montanistica Acad Sci Hung* 12:183–186
- Lemonnier C, Marquis G, Perrier F, Avouac JP, Chitraker G, Kafle B, Sapkota S, Gautam U, Tiware D, Bano M (1999) Electrical structure of the Himalaya of Central Nepal: high conductivity around the mid-crustal ramp along MHT. *Geophys Res Lett* 26:3261–3264
- Li S, Unsworth MJ, Booker JR, Wei W, Tan H, Jones AG (2003) Partial melt or aqueous fluid in the midcrust of Southern Tibet? Constraints from INDEPTH magnetotelluric data. *Geophys J Int* 153:289–304
- Novák A (2009) Electromagnetic imaging in geophysics with tensorial invariants: from the near surface to Transdanubian deep structures. PhD dissertation (in Hungarian), Sopron (Hungary), 187 pp
- Ogawa Y (2007) Correlation of resistivity structure, seismicity and deformation from some case studies in Japan. In: Lecture at the JAS003 conference of the IUGG General Assembly in Perugia
- Parkinson WD (1959) Direction of rapid geomagnetic fluctuation. *Geophys J R Astron Soc* 2:1–14
- Posgay K (1975) Mit Reflexionsmessungen bestimmte Horizonte und Geschwindigkeits Verteilung in der Erdkruste und Erdmantel. *Geophys Trans* 23:13–18
- Prácser E, Szarka L (1999) A correction to Bahr's "phase deviation" method for tensor decomposition. *Earth Planet Space* 51:1019–1022
- Rodi W, Mackie RI (2001) Nonlinear conjugate gradients algorithm for 2-D magnetotelluric inversion. *Geophysics* 66:174
- Rokityansky II (1982) *Geoelectromagnetic investigation of the Earth's crust and mantle*. Springer, Heidelberg, p 381
- Schmucker U (1959) *Erdmagnetische Tiefenondierung in Deutschland 1957/1959*. Magnetogramme und erste Auswertung. Abhandlungen der Akademie der Wissenschaften in Göttingen. Mathematisch-Physikalische Klasse. I, H5, 51. Vandenhoeck & Ruprecht, Göttingen.
- Schmucker U (1970) Anomalies of geomagnetic variations in the southwestern United States. *Bull Scripps Inst Oceanogr* 13:165
- Schmucker U (2003) Horizontal spatial gradient sounding and geomagnetic depth sounding in the period range of daily variation. In: Hördt A, Stoll J (eds) *Protokoll über das Kolloquium elektromagnetische Tiefenforschung*. Bibliothek des Wiss. Albert Einstein, Potsdam, pp 228–237
- Schultz A, Larsen JC (1983) Analysis of zonal field morphology and data quality for a global set of magnetic observatory daily main values. *J Geomagn Geoelectr* 35:835–846
- Simpson F, Bahr K (2005) *Practical magnetotellurics*. Cambridge University Press, Cambridge, 270 pp
- Siripunvaraporn W, Egbert GD (2000) An efficient data subspace inversion for two-dimensional magnetotelluric data. *Geophysics* 65:791–803
- Szarka L, Menvielle M (1997) Analysis of rotational invariants of the magnetotelluric impedance tensor. *Geophys J Int* 129:133–142
- Szarka L, Ádám A, Menvielle M (2005) Field test of a quicklook imaging method based on rotational invariants of the 3D magnetotelluric tensor. *Geophys Prospect* 53:325–334
- Tikhonov AN (1950) On investigation of electrical characteristics of deep strata of Earth's crust (in Russian). *Doklad Akademii Nauk SSSR* 73:295–297
- Vozar J, Semenov VY (2010) Compatibility of induction methods for mantle soundings. *J Geophys Res* 115(B03101):9. <https://doi.org/10.1029/2009JB006390>
- Wait JR (1962) *Geoelectromagnetism*. Academic, New York. 268 pp
- Wallner Á (1977) The main features of the induction arrows in the area of the Transdanubian conductivity anomaly. *Acta Geodaetica Geophysica et Montanistica* 12:145–150
- Weaver JT (1994) *Mathematical methods for geoelectromagnetic induction*. Wiley, Taunton, p 316
- Weaver JT, Agarwal AK, Lilly FE (2000) Characterization of the magnetotelluric tensor in terms of its invariants. *Geophys J Int* 141:321–336
- Weaver JT, Agarwal AK, Lilly FE (2006) The relationship between the magnetotelluric tensor invariants and the phase tensor of Caldwell, Bibby and Brown. *Explor Geophys* 37:261–267
- Wiese H (1965) *Geomagnetische Tiefentellurik*. Akad. Verlag, Berlin. 146 pp
- Yu Semenov V, Ádám A, Hvozda M, Wesztergom V (1997) Geoelectrical structure of the Earth's mantle in the Pannonian Basin. *Acta Geodaetica et Geophysica Hungarica* 32(1–2): 151–168
- Yu Semenov V, Pek J, Ádám A, Józviak W, Ladanyvskyy B, Logvinov IM, Pushkarev P, Vozar I (2008) Electrical structure of the upper mantle beneath Central Europe: results of the CEMES project. *Acta Geophysica* 56:957–981
- Zhdanov MS (2009) *Geophysical electromagnetic theory and methods*. Elsevier, Amsterdam, p 848
- Zhdanov MS, Golubev NG, Iv Varentsov M, Abramova LM, Schneer MN, Berdishevsky MN, Zhdanova ON, Gordienko VV, Kulik SN, Bilinsky AI (1986) 2D model fitting of a geomagnetic anomaly in the Soviet Carpathians. *Ann Geophys B4(3):335–342*
- Zsíros T (1985) An estimation of seismic hazard in Hungary. *Gerlands Beiträge zur Geophysik Leipzig* 94:111–122

Geoid

Paramesh Banerjee

Earth Observatory of Singapore, Prayukti Geomatics Pte Ltd, Singapore, Singapore

Synonyms

Equipotential surface at mean sea level

Definition

- Geoid** The gravitational equipotential surface of the Earth that best approximates the “mean sea level.”
- Tectonic Forces** Forces originating from within the Earth that alters surface configuration of the Earth through uplift, movement, and deformation of the parts of the Earth's crust.

Introduction

Geoid defines the physical shape of the earth and serves as a globally consistent reference surface for measuring heights. The concept of geoid was introduced by Listing (1873). Like any homogeneous liquid in hydrostatic equilibrium does, ocean water surface also tends to follow a gravity equipotential surface. The equipotential surface that best matches the ocean water surface on a global scale is defined as the geoid. Because of variation in salinity, particle content, water temperature, ocean currents, climate-related sea level rise, tectonic motions, etc., ocean water surface does not conform perfectly to the geoid surface. A best fit surface is thus adopted and maximum deviation of ± 2 m. has been observed globally. The geoid is locally realized as “mean sea level,” by averaging water levels at tide gauge stations. This “mean sea level” is traditionally used as a reference for measuring heights over lands using leveling instruments. In surveying parlance, geoid and “mean sea level” are thus synonymous. Extension of that “sea level” inside continents is done either by leveling traverses or by modeling geoid using gravimetric data.

Geoid and Heights

Heights measured referenced to the geoid is called orthometric height (H) (Fig. 1) and is the most commonly used height in engineering and mapping all over the world. Realization of the geoid surface is not always a straight forward process and requires careful processing of surface gravity data. To overcome problems associated with unknown

subsurface density inhomogeneity that are required in computing geoid from surface gravity measurements, a new concept of “quasi-geoid” was introduced by Molodensky et al. (1960). Some European and former Soviet Union countries use quasi-geoid, instead of geoid, as the reference surface to the height. Quasi-geoid is a superficial surface, and the referenced height is known as “normal height.” With more subsurface and crustal structures being explored using geophysical techniques, the geoid remains most widely used and robust reference surface.

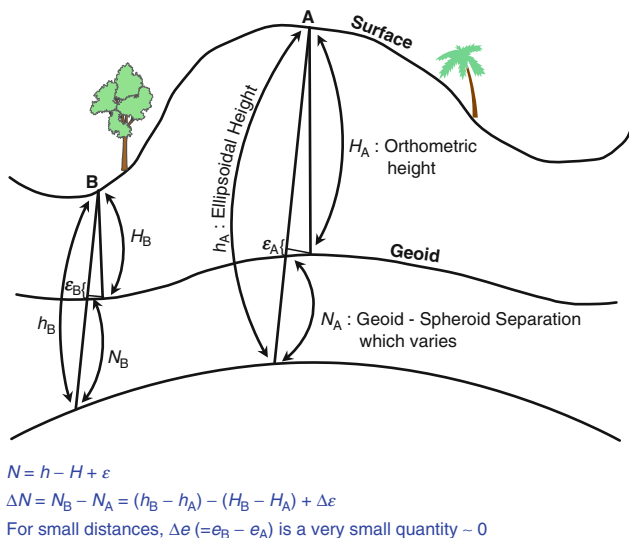
Geoid defines the physical shape of the earth. It is a real surface with irregularities and undulations caused by subsurface density inhomogeneities. A geocentric ellipsoid aligned with the rotational axis of the earth and best fitting the geoid defines the geometric shape of the earth. The reference ellipsoid provides a smooth, mathematically defined surface and has no physical existence. But it serves as the basis of the reference system for defining horizontal coordinates of points on the earth’s surface. Geoid forms the reference for the vertical datum. Vertical separation between the geoid and ellipsoid corresponding to a point on the surface of the earth is known as geoidal height (N), geoid anomaly, or geoid undulation, and varies between $+100$ m and -100 m globally. All positioning measurements made by satellite-based geodetic system, e.g., GNSS (Global Navigation Satellite System), including GPS (Global Positioning System), are referred to the reference ellipsoid. Height of any point referenced to the ellipsoid is called ellipsoidal or geodetic height (h). It has no practical use unless converted to the orthometric height. Relationship between the ellipsoidal (h), orthometric (H), and geoid (N) heights are given by Eq. (1). The relationship is approximate, but precise enough for most of the practical purposes, particularly when used in differential mode.

$$N \approx h - H \quad (1)$$

Because of the recent advances in space geodetic techniques, ellipsoidal heights can be measured with sub-cm level accuracy. Whereas, orthometric height measurements using terrestrial leveling method could attain 10–15 cm level accuracy even at the best measured first-order leveling networks. A geoid model (N) enables to convert high accuracy ellipsoidal heights into orthometric heights using Eq. (1). Determination of a highly accurate geoid model has thus become a prerequisite to derive benefits of space geodetic techniques which is easy, highly accurate, globally consistent, and economic.

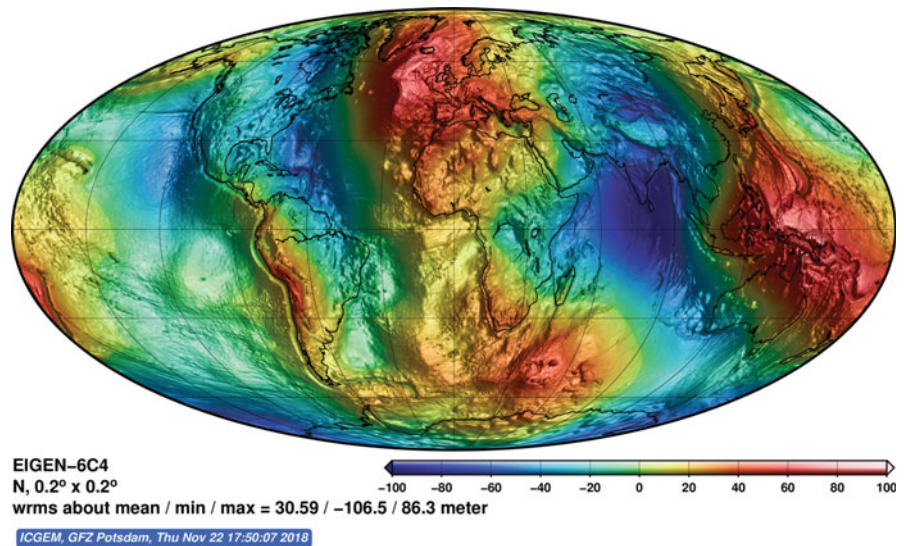
Geoid Modeling from Gravity Data

One simple way of estimating geoid height N is measuring a leveling benchmark with GNSS instrument, typically known as



Geoid, Fig. 1 A simplified cartoon showing geoid and spheroid levels both over the continents and oceans. Geoid undulation varies between $+100$ m to -100 m. Indian Ocean has the lowest geoid anomaly

Geoid, Fig. 2 A 3-D visualization of global geoid undulation model from the EIGEN-6C4 combined gravity data. For details, see Ince et al. (2019)



After E. S. Ince et al.: 15 years of ICGEM service, *Earth Syst. Sci. Data*, 11, 647–674, 2019
<https://doi.org/10.5194/essd-11-647-2019>
 © Author(s) 2019. This work is distributed under the Creative Commons Attribution 4.0 License.

“GPS-on-Leveling” technique, and applying Eq. (1) to get N . Sparse measurements can be interpolated to other points within nearby region where geoid can be assumed to be linear. But the accuracy of the geoid height is limited by terrestrial leveling measurements which is error prone over long distance, expensive, and slow. These problems are more severe in mountainous region, places far away from tide gauge stations, and tectonically active zones. Measurement error can accumulate to the order of meter or more.

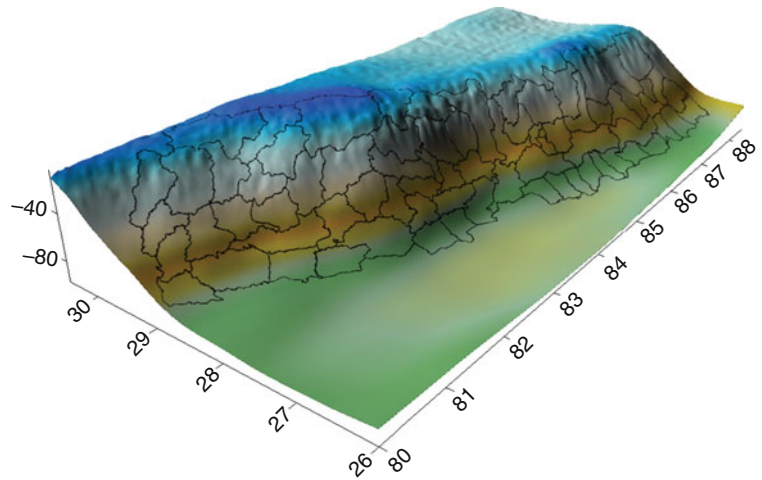
Geoid can be modeled from gravity data which are available in plenty and easy to acquire, by surface measurements, airborne surveys, and satellite tracking. Gravity field being the vertical gradient of the gravity potential, integration of the properly reduced gravity anomalies using the classical Stokes Integral (Stokes 1849) leads to the gravity potentials and geoid model. Typically satellite gravity data are used to develop global geoid models. To attain higher resolution at regional scale, both airborne and surface gravity measurements along with high-resolution topographic data are additionally used. With further developments in processing techniques, availability of refined global gravity model like EGM08, high-resolution geopotential models derived from GRACE (Tapley et al. 2004) and GOCE (Rummel 2010) missions, high-resolution Digital Elevation Models (SRTM, ETOPO, Lidar, etc.), and availability of more and more surface and airborne gravity measurements, objectives of developing a national geoid model with 1 cm precision seems more and more achievable. In mountainous regions, dilution of precision to one order of magnitude is expected. Developing a high accuracy, robust, and GNSS-friendly

height system based on a precise geoid model has become a priority in many countries in recent years. A recent high-resolution global geoid model is shown in Fig. 2 (Ince et al. 2019).

Measuring Height of the Everest

Measuring the height of the Mt. Everest, the highest topographic feature of the earth, is an interesting example of use of geoid in recent time. Measuring Mt. Everest has always remained a challenge to the geodesists. Starting with the classic triangulation method in 1855 by India, a series of measurements by China, USA, Europe, and Nepal gradually improved the precision of the measurements. The latest effort initiated by the Survey Department, Nepal, is to use GNSS measurements at the top and use a precise geoid model to get the orthometric height of the peak. In May, 2019, one Nepal Survey Dept. officer scaled the peak with a GNSS receiver and could collect data for around 30 min. A country-wide airborne gravity survey was carried out in Nepal (Forsberg et al. 2011) during 2010 and a precise geoid model was developed by Forsberg et al. (2011) with 7 cm precision. A 3D depiction is shown in Fig. 3. More surface gravity measurements have been recently added to further refine Forsberg’s geoid. The challenge of measuring the height of Mt. Everest has essentially boiled down to the challenge of developing most refined geoid map of the region. Official declaration from the Nepal Survey Dept. of the latest height is being expected soon.

Geoid, Fig. 3 Geoid model of Nepal from airborne gravity data (Forsberg et al. 2011). Starting from the Indian peninsula (−60 m), geoid surface goes up to −20 m in High Himalaya



Geoid Anomaly and Tectonics

Deviation of the actual geoid surface from the reference spheroid is known as geoid anomaly. Figure 4 shows the global geoid height anomaly (Bowin 2000). The largest geoid anomaly is a 100 m deep “hole” south of India, in the Indian Ocean. This huge geoid “low” is caused by a mass deficiency in the mantle underneath the sea. Other short wavelength geoid anomalies exist along the oceanic trenches where one tectonic plate subsides underneath another plate or along the mid-oceanic ridges where hot mantle material swells up to produce new oceanic crust. Geoid anomaly generally has a positive correlation with free-air gravity anomaly. However, gravity anomaly is more sensitive to the shallower features, whereas geoid anomaly is more sensitive to the deeper mass inhomogeneity.

Study of the geoid anomaly helps to understand the distribution of large scale density inhomogeneity, the state of isostatic compensation, and also about the flexural property of the lithosphere. Global geoid anomaly data indicates that, when loaded (or unloaded) with a large dimension surface load (e.g., a mountain formation or glaciation–deglaciation process) during a prolonged period, mantle flows and adjusts and the surface load eventually attains hydrostatic equilibrium. In contrast, short-period or small-scale surface load is not isostatically compensated. These are explained by a visco-elastic lithosphere which is strong enough to support at a smaller spatial and temporal scale, but has a finite viscosity and flows like a fluid at a larger spatial and temporal domain. Study of spatial as well as temporal variation of the geoid anomaly can thus be a very useful tool to understand mechanical properties of the lithosphere. It may be noted that isostatically balanced mass distribution does not cause zero geoid anomaly. An isostatically compensated mountain range of 3 km height causes a positive geoid anomaly of ~16 m, where a 5-km deep ocean basin would result a negative geoid anomaly of about 9 m (Fowler 1996). As seen in

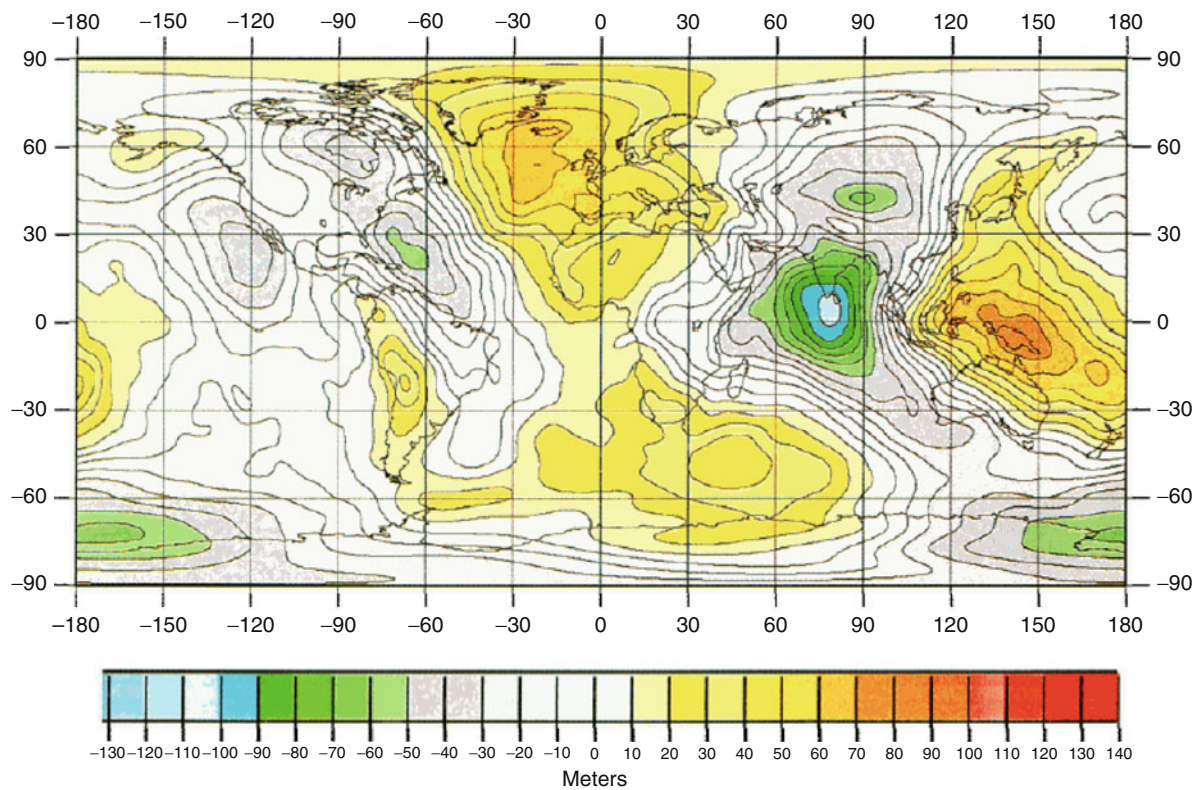
Fig. 3, Himalayan topography of around 6000 m average height caused a geoid anomaly of around 35 m (relative to stable India). As geoid anomaly is more sensitive to lateral variation in lithospheric density structure, surface measurements of geoid anomaly can provide a very useful tool to study subcrustal configuration across a plate boundary (Banerjee et al. 1999).

GRACE and Coseismic Deformation

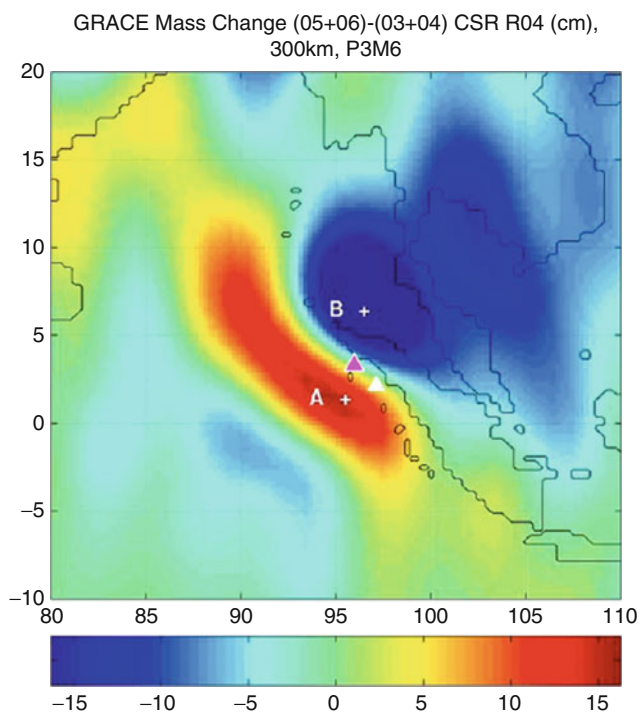
GRACE (Gravity Recovery and Climate Experiment) twin satellite system has made it possible to measure temporal variation of geoid height at a global scale with an accuracy of 2–3 mm (Tapley et al. 2004). The 2004 Sumatra Andaman earthquake (Mw 9.2) provided an opportunity to measure coseismic as well as post-seismic gravity and mass variations using GRACE data (Han et al. 2006; Chen et al. 2007). Sudden change in lithospheric density structure as well as vertical movement could be detected in GRACE data (Fig. 5). With this leap in space geodesy, remote sensing of earth’s geoid provides another important tool to study tectonic processes.

Tectonic Forces

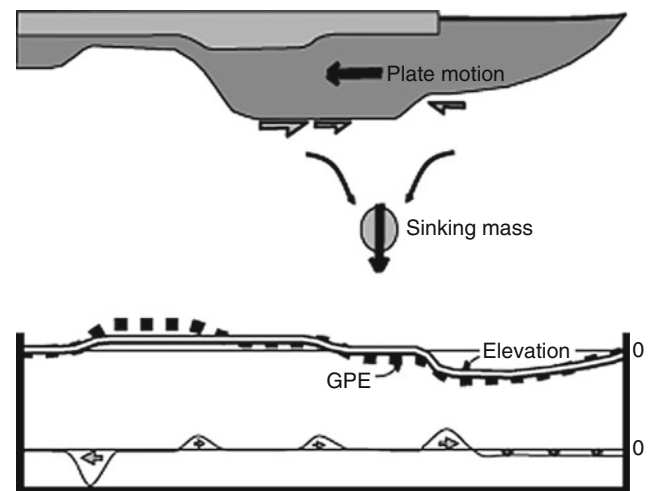
Tectonic forces acting on a lithospheric plate results in plate motion as well as accumulation of stresses, both in vertical and horizontal directions. Surface deformation and earthquakes are direct consequences of these stresses. Origin of the tectonic forces mostly lie in the flow in the asthenosphere which may either drive or resist the motion of the overlying plate. In the subduction zone, gravitational pull of the down going slab (slab pull) plays a more important role in driving the plate motion than the “ridge pushes” where slabs are driven away from the mid-oceanic ridges. Plate motion also creates traction on the base of the overlying continents



Geoid, Fig. 4 The geoid height anomaly map of the world. The contour line interval is 10 m. Plate 1; Bowin (2000). (Reproduced by permission of American Geophysical Union. © 2000, American Geophysical Union)



Geoid, Fig. 5 A change in crust-mantle mass distribution caused by 2004 Sumatra-Andaman earthquake as detected by GRACE gravity mission (Chen et al. 2007)



Geoid, Fig. 6 Different tectonic forces in play in a subduction zone where the subsiding slab sinks in weaker asthenosphere (slab pull), which excites mantle flow, thus creating a traction at the base of the plate (shown by small grey arrows at the base of the plate). Continental crust and oceanic crust are shown in light grey and black respectively, whereas mantle lithosphere is in dark grey

(Humphreys and Coblenz 2007). Another important source of force is generated by the lateral variation in density and thickness in the lithosphere which generates a gradient in gravitational potential energy (GPE). Figure 6 illustrates

how a subducting slab generates GPE gradient and basal traction force. Tibetan plateau is an example of high GPE where continental collision produced high topography, thickened crust, extensional stresses on the plateau, and compression at the edges.

Geoid height variations are linearly proportional to the GPE gradients in an isostatically balanced lithosphere (Turcotte and Schubert 2002). Geoid gradients can thus be used to study lithospheric stress gradients. Lithosphere-scale density inhomogeneities are reflected in shorter wavelength geoid anomalies. This, in combination with integrated density moment computed using a crustal density model at longer wavelength, can be used to produce GPE (Humphreys and Coblentz 2007).

Bibliography

- Banerjee P, Foulger GR, Satyaprakash, Dabral CP (1999) Geoid undulation modeling and interpretation at Ladakh, NW Himalaya using GPS and leveling data. *J Geod* 73:79–86
- Bowin C (2000) Mass anomaly structure of the Earth. *Rev Geophys* 38(3):355–387
- Chen JL, Wilson CR, Tapley BD, Grand S (2007) GRACE detects coseismic and postseismic deformation from the Sumatra-Andaman earthquake. *Geophys Res Lett* 34:L13302. <https://doi.org/10.1029/2007GL030356>
- Forsberg R, Olesen AV, Einarsson I, Manandhar N, Shrestha K (2011) Geoid of Nepal from airborne gravity survey. In: *Proc. IAG symposia* 139. Springer, pp 521–528
- Fowler CMR (1996) *The solid Earth*. Cambridge University Press, Cambridge, pp 214–216
- Han SC, Shum CK, Bevis M, Ji C, Kuo C-Y (2006) Crustal dilatation observed by GRACE after 2004 Sumatra-Andaman earthquake. *Science* 313(5787):658–666. <https://doi.org/10.1126/science.1128661>
- <http://earthobservatory.nasa.gov/Features/GRACE>
- Humphreys ED, Coblentz DD (2007) North American dynamics and western U.S. tectonics. *Rev Geophys* 45(RG3001):1–30
- Ince ES, Barthelmes Franx, ReiBland Sven, Elger Kirschen, Forste Christoph, Flechtner Frank, Schuh Harald (2019) ICGEM – 15 years of successful collection and distribution of global gravitational models, associated services, and future plans. *Earth Syst Sci Data* 11:647–674. <https://doi.org/10.5194/essd-11-647-2019>
- Listing JB (1873) *Über unsere jetzige Kenntniss der Gestsalt und Grosse der Erde*, Nachrichten von der Königl. Gesellschaft der Wissenschaften und der G A Universität zu Göttingen 3:33–98
- Molodensky MS, Yeremeev VF, Yurkina MI (1960) *Methods for study of the External Gravitational Field and Figure of the Earth*. (Translated from Russian by the Israel Program for Scientific Translation, Office of Technical Services, Department of Commerce, Washington, D.C., 1962)
- Rummel R (2010) GOCE: Gravitational Gradiometry in a Satellite. In: Freeden W, Nashed M.Z., Sonar T (eds) *Handbook of Geomathematics*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-01546-5_4
- Stokes GG (1849) On the variation of gravity at the surface of the Earth. *Trans Camb Philos Soc* 8:672
- Tapley BD, Bettadpur S, Ries J, Thomson PF, Watkins MM (2004) GRACE measurements of mass variability in the Earth system. *Science* 305(5683):503–505
- Turcotte DL, Schubert G (2002) *Geodynamics*, 2nd edn. Cambridge University Press, Cambridge

Geoid Determination, Computational Methods

Michael G. Sideris

Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada

Definition

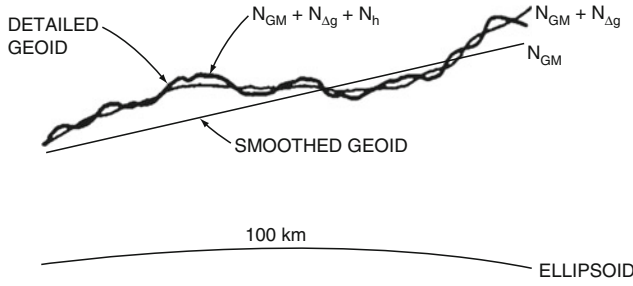
<i>The geoid</i>	is the equipotential surface of the Earth's gravity field, which best approximates the mean sea level at rest.
<i>Geoid determination</i>	encompasses the methodology and computational techniques used to calculate the values of the geoid undulations in particular points or regions.

Introduction

In this entry, we are presenting numerical methods for estimating geoid undulations and other relevant quantities, based on the theory presented in the article ► “[Geoid Determination, Theory and Principles](#)”. We will refer to formulas in that article by the equation number followed by Sideris and then by ► “[Geoid Determination, Theory and Principles](#)”. For example, Stokes's equation will be referred to as Eq. 17 in Sideris, ► “[Geoid Determination, Theory and Principles](#)”.

The Remove-Compute-Restore Technique

In practice, we can evaluate Stokes's equation by numerical integration using only discrete data; therefore, short-wavelength information will not be properly represented in the solution (it will be aliased). In addition, we use data only in a certain region, and therefore the long-wavelength contributions of the gravity field will not be present in the results and must be computed in another way. These problems can be remedied by using a global geopotential model (GM) and heights, H , in a digital elevation model (DEM) to obtain the long and short wavelengths of the field, respectively. This leads to the standard *remove-compute-restore technique*, in which the long and short wavelengths are first removed from the gravity anomalies in a preprocessing step, then the residual anomalies Δg_r are used to compute residual undulations $N_{\Delta g_r}$, and then the long- and short-wavelength contributions are restored in a post-processing step. The computation is based on the following formula:



Geoid Determination, Computational Methods, Fig. 1 Contributions of different data to regional geoid determination

$$N = N_{GM} + N_{\Delta g_r} + N_H, \Delta g_r = \Delta g - \Delta g_{GM} - \Delta g_H \quad (1)$$

where Δg_H denotes the terrain contribution and N_H the corresponding indirect effect on the geoid; see Fig. 1. It should also be noted that such residual anomalies are typically smaller and smoother and therefore appropriate for use in interpolation and in the creation of gridded values by, e.g., least-squares collocation.

Computation of the GM-Contributions

The geopotential model parts of Δg and N are usually computed by the spherical approximation of Eqs. 15 and 16 in Sideris, ▶ “Geoid Determination, Theory and Principles” from the coefficients of a GM of maximum degree $n_{\max}$, which is a function of the resolution of the data that were used to compute the coefficients C_{nm} , S_{nm} :

$$N_{GM_P} = R \sum_{n=2}^{n_{\max}} \sum_{m=0}^n [C_{nm} \cos(m\lambda_P) + S_{nm} \sin(m\lambda_P)] P_{nm}(\sin \phi_P) \quad (2)$$

$$\Delta g_{GM_P} = G \sum_{n=2}^{n_{\max}} (n-1) \sum_{m=0}^n [C_{nm} \cos(m\lambda_P) + S_{nm} \sin(m\lambda_P)] P_{nm}(\sin \phi_P) \quad (3)$$

G and R are the mean gravity and mean radius of the sphere, and ϕ_P , λ_P are the spherical latitude and longitude of the computation point. This representation causes an omission or truncation error, but there is also a commission error because the data used to estimate the coefficients, and therefore also the coefficients themselves, do contain errors. Naturally, the spherical approximation itself causes an additional error, and therefore Eqs. 15 and 16 in Sideris, ▶ “Geoid Determination, Theory and Principles” are preferable to Eqs. 2 and 3.

A comprehensive discussion on the determination and proper use of GM coefficients can be found in ▶ “Spherical Harmonic Analysis Applied to Potential Fields”, in

▶ “Gravity, Global Models”, and in Chapter 6 in Sansò and Sideris (2013). Information on the high-resolution EGM08 model is given in Pavlis et al. (2012), and a complete set of freely available GMs is provided by the ICGEM (International Centre for Global Earth Models) Service of the International Association of Geodesy at <http://icgem.gfz-potsdam.de/home>.

Computation of the Δg -Contribution

The contribution of gravity anomalies can be computed by use of either Stokes’s integral or least-squares collocation (LSC). For simplicity in the notation, we will often write just N and Δg instead of $N_{\Delta g_r}$ and Δg_r in the equations, respectively.

Evaluation of Stokes’s integral. Stokes’s integral can be evaluated either on the sphere or on the plane by numerical integration using data on circular sectors (ring integration method) or on regular grids. Since gravity data are typically available on grids, and because Stokes’s integral is a convolution integral on the plane and along parallels on the sphere, we will discuss here the integral’s evaluation on regular grids by use of the very efficient *fast Fourier transform* (FFT) *technique* (Bracewell 2000).

On the sphere, using Eqs. 17, 18, and 19 in Sideris, ▶ “Geoid Determination, Theory and Principles”, we can write Stokes’s integral explicitly as

$$N_P = \frac{R}{4\pi\gamma} \iint_{\sigma} \Delta g S(\phi_P, \phi, \lambda - \lambda_P) \cos \phi d\phi d\lambda \quad (4)$$

which is a convolution between $\Delta g \cos \phi$ and the Stokes kernel S only along parallels. For Δg_r given on $M \times N$ points on a spherical grid with spacings $\Delta \phi$, $\Delta \lambda$, Eq. 4 can be written for every parallel of constant latitude ϕ_l as

$$\begin{aligned} N_{\Delta g_r}(\phi_l, \lambda_k) &= \frac{R}{4\pi\gamma} \sum_{j=0}^{N-1} \left[\sum_{i=0}^{M-1} \Delta g_r(\phi_j, \lambda_i) \cos \phi_j S(\phi_l, \phi_j, \lambda_k - \lambda_i) \Delta \lambda \right] \Delta \phi \\ &= \frac{R}{4\pi\gamma} \sum_{j=0}^{N-1} \left[(\Delta g_r(\phi_j, \lambda_k) \cos \phi_j) * S(\phi_l, \phi_j, \lambda_k) \right] \Delta \phi, \\ \phi_l &= \phi_1, \phi_2, \dots, \phi_N \end{aligned} \quad (5)$$

The brackets in Eq. 5 contain a one-dimensional (1D) discrete convolution with respect to λ , i.e., along a parallel. Therefore, the above discrete convolution, denoted by $*$, can be evaluated by the 1D FFT (Strang van Hees 1990; Haagmans et al. 1993) as follows:

$$N_{\Delta g_r}(\phi_l, \lambda_k) = \frac{R}{4\pi\gamma} F_\lambda^{-1} \left\{ \sum_{j=0}^{N-1} F_\lambda \left\{ \Delta g_r(\phi_j, \lambda_k) \cos \phi_j \right\} \right. \\ \left. F_\lambda \left\{ S(\phi_l, \phi_j, \lambda_k) \right\} \right\} \Delta \phi, \\ \phi_l = \phi_1, \phi_2, \dots, \phi_N \quad (6)$$

where F_λ denotes the 1D direct Fourier transform and F_λ^{-1} the 1D inverse Fourier transform along parallels of constant latitude (Sideris 2016). Although Eq. 6 is rigorous, its evaluation is not as efficient as it could be, as it still requires numerical summation along meridians. By employing various approximations for the Stokes kernel function, such as, e.g., $\cos \phi_P \cos \phi = \cos^2 \bar{\phi}$, where $\bar{\phi}$ is the mean latitude of the computation area, it is possible to transform Eq. 6 into a two-dimensional (2D) convolution. In this case, accuracy is traded for efficiency. Detailed formulas can be found in Forsberg and Sideris (1993).

In planar approximation, which can be employed when the data is available in a limited area E with extent of no more than several degrees in each direction, the Stokes kernel can be approximated by its first term $1/\sin(\psi/2)$. Then the kernel and the differential area can be converted to their planar equivalents, i.e., $S(\psi) \approx 1/\sin(\psi/2) \approx 2/\psi \approx 2R/l$, and $Rd\sigma = dx dy$, yielding the following *planar form of Stokes's integral* (Schwarz et al. 1990; Sideris 2016):

$$N_P = \frac{1}{2\pi\gamma} \iint_E \frac{\Delta g}{l} dx dy = \frac{1}{2\pi\gamma} \Delta g_P * l_{N_P} \quad (7)$$

where l is the planar distance between data and computation point and l_N is the planar form of Stokes's kernel:

$$l_N = (x^2 + y^2)^{-1/2} \quad (8)$$

On a grid with spacings $\Delta x, \Delta y$, the above 2D convolution integral takes the discrete form

$$N_{\Delta g_r}(x_k, y_l) = \frac{1}{2\pi\gamma} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \Delta g_r(x_i, y_j) l_N(x_k - x_i, y_l - y_j) \\ \Delta x \Delta y = \frac{1}{2\pi\gamma} \Delta g_r(x_k, y_l) * l_N(x_k, y_l) \quad (9)$$

$$l_N(x_k - x_i, y_l - y_j) = \begin{cases} \left[(x_k - x_i)^2 + (y_l - y_j)^2 \right]^{-1/2}, & x_k \neq x_i \text{ or } y_l \neq y_j \\ 0, & x_k = x_i \text{ and } y_l = y_j \end{cases} \quad (10)$$

Geoid undulations can then be evaluated by the 2D FFT as follows (Sideris 2016):

$$N_{\Delta g_r}(x_k, y_l) = \frac{1}{2\pi\gamma} F^{-1} \{ F \{ \Delta g_r(x_k, y_l) \} F \{ l_N(x_k, y_l) \} \} \quad (11a)$$

or, in more compact form (which we will use for all convolutions from this point on)

$$N_P = \frac{1}{2\pi\gamma} F^{-1} \{ F \{ \Delta g_P \} F \{ l_{N_P} \} \} \quad (11b)$$

Because $S(\psi)$ and l_N are singular at the origin, the contribution of the gravity anomaly of the computation point must be evaluated separately and added to the results of the convolutions. On the sphere, Hofmann-Wellenof and Moritz (2006) give this contribution approximately as

$$dN_{\Delta g_r}(\phi_P, \lambda_P) \approx \gamma^{-1} s_o \Delta g_r(\phi_P, \lambda_P) \quad (12)$$

where s_o denotes the radius of a small area around the computation point in which Δg_r can be considered constant. On the plane, the corresponding equation would be

$$dN_{\Delta g_r}(x_P, y_P) \approx \gamma^{-1} \sqrt{\frac{\Delta x \Delta y}{\pi}} \Delta g_r(x_P, y_P) \quad (13)$$

Equation 11 employs the numerical evaluation of the spectrum of Stokes's kernel and is the one that should be used in practice (with proper zero padding of the data) if we want the results from FFT to be identical to those from numerical integration; see Li and Sideris (1993). This is in fact true for any of the other convolution integrals that we are discussing here. Nevertheless, it is well-known from Fourier theory that the inverse distance function l_N has an analytical spectrum, which is the inverse frequency q^{-1} in the frequency domain. Using u, v, w to denote the frequencies corresponding to x, y, z , this spectrum is

$$F \{ l_N(x, y) \} = F \left\{ (x^2 + y^2)^{-1/2} \right\} = (u^2 + v^2)^{-1/2} = q^{-1} \quad (14)$$

and therefore Eq. 7 can be evaluated as follows:

$$N_P = \frac{1}{2\pi\gamma} F^{-1} \{ F \{ \Delta g_P \} q^{-1} \} \quad (15)$$

This representation clearly shows that Stokes's integral is a smoothing, mainly low-frequency-dependent operator, and therefore geoid undulations are more sensitive to the long wavelengths of the gravity field.

The FFT method can use either mean or point data on a grid and computes simultaneously the geoidal heights at all grid points; see Sideris and Tziavos (1988) for the formulas with mean data. A large area can thus be covered very fast and with minimum cost. An interpolation technique such as LSC must be employed for predicting geoidal heights at points not on the grid and/or to grid irregular gravity anomalies. With gridded data, the FFT technique is by far the fastest of all available methods. Criticism that only the FFT suffers from leakage and aliasing effects are not valid since these effects depend on data density and coverage and thus affect all methods.

Given gridded gravity and topography data, the Molodensky solution of Eqs. 20, 21, and 22 in Sideris, ► “Geoid Determination, Theory and Principles” can also be efficiently evaluated by FFT. Details can be found in Sideris (1987) and Schwarz et al. (1990).

Computation by least-squares collocation. A solution can also be obtained using least-squares collocation (LSC). For example, in the simple case that we only want to predict residual geoid undulations from residual gravity anomalies, Eqs. 26 and 27 in Sideris, ► “Geoid Determination, Theory and Principles” reduce to

$$\hat{N}_{Ag} = \mathbf{C}_{N\Delta g} (\mathbf{C}_{\Delta g\Delta g} + \mathbf{C}_{nn})^{-1} \Delta \mathbf{g} \quad (16)$$

$$\mathbf{C}_{ee} = \mathbf{C}_{NN} - \mathbf{C}_{N\Delta g} (\mathbf{C}_{\Delta g\Delta g} + \mathbf{C}_{nn})^{-1} \mathbf{C}_{\Delta gN} \quad (17)$$

Given the covariance matrix for T , $\mathbf{C}_{TT}$, the other covariance matrices can be derived by covariance propagation using Eqs. 12 and 13 in Sideris, ► “Geoid Determination, Theory and Principles” as follows:

$$\begin{aligned} \mathbf{C}_{NN} &= \frac{1}{\gamma^2} \mathbf{C}_{TT}, \\ \mathbf{C}_{N\Delta g} &= \frac{1}{\gamma} \left(-\frac{\partial \mathbf{C}_{TT}}{\partial h} + \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} \mathbf{C}_{TT} \right), \\ \mathbf{C}_{\Delta g\Delta g} &= \left(-\frac{\partial}{\partial h} + \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} \right) \left(-\frac{\partial \mathbf{C}_{TT}}{\partial h} + \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} \mathbf{C}_{TT} \right) \end{aligned} \quad (18)$$

LSC’s drawback is the need for inverting large covariance matrices or simply solving large systems of linear equations, which in practice often prevents the utilization of all given data, and thus prevents the application of LSC over large regions. To circumvent this problem, if gridded data are available, LSC can be applied in the frequency domain. This is actually related directly to the use of FFT formulas with noisy, heterogeneous data. It will in fact be shown below that the FFT method can use heterogeneous data, provided that they are given on a grid, and can produce error estimates, provided that the power spectral density (PSD) functions, which are the Fourier transform of the covariance functions, of the data and their noise are known. In this case, the

technique is equivalent to frequency-domain collocation; see also Sideris (2016).

Computation by frequency-domain collocation. To illustrate how error propagation can be used with Stokes’s integral, we assume that the anomalies have errors n with known PSD P_{nn} . In this case, Stokes’s convolution of, e.g., Eq. 7 becomes

$$N = (\Delta g + n) * s + e \quad (19)$$

where $s = l_N/(2\pi\gamma)$ and e is the error of the estimated undulations.

Making use of the definition of the PSD function P_{fg} for two signals f and g (Bendat and Piersol 2010) as

$$P_{fg} = F\{f\}F\{g\}^* \quad (20)$$

where the superscript $*$ denotes complex conjugate, and assuming no correlation between the signal and noise, the PSD of N and the cross-PSD of Δg and N can be derived as follows (Sideris 2016):

$$P_{NN} = P_{ee} + S(P_{\Delta g\Delta g} + P_{nn})S^* \quad (21)$$

$$P_{\Delta gN} = P_{ee} + S(P_{\Delta g\Delta g} + P_{nn}) \quad (22)$$

where $S = F\{s\}$ and $P_{\Delta g\Delta g}$ is the PSD of the gravity anomalies. From the above equations, the spectrum of Stokes’s kernel and the output-noise PSD can be obtained as

$$S = \frac{P_{N\Delta g}}{P_{\Delta g\Delta g} + P_{nn}} \quad (23)$$

$$P_{ee} = P_{NN} - P_{N\Delta g}(P_{\Delta g\Delta g} + P_{nn})^{-1}P_{N\Delta g} \quad (24)$$

Now Eq. 19 can be evaluated in the frequency domain as follows:

$$F\{N\} = P_{N\Delta g}(P_{\Delta g\Delta g} + P_{nn})^{-1}F\{\Delta g\} \quad (25)$$

One can see immediately that Eqs. 24 and 25 are, formally, the frequency-domain equivalents of the LSC Eqs. 16 and 17. In fact, they can also be obtained starting from the general form of the equation for P_{ee} and estimating that $S = F\{s\}$ which makes P_{ee} minimum; such a derivation can be found in Sideris (1996). What makes them much more efficient, though, is that they require no matrix inversion as the -1 superscript in Eqs. 24 and 25 implies straightforward division at each frequency. Notice also that for noiseless data ($P_{nn} = 0$), Eq. 24 yields $P_{ee} = 0$, and Eq. 23 becomes $S = P_{N\Delta g}/P_{\Delta g\Delta g} = F\{N\}/F\{\Delta g\} = (2\pi\gamma q)^{-1}$. Note, however, that for the PSDs to be the Fourier transform of the covariance functions (CVs) used in collocation, these CVs have to be

stationary, which is not the case in practice with the noise CVs. Thus, the FFT method, although it is much more efficient than LSC, has to approximate nonstationary noise covariance functions (which are easily handled by LSC) by stationary ones; for details, see Sansò and Sideris (1997) and Kotsakis and Sideris (2001).

Computation of the Terrain Contributions

The terrain effect on Δg and the indirect effect on N are computed from approximations to Eqs. 32 and 33 in Sideris, ▶ “Geoid Determination, Theory and Principles”. Keeping only the terms for $r = 1$, these equations take the following form:

$$-\Delta g_{H_P} = c_P = \pi k \mathbf{L} [\rho (H - H_P)^2] \quad (26)$$

$$\delta N_{H_P} = -\frac{\pi k \rho}{\gamma} H_P^2 - \frac{\pi k}{3\gamma} \mathbf{L} [\rho (H^3 - H_P^3)] \quad (27)$$

With height data given in small area E around the computation point P , the planar approximation of the $\mathbf{L}$ operator can be used. In this case, the above equations become (Sideris 1990, 2016)

$$\begin{aligned} c_P &= \frac{1}{2} k \iint_E \rho \frac{(H - H_P)^2}{l^3} = \frac{1}{2} k \iint_E \frac{\rho H^2 - 2\rho H H_P + \rho H_P^2}{l^3} dx dy \\ &= \frac{1}{2} k [(\rho_P H_P^2) * l_P^{-3} - 2H_P [(\rho_P H_P) * l_P^{-3}] + H_P^2 (\rho_P * l_P^{-3})] \end{aligned} \quad (28)$$

$$\begin{aligned} \delta N_{H_P} &= -\frac{\pi k \rho}{\gamma} H_P^2 - \frac{k}{6\gamma} \iint_E \frac{\rho H^3 - \rho H_P^3}{l^3} dx dy \\ &= -\frac{\pi k \rho}{\gamma} H_P^2 - \frac{k}{6\gamma} [(\rho_P H_P^3) * l_P^{-3} - H_P^3 (\rho_P * l_P^{-3})] \end{aligned} \quad (29)$$

where $l_P^{-3} = (x_P^2 + y_P^2)^{-3/2}$. The above convolutions can be evaluated by FFT as follows:

$$\begin{aligned} c_P &= \frac{1}{2} k [F^{-1} \{F\{\rho_P H_P^2\} F\{l_P^{-3}\}\} + H_P^2 F^{-1} \{F\{\rho_P\} F\{l_P^{-3}\}\}] \\ &\quad - k H_P [F^{-1} \{F\{\rho_P H_P\} F\{l_P^{-3}\}\}] \end{aligned} \quad (30)$$

$$\begin{aligned} \delta N_{H_P} &= -\frac{\pi k \rho}{\gamma} H_P^2 - \frac{k}{6\gamma} [F^{-1} \{F\{\rho_P H_P^3\} F\{l_P^{-3}\}\} \\ &\quad - H_P^3 F^{-1} \{F\{\rho_P\} F\{l_P^{-3}\}\}] \end{aligned} \quad (31)$$

From the definition of the vertical derivative operator $\mathbf{L}$ in Eq. 21 in Sideris, ▶ “Geoid Determination, Theory and Principles” and the derivative property of the Fourier transform (Bracewell 2000; Sansò and Sideris 2013, Chapter 10, Appendix A), we can see that

$$F\{\mathbf{L}f\} = F\left\{\frac{\partial f}{\partial z}\right\} = 2\pi i w F\{f\} = -2\pi q F\{f\} \quad (32)$$

where i is the imaginary unit. We have also used the relationship $i w = -q$, which is a straightforward result of Laplace’s equation in the frequency domain:

$$F\{\nabla^2 f\} = [(i2\pi u)^2 + (i2\pi v)^2 + (i2\pi w)^2] F\{f\} = 0 \quad (33)$$

Now Eqs. 26 and 27 can also be evaluated by FFT using the analytical spectrum of the $\mathbf{L}$ operator as follows:

$$\begin{aligned} c_P &= k\pi [F^{-1} \{-2\pi q F\{\rho_P H_P^2\}\} - 2H_P F^{-1} \{-2\pi q F\{\rho_P H_P\}\} \\ &\quad + H_P^2 F^{-1} \{-2\pi q F\{\rho_P\}\}] \end{aligned} \quad (34)$$

$$\begin{aligned} \delta N_{H_P} &= -\frac{\pi k}{\gamma} \rho_P H_P^2 - \frac{\pi k}{3\gamma} [F^{-1} \{-2\pi q F\{\rho_P H_P^3\}\} \\ &\quad - H_P^3 F^{-1} \{-2\pi q F\{\rho_P\}\}] \end{aligned} \quad (35)$$

Such analytical representations of kernel functions, though not recommended for use in limited areas (Li and Sideris 1993), are often used in satellite altimetry applications, as discussed in the section that follows.

Evaluation of Satellite Altimetry Contributions

Like Stokes’s formula, the inverse Stokes, inverse Vening Meinesz, and deflection-geoid formulas employ kernels that are functions of the spherical distance ψ between the data and computation points; see Eq. 19 in Sideris, ▶ “Geoid Determination, Theory and Principles”. Therefore, when gridded data is available, they can be efficiently evaluated on the sphere by either the rigorous 1D FFT or the approximate 2D FFT formulas, completely analogous to Eqs. 4–9 used for Stokes’s integral. They will not be discussed further here, and the interested reader is referred to Hwang (1998) for detailed equations. Instead, we will concentrate here on the planar approximations with analytical kernel spectra.

On the plane, employing the same approximation for the sine and cosine of ψ as we did for Stokes’s formula, Eqs. 36–40 in Sideris, ▶ “Geoid Determination, Theory and Principles” simplify to

$$\Delta g_P = -\frac{\gamma}{2\pi} \iint_E \frac{N - N_P}{l^3} dx dy = -\gamma \mathbf{L}N = -\gamma \frac{\partial N}{\partial z} \quad (36)$$

$$\begin{aligned} \Delta g_P &= -\frac{\gamma}{2\pi} \iint_E \left(\xi \frac{y - y_P}{l^3} + \eta \frac{x - x_P}{l^3} \right) dx dy \\ &= -\frac{\gamma}{2\pi} (\xi_P * l_{\xi_P} + \eta_P * l_{\eta_P}) \end{aligned} \quad (37)$$

$$l_\xi = y/(x^2 + y^2)^{3/2} = -\partial l_N / \partial y, l_\eta = x/(x^2 + y^2)^{3/2} = -\partial l_N / \partial x \quad (38)$$

$$\begin{aligned} N_P &= -\frac{1}{2\pi} \iint_E \left(\xi \frac{y - y_P}{l^2} + \eta \frac{x - x_P}{l^2} \right) dx dy \\ &= -\frac{1}{2\pi} (\xi_P * l_{\xi_P} + \eta_P * l_{\eta_P}) \end{aligned} \quad (39)$$

$$l_y = y/(x^2 + y^2), l_x = x/(x^2 + y^2) \quad (40)$$

By making use of the derivative property of the Fourier transform, their evaluation by FFT is done as follows (Hwang 1998):

$$\Delta g_P = \gamma F^{-1} \{2\pi q F\{N_P\}\} \quad (41)$$

$$\Delta g_P = \gamma F^{-1} \left\{ \frac{i}{q} (vF\{\xi_P\} + uF\{\eta_P\}) \right\} \quad (42)$$

$$N_P = F^{-1} \left\{ \frac{i}{2\pi q^2} (vF\{\xi_P\} + uF\{\eta_P\}) \right\} \quad (43)$$

The planar form of Eq. 42 in Sideris, ► “Geoid Determination, Theory and Principles”, namely

$$\frac{\partial \Delta g}{\partial z} = -\gamma \left(\frac{\partial \xi}{\partial y} + \frac{\partial \eta}{\partial x} \right) \quad (44)$$

can also be evaluated by FFT from the expression

$$\left. \frac{\partial \Delta g}{\partial z} \right|_P = -2\pi \gamma F^{-1} \{ivF\{\xi_P\} + iuF\{\eta_P\}\} \quad (45)$$

Finally, Eq. 35 in Sideris, ► “Geoid Determination, Theory and Principles” that gives the deflections from altimetry-derived undulations is in planar approximation

$$\xi = -\frac{\partial N}{\partial y}, \eta = -\frac{\partial N}{\partial x} \quad (46)$$

and therefore its FFT-evaluation can be done via the expression

$$\xi_P = -2\pi F^{-1} \{ivF\{N_P\}\}, \eta_P = -2\pi F^{-1} \{iuF\{N_P\}\} \quad (47)$$

Summary

In this entry, we have outlined the steps in the remove-compute-restore technique that is employed in the numerical computations of geoid undulations. We have shown that most of the equations expressing the terrain and gravity effects on the geoid are convolution integrals, and we have described in detail how they can be efficiently evaluated by FFT given gridded data on the sphere or on the plane. In addition, computations by least-squares collocation have been outlined both in the space and in the frequency domain, in which the formula is equivalent to Wiener filtering and provides both noise filtering and accuracy estimates of the geoid through propagation of the data noise to the results. For a detailed treatment of FFT-based methods, the interested reader is referred to Chapter 10 in Sansò and Sideris (2013) and to Sideris (2016). The entry closes with the evaluation of various formulas used in gravity and geoid determination in the oceans using data from satellite altimetry.

Cross-References

- [Geodesy, Figure of the Earth](#)
- [Geodesy, Physical](#)
- [Geoid Determination, Theory and Principles](#)
- [Geoid Undulation, Interpretation](#)
- [Gravimeters](#)
- [Gravity Anomalies, Interpretation](#)
- [Gravity Data, Advanced Processing](#)
- [Gravity Field of the Earth](#)
- [Gravity Method, Satellite](#)
- [Gravity, Data to Anomalies](#)
- [Gravity, Global Models](#)
- [Spherical Harmonic Analysis Applied to Potential Fields](#)

Bibliography

- Bendat JS, Piersol AG (2010) Random data – analysis and measurement procedures, 4th edn. Wiley, Hoboken, New Jersey
- Bracewell RN (2000) The Fourier transform and its application, 3rd edn. McGraw-Hill, New York
- Forsberg R, Sideris MG (1993) Geoid computations by the multi-banding spherical FFT approach. Manuscr Geodaet 18:82–90
- Haagmans R, de Min E, van Gelderen M (1993) Fast evaluation of convolution integrals on the sphere using ID FFT, and a comparison with existing methods for Stokes’ integral. Manuscr Geodaet 18:227–241
- Hofmann-Wellenhof B, Moritz H (2006) Physical geodesy, 2nd edn. Springer Wien, New York

- Hwang C (1998) Inverse Vening Meinesz formula and deflection-geoid formula: applications to the predictions of gravity and geoid over the South China Sea. *J Geod* 72:304–312
- Kotsakis C, Sideris MG (2001) A modified Wiener-type filter for geodetic estimation problems with non-stationary noise. *J Geod* 75(12):647–660
- Li YC, Sideris MG (1993) Gravity field convolutions without windowing and edge effects. *Manuscr Geodaet* 67:107–118
- Moritz H (1980) *Advanced physical geodesy*. Karlsruhe: Wichmann (reprinted 2008; Columbus, OH: School of Earth Sciences, Ohio State University)
- Pavlis NK, Holmes SA, Kenyon SC, Factor JK (2012) The development and evaluation of the earth gravitational model 2008 (EGM2008). *J Geophys Res* 117:B04406
- Sansò F, Sideris MG (1997) On the similarities and differences between systems theory and least-squares collocation in physical geodesy. *Bollettino di Geodesia e Scienze Affini* LVI(2):173–206
- Sansò F, Sideris MG (eds) (2013) *Geoid Determination – Theory and Methods*, Lecture Notes in Earth System Sciences, vol 110. Springer, Berlin, Heidelberg
- Schwarz KP, Sideris MG, Forsberg R (1990) The use of FFT in physical geodesy. *Geophys J Int* 100:485–514
- Sideris MG (1987) Spectral methods for the numerical solution of Molodensky's problem. UCSE report no. 20024, Department of Surveying Engineering, The University of Calgary, Calgary, Alberta
- Sideris MG (1990) Rigorous gravimetric terrain modelling using Molodensky's operator. *Manuscr Geodaet* 15:97–106
- Sideris MG (1996) On the use of heterogeneous noisy data in spectral gravity field modelling methods. *J Geod* 70(8):470–479
- Sideris MG (2016) The FFT in local gravity field determination. In: Grafarend EW (ed) *Encyclopedia of Geodesy*. Springer International Publishing, Switzerland
- Sideris MG, Tziavos IN (1988) FFT-evaluation and applications of gravity-field convolution integrals with mean and point data. *Bulletin Geodesique* 62:521–540
- Strang van Hees G (1990) Stokes' formula using fast Fourier techniques. *Manuscr Geodaet* 15:235–239

Geoid Determination, Theory and Principles

Michael G. Sideris

Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada

Definition

The *geoid* is the equipotential surface of the Earth's gravity field, which best approximates the mean sea level at rest. Its determination requires (i) knowledge of the Earth's gravity potential and (ii) a way of geometrically representing the geoidal surface, which is done via its deviations – called geoid undulations N – from the surface of a known reference ellipsoid. The geometric representation is of particular importance given that the geoid is the reference surface, or datum, for the physically meaningful orthometric heights (see chapter ► “Height Systems, Vertical Datums and Their Unification”), often called heights above mean sea level.

Basic Principles: the Gravity, Normal, and Disturbing Potentials

The Earth's *gravity potential* W comprises the gravitational potential V and the centrifugal potential Φ :

$$W = V + \Phi = k \iiint_V \frac{\rho}{l} dv + \frac{1}{2} \omega^2 (x^2 + y^2) \quad (1)$$

V is a Newtonian potential due to the attraction of the masses, expressed as an integral of the mass density ρ over the volume v ; l is the Cartesian distance between the computation point and the mass element ρdv ; and k is Newton's gravitational constant. The centrifugal potential Φ is a function of the Earth's rotational velocity ω and the distance $p = (x^2 + y^2)^{1/2}$ from the rotation axis, namely, $\Phi = \omega^2 p^2/2$.

Due to our insufficient knowledge of the density distribution, Eq. 1 cannot be used directly to compute V and W . Nevertheless, a solution can be obtained by utilizing the fact that V is a harmonic function outside the masses, that is (a) it satisfies Laplace's equation

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0 \quad (2)$$

since the inverse distance in the integral of Eq. 1 clearly satisfies $\nabla^2 l^{-1} = 0$ and (b) it vanishes as $l \rightarrow \infty$. *Stokes's theorem* postulates that V can be determined everywhere in free space from its values on the boundary surface that encloses all masses and thus knowledge of the interior density distribution is not necessary.

The above problem formulation is called *Dirichlet's problem*, or first boundary value problem (BVP) of potential theory (Hofmann-Wellenhof and Moritz 2006). Its practical significance is rather limited since the potential is not a directly measurable quantity; it is important, however, for the expansion of V into a spherical harmonic series and for the relationship between V on the boundary surface and V in space (Poisson's integral). A more realistic BVP is *Neumann's problem*, or second BVP, in which the normal derivative $\partial V/\partial n$ is given on the boundary surface instead of V itself. This is because the first-order derivatives of W constitute the *gravity vector*

$$\mathbf{g} = \nabla W = \left(\frac{\partial W}{\partial x}, \frac{\partial W}{\partial y}, \frac{\partial W}{\partial z} \right)^T \quad (3)$$

which is observable. It is, however, the third BVP that is particularly important in practice; it uses observables on the boundary surface that are a linear combination of V and $\partial V/\partial n$. We will see in the following that the determination of geoid undulations from gravity anomalies – which are derived from the measurements taken with a gravimeter – is such a problem.

The determination of W is a nonlinear problem and is further complicated by the fact that W itself is not harmonic since $\nabla^2 W = \nabla^2 \Phi = 2\omega^2$. Both limitations can be overcome by linearization and approximating the shape, size, and gravity field of the Earth by an equipotential ellipsoid of revolution with the same mass M and same rotational velocity ω as the Earth. In this case, the gravitational potential of the reference ellipsoid V_e can be computed analytically, and its centrifugal potential Φ_e is the same as Φ . Its gravity potential, termed *normal potential* U , is

$$U = V_e + \Phi_e = V_e + \Phi \quad (4)$$

and its gradient is the *normal gravity vector*

$$\gamma = \nabla U = \left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial U}{\partial z} \right)^T \quad (5)$$

Both U and γ are analytical functions that can be computed from M , ω , and the semimajor and semiminor axes a , b of the reference ellipsoid; see Sansò and Sideris (2013), section 1.9. The difference between W and U at the same point is the *disturbing potential* T , which is obviously a harmonic function (since it is independent of the centrifugal potential, and since V_e , just like V , is harmonic being a function of the inverse distance):

$$T = W - U = V - V_e \quad (6)$$

$$\nabla^2 T = 0 \quad (7)$$

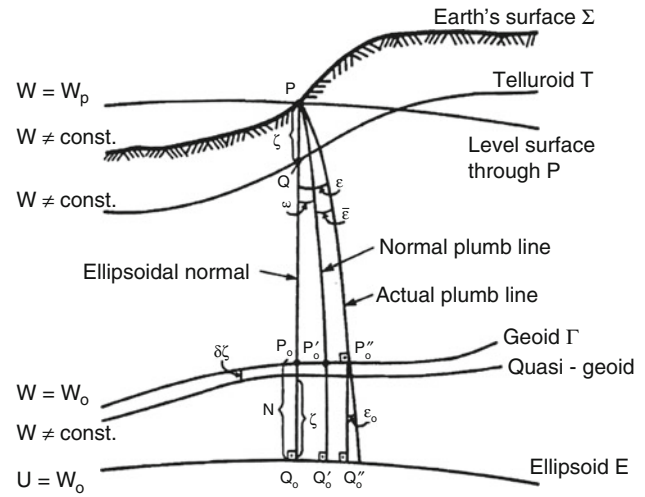
Now since U is known, the problem of determining W has been reduced to the problem of determining T , which is a third BVP provided that an appropriate observable would be available on the boundary surface. This is indeed the case, as it will be shown in the following.

The Stokes and Molodensky Geodetic BVPs

We select the geoidal surface Γ as the boundary surface that encloses all masses, we approximate Γ by the surface of the reference ellipsoid E , choose the normal potential U_o on E to be the same as the gravity potential W_o on the geoid, project a point P_o on the geoid on to a point Q_o on the ellipsoid along the ellipsoidal normal (see Fig. 1), and define the geoid undulation N as the distance $P_o Q_o$ and the *gravity anomaly* and *gravity disturbance* vectors as $\Delta \mathbf{g} = \mathbf{g}_{P_o} - \boldsymbol{\gamma}_{Q_o}$ and $\delta \mathbf{g} = \mathbf{g}_{P_o} - \boldsymbol{\gamma}_{P_o} = \nabla T$, respectively, with magnitudes

$$\Delta g = g_{P_o} - \gamma_{Q_o} \quad (8)$$

$$\delta g = g_{P_o} - \gamma_{P_o} \quad (9)$$



Geoid Determination, Theory and Principles, Fig. 1 Geometry of the Stokes BVP and the Molodensky BVP

We can now relate N to T , and Δg to T and its vertical derivative. Using Fig. 1, only first-order terms in Taylor expansions, and assuming – because the vertical deflection angle ε is very small – that the vertical derivatives along the ellipsoidal normal, normal plumb line, and actual plumb line are all the same, we can write

$$T_{P_o} = W_{P_o} - U_{P_o} = W_{P_o} - U_{Q_o} - \frac{\partial U}{\partial h} \Big|_{Q_o} \quad (10)$$

$$N = W_o - U_o + \gamma N = \gamma N$$

$$\delta g_{P_o} = -\frac{\partial T}{\partial h} \Big|_{P_o} = g_{P_o} - \gamma_{P_o} = g_{P_o} - \gamma_{Q_o} - \frac{\partial \gamma}{\partial h} \Big|_{Q_o} T \quad (11)$$

$$N = \Delta g - \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} \Big|_{Q_o} T$$

Plumb lines are lines intersecting perpendicularly the surfaces of equal potential; that is, at every point, the gravity vector is tangent to the plumb line through that point.

The last two equations can now simply be written as

$$N = \frac{T}{\gamma} \quad (12)$$

$$\Delta g = \delta g + \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} T = -\frac{\partial T}{\partial h} + \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} T \quad (13)$$

Equation 12 is called *Bruns's equation* and provides N given T , while Eq. 13 holds only on Γ and it is called the *fundamental equation of physical geodesy*. The normal gravity on an ellipsoid with semi-axes a , b and normal gravity values at the equator and pole γ_a , γ_b , respectively, can be

computed from Somigliana's formula as a function of the point's latitude ϕ :

$$\gamma = \frac{a\gamma_a \cos^2 \phi + b\gamma_b \sin^2 \phi}{(a^2 \cos^2 \phi + b^2 \sin^2 \phi)^{1/2}} \quad (14)$$

Now *Stokes's BVP* can be defined as the problem of determining T , harmonic outside Γ , from measurements of Δg on the geoid that satisfy Eq. 13. This is a third BVP of potential theory, mathematically described by the partial differential equation of Eq. 7 with the boundary condition of Eq. 13, and it provides T on and outside Γ . Because of the assumption that the geoid encloses all masses, the influence of masses outside Γ must be computed and removed mathematically from the measurements via various types of terrain reductions. These require knowledge of the density distribution of the topographic masses and are described in this chapter in section ► [“TOPO-EUROPE: From the Deep Earth to the Surface of Continental Europe and Its Margins”](#).

Since all gravity measurements are taken on the surface of the Earth Σ , a more realistic way of defining the problem is to select Σ as the boundary surface rather than the geoid. This requires no terrain reductions but results in a mathematically more complicated BVP because Σ is not an equipotential surface and the gravity vector is not perpendicular to it. The determination of the Earth's surface from gravity anomalies on it is an oblique derivative problem called *Molodensky's problem* (Molodensky et al. 1962).

In *Molodensky's BVP*, the Earth's surface is approximated by a surface called the *telluroid* T , W on Σ is approximated by the normal potential U on T , and the gravity anomalies now are the difference between the gravity at the Earth's surface and the normal gravity at the telluroid, that is, $\Delta g = g_P - \gamma_Q$; see Fig. 1. The surface of T is defined as the set of points Q above the ellipsoid that correspond to points P on Σ having $U_Q = W_P$, $\phi_Q = \phi_P$, $\lambda_Q = \lambda_P$. This is called the *Marussi telluroid* and the ellipsoidal heights of its points are called *normal heights* H^* . The separation PQ between Σ and T is called *height anomaly* ζ . The height anomaly is the solution to the Molodensky BVP, as defined by Eqs. 7 and 13, where now the boundary condition holds on the telluroid rather than the geoid and Δg refers to the Earth's surface rather than the geoid.

Analytical Solutions to the Geodetic BVPs

Since $\nabla^2 T = 0$, T and therefore N can be expressed as a spherical harmonic series (see chapters ► [“Gravity, Global Models”](#) and ► [“Spherical Harmonic Analysis Applied to Potential Fields”](#))

$$N = \frac{kM}{\gamma r} \sum_{n=2}^{\infty} \left(\frac{a}{r}\right)^n \sum_{m=0}^n [C_{nm} \cos m\lambda + S_{nm} \sin m\lambda] P_{nm}(\sin \phi) \quad (15)$$

the coefficients of which, S_{nm} and C_{nm} , can be determined from, for example, gravity anomalies on the geoid using the orthogonality property of spherical harmonics. P_{nm} are Legendre functions and r , ϕ , λ are the coordinates of the computation point. Applying Eq. 15 to Eq. 13, the spherical harmonic series for Δg is

$$\Delta g = -\frac{kM}{r^2} \sum_{n=2}^{\infty} \left(\frac{a}{r}\right)^n (n-1) \sum_{m=0}^n [C_{nm} \cos m\lambda + S_{nm} \sin m\lambda] P_{nm}(\sin \phi) \quad (16)$$

Another analytical solution of this problem is given by *Stokes's integral* (see chapter ► [“Gravity Field of the Earth”](#)):

$$N = \frac{R}{4\pi\gamma} \iint_{\sigma} \Delta g S(\psi) d\sigma = \frac{1}{\gamma} \mathbf{S} \Delta g \quad (17)$$

where ψ is the spherical distance between data point and computation point, σ is the surface of a sphere of radius R (approximating the ellipsoid), and $\mathbf{S}$ is the Stokes integral operator. $S(\psi)$ is *Stokes's kernel function*

$$S(\psi) = \frac{1}{\sin(\psi/2)} - 6 \sin \frac{\psi}{2} + 1 - 5 \cos \psi - 3 \cos \psi \ln \left(\sin \frac{\psi}{2} + \sin^2 \frac{\psi}{2} \right), \quad (18)$$

$$\sin^2 \frac{\psi}{2} = \sin^2 \frac{\phi_P - \phi}{2} + \sin^2 \frac{\lambda_P - \lambda}{2} \cos \phi_P \cos \phi \quad (19)$$

Molodensky's problem can be solved in various ways. We give here the final solution by the *method of analytical continuation to point level*; detailed derivations can be found in Moritz (1980). In this method, the gravity anomalies are reduced to the equipotential surface passing through the computation point P ; see Fig. 1. They are denoted by $\Delta g'$ and are obtained by the formula

$$\Delta g' = \sum_{n=0}^{\infty} g_n,$$

$$g_n = - \sum_{m=1}^n (h - h_P)^m \mathbf{L}_m g_{n-m}, \quad (20)$$

$$g_o = \Delta g$$

$\mathbf{L}$ is the *vertical derivative operator* expressed in terms of surface values as follows:

$$\mathbf{L}_m = \frac{1}{m!} \mathbf{L}^m = \frac{1}{m!} \frac{\partial^m}{\partial h^m}, \quad (21)$$

$$\mathbf{L}f = -\frac{f}{R} + \frac{R^2}{2\pi} \iint_{\sigma} \frac{f - f_P}{l^3} d\sigma$$

where l is now the straight distance between the running point and the computational point P. Now, ζ at point P can be obtained by applying Stokes's integral operator $\mathbf{S}$ to $\Delta g'$:

$$\zeta = \frac{1}{\gamma} \mathbf{S} \Delta g' = \frac{1}{\gamma} \mathbf{S} \Delta g + \frac{1}{\gamma} \sum_{n=1}^{\infty} \mathbf{S} g_n = \sum_{n=0}^{\infty} \zeta_n \quad (22)$$

When the height anomaly is plotted above the ellipsoid, the resulting surface is called the *quasi-geoid*; see Fig. 1. Since $h = H^* + \zeta \approx H + N$, the geoid undulation can be estimated from the height anomaly by the following expression (Sansò and Sideris 2017):

$$N = \zeta + \delta\zeta, \quad (23)$$

$$\delta\zeta = H^* - H = \frac{\bar{g} - \bar{\gamma}}{\bar{\gamma}} H \approx \frac{\Delta g_B}{\gamma} H$$

where $\bar{g}$ and $\bar{\gamma}$ are the mean values of gravity and normal gravity along the plumb line between the geoid and Earth surface and the normal plumb line between the ellipsoid and the telluroid, respectively, and $\Delta g_B = \Delta g - A_B + F$ is the Bouguer gravity anomaly at the computation point; the definitions of A_B and F are given after Eq. 28. $H = \text{PP}_o''$ and $H^* = \text{QQ}_o$ are the *orthometric height* and *normal height*, respectively, of point P.

The Operational Solution to the Geodetic BVPs

In contrast to the analytical solution which is model-driven, the operational solution is data-driven. It starts from the data and information about their noise and the first and second moments (covariance functions) and estimates T as the “best” unbiased linear combination of the measurements. Besides being able to propagate the data noise into the results, it is also able to utilize as input, as well as predict, any functional of T . The harmonicity of T and the analytical relations between T and its functionals $f(T)$ are only used to derive the required covariance functions.

Mathematically, we want to estimate T , or more generally a signal vector $\mathbf{s}$ of functionals of T , from a vector of measurements

$\mathbf{l}$, which can contain Δg or any other $f(T)$, such as deflections of the vertical, second-order gradients, geoid undulations, etc.

$$\hat{\mathbf{s}} = \mathbf{H} \mathbf{l} \quad (24)$$

In order to estimate the matrix $\mathbf{H}$, we require that the variances of the estimation errors $\mathbf{e} = \hat{\mathbf{s}} - \mathbf{s}$ be minimum or that the error covariance matrix $\mathbf{C}_{ee} = E\{\mathbf{e}\mathbf{e}^T\}$ be minimum. By definition,

$$\begin{aligned} \mathbf{C}_{ee} &= E\{\mathbf{e}\mathbf{e}^T\} = E\{(\mathbf{H}\mathbf{l} - \mathbf{s})(\mathbf{H}\mathbf{l} - \mathbf{s})^T\} \\ &= \mathbf{H}E\{\mathbf{l}\mathbf{l}^T\}\mathbf{H}^T - E\{\mathbf{s}\mathbf{l}^T\}\mathbf{H}^T - \mathbf{H}E\{\mathbf{l}\mathbf{s}^T\} + E\{\mathbf{s}\mathbf{s}^T\} \\ &= \mathbf{H}\mathbf{C}_{ll}\mathbf{H}^T - \mathbf{C}_{sl}\mathbf{H}^T - \mathbf{H}\mathbf{C}_{ls} + \mathbf{C}_{ss} \\ &= \mathbf{C}_{ss} - \mathbf{C}_{sl}\mathbf{C}_{ll}^{-1}\mathbf{C}_{ls} + (\mathbf{H} - \mathbf{C}_{sl}\mathbf{C}_{ll}^{-1})\mathbf{C}_{ll}(\mathbf{H} - \mathbf{C}_{sl}\mathbf{C}_{ll}^{-1})^T \end{aligned} \quad (25)$$

The last term in the right-hand side of the above equation is a nonnegative quadratic term, and therefore $\mathbf{C}_{ee}$ can only be minimum when that term is zero, that is, when $\mathbf{H} = \mathbf{C}_{sl}\mathbf{C}_{ll}^{-1}$. Therefore,

$$\hat{\mathbf{s}} = \mathbf{C}_{sl}\mathbf{C}_{ll}^{-1}\mathbf{l} \quad (26)$$

$$\mathbf{C}_{ee} = \mathbf{C}_{ss} - \mathbf{C}_{sl}\mathbf{C}_{ll}^{-1}\mathbf{C}_{ls} \quad (27)$$

When the observations $\mathbf{l}$ contain noise $\mathbf{n}$, that is, consist of a signal component $\mathbf{t}$ and a noise component $\mathbf{n}$, then assuming that signal and noise are uncorrelated, the covariance matrices in the above equations become $\mathbf{C}_{ll} = \mathbf{C}_{tt} + \mathbf{C}_{nn}$, $\mathbf{C}_{ls} = \mathbf{C}_{ts}$, and $\mathbf{C}_{sl} = \mathbf{C}_{st} = \mathbf{C}_{ls}^T$. This method is generally known by the name of *least-squares collocation* (LSC). Knowledge of the covariance matrix of only one gravity field quantity is sufficient. From this, all other covariance matrices can be derived by covariance propagation. A complete discussion of LSC can be found in Moritz (1980) and in Chapter 5 in Sansò and Sideris (2013).

The Treatment of the Topography

The undulation of the geoid N can be obtained from the solution of Stokes's BVP provided that there are no masses outside the geoidal surface. There are many ways to mathematically account for the topographic masses, called *terrain reductions*, which preserve the total mass of the Earth but shift the masses above the geoid below it. *Helmert's second condensation reduction* will be briefly presented here as a representative from a number of possible terrain reductions. The masses above the geoid are basically condensed onto a layer on the geoid by (1) computing at point P and removing the attraction A_P of all masses above the geoid; (2) restoring at

P the attraction A_P^c of the masses condensed on a layer on the geoid with density $\sigma = \rho H$; and (3) downward-continuing the resulting gravity anomalies from P to P_o (see Fig. 2). This procedure gives Δg on the geoid computed from the approximate expression

$$\Delta g \approx \Delta g_P - A_P + A_P^c - \frac{\partial \gamma}{\partial h} H_P = \Delta g_P + \delta A + F \quad (28)$$

where the free-air gravity gradient $-\partial \gamma / \partial h$ is often used to approximate the gradient of the Helmert gravity anomalies, $(\Delta g_P + F)$ is the free-air gravity anomaly at P, and δA is the resulting attraction change. For a completely rigorous treatment of Helmert's second condensation reduction, Chapter 4 in Sansò and Sideris (2017) should be consulted.

The A and A^c terms in Eq. 28 are obtained by applying Newton's integral in Eq. 1 over the volume v of the relevant masses. For example, when the attraction A of all topographic masses above the geoid is computed, v is the volume of the masses between the geoidal surface and the surface of the topography. From Fig. 2, it can be seen that $A = A_B - c$, where $A_B = 2\pi k \rho H$ (for $\rho = \text{const}$) is the attraction of the infinite Bouguer plate through P called *Bouguer correction*, and c is the attraction of the residual topographic masses above and below the Bouguer plate, called the classical *terrain correction*. A complete set of formulas for the attractions, as well as the potentials, of the topography and the condensed topography can be found in Heck (2003).

δA is not the only change associated with this reduction. Due to the shifting of masses, the potential changes as well by an amount δT called the *indirect effect on the potential*:

$$\delta T = T_{P_o} - T_{P_o}^c \quad (29)$$

where T_{P_o} is the potential of the topographic masses at P_o and $T_{P_o}^c$ is the potential of the condensed masses at P_o . The corresponding *indirect effect on the geoid* is $\delta N = \delta T / \gamma$. Due to this potential change, the use of Δg from Eq. 28 into Eq. 17 produces not the geoid but a surface called the *co-geoid*, depicted by the *co-geoidal height* N^c above the reference ellipsoid. Thus, before applying Stokes's equation, the gravity anomalies must be transformed from the geoid to the co-geoid by applying a small correction $\delta \Delta g$ called the *indirect effect on gravity*

$$\delta \Delta g = -\frac{1}{\gamma} \frac{\partial \gamma}{\partial h} \delta T = -\frac{\partial \gamma}{\partial h} \delta N \quad (30)$$

The final expression giving N can now be written as

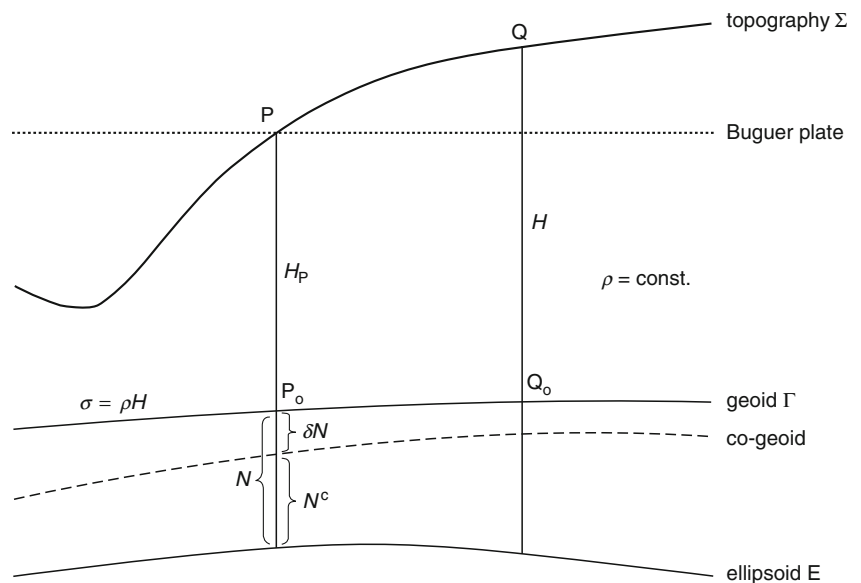
$$N = \frac{1}{\gamma} \mathbf{S}(\Delta g + \delta \Delta g) + \frac{1}{\gamma} \delta T = N^c + \delta N \quad (31)$$

where N^c is the co-geoidal height and δN is the indirect effect on the geoid; see Fig. 2.

Following Sideris (1990), δT and δA can formally be expressed using the planar approximation of the vertical derivative operator $\mathbf{L}$ of Eq. 21, that is, without the term $-f/R$, which then becomes zero since $R \rightarrow \infty$ when the spherical surface is approximated by the tangent plane. The potential change, for non-constant density ρ , is

$$\delta T = -\pi k \rho H_P^2 - 2\pi k \sum_{r=1}^{\infty} \frac{1}{(2r+1)!} \mathbf{L}^{2r-1} [\rho (H^{2r+1} - H_P^{2r+1})] \quad (32)$$

Geoid Determination, Theory and Principles, Fig. 2 Actual and condensed topography



and the attraction change is, to very good approximation (Martinec et al. 1993; Heck 2003), equal to the classical terrain correction c :

$$\delta A \approx c = 2\pi k \sum_{r=1}^{\infty} \frac{1}{(2r)!} \mathbf{L}^{2r-1} [\rho(H - H_P)^{2r}] \quad (33)$$

These series expansions are valid for low slopes of the terrain, namely, for $(H - H_P)/l_o \leq 1$, where l_o denotes the planar distance between running and computation point.

Sansò and Sideris (2017) have shown the equivalence between the solutions to the Molodensky BVP and the Stokes BVP with Helmert's second condensation reduction.

Another commonly used terrain reduction is the *residual terrain model* (RTM) reduction (Forsberg and Tscherning 1981; Sansò and Sideris 2013, Chapter 8). It computes at point P and removes only the high-frequency contributions of the topographic masses to gravity by, for example, computing the attraction of the masses between the actual topography and a very smooth “average topography” surface. The residual “geoid” produced by such reduced gravity anomalies is finally corrected by restoring the effect of the topographic masses that were removed. The RTM reduction yields a quasi-geoid rather than the geoid and, therefore, a final correction like the one in Eq. 23 must be applied in the end in order to recover the actual geoid undulation.

The Oceanic Geoid and Gravity from Satellite Altimetry

The only areas on Earth where, by definition, the geoid is almost directly observable are the oceans. From the difference between the ellipsoidal heights of altimetry satellites and the radar measurements to the surface of the sea, the ellipsoidal height h_s of the sea surface can be derived. h_s consists of the geoid undulation N and the sea surface topography (SST) H_s above the geoid. Thus, in principle, the geoid in the oceans can be derived from the expression

$$N = h_s - H_s \quad (34)$$

provided that a good model for the SST is available from oceanographic information. In addition, to minimize possible long-wavelength data biases, the *deflections of the vertical* are often derived numerically over the oceans as the slopes of the altimetry-derived geoid:

$$\xi = -\frac{\partial N}{r \partial \phi}, \quad \eta = -\frac{\partial N}{r \cos \phi \partial \lambda} \quad (35)$$

Although theory exists for the solution of overdetermined and altimetry-gravimetry BVPs, where given are gravity anomalies on the continents and geoid from altimetry and/or

gravity anomalies in the oceans, often in practice the oceanic geoid is also computed simply from Stokes's formula or from LSC using gravity data. Therefore, there is a need to obtain gravity anomalies in the oceans using as data altimetry-derived N and/or ξ, η .

Naturally, knowing N , one can use the following *inverse Stokes formula* to derive gravity anomalies in the oceans:

$$\Delta g_P = \frac{\gamma N_P}{R} - \frac{\gamma}{16\pi R} \iint_{\sigma} \frac{N - N_P}{\sin^3(\psi/2)} d\sigma \quad (36)$$

and then combine these gravity anomalies with the ones over land to produce a global set of gravity anomaly data which could be used in Stokes or LSC formulas to compute the global geoid. Note that the above equation has basically the form of the $\mathbf{L}$ operator acting on N , which, as expected, indicates that $\mathbf{L}$ and $\mathbf{S}$ are inverse operators. This can be proven easily, though rigorously, in planar approximation in the frequency domain; see details in chapter ► “Geoid Determination, Computational Methods”.

Gravity anomalies in the oceans can also be computed from the deflections of the vertical, using the *inverse Vening Meinesz formula*

$$\Delta g_P = \frac{\gamma}{4\pi} \iint_{\sigma} (\xi \cos \alpha + \eta \sin \alpha) H(\psi) d\sigma, \quad (37)$$

$$H(\psi) = \frac{\cos(\psi/2)}{2 \sin(\psi/2)} \left(-\frac{1}{\sin(\psi/2)} + \frac{3 + 2 \sin(\psi/2)}{1 + \sin(\psi/2)} \right) \quad (38)$$

where α is the azimuth of the line between the computation and the data point. A similar formula, termed the *deflection-geoid formula* (Hwang 1998), also exists that gives N from ξ, η :

$$N_P = \frac{R}{4\pi} \iint_{\sigma} (\xi \cos \alpha + \eta \sin \alpha) C'(\psi) d\sigma, \quad (39)$$

$$C'(\psi) = -\cot(\psi/2) + \frac{3}{2} \sin \psi \quad (40)$$

Another often used formula is the one that relates the vertical derivative of Δg to the horizontal derivatives of ξ and η . Starting from the spherical form of Laplace's equation

$$\begin{aligned} \nabla^2 T &= \frac{\partial^2 T}{\partial r^2} + \frac{2}{r} \frac{\partial T}{\partial r} + \frac{1}{r^2} \frac{\partial^2 T}{\partial \phi^2} - \frac{\tan \phi}{r^2} \frac{\partial T}{\partial \phi} + \frac{1}{r^2 \cos^2 \phi} \frac{\partial^2 T}{\partial \lambda^2} \\ &= \frac{\partial \Delta g}{\partial r} + \gamma \frac{\partial \xi}{r \partial \phi} - \frac{\tan \phi}{r} \gamma \xi + \gamma \frac{1}{r \cos \phi} \frac{\partial \eta}{\partial \lambda} = 0 \end{aligned} \quad (41)$$

and neglecting the small third term on the right-hand side, we obtain

$$\frac{\partial \Delta g}{\partial r} = -\gamma \left(\frac{\partial \xi}{r \partial \phi} + \frac{1}{r \cos \phi} \frac{\partial \eta}{\partial \lambda} \right) \quad (42)$$

It is clear that since Δg , ξ , η , N , and their various derivatives are all functionals of T , they could all be used as input in LSC to predict both Δg and N in the oceans. Therefore, collocation could be used either to predict directly undulations and anomalies in the oceans from any gravity field quantity, or to densify or interpolate data necessary for the efficient evaluation of integrals like inverse Stokes or inverse Vening Meinesz. Details can be found in chapter ► [“Geoid Determination, Computational Methods”](#).

Summary

We have discussed briefly how the geoid is defined and how it can be estimated either analytically as the solution of the third BVP (or Molodensky’s BVP) from gravity anomalies all over the Earth or as a “best” unbiased linear combination of all available gravimetric data using least-squares collocation. In either case, data have to be reduced to the geoid via an appropriate terrain reduction. The methodology used in the numerical determination of geoid undulations is described in chapter ► [“Geoid Determination, Computational Methods”](#). Readers interested in obtaining a more comprehensive view of gravity field approximation and geoid determination are referred to Sansò and Sideris (2013, 2017) and are encouraged to also read the other chapters on geoid and gravity included in this volume. Finally, with dedicated gravity satellites like GRACE and its follow-on mission GRACE-FO, the computation of temporal variations of the geoid and its application to other geosciences is now possible; see chapter ► [“Gravity Field, Temporal Variations from Space Techniques”](#). This, however, is beyond the scope of this short entry and will not be discussed here.

Cross-References

- [Geodesy, Figure of the Earth](#)
- [Geodesy, Physical](#)
- [Geoid Determination, Computational Methods](#)
- [Geoid Undulation, Interpretation](#)
- [Gravimeters](#)
- [Gravity Anomalies, Interpretation](#)
- [Gravity Data, Advanced Processing](#)
- [Gravity Field of the Earth](#)
- [Gravity Field, Temporal Variations from Space Techniques](#)
- [Gravity Field, Time Variations from Surface Measurements](#)

- [Gravity Method, Satellite](#)
- [Gravity, Data to Anomalies](#)
- [Gravity, Global Models](#)
- [Height Systems, Vertical Datums and Their Unification](#)
- [Spherical Harmonic Analysis Applied to Potential Fields](#)

Bibliography

- Forsberg R, Tscherning CC (1981) The use of height data in gravity field approximation by collocation. *J Geophys Res* 86(B9):7843–7854
- Heck B (2003) On Helmert’s methods of condensation. *J Geod* 77:155–170
- Hofmann-Wellenhof B, Moritz H (2006) *Physical Geodesy*, 2nd edn. Springer Wien, New York
- Hwang C (1998) Inverse Vening Meinesz formula and deflection-geoid formula: applications to the predictions of gravity and geoid over the South China Sea. *J Geod* 72:304–312
- Martinec Z, Matyska C, Grafarend EW, Vaniček P (1993) On Helmert’s 2nd condensation method. *Manuscr Geodaet* 18:417–421
- Molodensky MS, Eremeev VF, Yurkina MI (1962) *Methods for the study of the external gravity field and figure of the earth*. Israel Program of Scientific Translations, Jerusalem. (Russian original, 1960)
- Moritz H (1980) *Advanced physical geodesy*. Herbert Wichmann Verlag, Karlsruhe. (Reprint 2008 by School of Earth Sciences, Ohio State University, Columbus, Ohio)
- Sansò F, Sideris MG (eds) (2013) *Geoid Determination – Theory and Methods*, Lecture Notes in Earth System Sciences, vol 110. Springer, Berlin Heidelberg
- Sansò F, Sideris MG (2017) *Geodetic boundary value problem: the equivalence between Molodensky’s and Helmert’s solutions*. Springer Briefs in Earth Sciences, Springer International Publishing AG, Cham, Switzerland
- Sideris MG (1990) Rigorous gravimetric terrain modelling using Molodensky’s operator. *Manuscr Geodaet* 15(2):97–106

Geoid Undulation, Interpretation

Petr Vaniček

Department of Geodesy and Geomatics Engineering,
University of New Brunswick, Fredericton, NB, Canada

Synonyms

Depth of Moho surface is the same as the thickness of the lithosphere; Direct gravimetric problem, aka forward modeling; Geoidal height, aka geoid undulation; Moho surface, aka compensation level for the Airy-Heiskanen isostatic hypothesis; Ocean bottom depth, aka bathymetry; Schiaparelli’s body, aka body of nil external potential; Sea surface topography (SST), aka dynamic topography; Spectral transfer function, aka admittance; Stokes’s integral truncated to the spherical distance of ψ_0 , aka inner zone contribution to Stokes’s integral

Definition

Direct gravimetric problem is defined as follows: “Given the distribution of density within the Earth, derive the (external) gravity field the density generates.”

The *inverse gravimetric problem* (GIP) reads: “Given the external gravity field of the Earth, derive the distribution of density implied by the field.”

Introduction

The main goal of geophysics is to learn about the physical processes as well as properties of the material within the Earth. These are studied by means of measurements (data) collected on the surface of the Earth and above it and, to a very limited degree, also in very shallow probes inside the Earth. The discussion of the data to be collected and the techniques to be used to study the physical processes and material properties within the earth are what this book is all about.

The focus of this section is on gravity field, the geoid in particular, to the exclusion of all other, such as seismological, magnetic, electrical, heat flow data. The only attribute of the Earth that can be studied by means of gravity is the mass density variations as gravity is the acceleration caused by mass attraction. Three aspects of gravity field can be used for these studies: gravity acceleration, the deflection of the vertical, and gravity potential. Gravity acceleration, in the guise of gravity anomalies, are discussed in *Gravity Anomalies, Interpretation*, the deflections of the vertical are introduced in *Geodesy, Physical* and for an in-depth view of the definition of geoid the reader is advised to consult *Geoid*.

One of the general properties of the gravity field is that deep-seeded density anomalies have long wavelength signature in the gravity field while shallow density anomalies have short wavelength signature. Consequently, if one wishes to study shallow density features, one looks at the short wavelength part of gravity field and if one wishes to study deep-seeded features, one looks at the long wavelength part. This is, of course, true only as a general principle as there exist also shallow features of large lateral extent that have long wavelength signatures.

When it comes to the interpretation of gravity field, the data of choice by exploration geophysicists, that is, people looking for ore, hydrocarbon, and other deposits, are gravity anomalies. The reason for this is that these data are bountiful on land, relatively cheap, and they contain all the short wavelength information. Observed deflections of the vertical as supplied by national geodetic offices are less popular as a source of information as they are not very accurate, and they are few and far apart. But we may live to see this situation change with the advent of Digital Zenith Cameras (Hirt et al. 2010).

Gravity potential data, such as the values of disturbing potential – divided by the value of gravity to convert them to the geoid – are much smoother than the first two kinds. This means that the short wavelength content is missing. We know that all three descriptors of gravity field we discuss here are functionally related: gravity acceleration at a point is a total gradient of the potential at that point, the deflection of the vertical at a point is a horizontal gradient of the equipotential surface passing through that point. Thus, there is nothing to stop us from using the potential data, the geoid, in such a way as to evaluate either the gravity anomalies or the deflections of the vertical and use these for the interpretation, is there?

That can be done of course, but this procedure cannot reconstitute the high frequency content as it has not been there to begin with. Thus, the conversion would not help and it is better to use the geoid as is; this is the thesis of the book (Vaníček and Christou 1993). The direct use of the geoid has caught an interest of interpreters as the marine geoid data from satellite altimetry became quite reliable and freely available in the past 25 years.

Direct Problem

Suppose we know the density distribution $\rho(\mathbf{r})$ within the Earth. Then we can determine the potential $V(\mathbf{r})$ outside the earth in two different ways. We can use the well-known Newton's integral to get the gravitational potential in a spatial form at any point external to the Earth as (Vaníček and Krakiwsky 1986, Equation 20.48)

$$V(r, \Omega) = G \int_{\Omega_0}^{r_t(\Omega)} \frac{\rho(r', \Omega')}{|\mathbf{r} - \mathbf{r}'|} (r')^2 dr' d\Omega', \quad (1)$$

where G is Newton's gravitational constant, r is the radial distance from the center of the Earth, Ω is the geocentric direction composed of colatitude θ and longitude λ , Ω_0 is the full geocentric solid angle, r_t is the distance from the center of the Earth to the topographical surface of the Earth.

Alternatively, the reciprocal distance $\frac{1}{|\mathbf{r} - \mathbf{r}'|}$ can be developed into a series of fully normalized spherical harmonic complex functions Y_{jm} to yield

$$V(r, \Omega) = \frac{GM}{a} \sum_{j=0}^{\infty} \left(\frac{a}{r}\right)^{j+1} \sum_{m=-j}^j A_{jm} Y_{jm}(\Omega), \quad (2)$$

where M is the mass of the Earth, a is the major semiaxis of the best fitting Earth ellipsoid, and A_{jm} are the gravity potential coefficients. The potential coefficients can be evaluated from the known density $\rho(\mathbf{r}) = \rho(r, \Omega) = \rho(r, \theta, \lambda)$ within the Earth as

$$A_{jm} = \int_{\Omega_0} \int_{r=0}^{r_i(\Omega)} \rho(r, \Omega) G_{jm}(r, \Omega) dr d\Omega, \quad (3) \quad \text{where}$$

where the integration kernel G_{jm} is given by Martinec in Vaníček and Christou (1993, Chap. 7, Equations 13 and 14)

$$G_{jm}(r, \Omega) = \frac{4\pi}{Ma^j(2j+1)} r^{j+2} Y_{jm}^*(\Omega). \quad (4)$$

Here $*$ denotes a complex conjugate. This is the spherical harmonic form of potential V , a dual form to the spatial form (Eq. 1 above) – these two forms can be transformed into each other, back and forth any time. The relation between potential V and the geoid undulation N is explained in *Geodesy, Physical*.

The direct approach, also known as forward modeling, can be used for solving some geophysical problems. Notably, it can be used to determine the depth of Mohorovicic's surface, the surface that divides the asthenosphere (density about 3.2 g cm^{-3}) from the lithosphere (density about 2.67 g cm^{-3}). The Moho surface is also the compensation level for the Airy-Heiskanen isostatic hypothesis. The depth of the Moho surface, or the lithospheric (crustal) thickness derived this way, assuming the most realistic asthenospheric density and described by a spherical harmonic series of a limited degree and order is shown on a web page <http://geology.com/articles/mohorovicic-dis-continuity.shtml>. The displayed map was compiled by US Geological Survey in 1999.

Inverse Problem

The determination of density $\rho(\mathbf{r})$ as a function of position $\mathbf{r}$ from any of the gravity field attributes, that is, gravity anomaly $\Delta g(\mathbf{r})$, the components of the deflection of the vertical ($\xi(\mathbf{r}), \eta(\mathbf{r})$), or the geoidal height $N(\mathbf{r})$, is known as the inverse gravity problem (GIP). The non-uniqueness of GIP was known already to Schiaparelli, who in 1875–1876 was speaking of “gravitationally transparent bodies,” specific distributions of density that do not generate any gravitational attraction outside the Earth. A complete mathematical description of a Schiaparelli's body was formulated by Vaníček and Kleusberg (1985) as follows: for arbitrary values of coefficients $c_{ljk}^{(i)}$ the following “body”:

$$\begin{aligned} \theta \in < -\frac{\pi}{2}, \frac{\pi}{2} >, \lambda \in < 0, 2\pi >, r \in < 0, 1 >: \\ \tilde{\rho}(\mathbf{r}) = \sum_{i=0}^{\infty} \sum_{l=0}^i \sum_{j=0}^{\infty} \sum_{k=0}^j c_{ljk}^{(i)} \phi_{ijk}(\mathbf{r}), \end{aligned} \quad (5)$$

$l \neq j$

$$\begin{aligned} \forall i, j = 0, 1, \dots, k \leq j : \phi_{ijk}(\mathbf{r}) \\ = J_i(3, 3, r) \left[Y_{jk}^c(\Omega), Y_{jk}^s(\Omega) \right], \end{aligned} \quad (6)$$

J_i are normalized Jacobi polynomials (Abramowitz and Stegun 1964) and Y_{jk}^c, Y_{jk}^s are the cosine and sine parts of spherical harmonic functions, has a nil external potential outside the ball of radius 1. The consequence of this behavior is that if one tries to interpret the external gravity potential in terms of internal density distribution, any distribution that conforms to Eq. 5 can be randomly and legitimately added to the solution.

If the GIP does not have a unique solution, does it make any sense to even think about using gravity in studying the density distribution? True, it is not possible to take the gravity field outside the Earth and invert it uniquely into the density distribution within the Earth. But the external gravity field, particularly the geoid in our context here, can be used together with other data to get useful information on density.

Truncated Geoid

There is a variety of techniques used in geophysical exploration to interpret gravity information obtained on the surface of the Earth in terms of anomalous density. One interesting approach uses the truncated geoid to determine the depth of a spherical, i.e., ball-like, density anomaly, where by “truncated geoid” we understand the following integral:

$$N^{\psi_0}(P) = \frac{R}{4\pi\gamma} \int_{C_{\psi_0}} \Delta g(Q) S(P, Q) d\sigma, \quad (7)$$

which represents the contribution of gravity to the geoid within the spherical cap C of a radius ψ_0 . It can be viewed as the Stokes integral over the whole Earth, truncated to the spherical distance ψ_0 , where P is the point of interest, Q is the dummy point in the integration, $S(P, Q) = S(\psi)$ is the Stokes spherical kernel and $d\sigma$ is the surface element of the integration sphere of radius R .

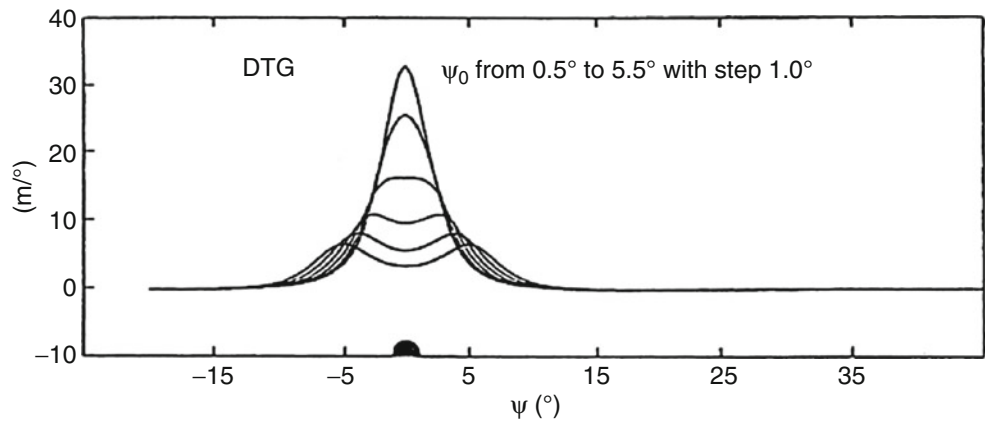
It is the change of the truncated geoid with respect to the truncation parameter ψ_0 ,

$$\frac{dN^{\psi_0}}{d\psi_0} = \lim_{\Delta\psi_0 \rightarrow 0} \frac{N^{\psi_0+\Delta\psi_0} - N^{\psi_0}}{\Delta\psi_0}, \quad (8)$$

that is of real interest. At the location of the density anomaly, the sequence of surfaces or profiles $\frac{dN^{\psi_0}}{d\psi_0}$ displays a dimple for a specific value of ψ_0 – see Fig. 1 (from Vajda and Vaníček

Geoid Undulation, Interpretation,

Fig. 1 Sequence of profiles of the change of the truncated geoid with respect to the truncation parameter ψ_0 that shows the onset of the dimple for a value of $\psi_0 = 2.5^\circ$



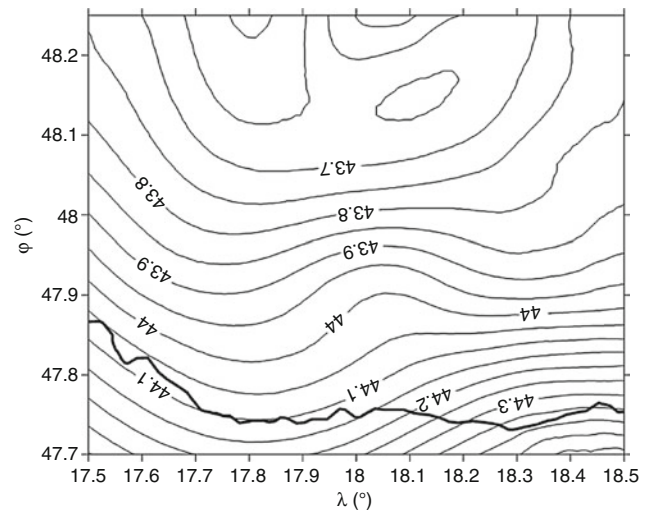
1999). This specific value of ψ_0 is almost linearly related to the depth of the anomaly.

In Fig. 2, geoid from an area in Southern Slovakia is shown. Just northeast of the geographical center of the area, there is a local geoid low, known as “Kolarovo anomaly,” that has been intriguing geophysicists for some time. It has been postulated by several authors that the center of mass of the density anomaly responsible for the shown pattern is about 9.5 km bellow the Earth surface. The application of the truncated geoid technique shows an estimate of 8.7 km (Vajda et al. 2002).

Some Other Applications

In addition to the above-discussed applications, geoid can be used in many other ways; one of those applications is studying tectonic forces. Interested reader may read about the involved concepts in *Gravity, Global Models*. Bowin, in his earlier studies (1991), postulated that the $4\text{--}10^\circ$ geoid undulations are consistent with the present zones of subducted tectonic slabs. His postulate has more recently been confirmed by Ricard et al. (2006). These authors also agree with Bowin’s speculation (1983) that only the deepest Earth’s mass anomalies cause geoid undulations of $2\text{--}10^\circ$.

Casenave in Vaniček and Christou (1993, Chap. 13) discusses the use of the geoid in studying different isostatic compensation models using the spectral transfer function, aka admittance in the “theory of systems” parlance. She studies the mechanical behavior of oceanic tectonic plates (flexure under seamount loading) and thermal evolution of oceanic lithosphere. Ricard et al. (2006) show that the elasticity of the lithosphere affects the geoid between 200° and 250° . Casenave also argues that the signatures of oceanic hot spots in the geoid are in the medium wavelength of $n \in (5, 20)$ and in 1–10 m in range.



Geoid Undulation, Interpretation, Fig. 2 Geoid in the area of “Kolarovo anomaly” in Southern Slovakia. (Courtesy of J. Janak, Slovak University of Technology)

It may be worth mentioning that meteorite craters are often clearly discernible in detailed geoid maps. The Manicouagan crater in Eastern Canada (location: $N 51^\circ 23'$ and $W 68^\circ 42'$) makes a very clear imprint on the geoidal map.

The last applications of the geoid we want to show here concern physical oceanography and the oceans in general. A very good overview of this topic by Nerem and Koblinsky is in Vaniček and Christou (1993, Chap. 16). Geoid, being an equipotential (level) surface, should be represented by the mean sea level if seawater were homogeneous, which, of course, it is not. As a result, the mean sea level is not in equilibrium and currents develop. Some time after the initiation of this motion, these currents reach a steady state, called geostrophic, and swirl around the globe in a perpetual motion attenuated only by friction.

The equation that describes the geostrophic motion is ascribed to Navier and Stokes, and it reads as follows:

$$\mathbf{V} + (\mathbf{V} \cdot \nabla)\mathbf{V} + 2\boldsymbol{\Omega} \times \mathbf{V} = \frac{-\nabla p}{\rho} + \mathbf{g} + \mathbf{F}, \quad (9)$$

where $\mathbf{V}$ is the flow velocity vector, $\boldsymbol{\Omega}$ is the rotation vector of the Earth, p is the hydrostatic pressure, ρ is the seawater density, $\mathbf{g}$ is the gravitation vector, and $\mathbf{F}$ is the vector representing frictional forces (ibid. Eq. 1). This equation, together with measurements taken at sea, is then used to derive the height of the sea surface above the geoid, which is called the “dynamic topography.”

As discussed above, the instantaneous sea surface has been monitored by a technique called satellite altimetry for years. This surface represents, of course, the sum of the geoidal height and the dynamic topography. Subtracting the geoid from the measured sea surface yields the dynamic topography that can be easily converted into information about currents. Conversely, subtracting the dynamic topography from the measured surface yields the geoid. Interested reader can refer to following URL <http://www.aviso.oceanobs.com/en/data/products/auxiliary-products/index.html> for more detailed information.

Last but not least, the geoid obtained from satellite altimetry can be used to calculate the ocean bottom depth, aka bathymetry (Sandwell et al. 2006). The principle used in these calculations is the same as the one used in the determination of the Moho surface. Acknowledging that the ocean bottom is a surface that divides the crust (of an average density of 2.67 g cm^{-3}) from the layer of water (of an average density of 1.03 g cm^{-3}), forward modeling it then used to yield the ocean bottom surface.

Summary

The geoidal height (geoid undulation) is a quantity that can be quite readily interpreted in terms of density distribution within the Earth. The main application is in studying the processes that take place deep within the Earth, but even shallower density anomalies have an impact on a geoid map. Nowadays, when sea surface height measured by satellite altimetry is freely available, marine geoid has become a particularly handy tool to use.

Cross-References

- [Earth, Density Distribution](#)
- [Geodesy, Physical](#)
- [Geoid](#)
- [Geoid Determination, Theory and Principles](#)
- [Gravity Anomalies, Interpretation](#)
- [Gravity Field of the Earth](#)

Bibliography

- Abramowitz M, Stegun IA (1964) Handbook of mathematical functions. Dover, St. Mineola. Reprint
- Bowin C (1983) Depth of principal mass anomalies contributing to the Earth's geoidal undulations and gravity anomalies. *Mar Geod* 7:61
- Bowin C (1991) Speculations on a mechanism for maintenance of absolute plate motion, AGU abstract. *Eos Trans AGU* 72(44):244
- Hirt C, Burki B, Somieski A, Seeber G (2010) Modern determination of vertical deflections using digital zenith cameras. *J Surv Eng* 136(1):1–12
- Ricard Y, Chambat F, Lithgow-Bertelloni C (2006) Gravity observations and 3D structure of the earth. *C R Geosci* 338:992–1001
- Sandwell DT, Smith WHF, Gille S, Kappel E, Jayne S, Soofi K, Coakley B, Geli L (2006) Bathymetry from space: rationale and requirements for a new, high-resolution altimetric mission. *C R Geosci* 338:1049–1062
- Schiaparelli GV (1875–1876) *Lezioni di Meccanica Celeste* (Appunti). University of Pavia, Italy
- Vajda P, Vaníček P (1999) Truncated geoid and gravity inversion for one point-mass anomaly. *J Geod* 73:58–66
- Vajda P, Bielik M, Pohanka V (2002) Testing the application of the truncation filtering methodology in interpreting real gravity data: the Kolárovo gravity anomaly. *Contrib Geophys Geod* 32(1):57–66
- Vaníček P, Christou NT (1993) *Geoid and its geophysical interpretations*. CRC Press, Boca Raton
- Vaníček P, Kleusberg A (1985) What an external gravitational potential can really tell us about mass distribution. *Boll Geofis Teor Appl* XXCII(108):243–250
- Vaníček P, Krakiwsky EJ (1986) *Geodesy: the concepts*, 2nd edn. North Holland, Amsterdam

Geomagnetic Excursions

Martha Schwartz

Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA

Synonyms

Excursions were initially observed as “tiny wiggles” in marine magnetic anomalies (LaBreque et al. 1977) and were termed “cryptochrons” by Cande and Kent (1992a, b).

Definition

Geomagnetic Excursions are periods of unusually high-amplitude secular variation (see ► [“Geomagnetic Field, Secular Variation”](#)). They are usually defined as short-term deviations of the Virtual Geomagnetic Dipole (VGP) more than $40\text{--}45^\circ$ from the geographic poles (see ► [“Geomagnetic Field, Global Pattern”](#)).

Geographical and Temporal Extent

Excursions last for up to a few thousand years during times of otherwise stable field polarity. The unusual directions, which sometimes represent regionally complete reversals (see ► “[Geomagnetic Field, Polarity Reversals](#)”), are always accompanied by globally correlative low levels of field intensity (see ► “[Paleomagnetic Field Intensity](#)”). Low intensity starts earlier than the anomalous directions and outlasts them. There is some emerging evidence that excursions are global in extent. However, uncertainties in age control limit the ability to establish synchronicity among directional records recovered from different regions. Inadequate paleomagnetic recording may also cause excursions to be missed at certain localities. Lava flows do not provide a continuous record of field behavior, and sediment records of short-term directional field behavior may be smeared due to low sedimentation rates (see ► “[Geomagnetic Field, Measurement Techniques](#)”).

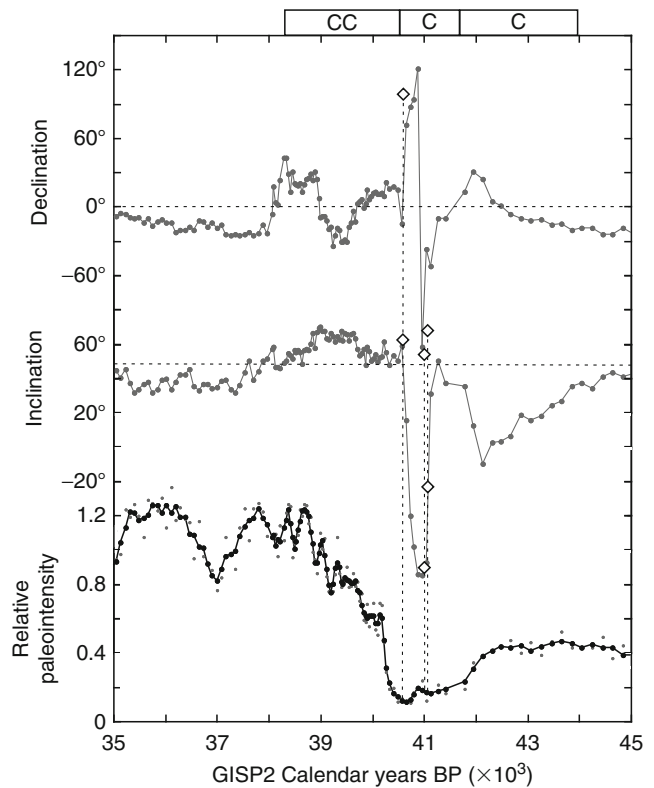
The frequency and duration of geomagnetic excursions are also open issues. Answers depend on recovering further high-quality replicated records, as well as better understanding of their regional versus global extent. There is reproducible evidence for 17 time intervals within the Brunhes Chron with at least regionally verifiable magnetic field excursions (Lund et al. 2006). Two well-recorded examples of excursions are the Laschamp Excursion, first noted in French lavas (Bonhommet and Babkine 1967), which occurred about 40 ka, and the Blake Event (first observed by Smith and Foster 1969) from about 120 ka.

The Fig. 1, reproduced from Lund et al. (2005), is a high-resolution record of the Laschamp Excursion, as recovered from the Blake Outer Ridge in the North Atlantic Ocean.

Field Configuration During Excursions

Complete understanding of the mechanism causing excursions depends on improved mapping of the global field geometry during excursive intervals. Roberts (2008) summarized three suggested excursionary field configurations. (1) The field remains largely dipolar but wanders away from the rotational axis. In this case VGP's would be consistent regardless of sampling site. (2) The ratio of non-dipole and dipole field components becomes very high. In this case VGPs would depend strongly on the sampling site. (3) Relatively localized sources in the outer-core produce regional excursions (see ► “[Geomagnetic Field, Theory](#)”).

Several numerical dynamo models are able to simulate magnetic field excursions for certain choices of boundary conditions. But model numerical simulations are limited in spatial or temporal resolution due to presently limited computer technology (Glatzmaier 1999).



Geomagnetic Excursions, Fig. 1 Paleomagnetic secular variation and relative paleointensity (small dots are original data, large dots are 3-pt running average) records from Blake Outer Ridge core JPC14 for the time interval of the Laschamp Excursion. NRM after 20 mT demagnetization was used outside the excursion interval and NRM characteristic remanences were used in the excursion interval. These records show the same distinctive interval of low paleointensity seen in other cores (Figs. 2 and 3). But it contains a much more dramatic pattern of directional variability that we consider to be a more accurate and high-resolution record of the Laschamp Excursion as seen on the Blake Outer Ridge. The dashed lines indicate three time intervals where directions in this record are almost identical to the Laschamp Excursion directions seen in Europe (open diamonds). These three intervals also have similar paleointensity lows. Boxes at top indicate intervals of clockwise (C) or counterclockwise (CC) looping (circularity) in the directional data

Origin

The most widely accepted explanation for the origin of geomagnetic excursions comes from Gubbins (1999). According to this theory, the outer core reverses briefly but the solid inner core, which has a magnetic diffusion time on the order of 3000 years, does not reverse. In the case in which the inner core did have time to turn around, the event would be recorded as a full field reversal (see ► “[Geomagnetic Field, Polarity Reversals](#)”).

Summary

Geomagnetic excursions are brief periods of anomalous directions coupled with low field intensity. Their spatial

extent, duration, and causative mechanism are under study. A review paper by Laj and Channell (2007) provides an excellent summary of the research to date.

Cross-References

- [Geomagnetic Field, Global Pattern](#)
- [Geomagnetic Field, Measurement Techniques](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Geomagnetic Field, Theory](#)
- [Paleomagnetic Field Intensity](#)

Bibliography

- Bonhommet N, Babkine J (1967) Sur la presence d'aimantation inverse dans la Chaîne des Puys. *CR Acad Sci Ser B* 264:92–94
- Cande SC, Kent DV (1992a) A new geomagnetic polarity timescale for the late cretaceous and cenozoic. *J Geophys Res* 97:13917–13951
- Cande SC, Kent DV (1992b) Ultrahigh resolution marine magnetic anomaly profiles: a record of continuous paleointensity variations? *J Geophys Res* 97:15075–15083
- Gubbins D (1999) The distinction between geomagnetic excursions and reversals. *Geophys J Int* 137:F1–F4
- Glatzmaier GA (1999) The role of the Earth's mantle in controlling the frequency of geomagnetic reversals. *Nature* 401:885–990
- LaBrecque JL, Kent DV, Cande SC (1977) Revised magnetic polarity time-scale for the late cretaceous and cenozoic time. *Geology* 5:330–335
- Laj C, Channell JET (2007) Geomagnetic excursions. In: Kono M (ed) *Treatise on geophysics, geomagnetism*, vol 5. Elsevier, Amsterdam, pp 373–416
- Lund SP, Schwartz M, Keigwin L, Johnson T (2005) Deep-sea sediment records of the Laschamp geomagnetic field excursion (~41,000 calendar years before present). *J Geophys Res* 110:B04101. <https://doi.org/10.1029/2003JB002943>.
- Lund S, Stoner JS, Channell JET, Acton G (2006) Brunhes paleomagnetic field variability recorded in ocean drilling program cores. *Phys Earth Planet Inter* 156:194–205
- Roberts A (2008) Geomagnetic excursions: knowns and unknowns. *Geophys Res Lett* 35:L17307
- Smith JD, Foster JH (1969) Geomagnetic reversal in the Brunhes normal polarity epoch. *Science* 163:565–567

Geomagnetic Field, Global Pattern

Susan Macmillan
Geomagnetism Team, British Geological Survey,
Edinburgh, UK

Definition

The geomagnetic field is the Earth's magnetic field and in this chapter we describe how its pattern is mapped at the surface of the Earth through time.

Introduction

The geomagnetic field is generated in the fluid outer core region of the Earth by electrical currents flowing in the slowly moving molten iron. In addition to sources in the Earth's core, the geomagnetic field observable at the Earth's surface has sources in the crust and in the ionosphere and magnetosphere. The signal from the core dominates, accounting for over 95% of the field at the Earth's surface. The geomagnetic field varies on a range of scales, both temporal and spatial; the description of the variations made here concentrates on the recent spatial and temporal variations of the field with origins in the Earth's core that can be surmised from observations made over the last four centuries.

Observations

The geomagnetic field is a vector quantity, having magnitude as well as direction. The direction of the geomagnetic field has been directly observed and recorded for at least 400 years beginning with observations of the angle in the horizontal plane between true north and the magnetic field vector, known as declination or magnetic variation. In 1634 it was realized by Henry Gellibrand that declination in London was changing with time when he compared his observations with those made by Edmund Gunter 12 years earlier. Since then this important discovery of secular variation has ensured regular observations of the magnetic field through time. Measurements of declination were important for navigation across the oceans, and a source of early observations is ships' log-books (Jackson et al. 2000). Before long it was also realized that other elements of the geomagnetic field were of interest, in particular the angle of dip of the magnetic field vector from the horizontal known as inclination, and this was also measured. However, it was not till 1832 when Carl Friedrich Gauss established his technique for measuring absolute intensity did we have accurate measurements of the magnitude of the geomagnetic field.

As demands on accuracy and interest in the geomagnetic field increased, permanent observatories began to be established. Since the 1840s the number of observatories around the world has slowly increased to a peak of about 180 in the late 1980s and early 1990s but is about 160 in recent times. The advent of the proton precession magnetometer and the fluxgate magnetometer in the twentieth century considerably eased the automation of observatories during this time. However, regular manual absolute observations, nowadays of declination and inclination using a fluxgate theodolite, are necessary to maintain accurate baseline control over long periods of time. There are also networks of repeat stations providing data over extensive areas from which country-wide magnetic charts could be derived. More recently, magnetic measurements made by satellites have

become important in determining the pattern of the geomagnetic field. In particular, vector data from Magsat (1979–1980), Ørsted (1999–2013), CHAMP (2000–2010), and Swarm (2013–) have all been utilized in the production of recent spherical harmonic models of the geomagnetic field.

Spherical Harmonic Analysis

Till the 1950s magnetic charts at a given epoch for use in navigation depended on manually drawing contours through the observations (Malin 1971). Although spherical harmonic analysis had been developed in 1839 by Carl Freidrich Gauss, it was not routinely used to fit mathematical models to the observations till the advent of computers. Before computers spherical harmonic analysis required that the data be in the form of values of one or more of the orthogonal components X , Y , and Z of the geomagnetic field (magnetic intensities in the northerly, easterly, and vertically down directions) at points regularly spaced in latitude and longitude (Barraclough 1978). This involved the intermediate stage of interpolating values from manually drawn contour charts.

The mathematical and physical basis for spherical harmonic analysis is now given. In a source-free region, the Earth's magnetic field $\mathbf{B}$ is the negative gradient of a magnetic potential V that satisfies Laplace's equation:

$$\mathbf{B} = -\nabla V \quad \text{where} \quad \nabla^2 V = 0$$

A solution to Laplace's equation for the magnetic potential arising from sources inside the Earth at a given epoch is given by:

$$V(r, \theta, \lambda) = a \sum_{n=1}^{n_{\max}} \left(\frac{a}{r}\right)^{n+1} \sum_{m=0}^n \times (g_n^m \cos m\lambda + h_n^m \sin m\lambda) P_n^m(\theta)$$

In this equation r , θ , λ are geocentric coordinates (r is the distance from the center of the Earth, θ is the colatitude, i.e., $90^\circ - \text{latitude}$, and λ is the longitude), a is a reference radius for the Earth (6371.2 km), g_n^m and h_n^m are the spherical harmonic Gauss coefficients of degree n and order m , $P_n^m(\theta)$ are the Schmidt semi-normalized Legendre functions, and $n_{\max}$ is the maximum degree of the spherical harmonic expansion. If the observations of the magnetic field are spread over time, the magnetic potential and corresponding Gauss coefficients are also dependent on time with splines being commonly used.

Maps of the Geomagnetic Field

It is worth noting here that the first published map of the geomagnetic field is that of declination made in 1701 by

Edmond Halley. It was based on declination observations made during a magnetic survey expedition under naval command and covered the Atlantic Ocean (Clark and Barraclough 2001).

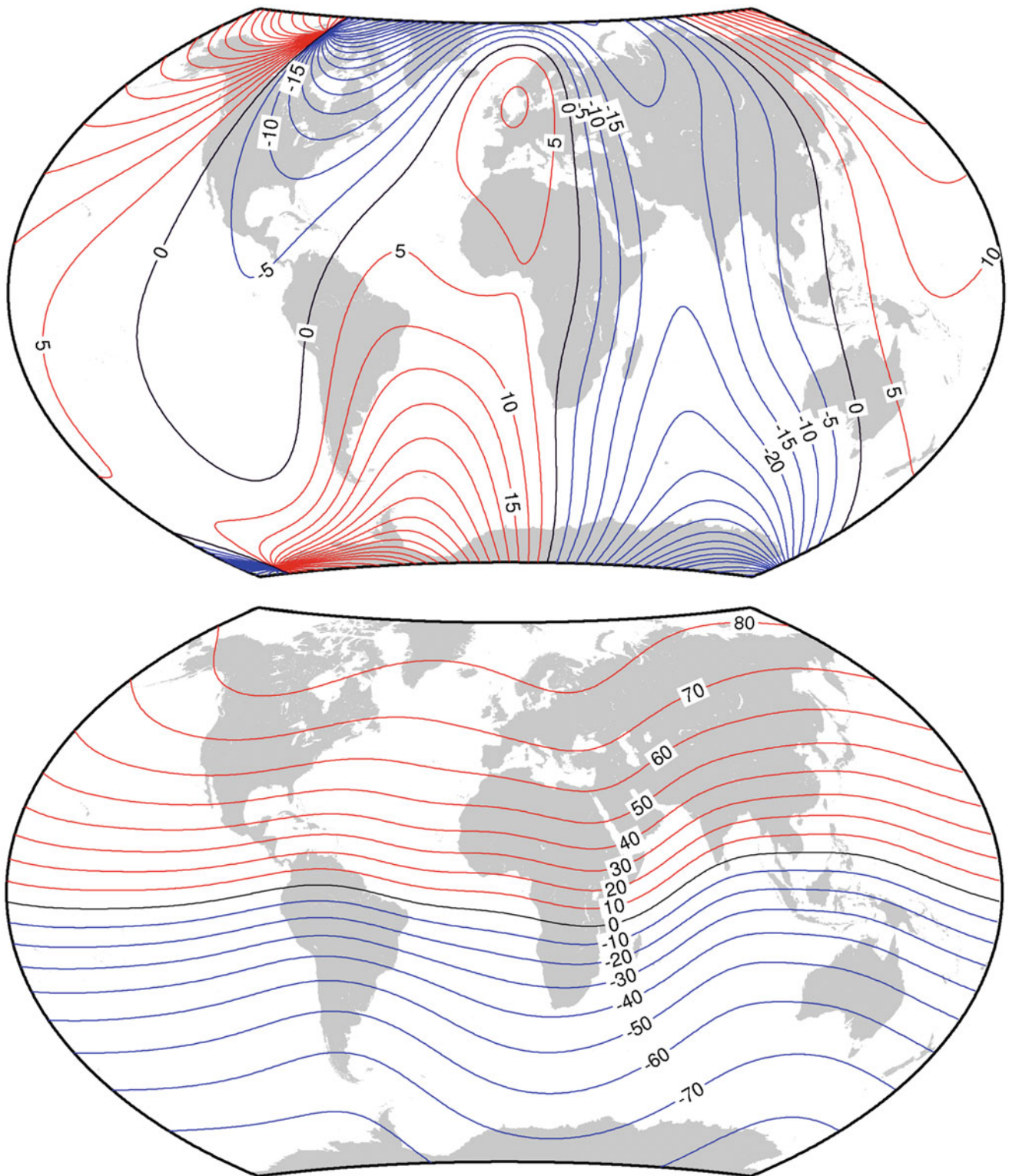
The maps that we show here are based on spherical harmonic models. Nowadays there are many such models of the geomagnetic field and here we look at two examples. One is the *gufm1* model which is derived from four centuries of magnetic field observations (Jackson et al. 2000) and extends to spherical harmonic degree 14 with B-splines being used to fit the variations in time from 1590 to 1990. It has smoothness imposed in both the spatial and temporal domains. The other model is the International Geomagnetic Reference Field (Thébault et al. 2015) produced under the auspices of the International Association of Geomagnetism and Aeronomy (IAGA). It extends to spherical harmonic degree 10 till 1995 and thereafter extends to degree 13. The variations in time from 1900 to 2020 in the IGRF are assumed to be piecewise linear and there is effectively no smoothness imposed in either the spatial or temporal domains. The final IGRF coefficients are made up of weighted averages of candidate sets of coefficients from (mostly) unregularized models.

Using the *gufm1* model, plots of declination and inclination at 1600, declination, inclination, and total intensity at 1950 are shown in Figs. 1 and 2. Using the 12th generation IGRF, plots of the magnetic elements at 2015 and their rates of change are shown in Figs. 3, 4, 5, 6, 7, 8, and 9. The rates of change plots are derived from a predictive secular variation model which extends to spherical harmonic degree 8. Up-to-date maps, and on-line calculators are available from a number of websites, for example, <https://geomag.bgs.ac.uk/navigation.html> and <https://ngdc.noaa.gov/geomag/WMM/DoDWMM.shtml>.

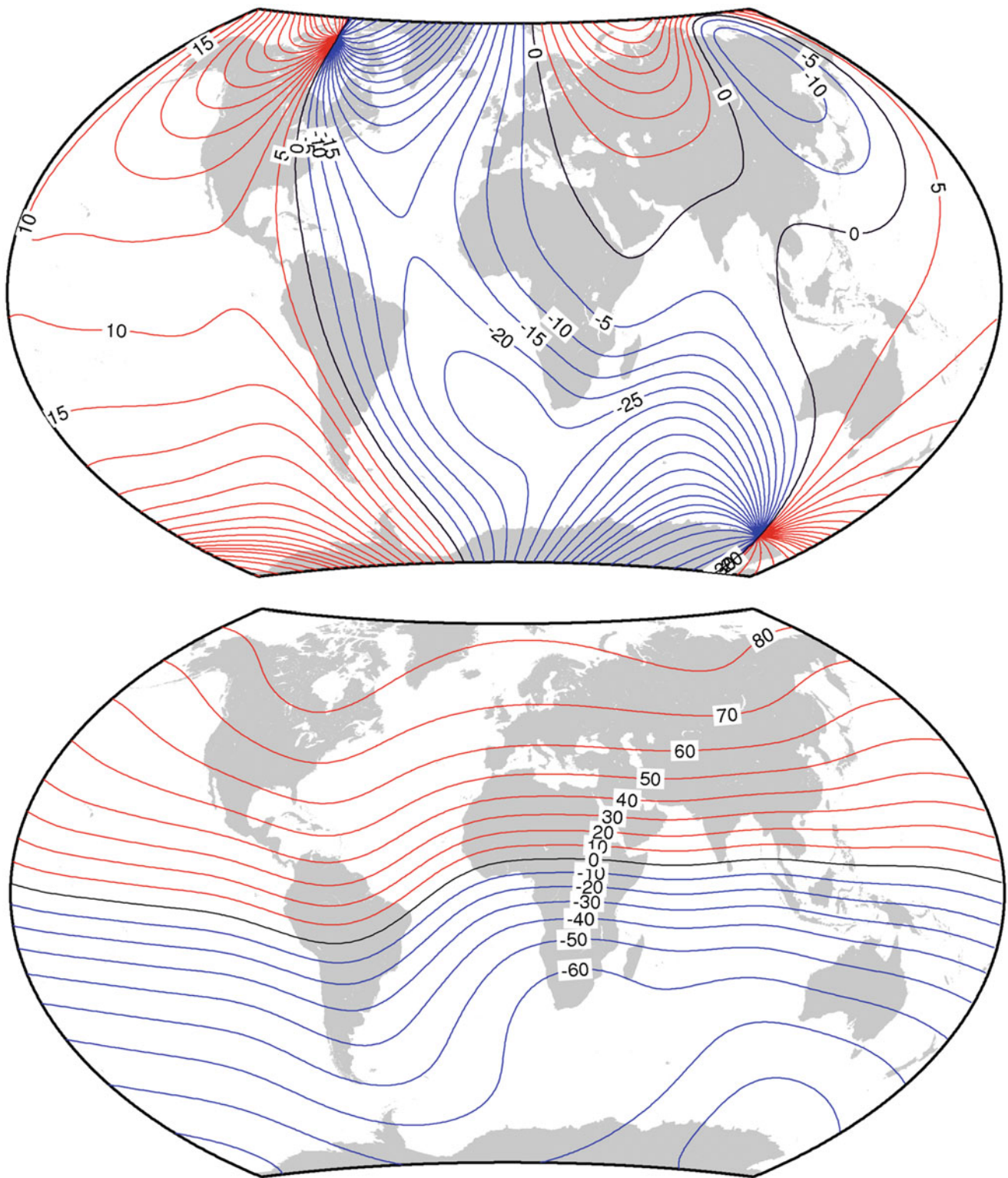
Features of note on these maps are the dip equator and dip poles in the maps of inclination (Figs. 1, 2, and 8) and the area of weak magnetic field in the South Atlantic, and strong magnetic field near the poles in Figs. 2 and 9. The dip equator is where the magnetic field is horizontal (inclination = 0°) and a current system is set up in the upper atmosphere called the equatorial electrojet. The dip poles (also called magnetic poles) are locations where the magnetic field is, on average, vertical. Another set of poles is the dipole poles or geomagnetic poles. Their positions can be derived from the degree 1 spherical harmonic coefficients (g_1^0 , g_1^1 , and h_1^1). Associated with the geomagnetic poles, in approximately oval shaped loci, are the auroral electrojets in the upper atmosphere. The electrojets, particularly the auroral electrojets, generate rapid time-varying magnetic fields of significant amplitude.

Recent Changes in the Geomagnetic Field

From a comparison of the declination maps in Figs. 2 and 7, it can be seen that the zero contour (the agonic line) is



Geomagnetic Field, Global Pattern, Fig. 1 Declination and inclination at 1600.0 (degrees) computed from the *gufm1* model

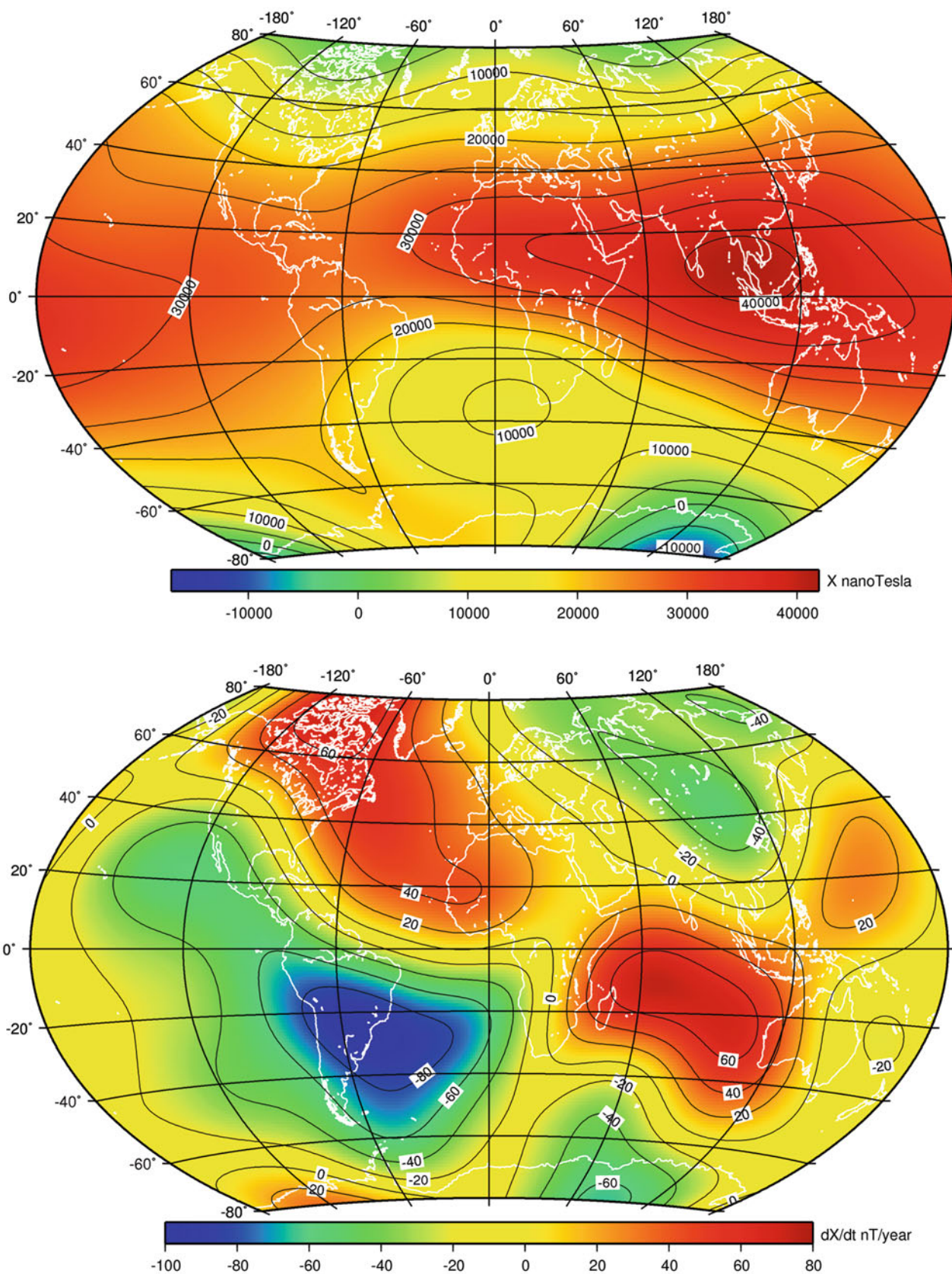


Geomagnetic Field, Global Pattern, Fig. 2 Declination, inclination (degrees), and total intensity (nT) at 1950.0 computed from the *gufm1* model

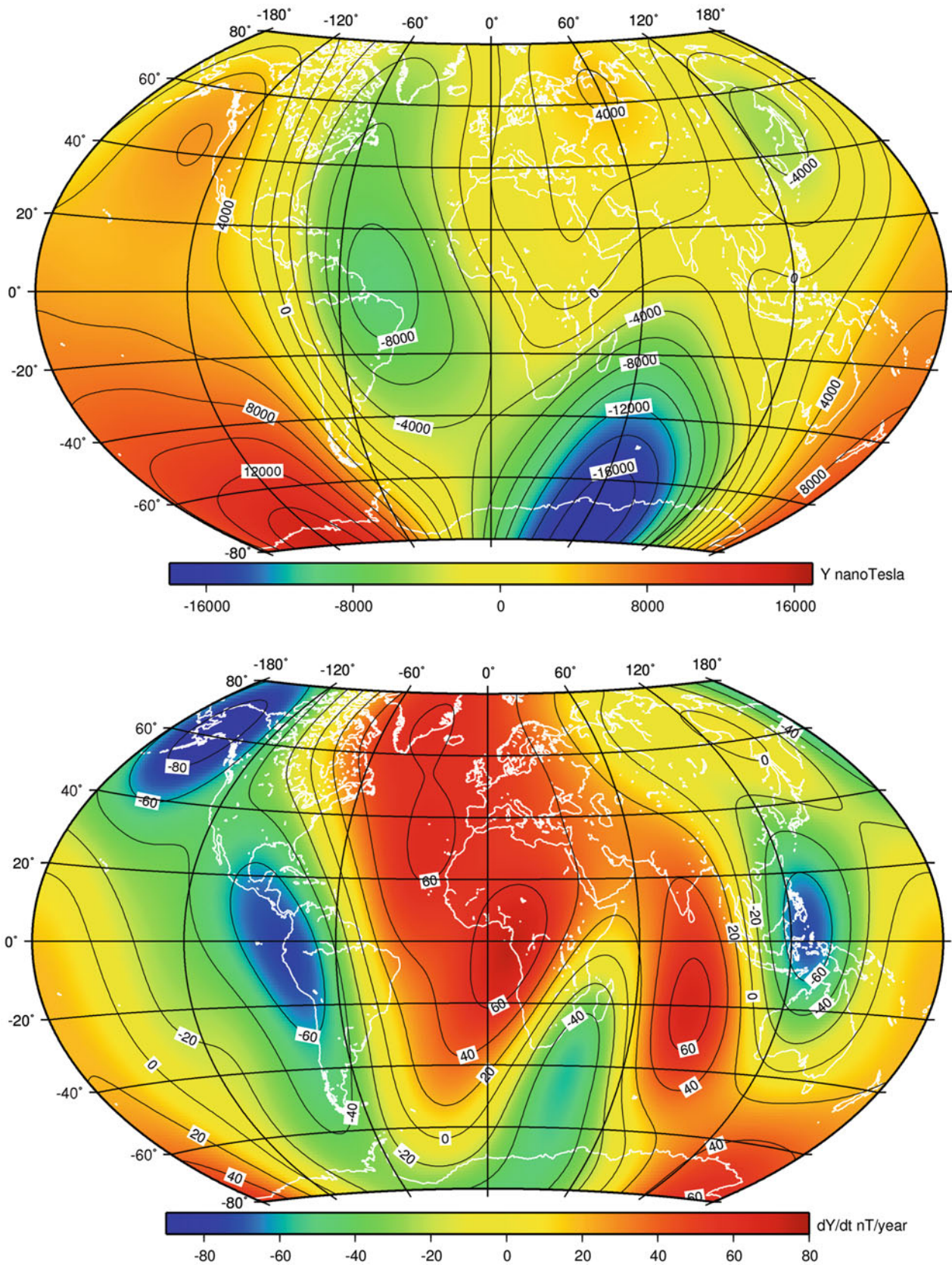
moving slowly westwards with time. This westwards movement of the magnetic field pattern at the Earth's surface, particularly prevalent in the Atlantic hemisphere, is related to the motion of fluid at the core surface

slowly westwards, dragging with it the magnetic field lines.

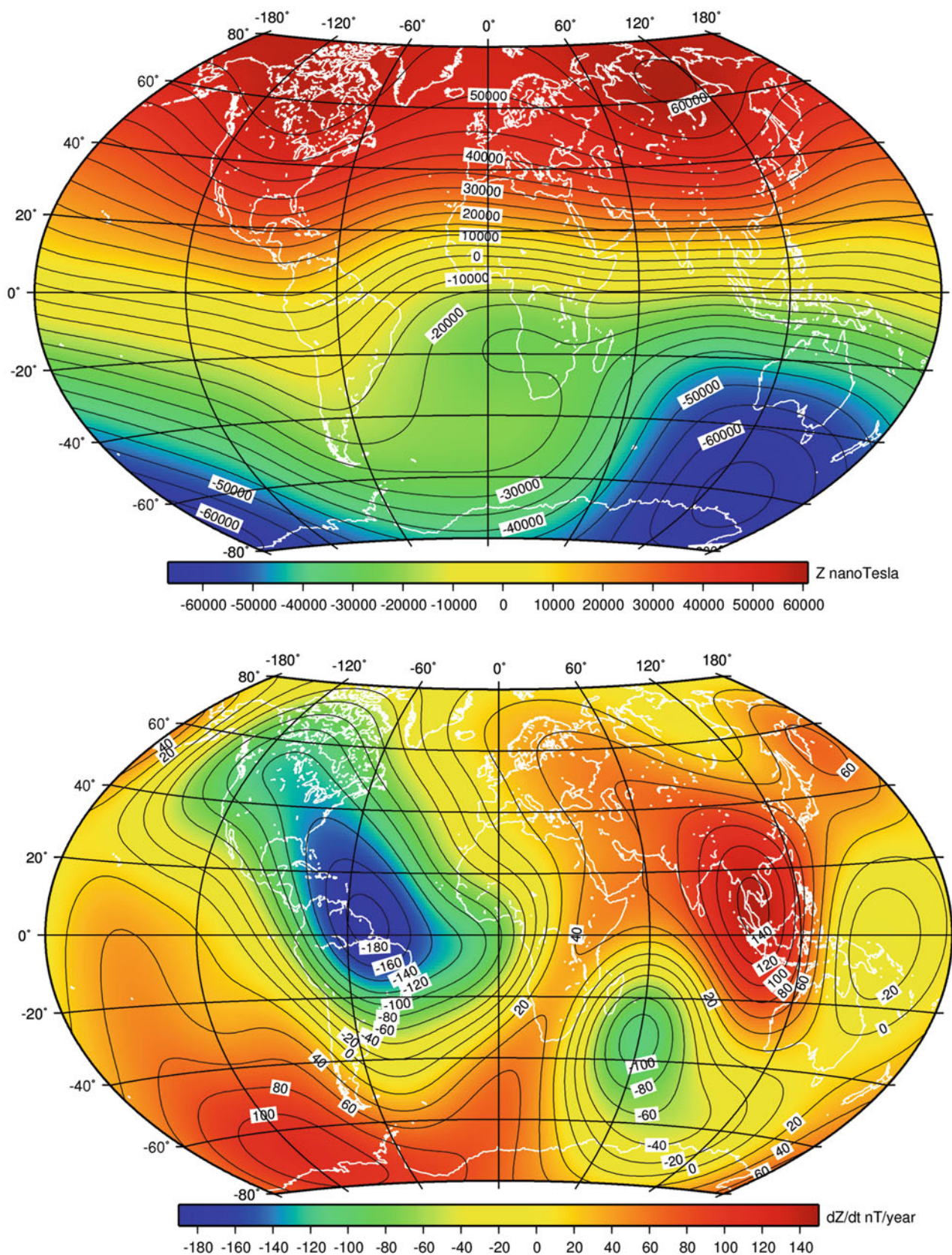
At high northern latitudes from 2000 to 2020, the motion of the north dip pole has been more than 50 km per year and



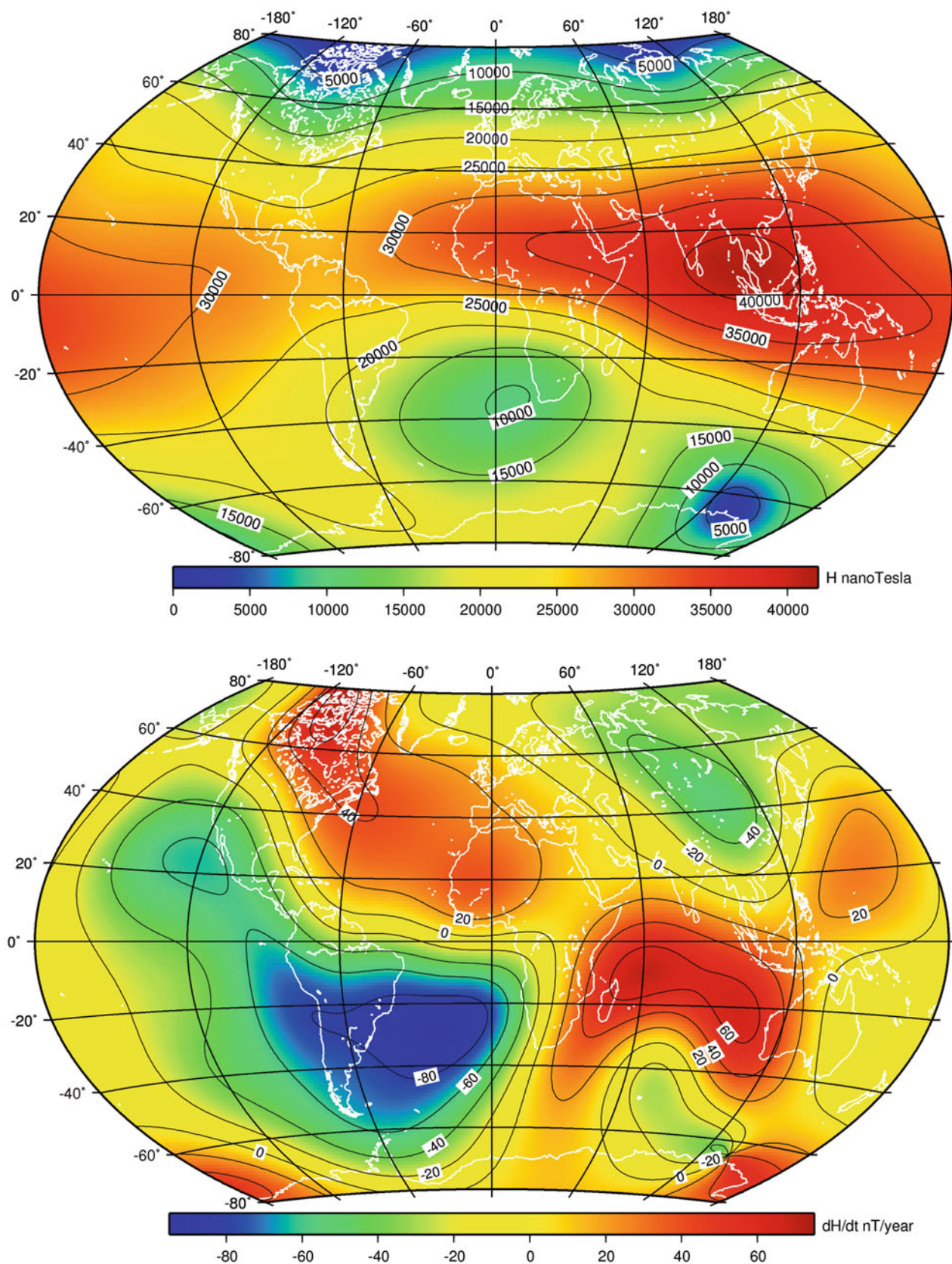
Geomagnetic Field, Global Pattern, Fig. 3 Northern intensity (nT) at 2015.0 and its rate of change (nT/year) for 2015.0–2020.0 computed from IGRF-12



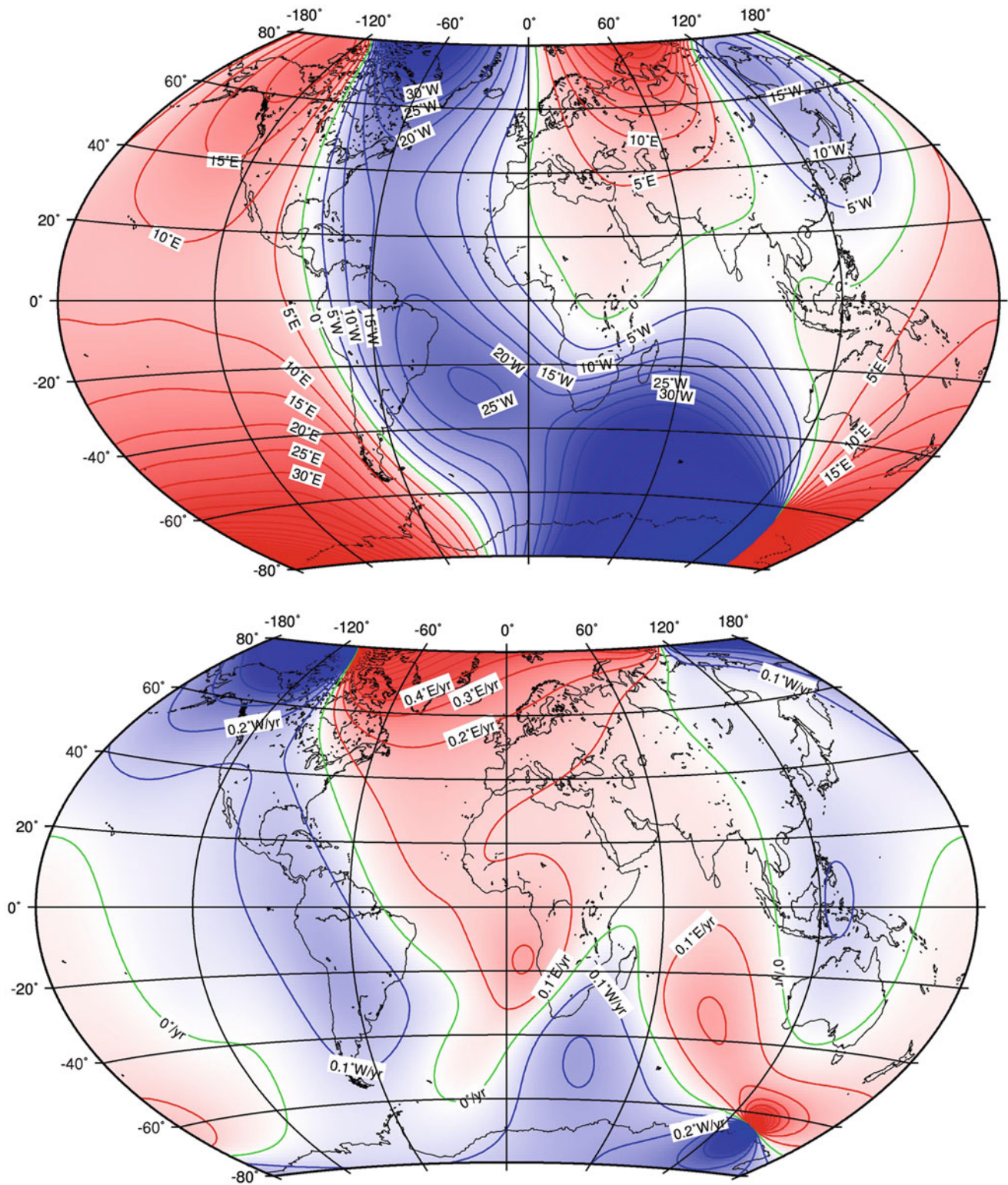
Geomagnetic Field, Global Pattern, Fig. 4 Easterly intensity (nT) at 2015.0 and its rate of change (nT/year) for 2015.0–2020.0 computed from IGRF-12



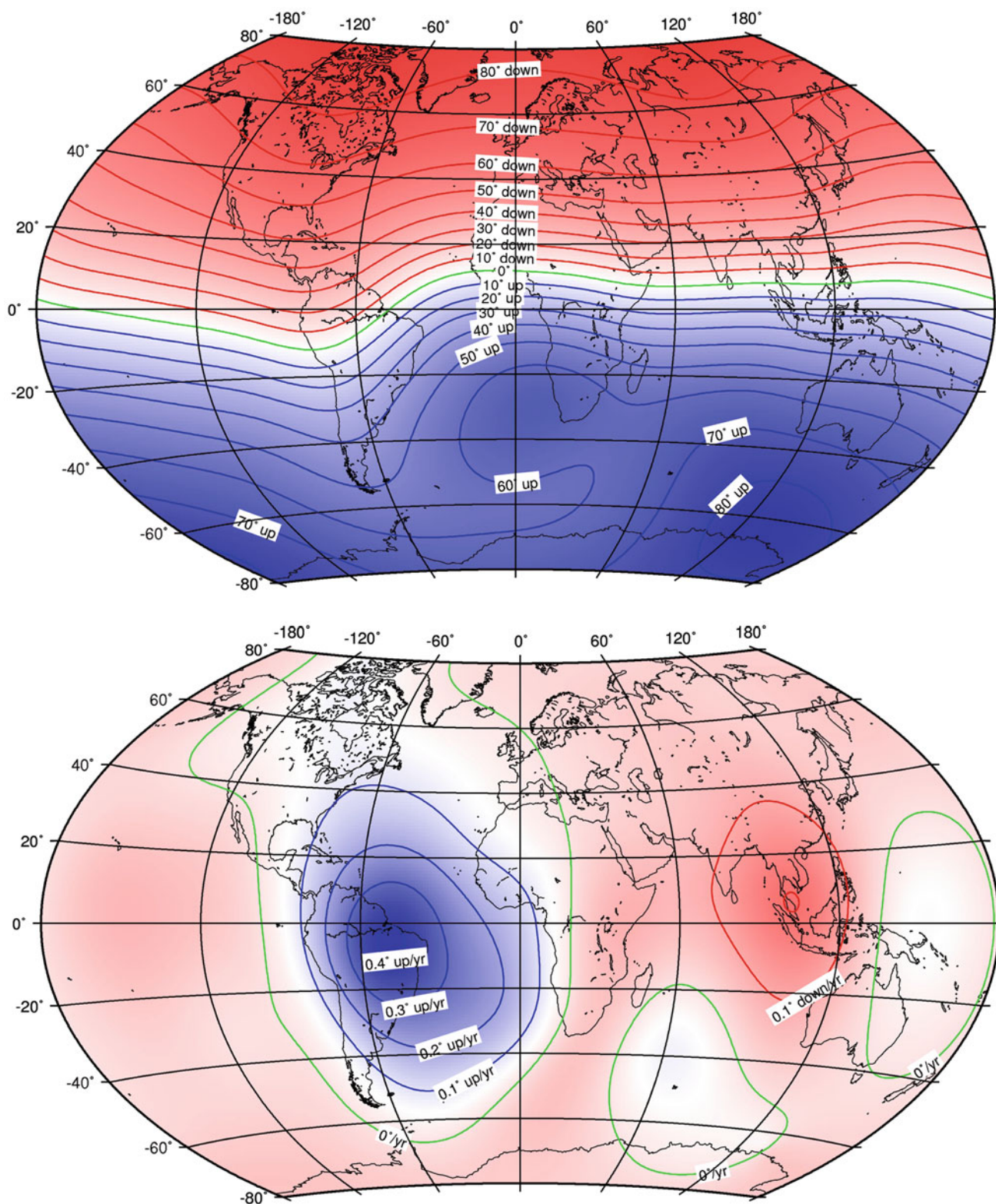
Geomagnetic Field, Global Pattern, Fig. 5 Vertical intensity (nT) at 2015.0 and its rate of change (nT/year) for 2015.0–2020.0 computed from IGRF-12



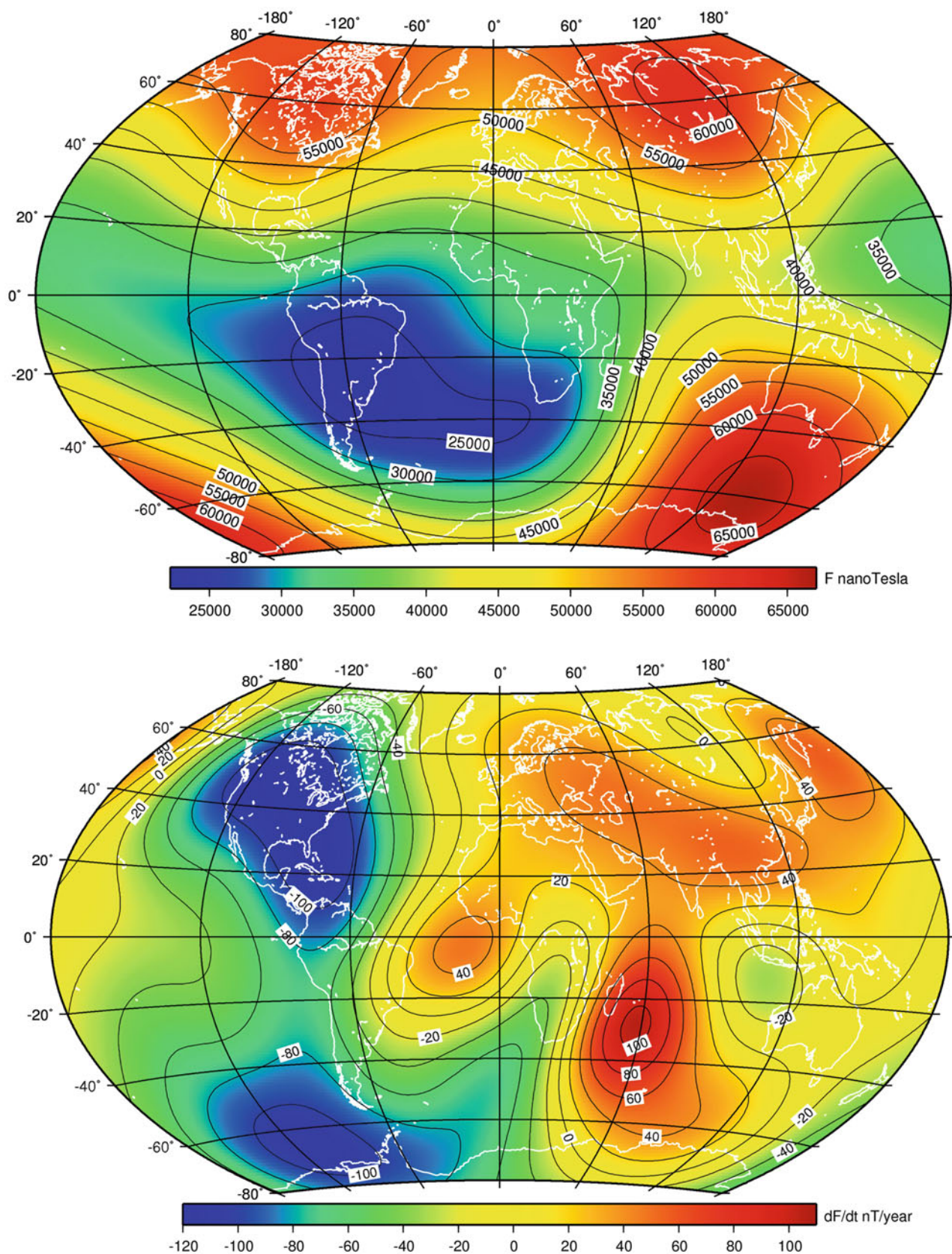
Geomagnetic Field, Global Pattern, Fig. 6 Horizontal intensity (nT) at 2015.0 and its rate of change (nT/year) for 2015.0–2020.0 computed from IGRF-12



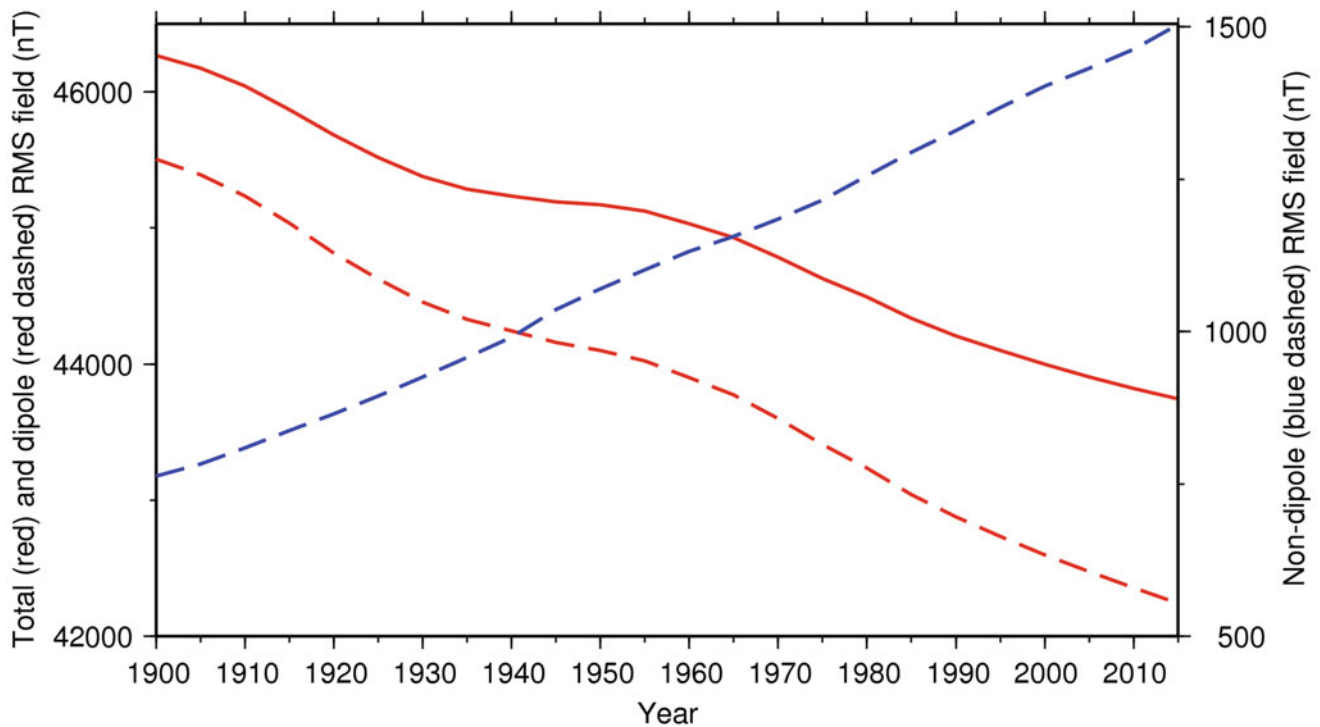
Geomagnetic Field, Global Pattern, Fig. 7 Declination (degrees) at 2015.0 and its rate of change (arc-minutes/year) for 2015.0–2020.0 computed from IGRF-12



Geomagnetic Field, Global Pattern, Fig. 8 Inclination (degrees) at 2015.0 and its rate of change (arc-minutes/year) for 2015.0–2020.0 computed from IGRF-12



Geomagnetic Field, Global Pattern, Fig. 9 Total intensity (nT) at 2015.0 and its rate of change (nT/year) for 2015.0–2020.0 computed from IGRF 12



Geomagnetic Field, Global Pattern, Fig. 10 Decline of the whole, and dipolar part of the geomagnetic field at the Earth's surface and the growth of the nondipolar part since 1900, computed from IGRF-12

sometime during 2017/2018 it reached its highest latitude (86.6°N). According to the latest IGRF, we are now seeing a deceleration of this motion, but the pole is still moving at roughly 40 km per year, and heading southwards. In contrast, the southern dip pole, near Antarctica, has moved relatively little in the past few hundred years.

Using IGRF-12 to compute the root mean square magnetic field vector at the Earth's surface arising from all spherical harmonic terms ($n \leq 10$), the centered dipole terms ($n = 1$), and the nondipole terms ($1 < n \leq 10$), it can be seen in Fig. 10 that since 1900 the geomagnetic field is weakening overall by becoming less dipolar. However, the nondipolar part is strengthening. This may have consequences for the trajectories of energetic charged particles that enter the Earth's magnetosphere. One manifestation of this is the deepening, and westwards movement, of the South Atlantic Anomaly, the region where the geomagnetic field is weaker than elsewhere (see Figs. 2 and 9). Energetic charged particles are able to penetrate closer to the Earth and cause a radiation hazard for satellites passing through this region.

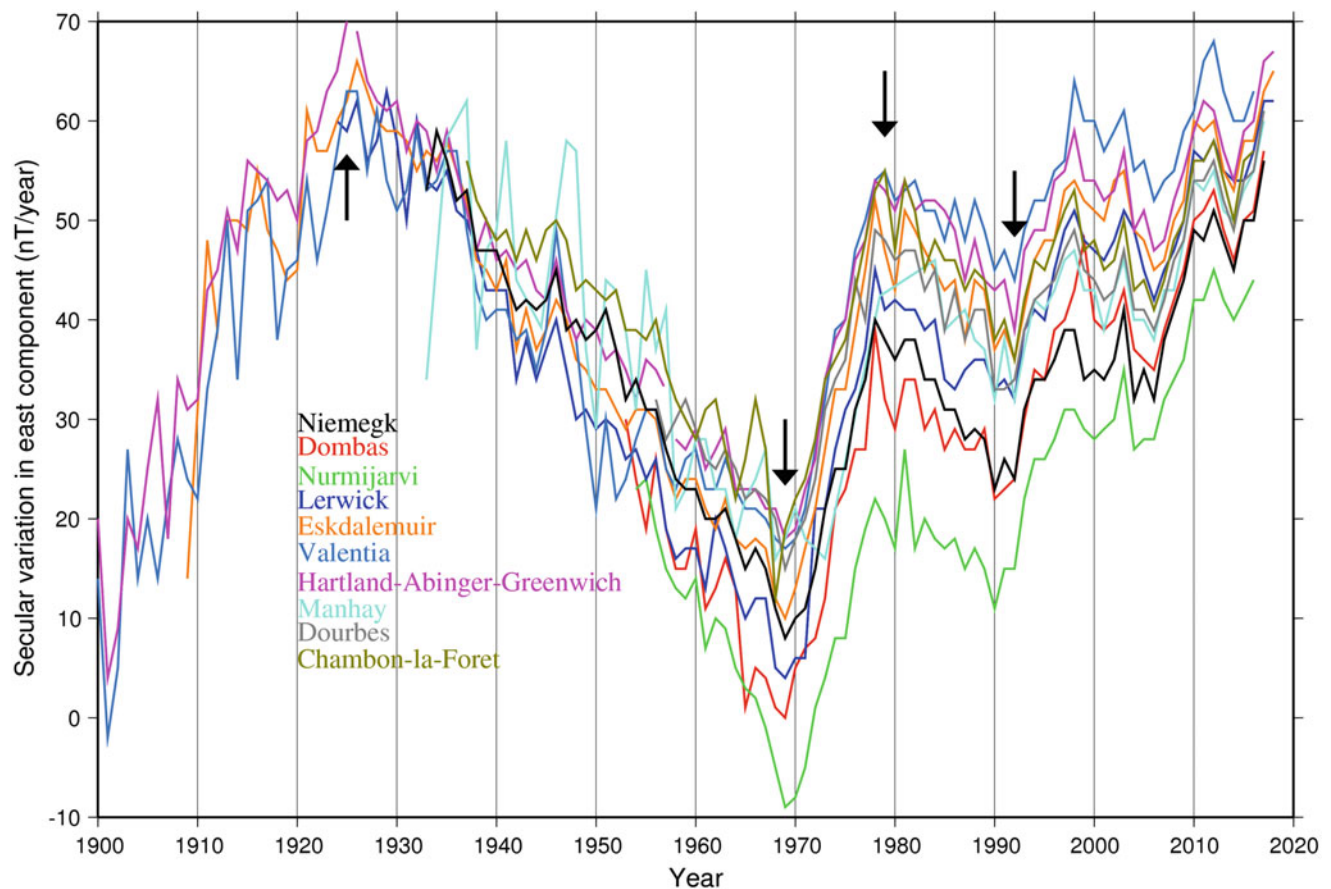
Superimposed on these gradual changes of the magnetic field are the so-called geomagnetic jerks. They are relatively abrupt (duration of months) changes in the second time derivative, or secular acceleration, of the magnetic field. The first observed geomagnetic jerk was that around 1969, and since the late nineteenth century when direct and continuous measurements of the Earth's magnetic field have been available, geomagnetic

jerks have also been widely accepted to have occurred around 1925, 1978, and 1991. These jerks are most readily observed in the first time derivative of the easterly intensity at European observatories (Fig. 11). More recent jerks are also being detected, sometimes with help of data from satellites.

Various analysis techniques have been applied to jerks to investigate specific aspects of their temporal and spatial characteristics. For example, Brown et al. (2013) used minimal a priori information to accurately time the occurrence of jerks around the world and establish their amplitudes using observatory data. The variable timing is particularly interesting for studies of the geodynamo though the effects of varying conductivity in the deep mantle on the arrival time of the jerk signals at the Earth's surface cannot be discounted (Pinheiro et al. 2015). Understanding the origin of jerks is also important for improving time-dependent models of the geomagnetic field and for the strictly practical purpose of forecasting its future behavior, for example, as used in navigation and orientation.

Summary

In this chapter, we have described the global mapping of the geomagnetic field for the 400+ year period during which direct field observations have been available. The mapping technique involving the fitting of spherical harmonic models



Geomagnetic Field, Global Pattern, Fig. 11 Geomagnetic jerks as seen in the secular variation of the east component of the geomagnetic field observed at European observatories. Times of jerks are shown by arrows

is described, and maps of the various components of the magnetic field are provided for 1600, 1950 and the present day. The geomagnetic field is constantly changing on a range of temporal and spatial scales, and these changes provide scientists with important information about processes deep inside the Earth.

Cross-References

- [Geomagnetic Field, IGRF](#)
- [Geomagnetic Field, Measurement Techniques](#)
- [Geomagnetic Field, Secular Variation](#)
- [Geomagnetic Field, Theory](#)
- [Magnetic Anomaly Map, Global](#)
- [Magnetic Methods, Principles](#)

Bibliography

- Barracough DR (1978) Spherical harmonic models of the geomagnetic field. *Geomagnetic Bulletin Institute of Geological Sciences* No 8
- Brown WJ, Mound PW, Livermore PW (2013) Jerks abound: an analysis of geomagnetic observatory data from 1957 to 2008. *Phys Earth Planet Int* 223:62–76

- Clark T, Barracough D (2001) The first magnetic chart. *Astron Geophys* 42:1.23–1.25
- Malin SRC (1971) British world magnetic charts. In: Zmuda AJ (ed) *World magnetic survey 1957–1969*. IAGA Bulletin No 28
- Pinheiro KJ, Jackson A, Amit H (2015) On the applicability of Backus' mantle filter theory. *Geophys J Int* 200:1336–1346
- Thébault et al (2015) International geomagnetic reference field: the 12th generation. *Earth Planets Space* 67–79. <https://doi.org/10.1186/s40623-015-0228-9>

Geomagnetic Field, IGRF

Aude Chambodut

Institut de Physique du Globe de Strasbourg, Université de Strasbourg/Ecole et Observatoire des Sciences de la Terre, CNRS, Strasbourg Cédex, France

Definition

The International Geomagnetic Reference Field (IGRF) is a series of mathematical models of the Earth's main field and its annual rate of change (secular variation).

Each generation of the IGRF comprises three constituent models:

1. The IGRF model that is a prediction of the state of the main field at an epoch t (considering that the epoch of IGRF model elaboration is approximately between $[t - 3 \text{ years}]$ and $[t - 1 \text{ year}]$).
2. The secular variation model that is also a prediction of the annual variation rate of the main field between epoch t and $(t + 5 \text{ years})$.
3. The Definitive Geomagnetic Reference Field model (DGRF), which is the revision of the previous generation IGRF model 5 years before the epoch $(t - 5 \text{ years})$.

The constituent models are designated provisional or definitive. Once a constituent model is designated definitive, it is not revised in subsequent generations of the IGRF.

Datasets

In Table 1, the periodicity of IGRF's elaboration varies with time. In addition, upgrades are regularly incorporated. For example, the truncation degree/order of the main field spherical harmonic analysis was extended from 10 up to 13 after 1995 (IAGA, Division V, Working Group 8 et al. 2003). Similarly, the precision of the coefficients was extended from 1 nT to 0.1 nT after 1995 (IAGA, Division V, Working Group 8 et al. 2003), and then to 0.01 nT after 2000 (IAGA, Working Group V-MOD et al. 2010, 2015, 2020).

These two changes were made possible by technological advances. Indeed, before 1979, only near-surface data (permanent observatories, repeat stations, land, airborne, and marine surveys; see ► “Geomagnetic Field, Measurement Techniques” and ► “Magnetic Methods, Airborne”) were used (Peddie 1983; Zmuda 1971), and consequently, the spatial distribution of data over the globe remained poor. The International Decade of Geopotential Field Research allowed modelers to take advantage of the excellent spatial coverage and high quality of measurements made by near-polar satellites (see ► “Magnetic Methods, Principles”). The POGO satellites (1965–1971), Magsat (1979–1980), POGS (1990–1993), Oersted (1999–2014), SAC-C (2000–2013), and CHAMP (2000–2010) have all been utilized in the production of the IGRF.

Today, data from permanent magnetic observatories are still the most important source of information about time-varying fields, especially for secular variation constituent models (see ► “Geomagnetic Field, Secular Variation”), while magnetic satellite missions, as the Swarm constellation (2013–), allow the acquisition of homogeneously distributed data over the entire globe during a short period of time (as compared to magnetic observatories time series). The two datasets remain highly complementary.

Basic Mathematical Expression

Each mathematical model consists in a set of coefficients, called Gauss coefficients, associated with solid spherical harmonic functions.

Indeed, in source-free regions (the spherical shell bounded by the Earth's surface and the base of the ionosphere at an altitude around 100 km), the main field, generated by fluid motions in the Earth's electrically conducting liquid outer core, may be expressed as the negative gradient of a scalar potential V , which is represented by a truncated (up to degree N) series expansion (see ► “Geomagnetic Field, Theory”):

$$V(r, \theta, \phi) = R \sum_{n=1}^N \left(\frac{R}{r}\right)^{n+1} \sum_{m=0}^n (g_n^m \cos(m\phi) + h_n^m \sin(m\phi)) P_n^m(\theta)$$

where r , θ , and ϕ are geocentric coordinates according to a reference coordinate system.

r is the distance from the center of the Earth, θ is the colatitude (i.e., $90^\circ - \text{latitude}$), and ϕ is the longitude, R is a reference radius (Earth mean radius) set to 6371.2 km; g_n^m and h_n^m are the Gauss internal coefficients, and $P_n^m(\theta)$ are the Schmidt quasi-normalized associated Legendre function of degree n and order m .

The IGRF coefficients are time dependent. In the series of IGRF mathematical models, the variations are assumed to be linear on 5-year intervals. For the upcoming 5-year epoch, the rate of change is given by predictive secular variation coefficients (see ► “Geomagnetic Field, Secular Variation”).

IGRF Evolving

One of the first suggestions to develop an International Geomagnetic Reference Field model was made in 1954 at the International Association of Geomagnetism and Aeronomy (IAGA) Scientific General Assembly in Toronto. The aim in developing such reference models was to have a common basis for calculating the main geomagnetic field in order to homogenize the results of academical/fundamental studies (on fields as diverse as the crustal and external fields) and to provide the industrial world (Meyers and Davis 1990) with a well-identified description of the Earth's core magnetic field.

Thus, an internationally coordinated effort was carried out under the aegis of IAGA relying on the cooperation between magnetic field modelers, institutes, and agencies responsible for collecting and publishing geomagnetic field data.

Since 1969, date of the first adoption of an IGRF (IGRF-1 valid for 1955–1975), the IGRF has evolved into a series of mathematical models of the Earth's main field (Langel 1987) and its annual rate of change at 5-year intervals covering the period 1900–2025 (Table 1).

Geomagnetic Field, IGRF, Table 1 Summary of IGRF history

Full name	Short name	Valid for	Definitive for
IGRF 13th generation	IGRF-13	1900–2025	1945–2015
IGRF 12th generation (revised 2019)	IGRF-12	1900–2020	1945–2010
IGRF 11th generation (revised 2009)	IGRF-11	1900–2015	1945–2005
IGRF 10th generation (revised 2004)	IGRF-10	1900–2010	1945–2000
IGRF 9th generation (revised 2003)	IGRF-9	1900–2005	1945–2000
IGRF 8th generation (revised 1999)	IGRF-8	1900–2005	1945–1990
IGRF 7th generation (revised 1995)	IGRF-7	1900–2000	1945–1990
IGRF 6th generation (revised 1991)	IGRF-6	1945–1995	1945–1985
IGRF 5th generation (revised 1987)	IGRF-5	1945–1990	1945–1980
IGRF 4th generation (revised 1985)	IGRF-4	1945–1990	1965–1980
IGRF 3rd generation (revised 1981)	IGRF-3	1965–1985	1965–1975
IGRF 2nd generation (revised 1975)	IGRF-2	1955–1980	–
IGRF 1st generation (revised 1969)	IGRF-1	1955–1975	–

New candidate models are carefully produced and widely documented as are their evaluation procedures. The final decision is made by the IAGA Working Group-V Mod. Most of the time, the elected DGRF and IGRF models of the current generation are derived by taking means (sometimes weighted) of the coefficients of selected candidate models. Indeed, one particular selected candidate model cannot be chosen preferably to another as the various datasets, selection criteria of the data, applied corrections, and even modeling methods have their own strengths and weaknesses. The robustness of IGRF lies in the fact that even if there are significant differences between the candidate models in every part of the development chain (datasets, data selections, algorithms, and regularizations), the results are fairly consistent at the end.

General Information about the IGRF and coefficients are available at: <http://www.ngdc.noaa.gov/IAGA/vmod/igrf.html>.

Additional figures and programs to compute the field components for a specific location and time are available at: <http://wdc.kugi.kyoto-u.ac.jp/igrf/index.html> and http://www.geomag.bgs.ac.uk/data_service/models_compass/igrf_form.shtml.

Summary or Conclusions

The IGRF model is the product of a collaborative effort between magnetic field modelers and the institutes involved in collecting and disseminating magnetic field data from

ground observatories and surveys around the world as well as from satellites. IGRF provides a description of the Earth's internal magnetic field.

The IGRF model is a widely used common reference in many academic (from solid earth to space weather) and applied (navigation, localization, communication, etc.) scientific fields.

Cross-References

- [Geomagnetic Field, Measurement Techniques](#)
- [Geomagnetic Field, Secular Variation](#)
- [Geomagnetic Field, Theory](#)
- [Magnetic Methods, Airborne](#)
- [Magnetic Methods, Principles](#)

Bibliography

- International Association of Geomagnetism and Aeronomy (IAGA) Division V, Working Group 8, Macmillan S et al (2003) The 9th generation international geomagnetic reference field. *Earth Planets Space* 55:i–ii
- International Association of Geomagnetism and Aeronomy (IAGA), Working Group V-MOD, Finlay CC et al (2010) International geomagnetic reference field: the eleventh generation. *Geophys J Int* 183(3):1216–1230. <https://doi.org/10.1111/j.1365-246X.2010.04804.x>
- International Association of Geomagnetism and Aeronomy (IAGA), Working Group V-MOD, Thébault E et al (2015) International geomagnetic reference field: the 12th generation. *Earth Planet Space* 67:79. <https://doi.org/10.1186/s40623-015-0228-9>
- International Association of Geomagnetism and Aeronomy (IAGA), Working Group V-MOD, Alken P et al (2020) International geomagnetic reference field: the thirteenth generation. *Earth Planet Space*. in press. <https://earth-planets-space.springeropen.com/igrf13>
- Langel RA (1987) The main field. In: Jacobs JA (ed) *Geomagnetism*, vol 1. Academic, London, pp 249–512
- Meyers H, Davis WM (1990) A profile of the geomagnetic model users and abusers. *J Geomagn Geoelectr* 42:1079–1085
- Peddie NW (1983) International geomagnetic reference field – its evolution and the difference in total field intensity between new and old models for 1965–1980. *Geophysics* 48:1691–1696
- Zmuda AJ (1971) The international geomagnetic reference field: introduction. *Bull Int Assoc Geomagn Aeron* 28:148–152

Geomagnetic Field, Measurement Techniques

Mioara Manda¹ and Anca Isac²

¹CNES – Centre National d'Etudes Spatiales, Paris, France

²Geological Institute of Romania, Surlari National Geomagnetic Observatory, Bucharest, Romania

Introduction

At any point in space, the magnetic field is a vector quantity, characterized by a direction associated with a field strength.

A complete representation of the geomagnetic field needs not only the value of its intensity, but also its direction. Continuous monitoring, modeling, and interpretation of the geomagnetic field are the prerequisites to understand our magnetic Planet. Based on geomagnetic models (► [“Geomagnetic Field, Global Pattern”](#), S. Macmillan), covering very short to longer temporal scales, it makes it possible to understand core dynamics (► [“Core Dynamo”](#), U. Christensen), lithospheric magnetic signals (► [“Magnetic Anomaly Map, Global”](#), H. Kumar; ► [“Magnetic Anomalies: Interpretation”](#), E. Thebaud), and the geomagnetic environment. This represents an important challenge in understanding the System Earth. Clearly, geomagnetic measurements from ground and space represent the main ingredient in modeling and understanding the Earth’s magnetic field (e.g., Manda and Purucker 2018; ► [“Geomagnetic Field, IGRF”](#), A. Chambodut). Furthermore, magnetic measurements are also used in connection with different geophysical surveys (ground and/or airborne) for detecting mineralization and/or geological and tectonic structures. They are also used to detect archaeological sites, shipwrecks, and other buried or submerged objects. Moreover, magnetic techniques are used at drilling sites for navigating well-bores, or in monitoring the geomagnetically induced currents which can result in damage to transformers, a potentially major expense to power companies. Finally, prediction of space weather is crucial considering its potential impact on man-made technologies on Earth and in space (satellites, electricity power grids, pipelines, radio and telephone communications, geophysical exploration, etc.).

Two types of measurements are needed to characterize the geomagnetic field: scalar and vector. Scalar measurements are based on the transformation of a frequency – a very accurate measurement – into intensity of the magnetic field. In the spectral domain, the measurement is indeed the location of the center of the spectral line. In a proton magnetometer, protons in a liquid (such as water or kerosene) are first polarized; when the polarizing field is switched off, the protons precess around the magnetic field direction with a frequency which is proportional to the magnetic field intensity. Overhauser magnetometers are typically installed at magnetic observatories, having been also used for some other ground observations or on board of the Ørsted and CHAMP magnetic satellites.

In contrast, vector measurements made with a fluxgate magnetometer are based on the nonlinearity of the magnetization of soft magnetic materials. The very sensitive part of this instrument is an easily saturable core made of a material with high permeability. This core is wrapped by two windings: an excitation coil and a pick-up coil. An alternating excitation current is passed through the excitation coil, driving the core through an alternating cycle of magnetic saturation. This changing field induces an electrical current in the pick-up coil, and this output current can be accurately

measured. The fluxgate magnetometers are subject to an instrument drift. To minimize this drift in the final data, different approaches are used to calibrate the fluxgate magnetometers used for ground and satellite observations.

The instruments used to measure the magnetic field on ground or in space are based on the same physical concepts, and they are basically the same. Of course, for space missions constraints are imposed by weight and consuming energy. More details on magnetometers used in ground-based observations and newly studies on absolute calibration of a vector magnetometers are given in section [“Earth’s Surface Magnetic Field Measurements”](#). Measurements carried out by satellite missions are presented in section [“Near-Earth Magnetic Field Measurements”](#). Finally, in section [“Summary”](#) we offer a perspective on promising avenues in measuring the magnetic field.

Earth’s Surface Magnetic Field Measurements

General Observatory Setting

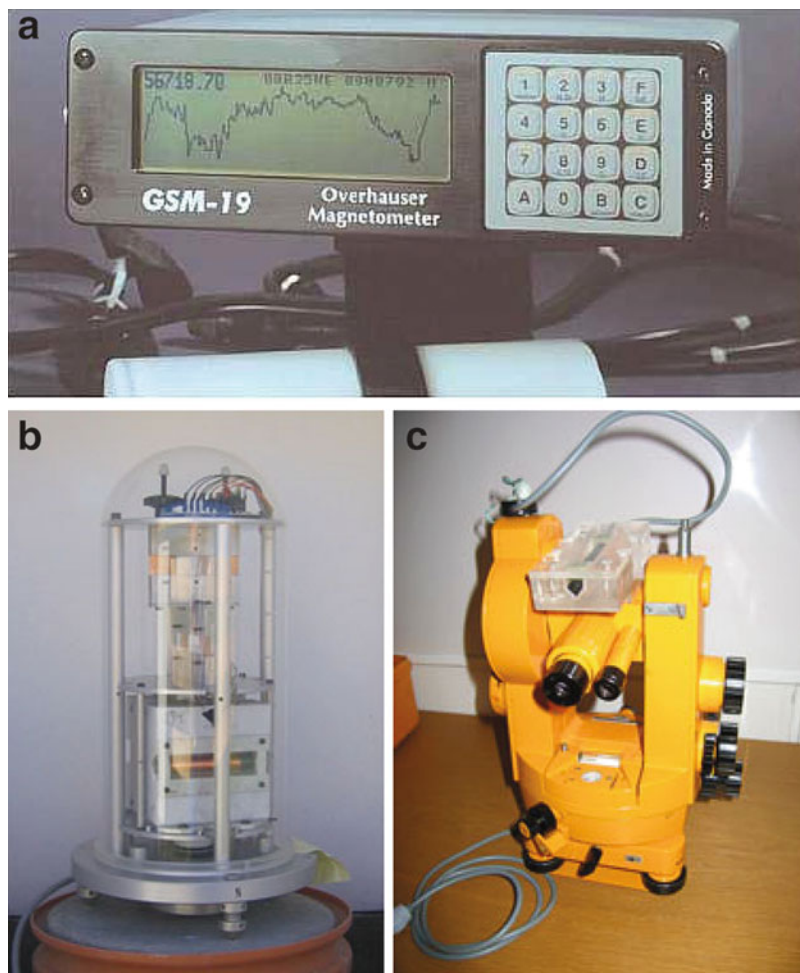
Historically, magnetic observatories were established to monitor the secular change of the Earth’s magnetic field, and this remains one of their major functions (► [“Geomagnetic Field, Secular Variation”](#), M. Korte). This, generally involves *absolute measurements* that are sufficient in number to monitor instrumental drift of fluxgate magnetometers giving *variation measurements* of the three field components (X – North, Y – East and Z – vertical), in arbitrary units (► [“Magnetometers”](#), I. Hrvoic).

Absolute measurements. A scalar measurement of the field intensity obtained commonly by a proton magnetometer is absolute: as noticed before it depends only upon our knowledge of a physical constant and a measurement of frequency. The measurements achieved with such an instrument have a great accuracy (in excess of 10 ppm). Fig. 1a shows an Overhauser magnetometer.

Scalar magnetometers make measurements of the strength of the magnetic field only and provide no information about its direction. It is also possible to make an absolute measurement of direction of the geomagnetic field, by measuring the angle I (*inclination*) with respect to the horizontal plane, and the angle D (*declination*) in the horizontal plane between the magnetic North and true North (the last one determined by reference to a fixed mark of known azimuth, obtained astronomically or by using a gyro). A fluxgate-theodolite (DI-flux), as shown in Fig. 1b, used commonly to measure D and I , requires manual operation. The DI-flux consists of a nonmagnetic theodolite and a single-axis fluxgate sensor mounted on top of a telescope. The DI-flux is considered to be an absolute instrument, which means that the angles measured by the instrument do not deviate from the true values D and I . This is achieved by using an observation procedure that eliminates unknown parameters such as sensor offset,

Geomagnetic Field, Measurement Techniques,

Fig. 1 Geomagnetic observatory sensor package: proton magnetometer (a), fluxgate magnetometer (b), theodolite with fluxgate coupled to the telescope (c)



collimation angles, and theodolite errors. For example, collimation errors between the fluxgate sensor and the optical axis of the theodolite, and within the theodolite, are minimized by taking readings from four telescope positions. With the fluxgate sensor operating in the null-field mode, the stability and sensitivity of the sensor and its electronics are maximized. In a land-based observatory, such absolute measurements are typically performed once/twice a week, manually, and are used to monitor the drift of the fluxgate variometers.

Over the last decade, high temporal and field resolution data are requested from observatories to complement the magnetic satellite magnetic. One option comes from the dIdD (delta Inclination/delta Declination) magnetometers recording the field components by using a coil system around a proton magnetometer (e.g., Csontos et al. 2012). A second type of instruments aiming to provide continuous (quasi-) absolute recordings are the optically pumped magnetometers, where the sensor contains a gas, helium, or an alkali vapor such as cesium, rubidium, or potassium. These instruments have higher measurement rates and lower noise than Overhauser magnetometers, being suitable for fast moving platforms, as airplanes and satellites.

Variation measurements. A vector measurement is made with a fluxgate magnetometer which operates with minimal manual intervention. These instruments require temperature-controlled environments and installation on extremely stable platforms (however, some modern systems are suspended and therefore compensate for platform tilt). Even with these precautions they can still be drifting subject arising from sources both within the instrument (temperature effects) and also the stability of the instrument mounting. Because these measurements are not absolute, they are referred to as variation measurements, they have to be calibrated against standard instruments, and then the instruments are known as *variometers*. One of the most widely used variometers is the FGE fluxgate (see Fig. 1c). The sensor unit consists of three orthogonally mounted sensors on a marble cube. In order to improve long-term stability, these sensors have compensation coils wound on quartz tubes, resulting in a sensor drift of only a few nT per year. The marble cube is suspended by two strips of crossed phosphor-bronze working as a cardanic suspension to compensate for pillar tilting that might cause baseline drift. The box containing the electronics is almost magnetic free and is placed several meters from the sensor to further

minimize its effect on the readings. The magnitude and directional response of these instruments needs to be calibrated against accurately known sources.

All modern land-based magnetic observatories use similar instrumentation and process the data in a comparable way to produce similar data products. For a full description, see the INTERMAGNET web site (<http://www.intermagnet.org>; accessed June 2020). The fundamental measurements recorded are averaged one-minute values of the vector components and of scalar intensity. From the one-minute data, hourly, daily, monthly, and annual mean values are produced.

Due to developments in observatory instrumentation and data acquisition, filtered one-second data product have become available through the INTERMAGNET network. Since 2014 definitive one-second data are provided to the user community. The minimum standard of magnetometer performance is actually given by high stability and low noise, good linearity, high resolution – <0.01 nT, and fast dataloggers with an accuracy of the time-stamp – within 0.01 s, and sharp filtering. All the parameters describing the specifications of a complete observatory system for a one-second vector data are resumed in Turbitt et al. (2013).

Finally, let us note that data processing consists in a weekly determination of baseline values from absolute measurements. The processing has to include the determination of all systematic errors, and the quality of measurements depends strongly on the experience and accuracy of the observer. Because of this subjective parameter in the final data products and because regular absolute measurements lack in remote areas, several attempts have been made to operate absolute vector field instruments automatically. Already, some 50 years ago, Alldredge planned an automatic standard magnetic observatory (Alldredge and Saldukas 1964), that is, a device providing at each time the absolute values of the geomagnetic field (without extra independent absolute measurements, at least for a long enough time span). This idea has remained in the geomagnetism community and, over the last years, attempts have been made to automate a DI-theodolite (van Loo and Rasson 2007), or to build a device which can perform discrete absolute measurements automatically (Auster et al. 2007). One important issue remains the orientation problem for an automatic magnetic observatory, and here we briefly summarize it.

Absolute Calibration of a Vector Magnetometers

Gravrand et al. (2009) addressed the question of the internal calibration of a vector (or tri-axial) magnetometer, such as the ^4He pumped magnetometer built by the Laboratoire d'électronique et de technique de l'information (LETI/CEA), or the Oersted space mission fluxgate magnetometer, built by the Institute for automation of the Danish Technical University (DTU). The problem of the internal calibration is to determine the six (^4He magnetometer) or nine (fluxgate

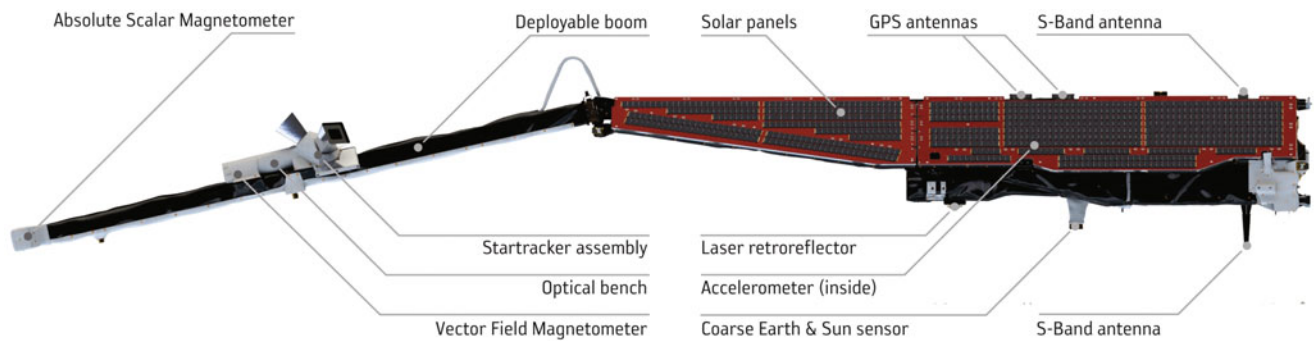
magnetometer) intrinsic parameters: the three scale values (one for each modulation coil physical axis in the ^4He case, or each fluxgate sensor axis in the fluxgate case), the three angles between the three physical axis in both cases, and the three offset values in the case of the fluxgate variometer.

The calibration problem can be solved by rotating in space the triaxial magnetometer while making simultaneously absolute intensity measurements with a proton or optical pumping magnetometer (Olsen et al. 2003), and it is essentially linear (Gravrand et al. 2009). The calibration algorithm was developed for ground operation and extended to in flight calibration. This internal calibration is not enough when installing somewhere a genuine automatic magnetic observatory. Gravrand et al. (2009) noted that an absolute calibration should not be too difficult in a place where standard absolute measurements can be performed. A quite practical interest is linked to this statement, which was addressed by Khokhlov et al. (2013), who showed that the above statement is both valid and invalid, depending on the objective, and demonstrated that only the determination of the scale factors provided by the internal calibration process are of interest for the absolute calibration. Due to a lack of space, we do not detail this algorithm here and the interested reader can refer to Khokhlov et al. (2013).

Near-Earth Magnetic Field Measurements

High accuracy and precision measurements of the near-Earth geomagnetic field from space began with NASA's MAGSAT satellite that flew for 6 months from 1979 to 1980. This was followed in the first decade of the twenty-first century by the Ørsted ([https://en.wikipedia.org/wiki/Ørsted_\(satellite\)](https://en.wikipedia.org/wiki/Ørsted_(satellite)); accessed June 2020), CHAMP (<http://www.gfz-potsdam.de/champ/>; accessed June 2020), and SAC-C magnetic field satellites. These missions have been very important for modeling and a better understanding of the geomagnetic field and its spatial and temporal variations. The huge step in measuring the magnetic field from space has arrived with the launch of the Swarm mission.

Swarm (http://www.esa.int/Our_Activities/Observing_the_Earth/Swarm; accessed June 2020), the 3-satellite ESA constellation, is the most current and most advanced geomagnetic observatory in space. All three satellites are identical and the setting of the magnetic instruments on the boom is given in Fig. 2. One of the novelties of the mission is not only the constellation concept, but also the new absolute scalar magnetometer, an optically pumped ^4He magnetometer, which provides absolute scalar measurements of the magnetic field with high accuracy and stability for the calibration of the vector field magnetometer (Leger et al. 2009). The mission was designed to derive the first global representation of the geomagnetic field variations on time-scales from an hour to several years, addressing the crucial problem of source



Geomagnetic Field, Measurement Techniques, Fig. 2 The Swarm satellite. The main magnetic instruments are installed on the 4-m boom, the absolute scalar magnetometer (at the very end) and the vector magnetometer, and the three star trackers (halfway along the boom). (Credit ESA)

separation. The three Swarm satellites are identical, and the constellation consists of two satellites (A and C) flying almost side-by-side at an altitude close to 460 km (April 2016), longitude separation of 1.4° and on circular and almost polar orbits with an inclination of 87.4° . The third satellite (B) flies higher, close to 510 km (April 2016), on a more polar orbit (inclination of 87.8°) and is progressively separated in LT with respect to A and C.

On Swarm satellites, the magnetic instruments are mounted on the boom, which accounts for almost half the length of the satellite and trails at the back. This solution is needed as the front surface is used for the electric field instrument to collect and measure the speed and direction of incident ions along the orbital path. The optical bench holding the vector magnetometer and the three star trackers is mounted halfway along the boom. The vector field magnetometer is the main instrument, providing high-precision measurements of the magnitude and direction of the magnetic vector field. This latest generation of instruments was developed and manufactured at the Technical University of Denmark. The absolute scalar magnetometer measures the strength of the magnetic field to greater accuracy than any other magnetometer. It was developed and manufactured by LETI/CEA and provides scalar measurements of the magnetic field to calibrate the vector field magnetometer.

Swarm magnetometers measure a superposition of magnetic fields produced by a number of different sources: core, lithosphere, ionosphere, and magnetosphere. Note that contributions due to external sources depend largely on latitude, LT, and solar activity, and that other satellites have the capability of making highly precise and accurate magnetic field measurements with fluxgate magnetometers within the magnetosphere or ionosphere. Examples include Cluster (<http://sci.esa.int/cluster/>; accessed June 2020), MMS (<http://mms.gsfc.nasa.gov>; accessed June 2020), and the Van Allen Probes (http://www.nasa.gov/mission_pages/rbsp/mission/; accessed June 2020). These can augment the results of the Swarm mission in the distant magnetosphere and illuminate the process of magnetic reconnection for example.

Summary

Here, we concentrate primarily, but not exclusively, on geomagnetic instruments and techniques used in a magnetic observatory settings or in a near-Earth magnetic satellites. We have briefly described the basic instruments, their mode of operation, their relative and/or absolute accuracy, and the related sources of error. It is more likely that some instrumental advances will be made towards the aim to stabilize vector magnetometers. Advances will be made in power reduction, data storage and telemetry with the goals of increased automation and cutting costs. The coming age of the automated absolute instruments makes it possible to consider once again the possibility of a true underwater observatory. The future needs of the scientific and commercial users will include ground-based and space measurements. Increasing resolution and timing accuracy will be crucial. However, to achieve these requirements will demand more efforts from instrument makers and spacecraft designers.

The Swarm mission may continue to operate until the 2030s, when the lower two satellites will re-enter the atmosphere, leaving only the higher altitude satellite. In order to be assured of high-quality geomagnetic field data in the future, it is important to develop concepts for future missions now. One possibility is the use of one or more CubeSats with scalar ^4He magnetometers, a concept under development in France. This implies new instrument developments, including the need to miniaturize the optically pumped ^4He magnetometer. Diamond vector magnetometry, using nitrogen vacancy centers in diamond, is another promising class of absolute scalar and vector magnetometers that may see use in observatories and satellite missions of the future.

Cross-References

- Core Dynamo
- Geomagnetic Field, Global Pattern
- Geomagnetic Field, IGRF
- Geomagnetic Field, Secular Variation

- [Magnetic Anomalies: Interpretation](#)
- [Magnetic Anomaly Map, Global](#)
- [Magnetic Methods, Principles](#)
- [Magnetometers](#)

Bibliography

- Allredge LR, Saldukas I (1964) An automatic standard magnetic observatory. *J Geophys Res* 69(10):1963–1970
- Auster HU, Manda M, Hemshorn A, Pulz E, Korte M (2007) Automation of absolute measurement of the geomagnetic field. *Earth Planets Space* 59:1007–1014
- Chulliat A (2019) Geomagnetic measurements. In: Manda M, Korte M, Yau A, Petrovsky E (eds) *Geomagnetism, aeronomy and space weather: a journey from the Earth's core to the sun*. Part of special publications of the International Union of Geodesy and Geophysics. Cambridge University Press, 344pp. ISBN: 9781108418485
- Csontos A, Šugar D, Brkic M, Kovács P, Hegymegi L (2012) How to control a temporary DIDD based observatory in the field? *Ann Geophys* 55(6):1085–1094. <https://doi.org/10.4401/ag-5447>
- Gravrand O, Khokhlov A, Le Mouél J-L, Leger J-M (2009) On the calibration of a vectorial 4He pumped magnetometer. *Earth Planets Space* 53:949–958
- Khokhlov A, Le Mouél J-L, Manda M (2013) Contribution to solving the orientation problem for an automatic magnetic observatory. *Geosci Instrum Method Data Syst* 2:1–9
- Leger MJ, Bertrand F, Jager T, Le Prado M, Fratter I, Lalaurie JC (2009) Swarm absolute scalar and vector magnetometer based on helium 4 optical pumping. In: *Proceedings of the Eurosensors XXIII conference*, pp 634–637. <https://doi.org/10.1016/j.proche.2009.07.158>
- Manda M, Purucker M (2018) The varying core magnetic field from a space weather perspective. *Space Sci Rev* 214:11. <https://doi.org/10.1007/s11214-017-0443-8>
- Olsen N, Toffner-Clausen L, Sabaka T, Brauer P, Merayo JM, Jorgensen JL, Leger J, Nielsen OV, Primdahl F, Risbo T (2003) Calibration of the Oersted vector magnetometer. *Earth Planets Space* 55(1):11–18
- Turbitt CW, Matzka J, Rasson J, St-Louis B, Stewart D (2013) An instrument performance and data quality standard for INTERMAGNET one-second data exchange. In: *Proceedings of the XVth IAGA workshop on geomagnetic observatory instruments, data acquisition, and processing*, pp 186–188
- van Loo SA, Rasson JL (2007) Presentation of the prototype of an automatic DI-flux. In: *XIIth IAGA workshop on geomagnetic observatory instruments, data acquisition and processing*. Publications of the Institute of Geophysics, Polish Academy of Sciences, C-99 (398)

Geomagnetic Field, Polarity Reversals

Carlo Laj
Laboratoire des Sciences du Climat, Unité mixte
CEA-CNRS-UVSQ, Gif-sur-Yvette, France

Introduction: The Discovery of Geomagnetic Reversals

Bernard Brunhes (1906) was the first to measure magnetization directions in rocks that were approximately antiparallel to the present Earth's field. Brunhes (1906) recorded

magnetizations in baked sedimentary rocks that were aligned with reverse magnetization directions in overlying Miocene lavas from central France (Puy de Dome). In so doing, Brunhes (1906) made first use of a field test for primary thermal remanent magnetization (TRM) that is now referred to as the “baked contact” test. Matuyama (1929) was the first to attribute reverse magnetizations in (volcanic) rocks from Japan and China to reversal of geomagnetic polarity, and to differentiate mainly Pleistocene lavas from mainly Pliocene lavas based on the polarity of the magnetization. In this respect, Matuyama (1929) was the first person to use the sequence of geomagnetic reversals as a means of ordering rock sequences.

The reality of geomagnetic reversals was then progressively established with the studies of Hospers (1951, 1953) in Iceland, and Roche (1950, 1951, 1956) in the Massif Central of France. The work of Hospers on Icelandic lavas was augmented by Rutten and Wensink (1960) and Wensink (1966) who subdivided Pliocene-Pleistocene lavas in Iceland into three polarity zones from young to old: N-R-N. Magnetic remanence measurements on basaltic lavas combined with K/Ar dating, pioneered by Cox et al. (1963) and McDougall and Tarling (1963a, b, 1964), resulted in the beginning of development of the modern geomagnetic polarity timescale (GPTS). These studies, and those that followed in the mid-1960s, established that rocks of the same age carry the same magnetization polarity, at least for the last few million years. The basalt sampling sites were scattered over the globe. Polarity zones were linked by their K/Ar ages, and were usually not in stratigraphic superposition. Doell and Dalrymple (1966) designated the long intervals of geomagnetic polarity of the last 5 Myrs as magnetic epochs, and named them after pioneers of geomagnetism (Brunhes, Matuyama, Gauss, and Gilbert).

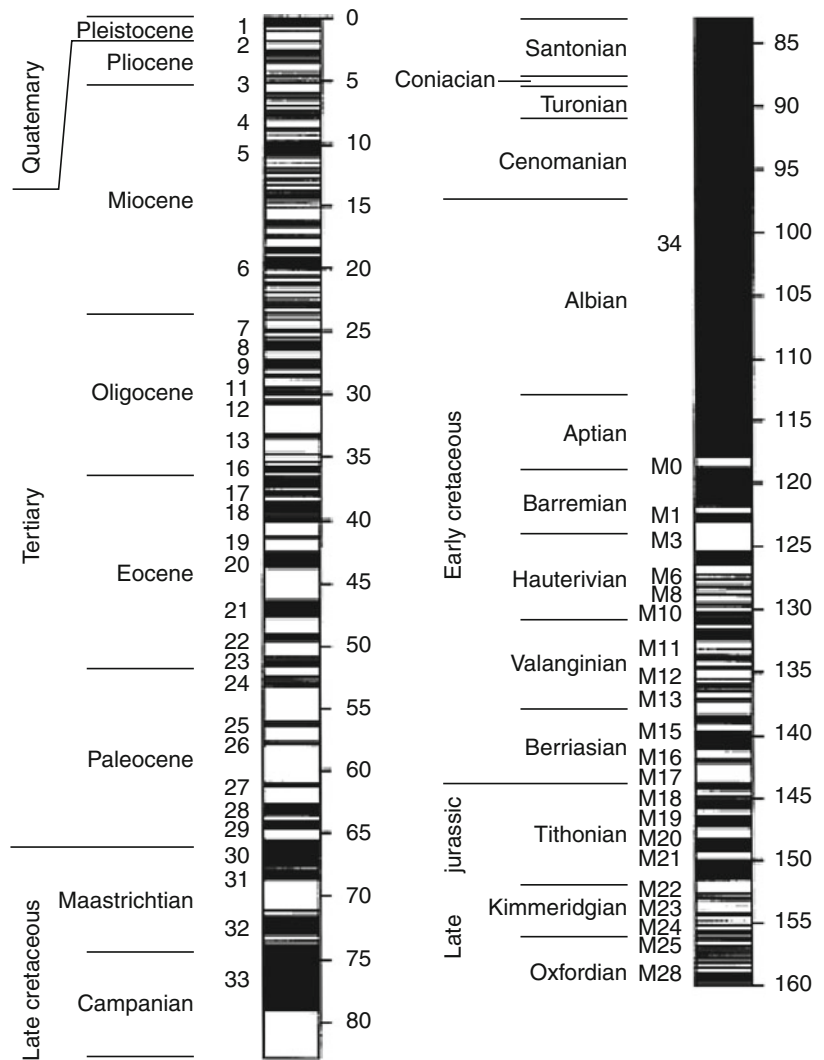
Then, the discovery of marine magnetic anomalies confirmed seafloor spreading (Vine and Matthews 1963), and the GPTS was extended to older times (Vine 1966; Heirtzler et al. 1968; Lowrie and Alvarez 1981). Since then, the succession of polarity intervals has been extensively studied and used to construct magnetostratigraphic timescales linking biostratigraphies, isotope stratigraphies, and absolute ages (see Opdyke and Channell 1996, “Magnetic stratigraphy”, for a review).

The Geomagnetic Polarity Timescale

The fit of the land-derived polarity timescale, from paleomagnetic and K/Ar studies of exposed basalts, with the polarity record emerging from marine magnetic anomalies (MMA) (Vine and Matthews 1963; Vine 1966; Pitman and Heirtzler 1966; Heirtzler et al. 1968) resulted in a convincing argument for synchronous global geomagnetic polarity reversals, thereby attributing them to the main axial dipole. This intense research effort has culminated with the work of Cande and

Geomagnetic Field, Polarity Reversals, Fig. 1

The geomagnetic polarity timescale for the last 160 My. Numbers indicate the marine magnetic anomalies



Kent (1995) based on the Marine Magnetic Anomalies record, using the South Atlantic as the fundamental template with inserts from faster spreading centers in the Indian and Pacific oceans. Cande and Kent (1995) adopted the astrochronological age estimates for the Pliocene-Pleistocene polarity reversals and fixed the age tie-point at the Cretaceous-Tertiary boundary at 65 Ma (Fig. 1).

The Changing Frequency over Time and the Duration of Geomagnetic Reversals

Examination of the geomagnetic polarity timescale reveals some of the major characteristics of geomagnetic reversals. First, a clear characteristic is the widely different duration of periods of stable polarity separating geomagnetic reversals. Long periods of time during which there were no reversals have been clearly identified, such as the Cretaceous Long Normal Chron (also called the Cretaceous Superchron,

about 118–183 million years ago) or the Kiaman Superchron, which lasted approximately from the late Carboniferous to the end of the Permian.

The frequency of reversals decreased prior to the Cretaceous Superchron, and slowly increased over the 80 million years from the end of this Superchron to the present. It has been suggested (Mazaud et al. 1983; Mazaud and Laj 1991) that in this last period the frequency does not increase randomly but presents a dominant frequency corresponding to a period of approximately 15 million years. This suggestion has, however, been received with some skepticism (McFadden 1987). The most recent analyses seem to converge toward a nonstationary Poisson process or at least a gamma process with k only slightly greater than 1 (Lowrie and Kent 2004).

Another characteristic, revealed from the very beginning by the studies of reversals, is that the Earth's magnetic field strength drops to low levels during polarity reversals. There has been some debate on whether this drop is symmetrical or

not on each side of the reversal. Valet and Meynadier (1993) suggested that the dipole intensity slowly decreases over a long period of time preceding a reversal and rapidly increases immediately after (the so-called sawtooth hypothesis). These authors also suggested that there is a positive correlation between the magnitude of the increase in intensity and the length of the subsequent polarity interval.

The sawtooth pattern, however, is not present in many sedimentary records (Tauxe and Shackleton 1994; Tauxe and Hartl 1997; Laj et al. 1996; Channell and Kleiven 2000; Channell et al. 2009). Interpretations of the sawtooth pattern, other than true geomagnetic behavior, have been given, either assuming sedimentary postdepositional artifacts (Mazaud 1996) or the presence of a cumulative viscous component not removed by alternating field demagnetization (Kok and Tauxe 1996a, b). Moreover, McFadden and Merrill (1998) have strongly questioned the suggestion that the duration of polarity intervals could be determined by the amplitude of the field recovery after the reversal. The fast recovery of the field intensity after the reversal does not appear either in a recent reconstruction of the Brunhes–Matuyama transition, based on four independent records, where the pre- and posttransitional intensity records appear symmetric about the reversal (Leonhardt and Fabian 2007). In summary, the experimental evidence for the sawtooth hypothesis is controversial.

Also clearly apparent from the geomagnetic polarity time-scale is the extremely short duration of polarity reversals compared to periods of stable polarity. The time it takes for the reversal to happen has been roughly estimated from the early works (Harrison and Somayajulu 1966; Cox et al. 1963, 1964; Cox 1969) to be between 10^3 and 10^4 years, with some estimates up to 28,000 years. In a recent article, Clement (2004) presents an analysis of available sedimentary records and shows that these records yield an average estimate of 7000 years for the duration of a reversal. Moreover, the duration varies with site latitude, with a shorter duration observed at low-latitude sites. This short duration has been one of the most delicate factors to overcome in studying the detailed structure of reversals.

The Morphology of Geomagnetic Reversals

Because of its importance to dynamo theory, the morphology of the geomagnetic field during reversals and whether or not it displays statistical asymmetries has been a subject of continuous research for many years. Records of geomagnetic field changes during a polarity transition have, however, been difficult to obtain, not only because of their short duration mentioned above, but also because the magnetizations of the rocks is weak due to the low intensity of the transitional field. It is only with the advent of sensitive magnetometers that the number of published transitional records has rapidly increased.

In probably the most famous of the early records, Hillhouse and Cox (1976) showed that the Virtual Geomagnetic Pole (VGP) (The Virtual Geomagnetic Pole is defined as the pole of the geocentered axial dipolar field that would give the observed direction of the field at the observation site. If the VGPs observed at several sites at a given time coincide, then the field has a dipolar structure.) path for the Brunhes–Matuyama transition recorded in sediments from Lake Tecopa (California) was different from the one previously obtained by Niitsuma (1971) from the Boso Peninsula in Japan, thereby providing evidence that the transitional field is not dipolar. Although in a more recent article Valet et al. (1988a) showed that the AF demagnetization used by Hillhouse and Cox (1976) did not completely remove a strong overprint, the conclusion that the transitional field was not dipolar was maintained.

Another famous record was obtained from a Miocene lava sequence at the Steens Mountain (Mankinen et al. 1985; Prévot et al. 1985). The record displays several phases with large swings, one of which occurring at the extremely rapid rate of $50 \pm 21^\circ$ per year, while the intensity change was $6700 \pm 2700 \mu\text{T}$ per year, about 10–50 times larger than the maximum rate of change observed for the non-dipole field in historical records. These amazing rates of change are still a matter of discussion among geophysicists.

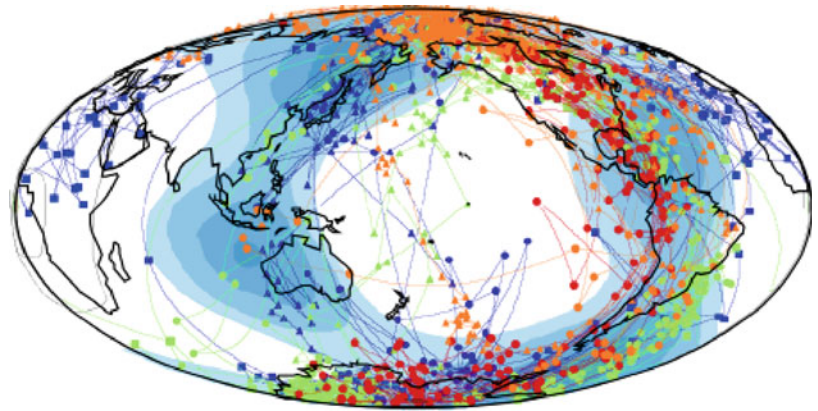
Other transitional records were obtained from lava sequences in French Polynesia (Roperch and Chauvin 1987; Roperch and Duncan 1990; Chauvin et al. 1990). These records appear to indicate that a zonal field dominates the beginning of the transition, then non-dipolar terms emerge without any preferred axi-symmetry, consistent with an earlier suggestion by Hoffman (1982).

Transitional records from sequences of relatively high accumulation rate marls in Western Crete (Greece), were reported in a series of papers by Valet and Laj (1981, 1984) and by Valet et al. (1983, 1986, 1988b). The results from multiple records indicated that the reversal process observed for four sequential transitions remained unchanged for about 1.3 Ma. In one record, obtained from a particularly high sedimentation rate section, fast directional fluctuations, reminiscent of the Steens Mountain record, were observed whose amplitude significantly increases during the transition. Slight smoothing of the data reveals that almost periodic fluctuations occurred on a longer timescale during the transition (Valet et al. 1986).

Two landmark papers were published almost simultaneously and independently in 1991. Clement (1991) examined records of the Brunhes–Matuyama transition obtained from the northern, equatorial, and southern latitudes in the Atlantic sector. The VGP paths from the mid-latitudes in the northern and southern hemispheres are nearly coincident and extend over the Americas. Paths from the equatorial site track almost antipodal to the other paths. These results suggest that during the reversal VGPs tend to fall along two longitudinal bands.

Geomagnetic Field, Polarity Reversals, Fig. 2

The compilation of sedimentary records for the last 12 My (Laj et al. 1991) indicating the existence of preferred longitudinal bands for transitional VGPs. The blue shaded zones are characterized by fast seismic velocities in the lower mantle



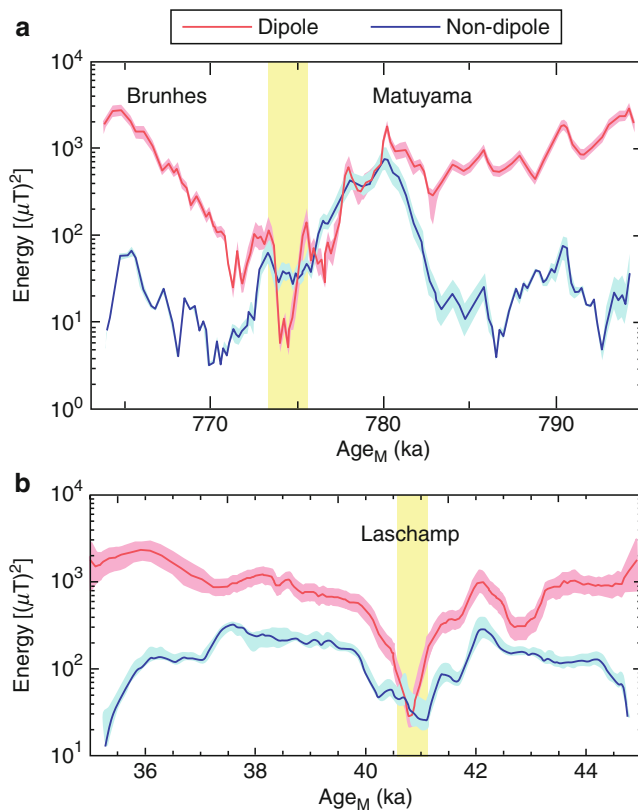
From a compilation of all the available sedimentary records from several sites and spanning about 12 million years, Laj et al. (1991) suggested that reversals and excursions exhibit geometric regularity: the VGP paths of these transitions tend to fall along either American or Asian-Australian longitudes, that is, the same as for the Brunhes-Matuyama reversal (Fig. 2). As remarked by the authors a persistent geometry over 12 Ma has important geomagnetic consequences: since the duration between transitions is longer than both the core's convective and magnetic diffusion time-scales, the most plausible mechanism by which the core could retain a memory of previous transitions, particularly preferred longitudinal bands, is through some form of core-mantle coupling. This argument was reinforced by Laj et al. (1991) who noted the apparent correlation between preferred longitudes and lower mantle seismic velocity variations, possibly arising from thermal heterogeneities.

The suggestion of preferred longitudinal bands for transitional VGPs has, however, been met with some skepticism, some authors questioning the adequacy of the geographical distribution of the sites (Valet et al. 1992), or the reliability of transitional records from sediments (Langereis et al. 1992; Barton and McFadden 1996). Statistical analyses of the distribution of path have not given a clear answer (Weeks et al. 1992; McFadden et al. 1993; Quidelleur and Valet 1994). An initial analysis of transitional records from lavas (which are devoid of many of the problems of sedimentary magnetizations) appeared to indicate that transitional VGPs are statistically axi-symmetric (Prévot and Camps 1993). However, a more recent analysis by Love (1998, 2000) using a mathematical approach in which the records are normalized taking into account differences in recording density of the paleomagnetic directions, has concluded that the volcanic transitional VGPs indeed show a significant tendency to fall along American and Asian-Australian longitudes, consistent with the sedimentary data. The existence of preferred longitudinal bands for transitional VGPs must, therefore, be considered as a realistic suggestion.

Alongside with the observations of records of geomagnetic reversals, phenomenological models have been developed to account for at least the main characteristics of the records. As early as 1976, Hillhouse and Cox (1976) suggested that if the usual non-dipole drift remained unchanged during the reversal, then one should observe large longitudinal swings in the reversal PGV paths. As this was not observed, these authors proposed that an invariant component dominates the field when the usual dipole has vanished (the so-called standing field hypothesis). Subsequently, Hoffman and Fuller (1978) and Hoffman (1979) proposed a model in which reversals originate in particular regions of the core and progressively extend to other regions (the so-called flooding model). None of these models was, however, capable of predicting VGP paths from particular sites on the Earth.

A significant step forward has been made recently, with the development by Leonhardt and Fabian (2007) of a Bayesian inversion method devised to reconstruct the spherical harmonic contents of the geomagnetic field during a reversal from paleomagnetic data. The validity of this inversion technique is proven, in particular, by iteratively combining four geographically distributed high quality paleomagnetic records of the Brunhes-Matuyama reversal into a single scenario without assuming an a priori common age model. The obtained results successfully describe most independent Brunhes-Matuyama transitional records.

Leonhardt and Fabian (2007) were able to discuss many questions about the transitional field, which had been controversially discussed (see, e.g., Merrill et al. (1996)) so far: primarily the question of contribution of dipole versus non-dipole fields during the reversal process. The results of their modeling indicates a strong increase of non-dipolar energy a few kyrs before the reversal. The non-dipole energy prevails on the dipole some 5 kyrs before the dipole minimum, after which both the non-dipole and dipolar energies decrease. Following the minimum in dipolar energy, the non-dipolar energy drops further, while the dipolar energy increases first sharply then progressively recovers (Fig. 3a). The model also



Geomagnetic Field, Polarity Reversals, Fig. 3 (a) Evolution of the dipole and non-dipole energies during the Matuyama–Brunhes polarity transition. (b) Evolution of the dipole and nondipole energies during the Laschamp excursion

shows that, during the reversal, the inverse flux penetrates the tangent cylinder (the imaginary cylinder aligned with the Earth's rotation axis and tangent to the inner core).

Dynamo Mechanisms and Reversals

For many years, scientific opinion has been divided on what causes geomagnetic reversals. Initial suggestions have been made that geomagnetic reversals are not due to spontaneous processes in the Earth's dynamo, but rather that they are triggered by external factors (such as arrival of continental slabs near the core-mantle boundary as a result of plate tectonics, or even mantle-core shear forces linked to external impacts).

Today, most geophysicists believe that reversals are inherent aspects of the dynamo, which generates the geomagnetic field. Recent developments in numerical dynamo models, which take advantage of modern supercomputers, and also an experiment with a laboratory dynamo, point strongly that this is the only sustainable mechanism. In these models the equations describing the time-dependent thermal, compositional, velocity and magnetic field are solved simultaneously

in the three-dimensional spherical geometry appropriate to the core. There is no a priori prescription for reversals, so, if they happen, they are an intrinsic property of the mechanisms of the geodynamo.

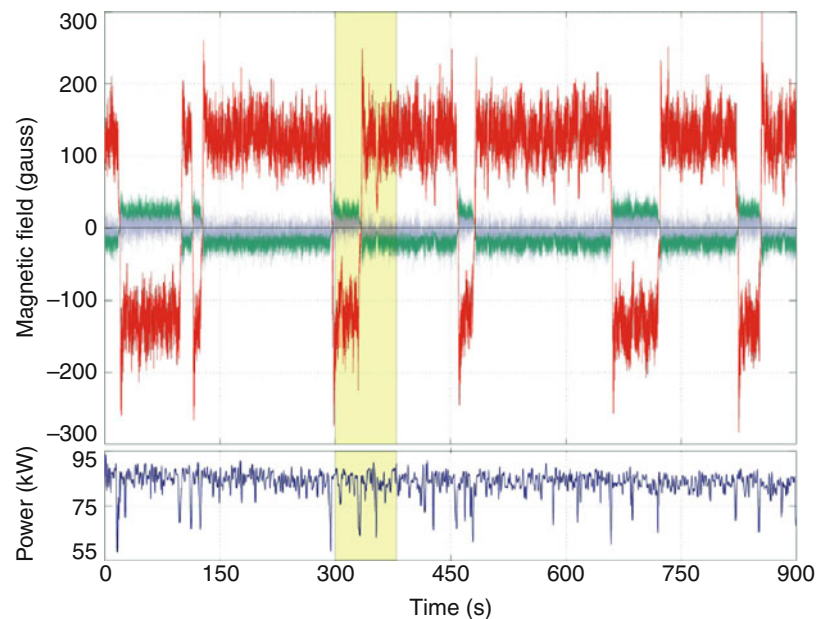
The first of these numerical models aiming to simulate Earth-like characteristics was published by Glatzmaier and Roberts (1995). This three-dimensional, selfconsistent model was able to maintain a magnetic field for over 40,000 years. The model, considers a finitely conducting solid inner core, in agreement with findings by Hollerbach and Jones (1993) that a solid inner core tends to stabilize the geodynamo. Over the period of 40,000 year the model undergoes several polarity excursions and finally a reversal is observed. The model, therefore, shares some similarities with real reversals of the geomagnetic field, and therefore may provide insight into the mechanisms of geomagnetic reversals.

In this respect, Glatzmaier et al. (1999) and Coe et al. (2000) have run numerical simulations of the Glatzmaier–Roberts dynamo using a variety of thermal boundary conditions. The model displays a range of behavior that resembles records of real reversals of the Earth's magnetic field. Reversals occurred during both the homogeneous and tomographic simulation (in this last case the heat flux at the Core-Mantle Boundary is patterned after a seismic velocity model from tomographic investigation of the lower mantle). All reversals appear spontaneously, with no external triggering required. The tomographic model of Glatzmaier et al. (1999) also appears to offer some support for the hypothesis that VGPs during reversals correlate with areas of higher than average heat flux.

The question of how well the geodynamo model represents the Earth's field is of course fundamental. Indeed, if the Earth's values for the radius and the rotation of the core are used, then, owing to computation limitations, viscous diffusivity values at least three to four orders of magnitude larger than those of the Earth must be used. So, while some modelers have argued that the field produced by models should be fairly realistic, other modelers are less confident that the results can be directly extrapolated to the Earth, even at the large scales. In the author's opinion, numerical models have very largely improved our understanding of the mechanisms inherent to the production of the Earth's magnetic field.

Finally, a benchmark experiment showing how magnetic field reversals can be generated without any external triggering in a turbulent conducting field has been obtained recently in a laboratory environment, the so-called VKS experiment (this acronym stands for Von Karman Sodium) (Berhanu et al. (2007)). In 2000 experiments in Riga (Gailitis et al. 2001) and Karlsruhe (Stieglitz and Müller 2001) showed that fluid dynamo could be generated by fluid sodium flows, when a favorable geometry is imposed to the flow. In the more recent VKS experiment, a turbulent flow is produced in a copper cylinder filled with liquid sodium whose flow is created by

Geomagnetic Field, Polarity Reversals, Fig. 4 Evolution of the magnetic field measured inside the flow vessel during the VKS experiment. The main component (red) is the azimuthal one. Note that all components decay to zero at a reversal. The bottom graph shows the evolution of the power driving the flow. (Adapted from Berhanu et al. 2007)



two counter-rotating impellers. When the speed of the impellers exceeds a critical value, the flow spontaneously generates a magnetic field that is self-sustained by the flow itself. This field undergoes large fluctuations, originating from the high level of turbulence. The most spectacular result was observed when the two impellers were counter-rotated at different speeds: in this case the temporal evolution of the field is characterized by a series of erratic reversals of polarity. These reveal remarkable similarities with the Earth's magnetic field, in the sense that the duration of the transitions is very short compared to the periods of stable polarity. In addition, excursions are also present, during which the field decays to very low values before growing again without a change in polarity (Fig. 4). Although similarity to the Earth's magnetic field should not be overstressed, this laboratory experiment clearly shows that reversals can appear spontaneously, as a consequence of the complexity of the fluid flow, without any necessity for an external triggering.

Outlook: The Future of Geomagnetic Reversals

The geomagnetic dipole has decreased in recent historical times at a rate of 5% per century, which is about 10–12 times the rate of free Ohmic decay of the geomagnetic dipole field in the core. Additional evidence for this rapid decrease is the poleward migration of large patches of reverse magnetic field at the core-mantle boundary that have largely contributed to the historical dipole field drop (Olson and Amit 2006). Therefore, this is a very rapid change for the geodynamo, which has led to speculations about the possibility of an impending field reversal (Constable and Korte (2006); Olson and Amit (2006)). Because an intensity drop is also

associated with excursions this second possibility should also be considered.

The inverse model of Leonhardt and Fabian (2007) may allow to have a deeper insight on this scenario. Indeed, when applied to the study of the Laschamp excursion (Leonhardt et al. 2009) the model indicates that both dipolar and non-dipolar energies at the Earth's surface decrease at approximately the same time with similar rates. The dipolar energy reaches its minimum slightly later than the non-dipolar energy. Then, for a brief interval the non-dipolar energy prevails, contrary to the Brunhes–Matuyama reversal discussed above where the non-dipole energy prevails on the dipole some 5 kyrs before the dipole minimum, after which both the non-dipole and dipolar energies decrease (Fig. 3b). The time evolution of the dipole versus non-dipole fields is different for a reversal or an excursion and could, therefore, be an indicator for an incoming reversal or excursion.

At present, however, the non-dipole field energy is about one order of magnitude smaller than the dipolar term. This latter is still larger than the dipole energy calculated before either a reversal or an excursion using the Leonhardt and Fabian approach. So despite the decreasing trend of dipole energy, there is no indication in the present field that would point toward either type of instability. Together with Olson and Amit (2006), Leonhardt et al. (2009) therefore consider that it is not justified to consider the present decrease in dipole energy and an indication of an imminent reversal. We will have to wait some more time to know whether we are heading toward a reversal, an excursion, or whether we are simply undergoing an intensity fluctuation.

Acknowledgments I thank Alain Mazaud for helpful discussions during the writing of the manuscript.

Bibliography

- Barton CE, McFadden PL (1996) Inclination shallowing and preferred transitional VGP paths. *Earth Planet Sci Lett* 140:147–157
- Berhanu M, Monchaux R, Fauve S, Mordant N, Pétélis F, Chiffaudel A, Daviaud F, Dubrulle B, Marié L, Ravelet F, Bourgoin M, Odier PH, Pinton J-F, Volk R (2007) Magnetic field reversals in an experimental turbulent dynamo. *Europhys Lett* 77:59001–59005
- Brunhes B (1906) Recherches sur la direction de l'aimantation des roches volcaniques. *J Phys* V:705–724
- Cande SC, Kent DV (1995) Revised calibration of the geomagnetic polarity timescale for the late cretaceous and cenozoic. *J Geophys Res* 100:6093–6095
- Channell JET, Kleiven HF (2000) Geomagnetic palaeointensities and astrochronological ages for the Matuyama-Brunhes boundary and the boundaries of the Jaramillo Subchron: palaeomagnetic and oxygen isotope records from ODP site 983. *Phil Trans R Soc A* 358:1027–1047
- Channell JET, Hodell DA, Lehman B (1997) Relative geomagnetic paleointensity and $\delta^{18}\text{O}$ at ODP site 983 (Gardar Drift, North Atlantic) since 350 ka. *Earth Planet Sci Lett* 153:103–118
- Channell JET, Xuan C, Hodell DA (2009) Stacking paleointensity and oxygen isotope data for the last 1.5 Myr (PISO-1500). *Earth Planet Sci Lett* 283:14–23
- Chauvin A, Duncan RA, Bonhommet N, Levi S (1989) Paleointensity of the Earth's magnetic field and K-Ar dating of the Louchadière volcanic flow (central France): new evidence for the Laschamp excursion. *Geophys Res Lett* 16:1189–1192
- Chauvin A, Roperch P, Duncan RA (1990) Records of geomagnetic reversals from volcanic islands of French polynesia, 2-paleomagnetic study of a flow sequence (1.2 to 0.6 Ma) from the Island of Tahiti and discussion of reversal models. *J Geophys Res* 95:2727–2752
- Clement BM (1991) Geographical distribution of transitional V. G.P.'s: evidence for non-zonal equatorial symmetry during the Matuyama-Brunhes geomagnetic reversal. *Earth Planet Sci Lett* 104:48–58
- Clement BM (2004) Dependence of the duration of geomagnetic polarity reversals on site latitude. *Nature* 428:637–640
- Coe RS, Hongre L, Glatzmaier GA (2000) An examination of simulated geomagnetic reversals from a paleomagnetic perspective. *Phil Trans R Soc A* 358:1141–1170
- Constable CG, Korte M (2006) Is the Earth's magnetic field reversing? *Earth Planet Sci Lett* 246:1–16
- Cox A (1969) Geomagnetic reversals. *Science* 163:237–245
- Cox A, Doell RR, Dalrymple GB (1963) Geomagnetic polarity epochs and Pleistocene geochronometry. *Nature* 198:1049–1051
- Cox A, Doell RR, Dalrymple GB (1964) Reversals of the Earth's magnetic field. *Science* 144:1537–1543
- Doell RR, Dalrymple GB (1966) Geomagnetic polarity epochs: a new polarity event and the age of the Brunhes-Matuyama boundary. *Science* 152:1060–1061
- Gailitis A, Lielausis O, Platācis E, Dement'ev S, Cifersons A, Gerbeth G, Gundrum T, Stefani F, Christen M, Will G (2001) Magnetic field saturation in the Riga dynamo experiment. *Phys Rev Lett* 86:3024–3027
- Glatzmaier GA, Roberts PH (1995) A three-dimensional self-consistent computer simulation of a geomagnetic field reversal. *Nature* 377:203–209
- Glatzmaier GA, Coe RS, Hongre L, Roberts PH (1999) The role of the Earth's mantle in controlling the frequency of geomagnetic reversals. *Nature* 401:885–890
- Gubbins D (1999) The distinction between geomagnetic excursions and reversals. *Geophys J Int* 137:F1–F3
- Harrison CGA, Somayajulu BLK (1966) Behaviour of the earth's magnetic field during a reversal. *Nature* 212:1193–1195
- Heirtzler JR, Dickson GO, Herron EM, Pittman WC III, LePichon X (1968) Marine magnetic anomalies, geomagnetic field reversal and motions of the ocean floor and continents. *J Geophys Res* 73:2119–2136
- Hillhouse J, Cox A (1976) Brunhes-Matuyama polarity transition. *Earth Planet Sci Lett* 29:51–64
- Hoffman KA (1979) Behaviour of the geodynamo during a reversal: a phenomenological model. *Earth Planet Sci Lett* 44:7–17
- Hoffman KA (1982) The testing of geomagnetic reversal models/recent developments. *Philos Trans R Soc Lond A* 306:147–159
- Hoffman KA, Fuller M (1978) Polarity transition records and the geomagnetic dynamo. *Nature* 273:715–718
- Hollerbach R, Jones CA (1993) Influence of the Earth's inner core on geomagnetic fluctuations and reversals. *Nature* 365:541–543
- Hospers J (1951) Remanent magnetization of rocks and the history of the geomagnetic field. *Nature* 168:1111–1112
- Hospers J (1953–1954) Reversals of the main geomagnetic field I, II, and III. *Proc K Ned Akad Wet C* 56:467–491; 57:112–121
- Kok YS, Tauxe L (1996a) Saw-toothed pattern of relative paleointensity records and cumulative viscous remanence. *Earth Planet Sci Lett* 137:95–100
- Kok YS, Tauxe L (1996b) Saw-toothed pattern of sedimentary paleointensity records explained by cumulative viscous remanence. *Earth Planet Sci Lett* 144:9–14
- Laj C, Mazaud A, Fuller M, Herrero-Bervera E (1991) Geomagnetic reversal paths. *Nature* 351:447
- Laj C, Kissel C, Garnier F, Herrero-Bervera E (1996) Relative geomagnetic field intensity and reversals for the last 1.8 My from a central Pacific core. *Geophys Res Lett* 23:3393–3396
- Langereis CG, Van Hoof AAM, Rochette P (1992) Longitudinal confinement of geomagnetic reversal paths: as a possible sedimentary artefact. *Nature* 358:226–230
- Leonhardt R, Fabian K (2007) Paleomagnetic reconstruction of the global geomagnetic field evolution during the Matuyama/Brunhes transition: iterative Bayesian inversion and independent verification. *Earth Planet Sci Lett* 253:172–195
- Leonhardt R, Fabian K, Winklhofer M, Ferk A, Laj C, Kissel C (2009) Geomagnetic field evolution during the Laschamp excursion. *Earth Planet Sci Lett* 278:87–95
- Love JJ (1998) Paleomagnetic volcanic data and geometric regularity of reversals and excursions. *J Geophys Res* 103:12435–12452
- Love JJ (2000) Statistical assessment of preferred transitional VGP longitudes based on paleomagnetic lava data. *Geophys J Int* 140:211–221
- Lowrie W, Alvarez W (1981) One hundred million years of geomagnetic polarity history. *Geology* 9:392–397
- Lowrie W, Kent DV (2004) Geomagnetic polarity timescales and reversal frequency regimes. In: Channell JET et al (eds) *Timescales of the geomagnetic field*. AGU geophysical monograph series, vol 145, pp 117–129
- Mankinen EA, Prévot M, Grommé CS, Coe RS (1985) The Steens Mountain (Oregon) geomagnetic polarity transition. 1. Directional history, duration of the episodes, and rock magnetism. *J Geophys Res* 90:393–416
- Matuyama M (1929) On the direction of magnetization of basalts in Japan, Tyosen and Manchuria. *Proc Imperial Acad* 5:203–205. (Tokyo)
- Mazaud A (1996) Sawtooth variation in magnetic intensity profiles and delayed acquisition of magnetization in deep sea cores. *Earth Planet Sci Lett* 139:379–386
- Mazaud A, Laj C (1991) The 15 m.y. geomagnetic reversal periodicity: a quantitative test. *Earth Planet Sci Lett* 107:689–696
- Mazaud A, Laj C, De Seze L, Verosub KL (1983) 15-Myr periodicity in the frequency of geomagnetic reversals since 100 Myr. *Nature* 304:328–330

- McDougall I, Tarling DH (1963a) Dating of reversals of the Earth's magnetic field. *Nature* 198:1012–1013
- McDougall I, Tarling DH (1963b) Dating of polarity zones in the Hawaiian Islands. *Nature* 200:54–56
- McDougall I, Tarling DH (1964) Dating geomagnetic polarity zones. *Nature* 202:171–172
- McFadden PL (1987) Comment on “a periodicity of magnetic reversals?”. *Nature* 330:27
- McFadden PL, Merrill RT (1998) Sawtooth paleointensity and reversals of the geomagnetic field. *Phys Earth Planet Inter* 103:247–252
- McFadden PL, Barton CLE, Merrill RT (1993) Do virtual geomagnetic poles follow preferential paths during geomagnetic reversals? *Nature* 361:342–344
- Merrill RT, McElhinny M, McFadden PL (1996) The magnetic field of the earth, international geophysics series, vol 63. Academic, San Diego, p 531
- Meynadier L, Valet J-P, Weeks R, Shackleton NJ, Hagee VL (1992) Relative geomagnetic intensity of the field during the last 140 ka. *Earth Planet Sci Lett* 114:39–57
- Meynadier L, Valet J-P, Shackleton NJ (1995) Relative geomagnetic intensity during the last 4 MY from the equatorial Pacific. *Proc ODP Sci Res* 138:779–793
- Niituma N (1971) Detailed studies of the sediments recording the Matuyama-Brunhes geomagnetic reversal. *Tohoku Univ Sci Rpt 2nd Ser (Geology)* 43:1–39
- Olson P, Amit H (2006) Changes in the Earth's dipole. *Naturwissenschaften* 93:519–542
- Opdyke ND, Channell JET (1996) Magnetic stratigraphy. Academic, San Diego, p 346
- Pitman WC III, Heirtzler JR (1966) Magnetic anomalies over the Pacific-Antarctic Ridge. *Science* 154:1164–1171
- Prévot M, Camps P (1993) Absence of preferred longitudinal sectors for poles from volcanic records of geomagnetic reversal. *Nature* 366:53–57
- Prévot M, Mankinen E, Coe RS, Grommé CS (1985) The steens mountain (Oregon) geomagnetic polarity transition. 2. Field intensity variations and discussion of reversal models. *J Geophys Res* 90:10417–10448
- Quidelleur X, Valet J-P (1994) Paleomagnetic records of excursions and reversals: possible biases caused by magnetization artefacts. *Phys Earth Planet Inter* 82:27–48
- Roche A (1950) Sur les caracteres magnétiques du système éruptif de Gergovie. *C R Acad Sci* 230:113–115
- Roche A (1951) Sur les inversions de l'aimantation remanente des roches volcaniques dans les monts d'Auvergne. *C R Acad Sci* 223:1132–1134. Paris
- Roche A (1956) Sur la date de la dernière inversion du champ magnétique terrestre. *C R Acad Sci* 243:812–814. Paris
- Roperch P, Chauvin A (1987) Transitional geomagnetic field behavior: volcanic records from French Polynesia. *Geophys Res Lett* 14:151–154
- Roperch P, Duncan RA (1990) Records of geomagnetic reversals from volcanic islands from French polynesia I. Paleo-magnetic study of a polarity transition on a lava sequence from the island of Huahine. *J Geophys Res* 95(B3):2713–2726
- Rutten MG, Wensink H (1960) Paleomagnetic dating glaciations and chronology of the Plio-Pleistocene in Iceland. *Int Geol Congress* IV:62–70. Sess 21
- Stieglitz R, Müller R (2001) Experimental demonstration of a homogeneous two-scale dynamo. *Phys Fluids* 13:561. <https://doi.org/10.1063/1.1331315>
- Tauxe L, Hartl P (1997) 11 million years of oligocene-geomagnetic field behavior. *Geophys J Int* 128:217–229
- Tauxe L, Shackleton NJ (1994) Relative paleointensity record from the Ontong-Java plateau. *Geophys J Int* 117:769–782
- Valet J-P, Herrero-Bervera E (2003) Some characteristics of geomagnetic reversals inferred from detailed volcanic records. *C R Geosci* 335:79–90
- Valet J-P, Laj C (1981) Paleomagnetic record of two successive Miocene geomagnetic reversals in Western Crete. *Earth Planet Sci Lett* 54:53–63
- Valet J-P, Laj C (1984) Invariant and changing transitional field configurations in a sequence of geomagnetic reversals. *Nature* 311:552–555
- Valet J-P, Meynadier L (1993) Geomagnetic field intensity and reversals during the last four million years. *Nature* 366:234–238
- Valet J-P, Laj C, Langereis CG (1983) Two different R-N geomagnetic reversals with identical VGP paths recorded at the same site. *Nature* 304:330–332
- Valet J-P, Laj C, Tucholka P (1986) High resolution sedimentary record of a geomagnetic reversal. *Nature* 322:27–32
- Valet J-P, Laj C, Langereis CG (1988a) Sequential geomagnetic reversals recorded in upper tortonian marine clays in Western Crete (Greece). *J Geophys Res* 93:1131–1151
- Valet J-P, Tauxe L, Clark DR (1988b) The Matuyama-Brunhes transition recorded from Lake Tecopa sediments 5 (California). *Earth Planet Sci Lett* 87:463–472
- Valet JP, Tucholka P, Courtillot V, Meynadier L (1992) Paleomagnetic constraints on the geometry of the geomagnetic field during reversals. *Nature* 356:400–407
- Vine FJ (1966) Spreading of the ocean floor: new evidence. *Science* 154:1405–1415
- Vine FJ, Matthews DH (1963) Magnetic anomalies over oceanic ridges. *Nature* 199:947–949
- Weeks R, Fuller M, Laj C, Mazaud A, Herrero-Bervera E (1992) Sedimentary records of reversal transitions – magnetization smoothing artefact or geomagnetic field behaviour? *Geophys Res Lett* 19:2007–2010
- Wensink H (1966) Paleomagnetic stratigraphy of younger basalts and intercalated Plio-Pleistocene tillites in Iceland. *Geol Rundsch* 54:364–384

Geomagnetic Field, Secular Variation

Monika Korte

Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany

Definition

The geomagnetic field observed at and above the Earth's surface varies on a broad range of timescales, from less than seconds to millions of years. In general, fast variations with periods less than a year originate from sources external to the Earth. Secular variation is the continuous variation of direction and intensity of the geomagnetic core field, associated to changes of fluid flow in the outer core (see, e.g., Jackson and Finlay 2015 for a detailed review on secular variation). The name originates from the Latin “seculum,” referring to a time interval of about a century. Secular variation describes the velocity of field change or the first temporal derivative of a time series of field observations. The second derivative, or

rate of change of secular variation, is called secular acceleration.

Secular variation consists of mostly nonperiodic, non-predictable field changes. The high-frequency end of the temporal secular variation spectrum is not well-understood. Very rapid variations that might occur in the core are damped and filtered out when passing through the mantle before being observed. The fastest measured secular variation features are named geomagnetic jerks and rapid fluctuations. They are caused by rapid secular acceleration changes and occur on timescales on the order of 1 year (Mandea and Olsen 2009). It is challenging to discriminate between internal and external variations in observational data on these timescales. Long-term external field changes, linked in particular to the approximately 11-year sunspot cycle, are also reflected in any measured time series.

On several thousands to million year timescales, the usually moderate secular variation is interrupted by excursions and complete polarity reversals of the geomagnetic core field (see ► “[Geomagnetic Excursions](#)” and ► “[Geomagnetic Field, Polarity Reversals](#)”). It is unclear whether these should be seen as part of the secular variation spectrum, or whether they have a distinctly different physical mechanism. The maximum possible rates of secular variation are recently debated since rapid high intensity “spikes” have been found in archaeomagnetic data (Shaar et al. 2011; Livermore et al. 2014, Korte and Constable 2018).

Geomagnetic secular variation is not globally uniform. The geomagnetic dipole moment, a measure of the overall global field strength, has been decreasing at a rate of 6% per century since the beginning of systematic magnetic field intensity observations around 1840. The present spatial distribution of intensity secular variation is characterized by large areas of intensity decrease over North America, large parts of the Atlantic and the southern Pacific, and smaller areas of intensity increase over the Indian Ocean, northern Africa, Europe, and Asia (Thébault et al. 2015). Present, historical and paleomagnetic data indicate that secular variation has always been weak in the Pacific compared to the Atlantic hemisphere, which might be a consequence of lateral variations in lower mantle temperature (see, e.g., Gubbins and Gibbons 2004). Higher secular variation in the southern than northern hemisphere characterizes at least the Holocene geomagnetic field evolution (Constable et al. 2016).

Cross-References

- [Core Dynamo](#)
- [Geomagnetic Excursions](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Geomagnetic Field, Theory](#)

Bibliography

- Constable C, Korte M, Panovska S (2016) Persistent high paleosecular variation activity in southern hemisphere for at least 10 000 years. *Earth Planet Sci Lett* 453:78086
- Gubbins D, Gibbons S (2004) Low Pacific secular variation. In: Channell JET, Kent DV, Lowrie W, Meert JG (eds) *Timescales of the geomagnetic field*. Geophysical monograph series 145. American Geophysical Union, Washington, DC, pp 279–286
- Jackson A, Finlay C (2015) Geomagnetic secular variation and its application to the core. In: Schubert G (ed) *Treatise on geophysics*, vol 5, 2nd edn. Elsevier, Amsterdam, pp 137–184
- Korte M, Constable CG (2018) Archeomagnetic intensity spikes: global or regional geomagnetic field features? *Front Earth Sci* 6:17. <https://doi.org/10.3389/feart.2018.00017>
- Livermore PW, Fournier A, Gallet Y (2014) Core-flow constraints on extreme archeomagnetic intensity changes. *Earth Planet Sci Lett* 387:145–156
- Mandea M, Olsen N (2009) Geomagnetic and archeomagnetic jerks: where do we stand? *EOS Trans AGU* 90:24
- Thébault E, Finlay C, Beggan C, Alken P, Aubert J, Barrois O, Bertrand F, Bondar T, Boness A, Brocco L, Canet E, Chambodut A, Chuillat A, Coisson P, Civet F, Du A, Fournier A, Fratter I, Gillet N, Hamilton B, Hamoudi M, Hulot G, Jager T, Korte M, Kuang W, Lalanne X, Langlais B, L  ger JM, Lesur V, Lowes FJ, Macmillan S, Mandea M, Manoj C, Maus S, Olsen N, Petrov V, Rother M, Sabaka TJ, Saturnino D, Schachtschneider R, Sirol O, Tangborn A, Taylor V, Thomson A, T  ffner-Clausen L, Vigneron P, Wardinski I, Zvereva T (2015) International geomagnetic reference field: the 12th generation. *Earth Planets Space* 67:79. <https://doi.org/10.1186/s40623-015-0228-9>

Geomagnetic Field, Theory

Friedrich H. Busse

Institute of Physics, University of Bayreuth, Bayreuth, Germany

Definition

Motions of an electrically conducting fluid without a magnetic field may become unstable and develop a magnetic field if the magnetic Reynolds number is sufficiently high. This process is called a dynamo. The magnetic Reynolds number is proportional to the electrical conductivity of the fluid, the amplitude of motion, and the typical scale of motion. The simplest example of a dynamo process is the disk dynamo first considered by Faraday. The exact nature of the motions driving the geodynamo in the liquid iron core of the Earth is still a subject of debate.

Introduction

The forces exerted by magnetized rocks or lodestones have been known to mankind for many centuries before Christ, but

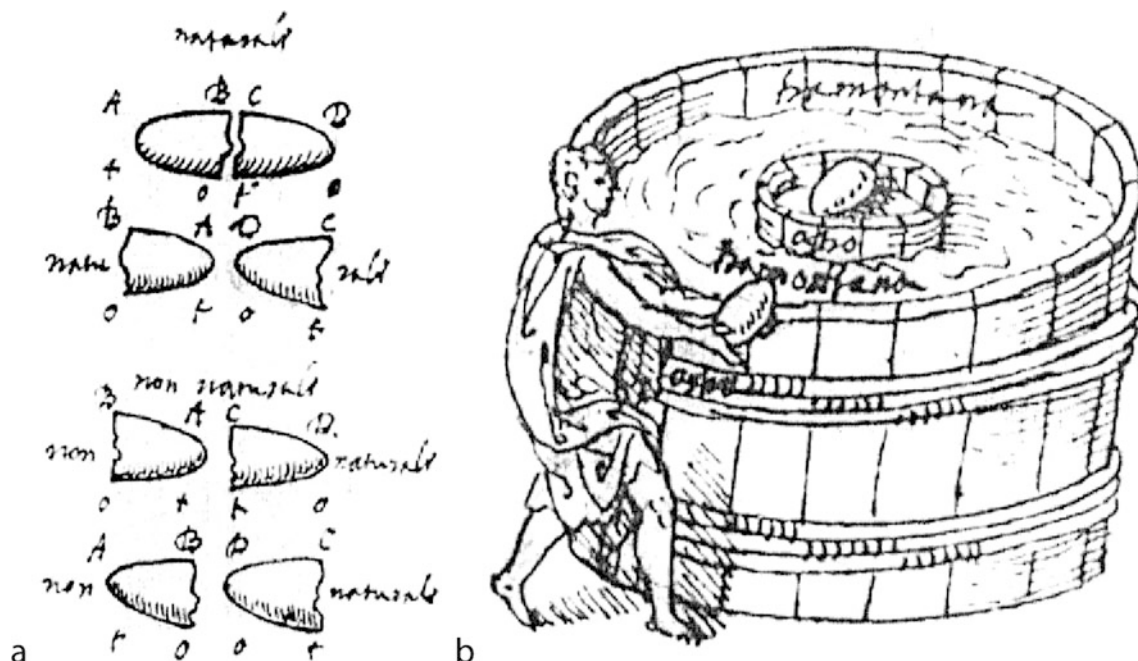
the directivity of the lodestone appears to have been first observed by the Chinese in the first century B.C. It is known that at this time Chinese geomancers used a freely turnable spoon-like lodestone on a polished plate as a compass, not for navigational purposes, but probably for aligning tombs and houses. As navigational instrument the compass was introduced to European sailors by Arabic merchants in the twelfth century A.C. A first systematic experimental study of the forces exerted by lodestones was published by Petrus Peregrinus in 1269. Some of his results are shown in Fig. 1a. More refined experiments were carried out and described in his book *De Magnete* by William Gilbert (1600). He concludes that the Earth itself is a magnet like a lodestone.

In 1839 Carl-Friedrich Gauss introduced the representation in terms of spherical harmonics for the observed geomagnetic field. Through this representation, it became possible to separate the magnetic fields of internal and external origin. The latter, being much smaller than the former and much more variable in time, is caused by ionospheric and magnetospheric currents. Only during magnetic storms can the externally generated magnetic field reach a significant fraction of the total strength of the Earth's magnetic field (see ► “Electromagnetic Pulsations and Magnetic Storms”). Together with Alexander von Humboldt Gauss founded in 1836 the Göttinger Magnetischer Verein (Göttingen Magnetic Union) in order to create a worldwide network of observatories for the measurement of the magnetic field and its

variations in time. Since that time the knowledge about the geomagnetic field and its time dependence has improved continuously (see ► “Geomagnetic Field, Global Pattern”).

The idea that the geomagnetic field originates from the ferromagnetism of the Earth's interior prevailed until the end of the nineteenth century even though the phenomenon of secular variation – known since the seventeenth century – was difficult to reconcile with the hypothesis of remanent magnetism. In early years of the twentieth century, it became clear that the Curie temperature of ferromagnetic material is exceeded within the entire Earth except for the uppermost few tens of kilometers and that the magnetization of the Earth's crust is far too weak to act as a source of geomagnetism. Numerous other hypotheses have been proposed as origin of geomagnetism and rejected (Rikitake 1966) except for the dynamo hypothesis. In the meantime, magnetic fields had been observed on the sun and in 1919 Larmor proposed the dynamo mechanism for the origin of the magnetic field of sunspots.

According to the dynamo hypothesis motions within the outer liquid-iron core of the Earth can amplify a magnetic field from arbitrarily small amplitudes and sustain it at the observed finite strength against Ohmic dissipation. The feasibility of this mechanism in a simply connected region of nearly uniform electrical conductivity appeared to be doubtful for a long time, and the dynamo hypothesis experienced a setback when Cowling (1934) demonstrated that axisymmetric magnetic fields cannot be generated by the dynamo



Geomagnetic Field, Theory, Fig. 1 Lodestone experiments of Petrus Peregrinus. (a) Experiments showing the attracting and repelling forces of a broken piece of lodestone. Upper configuration: “naturale,” lower

configuration: “nonnaturale.” (b) A simple compass in the form of an afloat lodestone. (After Stefani et al. 2008)

mechanism. Since the geomagnetic field is axisymmetric in first approximation, Cowling's theorem stimulated the search for alternative mechanisms. It was not until the late 1950s that Backus (1958) and Herzenberg (1958) demonstrated independently that dynamos are indeed possible in spheres of uniform conductivity.

Since that time dynamo theory has developed rapidly and it has become generally accepted that a geodynamo operates in the liquid outer core of the Earth. While in the 1960s and 1970s of the past century analytical models of the geodynamo have prevailed, the strong growth of computer capacities in recent decades has permitted numerical simulations with ever increasing details of convection driven dynamos in Earth's core-like rotating spherical shells. The principles of the dynamo process in such an environment are now reasonably well understood. The remaining uncertainties are caused by an insufficient knowledge of conditions in the Earth's core and lower mantle and the inability to model the dynamo process under the highly turbulent conditions of flow in the outer core.

Before outlining the concepts of dynamo theory, we briefly mention the role of the crustal magnetic field which originates from ferromagnetic remanence of material below its Curie temperature in the lithosphere. When the observed magnetic field is represented in terms of spherical harmonics (see ► [“Geomagnetic Field, IGRF”](#)), it is found that crustal contributions dominate in coefficients of the degree 14 and higher. The crustal contributions to coefficients of lower degrees are still substantial, of course, but they do not exceed those originating from the core in general (see ► [“Magnetic Anomaly Map, Global”](#)). Crustal magnetic fields are often correlated with geological features and may indicate iron ore deposits such as the famous Kursk formation in the Ukraine. Other important crustal magnetic features are the magnetic stripes of the ocean floors extending parallel to the mid-oceanic ridges which represent the most convincing evidence for the concept of plate tectonics (see ► [“Magnetic Anomalies: Interpretation”](#)).

The following sections are devoted to the dynamo theory of geomagnetism and its applications to the Earth's core. We shall start with the simplest dynamo device that converts mechanical energy into magnetic energy and then proceed to the concepts of the geodynamo. The current state of numerical models of the geodynamo will be described in another entry under the heading ► [“Core Dynamo”](#).

Kinematic Dynamo Theory

The simplest description of a dynamo is obtained for a fixed velocity field, that is, when the feedback of the generated magnetic field on the flow through the action of the Lorentz force is neglected. A single equation of magnetic induction

describing the time dependence of the magnetic flux density $\vec{B}$ can be derived from Maxwell's equations in the magneto-hydrodynamic approximation, in which the displacement current is neglected, and from Ohm's law which are given by

$$\begin{aligned} \nabla \cdot \vec{B} &= 0, & \frac{\partial \vec{B}}{\partial t} &= \nabla \times \vec{E}, & \nabla \times \vec{B} &= \mu \vec{j}, \\ \vec{j} &= \sigma (\vec{E} + \vec{v} \times \vec{B}) \end{aligned} \quad (1)$$

where $\vec{B}$ is the magnetic flux density, $\vec{E}$ is the electric field and $\vec{j}$ is the current density. σ , μ and $\vec{v}$ denote the electrical conductivity, the magnetic permeability and the velocity vector, respectively. Ohm's law has been written in the form valid for a moving conductor. Accordingly Eqs. 1 are invariant with respect to a Galilei transformation; in other words, the same Eqs. 1 hold in a system moving with constant velocity $\vec{V}$ with respect to the original system of reference when $\vec{v}$, $\vec{E}$ and $\vec{B}$ obey the relationships

$$\vec{v}' = \vec{v} - \vec{V}, \quad \vec{B}' = \vec{B}, \quad \vec{E}' = \vec{E} + \vec{V} \times \vec{B} \quad (2)$$

where the primed quantities refer to the new system. It is convenient to eliminate $\vec{E}$ and $\vec{j}$ from Eqs. 1 and to write

$$\frac{\partial \vec{B}}{\partial t} = \nabla \times (\lambda \nabla \times \vec{B}) = \nabla \times (\vec{v} \times \vec{B}) \quad (3)$$

where $\lambda \equiv 1/(\mu\sigma)$ is the magnetic diffusivity. Because of the absence of ferromagnetic effects in the Earth's core, the permeability is usually replaced by the permeability of the vacuum. From the form of Eq. 3 it is obvious that $\nabla \cdot \vec{B} = 0$ remains satisfied if this equation is satisfied by the initial conditions. It is remarkable that Eq. 3 is not only invariant with respect to Galileo transformations and thus compatible with the Navier–Stokes equations of motion, but is also invariant with respect to a transformation to a rotating system.

Since the Earth's outer core can be regarded as an incompressible fluid in first approximation, Eq. 3 can be simplified further by the use of equations $\nabla \cdot \vec{v} = 0$, $\nabla \cdot \vec{B} = 0$,

$$\left(\frac{\partial}{\partial t} + \vec{v} \cdot \nabla \right) \vec{B} - \lambda \nabla^2 \vec{B} = \vec{B} \cdot \nabla \vec{v}, \quad (4)$$

where for simplicity $\lambda = \text{const.}$ has been assumed. This form of the equation of induction is revealing since it resembles the heat equation with the term $\vec{B} \cdot \nabla \vec{v}$ on the right hand side assuming the role of a heat source. This term tends to strengthen the magnetic field when a fluid particle is stretched in the direction of the local magnetic field line. Such an interpretation is especially useful for the dynamo problem.

In order that a magnetic field $\vec{B}$ may grow, the term on the right hand side of Eq. 4 must overcome the effect of the magnetic diffusion term on the left hand side. Using a typical velocity U and a typical length scale d , the ratio of the two terms can be described by the magnetic Reynolds number Rm ,

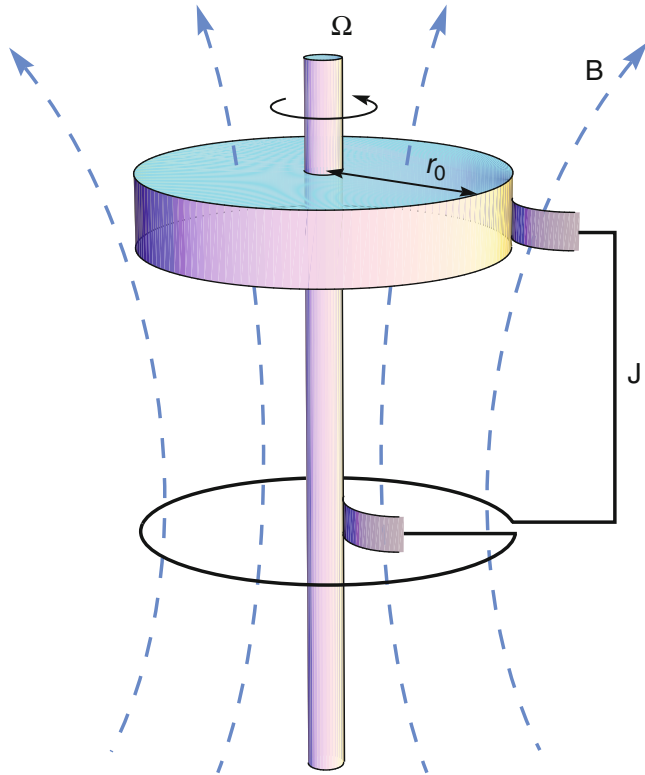
$$Rm \equiv Ud/\lambda. \quad (5)$$

Only when Rm is of the order of 1 or larger may growing magnetic fields become possible. Necessary conditions of this kind for a dynamo have been derived rigorously by Backus and Childress for a fluid sphere of constant magnetic diffusivity,

$$Rm \equiv Ur_o/\lambda > \pi, \quad (6)$$

where r_o is the radius of the sphere. In Childress's formulation U refers to the maximum velocity, whereas in Backus's formulation $U/\pi r_o$ is the maximum rate of strain of the velocity field.

The simplest dynamo is the disk dynamo originally invented by Faraday and shown in Fig. 2. A metal disk is rotating about its axis with the angular velocity Ω . When the disk is permeated by an initial magnetic field $\vec{B}_0$ parallel to the axis, an electromotive force U_e is generated between axis and rim of the disk,



Geomagnetic Field, Theory, Fig. 2 The disk dynamo

$$U_e = \Omega \int_{r_1}^{r_0} \vec{B}_0 r dr \quad (7)$$

where r_1 and r_0 denote the radii of axis and disk. The electromotive force U_e can be used to drive an electric current J through the circuit indicated in the figure. Denoting by L and R the inductivity and the ohmic resistance of the circuit we obtain for the current J the equation

$$L \frac{dJ}{dt} + RJ = U_e \quad (8)$$

The current J flowing through the circuit is associated with a magnetic field $\vec{B}_1$ which may replace the initial field $\vec{B}_0$. The integral $2\pi \int_{r_1}^{r_0} \vec{B}_1 dr/J$ describes the mutual inductivity M between circuit and disk. Equation 8 for the self-excited disk dynamo can thus be written in the form

$$L \frac{dJ}{dt} = (\Omega M/2\pi - R)J \quad (9)$$

which allows for exponentially growing solutions once the dynamo condition

$$\Omega > 2\pi R/M \quad (10)$$

is satisfied. The exponential growth will be modified, of course, as soon as the Lorentz force $\vec{j} \times \vec{B}$ in the disk has grown sufficiently to affect the angular velocity Ω . A larger torque will then be needed to sustain the rotation speed.

The disk dynamo is an inhomogeneous dynamo since it depends on an inhomogeneous, multiply connected distribution of electrical conductivity as given by the wiring of the circuit. The dynamo would not work if the sense of wiring around the axis would be opposite to that shown in the figure or if it would be immersed in a highly conducting fluid. In this respect, the disk dynamo differs from the homogeneous dynamo operating in the Earth's core in a singly connected finite domain of essentially homogeneous electrical conductivity. Velocity fields required for driving a homogeneous dynamo are necessarily more complex than the simple rotation velocity of the disk dynamo. It can even be proven that a simple form of motion, such as that of the disk dynamo, cannot drive a dynamo in a sphere if λ is a function of only the radial coordinate. More generally, the toroidal theorem states that growing solutions of Eq. 3 do not exist if $\vec{v}$ and λ can be written in the form

$$\vec{v} = \nabla \Psi \times \vec{r}, \quad \lambda = \lambda(|\vec{r}|). \quad (11)$$

In spite of their differences homogeneous dynamos and the disk dynamo share a number of properties:

- (i) For given external parameters there exists always a solution without magnetic field besides the dynamo solution, just as the disk of Fig. 2 can rotate with $\Omega > \Omega_c$ in the absence of any initial field $\vec{B}_0$.
- (ii) The existence of a dynamo solution requires that the magnetic Reynolds number Rm exceeds a critical value Rm_c which corresponds to the quantity $\Omega M/R$ in the case of the disk dynamo. The nonmagnetic solution exists, but it is unstable for $Rm > Rm_c$.
- (iii) Magnetohydrodynamic dynamos always occur in two forms which are identical except for the sign of the magnetic field $\vec{B}$. This property reflects the fact that the Lorentz force is given by an expression that is quadratic in $\vec{B}$.

Property (iii) is the basic reason that the geomagnetic field has often switched its polarity in the geologic past. These “reversals” have occurred nearly randomly about every 200,000 years in the time average. In contrast, the solar magnetic field reverses every 11 years in a surprisingly periodic fashion.

For the description of magnetic fields associated with a spherical system one often uses the general representation for a solenoidal vector field,

$$\vec{B} = \nabla \times (\nabla h \times \vec{r}) + \nabla g \times \vec{r} \quad (12)$$

in terms of a poloidal and a toroidal component each of which is described by a scalar function, the poloidal function h , and the toroidal function g . Without loss of generality, the averages of h and g over surfaces $|r| = \text{const.}$ can be assumed to vanish. A homogeneous dynamo usually requires the interaction of both components. The functions h and g can be separated into their axisymmetric parts $\bar{h}$ and $\bar{g}$ and non-axisymmetric parts, $h = h - \bar{h}$ and $g = g - \bar{g}$. Because of Cowling’s theorem the latter parts play an essential role in the dynamo process. The component $\bar{g}$ can easily be generated in a spherical dynamo through a stretching of the axisymmetric poloidal field by a differential rotation. This process is known as the ω -effect. But the amplification of $\bar{h}$ requires the interaction of the nonaxisymmetric components of magnetic fields and velocity fields. This is often called the α -effect. This latter effect can also be used, of course, for the generation of $\bar{g}$ in the absence of a differential rotation. One thus distinguishes $\alpha\omega$ - and α^2 -dynamos. These concepts were originally introduced within the framework of mean-field magnetohydrodynamics for which we refer to the book of Krause and Raedler (1980).

Energetics of the Geodynamo

For a prescribed stationary velocity field, Eq. 3 admits solutions with an exponential time dependence, of which the

growing ones are called dynamos. As the magnetic field grows, the action of the Lorentz force on the velocity field can no longer be neglected. The work done by the velocity field against the Lorentz force equals the energy consumed by ohmic heating. Thus, the sustenance of a finite-amplitude magnetic field against ohmic dissipation requires a source of mechanical energy. A lower bound of about $10^8 W$ on the ohmic heating in the Earth’s core can be obtained by a calculation of the minimum dissipation corresponding to the observed geomagnetic field, while an upper bound of about $10^{13} W$ is given by the observed geothermal heat flux. A realistic value is likely to be within the range 10^{11} to $10^{12} W$.

Numerous sources of energy for the geodynamo have been considered in the history of dynamo theory. The major ones that cannot be excluded right away are the precession of the Earth, thermal heat sources, and compositional buoyancy. The latter source first proposed by Braginsky in 1963 is probably the largest contributor, since the turbulence induced by precession appears to be marginal and since the Carnot efficiency of heat sources is relatively low (Gubbins et al. 2004). Compositional buoyancy is generated by the growing solid inner core. Iron and perhaps nickel crystallize at the inner core-outer core boundary while lighter elements stay in solution. The decrease in density caused by the increase in the concentration of light elements provides a source of potential energy that can drive convection in the outer core. Since the outer core contains about 10% light elements by weight, most likely silicon, sulfur and oxygen, the buoyancy provided by this process appears to be more than adequate to drive the geodynamo. The compositional buoyancy mechanism has the advantage that it could explain the apparent absence of a dynamo in the case of Venus (Stevenson et al. 1983; Nimmo 2002). Since Venus resembles the Earth in most respects, the absence of a measurable magnetic field has posed a puzzle. The somewhat higher pressure at the Earth’s center and the more efficient heat transport out of the planet because of plate tectonics seem to have allowed the inner core to form, whereas Venus will reach this stage of its thermal history much later. Estimates for the age of formation of the inner core of the Earth range from 1 to 2 b.y. Since evidence for geomagnetism exist for earlier times in the paleomagnetic record when the inner core had probably not yet nucleated, thermal buoyancy generated by cooling and possibly by radioactive elements is likely to have played a role in driving the geodynamo in the Earth’s early history. In this respect the Earth resembles the major planets for which thermal buoyancy appears to be the only viable source of energy to drive the planetary dynamo.

The way in which compositional buoyancy drives convective motion in the outer core is not well understood. It is even possible that the outer part of the outer core is stably stratified. Another peculiar property of the solidification process at the inner core is the so-called mushy zone in which liquid and

Geomagnetic Field, Theory, Table 1 Properties of the liquid outer core of the Earth

Property	Symbol	Value	Unit
Angular velocity of rotation	Ω	$7.27 \cdot 10^{-5}$	s^{-1}
Mean outer radius	r_o	3485 ± 3	km
Mean inner radius	r_i	1225 ± 15	km
Density	ρ	$9.9 - 12.2$	$10^3 kg/m^3$
Coefficient of thermal expansion	α	10^{-5}	K^{-1}
Thermal diffusivity	κ	$5 \cdot 10^{-6}$	m^2/s
Magnetic diffusivity	λ	2	m^2/s
Kinematic viscosity	ν	10^{-6}	m^2/s
Permeability of vacuum	μ_0	$4\pi \cdot 10^{-7}$	H/m

solid phases coexist (see ► “Earth’s Structure, Core”). But those details have little influence on the theory of the geodynamo in its current state. Since the form of motion in the outer core is primarily determined by Coriolis and Lorentz forces the theory of geomagnetism has proceeded without distinguishing much between thermal and compositional buoyancy.

Some typical values of parameters used in the theory of the geodynamo are given in Table 1.

Magnetohydrodynamic Theory of the Geodynamo

The theory of the geodynamo requires the inclusion of the equations of motion for the rotating liquid core for several reasons. First of all, constraints on the velocity fields used in the dynamo equation are needed since rather arbitrary velocity fields are capable of generating fields that resemble the observed geomagnetic field. The velocity fields driven by geophysically realistic forces share special properties, such as the approximate alignment of the vorticity with the axis of rotation, that make them useful in the interpretation of certain features of the observed magnetic field. The second reason for including the equations of motion stems from the necessity of taking into account the Lorentz force without which an equilibrium amplitude of the magnetic field cannot be determined. Thirdly, the time dependence of the geomagnetic field cannot be understood without a detailed knowledge of the dynamics of the velocity field in the core. The basic equations of motion for the velocity vector $\vec{v}$, the heat equation for the deviation Θ from the static temperature distribution, and the equation of induction for the magnetic flux density $\vec{B}$ are usually written in the following form:

$$\nabla \cdot \vec{v} = 0, \quad \nabla \cdot \vec{B} = 0, \quad (13)$$

$$\begin{aligned} & \left(\partial_t + \vec{v} \cdot \nabla \right) \vec{v} + 2\Omega \vec{k} \times \vec{v} \\ & = -\nabla \pi + \alpha \Theta g \vec{r} + \nu \nabla^2 \vec{v} + \vec{B} \cdot \nabla \vec{B} / \rho \mu, \end{aligned} \quad (14)$$

$$\partial_t \Theta + \vec{v} \cdot \nabla \Theta = \vec{v} \cdot \nabla T_s + \kappa \nabla^2 \Theta, \quad (15)$$

$$\lambda \nabla^2 \vec{B} = \partial_t \vec{B} + \vec{v} \cdot \nabla \vec{B} - \vec{B} \cdot \nabla \vec{v}, \quad (16)$$

where ∂_t denotes the partial derivative with respect to time t and where $-g \vec{r}$ denotes the force of gravity. ν is the kinematic viscosity, $\vec{k}$ is the unit vector of the axis of rotation and Ω is the angular velocity of rotation. The potential temperature distribution T_s describes the difference between the temperature distribution for the static solution of the problem and the adiabatic temperature distribution with the same temperature at the inner core boundary. As mentioned before, Eq. 15 remains essentially unchanged if T_s and Θ are interpreted a concentration of light elements. In that case the coefficient of thermal expansion α and the thermal diffusivity κ just need to be reinterpreted. All gradient terms in the equation of motion have been combined into $\nabla \pi$. In writing Eqs. 13, 14, 15, and 16, we have assumed the Boussinesq approximation implying constant material properties and a constant density ρ except in the gravity term.

Using a dimensionless description, the parameters of the problem can be reduced to four dimensionless groups, the Rayleigh number Ra , the Coriolis number τ , the Prandtl number Pr , and the magnetic Prandtl number Pm which are defined by

$$\begin{aligned} Ra &= \frac{\alpha g \Delta T d^4}{\nu \kappa}, & \tau &= \frac{2\Omega d^2}{\nu}, \\ Pr &= \frac{\nu}{\kappa}, & Pm &= \frac{\nu}{\lambda}, \end{aligned} \quad (17)$$

where the thickness d of the spherical shell has been used as length scale and ΔT is the difference of the potential temperature between the inner core-outer core boundary and the core-mantle boundary. Instead of the Coriolis number often its inverse, the Ekman number $E = \tau^{-1}$ is used. Since λ/ν is roughly 10^6 in the Earth’s core the main sink of mechanical energy is ohmic dissipation and it is generally believed that the term $\nu \nabla^2 \vec{v}$ is unimportant except close to the rigid boundaries of the liquid core. Moreover, the magnetostrophic approximation is often applied in which the inertial term $\left(\partial_t + \vec{v} \cdot \nabla \right) \vec{v}$ is neglected in comparison to the Coriolis force. Without these terms the important Taylor constraint can be derived: After integrating the azimuthal component of Eq. 14 over a coaxial cylindrical surface in the liquid core, one must conclude that an accelerating differential rotation occurs unless the condition

$$\int \int \left[(\nabla \times \vec{B}) \times \vec{B} \right]_\varphi s d\varphi dz = 0 \quad (18)$$

is satisfied for every distance s from the axis. In deriving Eq. 18, we used the fact that the inflow across the cylindrical surface is balanced by the outflow with the result that the contribution from the Coriolis force term vanishes. Different strategies have been designed to accommodate the Taylor constraint (Eq. 18). In Braginsky's (1978) model Z the friction of the viscous Ekman layer at the solid surface of the sphere provides a balance on the right side of Eq. 18. But this effect is small such that the constraint of Eq. 18 becomes nearly satisfied by a poloidal field that is extended parallel to the axis of rotation except for the region close to the boundary.

In modern numerical simulations of convection driven dynamos in rotating spherical shells the Taylor constraint (Eq. 18) is usually not approached because relatively high values of the kinematic viscosity must be assumed for numerical reasons and also because the inertial term $(\partial_t + \vec{v} \cdot \nabla) \vec{v}$ plays a larger role than it possibly plays in the geodynamo. Instead of an Earth-like ratio, values of the order unity for λ/ν are used to ensure numerical convergence. The high value of ν is interpreted as an eddy viscosity which is supposed to represent the dissipative action of small scale components of the velocity field which cannot be resolved in the numerical simulations.

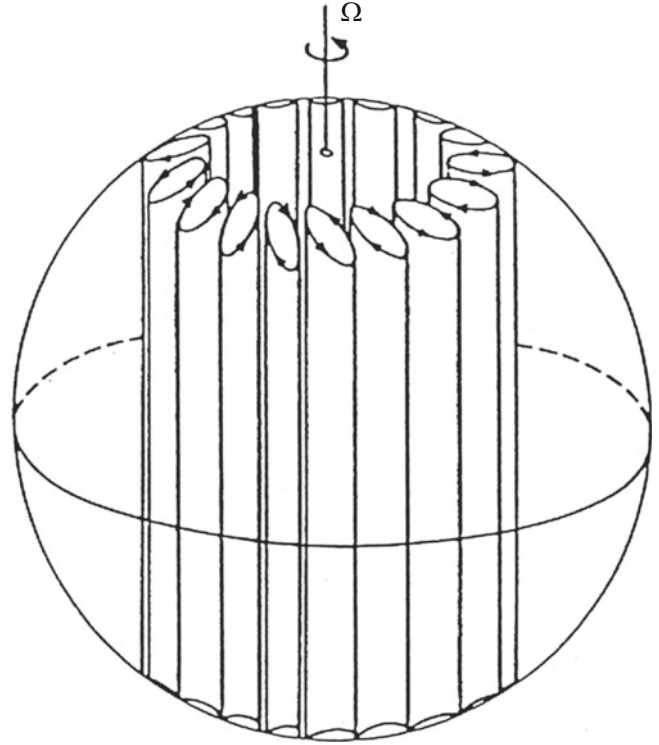
It is instructive to consider Eqs. 13, 14, 15, and 16 in the limit of a vanishing magnetic field. In that case the dominant terms are the Coriolis force and the pressure gradient, the balance of which,

$$2\Omega \vec{k} \times \vec{v} = -\nabla \pi \quad (19)$$

is called the geostrophic balance. The curl of Eq. 14 together with $\nabla \cdot \vec{v} = 0$ gives rise to the Proudman–Taylor theorem

$$\Omega \vec{k} \cdot \nabla \vec{v} = 0 \quad (20)$$

which states that steady motion of a rotating inviscid fluid must not depend on the coordinate in the direction of the axis of rotation. Convection in a sphere requires, however, a finite radial velocity component and therefore cannot fully satisfy condition (20). As a result nonmagnetic solutions of Eqs. 13, 14, 15, and 16 exhibit a time dependence in the form of a drift in the prograde azimuthal direction. In addition they are characterized by small length scales in the directions perpendicular to the axis of rotation whereby viscous friction is increased. A sketch of the motion at the onset of convection in a homogeneously heated sphere is shown in Fig. 3. For details on the dynamics of these convection columns and their dynamo action, we refer to the review article by Busse (2002).



Geomagnetic Field, Theory, Fig. 3 Sketch of convection columns in a rotating, internally heated sphere for Rayleigh numbers close to the critical value

Because of the small viscosity in the Earth's core, the Lorentz force offers a more effective release of the rotational constraint expressed by condition (20) than does viscous friction. The simplest approach to this problem is to consider the onset of convection in the presence of an azimuthal magnetic field of strength B_0 (Fearn 1979). The analysis shows that the scale of the preferred mode of convection increases dramatically as B_0 is increased and the critical value of the Rayleigh number and the drift rate decrease. The dimensionless parameter governing this effect is the Elsasser number

$$\Lambda = \frac{B_0^2}{2\Omega\mu_0\rho\lambda} \quad (21)$$

which describes the ratio between Lorentz and Coriolis forces. When the magnetic field becomes very strong such that Λ exceeds unity, it exerts an inhibiting influence on convection. It turns out that the minimum Rayleigh number for the onset of convection is achieved for $\Lambda \approx 1$. Eltayeb and Roberts suggested in 1970 that this condition is likely to determine the equilibrium strength of the magnetic field inside the Earth's core. While the condition $\Lambda \approx 1$ appears to fit the amplitude of the geomagnetic field reasonably well,

numerical simulations of spherical dynamos do not provide much support for such a relationship. In fact, the dynamical effect of a magnetic field generated by dynamo action is quite different from that of an imposed magnetic field. In present geodynamo models, it does not seem to affect the scale of convection significantly nor has it been found that it allows convection to exist at Rayleigh numbers below the critical value for onset of convection in the absence of a magnetic field.

There is still no general agreement about the relationship that governs the average strength of the magnetic field reached in the dynamo process. But it appears likely that the available energy for driving motions determines the average strength of the magnetic field as has been proposed by Christensen and Aubert (2006).

The velocity field displayed in Fig. 3 is symmetric with respect to the equatorial plane. Even in its turbulent state the convection flow retains approximately this symmetry outside the tangent cylinder which touches the inner core at its equator. Because of this symmetry, the solutions $\vec{B}$ of Eqs. 13, 14, 15, and 16 can be separated into two classes: a quadrupolar class that has the same symmetry with respect to the equatorial plane as the velocity field, and a dipolar class that exhibits the opposite symmetry. In both cases the Lorentz force is compatible with the symmetry of the velocity field. The observed geomagnetic field is primarily a dipolar field and so are most of the other planetary magnetic fields, but from a basic point of view quadrupolar dynamos are not less likely. The more complex structure of their poloidal field is balanced by the simpler form of their zonal component. The latter is symmetric with respect to the equatorial plane, whereas it is antisymmetric in the dipolar case. The observed geomagnetic field includes a substantial contribution from quadrupolar terms which is not surprising, since the symmetry property exhibited in Fig. 3 does not hold inside the tangent cylinder. Moreover, the thermal conditions at the core surface are not likely to be spherically homogeneous because of convection in the Earth's mantle.

The solution of Eqs. 13, 14, 15, and 16 becomes much simpler when the geometry of the spherical shell is replaced by that of a cylindrical annulus with a height decreasing with distance from the axis. An early analytical model (Busse 1975) demonstrates that the dynamo action of convection columns such as those shown in Fig. 3 depends on their helicity. The latter is defined as the azimuthal average over the vorticity multiplied by the velocity vector, $\vec{v} \cdot (\nabla \times \vec{v})$. The helicity is generated for example in the columns through the suction exerted by the Ekman layers at the upper and lower boundaries and by the finite change in height of the columns with distance from the axis. This helicity effect appears to dominate also in modern numerical simulations of the geodynamo.

Time Dependence of the Geodynamo

From observations of the secular variation and from the paleomagnetic evidence for reversals of the direction of the geomagnetic field, it is clear that the mechanism of the geodynamo is not steady. Since a complete model of the geodynamo does not yet exist, the various features that may lead to the observed time dependence cannot be derived from basic mechanisms. Thus, the theories of secular variation and of reversals rely on simplifying assumptions (see ► “[Geomagnetic Field, Secular Variation](#)” and ► “[Geomagnetic Field, Polarity Reversals](#)”). As simulations of the geodynamo become more refined, however, the observed secular variation and the increasing body of data on the process of excursions and reversals will be important for discriminating between different models of the geodynamo.

The time dependence of the nonaxisymmetric part of the magnetic field is a result of the basic dynamics discussed in the preceding section. Even without a magnetic field, the convection motion exhibits a time dependence in the form of a prograde or eastward drift at slightly supercritical Rayleigh numbers, to which a chaotic variability is added as convection becomes more vigorous. In the presence of a magnetic field, this drift is decreased substantially because the Lorentz force balances the ageostrophic part of the Coriolis force. The order of magnitude of the drift corresponds roughly to the observed westward drift of the nondipole part of the geomagnetic field. But since there is likely to be a zonal flow of the liquid core relative to the mantle of a similar order of magnitude, no direct comparison between theoretical and observed drift rates can be made without further knowledge about the velocity field in the core.

The geomagnetic time dependence of internal origin with the shortest duration is known as “geomagnetic jerk.” This term refers to rather abrupt changes in the second derivative with respect to time of the East-West component of the geomagnetic field over the time span of a year or so. Geomagnetic jerks are likely to be the result of magnetohydrodynamic waves in the outer core (Blokhin et al. 2002).

In contrast to the solar magnetic field, the geomagnetic field does not reverse in nearly periodic intervals. The time-scale of reversals is an order of magnitude larger than the timescale of about 2×10^4 years for the diffusive decay of Earth's magnetic field. This may suggest the geodynamo is basically a steady dynamo. Recent evidence (Roberts 2008) suggests, however, that geomagnetic excursions, in which the dipole moment of the Earth decreases to almost vanishing values for a time span of the order of a 1000 years, occur much more frequently than reversals. No consensus about the origin of excursions and reversals has yet been achieved.

Outlook

In the past two decades, the dynamo theory of geomagnetism has progressed enormously through increasing availability of computer resources; see ► [“Core Dynamo”](#). This development will continue in the future with the main goal of reaching higher (and thus more realistic) values of the Coriolis number. Ultimately the lack of data on properties of the Earth’s core and lower mantle will inhibit the progress in understanding the geodynamo.

For a more detailed exposition of the foundations dynamo theory and the associated fluid dynamical concepts, the reader is referred to the book edited by Dormy and Soward (2007) and the book by G. Glatzmayer (2014). A recent review of computational results related to the geodynamo can be found in the article by J. Wicht and S. Sanchez (2019).

Cross-References

- [Core Dynamo](#)
- [Core-Mantle Coupling](#)
- [Earth’s Structure, Core](#)
- [Energy Budget of the Earth](#)
- [Geomagnetic Excursions](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Geomagnetic Field, Secular Variation](#)
- [Paleomagnetism, Principles](#)

Bibliography

- Backus GE (1958) A class of self sustaining dissipative spherical dynamos. *Ann Phys* 4:372–447
- Bloxham J, Zatman S, Dumberry M (2002) The origin of geomagnetic jerks. *Nature* 420:65–68
- Braginsky SI (1963) Structure of the F-layer and reasons for convection in the Earth’s core. *Sov Phys Dokl* 149:8–10
- Braginsky SI (1978) Nearly axially symmetric model of the hydromagnetic dynamo of the earth. *Geomagn Aeron* 18:225–231
- Busse FH (1975) A model of the geodynamo. *Geophys J R Astron Soc* 42:437–459
- Busse FH (2002) Convection flows in rapidly rotating spheres and their dynamo action. *Phys Fluids* 14:1301–1314
- Christensen UR, Aubert J (2006) Scaling laws for dynamos in rotating spherical shells. *Geophys J Int* 166:97–114
- Christensen UR, Aubert J, Cardin P, Dormy E, Gibbons S, Glatzmaier GA, Grote E, Honkura Y, Jones C, Kono M, Matsushima M, Sakuraba A, Takahashi F, Tilgner A, Wicht J, Zhang K (2001) A numerical dynamo benchmark. *Phys Earth Planet Inter* 128:25–34
- Cowling TG (1934) The magnetic field of sunspots. *Mon Not R Astron Soc* 94:39–48
- Dormy E, Soward AM (2007) *Mathematical aspects of natural dynamos*. CRC Press, Boca Raton
- Fearn DR (1979) Thermal and magnetic instabilities in a rapidly rotating fluid sphere. *Geophys Astrophys Fluid Dyn* 14:103–126
- Gilbert W (1600) *De Magnete*. Chiswick Press, London

- Glatzmayer G (2014) *Introduction to modeling convection in planets and stars*. Princeton University Press, Princeton, N.J.
- Gubbins D, Alf D, Masters G, Price GD, Gillan M (2004) Gross thermodynamics of two-component core convection. *Geophys J Int* 157:1407–1414
- Herzenberg A (1958) Geomagnetic dynamos. *Philos Trans R Soc Lond A* 250:543–585
- Krause F, Raedler K-H (1980) *Mean-field Magnetohydrodynamics and dynamo theory*. Pergamon, Oxford
- Nimmo F (2002) Why does Venus lack a magnetic field? *Geology* 30:987–990
- Rikitake T (1966) *Electromagnetism and the Earth’s interior*. Elsevier, Amsterdam
- Roberts AP (2008) Geomagnetic excursions: knowns and unknowns. *Geophys Res Lett* 35:L17307
- Stefani F, Gailitis A, Gerbeth G (2008) Magnetohydrodynamic experiments on cosmic magnetic fields. *Z Angew Math Mech* 88:930–954
- Stevenson DJ, Spohn T, Schubert G (1983) Magnetism and thermal evolution of the terrestrial planets. *Icarus* 54:466–489
- Wicht J, Sanchez S (2019) Advances in geodynamo modelling. *Geophys Astrophys Fluid Dyn* 113:2–50

Geomagnetically Induced Currents

G. S. Lakhina¹, R. Hajra² and B. T. Tsurutani³

¹Indian Institute of Geomagnetism, New Panvel (W), Navi Mumbai, India

²Indian Institute of Technology Indore, Simrol, Indore, India

³Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Definition

Geomagnetically induced current (GIC)

Intense, low-frequency, quasi-direct currents flowing in the conductor systems owing to rapid changes of the geomagnetic fields

Space weather

Dynamic variations at the Sun and in interplanetary space disturbing the near-Earth space environment creating conditions that are hostile to space-borne and ground-based technological systems and can even impose danger to human life

Introduction

Geomagnetically induced currents (GICs) are intense, low-frequency (~ 0.001 – 1 Hz) currents flowing in Earth-ground conductor systems owing to rapid changes of the geomagnetic fields ($dB/dt > 1$ nT s^{−1}) (Campbell 1980; Akasofu and Aspinnes 1982). According to Faraday’s law of induction, a

fluctuating geomagnetic field induces geoelectric fields which drive GICs in the ground and electrical power systems. The induced currents are essentially quasi-DC as their frequencies are much less than the frequency of the electrical grids operating at 50 or 60 Hz.

Intense GICs are known to cause extensive and expensive damages to a variety of technological systems such as electric power systems, telecommunication cables, and natural gas pipelines. Strong GICs enter the neutral grounding of transformers (operating typically at ~ 50 Hz frequency) causing transformer core magnetic flux increase and consequent half-cycle saturation of the core. These in turn distort the AC waveform of the power signal, which can lead to transformer overheating, system-relay interference, reactive power loss, or even system damage.

Space Weather Events and Geomagnetically Induced Currents

Space weather generally refers to the physical conditions in the Sun-Earth system that can affect the performance of spacecraft and ground-based technological systems, including electric power distribution systems. The solar sources of the space weather are the solar activity transients such as solar flares and coronal mass ejections (CMEs), solar wind high-speed ($\sim 550\text{--}800$ km s $^{-1}$) streams (HSSs), and corotating interaction regions (CIRs) formed by the HSS interaction with slow ($\sim 300\text{--}400$ km s $^{-1}$) solar-wind streams. During all of the above phenomena, significant variability occurs in the solar-wind speed, plasma density, and strength and direction of the magnetic field. The dynamics of geospace, extending from the Earth's plasma environment called the ionosphere to the protecting magnetic bubble called the magnetosphere, is largely affected by this solar-wind variability.

Magnetic reconnection between the interplanetary magnetic field (IMF) and the dayside geomagnetic field is considered to be the main mechanism of energy transfer from solar wind to the Earth's magnetosphere (Dungey 1961). The process of magnetic reconnection works most efficiently when IMF is directed southward, opposite to the dayside magnetopause magnetic field direction (Tsurutani and Gonzalez 1997). Dayside magnetic reconnection leads to injection and storage of solar wind kinetic energy (energetic particles) in the magnetotail. This causes stretching of the magnetotail and formation of a near-Earth neutral line. Onset of magnetic reconnection at this neutral line gives rise to an explosive release of energy from the magnetotail toward the inner magnetosphere. This near-midnight plasma injection excites a substorm (Tsurutani and Meng 1972), and auroras in the high-latitude nightside ionospheric region. If the southward IMF is intense with hours duration, such as during the magnetic cloud portion of an interplanetary coronal mass ejection

(ICME) (Burlaga et al. 1981), the plasma injection is deeper and the particles get energized to $\sim 10\text{--}300$ keV. This leads to intensification of the ring current which produces a diamagnetic decrease in the Earth's magnetic field measured at near-equatorial magnetic stations. This latter effect is the *main phase* of the geomagnetic storm. After the storm main phase is over, the decay of the ring current starts the *recovery phase* of the magnetic storm. The energetic particles are lost by several physical processes, namely, charge-exchange, Coulomb collisions, wave-particle interactions, and convection out the front side of the magnetopause (Kozyra and Liemohn 2003).

The intensity of the magnetic storm is measured by the disturbance storm time (Dst) index, an hourly geomagnetic activity index, or by the SYM-H index with a temporal resolution of 1 min. The intensity of a substorm is measured by the auroral electrojet (AE) index. The intensities of both substorms and magnetic storms are expressed in nT. During intense magnetic storms, near-equatorial magnetic field disturbances of a few 100s nT are recorded for a few hours to days. This ground magnetic effect is due to the magnetospheric ring current which exists from $L = 2$ to $L = 8$ (L is the distance in Earth's radius R_E that a dipole magnetic field crosses the magnetic equator). During super magnetic storms, the magnetic disturbances could be higher, exceeding 1000 nT at times. Substorms cause high-latitude ground-magnetic disturbances from ~ 100 to 1000s nT lasting for a few minutes to hours. These magnetic effects are due to a combination of the auroral electrojet which is located in the ionosphere at a height of ~ 100 km and field-aligned currents coming from the magnetosphere to the ionosphere. The ring current effects of substorms are relatively minor compared to that of magnetic storms. During magnetic storm main phases, many substorms may also occur.

A second means of energy transfer from the solar wind to the magnetosphere is sudden solar wind ram pressure changes. One mechanism is the pressure pulse due to high plasma densities sunward of fast magnetosonic shocks (Kennel et al. 1985). These cause sudden impulses (SI $^+$ s), strong changes of the equatorial magnetic field of 10–100 nT with $\sim 1\text{--}10$ min durations (Araki 1977), the triggering of nightside substorms (Heppner 1955), and in extreme cases, the formation of new radiation belts (Blake et al. 1992).

Geomagnetic field disturbances associated with magnetic storms and substorms are known as the key factor for the generation of GICs. Strongest GIC events are recorded in the high ($>50^\circ$) geomagnetic latitude zone in association with large amplitude magnetic field fluctuations during intense and frequent auroral activities (Pirjola 2000). However, recent studies show evidence of mid-to-low latitude GICs in association with SI $^+$ events caused by solar wind ram pressure impingement on the dayside magnetopause of the Earth (Gaunt 2014; Carter et al. 2015).

Generation of GIC Events

A GIC event is generated by the high-frequency changes in geomagnetic fields due to space weather events. Geomagnetic storms, substorms, supersubstorms, magnetic pulsations, and sudden impulses (SI^+) are the major drivers for the generation of GICs (Pulkkinen et al. 2005; Viljanen et al. 2006; Ngwira et al. 2014; Tsurutani et al. 2020). In fact, a GIC event is essentially the ground manifestation of the complex space-weather chain starting from the solar eruptions, extending to the interplanetary space, and then to the surface of the Earth and below. Therefore, the GIC signals carry information about the processes involved in the Sun-Earth connection and space weather. Most models calculate GIC either from the measurements of geomagnetic field (B) and its time derivative dB/dt using magnetometers or from the magnetotelluric measurements of both B field and electric field E (Gaunt 2016; Pulkkinen et al. 2017).

Systematic studies on GICs and their effects on the electric power systems were started first in the high-latitude countries, e.g., Finland (Viljanen and Pirjola 1994) and Canada (Boteler 2001), where GICs might be expected to occur. Later, the GIC studies have been undertaken in the low- and mid-latitude regions such as Scotland, China, Japan, South Africa, Brazil, New Zealand, and Australia. An excellent overview of GIC impacts on electrical power systems is given by Gaunt (2016). Direct current (DC) blocking devices, such as neutral-point reactors and series capacitors, are being designed and installed in high voltage transmission grids to mitigate the flow of GICs through transformers (Kappenman et al. 1991; Thomson et al. 2010). The real-time monitoring of GICs provides a useful means to understand the power-system status, but it does not guarantee that the system will survive a future severe GIC event (Erinmez et al. 2002). At present, no reliable forecastable model for GICs exists (Gaunt 2016).

The severity of the GICs flowing through ground-based technological systems depends on several factors such as severity of the magnetic disturbances in the magnetosphere-ionosphere system, location and configuration of the system (i.e., power transformers, transmission lines, etc.), and conductivity of the Earth's surface and interior, coastal effects, etc. (Clilverd et al. 2018; Liu et al. 2018). The highest GIC levels occur for the fastest rate of change of the magnetic field. Therefore, dB/dt is considered as an important parameter to quantify the relative level of GIC threat. For example, the Québec power station blackout during the 13–14 March 1989 magnetic superstorm occurred when the peak dB/dt was ~ 480 nT/min. Recent study of the March 1989 magnetic storm by Boteler (2019) shows that the peak dB/dt occurred during a shock-induced substorm, much before this superstorm attained its peak $Dst = -589$ nT value. It is interesting to note that dB/dt values exceeding 2000 nT/min have been observed, and much higher values ~ 5000 nT/min might have

occurred during historic magnetic storms (Kappenman 2006). Recent studies show that compression of the magnetosphere caused by interplanetary shock impacts, discussed earlier, increases the magnetopause currents leading to large dB/dt variations in the mid- to low- latitudes as well as in the equatorial electrojets (Carter et al. 2015; Oliveira et al. 2018). Furthermore, in many magnetic storms, the GICs due to the SI^+ can be much larger than those recorded at the same station during the storm main phase (Zhang et al. 2015). Thus, even the power stations situated near the equatorial electrojet are susceptible to GIC risks. For maximum possible dB/dt values due to SI^+ s, we refer the reader to Tsurutani and Lakhina (2014).

Future Outlook

Impacts of GICs on electric power systems, natural gas pipelines, and telecommunications have direct socioeconomic impacts. This can be understood from some historical examples of large GIC events. The largest recorded magnetic storm at Earth with a peak $Dst = -1760$ nT was caused by a CME associated with an extremely intense solar flare on 1 September 1859 (Tsurutani et al. 2003). GICs on telegraph lines caused arcing that led to fires and electrical shocks (Loomis 1861). Two recent examples of GIC-related hazards are the collapse of the entire Hydro Québec power system during a superstorm (with peak $Dst = -589$ nT) on 13–14 March 1989 (Allen et al. 1989), and the transformer collapse near Malmö in Sweden during the 29–30 October 2003 superstorm (with peak $Dst \sim -400$ nT) (Pulkkinen et al. 2005; Wik et al. 2008). Due to growing dependence of modern society on electric technology (and greater vulnerability), large GIC impacts are now realized to be quite damaging to society as they can disrupt functioning of the power grids and communication cables, and cause erosion of oil and gas pipelines (Royal Academy of Engineering report 2013; Pulkkinen et al. 2017). Currently GICs are considered as the greatest space-weather hazard globally.

There has been tremendous progress in understanding the space weather processes and generation of GICs during the past two decades. An exhaustive review on the current status and future challenges of space-weather physics and solar-terrestrial physics (STP) is given by Tsurutani et al. (2020). The ultimate challenge for the space community is to predict the occurrence of extreme magnetic storms and their associated extreme GICs. Although extreme events are rare, yet they are the ones that can cause the maximum damage. Presently, there is a limited understanding of the upper limit for the geoelectric fields that drive GICs. Additionally, the prediction of GICs with a lead time of 1–3 days is very limited or unreliable (Gaunt 2016; Pulkkinen et al. 2017) as reliable estimates for the strength of magnetic storms a couple of

days in advance are not available. The major sources of uncertainty are prediction of the IMF southward components associated with the ICME and in the sheath region behind the interplanetary shock. Solar wind measurements taken upstream of the Earth at the Lagrange 1 (L1) point by NASA's Advanced Composition Explorer (ACE) and Deep Space Climate ObserVatoRy (DSCOVR) are being used to provide reliable estimates on the solar-wind drivers. This information is used in the physics-based and empirical models to calculate the currents in the magnetosphere and ionosphere to make real-time prediction of GICs with a lead time of ~30 min or so. For extreme events, such short-time notice is not sufficient to put any mitigation plan in place against hazardous extreme GICs. Since GICs depend on the ground conductivity and characteristics of regional power grids, it is a big challenge for the modelers to develop global GIC models that can predict the magnitude and duration of GIC events in different parts of the world.

Cross-References

- [Electromagnetic Pulsations and Magnetic Storms](#)
- [Gravity Field, Temporal Variations from Space Techniques](#)
- [Magnetotellurics, Crustal Imaging](#)

Acknowledgments GSL thanks Indian National Science Academy, New Delhi, for support under the INSA-Honorary Scientist Scheme. The work of RH is funded by the Science & Engineering Research Board (SERB), a statutory body of the Department of Science & Technology (DST), Government of India through Ramanujan Fellowship. Portions of the research has been performed at the Jet Propulsion Laboratory, California Institute of Technology under contract with NASA.

Bibliography

- Akasofu SI, Aspin JD (1982) Auroral effects on power transmission line systems. *Nature* 295:136–137. <https://doi.org/10.1038/295136a0>
- Allen J, Frank L, Sauer H, Reiff P (1989) Effects of the March 1989 solar activity. *EOS Trans* 70:1479–1488. <https://doi.org/10.1029/89EO00409>
- Araki T (1977) Global structure of geomagnetic sudden commencements. *Planet Space Sci* 25(4):373–384. [https://doi.org/10.1016/0032-0633\(77\)90053-8](https://doi.org/10.1016/0032-0633(77)90053-8)
- Blake JB, Kolassinski WA, Fillius RA, Mullen EG (1992) Injection of electrons and protons with energies of tens of MeV into $L < 3$ on March 24, 1991. *Geophys Res Lett* 19:821–824. <https://doi.org/10.1029/92GL00624>
- Boteler DH (2001) Assessment of geomagnetic hazard to power systems in Canada. *Nat Hazards* 23:101–120. <https://doi.org/10.1023/A:1011194414259>
- Boteler DH (2019) A 21st century view of the March 1989 magnetic storm. *Space Weather* 17:1427–1441. <https://doi.org/10.1029/2019SW002278>
- Burlaga LF, Sittler E, Mariani F, Schwenn R (1981) Magnetic loop behind an interplanetary shock: voyager, Helios, and IMP 8 observations. *J Geophys Res* 86:6673–6684. <https://doi.org/10.1029/JA086iA08p06673>
- Campbell WH (1980) Observation of electric currents in the Alaska oil pipeline resulting from auroral electrojet current sources. *Geophys J Int* 61:437–449
- Carter BA, Yizengaw E, Pradipta R, Halford AJ, Norman R, Zhang K (2015) Interplanetary shocks and the resulting geomagnetically induced currents at the equator. *Geophys Res Lett* 42:6554–6559. <https://doi.org/10.1002/2015GL065060>
- Clilverd MA, Rodger CJ, Brundell JB, Dalzell M, Martin I, MacManus DH et al (2018) Long-lasting geomagnetically induced currents and harmonic distortion observed in New Zealand during the 7–8 September 2017 disturbed period. *Space Weather* 16:704–717. <https://doi.org/10.1029/2018SW001822>
- Dungey JW (1961) Interplanetary magnetic field and the auroral zones. *Phys Rev Lett* 6:47–48. <https://doi.org/10.1103/PhysRevLett.6.47>
- Erinmez IA, Kappenman JG, Radasky WA (2002) Management of the geomagnetically induced current risks on the national grid company's electric power transmission system. *J Atmos Solar Terr Phys* 64:743–756
- Gaunt CT (2014) Reducing uncertainty – responses for electricity utilities to severe solar storms. *J. Space Weather Space Clim* 4:A01. <https://doi.org/10.1051/swsc/2013058>
- Gaunt CT (2016) Why space weather is relevant to electrical power systems. *Space Weather* 14:2–9. <https://doi.org/10.1002/2015SW001306>
- Heppner JP (1955) Note on the occurrence of world-wide SSCs during the onset of negative bays at college, Alaska. *J Geophys Res* 60:29–32. <https://doi.org/10.1029/JZ060i001p00029>
- Kappenman JG (2006) Great geomagnetic storms and extreme impulsive geomagnetic field disturbance events – an analysis of observational evidence including the great storm of May 1921. *Adv Space Res* 38:188–199. <https://doi.org/10.1016/j.asr.2005.08.055>
- Kappenman JG, Norr SR, Sweezy GA, Carlson DL, Albertson VD, Harder JE, Damsky BL (1991) GIC mitigation: a neutral blocking/bypass device to prevent the flow of GIC in power systems. *IEEE Trans Power Delivery* 6(3):1271–1281. <https://doi.org/10.1109/61.85876>
- Kennel CF, Edmiston JP, Hada T (1985) A quarter century of collisionless shock research. In: Stone RG, Tsurutani BT (eds) *Collisionless shocks in the heliosphere: a tutorial review*, Geophysical monograph series, vol 34. AGU, Washington, pp 1–36
- Kozyra JU, Liemohn MW (2003) Ring current energy input and decay. *Space Sci Rev* 109:105–131. <https://doi.org/10.1023/B:SPAC.0000007516.10433.ad>
- Liu CM, Wang X, Wang H, Zhao H (2018) Quantitative influence of coast effect on geomagnetically induced currents in power grids: a case study. *J Space Weather Space Clim* 8:A60. <https://doi.org/10.1051/swsc/2018046>
- Loomis E (1861) On the great auroral exhibition of Aug. 28th to Sept. 4, 1859, and on auroras generally. *Am J Sci* 82:318–335
- Ngwira CM, Pulkkinen A, Kuznetsova MM, Gloer A (2014) Modeling extreme “Carrington-type” space weather events using three-dimensional MHD code simulations. *J Geophys Res* 119:4456–4474. <https://doi.org/10.1002/2013JA019661>
- Oliveira DM, Arel D, Raeder J, Zesta E, Ngwira CM, Carter BA et al (2018) Geomagnetically induced currents caused by interplanetary shocks with different impact angles and speeds. *Space Weather* 16:636–647. <https://doi.org/10.1029/2018SW001880>
- Pirjola R (2000) Geomagnetically induced currents during magnetic storms. *IEEE Trans Plasma Sci* 28:1867–1873. <https://doi.org/10.1109/27.902215>
- Pulkkinen A, Lindahl S, Viljanen A, Pirjola R (2005) Geomagnetic storm of 29–31 October 2003: geomagnetically induced currents and their relation to problems in the Swedish high-voltage power

- transmission system. *Space Weather* 3:S08C03. <https://doi.org/10.1029/2004SW000123>
- Pulkkinen A, Bernabeu E, Thomson A, Viljanen A, Pirjola R, Boteler D, Eichner J, Cilliers PJ, Welling D, Savani NP, Weigel RS, Love JJ, Balch C, Ngwira CM, Crowley G, Schultz A, Kataoka R, Anderson B, Fugate D, Simpson JJ, Mac Alester M (2017) Geomagnetically induced currents: science, engineering, and applications readiness. *Space Weather* 15:828–856. <https://doi.org/10.1002/2016SW001501>
- Royal Academy of Engineering report (2013) Extreme space weather impacts on engineered systems and infrastructure. Royal Academy of Engineering, London
- Thomson AWP, Gaunt CT, Cilliers P, Wild JA, Opperman B, McKinnell LA, Kotze P, Ngwira CM, Lotz SI (2010) Present day challenges in understanding the geomagnetic hazard tonational power grids. *Adv Space Res* 45(9):1182–1190. <https://doi.org/10.1016/j.asr.2009.11.023>
- Tsurutani BT, Gonzalez WD (1997) The interplanetary causes of magnetic storms: a review. In: Tsurutani BT et al (eds) *Magnetic storms, Geophysical monograph series*, vol 98. AGU, Washington, DC, pp 77–89. <https://doi.org/10.1029/GM098p0077>
- Tsurutani BT, Lakhina GS (2014) An extreme coronal mass ejection and consequences for the magnetosphere and Earth. *Geophys Res Lett* 41. <https://doi.org/10.1002/2013GL058825>
- Tsurutani BT, Meng CI (1972) Interplanetary magnetic-field variation and substorm activity. *J Geophys Res* 77(16):2964–2970
- Tsurutani BT, Gonzalez WD, Lakhina GS, Alex S (2003) The extreme magnetic storm of 1–2 September 1859. *J Geophys Res* 108:1268. <https://doi.org/10.1029/2002JA009504>
- Tsurutani BT, Lakhina GS, Hajra R (2020) The physics of space weather/solar-terrestrial physics (STP): what we know now and what the current and future challenges are. *Nonlinear Processes Geophys* 27:75–119. <https://doi.org/10.5194/npg-27-75-2020>
- Viljanen A, Pirjola R (1994) Geomagnetically induced currents in the Finnish high-voltage power system. *Surv Geophys* 15:383–408. <https://doi.org/10.1007/BF00665999>
- Viljanen A, Tanskanen EI, Pulkkinen A (2006) Relation between substorm characteristics and rapid temporal variations of the ground magnetic field. *Ann Geophys* 24:725–733
- Wik M, Viljanen A, Pirjola R, Pulkkinen A, Wintoft P, Lundstedt H (2008) Calculation of geomagnetically induced currents in the 400kV power grid in southern Sweden. *Space Weather* 6:S07005. <https://doi.org/10.1029/2007SW000343>
- Zhang JJ, Wang C, Sun TR, Liu CM, Wang KR (2015) GIC due to storm sudden commencement in low-latitude high-voltage power network in China: observation and simulation. *Space Weather* 13:643–655. <https://doi.org/10.1002/2015SW001263>

Geophysical Well Logging

Miroslav Kobl
Institute of Hydrogeology, Engineering Geology and Applied Geophysics, Charles University in Prague, Prague, Czech Republic

Synonyms

Borehole geophysics

Definition

Borehole geophysics includes all methods that make continuous profiles or point measurements at discrete depth stations in a borehole. These measurements are made by lowering different types of probes into a borehole and by electrically transmitting data in the form of either analog or digital signals to the surface, where they are recorded as a function of depth. The measurements are related to the physical and chemical properties of the rocks surrounding the borehole, the properties of the fluid saturating the pore spaces in the formation, the properties of fluid in the borehole, the construction of the well, or some combination of these factors.

Introduction

The science of well logging was begun by Conrad Schlumberger in 1927 as an application of his work on resistivity measurements of the earth in surface exploration (see ► [“Electrical Resistivity Surveys and Data Interpretation”](#)). From the early measurement of electrical properties by means of normal and lateral resistivity curves, modern logging has progressed through such changes as the introduction of wall-contacting microspacing device, dipmeters, and conductive and inductive techniques. Parameters other than electrical have been utilized, such as radioactive properties of rocks both passive and forced as a result of bombardment by both gamma rays and neutrons; the measurement of acoustic properties, the measurements of other rock properties and auxiliary technique developments such as dielectric logging, temperature log, downhole viewers, remote sensors (dipmeter log, radar system, borehole gravimeter), etc.

The rapid advancement of well-logging techniques is observed in the 1980s and 1990s with the motivation of the petroleum industry to develop ever more cost-effective technology to evaluate thin-bed and nonconventional reservoirs. At the same time, geophysical technology also advanced rapidly with development of new techniques and improved mathematical methodology. The increased use of array tool design (multiple sources and receivers) provides improved vertical resolution, improved and multiple depths of investigations, and enables radial and azimuthal imaging by resistivity tools. Cased-hole resistivity logging may be realized, what could not be possible earlier. Acoustic-array designs, using either dipole or quadrupole sources, can acquire high-quality shear wave data in most types of reservoir rocks. A new generation of nuclear magnetic resonance (NMR) tools using pulseecho technology provides an improved measurement without the operational restrictions of earlier tools. Nuclear tools have been improved by the use of new high-density gamma-ray detectors that allow higher logging speeds and better repeatability and permit a slimhole, through-tubing

tool design. Renewed interest in neutron activation methods has produced new tools based on oxygen and silicon activation, like the pulsed-neutron source sensor. Improvements to existing resistivity and nuclear measurement-while-drilling (MWD) tools plus new caliper and acoustic porosity devices and measurement at the bit, ensure that MWD will continue to replace wireline logging in many situations. High angle, horizontal, and slimhole wells are now common as a result of advances in drilling technology. Wireline and MWD logging have been adapted for use in these wells; and new directionally focused resistivity tools have been developed; logging on coiled tubing is now routine. Improvements of downhole microprocessor-based technology and high-speed digital telemetry make possible the use of high data-rate array and imaging tools as well as increased possibilities of combinations of different tool types. Concurrent improvements in uphole computer processing combined with new visualization techniques made possible by the advent of powerful desktop computers enhance presentation and interpretation of the data acquired with these tools.

A chronological listing of these developmental steps is presented in literature (Johnson 1962; Segesman 1980; Prensky 1994). The applications for geophysical welllogging methods continue to expand away from exploration for natural resources (oil and gas, water, minerals, geothermics) into the field of civil engineering, rock mechanics, disaster prevention, and environmental preservation. Information about recent and emerging developments and trends in well-logging and rock-characterization technology offer a lot of publications from meetings of geophysical societies (proceedings) and geophysical journals and books.

The main objective of borehole geophysics is to obtain more information than can be obtained from conventional drilling, sampling, and testing. Drilling a borehole or well is expensive, but the borehole provides access to the subsurface for geophysical probes. The most obvious log is a description of extracted cores, to provide information of lithology, microfossils, permeability, porosity, and fluids. These logs are used for geological correlation between different wells across an oil field and for the quantitative interpretation of the geophysical logs. The ideal would be to obtain continuous core, but this is very slow and thus expensive, therefore, it is done only for formations of special interest, usually potential oil-producing formations when they can be identified in advance. After drilling, samples may be obtained by sidewall sampling. Samples of fluids from surrounding rocks are also taken with a formation tester. Borehole logs provide continuous records that can be analyzed in real time at the well site. Thus, they can be used to guide completion or testing procedures. Log data are readily reproduced over long periods. Logs also aid the lateral and vertical extrapolation of geologic and fluid sampling data or hydraulic test data obtained from wells. In contrast, samples of rock or fluid from a borehole provide data

only from sampled depth intervals and only after laboratory analysis. Data from some geophysical logs, such as acoustic velocity and resistivity, are also useful in interpreting surface geophysical surveys.

The manner in which a well is drilled affects the logging environment, through such factors as the resistivity of the fluid in the well bore, the diameter of the well, the roughness (rugosity) of the borehole wall, and the mudfiltrate invasion of porous formations surrounding the well. The thickness and composition of material between a logging probe and the borehole wall can have an effect on the data set. In general, the larger the borehole diameter, the poorer the quality of the geophysical logs obtained.

Wellbore Environment

The physical properties of the geological formations in the vicinity of a borehole are often affected by the drilling process. Further, the temperature, lithostatic stress, and fluid pressure increase with depth. Logging cable and tools must be constructed in such a way that can withstand the highest temperatures and pressures encountered in the well. Interpretation of well-log data may commonly require temperature correction.

Drilling is usually done using a rotating bit that grinds away the formations to produce cuttings from shales, sand from unconsolidated sandstones, or chips up to a few millimeters in size from solid rocks. The bit is fitted to the bottom of the drill stem, which typically consists of 10-m lengths of steel pipe screwed together. The bit may be turned by rotating the whole drill string of 10 m pipes from its top, but more commonly nowadays the bit is turned up by the mud being pumped through the turbine just behind the bit. Turbine drilling allows much greater control of drilling direction.

Down through the drill stem is pumped drilling fluid, commonly termed mud. It is a mixture of ingredients (liquid and solids) formulated to produce the required rheological properties such as density and viscosity. Drilling mud expedites the drilling operation by cooling, lubricating and cleaning the bit, reducing friction, and transporting cuttings from bit to the surface. Mud also serves to stabilize the wall of the hole, and – by its weight – to prevent any high-pressure gas encountered from blasting out of the hole, and also forms a mud cake, which is a layer made up of the solids in the mud deposited on the sides of hole opposite an intergranular permeable formation ensuring that the hydrostatic pressure of the drilling mud is higher than that of groundwater. The formation is invaded by the liquid of the mud, the mud filtrate, to a distance that depends on the porosity of the formation and properties of the mud. Nearest to the hole is the flush zone, in which the fluids are entirely mud filtrate (apart from some residual fluid that cannot be removed); next is the transition

zone, with decreasing replacement until a sufficient distance is the uninvaded zone, quite unaffected by the drilling. In very permeable and porous formations, the mudcake builds up quickly and forms a barrier to further invasion. In less-permeable, low-porosity formations, the buildup is slow and invasion can be very deep. Thus, the resistivity of permeable rocks changes markedly because of the origin of these radial zones (Schlumberger 1987). In pelitic formations (claystones, siltstones), the changes in elastic and mechanical (strength) properties occur along the borehole walls. Under the influence of drilling mud, the clay minerals swell and the cohesion of the rock is reduced; a low-velocity zone is formed along the borehole wall in the pelitic formations, being the more marked the longer the time interval between drilling and the moment of velocity measurement.

Calibration and Standardization of Logs

Standardization of logging measurements is the process of checking the response of logging probes in the field, using some type of portable field standard, usually before and after logging a well, to provide the basis for system drift correction in output with time. The Gamma-ray calibration pit and Neutron pit of the American Petroleum Institute (API) in the University of Houston are generally accepted over the world as a basic standard for nuclear logs because of definition of the API Gamma-ray unit and API Neutron unit.

Calibration is the process of converting measuring units of log response to units that represent rock characteristics. Calibration facilities should closely simulate the lithology and geometry to be logged. In many applications, it becomes difficult to utilize such calibration facility for several reasons (namely, the local lithologic environment substantially differs from lithology of calibrations blocks). Then, the calibration is performed by developing a regression between the measured geophysical property (the log) and the property of interest determined from laboratory tests using recovered core. The necessary condition is that the boreholes under study have been carefully cored. In principle, the quality of calibration is indicated by the correlation coefficient given from the regression. This calibration approach is also called “method of key beds.”

Benefits and Limitations of Logging

Well logging differs from most other geophysical methods, not only in being carried out down a borehole, but in relating the physical quantities measured more specifically to geological properties of interest, such as lithology, porosity, and fluid composition. Their main advantage is greater detail, the measurements are made within formations of interest and can be

compared with the geological information obtained from the same hole. Borehole logs usually provide continuous data that can be used to guide testing procedure and thus may reduce the number of samples needed and in this way the cost. The data are useful for interpreting surface geophysical surveys. Log data are repeatable over a long period. The repeatability and comparability provide the basis for measuring changes in exploitation wells. Changes in quality of casing, such as grade of corrosion, clogging, or changes in exploitation regime, may also be recorded.

Well logging cannot replace sampling completely. Background information is needed on each new geologic environment to aid log analysis. A log analysis cannot evaluate a set of logs properly without information on the local geology. Logs do not have a unique response, so it is necessary to combine more methods. And the limitation of logging measurements, which should be considered, is the borehole environment (diameter, mud, dimension of a probe, temperature, pressure, velocity of measurement). High technology and sophisticated interpretation procedures can overcome these limitations.

Principles

Well logging uses the principles of almost all methods in geophysical surface surveys: electrical, nuclear, seismic, geothermal, gravity, magnetic, and electromagnetic and additionally some procedures particular to drilling activity. The latter involve measurement of borehole diameter, bearing and deviation of the borehole from vertical, the strike and dip of beds, and the direction and velocity of fluid moving within a borehole. The instruments and measurements have to be adapted for the special conditions of the subsurface, particularly the confined space of a borehole, where they also have to overcome the alterations produced in the surrounding formation by the drilling. The drilled holes occur from a few meters depth for engineering purposes, to over 10 km for the crust at depth investigation. Most important are those for hydrocarbon exploration and extraction, which commonly reach depths of a few kilometers. The boreholes can be vertical, oblique, even horizontal and geophysical measurement can be also carried out between holes or between hole and the surface: nearwell and crosswell reflection imaging and tomography.

The variation of a property down a borehole is recorded against depth as a log. Most logs are run by lowering the tool (measuring device) on a cable, often called wireline. In some situations, especially if the borehole is deviated far from the vertical, or may be unstable, the tool is placed in a modified section of drill pipe, called a collar, and lowered into a hole at or near the end of the drilling string.

Many geophysical logging techniques were developed in the petroleum industry, which is the most important application in the exploration, evaluation, and production of oil and gas, by providing information on porosity, permeability, fluid content, and saturation of the formations penetrated by a borehole, but logging methods have also been applied in holes drilled for other purposes. Logging for uranium and coal is routinely used because it is relatively easy to identify and evaluate uranium and coal beds for their quality and thickness. Evergrowing applications of logs can be observed in water wells for evaluation of porosity and/or fractures, hydraulic parameters and water quality and groundwater dynamics. Shallow holes drilled for the characterization and remediation of contaminants has attracted some new logging techniques. Metal-mining applications continue to rely on the acquisition of core because borehole elemental assay methods have only been developed for a few specific cases. Core is also required for civil engineering (geotechnical) investigations: holes are generally shallow and core makes use of mechanical testing, whereas logging determines rock quality, namely, fractures under in situ conditions. Subsurface temperature measurements serve as critical input to many fields of engineering, exploration, and research such as gas and fluid production engineering, in well completions, in the exploration for hydrocarbons and ore minerals and for high-temperature geothermal resources as well, for testing hypotheses concerning the evolution of the Earth's crust and tectonic processes and for corrections of many temperature-sensitive borehole measurements. Various deep boreholes have been drilled for scientific purposes. Logging technology facilitates data generation for special geological studies concerning sedimentology and stratigraphic correlation, physical properties, mineralogy and geochemistry, fracture and stress, etc.

Instrumentation

Geophysical logging equipment appears complex, and the variety of instruments in use is very large. Any basic measuring system (measuring a natural physical field) consists of a sensor, signal conditioners, and recorders. Logging probes, also called sondes or tools, enclose the sensors, electronics for transmitting and receiving signals, and power supplies. Often there is an isolated artificial signal source in the lower part of the probe, which prevents the direct flow of the signal between source and sensor(s), so that the signal has to go through the rock surrounding the well. The probes (must be watertight and resistant to a high pressure and temperature) are connected to a cable by a cable head screwed onto the top of the probe. Some electric- and acoustic-logging devices are flexible; most are rigid. Probes vary in diameter from 25 mm to 150 mm. The probes are either axially symmetric (centralized) or use bowsprings or

springloaded arms to force the probe sensors pad or pads against the borehole wall (side collimation). Some tool arms may be hydraulically driven. All producers publish temperature, pressure, and diameter specifications for their sondes. Because the diameter is limited and much components must be installed in the sonde, in some cases sondes are very long (some meters) even if a large distance between a source and detector is not required. Usually, in deep holes, the probes are stacked to obtain multiple measurements for economical reasons. Measurements are usually made with the probe pulled upward by a winch fitted with a depth counter to insure steady tension on the logging cable when data are recorded. This results in more uniform speed and more accurate depth determination.

The most recent logging systems are fully digital, with control of the logging equipment implemented from a computer keyboard. Logs are displayed by the computer, and a field record is obtained from a printer. In logging of oil or gas wells, the measured data are telemetered to a computation center where they are promptly evaluated automatically.

The well-log equipments are mounted on vehicle chassis, lighter ones (with cable length ≤ 500 m) on minibus chassis, while the larger ones (with cable length 1,000–3,000 m) are truck-mounted and those for logging deep wells (to a depth over a 5,000 m) are mounted on heavy-duty trucks. In equipments of medium and large size, the winches are driven by the vehicle motors, in light equipments by separate electric motors or hydraulically. Portable devices are used for logging mine and engineering-geological boreholes of difficult access. They are provided with a logging cable up to 300 m long. The devices for logging in coal mines must be sealed to prevent electrically detonated explosions.

The most widely utilized logging methods are summarized in Table 1. The principles and applications of quoted methods in the Table 1 are explained at length in literature (Serra 2007; Serra and Serra 2004; Hearst et al. 2000; Asquith and Gibbon 1982; Keys 1997; Chapellier 1992; Schlumberger 1987/1989; Serra 1986; Rider 1986; Desbrandes 1985; Serra 1984; Pirson 1977, etc.)

Fundamentals of Log Analysis

The process of log analysis for a project can be simplified into several basic steps:

- Data processing, including depth matching and meeting all logs and other data from a single well and editing and smoothing the data set
- Correcting borehole effects and other errors
- Converting log measurements to geologic parameters of interest; this is best done with calibration data for each probe

Geophysical Well Logging, Table 1 The most commonly used logs

Type of log	Potential uses	Borehole conditions required	Limitations
Spontaneous potential SP	Lithology, correlation, water quality	Uncased, mud filled only	Needs contrast in solutes in borehole versus formation
Normal and/or lateral resistivity logs (non focused log), Ra	Lithology, correlation, water quality	Uncased, fluid filled only	Influenced by fluid resistivity and bed boundaries
Conductively focused-current logs, LL micro-focused logs, MLL	Lithology, correlation, water quality, monitoring formation fluid conductivity	Uncased, fluid filled only	Removes some but not all fluid and bed boundary effects
Inductively focused-current logs IL	Lithology, correlation, water quality, monitoring formation fluid conductivity	Uncased or cased with PVC or fiberglass, air or fluid filled	Poor response in massive resistive formations
Gamma-ray log GR spectral GR	Lithology, correlation	Most conditions	Unusual lithology interpretations in glauconite, arkose and phosphate sands
Gamma-gamma log GGL-D	Bulk density, porosity, moisture, lithology, correlation, well construction	Wide range	Mineralogy may influence calibration through Z effect
X-ray fluorescence log XRF	Elemental composition of ores	Slim uncased holes, constant diameter, pure water or dry	Very short radius of investigation, mineralogy affects the sensibility and accuracy
Neutron log NNL pulsed neutron porosity logging	Saturated porosity, moisture, lithology, correlation	Wide range, preferably in slim holes, fresh or brackish fluid filled	Ineffective porosity if clay minerals are included in total porosity
Spectroscopy of neutron-induced γ -rays	Mineralogy from elemental analysis	Wide range	Mineralogy may influence calibration
Nuclear magnetic resonance logging NMR	Amount of free fluids in the reservoir rocks	Uncased	Occurrence of magnetic minerals
Elastic wave propagation log (acoustic log) AL	Saturated porosity, lithology, correlation	Uncased, fluid filled only	Measures primary or matrix porosity rather than total porosity, soft formations
Cement bond log CBL	Fluid content differentiation, bonding of casing	Uncased or cased, fluid filled only	Response gives a circumferential average of bonding, may not detect channels
Acoustic televiewer AT	Character and orientation of fractures, solution openings, and bedding	Uncased, fluid filled, water or mud	Fracture interpretation may be affected by drilling damage to borehole wall
Borehole television BTV	Well construction, secondary porosity, lost objects	Uncased or cased, air or clear fluid	No information in opaque fluid, dark mineral bands mimic fractures
Caliper log CL	Borehole diameter, log correction, well construction, fractures, lithology	Most conditions	Probe performance often affected by borehole deviation
Temperature log TM	Fluid movement in the borehole and/or behind casing, gas inflow, heat flux determination	Water filled	Affected by disturbance to fluid column by probe motion
Fluid conductivity log FRL	Fluid movement in the borehole	Water filled	Affected by disturbance to fluid column by probe motion
Flowmeter FM	Vertical fluid movement	Fluid filled	Spurious results may be caused by convection, or by movement of fluid in annulus behind screens
Dipmeter and borehole image technologies	Dip of strata, features of structural geology	Uncased, fluid filled only	
Clinometer	Orientation of the dip log tool, deviation of the hole from vertical	Non metal casing, wide range	
Well completion methods	Well construction, numerous techniques for checking casing and completion materials, borehole deviation, etc.	All conditions	Labor and equipment intensive; completions may be irreversible and often depend on the specific interpretation

- Combining logs and other data from all wells on a project, so that the data can be extrapolated laterally in cross sections or on maps
- Writing a report describing the analytical procedures and results from the logs

The analysis of log data by computer offers a number of time-saving advantages: a large mass of data can be collated and displayed; logs can be corrected and replotted; scales can be changed, smoothing and filtering operations can be carried out; cross plots can be made between different kinds of logs

and core data and calibration curves; correlation functions and rations can be plotted, as well as cross sections and isopach maps. An excellent introduction into this matter is presented by Doveton (1994).

Cross Plots

Because valuable quantities can only be deduced from two or more physical properties, a number of logs are run; often several that will function together are combined in a single tool. Frequently used interpretation technique for more reliable estimation of resulting data is called cross plots. Fertl (1981) gives a good overview of the crossplot technique. The basic procedure is to plot log data for a given depth interval in a space defined by two log parameters. That is, one plots all pairs of observed values of neutron versus acoustic or density parameters, or density versus acoustic. Often, at each point in the plane of coordinate system corresponding to a measured pair of values, one plots a digit representing the number of times that pairs of values occur in the interval. Sometimes, the value of a third variable, such as the gamma-ray log, is measured at the same depth interval as the pair of data being crossplotted or is indicated by a color at the point in space; that is called a z-plot. This is particularly convenient in more complex lithologies.

Applications of Geophysical Well Logging

In the Oil Industry

The *spontaneous potential (SP) log* represents the measurements of the naturally occurring potential difference between two electrodes: one is movable in the well, the other is fixed (grounded) on the surface. It is measured in millivolts. Electric currents arising primarily from electrochemical factors within the borehole create the SP log response. These electrochemical factors are brought about by differences in salinities between mud filtrate (R_{mf}) and formation water resistivity (R_w) within permeable beds. Because a conductive fluid is needed in the borehole for the SP log to operate, it cannot be used in nonconductive (i.e., oil-based) drilling muds. The concept of static spontaneous potential (SSP) is important because SSP represents the maximum SP that a thick, shale-free, porous, and permeable formation can have for a given ratio between R_{mf}/R_w . SSP is determined by formula $SSP = -K \cdot \log(R_{mf}/R_w)$ [here R_{mf} is derived from the resistivity of mud, which is measured in the laboratory on a sample, while K has a value that depends on the temperature, it is 71 at the typical temperature of 25 °C] or chart and is a necessary element for determining accurate values of R_w and volume of shale. The record of SP shows permeable beds and their boundaries

and determines formation water resistivity and also volume of shale in permeable beds.

The *resistivity log* is used to determine the fraction of pore space that is occupied by water, called the water saturation, S_w (the fraction that is hydrocarbons $S_{hc} = 1 - S_w$). Because the rock's matrix or grains are commonly nonconductive, the ability of the rock to transmit a current is almost entirely a function of water in pores. Hydrocarbons are also nonconductive; therefore, as the hydrocarbon saturation of the pores increases, the rock's resistivity also increases. The most simple (standard) form of resistivity logging requires three electrodes that are lowered into a borehole, usually two current electrodes (a pair) A, B, and one measuring electrode M. The measuring electrode N is placed on the surface. The electrodes that are lowered into the borehole consist of lead rings fastened to the lower part of the cable or bar. Such an arrangement of electrodes is known as the resistivity logging probe. The contact between electrodes and formation is made through mud, which therefore must be conducting. The resulting curve is called apparent resistivity R_a . The true resistivity R_t would be measured by this simple sonde only in ideal case of homogeneous isotropic rock environment, which case in the borehole does not exist. The shape of resistivity curves obtained by logging is largely controlled by the configuration of electrodes. There are two basic types of simple resistivity probes, that is, *normal and lateral probes*. Normal probes have a distance between unpaired electrodes far smaller than between paired electrodes and represent the classical way of measurement. Measuring a set of R_a curves with different spacing is required to obtain a plot of values of R_a versus probe spacing. Connecting them by a line, we can correlate in this way obtained curves with theoretical curves (master curves) and thus to determine the true resistivity of the layer investigated, the resistivity and the diameter of the invaded zone. Since this process is extremely laborious, a method was sought which would replace this classic resistivity measurement and enable the true resistivity of rocks to be determined in a simpler way.

Research led to the development of a combination tool, known as laterologs, which consist of a central electrode and symmetrical placed other (shielding) electrodes. They have great advantage as the registered R_a values are closed to the true resistivity (Hearst et al. 2000; Schlumberger 1989; Serra 1986). Examples are the microlog, microlaterolog, proximity log, and induction log.

The *radioactivity logs* involve the use of radioactivity to identify lithologies and to help deduce density and porosity, particularly when a hole is cased because they will function with steel case in place. The simple *natural gamma ray-log* responds to total natural gamma ray activity of the surroundings rocks in the borehole, which is due mainly to potassium (K), thorium (Th), and uranium (U, resp. Ra). Of lithologies commonly present, usually shales have the highest activity;

sands are intermediate; limestones, dolomite, and quartz are low; while anhydrite, salt, coal, and chert are least active. Occasionally, potash beds, and accumulations of radioactive ores or minerals with higher activity than shales may be encountered. Although the gamma ray activity of shales varies widely on worldwide basis, it tends to be constant in a particular field. It is therefore used to estimate shale content, assuming the maximum value on the log means 100% shale and the minimum value means no shale. The *spectral gamma ray log* records separately the activities of K, Th, and U, so allowing more precise identification of shales, for these differ from other rocks in their relative proportions of the three elements. The *gamma–gamma or formation density log* (GG-D) has a strong source of gamma rays (mostly ^{137}Cs) at one end of the probe, at the other is a detector of the gamma rays that is shielded so that it cannot receive rays directly from the source, so only gamma rays that results from scattering from the surrounding rock are counted. The *neutron–neutron log* (NNL) is a record of the neutron flux after its passing from a source in the probe through a formation continuous with the borehole to the detector(s). The tool is equipped with a radioactive source that emits fast neutrons. Neutrons travelling through the formation only slow down significantly when they collide with atoms of a similar mass, that is, hydrogen atoms. That is why the measured value of NNL corresponds to the content of hydrogen (not only as its part of water or oil, but independently in the form of hydrogen occurrence in some minerals), referred to as hydrogen index. Other varieties of neutron–neutron log can be used, interrelated with application of different i) sources such as plutonium/beryllium or americium/beryllium or californium sources using neutron accelerators working in pulse regime and ii) detectors besides neutron tubes for epithermal and thermal neutrons using spectral detection of gamma rays produced as a result of few usual interactions of neutrons with formation (nonelastic collisions, neutron absorption, activation).

The *sonic log* measures a record of the elastic waves velocity versus depth in a borehole (see ► “[Seismic Properties of Rocks](#)”). The most simple tool design is single-transmitter tool, which commonly has a radial piezometric, electrostrictive, or magnetostrictive transducer placed along the axis of an acoustic tool. Receivers are mounted in much the same way. The borehole fluid must be a liquid for most tools to function. Care must be taken to acoustically isolate the transmitter from the receiver so signals are not transmitted through the tool (compare with *Seismic instrumentation*). This measuring system fails if the tool is not parallel to the borehole wall or if there are increases in hole diameter. To solve those problems, the *compensated acoustic log* was developed. It has two transmitters at each end, and for each transmitter there are two receivers positioned far enough along the tool that the first arrivals at both are rays refracted from the wall of the borehole. The choice of tool

frequency is critically important for the generation of useful acoustic logs. High frequency and short wavelength are characterized by detailed spatial resolution and strong attenuation. Low frequency and long wavelength are characterized by poor spatial resolution but low attenuation. Careful study of the effects of frequency on acoustic logging indicates that the most important requirement for logging tool performance is excitation of a single mode. Hearst et al. (2000) find that the most commonly used acoustic-logging frequencies (10–20 kHz) correspond with the source band most likely to excite the single fundamental pair of P and S peaks under typical logging conditions for 15–25 cm diameter boreholes. There are many variations of the continuous velocity log, but all rely more or less on the same principle, though the number and type of sources and receivers are variable. The advent of compact downhole electronics controlled by powerful computers in the logging truck has offered the collection of digital waveforms from logging tools consisting of many receivers. Thus, more information than just the first arrival can be presented, and several methods of presenting the information have been developed: the full-wave display, variable-intensity display. It is now common practice to record the entire waveform digitally for later processing (full-wave acoustic logging). Tool using this mode of acoustic propagation measure both the velocity and the amplitude of the various components of the full acoustic waveform, that is, compressional, shear, and tube (Stoneley) waves (Paillet et al. 1992; Chen 1989). Advances in characterization of full-waveform data facilitate fracture detection and evaluation, permeability measurement, and well completion. The measured data by continuous velocity log can help identify lithologies although the main application is to estimate the porosity. Other developments in acoustic logging are depicted in publication of Prenskey (1994): *reflexion (Pulse-echo) tool* which employs ultrasonic pulse-echo transducers (borehole-televue type) that measure signal velocity and amplitude and have application in both open and cased hole. Further advances in acoustic logging are improvements in techniques such as *Downhole seismic*, *Wellbore resonance*, *Behind-pipe flow*, etc.

The *temperature log* provides the values of temperature of fluids in a borehole (which is not necessarily in thermal equilibrium with the formation temperature) with depth (in $^{\circ}\text{C}$) or the temperature increment over a constant distance, that is, the temperature gradient (in $^{\circ}\text{C}.\text{m}^{-1}$). The instruments used are thermometers and differential thermometers. As a measuring element, temperature-dependent resistors are used. They are located in the probe so as to be in contact with the fluid with their entire surface, if possible, and to be simultaneously thermally insulated from the body of the measuring probe. The time constant of the device is thus reduced to a minimum. There are two types of temperature-dependent resistors: *thermally positive* (metal wires, e.g., platinum, copper, etc.) and *thermally negative* semiconductor

materials (thermistors). These thermally negative materials exhibit substantially higher sensitivity (by as much as 10 times) than the thermally positive materials, but their drawback is that their resistance varies with temperature nonlinearly (in contrast to metals). The temperature sensor independent of its material is usually inserted in a Wheatstone bridge so that temperature-dependent resistor is located in one branch of the bridge, and remaining three branches consist of resistors independent of temperature. For calibration, the logging thermometers are put into a water bath whose temperature is changed gradually and measured as accurately as possible. Desired accuracy is of the order of 0.01 °C, the effective accuracy of commercial temperature log is 0.5 °C. Very often, the temperature is recorded during the lowering of the probe so as to meet a thermal field with the minimum possible disturbance. The application of borehole-temperature data can be grouped in two broad categories: engineering uses and scientific uses (Prensky 1992). Engineering uses incorporate: (1) correction of resistivity logs because temperature normally increases with depth and resistivity is a function of temperature (and salinity); (2) reservoir engineering: temperature is required for calculations of hydrocarbon recovery factors, including pressure–volume–temperature relationships and gas–oil ratios; (3) well completion and production logging: this issue includes the production of hydrocarbons and geothermal energy, location of top of casing cement (owing an exothermic reaction of setting cement producing an obvious temperature anomaly), designing cement slurries, detecting channels in the cement, Fluid flow – temperature anomalies are used for identifying the depth of fluid and gas entry or exit, detection of casing leaks and intervals of lost circulation, fluid injection – the contrast in temperature between injected fluid and formation fluids can identify zones of fractures.

Scientific uses are based on heat-flow measurements in boreholes, which are made by combining sets of temperature measurements under a steady temperature regime and thermal conductivity data, following the formula $Q = K (dT/dz)$, where Q is heat flow, K is thermal conductivity (known from literature for a given lithology or measured in laboratory), and T is temperature at depth z (when the temperature of mud has equalized with the temperature of the formation). Temperature in a borehole also generally changes with time as it returns toward equilibrium after the disturbance caused by drilling, which can take long time and may range from a few days for a shallow (100–150 m) air-drilled hole to several months for deep mud-drilled oilwells. The heat-flow knowledge contributes to solving a lot of scientific work, such as study of the Earth's evolution, stratigraphic correlation (changes in thermal gradient may correspond to lithology variations), basin analysis and modeling, detection of zones of overpressuring, detecting of regional aquifer flow, exploration for geothermal reservoir, etc.

Borehole imaging technologies are based on three types of sensors: electrical, ultrasonic, and optical. The *dipmeter* can also be considered a form of imaging. The dipmeter is used to determine the dip angle and dip direction of planar features intercepted by the borehole.

Indirect-Imaging Technologies: Acoustic and Electrical

The electrical scanning tools rely on a resistivity contrast in the formation to obtain an image and thus limited to holes with conductive mud. Acoustic-imaging tools use a rotating, focused, ultrasonic transducer to measure reflectivity times and amplitude of the pulse-echo from the borehole wall or from the casing. Acoustic amplitude is sensitive to borehole irregularities, thus enabling determination of borehole shape, that is, borehole caliper. Prensky (1994) gives an excellent summary of the literature as regards the possibilities of these borehole imaging tools for identification and analysis of natural and induced fractures; borehole breakouts; thin beds; geologic structure, such as strike, dip, and faults; identification of sedimentary features; interpretation of depositional environment; and evaluation of casing corrosion.

Direct Imaging Technology: *Borehole video* is a wireline tool containing a miniature television camera and light source to obtain high-resolution, real-time visual images of the borehole or tubulars, in both air- or fluid-filled boreholes. The recent introduction of fiber-optic-based systems offers reduction in tool diameter and operation in greater depths. Applications include inspection of casing, tubing and perforations, detection of produced fluid entry, fracture deification, etc.

The *caliper logs* facilitate measurement of the true size and shape of a drilled hole. Most caliper tools are mechanical. The mechanical tool consists of several arms, which are pushed out against the wall of the hole by springs or hydraulic pressure; their extent is converted by an electromechanical device into an electrical signal suitable for recording (proportional to real diameter). Most independent calipers have three or four arms; some have as many as six. The diameter of the borehole is necessary to be known to introduce corrections in quantitative interpretation in almost all methods measured in the boreholes; to calculate the amount of cement needed to fix the casing; to give lithologic information, because washouts are indicative of formation properties – weak formations will tend to wash out, strong ones will not. Caliper logs can also give evidence of the presence of fractures.

The simplest relation between electrical conductivity and formation properties is given by an empirical relation between formation factor F , defined as the ratio of the resistivity of a rock fully saturated by water R_0 to the resistivity of the water R_w with which the rock medium is saturated, and porosity Φ of clean sands (known as Archie's law): $F = R_0/R_w = a/\Phi^m$, where the coefficient a 'and the exponent m ' (cementation factor) can be determined independently from laboratory measurements on core samples and are usually consistent within a lithologic unit. The Archie's formula has some

limitations dealing with conductivity of pore space, which is frequently discussed in literature, for example in Hearst et al. 2000. *Dielectric log* is an electromagnetic method using high frequencies, when dielectric properties begin to dominate the propagation characteristics of electromagnetic waves. The presence of water continues to govern the tool response in this high-frequency regime (100 kHz and greater) because the dielectric permittivity of water is significantly greater than that of most earth materials. *Pulsed neutron logging* measures the rate of decay of thermal neutrons in the interval between two pulses of a neutron generator. *Nuclear magnetic resonance logging* (NMR) has been available for many years, since 1960s, but the literature on the applications of borehole NMR has grown very rapidly in the 1990s, and is constantly increasing. Originally, the method is based on the measurement of nuclear spin and magnetic moment of a hydrogen nucleus providing data of the free fluid index – how large a volume of fluid can be obtained from the rock by pumping, and the character of the fluid at the selected depths. The total signal is proportional to the total number of protons present in both the free and bound fluid. NMR measures proton spin–lattice (longitudinal, T1) and spin–spin (transverse, T2) relaxation times; these relaxation times are proportional to pore size and to bound (clays) and free (pore) fluids. In petrophysical terms, these parameters provide information for quantitative determination of permeability, pore-size distribution lithology in reservoir rocks. Kenyon (1997) is an excellent summary of the application of NMR to petroleum industry.

Porosity derived from logs shows summary in the Table 2.

Hydrogeologic Applications

Geophysical well logs in water wells can be used to determine lithologic parameters with the aim to evaluate quality and geometry of aquifer system, both with intergranular porosity

and fracture and estimate the quality and amount of contained water in the surroundings rocks. In addition to the lithologic parameters, a second important category of information that can be derived from well logs includes data on the character and movement of water in the borehole and the formation. Characteristics of fluids in the well bore can be measured directly, but the data on formation fluids must be inferred from logs. *Logs of the conductivity and temperature* of fluid in the borehole can be related to formation-fluid characteristics, provided that the well has had time to attain chemical and thermal equilibrium with the groundwater reservoir, that an adequate hydraulic connection exists between the hole and the rocks penetrated, and that the inhole flow does not disturb these conditions. *Flowmeters* (Schlumberger 1993) are probes designed to measure groundwater flow rates along boreholes under ambient and forced gradient (during pump-test) conditions. Impellor flowmeters (Hill 1990; Keys and Sullivan 1979) require a significant velocity to rotate impellor blades, so that this probe cannot measure velocities less than about 10^{-1} m/s. Heat pulse (Hess 1986) and electromagnetic (Molz et al. 1994) flowmeters can detect velocities as low as 10^{-3} m/s.

Techniques based on *time sequence measurements of flow logs* are applied only in clean wells, where mud has been replaced by water and the filter cake does not affect aquifer permeability in the vicinity of the borehole. Under these conditions, changes in the physical properties of the borehole fluids under ambient or stressed conditions reflect the properties of the formation and the water contained in the formation. Various properties of water in boreholes, such as the conductivity, temperature, transparency, and/or radioactivity (using tracers) as a function of depth, may be used in this fluid column logging method (Chapellier 1992; Kobr et al. 2005). Many of these properties can be measured using conventional borehole logging equipment. Log-derived hydraulic parameters, such as hydraulic conductivities or transmissivities can be obtained from both flowmeter analysis and time sequence

Geophysical Well Logging, Table 2 Summary of porosity derived from logs

Log	Property measured	Response to the type of porosity	Extraneous effects	Borehole requirements
Gamma-gamma	Electron density	Total porosity, best for high porosity	Matrix, salinity, water saturation, well construction	Wet or dry, cased or open
Neutron	Hydrogen content	Total porosity, best for low porosity	Bound water, matrix chemistry, salinity, water saturation	Wet or dry, cased or open
Acoustic velocity	Average P-wave velocity	Total porosity, only primary and intergranular	Cycle skipping, poor in high-porosity sediments	Fluid filled and open, high pressure
Resistivity	Pore fluid and interconnected porosity	Effective porosity both primary and secondary	Pore shape, fluid chemistry, water saturation	Open or with perforated PVC casing
Gravity logging	Bulk density	Total porosity	Stationary measurement	Deep reading, >50 ft. (18 m)
Nuclear magnetic resonance	Relaxation time of protons	Amount of fluid in pores larger than some minimum	Bound water, very small pores, clay and magnetic minerals	Open hole

of fluid logs. The last named method is also used to study water flow in the rock medium. Movement of the groundwater flow in sandy aquifers is oriented in the direction of the hydraulic velocity, by the hydraulic conductivity and the azimuth (direction) of the flow. Properly interpreted borehole fluid log can be used to locate zones where water enters and/or leaves the well bore, to estimate the specific yields of these zones and all parameters cited above. Also, this methodology is advantageous in contamination studies, to check the quality of well completion and in shallow boreholes for engineering geology (Hess 1986; Tsung et al. 1990; Kelly and Mareš 1993; Mareš et al. 1994; Molz et al. 1994; Paillet and Crowder 1996; Keys 1997; Paillet 2000).

Mineral Exploration

Although boreholes are routinely cored to detect and quantify the presence of economic minerals, often the information is incomplete because of missing core, or misleading because of rapid lateral variation of the mineralization (hydrothermal veins). The depth of investigation may be increased by lowering down the hole closer to the ore bodies. The logging system is adapted to the conditions of a mineral exploration program (smaller holes as diameters – from 49–122 mm; as depths – often less than 100 m) using portable logging system with appropriate methodology. The physical properties of interest are density, average mass number, electrical conductivity, and magnetic susceptibility. For disseminated types of ores induced polarization method is useful. Sometimes, self potential method is also used. In this case, the spontaneous potential is of interest in connection with minerals that are electronic conductors, such as massive sulfides, magnetite, and graphite. If two boreholes are available at reasonable mutual distance a cross-hole tomographic survey can be carried out either with seismic or electromagnetic method. This later can map the conductivity distribution between boreholes in the exploration for minerals as well as in environmental and crustal studies (see ► “Instrumentation, EM”). The modern radioactive methods (X-ray fluorescence analysis, spectroscopy of neutron-induced γ -rays) enable implementation of elemental analysis of the rocks in situ conditions. Due to heavy expenses involved, these methods are not usually considered as the routine ones in mineral exploration. But in some special conditions they may be economic (for example, neutron activation in the borehole on fluorine can be used for calculation of reserves).

Civil Engineering, Rock Mechanics, and Environmental Preservation

Other well-logging applications for various geological studies can be found in literature (Deltombe and Schepers 2004;

Hurst et al. 1992; Hurst et al. 1990). Almost all methods discussed earlier can be applied, some modifications have been introduced. The most important methods for this reason are the acoustic methods: *sonic log*, *vertical seismic profiling*, *seismic tomography*. The main goal of this application is direct (in situ) determination of dynamic elastic properties because sonic logging and waveform analysis provide the means for obtaining continuous measurements of compression and shear velocities. More information can be found in Takahashi et al. (2006).

Summary

Geophysical logs are made by instruments in sondes, or tools, suspended mostly on a wire, with readings usually taken as they are pulled to the surface. Sondes often contain several instruments that can operate without mutual interference. The records of the variation of various quantities down boreholes are called logs, and are of both geophysical and non-geophysical quantities. Geophysical logs are often more valuable when combined than when used singly. Geophysical logging is not only cheaper than continuous coring but it provides information that cannot be obtained on cores, partly because of the alterations produced by the dribbling, and reduces the need for sampling.

Measurements of geophysical parameters in boreholes present the follows characteristics:

- Instruments have to be adapted for the dimensions of a borehole and usually to operate submerged in drilling mud.
- Measurements have to take into account the conditions (pressure, temperature, and technical conditions of the borehole) resulting from the drilling.

The most important application of well logging in the exploration, evaluation, and production of oil and gas, as well as other uses in different geological and engineering branches has been discussed.

Cross-References

- [Electrical Properties of Rocks](#)
- [Electrical Resistivity Surveys and Data Interpretation](#)
- [Gravity Method, Principles](#)
- [Heat Flow Determinations, Continental](#)
- [Instrumentation, EM](#)
- [Magnetic Methods, Principles](#)
- [Magnetometers](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Sedimentary Basins](#)
- [Seismic Instrumentation](#)
- [Seismic Properties of Rocks](#)

- [Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat](#)
- [Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity](#)

Bibliography

- Asquith G, Gibbon C (1982) Basic well log analysis for geologists. Association of Petroleum Geologists, Tulsa
- Chapellier D (1992) Well logging in hydrogeology. A. A. Balkena, Rotterdam
- Chen ST (1989) Shear wave logging with multi-pole sources. *Geophysics* 54:509–517
- Deltombe JL, Schepers R (2004) New developments in real time processing of full waveform acoustic televiewer data. *J Appl Geophys* 55:161–168
- Desbrandes R (1985) Encyclopedia of well logging. Editions Technip, Paris
- Doveton JH (1994) Geologic log analysis using computer methods. AAPG, Tulsa
- Fertl WH (1981) Open hole crossplot – a powerful technique in well log analysis. *J Pet Technol* 33(3):535–549
- Hearst JR, Nelson PH, Paillet FL (2000) Well logging for physical properties. Wiley, Chichester
- Hess AE (1986) Identifying hydraulically conductive fracture with a slow velocity borehole flowmeter. *Can J Earth Sci* 23:69–78
- Hill AD (1990) Production logging-theoretical and interpretive elements. Society of petroleum engineers, Monograph 14, Richardson
- Hurst MA, Lovell MA, Morton AC (eds) (1990) Geological application of Wireline logs. The Geological Society, London
- Hurst MA, Griffiths CM, Worthington PF (eds) (1992) Geological application of wireline logs II. The Geological Society, London
- Johnson HM (1962) A history of well logging. *Geophysics* 27(4):507–527
- Kelly WE, Mareš S (1993) Applied geophysics in hydrogeological and engineering practice. Elsevier, Amsterdam
- Kenyon WE (1997) Petrophysical principles of applications of NMR logging. *Log Anal* 38(2):21
- Keys WS (1997) A practical guide to borehole geophysics in environmental investigations. CRC Press, Lewis Publisher, Boca Raton
- Keys WS, Sullivan JK (1979) Role of borehole geophysics in defining the physical characteristics of the raft river geothermal reservoir. *Geophysics* 44:1116–1141
- Kobr M, Mareš S, Paillet F (2005) Borehole geophysics for hydrogeological studies: principles and applications. In: Rubin Y, Hubard S (eds) *Hydrogeophysics*. Springer, Dordrecht, pp 291–331
- Mareš S, Zbořil A, Kelly WE (1994) Logging for determination of aquifer hydraulic properties. *Log Anal* 35(6):28–36
- Molz FJ, Bowman GK, Zouney SC, Waldrop WR (1994) Borehole flowmeters – field applications and data analysis. *J Hydrol* 163:347–371
- Paillet FL (2000) A field technique for estimating aquifer parameters using flow log data. *Ground Water* 38(4):510–521
- Paillet FL, Crowder RE (1996) A generalized approach for the interpretation of geophysical well logs in groundwater studies. Theory and application. *Ground Water* 34(5):883–898
- Paillet FL, Cheng CH, Pennington WD (1992) Acoustic-waveform logging-advances in theory and application. *Log Anal* 33(3):239–258
- Pirson SJ (1977) Geologic well log analysis. Gulf Publishing, Houston
- Presnky S (1992) Temperature measurements in boreholes: an overview of engineering and scientific applications. *Log Anal* 33(3):313–333
- Presnky SE (1994) A survey of recent developments and emerging technology in well logging and rock characterization. *Log Anal* 35(2):15–45. No. 5, pp. 78–84
- Rider MH (1986) The geological interpretation of well logs. Halsted Press, Blackie/Glasgow/London/New York
- Schlumberger (1987/1989) Log interpretation: principles/applications. SMP/7017. Schlumberger Educational Services, Houston
- Schlumberger (1993) Production log interpretation. Schlumberger Educational Services, Houston
- Segesman FF (1980) Well-logging method. *Geophysics* 45(11):1667–1684
- Serra O (1984) Fundamentals of well-log interpretation. Vol. 1: The acquisition of logging data, Developments in petroleum science, vol 15A. Elsevier, Amsterdam
- Serra O (1986) Fundamentals of well-log interpretation. Vol. 2: The interpretation of logging data, Developments in petroleum science, vol 15B. Elsevier, Amsterdam
- Serra O (2007) Well logging and reservoir evaluation. Editions Technip, Paris
- Serra O, Serra L (2004) Well logging: data acquisition and interpretation. Serralog
- Takahashi T, Takeuchi T, Sassa K (2006) ISRM suggested methods for borehole geophysics in rock engineering. *Int J Rock Mech Min Sci* 43(3):337–368
- Tsung CF, Hufschmied P, Hale FV (1990) Determination of fracture inflow parameters with a borehole fluid conductivity logging method. *Water Resour Res* 26:561–578

Geothermal Heat Pumps

Ladislaus Rybach

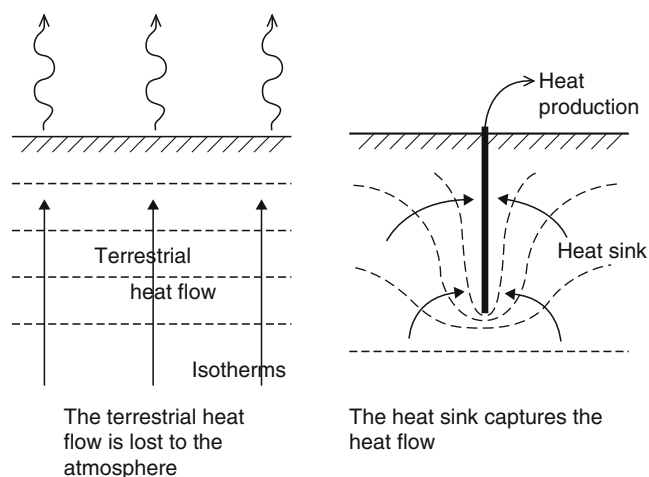
Institute of Geophysics, ETH Zurich, Zurich, Switzerland

Definition

A geothermal heat pump (GHP) is a heating and/or cooling system that moves heat to or from the ground. It uses the earth as a heat source (in the winter) or a heat sink (in the summer). Geothermal heat pumps are also known by a variety of other names, including ground-source, geexchange, earth-coupled, earth energy, or water-source heat pumps. They can be designed and installed in sizes from a few thermal kW to several MW capacity (the latter in modular assemblage). Currently, GHPs represent the largest share in geothermal direct use and account mainly also for the global growth.

The Resource

Shallow geothermal resources (the heat content of rocks in the top few meters of the continental crust) represent a major and ubiquitous energy source. The earth as planet can afford to give off heat by a thermal power of 40 million MW, without cooling down. Without utilization, the terrestrial heat flow is lost to the atmosphere. In this case, the isotherms run parallel to the earth's surface (i.e., horizontal in flat terrain) and the heat flow lines are



Geothermal Heat Pumps, Fig. 1 Principle of geothermal heat extraction and production

perpendicular to them. If, instead, the heat flow can be captured, e.g., by a heat extraction device like a borehole heat exchanger (BHE) (see later), the isotherms are deformed and the heat flow lines can be diverted toward heat sinks (Fig. 1).

Geothermal Heat Pump (GHP) Technology

A new chapter in the history of geothermal applications in buildings opened with the advent of geothermal heat pumps. This technology enables space heating, cooling, and domestic warm water production with the same installation. The GHP application is now in the focus of private, public, and municipal interest (Banks 2008). Geothermal heat pumps (GHPs) are one of the fastest growing applications of renewable energy in the world and definitely the fastest growing segment in geothermal technology, in an increasing number of countries (Rybach 2005). GHP represent a rather new but already well-established technology, utilizing the immense renewable storage capacity of the ground. GHPs use the relatively constant temperature of the ground to provide space heating, cooling, and domestic hot water for homes, schools, factories, public buildings, and commercial buildings. The applicational size can vary from single-family homes with 1–2 borehole heat exchangers (BHEs) to large-scale complexes with hundreds of BHEs. The decentralized systems can be tailor-made, taking into account the local conditions. It is essential to employ proper installation design that takes into account meteorologic, ground property, and technical supply conditions. By these means, reliable long-term operation can be secured. Of the local conditions, the thermal conductivity of ground materials and the groundwater properties are of key importance.

There exist mainly two types of geothermal heat pumps (Fig. 2). In ground-coupled systems, a *closed loop* of plastic

pipe is placed in the ground, either horizontally at 1–2 m depth or vertically in a borehole down to 50–300 m depth. A water-antifreeze solution is circulated through the pipe. Thus, heat is collected from the ground in the winter and optionally heat is rejected to the ground in the summer. An *open loop* system uses groundwater or lake water directly as a heat source in a heat exchanger and then discharges it into another well, a stream, or lake, or even into the ground. The installation of horizontal coils needs relatively large surface area and extensive earthworks (digging the ground down to the level of coil layout); the prerequisite for extracting the heat of groundwater is the presence of a shallow water table. For these reasons, the most widespread technology of shallow heat extraction is by BHEs. Heat extraction is established by closed-circuit fluid circulation (a few m³/h of pure water or with an antifreeze additive) through the BHE and the evaporator side of the heat pump.

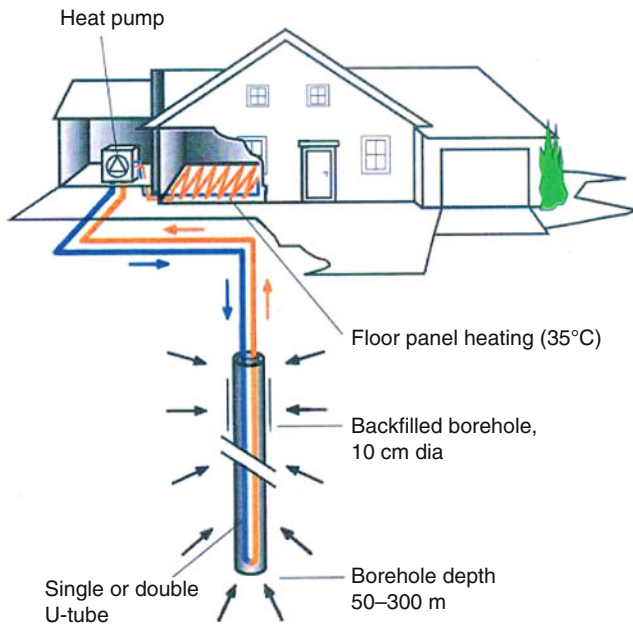
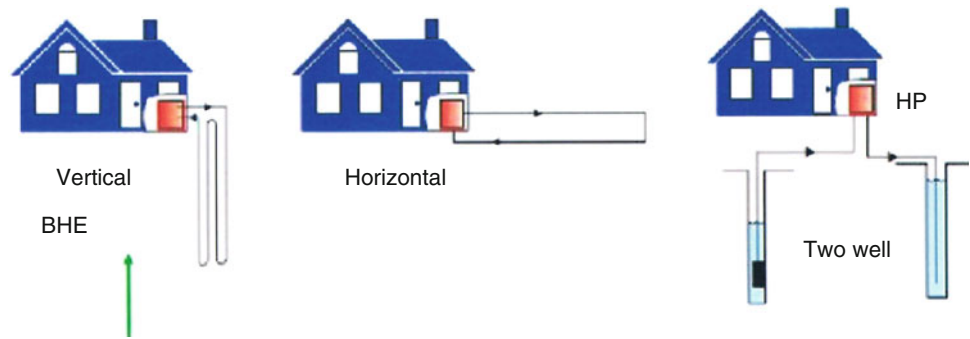
Three basic components make up a GHP system: (1) the heat extraction/storage part in the ground, (2) the central heat pump, and (3) the heat distributor/collector in the building (e.g., floor panel). The key component is the heat pump. In essence, heat pumps are nothing more than refrigeration units that can be reversed. In the heating mode, the efficiency is described by the coefficient of performance (COP) which is the heat output divided by the electrical energy input for the heat pump. Typically, this value lies between 3 and 4 (Rybach 2005). Except for larger, singular applications where gas-driven heat pumps are used, most heat pumps use electricity for their compressors. Therefore, GHPs are electricity consumers. The source of electricity varies from country to country; it would be elegant if the electricity to drive the GHP heat pumps would originate from renewable sources like solar, wind, or even geothermal! The principal components of a BHE-based GHP system are depicted in Fig. 3.

Heating and Cooling with GHPs

As mentioned above, GHP systems can provide space cooling also. In moderate climate, the ground below about 15 m depth is significantly colder than outside air in summer. Thus, a large geothermal store with favorable heat capacity is available where the heat can be exchanged (extracted from the building and deposited in summer, extracted from the ground store and supplied to the building in winter). The thermal capacity of the system depends, besides the volume, on the thermal and hydrogeologic characteristics of the installation site; these must be carefully considered in system dimensioning. In summer, most of the time the heat pump can be bypassed and the heat carrier fluid circulated through the ground by the BHEs and through the heating/cooling distribution (e.g., floor panels).

Geothermal Heat Pumps,

Fig. 2 Closed loop and open loop (groundwater) heat pump systems. The arrow indicates the most common system, with borehole heat exchangers (BHE). The heat pump (HP) is the red box in the house. (Modified from Lund et al. 2003)



Geothermal Heat Pumps, Fig. 3 Sketch of a GHP system with a single BHE. Horizontal/vertical arrows indicate circulation in the U-tube heat exchanger, oblique arrows indicate heat extraction from the ground (heating mode in winter). In summer, the oblique arrows are reversed; heat is extracted from the building and stored in the ground

By these means, the heat is collected from the building and deposited in the ground for extraction in the next winter (Lund et al. 2003).

Design, Costs

BHEs can be installed in nearly all kinds of geologic material. Design and installation must consider numerous influence factors. The sizing must consider the demand characteristics of the object to be supplied (size, extension; heating alone, or heating + domestic water, combined heating/cooling) as well as the local site conditions (climate, ground properties). The key ground property is the thermal conductivity of the ground surrounding the BHE. The higher the rock thermal

Geothermal Heat Pumps, Table 1 Borehole heat exchanger performance (single BHE, depth ≈ 150 m) in different rock types. (From Rybach 2001)

Rock type	Thermal conductivity ($\text{W m}^{-1} \text{K}^{-1}$)	Specific extraction rate (W per m)	Energy yield ($\text{kWh m}^{-1} \text{a}^{-1}$)
Hard rock	3.0	Max. 70	100–200
Unconsolidated rock, saturated	2.0	45–50	90
Unconsolidated rock, saturated	1.5	Max. 25	50

conductivity (W/m,K), the higher the specific heat extraction rate (W/m) and the energy yield (kWh/m,a) per unit BHE length (see Table 1). The installation and operation of a BHE/HP system needs installation and running costs. BHE drilling and installation costs amount to about 50 US\$ per meter in the USA and about 70 € per meter length in Europe. The operational cost is significantly less than with fossil-fueled systems. Currently, the price of heat and cold provided by GHP systems is comparable with conventional systems; GHP installation cost payback (the period of years before the savings in annual operating costs exceed the initial premium of an installation) ranges from 4 to 10 years, depending on the size and complexity of the system.

GHPs represent by far the fastest growing market in geothermal development and are also one of the fastest growing renewable technologies: the annual growth rate of globally installed capacities between 1995 and 2015 was 20% (Rybach and Sanner 2017).

Licensing, Environmental Benefits

In general, GHP installation needs permits. In most countries, the water protection agencies (local, regional, national) are providing permits. Regulation varies from country to country, in some cases even within the same country. Usually, some standard forms need to be filled and submitted. It is common

practice that in groundwater protection areas, the installation of GHP systems is limited or forbidden.

GHPs operate with little or no greenhouse gas (GHG) emissions since no burning processes are involved. GHPs driven by fossil-fueled electricity reduce the CO₂ emission by at least 50% compared with fossil-fuel-fired boilers. If the electricity that drives the geothermal heat pump is produced from a renewable energy source like hydropower or geothermal energy, the emission savings are up to 100%. Reducing current CO₂ emissions is the central option in all the efforts to mitigate global warming. Here, a clear distinction must be made between actual emission reduction and merely avoidance of additional emission: by new GHPs, only additional CO₂ emission can be avoided (“saving”), not a reduction of actual emissions. When GHPs are installed in refurbishment (to replace fossil-fueled systems), actual emission reduction can be achieved. Emission reduction is also evident when electric heater/cooler installations, driven by fossil-based electricity, are replaced by GHP systems.

Applications, Development Trends

There is a growing number of successful examples of GHP applications, for all kinds of buildings of highly varying sizes. The largest installation currently under construction is the geothermal store at the Bell State University (Indiana, USA) campus, with 4,100 BHEs, each 120 m deep (IGSHPA 2009). Several international organizations cover and/or promote the applications of GHPs. Here, only a few links are provided.

- International Ground Source Heat Pump Association (IGSHPA) <http://www.igshpa.okstate.edu/>
- Geothermal Heat Pump Consortium (GHPC) <http://www.geothermalheatpumpconsortium.org/>
- International Energy Agency Heat Pump Centre <http://www.heatpumpcentre.org/>
- European Heat Pump Association (EHPA) <http://www.ehpa.org/>

Besides, there are national associations active in many countries.

Most GHP systems operate with borehole heat exchangers (BHEs). In the early days of BHE applications, the boreholes have been drilled outside the objects to be served by the GHP installations. Nowadays, also dictated by the high prices of land, the boreholes are more and more located beneath the buildings to be erected.

Energy piles are now becoming increasingly popular. At many locations, the ground properties are such that load-

bearing rock formations occur only at some depth. Therefore, foundation piles need to be placed to secure firm standing of buildings. These piles can be equipped with heat exchanger pipes; the assemblage is then called “energy pile.” Special care must be taken to secure that the temperature changes caused by heat extraction and storage does not affect the pile/ground contact. A special handbook has been published in Switzerland that describes design, installation, and operation of energy pile systems (SIA 2005).

Summary

Currently, geothermal heat pumps (GHPs) represent the fastest growing branch in the utilization of geothermal energy. There are various factors that can lead to the successful dissemination of the GHP technology. Promoting factors are manifold: technical, economic, environmental awareness, governmental support, etc. The most important driving force is simply knowledge and know-how. Besides, various applications and wide-scale realizations, high-level quality assurance, and successful demonstration facilities are needed.

The use of GHP systems represents an important contribution of renewable energies to the mitigation of climatic warming: the increasing use of the environmentally friendly geothermal technology enables the saving of great amounts of fossil fuels, avoid corresponding CO₂ emissions and, in the case of the application of GHP systems in building renovation, it results in a significant emission reduction. It can be expected that GHP systems, along with other geothermal technologies (especially for power generation), will further increase the benefits to future generations.

Cross-References

- Heat Flow, Continental
- Lithosphere, Continental: Thermal Structure
- Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
- Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

Bibliography

- Banks D (2008) An introduction to thermogeology – ground source heating and cooling. Blackwell, Oxford, p 349
- EGEC (2009) Geothermal heat pumps – ground source heat pumps. Brussels, European Geothermal Energy Council, p 16
- IGSHPA (2009) GeoOutlook Magazine 6(2)
- Lund J, Sanner B, Rybach L, Curtis R, Hellström G (2003) Ground source heat pumps – a world review. *Renew Energy World* 6:218–227

- Rybach L (2001) Design and performance of borehole heat exchanger/heat pump systems. In: Proceedings of the European Summer School on Geothermal Energy Applications (CD-ROM), Oradea/Romania
- Rybach L (2005) The advance of geothermal heat pumps worldwide. IEA Heat Pump Center Newsl 23(4):13–18
- Rybach L, Sanner B (2017) Geothermal heat pump development trends and achievements in Europe. In: Bertani R (ed) Perspectives for geothermal energy in Europe. Imperial College Press, London, pp 215–253
- SIA (2005) Nutzung der Erdwärme mit Fundationspfählen und anderer erdberührten Betonebauteilen – Leitfaden zu Planung, Bau und Betrieb. SIA D 0190. Schweizerischer Ingenieur- und Architektenverein, Zurich. 101 p., ISBN 3-908483-58-1

Geothermal Record of Climate Change

Michael G. Davis¹, David S. Chapman² and Robert N. Harris³

¹Department of Physical Sciences, Arkansas Tech University, Russellville, AR, USA

²Department of Geology and Geophysics, The University of Utah, Salt Lake City, UT, USA

³College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA

Definition

Temperature changes at the Earth's surface diffuse slowly into the subsurface where they perturb the normal geothermal regime. The geothermal record of climate change utilizes precise temperature-depth measurements of these perturbed temperatures from boreholes to obtain a long-term, low-pass record of temperatures at the Earth's surface as interpreted using the one-dimensional heat diffusion equation.

Introduction

Because changes in temperature at the Earth's surface are transferred to the subsurface through the process of thermal diffusion, past changes in decadal and centennial surface temperature trends are now evident in subsurface temperatures. Thus, temperature-depth measurements in boreholes register not only the heat flowing out of the Earth but also transient departures from the background attributable to past surface temperature changes (e.g., Lachenbruch and Marshall 1986; Pollack and Chapman 1993; Pollack and Huang 2000; Beltrami 2002a; Harris and Chapman 2005). Identifying and interpreting these transient departures in terms of the past few hundred years of surface temperature change is now known as the geothermal method of climate reconstruction. Because thermal diffusion behaves as a low pass filter, such that high

frequency changes in surface temperature are attenuated more strongly and at shallower depths than longer-term changes, this method is most useful in identifying centennial scale climatic trends. The geothermal method of climate reconstruction has several benefits including (1) bypassing the empirical temperature calibration inherent with proxy records of climate change; (2) extending coverage to remote and widely dispersed continental sites where meteorological instrumental records are not available; and (3) establishing a surface baseline temperature prior to the Industrial Revolution when anthropogenic greenhouse gases started their unprecedented growth phase.

The Geothermal Method of Temperature Reconstruction

The geothermal method of temperature reconstruction is most easily understood in terms of the one-dimensional heat diffusion equation,

$$\frac{\partial T(z, t)}{\partial t} = \alpha \frac{\partial^2 T(z, t)}{\partial z^2}, \quad (1)$$

where T is temperature, z is depth, t is time, and α is thermal diffusivity. This equation shows how temporal variations in surface temperature lead to curvature in temperature-depth profiles. Time and depth are scaled nonlinearly through the thermal diffusivity such that a surface temperature change at time t prior to the present is largely captured within a depth, l , called the thermal length,

$$l = \sqrt{4\alpha t}. \quad (2)$$

Because the thermal diffusivity of rocks is about $1 \times 10^{-6} \text{ m}^2 \text{ s}^{-1}$, the majority of the past 100 years of surface temperature change is stored within the upper 150 m of the Earth; the majority of the last 1000 years of surface temperature change is captured in the uppermost 500 m. Careful analysis of curvature in the upper few hundred meters of a temperature-depth profile, therefore, can be used to reconstruct surface temperature change over the past millennium (e.g., Huang et al. 2000; Harris and Chapman 2001; Beltrami 2002b; Beltrami and Boulton 2004; Pollack and Smerdon 2004).

Useful analytical solutions to Eq. 1 for a temperature-depth profile resulting from simple functional changes of temperature at the Earth's surface include the following (Carslaw and Jaeger 1959; Lachenbruch and Marshall 1986; Chisholm and Chapman 1992):

- (a) single step ΔT change in surface temperature at time τ before present,

$$\Delta T(z) = \Delta T \operatorname{erfc}\left(\frac{z}{\sqrt{4\alpha\tau}}\right), \quad (3)$$

- (b) multiple step ΔT_i changes in surface temperature at respective times τ_i before present,

$$\Delta T(z) = \sum_{i=1}^n \Delta T_i \operatorname{erfc}\left(\frac{z}{\sqrt{4\alpha\tau_i}}\right), \quad (4)$$

- (c) a single ramp change in surface temperature starting at time τ before present and with amplitude ΔT when the temperature log is obtained,

$$\Delta T(z) = 4\Delta T i^2 \operatorname{erfc}\left(\frac{z}{\sqrt{4\alpha\tau}}\right), \quad (5)$$

where the complementary error function is denoted by erfc and $i^2 \operatorname{erfc}$ indicates the second integral of the complementary error function.

Interpretations of surface temperature change from the geothermal method of climate reconstruction are most confident when curvature in the temperature-depth profile can be shown to be transient and fully attributable to a changing surface temperature in the past. Because the Earth is not homogeneous, phenomena and processes other than a changing surface temperature can also cause curvature in temperature-depth profiles. Steady-state sources of curvature include variations in subsurface thermal conductivity, radioactive heat production, heat refraction due to topography, and differential solar insolation due to variations in slope and azimuth. Transient sources of curvature include changes in ground surface temperature in response to changes in surface air temperature (SAT), spatial changes in surface temperature around the borehole caused by temporal variations in albedo, non-isothermal groundwater flow, precipitation, and other microclimate effects (e.g., Beck 1982; Chisholm and Chapman 1992; Lewis and Wang 1992; Harris and Chapman 1995; Bartlett et al. 2004). Chisholm and Chapman (1992) and Harris and Chapman (1995) explore the magnitude of many of these effects quantitatively. While most studies of climate change inferred from borehole temperature profiles attempt to select sites that minimize these non-climatic sources of borehole temperature profile curvature, it is often difficult to partition curvature between steady-state and transient sources of curvature. Repeated temperature measurements in the same borehole can help improve the confidence in separating and interpreting steady-state and transient sources of curvature (e.g., Davis et al. 2010; Majorowicz et al. 2012).

Tutorial

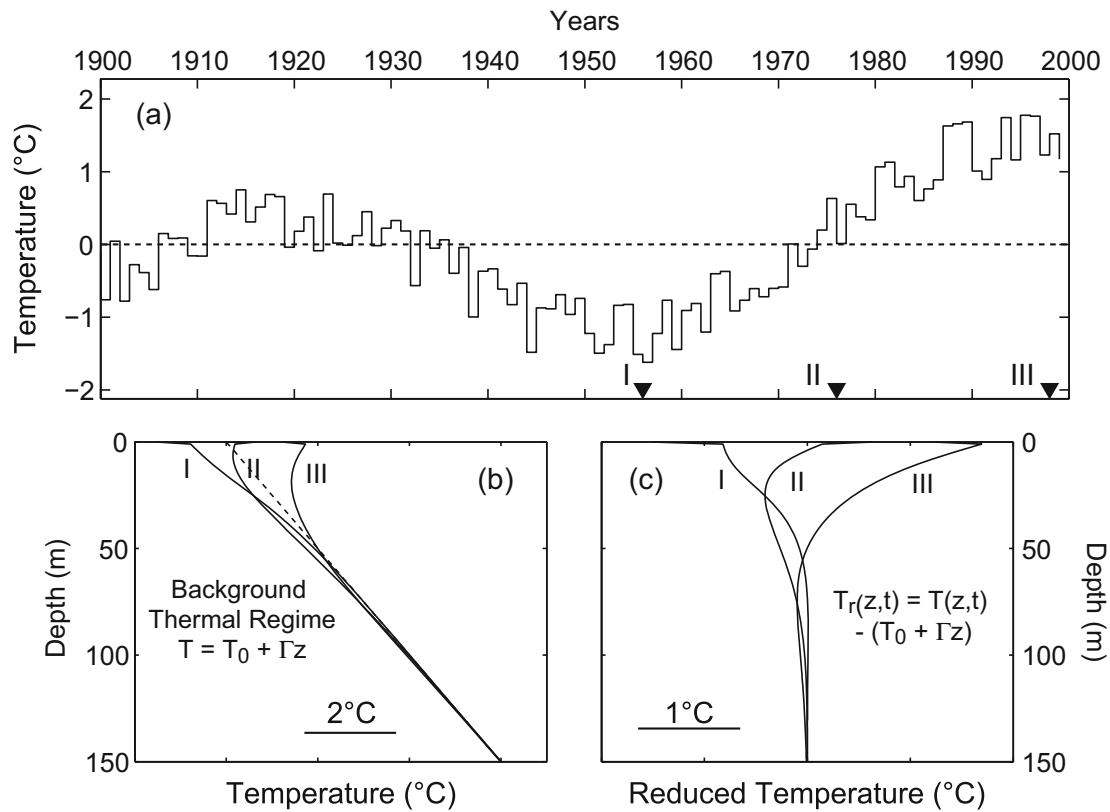
Important elements of the geothermal method of climate reconstruction are illustrated in Fig. 1. First, consider a

ground surface temperature (GST) that is essentially constant in time (dashed line; Fig. 1a). In this case, the measured borehole profile has no transient curvature, and the surface temperature intercept is equal to the GST (Fig. 1b). However, if the GST is not constant, but fluctuating in some way such that there is a deviation from the mean GST, it will impart curvature to the subsurface according to Eq. 1 (Fig. 1b). For interpretations of climate change based on temperature profiles, one is interested in the departures from the background thermal regime. To highlight these departures, we compute reduced temperatures expressed as

$$T_r(z) = T_{obs}(z) - (T_o + \Gamma z), \quad (6)$$

where T_o is the reducing intercept and Γ is the reducing gradient. In Fig. 1c, the background gradient has been removed, making the anomalous temperature strikingly apparent. The transient nature of the anomaly is illustrated by measuring the borehole temperatures at several times. The three temperature-depth profiles marked I, II, and III in Fig. 1b, c were measured at three different times shown by the solid triangles in Fig. 1a. Cooling from 1915 to 1955 creates a negative transient in the 1955 log I. Surface warming since 1955 results in warming features in the temperature-depth logs II and III that propagate deeper with time. This example also illustrates the low pass smoothing of thermal diffusion. While trends in subsurface curvature can be related to trends in GST, individual annual means are not apparent. Further, the resolution of the magnitude of GST change is much better resolved than the timing of the change, consistent with the physics of diffusion (Chisholm and Chapman 1992).

Because SATs represent the gold standard of temperature change, and because climatic proxies such as tree rings are empirically calibrated to SATs, a final step is often taken whereby derived changes in GST are related to SATs. Many borehole temperature-depth profiles are located in areas where SAT records exist, and the combination of datasets offers additional information about surface temperature change. Techniques to compare or use borehole temperatures with surface records include overlaying a GST reconstruction on an appropriate SAT curve (e.g., Huang et al. 2000) or using an appropriate SAT curve as a forcing function and comparing the resulting transient temperature depth curve to a reduced temperature profile (e.g., Harris and Chapman 2001). In this second technique, the best fitting initial temperature, sometimes called the pre-observational mean (POM; Chisholm and Chapman 1992; Harris 2007), is found by minimizing the fit between the reduced temperatures and the synthetic transient calculated from the SAT time series. An obvious uncertainty in this step is the nature of coupling between air and ground temperatures. Regional comparisons at the centennial time scale generally show good agreement between changes in air and ground temperature (Huang et al. 2000; Harris and Chapman 2001; Beltrami 2002b; Pollack and Smerdon 2004;



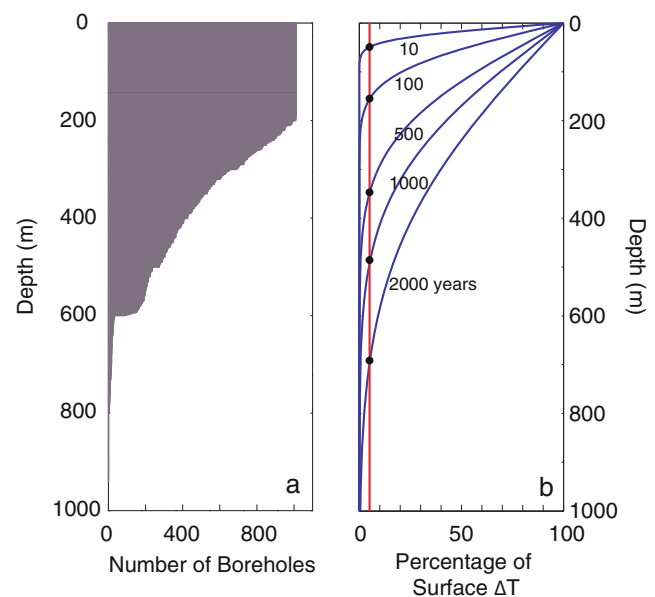
Geothermal Record of Climate Change, Fig. 1 Basic aspects of the geothermal method of climate reconstruction. (a) An arbitrary time series of surface temperatures for the last century. Dashed line represents a climatic history with no temperature change. Triangles indicate times of borehole temperature logs illustrated in (b) and (c); log I in 1955, log II in 1975, and log III in 1998. The method asserts that changes in ground surface temperature are

manifested as transient departures from the background, steady-state thermal regime for which a constant temperature gradient facilitates the conduction of heat from the Earth's interior. (b) Temperature-depth profiles for times I, II, and III. The linear profile results from the zero change scenario in (a)—dashed line. (c) Reduced temperature-depth profiles with background removed at times I, II, and III

Pickler et al. 2018). Modeling studies using general circulation models allow comparisons between changes in air and ground temperatures at longer time scales and also suggest good agreement between these measures (Gonzales-Rouco et al. 2003, 2006, 2009; Garcia-Garcia et al. 2016). Questions regarding changes in the relationship between air and ground temperatures have prompted detailed investigations often in combination with other meteorological parameters (e.g., Smerdon et al. 2003, 2004, 2006; Bartlett et al. 2006; Chudinova et al. 2006; Stieglitz and Smerdon 2007; Cheng et al. 2010). These studies have found that variations in air and ground temperatures generally track each other over the time period of study.

A Global Geothermal Climate Change Database

An archive of borehole temperature profiles at 1012 sites (Huang et al. 2000) suitable for studying climate change has been compiled at the University of Michigan and the National Centers for Environmental Information of the National Oceanic and Atmospheric Administration. Individually the borehole temperature profiles exhibit many characteristics of the



Geothermal Record of Climate Change, Fig. 2 (a) The number of boreholes in the global climate borehole database with temperature observations to a given depth (shaded). (b) Temperature versus depth at times following a step change ΔT in surface temperature. Solid dots mark depths where the signal is 5% of the surface change

meteorological record of climate change: hemispheric scale variation, regional variation, local effects, and noise. Collectively, however, borehole temperatures show clear evidence of significant twentieth-century warming and a potential for inferring longer-term climate change (Huang et al. 2000). Figure 2a shows a histogram of the number of boreholes in the global climate borehole database with temperature observations to a given depth. Boreholes at all 1012 sites extend to at least 200 m depth, one criterion for inclusion in the archive. Half of the boreholes are deeper than 350 m but fewer than 10% extend more than 600 m into the subsurface. The depth distribution of boreholes is important because the depth of borehole temperature anomalies is related to the timing of surface temperature changes (Eq. 2), as shown in Fig. 2b. The resolution of borehole temperature measurements can

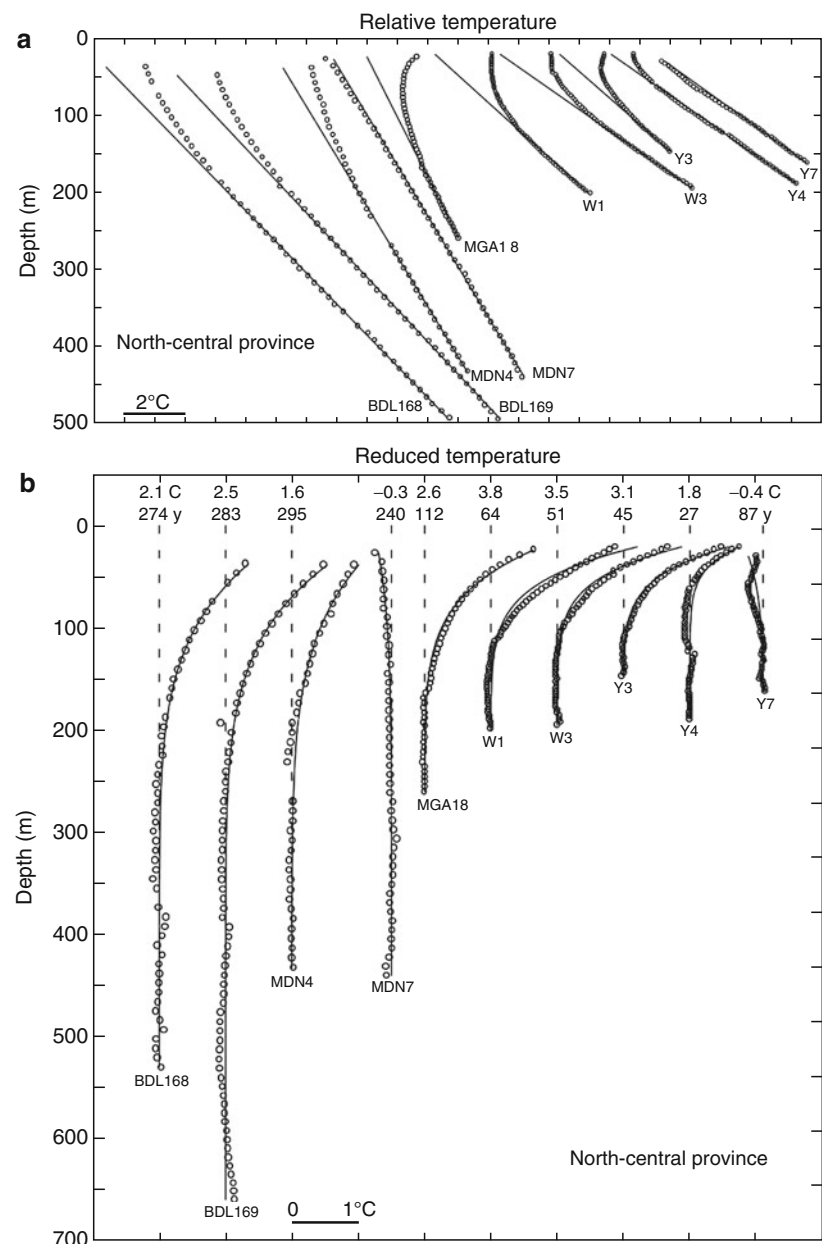
generally resolve a surface temperature signal that is greater than 5% of the surface temperature change. Figure 2b shows that to reconstruct past surface temperature adequately for the entirety of the last millennium, one must collect temperature-depth measurements down to at least 500 m depth. In practice these records must be deeper than 500 m so that the background thermal gradient and surface temperature intercept can be adequately estimated (Eq. 6).

Temperature-Depth Profiles

An ensemble of climatically perturbed temperature-depth profiles from India (Roy et al. 2002) is shown in Fig. 3a. The boreholes are from the north-central climatic province of

Geothermal Record of Climate Change, Fig. 3

(a) Plots of temperature-depth data (open circles) obtained at 10 selected sites from the north-central climatic province of the Indian peninsula. (Modified from Roy et al. 2002). The background temperature profiles (solid lines) are obtained by fitting a line just below the curved upper portion in each temperature-depth profile. Temperatures below 500 m are not displayed. (b) Reduced temperature profiles (open circles) for the 10 boreholes shown in (a). Each profile is offset to avoid overlap and plotted relative to zero reduced temperature (dashed line). Solid lines show the best fitting ramp models for the reduced temperature profiles. The magnitude and onset time of the ramp model are shown at the top of the profiles. (See Roy et al. (2002) for more details)



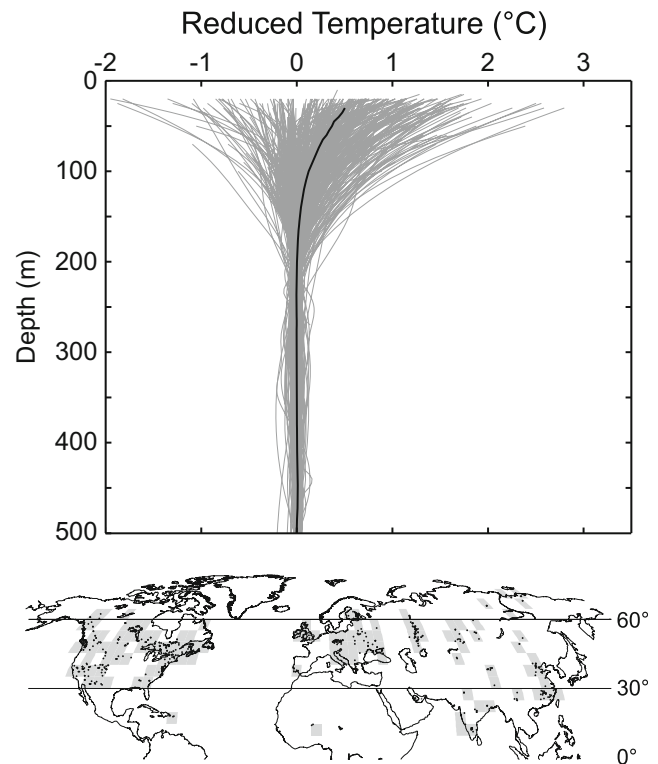
India. The first-order effect shown in these profiles is the general increase in temperature with depth, consistent with an upward flux of heat toward the Earth's surface. The deepest part of the borehole temperature profiles has a constant background thermal gradient. Figure 3b shows the reduced temperature profiles for these sites after the background gradient and surface temperature intercept have been removed (Eq. 6), illustrating the transient temperature departures in the uppermost 200 m of the boreholes. The transient departure from a steady-state background signal is most consistent with a ramp change of amplitude -0.3 to 3.8 °C starting 27 to 295 years ago. This example is notable because it shows the predominantly warming signal shown in ground temperatures consistent with surface temperature warming over the past two centuries, as well as the site-specific climatic conditions recorded in subsurface temperatures.

Repeat Temperature Logs

The regional analysis of Davis et al. (2010) employed multiple repeat temperature-depth profiles from three individual boreholes in northwestern Utah over nearly 30 years to provide observational constraints for understanding the relationship between ground and air temperatures. A key part of the study was to calculate and compare synthetic temperature profiles from nearby SAT records to the observed temperature profiles. The synthetic profiles closely fit observed temperature-depth profiles and match both the amplitude and pattern of subsurface transient changes. Further, forward modeled ground temperatures and differences between temperature logs isolate transient variations and indicate that the observed profiles can be ascribed to changes in GST (Davis et al. 2010). Such observations of transient temperatures in boreholes offer strong support for using GST histories to complement SAT data and multi-proxy reconstructions in climate change studies.

Temperature Reconstruction Analysis

Several groups have used boreholes to analyze past temperatures on both regional (e.g., Lachenbruch and Marshall 1986; Beltrami and Mareschal 1991; Chisholm and Chapman 1992; Harris and Chapman 1995; Roy et al. 2002; Davis et al. 2010; Pickler et al. 2018) and global (e.g., Huang et al. 2000; Harris and Chapman 2001; Beltrami 2002b) scales. The temperature reconstruction of Huang et al. (2000) utilized the global database and a Bayesian estimation technique, similar to the functional space inversion of Shen and Beck (1991). The method parameterized the reconstruction into century-long rates over the past 500 years, resulting in a 1.1 °C increase over the entire period (Huang et al. 2000). Further analysis of this dataset partitioned the borehole into both hemisphere and



Geothermal Record of Climate Change, Fig. 4 (Top) Northern Hemisphere reduced temperature profiles from Harris and Chapman (2001) and Roy et al. (2002). Lines represent 509 profiles whose locations are shown on the map (bottom). Bold line shows average reduced temperature profile. (Figure modified from Harris and Chapman (2001))

continental reconstructions. Huang et al. (2000) found that warming was greatest in the Northern Hemisphere, and greatest in the twentieth century.

Another analysis used a Northern Hemisphere subset of the global database (Harris and Chapman 2001) to obtain an average reduced temperature profile (Fig. 4). The surface temperature record was then used as the forcing function, calculating the best fitting synthetic profile. This process requires determination of a baseline or POM temperature prior to the beginning of the surface temperature record (Chisholm and Chapman 1992; Harris 2007). The result, which combined the POM and the instrumental surface temperature record, indicated that there was approximately 1 °C of ground warming over the past 100–200 years (Harris and Chapman 2001), consistent with proxy methods of paleoclimate reconstruction (e.g., NRC 2006, and references therein) and comparable to other borehole analyses (Huang et al. 2000; Beltrami 2002b).

Summary

Temperature-depth profiles measured in boreholes are an important record of surface temperature change and can be summarized as follows:

1. Thermal diffusion in the solid Earth results in low pass filtering of temperature changes at the Earth's surface. Whereas annual mean temperatures are not resolved, the curvature in the temperature-depth profiles in boreholes reveals a surface temperature history at the site. The depth distribution of borehole temperature anomalies is related to the timing of surface temperature changes.
2. Both globally and regionally, ground temperatures show a predominantly warming signal consistent with surface temperature warming over the past two centuries.
3. Repeated temperature measurements in the same borehole can help improve the confidence in separating and interpreting steady-state and transient sources of curvature.
4. Geothermal climate change reconstruction techniques show approximately 1 °C of ground warming over the past one to two centuries, consistent with proxy methods of paleoclimate reconstruction.

Cross-References

- [Heat Flow Determinations, Continental](#)
- [Lithosphere, Continental: Thermal Structure](#)

Bibliography

- Bartlett MG, Chapman DS, Harris RN (2004) Snow and the ground temperature record of climate change. *J Geophys Res* 109:F04008. <https://doi.org/10.1029/2004JF000224>.
- Bartlett MG, Chapman DS, Harris RN (2006) A decade of ground-air temperature tracking at Emigrant Pass Observatory, Utah. *J Clim* 19(3):722–3,731
- Beck AE (1982) Precision logging of temperature gradients and the extraction of past climate. *Tectonophysics* 83:1–11
- Beltrami H (2002a) Earth's long-term memory. *Science* 297:206–207
- Beltrami H (2002b) Climate from borehole data: energy fluxes and temperatures since 1500. *Geophys Res Lett* 29(23):2111. <https://doi.org/10.1029/2002GL015702>.
- Beltrami H, Bourlon E (2004) Ground warming patterns in the northern hemisphere during the last five centuries. *Earth Planet Sci Lett* 227:169–177
- Beltrami H, Mareschal JC (1991) Recent warming in Eastern Canada inferred from geothermal measurements. *Geophys Res Lett* 18:605–660
- Carslaw HS, Jaeger JC (1959) *Conduction of heat in solids*, 2nd edn. Oxford University Press, New York
- Cheng Y, Stieglitz M, Pan F (2010) Simple method to evolve daily ground temperatures from surface air temperatures in snow-dominated regions. *J Hydrometeorol* 11:1395–1404
- Chisholm TJ, Chapman DS (1992) Climate change inferred from analysis of borehole temperatures; an example from western Utah. *J Geophys Res* 97:14,155–14,175
- Chudnina SM, Frauenfeld OW, Barry RG, Zhang T, Sorokovikov VA (2006) Relationship between air and soil temperature trends and periodicities in the permafrost regions of Russia. *J Geophys Res* 111:F02008. <https://doi.org/10.1029/2005JF000342>.
- Davis MG, Harris RN, Chapman DS (2010) Repeat temperature measurements in boreholes from northwestern Utah link ground and air temperature changes at the decadal time scale. *J Geophys Res* 115:12. <https://doi.org/10.1029/2009JB006875>
- García-García A, Cuesta-Valero FJ, Beltrami H, Smerdon JE (2016) Simulation of air and ground temperatures in PMIP3/CMIP5 last millennium simulations: implications for climate reconstructions from borehole temperature profiles. *Environ Res Lett* 11:044022. <https://doi.org/10.1088/1748-9326/11/4/044022>
- Gonzalez-Rouco JF, von Storch H, Zorita E (2003) Deep soil temperature as proxy for surface air-temperature in a coupled model simulation of the last thousand years. *Geophys Res Lett* 30:2116–2119
- Gonzalez-Rouco JF, Beltrami H, Zorita E, von Storch H (2006) Simulation and inversion of borehole temperature profiles in surrogate climates: spatial distribution and surface coupling. *Geophys Res Lett* 33:L01703. <https://doi.org/10.1029/2005GL024693>
- Gonzalez-Rouco JF, Beltrami H, Zorita E, Stevens MB (2009) Borehole climatology: a discussion based on contributions from climate modeling. *Clim Past* 5:97–127
- Harris RN (2007) Variations in air and ground temperature and the POM-SAT model: results from the Northern Hemisphere. *Clim Past* 3:611–621
- Harris RN, Chapman DS (1995) Climate change on the Colorado Plateau of eastern Utah inferred from borehole temperatures. *J Geophys Res* 100:6367–6381
- Harris RN, Chapman DS (2001) Mid-latitude (30°–60°N) climatic warming inferred by combining borehole temperatures with surface air temperatures. *Geophys Res Lett* 28:747–750
- Harris RN, Chapman DS (2005) Borehole temperatures and tree-rings: seasonality and estimates of extratropical Northern Hemispheric warming. *J Geophys Res* 110:F04003. <https://doi.org/10.1029/2005JF000303>.
- Huang SP, Pollack HN, Shen P-Y (2000) Temperature trends over the past five centuries reconstructed from borehole temperatures. *Nature* 403:756–758
- Lachenbruch AH, Marshall BV (1986) Climate change: geothermal evidence from permafrost in the Alaskan Arctic. *Science* 234:689–696
- Lewis TJ, Wang K (1992) Influence of terrain on bedrock temperatures. *Glob Planet Chang* 6:87–100
- Majorowicz J, Skinner W, Šafanda J (2012) Western Canadian Sedimentary Basin temperature–depth transients from repeated well logs: evidence of recent decade subsurface heat gain due to climatic warming. *J Geophys Eng* 9:127–137
- National Research Council (2006) *Surface temperature reconstructions for the last 2,000 years*. The National Academies Press, Washington, DC, 145 pp
- Pickler C, Gurza Fausto E, Beltrami H, Mareschal J-C, Suárez F, Chacon-Oecklers A, Blin N, Cortes MT, Montenegro A, Harris R, Tassara A (2018) Recent climate variations in Chile: constraints from borehole temperature profiles. *Clim Past* 14:559–575
- Pollack HN, Chapman DS (1993) Underground records of changing climate. *Sci Am* 268(6):44–50
- Pollack HN, Huang S (2000) Climate reconstruction from subsurface temperatures. *Annu Rev Earth Planet Sci* 28:339–365
- Pollack HN, Smerdon JE (2004) Borehole climate reconstructions: spatial structure and hemispheric averages. *J Geophys Res* 109:D11106. <https://doi.org/10.1029/2003JD004163>.
- Roy S, Harris RN, Rao RUM, Chapman DS (2002) Climate change in India inferred from geothermal observations. *J Geophys Res* 107(B7):2138. <https://doi.org/10.1029/2001JB000536>
- Shen P-Y, Beck AE (1991) Least squares inversion of borehole temperature measurements in functional space. *J Geophys Res* 96:19,965–19,979
- Smerdon JE, Pollack HN, Enz JW, Lewis MJ (2003) Conduction-dominated heat transport of the annual temperature signal in soil. *J Geophys Res* 108(B9):2431. <https://doi.org/10.1029/2002JB002351>
- Smerdon JE, Pollack HN, Cermak V, Enz JW, Kresl M, Šafanda J, Wehmler JF (2004) Air-ground temperature coupling and

- subsurface propagation of annual temperature signals. *J Geophys Res* 109:D21107. <https://doi.org/10.1029/2004JD005056>.
- Smerdon JE, Pollack HN, Cermak V, Enz JW, Kresl M, Safanda J, Wehmiller JF (2006) Daily, seasonal and annual relationships between air and subsurface temperatures. *J Geophys Res* 111: D07101. <https://doi.org/10.1029/2004JD005578>.
- Stieglitz M, Smerdon JE (2007) Characterizing land-atmosphere coupling and the implications for subsurface thermodynamics. *J Clim* 20:21–37. <https://doi.org/10.1175/JCLI3982.1>

GPS, Data Acquisition, and Analysis

Carine Bruyninx¹, Wim Aerts² and Juliette Legrand¹

¹Department of Reference Systems and Planetology, Royal Observatory of Belgium, Brussels, Belgium

²KU Leuven, Electrical Engineering (ESAT) TC, Diepenbeek Campus, Diepenbeek, Belgium

Definition

GPS, Global Positioning System, is a US satellite-based navigation system which can accurately determine position, velocity, and time in a common reference system, anywhere on or near the Earth on a continuous basis.

Other Satellite-Based Navigation Systems

Next to GPS, several other global satellite-based navigation systems (GNSS) are in operation. Examples include Europe's Galileo (2009), Russia's Global'naya Navigatsionnaya Sputnikovaya Sistema (GLONASS, Russian Space Systems 2016), and China's BeiDou Navigation Satellite System (China Satellite Navigation Office 2019). Each of the systems emit a specific set of radio signals and all provide positioning, navigation, and timing capabilities following the same principles as described in the section “[Satellite Positioning](#).” We will here concentrate on GPS, but its principles can also be applied to the other GNSS or a combination of several GNSS. A detailed overview of GNSS is provided in Teunissen and Montenbruck (2017).

Satellite Positioning

Point Positioning

Satellite-based point positioning consists in determining the three-dimensional position (X, Y, Z) of a receiver, using the radio signals emitted by a number of satellites with known position. A receiver, usually located on the Earth, measures the travel time τ of the radio signal. Then it reconstructs the distance D between the receiver and the satellite by

multiplying the travel time τ by the velocity of the radio signals. In the ideal case, neglecting the influence of the Earth's atmosphere, radio signals travel with the speed of light c (299,792,458 m/s) and the distance reads $D = \tau \cdot c$.

Because the receiver and satellite clock are not perfectly synchronized, the measured travel time is corrupted by synchronization errors causing an error on the measured distance with respect to the true satellite-receiver distance. The unknown synchronization error dT between the satellite clock and the receiver clock then becomes the fourth unknown that the receiver has to determine, next to the three-dimensional (X, Y, Z) receiver position. Consequently, to solve this problem with four unknowns, at least four satellites have to be tracked simultaneously.

Relative Positioning

Differential or relative satellite positioning is a technique developed to improve the accuracy of satellite point positioning by combining the measured travel time at different receivers. The basic principle consists in reducing the biases at a “remote receiver” using measured biases at a fixed “reference receiver” with known coordinates. This allows removal of the errors common to the travel times measured at both stations. The position of the remote receiver (or the baseline vector) is then determined relative to the reference receiver. The corrections may be used in real time or later, using post-processing techniques. The remote receiver can be stationary or moving.

System Description

Introduction

Development work on the NAVSTAR/GPS (Navigation by Timing And Ranging/Global Positioning System) started in 1973, when the US Department of Defense (DoD) applied its technical resources to develop a highly accurate, space-based navigation system to replace the TRANSIT system.

The first GPS satellite was launched on February 22, 1978, and GPS Full Operational Capability (FOC) was declared in July 1995. FOC status means that the system meets all the requirements specified in a variety of formal performance and requirements documents. The system consists of three major segments: a space segment, satellites that transmit radio signals; a ground segment, ground-based equipment to monitor the satellites and update their signals; and a user segment, equipment which passively receives the satellite signals and converts satellites signals into position and navigation information.

Satellites, Signals, and System Time

The GPS is designed as a constellation of 24 satellites distributed among 6 orbital planes of which the inclination

with respect to the equator is 55° . The orbital planes have a radius of 26,600 km. Each satellite makes two complete orbits each sidereal day. With this constellation, one can always see minimum four satellites from any point on the Earth. In the original GPS design, the satellite signal consists of:

- The L1 and L2 carrier waves, with a respective wavelength of 19 and 24 cm.
- Information, encoded in the form of bits on the carrier signals by a process known as phase modulation:
 - The Coarse/Acquisition code or C/A, which is freely available to the public, is a binary code sequence which repeats every 1 ms. It is transmitted on the L1 carrier at 1.023 Mbps (million bits per second) using a binary-phase shift keying (BPSK) modulation technique. The “chip length,” or physical distance between binary transitions (between digits +1 and −1), is 293 m. Each satellite has a different C/A code, so that it can be uniquely identified.

The Precision code or P-code is transmitted on both the L1 and L2 frequencies at 10.23 Mbps with a chip length of 29.3 m and using the same BPSK modulation technique. The P-code repeats every 7 days, and like for the C/A-code, each satellite has a unique P-code. Whereas C/A is a coarser code appropriate for initially locking onto the signal, the P-code is designed for more precise positioning. Though the P-code is not classified, since January 31, 1994, a process called “anti-spoofing” encrypts the P-code in order to prevent an enemy imitating the GPS signal. To that end, the P-code was modulated with the W-code, a special encryption sequence, to generate the Y-code.

- The navigation message is modulated at 50 bit/s on the L1 wave on top of both ranging codes. It is sent in blocks of 1500 bits (taking 30 s to transmit). One message block is subdivided in three parts or five subframes. The first part, or first subframe, contains GPS date and time, plus the satellite’s status and an indication of its health. The second part, consisting of the second and third subframe, contains orbit parameters (often called the “broadcast ephemeris”) and allows the receiver to calculate the position of the satellite in the earth-fixed WGS 84 system defined in IS-GPS-200J (2018). The last part, consisting of the fourth and fifth subframes, is partly for military use but also contain information on the ionosphere (8 coefficients of the Klobuchar model, see Klobuchar (1986)), UTC data (Universal Time Coordinated), and the almanac data part. The almanac data contain information and status concerning all the satellites, their locations, and PRN (Pseudo Random Noise) numbers. The first and second

parts are identical in all transmitted 30-s blocks. The third part differs as it contains only 1/25 of the entire almanac. Hence, to receive the almanac of the entire constellation, a total of 25 blocks must be received, which takes 12.5 min.

The GPS signal is transmitted with enough power to ensure a minimum signal power level of -160 dBW at the Earth’s surface. The orientation of the satellites is always changing, such that the solar panels face the sun, and the antennas face the center of the Earth. For more details on the GPS signals and the navigation message, refer to the GPS Interface Specification (IS-GPS-200J 2018).

New demands on the existing GPS system pushed for a gradual modernization with newer generations of GPS satellites containing satellite signals not present in the original design. The GPS satellites in operation in 2019 are:

- Block IIR: launched after 1996, they have a lifetime of 10 years. They contain three rubidium atomic clocks.
- Block IIR-M: launched since 2005, these satellites transmit a second civil signal L2C on the L2 frequency and the military M signal on the L1 and L2 frequencies. The L2C signal aims at improving navigation accuracy, providing an easy to track signal, and acting as a redundant signal in case of localized interference. The satellites contain three rubidium atomic clocks.
- Block IIF: launched since 2010, the satellites emit a new L5 civil signal providing a second safety-of-life signal that meets the technical requirements for enabling aircraft to make precision landings in high multipath environments. The satellites carry four atomic clocks (cesium and rubidium).
- Block III: will carry a new L1C civil code, which will be transmitted on L1 in addition to current signals. This seeks to maximize interoperability with the Open Service signal of the European Galileo satellite navigation system. Further, Japan’s Quasi-Zenith Satellite System (QZSS) will transmit a signal with virtually the same design as L1C. A first Block III test satellite has been launched end of 2018.

GPS system time is a paper clock based on the atomic clocks in the GPS ground control stations and the GPS satellites themselves. The GPS Control Segment maintains GPS system time to within $1 \mu\text{s}$ of UTC(USNO) (IS-GPS-200J 2018) which is the realization of the UTC kept by the USNO (United States Naval Observatory, Washington, USA). However, GPS system time is continuous and does not follow the UTC leap-seconds. UTC(USNO) itself is kept very close to the international benchmark UTC as maintained by the Bureau International des Poids et Mesures in Paris.

GPS Ground Instrumentation

High-end GPS ground instrumentation generally consists of an antenna with hemispherical radiation pattern, connected to a receiver through a coaxial cable. GPS devices for the consumer market generally use small integrated antennas, resulting in a single portable device. For fixed GPS stations, portability is not an issue, but performance in terms of accuracy is. Consequently for this application, better, but larger, antennas and more complex, but more power consuming, receivers are used. Only the high-end instrumentation relevant for scientific use is discussed in this entry.

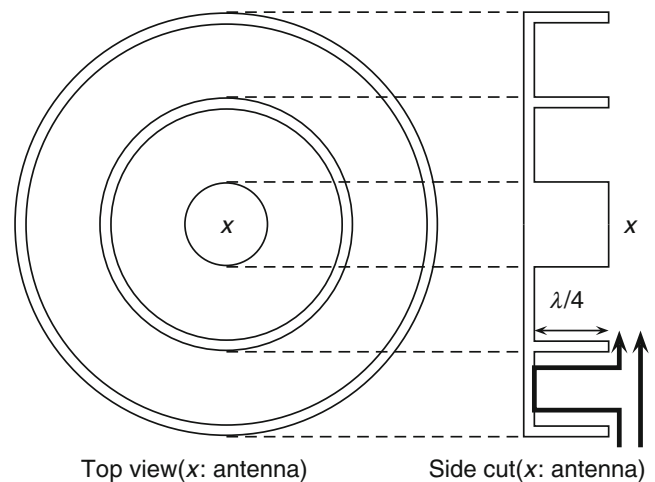
Antenna

Most GPS antennas in consumer electronics solely receive the primary L1 frequency. As a second frequency, however, allows for, e.g., correction of ionospheric delays which was already intended in the original system design (Spilker 1978), antennas for high-precision positioning will receive on both L1 and L2. From 2010 onwards, such antennas typically also receive the L5 frequency. This frequency resides in a band that is legally accredited for aeronautical use and hence enables dual frequency positioning in the world of aviation. Typically, these high-precision antennas will be mounted on locations with clear sky view in order to be able to track as many satellites as possible. This implies that the cable connecting the antenna to the receiver, somewhere inside a building, will have a length of several meters. To compensate for the signal loss in this cable, most high-precision antennas have a built-in low-noise amplifier (LNA). In most cases, the LNA is fed by a DC voltage put on the antenna cable using a bias-T, either inside the receiver or added in between antenna and receiver as an additional component.

As the fields emitted by the satellites are Right-Hand Circularly Polarized (RHCP) (IS-GPS-200J 2018), any antenna capable of receiving this polarization is able to receive the GPS signals. These can be wire antennas, such as helix antennas or crossed dipoles (e.g., the Dorne Margolin antennas included by many manufacturers), or circular patches.

In many cases, choke rings are added to the antenna in order to suppress the reflections (also known as multipath). These choke rings are cavities with a depth of approximately a quarter of a wavelength. As a consequence, multipath waves, generally coming in at negative elevation angles, will partly enter the cavities. After reflecting in the cavities, they will cancel out with those multipath waves that did not enter the cavity, thanks to a difference in path length of half a wavelength. This working principle is illustrated in Fig. 1. The concept of choke rings was first introduced by Jet Propulsion Laboratory (Tranquilla et al. 1989).

As the choke ring is typically a narrow band design, its performance can never be perfect on all (L1, L2 and L5)



GPS, Data Acquisition, and Analysis, Fig. 1 Working principle of the choke ring ground plane

frequencies. Therefore, many design improvements such as dual depth choke rings were proposed. Another problem linked with the antenna being multiband is the considerable variation of the phase center, which is the physical point where the signal is received. Fortunately, this can be measured and compensated for through an appropriate calibration procedure (Cohen and Parkinson 1991).

Receiver

A generic GPS receiver consists of a radio frequency (RF) part that demodulates the messages transmitted by the GPS satellites, and a positioning part that, based on the information received from the RF part, determines the position of the user.

Most GPS receivers allow storage of some variables (such as position and speed), and status information (such as system errors), and observables (such as code and phase measurements) in log files for further processing, mostly on removable memory cards. In combination with battery feeding, this should allow for autonomous operation for several days.

The RF block in the receiver demodulates the radio signals. In contemporary receivers, the RF block is made up of three separate RF front ends or receiving chains, in order to demodulate the signals of all satellites in view on the three different frequencies L1, L2, and L5. These signals are received by a single antenna, meaning that the receiver essentially gets a signal from the antenna that is the sum of all satellite signals on all carrier frequencies. Separation in the frequency domain is achieved by mixing down with different local oscillator (LO) signals in the different RF front ends. In order to distinguish between the signals from different satellites on a certain carrier frequency, the receiver uses code division multiplexing (CDM) techniques (Braasch and Van Dierendonck 1999). First, the receiver duplicates the antenna signal on a certain frequency into several identical copies.

Each of the identical copies will be sent into a different channel of the receiver. In every single receiving channel, a signature from a different GPS satellite will be sought for and the navigation message and the propagation delay of this satellite will be extracted. For cost reasons, this duplication is performed after down conversion of the original signal to a lower intermediate frequency (IF) and digitalization with an analog-to-digital Converter (ADC). Indeed, a digital signal can easily be copied to feed the distinct channels. Contemporary receivers have about 400 such channels.

Using a trial and error procedure by varying Doppler shift, carrier phase, and PRN time delays, the following steps are performed in each of the channels. The first step consists in removing the carrier by a numerical (digital) mixing of the IF signal with an internally generated carrier. The remaining signal only contains the PRN code and data bits. This signal is correlated with the satellite-specific PRN code generated by the receiver for varying time delays. In Fig. 2, the correlation with the received signal as a function of the time shift of the internally generated PRN is shown. Once the correlation of the internally generated replica of the PRN and the received signal succeeds, the Doppler shift, carrier phase, and code time delay of a certain satellite signal is determined. The channel is then said to be locked on to that specific satellite. Else, another PRN is tried.

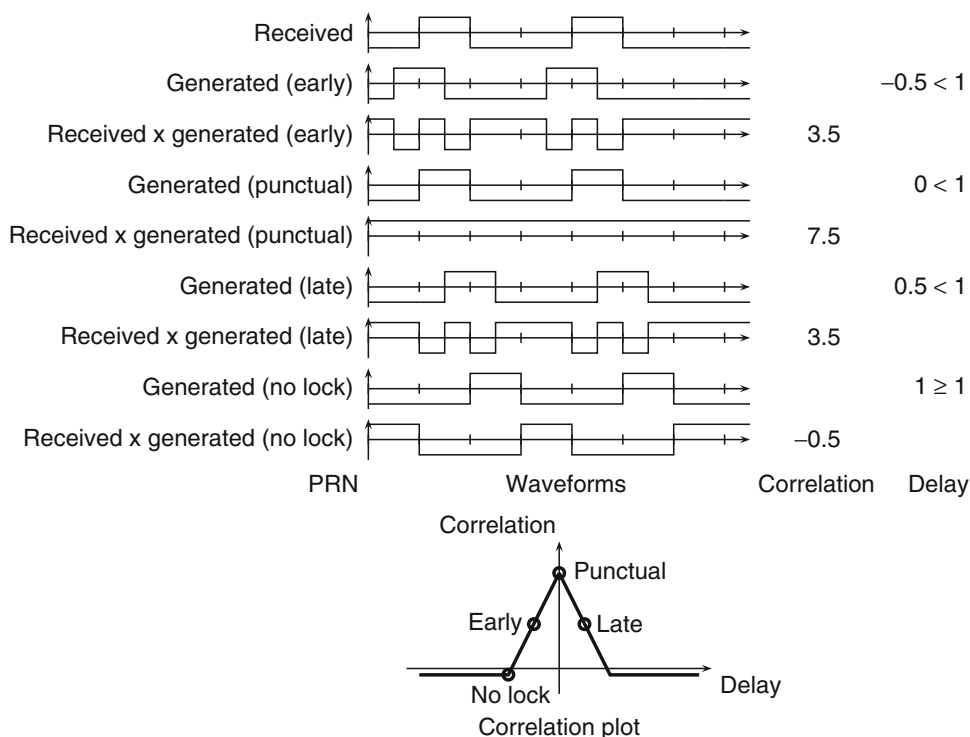
Obviously, the more stable the receiver internal clock (mostly a quartz crystal) is, the more accurately the Doppler shift, carrier phase, and code delay can be determined.

However, due to systematic errors, the timescale of the receiver's clock deviates from the GPS timescale. If the difference between both approaches 1 ms, the receiver's time is resynchronized to the GPS timescale, resulting in a time jump, but assuring time correspondence of measurements from different receivers. In addition, signal noise and noise in the channel control loops will cause minor errors in the determination of these variables. These errors, referred to as measurement noise, will limit the precision of the position determination.

Only the PRNs of the C/A code, sent on L1, are publicly known. The encrypted Y-code's PRNs cannot be generated by the civil GPS receivers. Consequently, the civil receivers cannot receive the messages that were encoded with the military CDM. However, as the Y-code is sent on both L1 and L2, techniques of cross-correlating the signals on both frequencies are used to still allow determination of difference in arrival time of the P-code at both frequencies (Woo 1999).

If the channel is locked to a satellite, the correlation between the received satellite and the internally generated PRN code sequence will be strong. Then, in order to decode the navigation message, the sign of this correlation is investigated. If the correlation is strongly positive, a 1 was sent by the satellite; if it is negative, a 0 was sent. This way, the bits sent by the satellite can be demodulated. The demodulated navigation message, together with the Doppler shift, carrier phase, and pseudo range of the channel are then passed to the positioning block that will combine all information from all

GPS, Data Acquisition, and Analysis, Fig. 2 Correlation with the received PRN as a function of time shift of the internally generated PRN



channels to determine the receiver's position. Besides using the information from the RF block, the positioning block of most contemporary receivers can also take into account signals from correction streams from neighboring GPS receivers (see section “Positioning Modes”) or inertia sensors.

Measurements

The basic principle of the GPS measurement is that the GPS receiver determines the time lapse for each satellite between the signal emission and the signal reception (see the section “Receiver” for more details). By multiplying this time difference with the signal travel speed, the distance between the GPS receiver and emitting antenna of each locked satellite can be reconstructed. As will be shown later, this measured distance does not correspond with the geometric distance separating the receiver and satellite and is therefore called “pseudo range.” As already outlined in the section “Receiver,” two main types of measurements can be done by the RF block of the GPS receiver:

- Range observations, based on the time shift of the internally generated PRN codes, also called “pseudo-range” measurements.
- Carrier phase observations, based on the phase shift in the numerical IF mixing step in each receiving channel. They are more precise than the pseudo-range type measurements but have a much higher degree of “ambiguity” than the code ranges.

Code Pseudo-range Model

The GPS code pseudo-range P_i^k measurement at receiver i to satellite k can be expressed according to the following simplified equation:

$$P_i^k = \rho_i^k + c \cdot (dt_i - dt_i^k) + I_i^k + T_i^k + \varepsilon_{P_i}^k \quad (1)$$

where ρ_i^k is the geometric range from receiver i to satellite k , dt_i and dt_i^k are, respectively, the receiver and satellite clock offsets from GPS system time, while I_i^k and T_i^k take respectively into account the effect of the ionosphere and troposphere on the signal. Finally, $\varepsilon_{P_i}^k$ stands for the measurement noise components, including multipath.

Using Pythagoras Theorem, we know that:

$$\rho_i^k = \sqrt{(x^k - x_i)^2 + (y^k - y_i)^2 + (z^k - z_i)^2} \quad (2)$$

with $\vec{x}_i = (x_i, y_i, z_i)$ and $\vec{x}^k = (x^k, y^k, z^k)$ respectively the receiver and satellite positions (at signal transmission time).

For positioning or navigation purposes, the user wishes to determine the receiver position $\vec{x}_i$. As Eq. 2 is a nonlinear

equation, it is linearized around the a priori receiver position $\vec{x}_{i0} = (x_{i0}, y_{i0}, z_{i0})$:

$$P_i^k(t_i) = \frac{(x_{i0} - x^k)}{\rho_{i0}^k} dx_i + \frac{(y_{i0} - y^k)}{\rho_{i0}^k} dy_i + \frac{(z_{i0} - z^k)}{\rho_{i0}^k} dz_i + c \cdot (dt_i - dt_i^k) + I_i^k + T_i^k + \varepsilon_{P_i}^k \quad (3)$$

where (dx_i, dy_i, dz_i) denotes the deviation of the unknown receiver position with respect to its a priori value and ρ_{i0}^k the geometric range from receiver i to satellite k computed using the a priori receiver position.

For most applications, the satellite position $\vec{x}^k$ and clock synchronization error dt_i^k can be considered as known. This leaves the following unknowns (dx_i, dy_i, dz_i, dt_i) in Eq. 3. As previously mentioned, a minimum of four GPS satellites has to be observed simultaneously to solve for these four unknowns.

Carrier Phase Model

The GPS carrier phase Φ_i^k measured at receiver i to satellite k can be modelled according to the following simplified equation (expressed in [m]):

$$\Phi_i^k = \rho_i^k + c \cdot (dt_i - dt_i^k) + \lambda N_i^k - I_i^k + T_i^k + \varepsilon_{\Phi_i}^k \quad (4)$$

This equation differs from the pseudo range (Eq. 1) as follows:

N_i^k the initial cycle ambiguity: an integer associated with the first measurement (fractional phase) of a satellite track (or after a tracking interruption). After the initial measurement (and as long as there is no tracking interruption), the receiver keeps track of the cumulative phase differences reflecting the changes in distance to the satellite.

λ the wavelength of the carrier wave (L1, L2, L5 or any of their linear combinations).

$\varepsilon_{\Phi_i}^k$ the measurement noise of the phase observable, including phase multipath.

But this equation can be linearized in exactly the same way as the code pseudo-range equation (Eq. 1).

Remarks

The following comments can be made with regard to Eqs. 1 and 4:

- These equations are generic and several of the terms in the equations differ depending on the observed code pseudo-range (C/A, P1, or P2) or carrier phase (L1, L2, or L5) signal.
- High-end GPS receivers are programmed to measure GPS satellites simultaneously at the epoch t_i (e.g., at the full

second). Even if the reception times are identical, respective signals leave the satellites at slightly different times because the satellite-receiver distances differ. The signal transmission time t_i^k corresponding to the reception time becomes thus receiver-dependent (hence the use of the double index). The receiver clock error is identical in all simultaneous measurements at receiver i from all satellites. Using simultaneous code pseudo-range measurements to at least four satellites, the absolute position of the receiver can be computed rather straightforward. This is not possible when using carrier phases as the initial cycle ambiguity is an additional satellite-receiver-dependent unknown which has to be dealt with.

- Note the different sign for the ionospheric effect in Eqs. 1 and 4. The ionosphere, being a dispersive medium, acts differently on the group velocity (code) than the phase velocity and is consequently delaying the code measurements while it advances the phases.

Single and Double Differences

For relative positioning between two stations (i, j) , the observations made simultaneously to the same satellite k are differenced:

$$\Delta P_{ij}^k \triangleq P_i^k - P_j^k = \Delta \rho_{ij}^k + c \cdot \Delta t_{ij} + \Delta I_{ij}^k + \Delta T_{ij}^k + \Delta \epsilon_{Pij}^k \quad (5)$$

$$\begin{aligned} \Delta \Phi_{ij}^k \triangleq \Phi_i^k - \Phi_j^k &= \Delta \rho_{ij}^k + c \cdot \Delta t_{ij} - \Delta I_{ij}^k + \Delta T_{ij}^k + \lambda \Delta N_{ij}^k \\ &+ \Delta \epsilon_{\Phi ij}^k \end{aligned} \quad (6)$$

$\Delta(\cdot)_{ij}^k$ denotes the single difference. Note that the satellite clock difference errors are eliminated from Eqs. 5 and 6. If in addition, the difference between the single differences to two satellites k and l is taken, then the so-called double differences are constructed (indicated with $\Delta(\cdot)_{ij}^{kl}$):

$$\Delta P_{ij}^{kl} \triangleq \Delta P_{ij}^k - \Delta P_{ij}^l = \Delta \rho_{ij}^{kl} + \Delta I_{ij}^{kl} + \Delta T_{ij}^{kl} + \Delta \epsilon_{Pij}^{kl} \quad (7)$$

$$\begin{aligned} \Delta \Phi_{ij}^{kl} \triangleq \Delta \Phi_{ij}^k - \Delta \Phi_{ij}^l &= \Delta \rho_{ij}^{kl} - \Delta I_{ij}^{kl} + \Delta T_{ij}^{kl} + \lambda \Delta N_{ij}^{kl} \\ &+ \Delta \epsilon_{\Phi ij}^{kl} \end{aligned} \quad (8)$$

where the receiver clock difference errors, which are the same for both single differences, are also eliminated.

Although the use of single and double differences has many advantages (see section “Errors and Biases”), the differencing process increases the measurement noise level by $\sqrt{2}$ for each difference and introduces correlations between the differenced equations.

The double difference integer ambiguity ΔN_{ij}^{kl} in Eq. 8 plays an important role in accurate positioning using carrier

phase double differences. When the ambiguity is estimated together with the other parameters as a real number, a so-called “float solution” is obtained. If the estimated ambiguities can, however, be successfully constrained to their integer value, then an “ambiguity fixed solution” is obtained. This process of estimating ambiguity parameters, and then selecting the likeliest integer values, is known as “ambiguity resolution.” Once the L1, L2, or L5 ambiguities are resolved, and can thus be removed from the Eq. 8, the double-differenced phase observations become precise pseudo-range observations.

Errors and Biases

Introduction

GPS measurements are corrupted by biases (measurement errors that cause true ranges to be different from measured ranges by a “systematic amount”) and errors (internal instrument noise or residual biases) which limit the positioning accuracy. Three groups of GPS errors and biases are generally distinguished: receiver-dependent, satellite-dependent, and signal propagation errors.

Receiver-Dependent Errors

Measurement Noise

The measurement noise is included in the terms $\epsilon_{P_i}^k$ and $\epsilon_{\Phi_i}^k$ within the code pseudo-range (Eq. 1) and carrier phase (Eq. 4) measurements. Without biases contaminating the measurements, low measurement noise is expected to result in high accuracy positioning. The “rule-of-thumb” states that the measurement noise is at the level of 1%, or better of the wavelength (or chip length) of the signal. In the case of code pseudo ranges, this results in a measurement noise of 3 m for the C/A code and 0.3 m for the P-code. With their shorter wavelength, carrier phase measurements are less noisy implying millimeter noise for carrier phase measurements.

Receiver Clock Errors

GPS receivers are equipped with relatively inexpensive quartz crystal oscillators. The receiver clock error dt_i contaminates all satellite-receiver ranges. As mentioned in the section “Code Pseudo-range Model,” the solution to this problem is to treat the clock error as an additional parameter in the code pseudo-range estimation procedure. This allows estimating the receiver clock error with a precision of 10^{-8} to 10^{-7} s which is in addition satisfactory to access the exact measurement epoch of both code and phase measurements. An alternative strategy uses single or double differences (see Eqs. 5, 6, 7 and 8) to eliminate the receiver clock errors dt_i .

Antenna-Specific Errors

The precise point whose position is measured by a GPS receiver is the electrical phase center of the GPS receiver's antenna which is not the physical center of the antenna. The phase center will change with the changing direction of the signal from a satellite. The a priori receiver position $\vec{x}_{i0}$ has to be corrected for this effect. This behavior is different for each antenna at each carrier frequency and modelling it requires precise antenna calibrations (Görres et al. 2006). GPS antenna calibrations consist of two parts: (1) an average phase center offset with respect to a physical feature of the antenna and (2) the phase center variation (PCV) with respect to the satellite elevation angle and azimuth. Ignoring these phase center variations can lead to serious (up to 10 cm) vertical positioning errors. Even if the antenna calibrations have submillimeter accuracy, the access to absolute high-precision positioning is still limited by site-specific near-field multipath errors which can introduce systematic biases up to 1–2 cm, mostly in the estimated height component.

Multipath effects, included in terms ϵ_{pi}^k and ϵ_{di}^k in respectively Eqs. 1 and 4, are propagation errors arising from the interference of the direct signal by reflected signals from water or metallic surfaces or nearby buildings. The combined direct and reflected signals will give rise to incorrect pseudo-range and phase measurements. Low elevation signals are more susceptible to multipath effects than those at high elevations. The size of pseudo-range multipath error can reach 10–20 m on the C/A code (Evans 1986), while for the carrier phases, it can have a maximum value of a quarter of a cycle (4.8 cm for L1).

Using specific antenna design (see section “Ground Instrumentation”), careful antenna mounting, and advanced signal processing to filter out multipath integrated in the receiver can reduce the multipath errors. Using differenced code or phase observables will not remove the effect of multipath as the effect is receiver- and satellite-dependent. However, when using stationary receivers, an appropriate multipath mitigation can be implemented by taking advantage of the multipath time correlation due to the repeating satellite ground tracks (Wu and Hsieh 2010).

Cycle Slips

If the receiver loses lock on a satellite signal, due to, e.g., obstructions or interference, then the carrier phase observables might suffer from cycle slips. These cycle slips are jumps in the carrier phase measurements caused by the fact that when signal lock is reacquired, the integer initial cycle ambiguity N_i^k of the phase observable will be re-initialized. Cycle slips can occur independently on L1, L2, and L5.

The detection and repair of cycle slips is an important carrier phase data preprocessing step. Each time the software will detect a cycle slip, the cycle slip will be repaired (to restore the continuity of carrier cycle count) or a new

integer initial ambiguity N_i^k will be setup. If this is properly done, the cycle slips are not contributing to the error budget anymore.

Satellite-Dependent Biases

Orbit and Satellite Clock Errors

Computing the a priori geometric range ρ_{oi}^k (Eq. 3) implies known satellite position. The navigation message broadcast by the GPS satellites provides the user in real-time information on the satellite position and clock error. Following the International GNSS Service (<http://igs.org/>), the accuracy of the GPS broadcast orbits is at the 1-m level while the satellite clock error is given at the 5 ns level. The effect of the satellite orbit and clock errors (dr_i^k in Eqs. 4 and 6) propagates directly into the point position results and amplifies the positioning error.

In the case of relative positioning, the orbit errors dr (in m) propagate into errors db (in m) in the coordinates of a baseline of length b following the rule of thumb: $\frac{|db|}{b} = \frac{|dr|}{r}$ where r is the receiver-satellite distance (in km), typically 20,000 km. So, for a receiver inter-distance (or baseline length) of 100 km and an orbit error of 1 m, the error on the baseline components will be of the order of 0.5 cm. Following Beutler (1996), this rule of thumb is a bit too pessimistic and actually gives mainly an approximation for the propagation of the orbit errors on the estimated height component. For the horizontal component, the error can be divided by a factor of 3.

Precise GPS positioning (see ► “GPS, Tectonic Geodesy”) requires orbits with a higher precision than the ones provided in the broadcast navigation message. For that purpose, the International GNSS Service (IGS, see ► “Geodesy, Networks, and Reference Systems”) uses the tracking data from a worldwide GNSS network to generate the most precise satellite orbits and clocks related to the International Terrestrial Reference system (ITRS). Their accuracy typically varies between 2.5 and 5 cm for the orbits and ± 75 to 150 ps for the satellite clocks.

Antenna-Specific Errors

Each GPS satellite has a transmitting antenna which, like the receiver antenna, also has a phase center which changes with elevation angle and is influenced by the local environment around each satellite. The scientific community, mainly driven by the International GNSS Service, has made considerable efforts to accurately determine (or collect information about) the satellite's antenna phase center offsets and variations. The correct location of this phase center with respect to the satellite's center of mass is critical for accurate orbit determination.

Signal Propagation Biases

When travelling from the GPS satellites to the receiver on Earth, the radio signals cross and interact with the Earth's atmosphere. The two atmospheric layers that influence the

most the propagation of the GPS signals are the troposphere and the ionosphere. The atmospheric refraction has consequently an ionospheric, I_i^k , and tropospheric, T_i^k , component, each with their own spatial and temporal characteristics.

Ionospheric Refraction

The ionosphere is stretching from a height of about 50 km to more than 1000 km over the surface of the Earth; it is named so because it is ionized by the Sun's ultraviolet light. When the GPS signals propagate through this medium dispersion occurs, bending the signal and changing its velocity (Klobuchar 1991). The signal bending can be neglected for satellite elevations larger than 5° .

The ionospheric delay I_i^k is proportional to the number of free electrons along the GPS signal path, called the total electron content (TEC) which varies as a function of the latitude of the receiver, the season, the time of day, and solar activity (approximately 11-year cycle). At first order, the ionospheric delay on the carrier phases reads: $I_i^k = -\frac{1}{\cos z} \times \frac{40.3}{f^2} \text{TEC}$, with z the satellite's zenith distance. It ranges from about 50 m for signals at the zenith to as much as 150 m for low-elevation observations.

To reduce the ionospheric effect, a correction (using the Klobuchar model, Klobuchar (1986)) is transmitted within the GPS navigation message. However, it is generally conceded that the broadcast correction model corrects only 50–60% of the ionospheric delays, which is considered as the performance level of the model. For single frequency receivers, the use of the correction model parameters is often the only option for point positioning. Another source of information on the behavior of the ionosphere is provided by the IGS from their Global Ionospheric Maps (GIM) which consist of a global TEC grid ($2.5^\circ \text{ lon} \times 2.5^\circ \text{ lat}$) available at 2-hourly intervals. The IGS GIM maps have a typical accuracy of 2–8 TEC units.

In relative positioning, the ionospheric error will induce a baseline shortening dependent of the baseline length. For example, at a mid-latitude site using an elevation cutoff angle of 20° , a horizontal scale bias of -0.63 ppm is incurred for each $1 \times 10^{-7} \text{ m}^{-2}$ of TEC not accounted for.

As the ionosphere acts as a dispersive medium, the ionospheric delay is frequency dependent. Hence, if dual-frequency receivers are available, a special combination of the L1 and L2 signals, called the ionosphere-free combination L3, can eliminate the first order of the ionospheric effect. The ionosphere-free carrier phase measurement is:

$$\Phi_{L3} = \frac{f_1^2}{f_1^2 - f_2^2} \Phi_{L1} - \frac{f_2^2}{f_1^2 - f_2^2} \Phi_{L2} \quad (9)$$

with f_1 and f_2 , respectively, the frequencies of the L1 and L2 signals and Φ_{L1} and Φ_{L2} the carrier phase observables

(converted in m) on L1 and L2. A similar equation can be derived to construct the ionosphere-free code measurements. Using the ionosphere-free combination removes 99% of the ionospheric error. For applications requiring the highest positioning precision, the usage of ionosphere-free differenced observables is the most effective way of mitigating the ionospheric effect.

Tropospheric Refraction

The troposphere is the lower part of the Earth's atmosphere, and it is the seat of all meteorological phenomena (clouds, rain, hydrometeors, etc.). It contains approximately 75% of the atmosphere's mass and 99% of its water vapor and aerosols. The troposphere extends from the surface of the Earth to less than 9 km over the poles to 16 km over the equator. For radio frequencies below 15 GHz, it is a nondispersive medium meaning that it delays all GPS codes and carriers by the same amount so that the effect cannot be removed by combining observations made on two frequencies.

The tropospheric refraction delay T_i^k , depends on atmospheric pressure, temperature, and water vapor pressure. In addition, it is the shortest in the zenith direction (2 m) and increases when going down to the horizon (about 20–28 m for signals at an 5° elevation angle) as the air mass traversed by the signal increases (Brunner and Welsch 1993).

The tropospheric delay consists of a dry and wet component. The dry (or hydrostatic) component, which depends on surface pressure and represents about 90% of the delay, can be easily modelled. The most commonly used a priori model is the Saastamoinen model (1972), which is also recommended by the IERS conventions (2010). Today, the necessary pressure values input to the Saastamoinen model can be derived from numerical weather model data from the European Centre for Medium-Range Weather Forecasts (ECMWF).

The wet component of the tropospheric delay accounts for the remaining 10% of the delay and depends on the amount of water vapor along the signal path; it is much more difficult to accurately model.

Two kinds of troposphere biases can affect relative positioning: relative and absolute tropospheric biases. Relative troposphere biases are caused by errors of (mismodelled) tropospheric refraction at one endpoint of a baseline relative to the other endpoint, while absolute ones are caused by errors common to both endpoints of a baseline. Relative troposphere biases primarily bias station heights. For example, a relative troposphere bias of only 1 cm leads to an error of approximately 3 cm in the estimated relative station height for an elevation cutoff angle of 20° . This error increases to 19 cm for an elevation cutoff angle of 3° . Absolute troposphere biases produce scale biases of the estimated baseline lengths. For example, an absolute troposphere bias of 10 cm induces a scale bias of 0.05 ppm for an elevation cutoff angle of 20° and of 0.3 ppm for a cutoff angle of 3° .

For high-precision applications where mm accuracy is desired, the tropospheric bias at zenith is parameterized in the final position solution (see IERS Conventions 2010). For that purpose, different mapping functions (Niell 2000; Boehm et al. 2006) are used to express the dependency between the tropospheric delay at zenith and the one experienced at satellite's zenith distance. The Vienna Mapping Function (VMF1, Boehm et al. 2006) is recommended by the IERS for the applications requiring the highest accuracy. It is derived from data from the European Centre for Medium-Range Weather Forecast (ECMWF).

In addition, as with the ionospheric bias, the tropospheric bias is spatially correlated over distances up to several tens of kilometers meaning that differential positioning is an effective strategy for mitigating the effect of the tropospheric bias on positioning results.

Summary

As we have seen, all biases exceeding the measurement noise level have a high probability to degrade the accuracy of the positioning and should therefore be accounted for somehow. Various strategies have been developed to account for them:

- They can be estimated as explicit (additional) parameters, e.g., the receiver clock error and the tropospheric refraction.
- Common-mode biases can be eliminated or mitigated through combination of observables or datasets, e.g., by using of the ionosphere-free observable or relative positioning to mitigate the ionospheric and tropospheric refraction as well as satellite clock and orbit errors.
- The biases can be directly measured, e.g., antenna calibrations.
- The biases can be considered known or adequately modelled, e.g., by using precise satellite orbits and clocks from the IGS.
- They can be ignored, e.g., the phase multipath error at moving receivers.

Positioning Modes

Two main methods of positioning with GPS exist: point positioning and relative positioning. In addition, the precision of the obtained positions depends on the observables used: code and/or carrier phase data, the interval over which the data are accumulated before being analyzed, and the sophistication of the error modelling.

Point Positioning

GPS point positioning employs the data from a single GPS receiver.

Using code pseudo-ranges from at least four satellites, Eq. 1 can be used to compute the instantaneous receiver position and receiver clock error. With the broadcast ephemeris and ionospheric correction, a typical position accuracy of 22 m (2D rms) can be achieved using C/A codes. Typically, code pseudo-range positioning is used for low-accuracy real-time land, sea, and air navigation. The distinguishing characteristic of navigation is the urgency with which positioning information is required.

A more sophisticated method of performing point positioning uses the ionosphere-free combination of dual-frequency code and carrier phase observables. It is known as Precise Point Positioning (PPP) and allows reaching millimeter- to centimeter-level precision for 24-h solutions in static mode. To reach this precision, it is necessary to use the most precise satellite orbit and clock information (e.g., from the IGS) as well as a sophisticated error modelling. Up to a couple of years ago, PPP was only implemented in scientific GNSS analysis software. PPP has the advantage that it eliminates the need to acquire simultaneous tracking data from a reference station although it still requires external products delivered by global networks such as the IGS. Thanks to the IGS efforts to provide real-time orbit and clock information, PPP is becoming more popular (Bisnath and Gao 2008). Especially PPP with integer ambiguity resolution (PPP-AR, Shi and Gao 2013) can significantly improve the accuracy of PPP over short observation sessions, which is beneficial for fieldwork in many industries as it can potentially be used instead of the RTK positioning mode (see next section). PPP-AR is implemented by applying additional satellite corrections to mitigate unmodelled satellite equipment delays. These products are available from several agencies, e.g., Collins et al. (2014); Laurichesse et al. (2014). PPP-AR can also benefit from the availability of three frequencies (Geng and Bock 2013).

PPP has quickly become very popular for high-precision applications. Chen et al. (2018) showed, for example, how PPP using combined GPS and GLONASS measurements has become an effective method for monitoring crustal deformation caused by earthquakes.

Relative Positioning

Relative positioning employs two GPS receivers, a remote and a reference receiver, simultaneously tracking the same satellites (minimally four) to determine their relative coordinates (see section “Satellite Positioning”). It is based on the ability to eliminate, or significantly reduce, the common measurement biases across observations made simultaneously by GPS receivers see (Eqs. 5, 6, 7, and 8 and section “Errors and Biases”). Depending on the accuracy requirements, code pseudo-range as well as carrier phase

measurements can be used in relative GPS positioning. In addition, relative positioning can be implemented with different latencies varying from real time to post-processing:

- Carrier phase-based relative positioning

The basis of high-precision relative positioning is carrier phase measurements because of their low measurement noise. The problem of processing the “ambiguous” carrier phase data can be overcome by collecting sufficient epochs of data (up to 1 h or more) to ensure ambiguity resolution and strengthen the baseline solution. The reliability of the ambiguity resolution depends on:

- Baseline length
- The number of satellites (the more, the better)
- The satellite-receiver geometry
- Whether observations are made on both frequencies (it is much easier to resolve ambiguities when dual-frequency observations are available)
- The length of the observation session (the longer, the better)

Throughout time, GPS ambiguity resolution methods have become more and more sophisticated. The LAMBDA (Least-squares AMBiguity Decorrelation Adjustment method by Teunissen (1995)) allows now even on-the-fly ambiguity resolution using few observation epochs. This is typically necessary when the receiver is in motion. In addition, the availability of triple frequency signals open new possibilities for ambiguity resolution Cocard et al. (2008). An overview of different approaches is given in Li (2018).

At one extreme, static GPS surveying is the most precise positioning technique in use today. Observation times vary between 20 min to a day, depending on the expected precision, the distance between the two receivers, the number of visible satellites, and the satellite geometry. Measurements are typically taken at intervals between 10 and 30 s. After completing the field measurements, the data are processed on a PC using different processing options depending on, e.g., observation time and baseline length. For long baselines (>20 km), typically ionosphere-free carrier phases will be processed. For very long baselines (>1000 km), the user will process the data using high-precision scientific GNSS analysis software. In this case, mm to cm precision can be reached over distances exceeding 1000 km. Typical applications are the surveying of crustal or intraplate deformations (see ► “GPS, Tectonic Geodesy”).

At the other extreme, real-time-kinematic (RTK) GPS is the less precise relative GPS positioning method based on carrier phases. The main differences to static GPS surveying is the drastic reduction of the observation time and the fact that the position of the remote (or rover) receiver is often computed in real time. RTK can be carried out in real

time if an appropriate communications link (UHF radio modem, satellite link, cell phone, etc.) is provided over which the carrier phase data collected at the reference receiver can be made available to the remote receiver’s onboard computer. In this way, the remote receiver can generate the double differences, resolve the initial ambiguities, and perform the position calculations. Typically, measurements are taken at a 1 Hz rate and initial ambiguities are determined on-the-fly. In order to mitigate as much as possible the error sources, baselines should not exceed 10–15 km to reach the expected positioning precision of the order of 2–5 cm. Cadastre surveying is one of the applications that uses RTK. More information on RTK can be found in Langley (1998).

Other relative positioning modes based on GPS carrier phases are rapid static positioning and stop-and-go surveying. They are lying between the above-mentioned extremes.

- Code-based relative positioning is mostly known as real-time Differential GPS (DGPS). There are two ways to implement DGPS: (1) the fixed reference station broadcasts the differences between the measured pseudo ranges and actual (internally computed) pseudo range so that it can correct its pseudo ranges by the same amount before determining its position (method of range corrections); (2) the fixed reference station compares its known position with its computed position and sends the difference to the remote receiver which uses this difference to correct its computed position. The correction signal is broadcast using similar means as uses RTK. Range corrections are the preferred mode for real-time DGPS navigation. DGPS can deliver meter-level positioning that typically ranges between 0.5 and 5 m.

The standard format used to exchange GPS data in real time is RTCM 104 (cf. RTCM 2013). For the exchange of GPS data targeting post-processing, the RINEX (Receiver INdependent EXchange) format (IGS RINEX WG and RTCM-SC104, 2018) is used.

High-Precision Positioning

Static relative GPS surveying and PPP using ambiguity resolution both enable the most precise positioning. Millimeter-level precisions are obtained when GPS receivers are installed on a fixed location and the data are processed in post-processing ingesting a full day of GPS observations. This enables a wide variety of applications, such as the mitigation of the effects of vertical land motion in tide gauges (Santamaría-Gómez et al. 2012), the demonstration of correlations between water pumping and vertical ground deformations (Argus et al. 2017), the quantification of ground movements caused by the glacial isostatic adjustment

(Kreemer et al. 2018), or the study of volcanic processes (Fernández et al. 2017).

Conclusion

GPS positioning is based on measuring the travel times of radio signals emitted by satellites to a receiver mostly located on the ground. These travel times can be obtained by measuring on the precise, but ambiguous, GPS carrier phases or on the, less precise, GPS PRN codes. By multiplying the travel time by the speed of light, the distances receiver–satellites are reconstructed. Using minimally four satellites, the receiver can compute its position. Errors that influence GPS positioning are receiver-related (measurement noise, multipath, receiver clock errors, antenna phase center variation), satellite-related (satellite orbit and clock error, antenna phase center variation) or related to the signal propagation (ionospheric and tropospheric effect). As a lot of these errors are similar for GPS receivers distant of a few kilometers, relative GPS positioning was developed to improve the point positioning accuracy.

GPS is active since the mid-1980s. It has grown from a purely military navigation system to a scientific tool allowing today to measure deformations of the Earth's crust with millimeter-level precision. Today, GPS is mostly used together with other GNSS leading to even more robust and precise positioning.

Cross-References

- Geodesy, Networks, and Reference Systems
- GPS, Tectonic Geodesy

Bibliography

- Argus DF, Landerer FW, Wiese DN, Martens HR, Fu Y, Famiglietti JS, Watkins MM (2017) Sustained water loss in California's mountain ranges during severe drought from 2012 to 2015 inferred from GPS. *J Geophys Res Solid Earth* 122:10559–10585. <https://doi.org/10.1002/2017JB014424>
- Beutler G (1996) GPS satellite orbits. In: Kleusberg A, Teunissen JG (eds) *GPS for geodesy. Lecture notes in earth sciences*, vol 60. Springer, Berlin, pp 37–101
- Bisnath S, Gao Y (2008) Current state of Precise Point Positioning and future prospects and limitations. In: Sideris M (ed) *International association of geodesy symposia "observing our changing earth"*, vol 133. Springer, Berlin, pp 615–623
- Boehm J, Werl B, Schuh H (2006) Troposphere mapping functions for GPS and very long baseline interferometry from European Centre for Medium-Range Weather Forecasts operational analysis data. *J Geophys Res* 111:B02406. <https://doi.org/10.1029/2005JB003629>
- Braasch MS, Van Dierendonck AJ (1999) GPS receiver architectures and measurements. *Proc IEEE* 87(1):48–64. <https://doi.org/10.1109/5.736341>
- Brunner FK, Welsch WM (1993) Effect of the troposphere on GPS measurements. *GPS World* 4(1):42–51
- Chen X, Lu C, Guo B, Guo F, Ge M, Li X, Schuh H (2018) GPS/GLONASS combined Precise Point Positioning with the modeling of highly stable receiver clock in the application of monitoring active seismic deformation. *J Geophys Res Solid Earth* 123:4025–4040. <https://doi.org/10.1029/2017JB015060>
- China Satellite Navigation Office (2019) BeiDou navigation satellite system signal in space interface control document, open service signal (version 3.0)
- Cocard M, Bourgon S, Kamali O, Collins P (2008) A systematic investigation of optimal carrier-phase combinations for modernized triple-frequency GPS. *J Geod* 82(9):555–564
- Cohen CE, Parkinson BW (1991) Mitigating multipath error in GPS Based Attitude Determination. In: *Proceedings of the annual rocky mountain guidance and control conference*, Keystone, February 2–6, 1991 (A93–15576 03–12), pp 53–68.
- Collins P, Bisnath S, Lahaye F, Héroux P (2014) Undifferenced GPS ambiguity resolution using the decoupled clock model and ambiguity datum fixing. *Navigation*. <https://doi.org/10.1002/j.2161-4296.2010.tb01772.x>
- Evans AG (1986) Comparison of GPS pseudorange and biased Doppler range measurements to demonstrate signal multipath effects. In: *Proceedings of the fourth international geodetic symposium on satellite positioning*, vol 1, University of Texas, Austin, pp 761–587
- Fernández J, Pepe A, Poland M, Sigmundsson F (2017) Volcano geodesy: recent developments and future challenges. *J Volcanol Geotherm Res* 344:1. <https://doi.org/10.1016/j.jvolgeores.2017.08.006>
- Galileo (2009) Report from the commission to the European Parliament and the Council on the implementation of the GNSS programmes and on future challenges. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52009DC0302&from=EN>
- Geng J, Bock Y (2013) Triple-frequency GPS precise point positioning with rapid ambiguity resolution. *J Geod* 87(5):449–460
- Görres B, Campbell J, Becker M, Siemes M (2006) Absolute calibration of GPS antennas: laboratory results and comparison with field and robot techniques. *GPS Solutions* 10(2):136–145. <https://doi.org/10.1007/s10291-005-0015-3>
- IERS Conventions (2010) In: Petit G, Luzum B (eds) *IERS technical note 36*. Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt am Main, 179 pp. ISBN: 3-89888-989-6
- IGS RINEX WG, RTCM-SC104 (2018) RINEX – the Receiver Independent EXchange format, Version 3.04. <ftp://igs.org/pub/data/format/rinex304.pdf>
- IS-GPS-200J (2018) GPS Interface Specification Revision J. <https://www.gps.gov/technical/icwg/IS-GPS-200J.pdf>
- Klobuchar JA (1986) Design and characteristics of the GPS ionospheric time delay algorithm for single frequency users. In: *Proceedings IEEE plans 1986 position, location and navigation symposium*, IEEE, Las Vegas, pp 280–286
- Klobuchar JA (1991) Ionospheric effect on GPS. *GPS World* 2(4):48–51
- Kreemer C, Hammond WC, Blewitt G (2018) A robust estimation of the 3-D intraplate deformation of the North American plate from GPS. *J Geophys Res Solid Earth* 123:4388–4412. <https://doi.org/10.1029/2017JB015257>
- Langley RB (1998) RTK GPS. *GPS World* 4(4):70–76
- Laurichesse D, Mercier F, Berthias J-P, Broca P, Cerri L (2014) Integer ambiguity resolution on undifferenced GPS phase measurements and its application to PPP and satellite precise orbit determination. *Navigation*. <https://doi.org/10.1002/j.2161-4296.2009.tb01750.x>
- Li BJ (2018) Review of triple-frequency GNSS: ambiguity resolution, benefits and challenges. *Glob Position Syst* 16:1. <https://doi.org/10.1186/s41445-018-0010-y>
- Niell AE (2000) Improved atmospheric mapping functions for VLBI and GPS. *Earth Planets Space* 52:699–702

- RTCM (2013) Radio Technical Commission for Maritime Services (RTCM) standard 10403.2, Differential GNSS (Global Navigation Satellite Systems) Services, Version 3 with Amendment 2, 7 Nov 2013. <http://www.rtcn.org/>
- Russian Space Systems (2016) GLONASS Interface Control Document. <http://russianspacesystems.ru/wp-content/uploads/2016/08/ICD-GLONASS-CDMA-General-Edition-1.0-2016.pdf>
- Saastamoinen J (1972) Atmospheric correction for the troposphere and stratosphere in radio ranging of satellites. In: Henriksen SW, Mancini A, Chovitz BH (eds) The use of artificial satellites for geodesy. Geophysical monograph series, vol 15. AGU, Washington, DC, pp 247–251
- Santamaría-Gómez A, Gravelle M, Collilieux X, Guichard M, Martin Miguez B, Tiphaneau P, Woppelmann G (2012) Mitigating the effects of vertical land motion in tide gauge records using state-of-the-art GPS velocity field. *Global Planet Change* 98–99:6–17. <https://doi.org/10.1016/j.gloplacha.2012.07.007>
- Shi J, Gao Y (2013) A comparison of three PPP integer ambiguity resolution methods. *GPS Solutions* 18:519–528. <https://doi.org/10.1007/s10291-013-0348-2>
- Spilker J (1978) GPS signal structure and performance characteristics. *J Inst Navig* 25:121–146
- Teunissen PJG (1995) The least-squares ambiguity decorrelation adjustment: a method for fast GPS integer ambiguity estimation. *J Geod* 70(1–2):65–82
- Teunissen PJ, Montenbruck O (eds) (2017) Handbook of global navigation satellite systems. Springer handbooks. Springer, Cham. https://doi.org/10.1007/978-3-319-42928-1_33
- Tranquilla JM, Colpitts BG, Cam JP (1989) Measurement of low-multipath antennas for topex. In: Proceedings 5th international geodetic symposium on satellite positioning, New Mexico State University, Las Cruces, pp 356–361
- Woo KT (1999) Optimum semi-codeless carrier phase tracking of L2. In: Proceedings 12th international technical meeting of the satellite division of the institute of navigation, Nashville, September 14–17, pp 289–305
- Wu J, Hsieh CH (2010) Statistical modeling for the mitigation of GPS multipath delays from day-to-day range measurements. *J Geod* 84(4):223–232. <https://doi.org/10.1007/s00190-009-0358-6>
- Zumberge J, Heflin M, Jefferson D, Watkins M, Webb F (1997) Precise point positioning for the efficient and robust analysis of GPS data from large networks. *J Geophys Res* 102:5005–5501

GPS, Tectonic Geodesy

Jeffrey T. Freymueller

Department of Earth and Environmental Sciences, Michigan State University, East Lansing, MI, USA

Synonyms

Crustal deformation

Definition

Earthquake cycle A conceptual model that explains the buildup of stress that causes earthquakes, through a period of steady tectonic loading that culminates in sudden slip in an earthquake.

Slip deficit	The difference between the slip that has occurred over a period of time on a fault and the amount expected from the long-term fault slip rate; a measure of the accumulated slip “available” for an earthquake on the fault.
Displacement	The change in position of a point on the Earth’s surface.
Coseismic displacement	The final, static displacement of the surface caused by slip in an earthquake.
Postseismic displacement	Displacements caused by postseismic processes, which are transient deformation processes that follow large earthquakes.
Creep	Steady gradual fault slip that is aseismic (does not result in an earthquake).
Volcanic inflation	Dilatational deformation caused by pressurization of a volcanic system.
Glacial isostatic adjustment (GIA)	The deformation, mainly uplift, that results from the removal of glacial surface loads and the isostatic response of the crust and mantle.

Introduction

Tectonic geodesy uses measured displacements of points on Earth’s surface to study active tectonics. These motions mainly result from plate motions, slip on active faults (including earthquakes), and other earthquake cycle-related deformation, including transient fault slip. Additional deformation signals may result from the pressurization or depressurization of magma systems beneath volcanoes, hydrological deformation caused by aquifer and oil use, and changes in surface loads such as glacial isostatic adjustment (GIA), and these must be recognized and modeled so that they are not misinterpreted as tectonic signals. Geodesy is one of the most important tools for the study of plate boundary zones, the regions of distributed deformation that are commonly found at the boundaries of the major tectonic plates. The primary measurement tools today are GPS (now also including data from other global navigation satellite systems) and interferometric synthetic aperture radar (InSAR) (see ► [“GPS, Data Acquisition, and Analysis,”](#) and ► [“SAR Interferometry”](#)). Other geodetic techniques can provide similar data, but the spatial density and wide coverage of GPS and InSAR measurements are of critical importance.

Measurements of displacements depend on a precise time series of positions in a stable terrestrial reference system (see ► [“Geodesy, Networks, and Reference Systems”](#)). Usually, GPS time series are dominated by steady, linear motion over time, so site velocities are frequently used to describe site motions. However, as measurement capabilities have improved, we have found

more and more deviations from linearity in data from tectonically active regions due to fault creep events, post-seismic deformation, and non-tectonic sources like volcanic inflation and surface loading changes.

History of Tectonic Geodesy

Early work in tectonic geodesy dates back more than a century. The Dutch surveyor J. J. A. Müller reported ground movements associated with the 1892 Tapanuli earthquake in Sumatra, which caused changes in observed angles (Bonafede et al. 1992). H. F. Reid used displacements computed from geodetic surveys before and after the 1906 San Francisco earthquake as the basis for the elastic rebound hypothesis (Reid 1910, 1911). These observations supported the hypothesis proposed by G. K. Gilbert a few decades earlier that earthquakes were the result of a sudden release of strain that had built up over a long period of time (Gilbert 1884). Numerous studies over the succeeding decades used many kinds of geodetic measurements to document both sudden displacements due to earthquakes and the slow, steady deformation between them. By the 1970s, efforts were underway in the USA, Japan, and a few other countries to make extensive and systematic terrestrial geodetic measurements for the purpose of measuring crustal deformation. Although restricted in spatial coverage (usually to individual faults or fault systems), these data prompted the development of most of the conceptual and numerical models used in tectonic geodesy today.

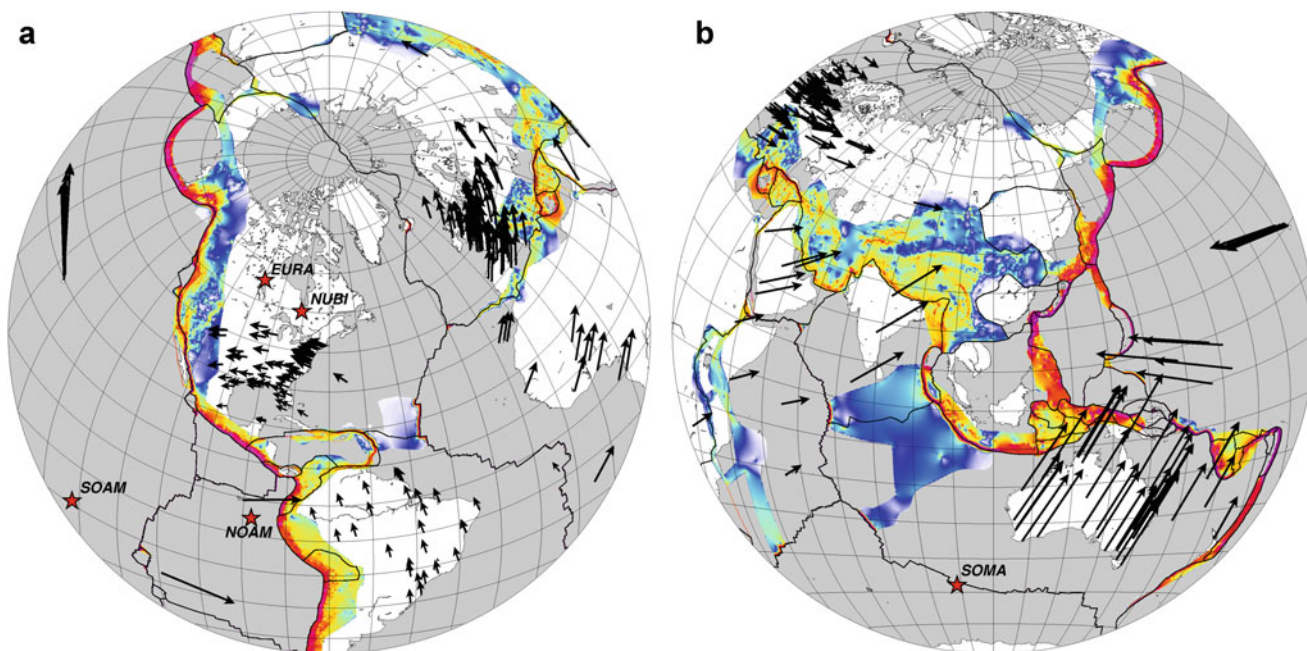
Measurement of tectonic motions over large spatial scales began in the 1980s with NASA's Crustal Dynamics Project and related international efforts (e.g., the WEGENER project in Europe). This work started with a network of fixed (and a few mobile) very long baseline interferometry (VLBI) and satellite laser ranging (SLR) stations (see ► [“Very Long Baseline Interferometry”](#)) and later added GPS measurements. Today, data from more than 10,000 GPS stations are available and used by researchers around the world (e.g., <http://geodesy.unr.edu/PlugNPlayPortal.php>). High-accuracy GPS satellite orbits are available from the International GNSS Service (IGS), and the stations that contribute to the IGS tracking network form a reference network that any researcher can use to determine station positions in the International Terrestrial Reference Frame (ITRF) (Altamimi et al. 2016, 2017). Modern studies in tectonic geodesy are based on motions derived from time series of site positions expressed in the ITRF, on regional to global scales. Global compilations of site velocities from GPS have been used to measure plate motions and to develop a global strain rate map (Kreemer et al. 2014) and global elastic block models (e.g., Graham et al. 2018).

Plate Motions

Plate tectonic theory holds that the lithosphere can be divided into a set of rigid plates moving relative to each other (Fig. 1), with tectonic strain concentrated at their boundaries and caused by their relative motions. The motion of a rigid plate on the surface of a sphere is a rotation about an axis passing through the geocenter, so that the motions of all points on that plate are described by a single angular velocity vector. The surface projection of this axis of rotation is called the pole of rotation or Euler pole. Models of plate motions thus consist of angular velocities of plates relative to some reference frame or angular velocities of relative plate motions. Outside of known regions of tectonic activity, the plate interiors are rigid to a level comparable to the measurement precision of GPS (Calais et al. 2006; Argus et al. 2010), once non-tectonic effects like GIA have been removed (e.g., Kreemer et al. 2018; Ding et al. 2019).

In the early history of tectonic geodesy, geodetic results were validated to models of geologically current plate motions, as those were more precise at the time and their level of accuracy was understood. Magnetic anomalies at spreading centers and transform fault azimuths record the rates and directions of past relative plate motions, and estimates of current plate motions are determined from these data (averaged over periods of up to a few million years). In the early 1990s, comparisons showed that geologic and geodetic plate motion models were very well correlated once the geologic rates had been rescaled to match a then-recent revision of the geomagnetic timescale (Robbins et al. 1993; DeMets et al. 1994). Today, geodetic estimates are more precise than the geologic estimates. The first global plate motion models based on GPS data alone were Argus and Heflin (1995) and Larson et al. (1997). Larson et al. (1997) found that the motion of the Nazca plate was significantly slower than predicted by the geologic models, a result later confirmed by Norabuena et al. (1998, 1999), who found that the rate of subduction of the Nazca plate beneath South America has declined steadily over the last 20 million years. The most recent work (DeMets et al. 2010; Argus et al. 2010) confirms that geologic and geodetic plate motion rates are very similar despite the different timescales (a few decades vs. 780,000 years), although the differences in plate motions are statistically significant. This implies that changes in plate motions of a few to several percent have occurred over the last three million years (Argus et al. 2010).

Geodetic motions are understood most easily in terms of geologic structures when the velocities of sites are expressed relative to one of the major plates. But plate boundaries are always zones of distributed deformation (Figs. 1 and 2), which can extend over hundreds or thousands of kilometers and involve many active structures, especially within the continental crust (Thatcher 2003, 2009). These may include



GPS, Tectonic Geodesy, Fig. 1 Velocities of geodetic sites can measure both rigid plate motion and plate boundary deformation. The two panels show views of roughly opposite sides of the planet. Velocities on plate interiors are from Altamimi et al. (2017), in the ITRF2014 reference frame. The colored regions in plate boundary zones show the second invariant of the strain rate tensor from the Global Strain Rate Model v2.1 of Kreemer et al. (2014), with hotter colors indicating regions of higher strain rate. The stars show the poles of rotation

(in ITRF) for the North American (NOAM), South American (SOAM), Eurasian (EURA), Nubian (NUBI), and Somalian (SOMA) plates, also derived from Altamimi et al. (2017). Other plate poles of rotation are located outside of the view of figure or in locations that would obscure other features shown in the figure. The plate rotation is counterclockwise about the pole, except for the Somalian plate where the pole for a clockwise rotation is shown. Major plate boundaries are shown in red, based on Bird (2003)

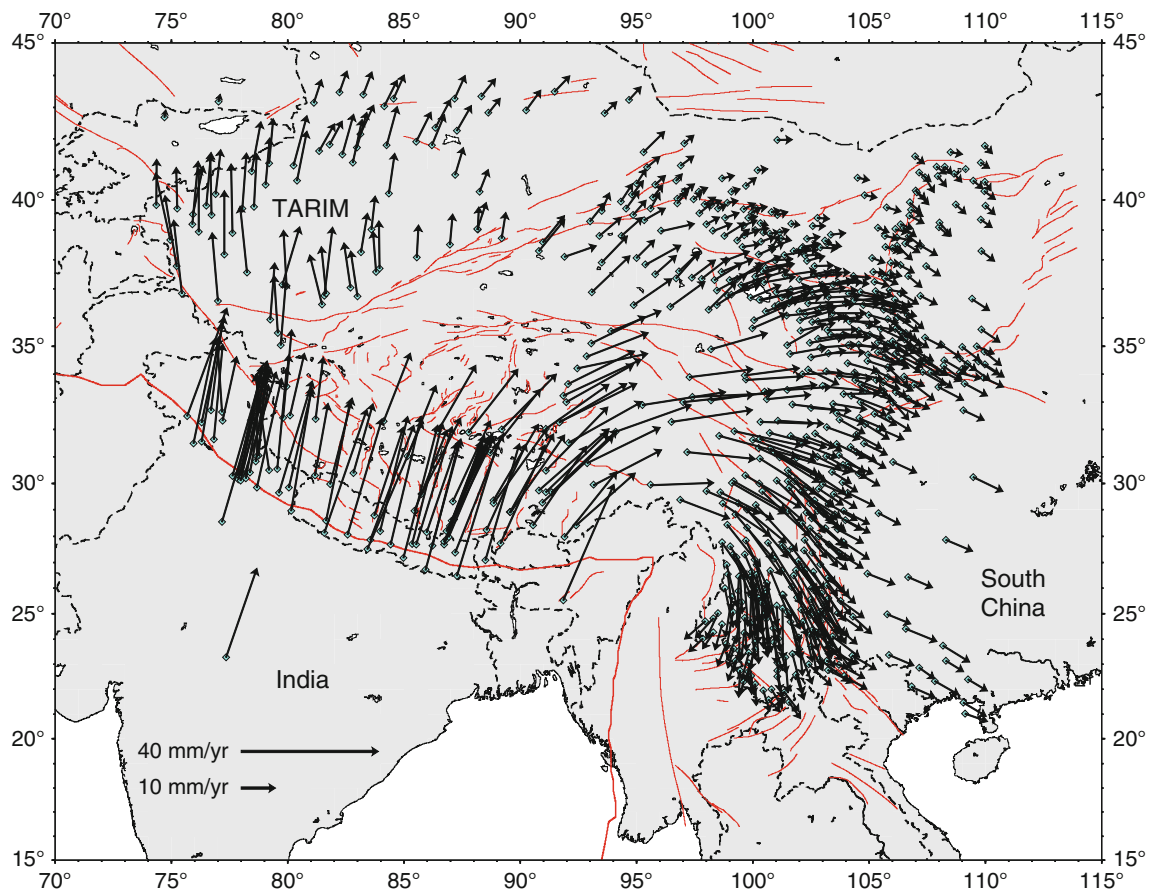
rigid blocks or microplates, which often can be identified directly from the data as low-strain regions (Fig. 1; also, Kreemer et al. 2014), such as the Tarim and South China in the broad and complex India–Eurasia boundary zone (Fig. 2). Plate boundary zone deformation in Asia and North America has received a great deal of study, and work on these plate boundary zones requires an integration of the concepts of plate tectonics with the earthquake cycle and interseismic deformation, along with other causes of deformation detailed later in this article.

Steady “Interseismic Deformation”

Time series of site positions are typically dominated by steady, quasi-linear motion in time, aside from periodic seasonal variations that are generally related to surface loading (especially in the vertical component). This is true for sites both in plate interiors and often within plate boundary zones. As a result, many models reduce these time series to a site velocity based on a linear fit. A linear velocity fit needs to take account of any discontinuities or known time-dependent variations in the time series, such as those due to earthquakes, postseismic deformation, equipment changes, and other effects, as discussed in later sections.

The steady, linear motion of the tectonic plates leads to the concept of an “earthquake cycle,” in which tectonic stresses increase linearly with time until the fault undergoes shear failure, at which point an earthquake occurs and the stored stresses are relieved by fault slip (e.g., Scholz 1990). The basic concept goes back to Reid’s elastic rebound hypothesis (Reid 1910), in which he proposed that all of the strain that accumulates around a fault before an earthquake is accommodated along the fault when it suddenly slips, so that the net result of one full cycle is a block-like offset along the fault, with no strain off of it – much like what would be observed by geologists (Fig. 3). Assuming that the Earth has a linear elastic rheology, then strain and deformation should occur at a constant linear rate between earthquakes. Elastic earthquake cycle models thus imply a long “interseismic” period of linear deformation with time, punctuated by offsets during large earthquakes. Although there is unambiguous observational evidence for transient postseismic deformation following earthquakes, the concept of “interseismic deformation” remains useful.

Numerical models for the interseismic period follow directly from the earthquake cycle model. Large earthquakes within continental crust rupture a limited range of depths, from at or near the surface to a lower depth limit of (commonly) 10–20 km. Below this, observations of exhumed



GPS, Tectonic Geodesy, Fig. 2 Velocities of sites in the India-Eurasia plate boundary zone, relative to the Eurasian plate. GPS velocities are from Gan et al. (2007), and faults (red lines) are from Taylor and Yin (2009). Rigid rotation of two microplates or small plates, Tarim and

South China, is clearly visible in the velocity field at this scale. Most of the motions shown here can be explained by a combination of rigid block motion, including a number of smaller blocks and elastic deformation that results from the shallow locked zones of block-bounding faults

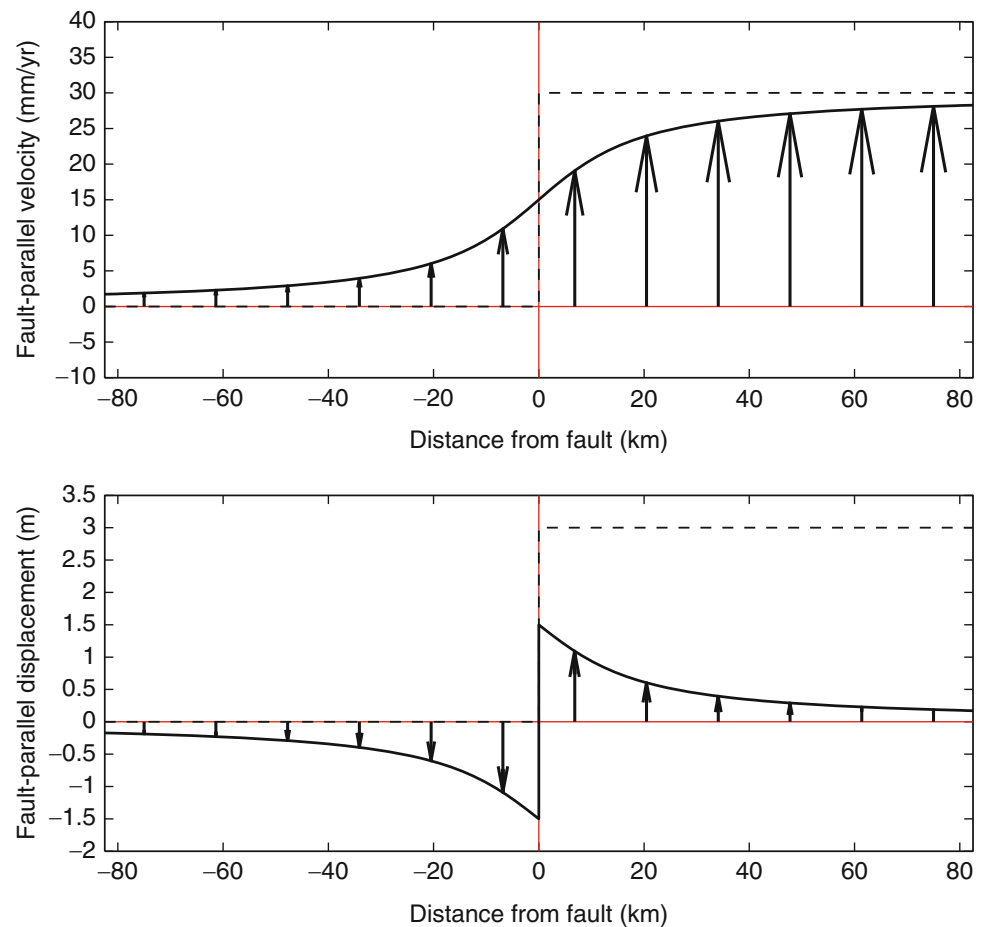
deep fault zones show evidence for ductile shear, presumed to occur continuously with time (summarized in Scholz 1990). Interseismic models commonly express this two-layer rheology using elastic dislocation theory. In the widely used model of Savage and Burford (1970), the fault zone creeps continuously at the long-term fault slip rate (accruing no strain) except for the shallow part, which remains locked by friction except in earthquakes. This is represented numerically by an elastic dislocation plane that creeps steadily from a locking depth d to infinite depth, embedded in an elastic half-space, which models the entire Earth as an elastic body. Alternatively, the slip on the fault as a function of depth can be represented by a linear superposition of uniform slip at all depths and backward slip, or backslip, on the shallow part of the fault. The fault does not actually slip backward; the backslip is a mathematical construct, one component of the linear superposition, and in fact represents the *slip deficit*, the degree to which the slip on the shallow part of the fault is interseismically trailing behind the long-term average slip rate. In the simplest earthquake cycle model, the slip in an earthquake would exactly compensate the accumulated slip

deficit over the interval between earthquakes, but in reality this is true only in a long-term average sense. Nevertheless, stress accumulates proportional to the slip deficit rate in all cases. In the two-dimensional case of an infinitely long strike-slip fault, the equations describing the dislocation model are very simple, with the velocity of a site located a distance x from the fault $V = (s/\pi) \cdot \arctan(x/d)$, where s is the average slip rate and d is the locking depth (Fig. 3). This velocity is expressed in a frame in which the velocity is zero at the fault and reaches $\pm 1/2$ the long-term fault slip rate far from the fault on either side. When the locking depth is shallow, the interseismic strain is concentrated close to the fault, and when it is deep, the strain is distributed over a larger region.

This simple model has proven to be extremely successful in a wide variety of tectonic settings, and the same conceptual model has been applied to normal faulting and subduction zones (Savage 1983). Savage and Burford (1970) and Savage (1983) assumed that steady slip would result in block-like or plate-like motion with all offset occurring as slip on the fault and no displacement or strain occurring in the surrounding rock, but this is strictly true only for a simplified fault

GPS, Tectonic Geodesy,

Fig. 3 Velocities and coseismic displacements expected from a simple elastic earthquake cycle model. The fault in this model is an infinitely long strike-slip fault (left-lateral) with a slip rate of 30 mm/yr., which has an earthquake after 100 years with 3 meter slip. During the interseismic period, the fault is locked from the surface to 15 km depth and slipping steadily below that. The interseismic deformation pattern (*top panel*) is shown by the arrows and solid line, relative to the far field on the left side of the fault. Compared to the uniform block motion that would result if the fault slipped uniformly at all depths (dashed line), the elastic response of the Earth spreads the fault shear over a broad area. The coseismic displacement (*bottom panel*) is antisymmetric about the fault, with maximum displacement at the fault (each side moves by half the coseismic slip, in opposite directions). The sum of 100 year of interseismic deformation and the coseismic displacement is a uniform block motion, shown by the dotted line



geometry such as a single planar fault. In other cases, the displacements due to the forward slip need to be calculated and included in the superposition (e.g., Cross and Freymueller 2008; Kanda and Simons 2010, 2012). Kanda and Simons (2010) further proposed a finite-plate elastic dislocation model for subduction zones, a generalization of the Savage (1983) approach.

When the locking depth varies along strike, the pattern of strain will also vary along strike; in the extreme limit of a locking depth of zero, the fault creeps at the surface at its long-term slip rate. Observations show a wide range of fault locking depths, generally in agreement with depths of maximum earthquake slip or depths of seismicity. Locking depths sometimes vary abruptly in the along-strike direction, for example, along subduction zones (Wallace et al. 2004, 2007; Cross and Freymueller 2008; Freymueller et al. 2008; McCaffrey 2008; Chlieh et al. 2008; Loveless and Meade 2010; Moreno et al. 2011; Nocquet et al. 2014; Chlieh et al. 2014; Protti et al. 2013). The parts of the fault determined (from geodetic data) to be locked by friction and accumulating slip deficit generally correspond well to the rupture areas of past great earthquakes.

The elastic dislocation approach to modeling interseismic deformation has been extended to three dimensions using elastic block modeling (e.g., McCaffrey 2002; Meade and Hager 2005a; Loveless and Meade 2011). In this approach, which blends the Euler pole description of plate motions with an elastic interseismic model, the model domain is first divided into a set of rigid blocks or microplates separated by block boundaries that represent faults and accommodate all relative motion. The velocity at any point on the surface is then the sum of two components, the rotation of the block it lies upon (described by the block angular velocity), and the elastic deformation caused by the slip deficit on all faults in the model (computed using elastic dislocation theory). The elastic block model enforces self-consistency of block motions, fault slip rates, and elastic deformation, with the only estimated or assumed parameters being the block angular velocities and the fault locking depths. Common practice is to assume a single locking depth for all or most crustal faults, although slip deficit distributions can be estimated. The inverse problem is linear when only block angular velocities are estimated and nonlinear when the locking depth or slip deficit distribution is estimated as well.

The elastic block modeling approach has been successful in describing deformation worldwide and has been extended to include uniform strain within some blocks (McCaffrey 2002; Nyst and Thatcher 2004; Meade and Hager 2005a, 2005b; Reilinger et al. 2006; McCaffrey et al. 2007; Meade 2007; Loveless and Meade 2010, 2011; Elliott et al. 2010, 2013; Johnson 2013). Graham et al. (2018) recently published a global reference elastic block model, the first attempt to extend this plate tectonics-like approach to plate boundaries globally, and used this model to estimate the global and regional moment-rate accumulation budgets.

Despite the general success of this approach, there is no way to account for time-dependent deformation within an elastic block model, so it must be removed from the data first. There have been attempts to develop viscoelastic block models, in which the predicted deformation rates evolve with time throughout the earthquake cycle (Johnson et al. 2007; DeVries et al. 2016). A major challenge for such models is that they are sensitive to both the viscosity structure and the recent earthquake history over decades to centuries, and in many places, there is not enough known to fully constrain the unknown model parameters. However, in principle such models integrate interseismic, coseismic, and postseismic deformation into a single framework.

The block modeling approach provides geodetic estimates of fault slip rates, which are in most cases consistent with longer-term geologic estimates (Thatcher 2009; Meade et al. 2013; Zeng and Shen 2014). However, there are a number of exceptions, and these may result from errors in data, inadequacies in the assumed block geometry, fundamentally non-block-like behavior, and neglected effects from viscoelasticity. In Southern California, a region with a notably complex network of active faults, several studies have concluded that viscoelastic block models give results more consistent with observed geologic slip rates than do purely elastic models (Chuang and Johnson 2011; Johnson 2013; Tong et al. 2014). Meade et al. (2013) argued that a transient rheology such as a Burger's body was required, as a viscoelastic model with a linear (Newtonian) viscosity would not explain the common agreement of geologic and geodetic rate estimates. An alternative explanation is that there is significant distributed deformation outside of the major faults, so that blocks are internally straining (Johnson 2013; Zeng and Shen 2014). Models that incorporate internally deforming blocks have also been proposed for parts of the Tibetan Plateau and India-Eurasia collision zone (e.g., Chen et al. 2004; Loveless and Meade 2011).

A prominent example of a remaining unresolved discrepancy between geological and geodetic slip rates involves the Garlock fault and Eastern California shear zone in the Mojave Desert of California (see discussion in Zeng and Shen 2014). The Garlock fault is a left-lateral strike-slip fault that is in a conjugate orientation to the right-lateral San Andreas fault

and the faults of the Eastern California shear zone. The present geodetic slip rate estimated for the Garlock fault is lower than expected based on the recent geological slip rate (McGill et al. 2009), while the reverse is true for the adjacent Eastern California shear zone. Models that incorporate time-dependent postseismic deformation have smaller discrepancies than elastic models, but this alone does not resolve the discrepancy (Tong et al. 2014). Dolan et al. (2007) and Oskin et al. (2008) suggested that time-dependent behavior of ductile shear zones in the deep crust may result in time-dependent fault slip, which in this case might result in switching between two modes: elevated slip rate on the Eastern California shear zone with low slip on the Garlock fault (the present) and elevated slip on the Garlock fault with low slip on the Eastern California shear zone (much of the geologically recent past).

Transient Fault Movements

Faults exhibit a range of transient slip behavior, including both spontaneous and triggered creep events. Distinct creep events of various sizes have been identified on several faults within the San Andreas fault system in California (e.g., Gladwin et al. 1994; Bürgmann et al. 1997; Lienkaemper et al. 1997; Taira et al. 2014). In recent years, particular focus has been placed on slow slip events (SSE) on subduction thrusts, also called episodic tremor and slip (ETS) events because the slip is usually accompanied by the emission of seismic tremor (Rogers and Dragert 2003). These transient fault movements are further evidence that the steady, elastic earthquake cycle model is a significant simplification of reality: even between earthquakes, there can be small variations in deformation. Mavrommatis et al. (2015) identified a decadal-scale slip event that preceded the 2011 Tohoku earthquake in Japan, which they interpreted as resulting from increasing slip rate and updip migration of deep creep down-dip of the eventual earthquake rupture. Recent lab experiments have produced spontaneous slow slip events in natural geological materials, opening up the possibility of testing hypotheses for the nucleation and evolution of these events in the lab (Ikari 2019).

SSE and ETS events were first observed in Japan (Heki et al. 1997) and at the Cascadia subduction zone (Dragert et al. 2001) in North America. A large number of these events have now been identified worldwide, and in most cases, the slip events are accompanied by emission of seismic tremor (Schwartz and Rokosky 2007). These events come in a range of sizes, many with slip and area equivalent to magnitude 6–7.5 earthquakes. The upper size limit for such events is not clearly defined. SSEs of decadal length have been detected or inferred from geodetic or paleo-geodetic (coral) data (Tsang et al. 2015; Li et al. 2016), and Meade and Loveless (2009) noted that scaling relations suggest that

average slip velocity decreases with event duration, such that multi-decadal SSEs might have slip rates slower than plate motion rates, making them appear like a case of “partial locking.” Small tremor bursts are also observed, as are smaller slip events too small to detect with GPS (but observed on strainmeters or tiltmeters). These mostly aseismic slip events, including small bursts of tremor that do not have resolvable geodetic signals, appear to follow a self-similar (power law) scaling relation like that of regular earthquakes but with a different power law exponent (Ide et al. 2007). Most large events recognized in the decade after Ide et al. (2007) have been interpreted as being consistent with the proposed power law scaling. However, Frank et al. (2018) argued that the large SSE in Guerrero, Mexico, in 2006 is more accurately described as a cluster of transient events with some relocking of the interface between events, which would result in a different scaling exponent.

Studies of SSEs at the Hikurangi margin (North Island, New Zealand) have been particularly important in improving our understanding, as events with a wide range of depths and durations (McCaffrey et al. 2008; Wallace and Beavan 2010) have been observed using an excellent on-land geodetic network along with temporary offshore pressure gauges (Wallace et al. 2016; Yohler et al. 2019). The detailed time dependence of slip in these events has been determined, and in some cases, events have influenced the slip history of neighboring segments (Bartlow et al. 2014). The 2016 Kaikoura earthquake triggered slow slip along much of the Hikurangi margin, as much as 600 km away (Wallace et al. 2017, 2018).

Prior to the recognition of SSEs or ETS events, the typical assumption in tectonic geodesy was that all deviations from a linear trend represented noise in the data, except when there was a known cause of displacement such as an earthquake or postseismic deformation. Today, deviations from a linear trend are likely to be considered signal, although it remains uncertain how to reliably distinguish between transient tectonic signals and noise; a particular focus of current research is in the automatic detection of transient events. A variety of approaches are used to identify temporal variations in individual time series and to use the spatial coherence of deformation signals as part of the event identification or to reduce ambiguities in the detection process (e.g., Ji and Herring 2013; Riel et al. 2014). The identified transient signals at multiple stations can then be inverted for a source model. However, noise in geodetic time series can be spatially correlated as well, and the separation of small transient signals from noise remains an area of active research.

Coseismic Deformation

Coseismic displacements are the sudden displacements caused by earthquakes. In large earthquakes they can extend

a long distance from the fault due to the elastic response of the Earth (Fig. 3). The earliest observations in tectonic geodesy were of the angle changes between triangulation stations caused by coseismic displacements. Today, GPS and InSAR displacements provide precise and spatially dense measurements of coseismic displacements, allowing detailed models of the distribution of slip in earthquakes to be determined; it is now common to use data from hundreds of GPS sites and thousands of InSAR pixels even after downsampling. Very large earthquakes are followed by periods of postseismic deformation during which deformation is aseismic (except for aftershocks) but is perturbed from the pre-earthquake pattern (see next section). Removing pre- and post-earthquake deformation from the coseismic displacements was, and still is, a significant challenge. Even in the era of dense continuous networks, the problem of separating coseismic and postseismic deformation has been pushed down to shorter timescales, as postseismic deformation is observed to begin immediately after the earthquake rupture terminates (Melbourne et al. 2002; Savage and Langbein 2008).

One of the earliest important earthquakes in the GPS era was the 1992 Landers earthquake in southern California. This M_W 7.4 strike-slip earthquake ruptured multiple faults with a complex surface rupture pattern (Sieh et al. 1993) and caused displacements of five continuous GPS sites in the region, which were used to develop slip models for the earthquake (Bock et al. 1993; Blewitt et al. 1993). Later, a much more extensive set of displacements determined from repeated GPS surveys of many survey marks in the region provided almost 100 displacement vectors that were used to develop detailed models of the distribution of slip during the earthquake (e.g., Hudnut et al. 1994; Freymueller et al. 1994). Work on the Landers earthquake demonstrated that even a small number of geodetic displacements provide significant constraints on the total seismic moment of the earthquake and that an extensive displacement field could reveal important details of the slip distribution. This earthquake was also the first large event to be imaged using InSAR (Massonnet et al. 1993).

The problem of estimating the earthquake slip distribution from surface displacement data is similar to that of modeling interseismic deformation. As in the interseismic case, elastic dislocation theory provides the mathematical basis for predicting surface displacements given slip on a fault and the inverse problem of estimating slip given surface displacement. Many studies have used the simplest elastic model, a uniform elastic half-space, while others have included the effect of elastic layering and/or sphericity of the Earth. For the largest earthquakes, a model that treats the Earth as a layered elastic sphere is necessary (Banerjee et al. 2005). The detailed fault geometry must be estimated as part of the inverse problem if not known from other sources. What is most different about the coseismic slip problem is that spatial

variations in the slip always must be estimated, and because the displacements are generally much larger than in the interseismic case, the data can resolve spatial variations in slip to a greater extent.

The most common approach to estimating the slip distribution is to segment the fault plane or planes into a mesh of “subfaults” assumed to each have uniform slip and estimate those slip values. This is done in an inverse approach (in which each column of the matrix used is the predicted elastic displacements from slip on a subfault), subject to smoothing between subfaults or other regularization to make the problem invertible. The subfaults should in principle be smaller than the spatial resolution of the data, to avoid oversmoothing possible small-scale variations in slip by the artificial choice of where the subfault boundaries are. An alternate approach is to vary the subfault size spatially as a function of the data resolution, so that the model is smooth where the data are sparse but can be rough where data are dense (Pritchard et al. 2002; Barnhart and Lohman 2010). Earthquake slip inversions also commonly use nonlinear optimization techniques, from least squares with inequality constraints (to constrain the slip direction) to full nonlinear inversions that are required when the fault geometry must be estimated (e.g., Ragon et al. 2018).

Bayesian inversion approaches offer an alternative to regularization-based inversion, in which the objective is to estimate not the best-fitting value of each subfault’s slip but rather its probability density function (pdf), by iterating through the model space with a Monte Carlo approach (e.g., Fukuda and Johnson 2008; Minson et al. 2013; Bagnardi and Hooper 2018). The likelihoods and correlations of the parameter estimates can then be directly explored rather than assumed based on an ad hoc regularization. Different regularization approaches based on mathematical constructs, physical properties (e.g., stress drop), or assumptions (e.g., self-similarity of slip) also can be tested within this same framework (Fukuda and Johnson 2008, 2010; Amey et al. 2018).

Large earthquakes cause substantial displacements very far from the fault, up to 1000 km for the 2010 Maule, Chile, and 2011 Tohoku earthquakes and up to 4000 km away for the 2004 Sumatra-Andaman quake. These far-field displacements can bias velocity estimates if not accounted for and can cause distortions of the geodetic reference frame. Tregoning et al. (2013) showed that the large earthquakes of the twenty-first century produced cumulative global displacements >1 mm over almost the entire planet and that the global deformation from these events was large enough to bias site velocities globally at the >0.1 – 0.3 mm/yr. level, if not accounted for. Past realizations of the ITRF have not consistently accounted for even some ~ 10 -mm-level offsets, but the latest ITRF2014 includes the most comprehensive set of coseismic displacements to date, along with postseismic displacement models (Altamimi et al. 2016).

In recent years, there has been a dramatic increase in the use of kinematic (high-rate) GPS to measure the ground motions during earthquakes. Nikolaidis et al. (2001) first demonstrated that kinematic GPS could record displacements due to seismic waves, but they were limited by having only data recorded once every 30 seconds. Larson et al. (2003) measured the propagation of surface waves from the 2002 Denali fault earthquake using GPS data recorded at 1 Hz, which demonstrated the utility of kinematic GPS for “GPS Seismology.” GPS displacement records are much less precise than records from seismometers, but can record large displacements faithfully, are not subject to distortion in the presence of tilts, and do not need external information to resolve integration constants because they provide a displacement record directly. Miyazaki et al. (2004a) determined a time-dependent slip model for the 2003 Hokkaido Tokachi-Oki earthquake in Japan based entirely on kinematic GPS records. This model clearly showed that slip propagated downdip and to the northwest from the hypocenter and also revealed a secondary slip peak to the northeast, which might represent triggered slip. The spatial and temporal resolutions of slip models estimated from kinematic GPS data are comparable to models derived from near-field strong motion recordings and superior to models derived purely from teleseismic recordings (Ji et al. 2004).

Real-Time Coseismic Deformation

Kinematic GPS solutions are increasingly becoming real-time solutions as well, driven by potential applications such as earthquake early warning and tsunami warning (e.g., Ruhl et al. 2017). Given a real-time stream of positions from a geodetic network, earthquake source parameters can be estimated directly from the time series. In the classic formulation of Gutenberg (1945), earthquake magnitude is a function of the peak ground displacement and the distance between the earthquake source and observation point. This relationship has historically been the basis of local magnitude estimation using seismic networks, but the relationship applies to displacement records measured geodetically as well and up to the largest magnitudes yet observed (Crowell et al. 2013; Fang et al. 2014). Geodetic magnitude estimates made in this way can correctly determine the magnitude to within ± 0.1 (moment with $\sim \pm 10\%$), which is comparable to the best magnitude precision from seismology, and can be made much more rapidly for the largest earthquakes.

Great progress has been made in the last decade in the development of real-time source model estimation. This requires multiple steps, some of which may be provided by or aided by seismic networks: (1) event detection, (2) displacement estimation, and (3) model estimation. Ohta et al. (2012) demonstrated an end-to-end system for all three steps. In other systems, the event detection can be triggered by notification from a seismic earthquake early warning system, and developments then focus on the displacement estimation

and/or source model estimation (e.g., Crowell et al. 2012; Minson et al. 2014; Grapenthin et al. 2014). As of 2019, test operational systems have been developed to support early warning and rapid situational awareness, and they performed well during the 2019 Ridgecrest earthquake in California (Melgar et al., submitted manuscript).

Postseismic Deformation

Postseismic deformation is the general term used for any kind of post-earthquake deformation induced by the stress changes in the earthquake. Tectonic strain is accommodated both by earthquakes in the upper crust and concomitantly by gradual aseismic deformation processes elsewhere in the crust and upper mantle (e.g., ductile flow and creep). The stress changes caused by earthquake slip perturb these aseismic processes, resulting in a strain pattern at the surface that evolves in space and time as dictated by the rheology of the crust and upper mantle (Bürgmann and Dresen 2008). Therefore, given sufficiently good data, postseismic deformation provides an opportunity to study this rheology and with it the dynamics of tectonic deformation (whereas interseismic deformation allows only a kinematic description).

Postseismic deformation is thought to result from a superposition of three main physical mechanisms, which operate on different spatial and temporal scales: viscoelastic relaxation in the mantle and possibly lower crust, afterslip on the very shallow or deep parts of the fault zone, and poroelastic relaxation. Poroelastic relaxation – fluid flow-driven deformation that relieves pressure differences caused by the earthquake – is generally considered to be important only very close to the fault and near geometric complexities such as fault stepovers and bends where dilatational strains are large (and fluid flow changes will thus be most significant). Shallow afterslip is a frictional process, first observed after the 1966 Parkfield earthquakes (Smith and Wyss 1968) and observed after many subsequent earthquakes, including subduction zone events (Hsu et al. 2006; Kreemer et al. 2006). Depending on the rheology of deep fault zones, deep afterslip may be a frictional process, a ductile shear process, or some combination of the two. Deep afterslip generally occurs immediately down-dip of the coseismic rupture zone.

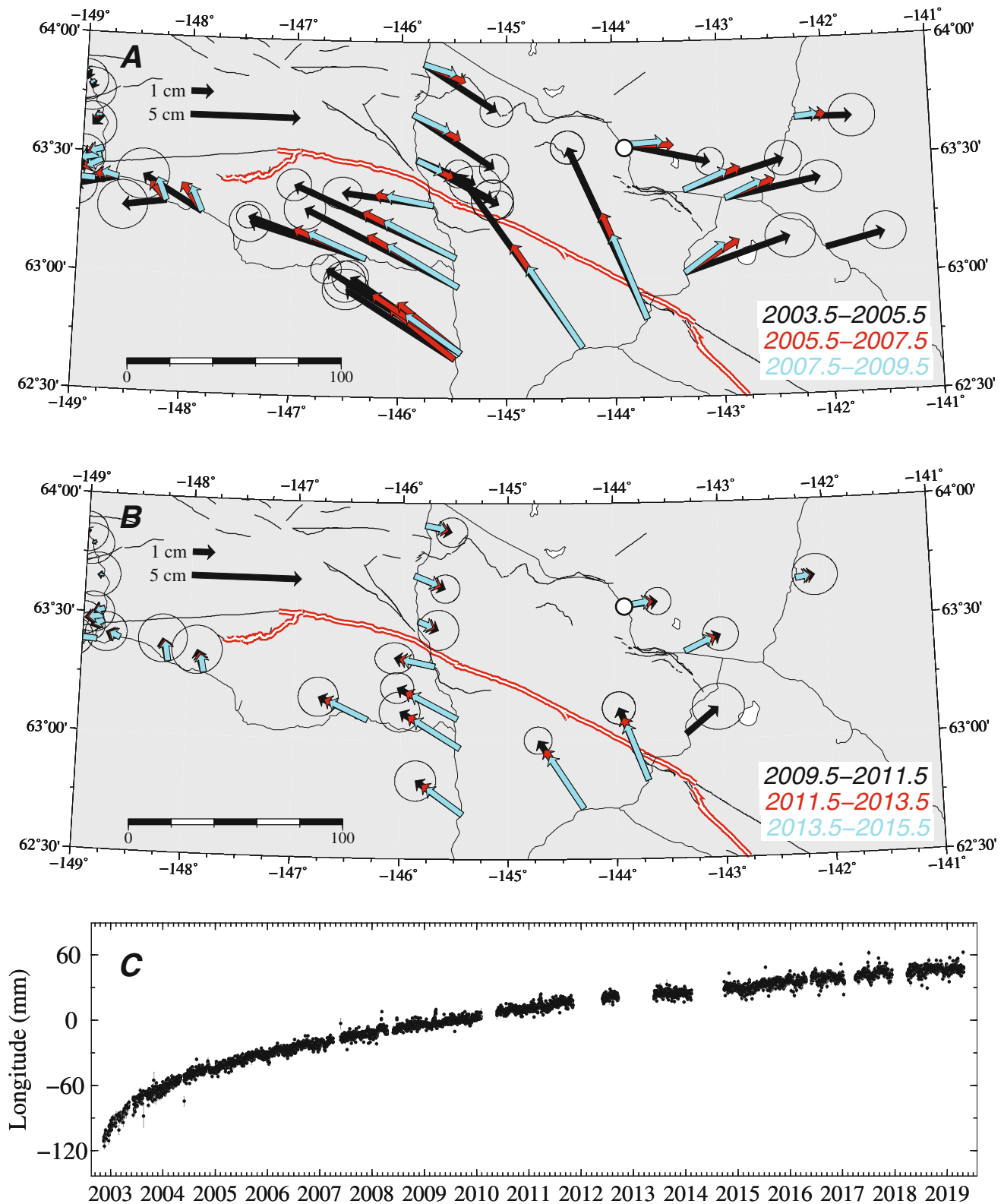
Postseismic deformation can be large and long-lasting, although there is great variation from earthquake to earthquake. Because it tends to occur deeper and often is more spatially distributed than coseismic deformation, postseismic displacements are usually significantly smaller than coseismic displacements close to the fault but can be comparable to them in the far field.

In the first few years after an earthquake, afterslip and viscoelastic relaxation can produce similar deformation pattern at the surface, especially in the horizontal components,

which can make them difficult to separate given limited data. As a result, controversy over postseismic deformation models can linger for years, even for well-studied earthquakes. Afterslip decays more rapidly with time than viscoelastic relaxation (e.g., Suito and Freymueller 2009), so after a long time viscoelastic relaxation can be expected to dominate the post-seismic signal. Empirical kinematic afterslip models can be estimated from data in the same way that coseismic slip models are estimated. Afterslip also can be simulated in dynamic models as creep that accommodates the coseismic stress changes as per an assumed physical creep rheology, such as rate and state friction, and these models can be combined with viscoelastic relaxation (e.g., Barbot et al. 2009; Johnson et al. 2009). Key decisions in this kind of modeling include which parts of the fault are allowed to experience stress-driven afterslip and the 3D elastic and/or viscosity structure. Stress-driven afterslip models can be compared with kinematic afterslip models to test and validate these decisions (e.g., Zhao et al. 2017; Wang and Fialko 2018). Another complication is that the appropriate rheological model varies from place to place, and parameters such as mantle viscosity are both heterogeneous and poorly constrained (Bürgmann and Dresen 2008). As a result of these factors, many postseismic studies still make assumptions that certain displacements result from only the effect of one mechanism or attempt to use the spatial or temporal character of the observed displacements to constrain the mechanisms separately.

The 2002 Denali fault earthquake provides a useful example. Ten continuous GPS sites were set up within a few weeks of the earthquake, and nearly 100 sites were surveyed repeatedly over the next several years, providing a rich data set in space and time (Fig. 4). Both the spatial and temporal variations in postseismic displacements require more than one mechanism to explain them. All sites show an initially rapid rate of deformation that decayed over time. In general, site velocities were ~20 times faster than the pre-earthquake rates in the first 1–2 years, remained several times higher at many sites for several years, and remain significantly different even 15 years after the earthquake. A combination of empirical relaxation functions can explain the temporal decay of the displacements quite well as long as the model fit to the time series has both a short relaxation component (characteristic decay time of ~months) and a long relaxation component (characteristic decay time of a few years).

Postseismic models for the 2002 Denali fault earthquake have called upon a variety of mechanisms to explain the deformation. Freed et al. (2006a) showed that no single mechanism could explain the observed displacements and proposed a model combining afterslip within the lithosphere (approximated as a narrow low viscosity shear zone) and viscoelastic relaxation in the asthenosphere. Freed et al. (2006b) analyzed the time dependence at far-field sites and



GPS, Tectonic Geodesy, Fig. 4 Postseismic displacements following the November 2002 Denali fault earthquake. (a, b) Displacements are shown for each 2-year interval during the time periods 2003.5–2009.5 (a, top panel) and 2009.5–2015.5 (b, middle panel), with the color of the vectors indicating the 2-year interval (not all sites have data for all

intervals). Error ellipses (95% confidence) are shown only for the first interval in each panel for clarity. The pre-earthquake motions have been subtracted from both figures, so the displacements represent the transient motion only. The 2002 surface rupture is shown in white outlined with red. (c, bottom panel) The time series of eastward motion of the site

concluded that a model with a nonlinear (power law) viscosity in the mantle, consistent with lab measurements of the rheology of olivine, explained these data better than a model with a linear Newtonian viscosity. Johnson et al. (2009) estimated a combined afterslip and viscoelastic model, using an assumed frictional model. However, none of these models have been very successful in predicting displacements into the future, beyond the time span of the data used to constrain them (Freymueller et al. 2009). These model shortcomings likely result from a combination of deficiencies in the estimated coseismic slip model, from a failure to separate the different mechanisms correctly or from additional processes not considered in the models, possibly including Denali earthquake-induced changes in the slip rates on the subduction interface in southern Alaska (Johnson et al. 2009).

Similar patterns of postseismic deformation have been observed in many other cases, but individual earthquakes have displayed a puzzling variety of postseismic responses. For some large subduction earthquakes, postseismic transients seem to have decayed away after only a few years, while in other cases, small transient signals have lingered for decades. After the 1995 M_w 8.1 Antofagasta (Chile) earthquake, there was significant afterslip but apparently little viscoelastic relaxation of the mantle, as postseismic deformation rates returned to pre-earthquake rates within a few years (Melbourne et al. 2002). The 2003 Hokkaido Tokachi-Oki earthquake caused afterslip downdip of the coseismic rupture, which also extended at depth for a significant lateral distance north of the rupture zone (Miyazaki et al. 2004b). In the 2006–2007 Kuril Island doublet, a subduction thrust earthquake followed by a similar-sized outer rise event, afterslip appears to have been limited to the first several months after the thrust earthquake (Steblov et al. 2008; Kogan et al. 2011). The cause of this wide range of behavior is not fully understood, but likely reflects heterogeneity in material properties. For the largest earthquakes, significant postseismic deformation clearly can last for decades (Wang 2007); postseismic deformation from the great 1960 Chile and 1964 Alaska earthquakes remains very large (rates of 1–2 centimeters per year) decades after those large events (e.g., Suito and Freymueller 2009; Wang et al. 2012).

The 2011 Tohoku earthquake is another key event, because of the unprecedented scope of observations of postseismic deformation (see Wang et al. 2018 for a more complete summary). The dense data sets available onshore and offshore, as well as from the Asian mainland, have allowed authors to develop detailed models of the 3D rheological structure,

including weak zones within the lithosphere of the Japanese islands, variations in the asthenospheric viscosity, the spatial extent of the mantle wedge, and other factors (e.g., Sun et al. 2014; Hu et al. 2016). However, even with an unprecedented data set, competing studies still disagree on a number of details, and continued work with additional years of data will be required to fully resolve the remaining conflicts.

Because of the long duration of transient postseismic deformation and the short length of precise geodetic measurements in most places, it is not clear whether postseismic transients decay away to a relatively steady background deformation rate or whether deformation rates are never truly steady in time. This question has important implications for the dynamics of Earth deformation. Decay to a steady background rate suggests that tectonic deformation in the mantle and lower crust is localized due to lateral variations in properties and/or nonlinear viscoelasticity. The long-term fault motion rate and background deformation rate would represent a balance between tectonic forces and the effective viscosity of the weak zone beneath the fault. On the other hand, a uniform linear viscoelastic mantle requires that shear in the mantle be broadly distributed rather than localized beneath faults. In this case, surface deformation would never be steady in time except when the earthquake recurrence time and mantle relaxation time are roughly equal, and transient deformation would always extend a great distance away from every active fault, even a long time after the last major earthquake (Savage and Prescott 1978). The apparently linear “interseismic” phase would be essentially an illusion caused by our short measurement history. However, our ability to measure steady rigid plate motions that are very close to long-term geological plate motion estimates argues in favor of some degree of localization of strain in the mantle beneath faults, as does the observation that interseismic motions are localized around faults like the North Anatolian Fault in Turkey regardless of the time since the last earthquake (Hearn et al. 2009). This is also supported by the frequent consistency of geologic estimates of fault slip rates and geodetic estimates based on simple linear models. However, this question is not yet settled and probably will remain a focus of research for some time.

Volcano Deformation

Unlike tectonic deformation, which results mainly from slip on faults (shear sources), volcanic deformation results mainly

GPS, Tectonic Geodesy, Fig. 4 (continued) JANL (shown with a white dot in panels **a** and **b**) relative to North America illustrates the nonlinear character of postseismic deformation. Although the time series

appears to be nearly linear several years after the earthquake, there is actually significant curvature of the time series over the entire time period

from volumetric or dilatational sources, such as the expansion or contraction of a dike or sill or other magma body. Injection of magma into the subsurface through intrusion, or removal of magma from the subsurface as it is erupted, causes subsurface pressure changes that result in deformation at the surface. Volcanic deformation commonly results in horizontal displacements that are oriented radially away from or toward the volcano and includes significant subsidence or uplift (Fig. 5). Some volcanoes have produced very large deformation signals associated with the intrusion of magma before eruption, while others have produced only small deformation signals even as the result of large eruptions. The heterogeneity of behavior is a result in part of magma properties that vary with magma composition and gas content (e.g., compressibility). Generally, basaltic volcanoes like Kilauea, Okmok, and Etna (e.g., Neri et al. 2009; Aloisi et al. 2018) produce larger and more predictable surface deformation than more silicic volcanoes.

The simplest volcanic source model is a spherical point pressure source, usually called the Mogi model after the Japanese scientist who first applied it (Mogi 1958). This model describes the deformation at the surface caused by a pressure change in a small spherical body in an elastic half-space; the point source approximation actually holds even for relatively large bodies. The Mogi model provides a surprisingly good fit to observed displacements from many volcanic sources. Models for finite spherical or elliptical sources have also been developed; the finite spherical source is quite similar to the Mogi source except for very shallow sources, but elliptical sources produce a different pattern of displacements, still radial but with asymmetry in the amplitudes. Dikes and sills, planar bodies of magma that are associated with the opening and filling of planar cracks, can be modeled using the same dislocation theory that is used for shear faulting, using the equations for opening-mode deformation.

Volcanic deformation is inherently time-dependent, and volcanoes display episodes of both inflation (increasing pressure) and deflation (decreasing pressure). Noneruptive deflationary episodes probably result from a pressure drop caused by the leaking out of exsolved gases or from the depressurization of a hydrothermal system. During eruption, most volcanoes show deformation associated with decreasing pressure, although in a few cases there have been observations of simultaneous deflation at depth and inflation close to the surface. One example of this is the 2000 eruption of Usu volcano in Japan, for which deformation and gravity change data were fit best by a model with a deflation source at ~4 km depth, connected to the surface by an intruded fissure (Jousset et al. 2003). If the magma is assumed to be incompressible, such as for a magma that contains no exsolved gas bubbles, then the volume of magma added or removed to the subsurface is proportional to the pressure change in the reservoir or the opening or closing of a dike or sill, and thus the

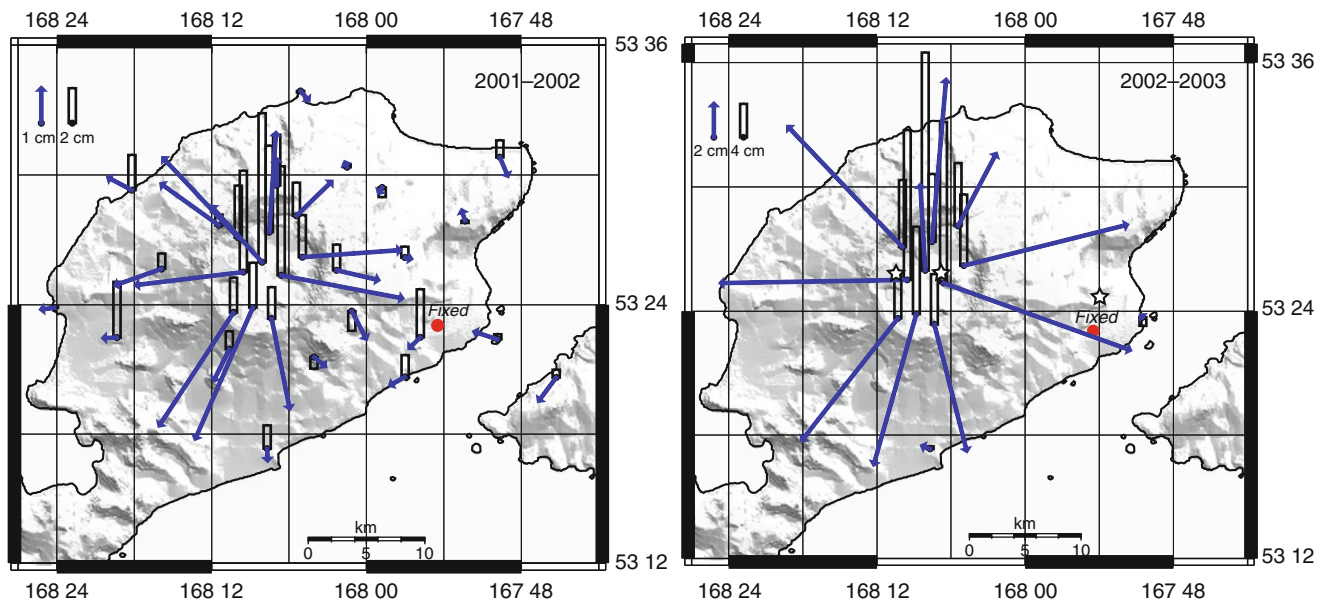
deformation magnitude is proportional to magma volume. In general the magma will be compressible, and the deformation magnitude will be a function of the magma volume change, magma compressibility, and the elastic properties of the surrounding rock.

Recent modeling approaches attempt to better characterize the chemical and physical state of the magma and surrounding material. Some models attempt to incorporate geodetic, petrologic, and other data along with physics-based relations to describe rates of magma supply and how the magma changes before or during eruptions (e.g., Anderson and Segall 2011; Anderson and Poland 2016).

Unlike the tectonic problem, it seems that a simple “eruption cycle” model does not apply to volcanoes. The idea is appealing: magma would accumulate in the subsurface until the pressure reaches some critical level, which causes failure of the overlying rock, opening of a conduit to the surface, and then eruption. However, observations show that the magma supply is not uniform in time (Dvorak and Dzurisin 1993), while tectonic loading is, and successive eruptions at the same volcano can be quite different in volume and style. Even at volcanoes that erupt frequently, there does not appear to be a clear relationship between the volume of magma that accumulates between eruptions and the volume that erupts. This reflects the fact that the crust beneath active volcanoes may be occupied by a number of bodies of crystal mush or partially solidified magma and also the importance of gases and volatiles in triggering magma rise and eruption. A small intrusion of fresh, hot magma into a cooler body can trigger exsolution of dissolved volatiles and remobilization of the magma, causing a pressure increase that may be capable of triggering an eruption.

Okmok volcano in the Eastern Aleutian arc (USA) experienced substantial accumulation of new magma at shallow depth (Fig. 5) following its 1997 eruption, causing as much as several decimeters of displacement at the surface (Lu et al. 2005; Fournier et al. 2009), but its 2008 eruption was preceded by a period of ~3 years of relative quiescence. Albright et al. (2019) used a data assimilation approach with finite element modeling of the Okmok magma system to develop a hindcast of the eruption. This approach tracks not only the volume and pressure within the magma reservoir but also the stress state of the surrounding material. It successfully identified potential points of failure within the surrounding rock that could explain why the 2008 eruption created a new vent close to the magma reservoir rather than following the path of earlier eruptions.

Similarly, although rapid deformation at a volcano is an indication that the volcanic system is active, many episodes of deformation may occur between eruptions. Some large calderas, such as Campi Flegrei in Italy, Yellowstone and Long Valley in the USA, and Rabaul in Papua New Guinea, have experienced episodes of substantial inflation, deflation, or



GPS, Tectonic Geodesy, Fig. 5 Displacements from Okmok volcano, Eastern Aleutian arc, Alaska, showing significant volcanic deformation. The left panel shows displacements from 2001–2002, and the right panel from 2002–2003. The scale for vertical motion is different from that of the horizontal, and there is a factor of two difference in scale between the

two plots, a result of the much larger deformation in 2002–2003. The blue arrows are horizontal displacements, and the white bars show vertical displacements. Data are from Miyagi et al. (2004). The deformation results from the pressurization of a magma body located about 3 km below the surface

both over a period of many years (Dvorak and Mastrolorenzo 1991; Dvorak and Dzurisin 1997). Some of this deformation may result from pressure changes in hydrothermal systems associated with the volcano, rather than from changes in the magma reservoirs themselves. Other recent eruptions had only subtle or nonexistent deformation prior to the eruption, including the eruptions of St. Helens in 2004, Augustine in 2006, and Redoubt in 2009 in the USA (Cervelli et al. 2006, 2010; Lisowski et al. 2008; Grapenthin et al. 2013). However, significant precursory deformation very close to the eruptive vent certainly must happen. A notable example came from a continuous GPS site that was installed on Stromboli volcano (Italy), which showed dramatic displacements in the minutes before it was destroyed by an eruption from a new vent located at a very short distance away (Mattia et al. 2004). Such deformation, associated with the final rise of magma to the surface, is very localized.

Kilauea volcano in Hawaii showed distinctive and reliable patterns of deformation during its most recent 1983–2018 eruption, in which eruptive activity was centered at Pu‘u ‘O‘o on Kilauea’s middle East Rift Zone and from 2008 onward also at the Kilauea summit. Eruptions at Pu‘u ‘O‘o were preceded by inflation at the summit, which then deflated as a pulse of magma was intruded into the rift zone and proceeded downrift (Dvorak and Dzurisin 1993, 1997). Pressure changes and lava lake levels were closely coupled between the summit and at Pu‘u ‘O‘o, indicating an efficient hydraulic connection (Patrick et al. 2019). In April–May 2018, an M6.9 earthquake occurred and a dike was intruded

downrift into the lower East Rift Zone, beyond the extent of the 1983–2018 activity, and a new and voluminous eruption began out of a vent that formed within a residential neighborhood (Neal et al. 2018). This eruption substantially drained the summit reservoir, producing $\sim 0.8 \text{ km}^3$ of new lava from the vent and a collapse structure of similar volume at the summit.

Glacial Isostatic Adjustment and Sea Level Studies

The Earth also deforms in response to changes in surface loads, for example, changes in the mass of glaciers, ice fields, and ice sheets on the surface. The best-known example of this deformation is the dramatic uplift that has occurred in the regions formerly covered by the continental ice sheets in North America and Europe. The ongoing vertical motions in these regions, termed glacial isostatic adjustment (GIA), are dominantly driven by the viscoelastic response of the mantle to the unloading as the ice sheets collapsed after the Last Glacial Maximum (LGM), about 23,000 years ago. While the ice sheets were growing, the mass of the ice depressed the surface and induced stresses that caused the underlying mantle to flow out of the way. Upon collapse of the ice sheets, the process was reversed and mantle is still flowing back underneath at present. When the ice was in place, a flexural bulge formed in a ring around the glaciated area; now that forebulge is collapsing, producing subsidence across regions like the Eastern USA. In addition to the long-lasting effects of

the great ice sheets, deformation due to present-day ice mass changes can cause even more rapid deformation even though the ice mass changes are much smaller. Additional constraints on present-day ice mass changes come from in situ observations and airborne or satellite measurements of surface height change or gravity change.

Displacements from load changes are predominantly vertical. Horizontal displacements are usually several times smaller than the vertical displacements but are still significant relative to present geodetic measurement errors (e.g., Kreemer et al. 2018; Ding et al. 2019). Horizontal displacements are more sensitive than the vertical displacements to assumptions made in the computations (compressibility vs. incompressibility, radially symmetric Earth structure vs. lateral variations, etc.). Thus, while model predictions for the vertical motions are in general agreement, the best models for the horizontal motions remain a matter of considerable debate.

Conceptually, it can be useful to break up the deformation due to load changes into two components: an instantaneous elastic response and a time-delayed viscoelastic response (GIA). However, in areas with low mantle viscosities such as Iceland, Alaska, Patagonia, and West Antarctica, the viscoelastic response is so rapid that these two parts must be modeled together (e.g., Larsen et al. 2005; Auriac et al. 2013; Nield et al. 2014; Richter et al. 2016; Barletta et al. 2018). The theory for computing the elastic response of the Earth dates back to Love in the early twentieth century; Farrell (1972) used this theory to compute the response of a realistic Earth model to both point loads and disk loads. The viscoelastic problem results in time-dependent deformation, which can be computed in a variety of ways, either in terms of viscoelastic Love numbers or viscoelastic Green's functions for point or finite loads or using finite element methods.

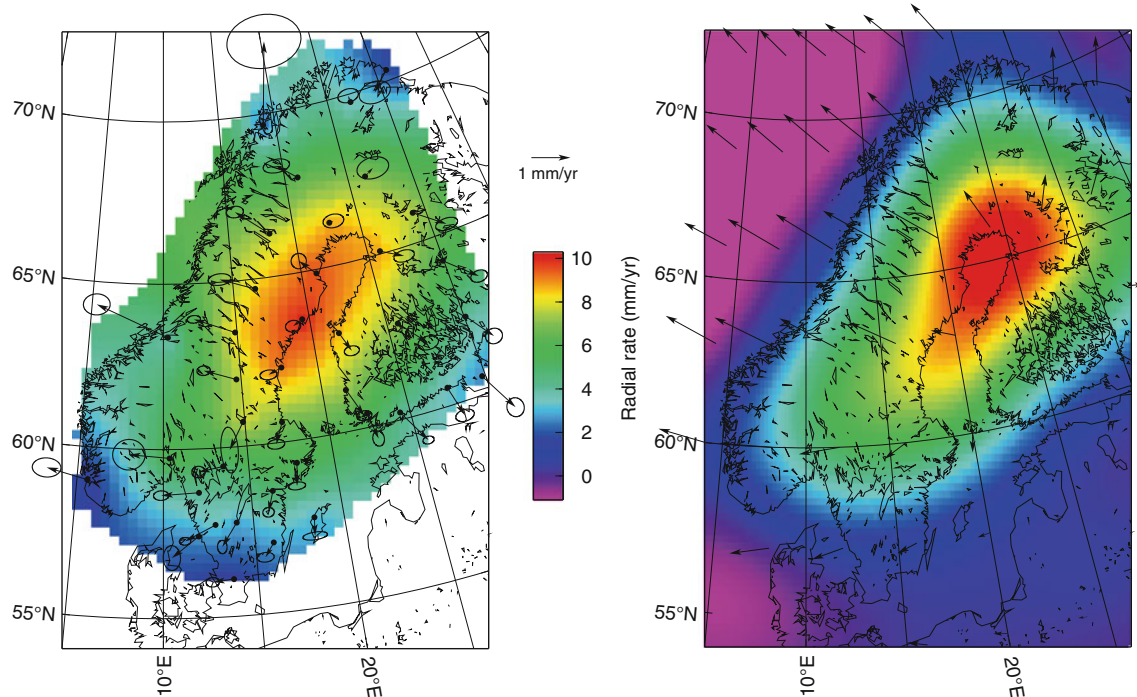
Global models of deglaciation and GIA date back to the 1970s, with the development of models for the viscoelastic response of a 1D (radially stratified) spheroid and the development of the sea level equation to account for the gravitationally consistent redistribution of ocean water (e.g., Peltier and Andrews 1976; Clark et al. 1978). Models of post-LGM GIA were initially constrained by relative sea level histories, and more recently including geodetic data such as the gravity field and gravity change, and present-day displacement rates. Recent model developments and improvements include accounting for rotational feedback in the shape of Earth and changes in the paleo-shoreline position for improved description of loading on continental shelves. The most influential and commonly used models over the last few decades have been the series of ICE models from the University of Toronto (currently, ICE-6GD) and their corresponding mantle viscosity models (Peltier et al. 2015, 2018). However, the viscosity structure of the deep mantle remains a matter of considerable dispute, with different families of viscosity models having

values that can differ by an order of magnitude in the deep mantle (e.g., Lau et al. 2016). Increasingly, new studies are moving toward 3D mantle viscosity models rather than 1D radial models (e.g., van der Wal et al. 2015).

The BIFROST project in Scandinavia was initiated in the early 1990s to study GIA resulting from the post-LGM collapse of the Fennoscandian ice sheet and demonstrated the capability of GPS to measure accurate vertical motions (Johansson et al. 2002; Milne et al. 2004). The observations revealed the expected pattern of deformation, with the largest uplift rates (12–13 mm/yr.) located on the northwest margin of the Baltic Sea where the LGM ice sheet had its maximum thickness (Fig. 6). There have been fewer geodetic studies of GIA from the Laurentide ice sheet in North America (e.g., Sella et al. 2007). More extensive vertical motion data were included in the global ICE-6G model (Peltier et al. 2015) and the regional ICE-7G_NA model (Roy and Peltier 2017). The scale of the GIA deformation in North America poses an interesting problem: the estimates of horizontal motion relative to North America depend on the estimated motion of the North American plate in ITRF, but it is difficult to measure the motion of North America without having that estimate biased by horizontal GIA (Kreemer et al. 2018; Ding et al. 2019). As a result, independent estimates of the motion of the North American plate tend to differ by more than would be expected from data noise alone, and this has also made it difficult to evaluate the horizontal predictions of GIA models.

The most rapid uplift rates from GIA come from regions of present-day ice mass loss. Ice mass loss in Alaska and surrounding parts of Canada in North America and in Patagonia in South America has been very large since the end of the Little Ice Age (LIA) ~200 years ago, large enough to account for 10–20% of twentieth century global sea level rise (Dyurgerov and Meier 2005; Larsen et al. 2007; Lemke et al. 2007; Berthier et al. 2010). In both of these regions, measured uplift rates exceed 30 mm/yr. at their peak, and significant uplift has been observed in a broad area surrounding the regional ice fields. In Alaska, the observed uplift depends significantly on both the ongoing regional ice mass loss and the nineteenth century deglaciation of Glacier Bay (Larsen et al. 2005). Deglaciation rate is changing with time, generally accelerating but with considerable variation in response from glacier to glacier (Larsen et al. 2015; Hu and Freymueller 2019).

Mass loss in the Greenland and Antarctic ice sheets has begun or accelerated within the last decade, and the changes in mass loss rate are reflected in changes in the vertical motions. Southeast Greenland made the transition from no or minimal uplift to significant uplift in the early 2000s, while the onset of uplift in northwest Greenland began a few years later (Khan et al. 2008, 2010). The uplift history measured by GPS is in agreement with the mass loss history estimated by gravity observations of the GRACE satellite mission and is



GPS, Tectonic Geodesy, Fig. 6 Observations and model predictions for GIA in the Fennoscandian region, after Milne et al. (2001) using the BIFROST solution of Lidberg et al. (2007). (*left panel*) Horizontal velocities relative to the Eurasian plate (arrows, with 95% confidence ellipses) and vertical velocities (background color). (*right panel*) Model

predictions based on ICE-3G with lithosphere thickness 120 km, upper mantle viscosity $5e20$ Pa s, and lower mantle $8e21$ Pa s. This figure was provided by James L. Davis, Harvard-Smithsonian Center for Astrophysics

corroborated by direct observation and measurement of glacier mass balance in the area. The geographic focus of mass loss in the Greenland ice sheet has evolved over the last two decades, and while mass loss before 2010 was driven dominantly by the impact of ocean warming on tidewater glaciers, more recent mass loss has been increasingly driven by surface melting (Bevis et al. 2019). Measurements in Antarctica show extremely rapid rates of uplift in the Amundsen Sea Embayment (ASE) and Antarctic Peninsula (e.g., Nield et al. 2014; Barletta et al. 2018). Uplift rates across the rest of the continent (e.g., Sasgen et al. 2018) are substantially lower.

As glaciers, ice fields and ice sheets lose mass, most of the water ends up in the ocean where it raises sea level. Tide gauges measure relative sea level, which is the level of the ocean surface relative to the land surface. Estimates of twentieth century global sea level rise depend on averaging over a global set of tide gauges where uplift or subsidence of the land is negligible, but more data could be used to analyze sea level variations if land uplift rates at tide gauges were precisely known. Wöppelmann et al. (2009) used GPS measurements at a globally distributed set of tide gauges to correct for vertical motions of the land and estimated an average global sea level rise of 1.6 ± 0.2 mm/yr., assuming that the vertical motions observed by GPS over the last decade are representative of the last 100 years. This estimate agrees with previous estimates

for twentieth century sea level rise. However, their uncertainty estimate did not account for the uncertainty in the geocenter stability of the ITRF, which was actually the dominant error source. Wöppelmann and Marcos (2016) updated these results but pointed out that the accuracy of the terrestrial reference frame currently limits our ability to measure global average sea level using GPS to the ~ 0.5 mm/yr. level.

Sea level does not rise uniformly around the globe when there is a significant redistribution of mass resulting from the melting of ice. Spatial variations in sea level rise result from changes in the geoid caused by the mass redistribution; sea level follows the geoid, so if the geoid height is increasing over a section of ocean, then local sea level would rise in a geocentric reference frame (Mitrovia et al. 2001; Tamisiea et al. 2001). The difference between relative sea level variations measured by tide gauges and vertical motions measured by GPS is a combination of vertical motion of the land, sea level rise due to the addition of water to the oceans, ocean thermal expansion, and the local geoid change. Geoid rate changes at the multiple mm/yr. level are predicted for areas of significant regional mass loss, such as Alaska and Patagonia, and significant geoid height trends have been measured using GRACE (Tamisiea et al. 2001, 2005). Much larger variations in the geoid would appear in the future if the Greenland and Antarctic ice sheets melt significantly (Mitrovia et al. 2009).

Future studies will integrate data and models for all of these effects to provide a complete and self-consistent picture of sea level, gravity change, and height change around the world's coastlines.

Summary

High-precision geodesy is an important tool for the study of active tectonics and other causes of crustal deformation. Motion and deformation of points on the surface of the Earth result from a combination of tectonic plate motions, elastic deformation from the slip deficit associated with faults that are locked by friction, and coseismic and postseismic displacements. Additional causes of deformation in some places include pressure changes within volcanic systems and past and present surface load changes, most importantly from the melting of ice in glaciers, ice fields, and ice sheets. Deformation due to several sources may be superposed in some locations, such as southern Alaska and Patagonia. Surface displacements can be related to the underlying sources using elastic or viscoelastic deformation models, the most important of which are based on dislocation theory.

Cross-References

- [Geodesy, Ground Positioning, and Levelling](#)
- [Geodesy, Networks, and Reference Systems](#)
- [GPS, Data Acquisition, and Analysis](#)
- [SAR Interferometry](#)
- [Very Long Baseline Interferometry](#)

Bibliography

- Albright JA, Gregg PM, Lu Z, Freymueller JT (2019) Hindcasting magma reservoir stability preceding the 2008 eruption of Okmok, Alaska. *Geophys Res Lett* 46. <https://doi.org/10.1029/2019GL083395>
- Aloisi M, Bonaccorso A, Cannavò F, Currenti GM (2018) Coupled short- and medium-term geophysical signals at Etna volcano: using deformation and strain to infer magmatic processes from 2009 to 2017. *Front Earth Sci*. <https://doi.org/10.3389/feart.2018.00109>
- Altamimi Z, Rebischung P, Métivier L, Collilieux X (2016) ITRF2014: a new release of the international terrestrial reference frame modeling nonlinear station motions. *J Geophys Res Solid Earth* 121:6109–6131. <https://doi.org/10.1002/2016JB013098>
- Altamimi Z, Métivier L, Rebischung P, Rouby H, Collilieux X (2017) ITRF2014 plate motion model. *Geophys J Int* 209:1906–1912. <https://doi.org/10.1093/gji/ggx136>
- Amey RMJ, Hooper A, Walters RJ (2018) A Bayesian method for incorporating self-similarity into earthquake slip inversions. *J Geophys Res Solid Earth* 123:6052–6071. <https://doi.org/10.1029/2017JB015316>
- Anderson KR, Poland MP (2016) Bayesian estimation of magma supply, storage, and eruption rates using a multiphysical volcano model: Kīlauea volcano, 2000–2012. *Earth Planet Sci Lett* 447:161–171. <https://doi.org/10.1016/j.epsl.2016.04.029>
- Anderson K, Segall P (2011) Physics-based models of ground deformation and extrusion rate at effusively erupting volcanoes. *J Geophys Res* 116(B7):20. <https://doi.org/10.1029/2010JB007939>
- Argus DF, Heflin MB (1995) Plate motion and crustal deformation estimated with geodetic data from the global positioning system. *Geophys Res Lett* 22:1973–1976
- Argus DF, Gordon RG, Heflin MB, Ma C, Eanes RJ, Willis P, Peltier WR, Owen SE (2010) The angular velocities of the plates and the velocity of Earth's centre from space geodesy. *Geophys J Int* 181:1–48. <https://doi.org/10.1111/j.1365-246X.2009.04463.x>
- Auriac A, Spaans KH, Sigmundsson F, Hooper A, Schmidt P, Lund B (2013) Iceland rising: solid Earth response to ice retreat inferred from satellite radar interferometry and viscoelastic modeling. *J Geophys Res Solid Earth* 118:1331–1344. <https://doi.org/10.1002/jgrb.50082>
- Bagnardi M, Hooper A (2018) Inversion of surface deformation data for rapid estimates of source parameters and uncertainties: a Bayesian approach. *Geochem Geophys Geosyst* 19:2194–2211. <https://doi.org/10.1029/2018GC007585>
- Banerjee P, Pollitz FF, Bürgmann R (2005) The size and duration of the Sumatra-Andaman earthquake from far-field static offsets. *Science* 308:1769–1772
- Barbot S, Fialko Y, Bock Y (2009) Postseismic deformation due to the mw 6.0 2004 Parkfield earthquake: stress-driven creep on a fault with spatially variable rate-and-state friction parameters. *J Geophys Res* 114:B07405. <https://doi.org/10.1029/2008JB005748>
- Barletta VR, Bevis M, Smith BE, Wilson T, Brown A, Bordon A, Willis M, Khan SA, Rovira-Navarro M, Dalziel I, Smalley R Jr, Kendrick E, Konfal S, Caccamise DJ II, Aster RC, Nyblade A, Wiens DA (2018) Observed rapid bedrock uplift in Amundsen Sea Embayment promotes ice-sheet stability. *Science* 360:1335–1339. <https://doi.org/10.1126/science.aao1447>
- Barnhart WD, Lohman RB (2010) Automated fault model discretization for inversions for coseismic slip distributions. *J Geophys Res* 115: B10419. <https://doi.org/10.1029/2010JB007545>
- Bartlow NM, Wallace LM, Beavan RJ, Bannister S, Segall P (2014) Time-dependent modeling of slow slip events and associated seismicity and tremor at the Hikurangi subduction zone, New Zealand. *J Geophys Res Solid Earth* 119:734–753. <https://doi.org/10.1002/2013JB010609>
- Bevis M, Harig C, Khan SA, Brown A, Simons FJ, Willis M, Fettweis X, van den Broeke MR, Madsen FB, Kendrick E, Caccamise DJ II, van Dam T, Knudsen P, Nylen T (2019) Accelerating changes in ice mass within Greenland, and the ice sheet's sensitivity to atmospheric forcing. *PNAS* 116(6):1934–1939. <https://doi.org/10.1073/pnas.1806562116>
- Bird P (2003) An updated digital model of plate boundaries. *Geochem Geophys Geosyst* 4(3):1027. <https://doi.org/10.1029/2001GC000252>
- Blewitt G, Heflin M, Hurst K, Jefferson D, Webb F, Zumberge J (1993) Absolute far-field displacements from the 28 June 1992 landers earthquake sequence. *Nature* 361:340–342
- Bock Y, Agnew DC, Fang P, Genrich JF, Hager BH, Herring TA, Hudnut KW, King RW, Larsen S, Minister JB, Stark K, Wdowinski S, Wyatt FK (1993) Detection of crustal deformation from the landers earthquake using continuous geodetic measurements. *Nature* 361:337–340
- Bonafede M, Strehlau J, Ritsema AR (1992) Geophysical and structural aspects of fault mechanics – a brief historical overview. *Terra Nova* 4:458–463
- Bürgmann R, Dresen G (2008) Rheology of the lower crust and upper mantle: evidence from rock mechanics, geodesy, and field observations. *Annu Rev Earth Planet Sci* 36:531–567. <https://doi.org/10.1146/annurev.earth.36.031207.124326>
- Bürgmann R, Segall P, Lisowski M, Svarc JL (1997) Postseismic strain following the 1989 Loma Prieta earthquake from GPS and leveling measurements. *J Geophys Res* 102:4933–4955

- Calais E, Han JY, DeMets C, Nocquet JM (2006) Deformation of the North American plate interior from a decade of continuous GPS measurements. *J Geophys Res* 111:B06402. <https://doi.org/10.1029/2005JB004253>
- Cervelli P, Fournier T, Freymueller JT, Power J (2006) Ground deformation associated with the precursory unrest and early phases of the January 2006 Eruption of Augustine Volcano, Alaska. *Geophys Res Lett* 33:L18304. <https://doi.org/10.1029/2006GL027219>
- Cervelli P, Fournier TJ, Freymueller JT, Power JA, Lisowski M, Pauk BA (2010) Geodetic constraints on magma movement and withdrawal during the 2006 Eruption of Augustine Volcano. In: Power JA, Coombs ML, Freymueller JT (eds) *The 2006 Eruption of Augustine Volcano, Alaska*, U.S. Geological Survey professional paper, vol 1769. U.S. Geological Survey, Reston
- Chen Q, Freymueller J, Wang Q, Yang Z, Xu C, Liu J (2004) A deforming block model for the present-day tectonics of Tibet. *J Geophys Res* 109(B1):B01403. <https://doi.org/10.1029/2002JB002151>
- Chlieh M, Avouac JP, Sieh K, Natawidjaja DH, Galetzka J (2008) Heterogeneous coupling of the Sumatran megathrust constrained by geodetic and paleogeodetic measurements. *J Geophys Res* 113: B05305. <https://doi.org/10.1029/2007JB004981>
- Chlieh M, Mothes PA, Nocquet J-M, Jarrin P, Charvis P, Cisneros D, Font Y, Collot J-Y, Villegas-Lanza J-C, Rolandone F, Vallée M, Regnier M, Segovia M, Martin X, Yepes H (2014) Distribution of discrete seismic asperities and aseismic slip along the Ecuadorian megathrust. *Earth Planet Sci Lett* 400(2014):292–301. <https://doi.org/10.1016/j.epsl.2014.05.027>
- Chuang R, Johnson K (2011) Reconciling geologic and geodetic model fault slip-rate discrepancies in Southern California: consideration of nonsteady mantle flow and lower crustal fault creep. *Geology* 39(7):627–630. <https://doi.org/10.1130/G32120.1>
- Clark JA, Farrell WE, Peltier WR (1978) Global changes in postglacial sea level: A numerical calculation. *Quat Res* 9(3):265–287. [https://doi.org/10.1016/0033-5894\(78\)90033-9](https://doi.org/10.1016/0033-5894(78)90033-9)
- Cross RS, Freymueller JT (2008) Evidence for and implications of a Bering plate based on geodetic measurements from the Aleutians and western Alaska. *J Geophys Res* 113:B07405. <https://doi.org/10.1029/2007JB005136>
- Crowell BW, Bock Y, Melgar D (2012) Real-time inversion of GPS data for finite fault modeling and rapid hazard assessment. *Geophys Res Lett* 39:L09305
- Crowell BW, Melgar D, Bock Y, Haase JS, Geng J (2013) Earthquake magnitude scaling using seismogeodetic data. *Geophys Res Lett* 40:6089–6094
- DeMets C, Gordon RG, Argus DF, Stein S (1994) Effect of recent revisions to the geomagnetic reversal time-scale on estimates of current plate motions. *Geophys Res Lett* 21:2191–2194
- DeMets C, Gordon RG, Argus DF (2010) Geologically current plate motions. *Geophys J Int* 181:1–80. <https://doi.org/10.1111/j.1365-246X.2009.04491.x>
- DeVries PMR, Krastev PG, Dolan JF, Meade BJ (2016) Viscoelastic block models of the North Anatolian fault: a unified earthquake cycle representation of pre- and postseismic geodetic observations. *Bull Seismol Soc Am* 107:403–417
- Ding K, Freymueller JT, He P, Wang Q, Xu C (2019) Glacial isostatic adjustment, intraplate strain, and relative sea level changes in the eastern United States. *J Geophys Res Solid Earth* 124:6056–6071. <https://doi.org/10.1029/2018JB017060>
- Dolan JF, Bowman DD, Sammis CG (2007) Long range and long-term fault interactions in Southern California. *Geology* 35:855–858. <https://doi.org/10.1130/G23789A.1>
- Dragert H, Wang K, James TS (2001) A silent slip even on the deeper Cascadia subduction interface. *Science* 292:1525–1528. <https://doi.org/10.1126/science.1060152>
- Dvorak JJ, Dzurlisin D (1993) Variations in magma supply rate at Kilauea volcano, Hawaii. *J Geophys Res* 98:22255–22268
- Dvorak JJ, Dzurlisin D (1997) Volcano geodesy: the search for magma reservoirs and the formation of eruptive vents. *Rev Geophys* 35:343–384
- Dvorak JJ, Mastrolorenzo G (1991) The mechanisms of recent vertical crustal movements in Campi Flegrei caldera, southern Italy. *Geological Society of America special paper*. Geological Society of America, Boulder, p 263
- Dyrgerov MB, Meier MF (2005) Glaciers and the changing earth system: a 2004 snapshot. INSTAAR occasional paper 58. Institute of Arctic and Alpine Research, University of Colorado, Boulder
- Elliott JL, Larsen CF, Freymueller JT, Motyka RJ (2010) Tectonic block motion and glacial isostatic adjustment in Southeast Alaska and adjacent Canada constrained by GPS measurements. *J Geophys Res* 115:B09407. <https://doi.org/10.1029/2009JB007139>
- Elliott J, Freymueller JT, Larsen CF (2013) Active tectonics of the St. Elias orogen, Alaska, observed with GPS measurements. *J Geophys Res Solid Earth* 118:5625–5642. <https://doi.org/10.1002/jgrb.50341>
- Fang R, Shi C, Song W, Wang G, Liu J (2014) Determination of earthquake magnitude using GPS displacement waveforms from real-time precise point positioning. *Geophys J Int* 196:461–472. <https://doi.org/10.1093/gji/ggt378>
- Farrell WE (1972) Deformation of the earth by surface loads. *Rev Geophys* 10:761–797
- Fournier T, Freymueller JT, Cervelli P (2009) Tracking magma volume recovery at Okmok volcano using GPS and an unscented Kalman filter. *J Geophys Res* 114:B02405. <https://doi.org/10.1029/2008JB005837>
- Frank WB, Rousset B, Lasserre C, Campillo M (2018) Revealing the cluster of slow transients behind a large slow slip event. *Sci Adv* 4: eaat0661
- Freed AM, Bürgmann R, Calais E, Freymueller J, Hreinsdóttir S (2006a) Implications of deformation following the 2002 Denali, Alaska earthquake for postseismic relaxation processes and lithospheric rheology. *J Geophys Res* 111:B01401. <https://doi.org/10.1029/2005JB003894>
- Freed A, Bürgmann R, Calais E, Freymueller J (2006b) Stress-dependent power-law flow in the upper mantle following the 2002 Denali, Alaska, earthquake. *Earth Planet Sci Lett* 252:481–489
- Freymueller JT, King NE, Segall P (1994) The co-seismic slip distribution of the landers earthquake. *Bull Seismol Soc Am* 84:646–659
- Freymueller JT, Woodard H, Cohen S, Cross R, Elliott J, Larsen C, Hreinsdóttir S, Zweck C (2008) Active deformation processes in Alaska, based on 15 years of GPS measurements. In: Freymueller JT, Haussler PJ, Wesson R, Ekstrom G (eds) *Active tectonics and seismic potential of Alaska*, AGU geophysical monograph, vol 179. AGU, Washington, DC, pp 1–42
- Freymueller, J. T., A. M. Freed, K. M. Johnson, R. Bürgmann, E. Calais, F. F. Pollitz, and J. Biggs, 2009. Denali fault earthquake Postseismic deformation models. *Eos transactions of the AGU*, 90, fall meet. Supplement, abstract G34A-05
- Fukuda J, Johnson KM (2008) A fully Bayesian inversion for spatial distribution of fault slip with objective smoothing. *Bull Seismol Soc Am* 98(3):1128–1146. <https://doi.org/10.1785/0120070194>
- Fukuda J, Johnson KM (2010) Mixed linear–non-linear inversion of crustal deformation data: Bayesian inference of model, weighting and regularization parameters. *Geophys J Int* 181:1441–1458. <https://doi.org/10.1111/j.1365-246X.2010.04564.x>
- Gan W, Zhang P, Shen Z-K, Niu Z, Wang M, Wan Y, Zhou D, Cheng J (2007) Present-day crustal motion within the Tibetan plateau inferred from GPS measurements. *J Geophys Res* 112:B08416. <https://doi.org/10.1029/2005JB004120>
- Gilbert GK (1884) A theory of the earthquakes of the Great Basin with a practical application. *Am J Sci* 27:49–53. Reprinted from the *Salt lake Tribune* of 20 Sept. 1883

- Gladwin MT, Gwyther RL, Hart RHG, Breckenridge KS (1994) Measurements of the strain field associated with episodic creep events on the San Andreas fault at San Juan Bautista, California. *J Geophys Res* 99(B3):4559–4565. <https://doi.org/10.1029/93JB02877>
- Graham SE, Loveless JP, Meade BJ (2018) Global plate motions and earthquake cycle effects. *Geochem Geophys Geosyst* 19:2032–2048. <https://doi.org/10.1029/2017GC007391>
- Grapenthin R, Freymueller JT, Kaufman AM (2013) Geodetic observations during the 2009 eruption of redoubt volcano, Alaska. *J Volcanol Geotherm Res* 259:115–132, electronic access at. <https://doi.org/10.1016/j.jvolgeores.2012.04.021>
- Grapenthin R, Johanson IA, Allen RM (2014) Operational real-time GPS-enhanced earthquake early warning. *J Geophys Res* 119(10):7944–7965. <https://doi.org/10.1002/2014JB011400>
- Gutenberg B (1945) Amplitudes of surface waves and magnitudes of shallow earthquakes. *Bull Seismol Soc Am* 35:3–12
- Hearn EH, McClusky S, Ergintav S, Reilinger RE (2009) Izmit earthquake postseismic deformation and dynamics of the North Anatolian Fault Zone. *J Geophys Res* 114:B08405. <https://doi.org/10.1029/2008JB006026>
- Heki K, Miyazaki S, Tsuji H (1997) Silent fault slip following an interplate thrust earthquake at the Japan Trench. *Nature* 386:595–598
- Hsu YJ, Simons M, Avouac J-P, Galetzka J, Sieh K, Chlieh M, Natawidjaja D, Prawirodirdjo L, Bock Y (2006) Frictional Afterslip following the 2005 Nias-Simeulue earthquake Sumatra. *Science* 312:1921–1926. <https://doi.org/10.1126/science.1126960>
- Hu Y, Freymueller JT (2019) Geodetic observations of time-variable glacial isostatic adjustment in Southeast Alaska and its implications for earth rheology. *J Geophys Res*. 2018JB017028. <https://doi.org/10.1029/2018JB017028>
- Hu Y, Bürgmann R, Uchida N, Banerjee P, Freymueller JT (2016) Stress-driven relaxation of heterogeneous upper mantle and time-dependent afterslip following the 2011 Tohoku earthquake. *J Geophys Res Solid Earth* 121:385–411. <https://doi.org/10.1002/2015JB012508>
- Hudnut KW, Bock Y, Cline M, Fang P, Feng Y, Freymueller J, Ge X, Gross WK, Jackson D, Kim M, King NE, Langbein J, Larsen SC, Lisowski M, Shen Z-K, Svarc J, Zhang J (1994) Co-seismic displacements of the 1992 landers earthquake sequence. *Bull Seismol Soc Am* 84:625–645
- Ide S, Beroza GC, Shelly DR, Uchide T (2007) A scaling law for slow earthquakes. *Nature* 447:76–79. <https://doi.org/10.1038/nature05780>
- Ikari M (2019) Laboratory slow slip events in natural geological materials. *Geophys J Int* 218:354–387. <https://doi.org/10.1093/gji/ggz143>
- Ji KH, Herring TA (2013) A method for detecting transient signals in GPS position time-series: smoothing and principal component analysis. *Geophys J Int* 193(1):171–186. <https://doi.org/10.1093/gji/ggt003>
- Ji C, Larson KM, Tan Y, Hudnut KW, Choi K (2004) Slip history of the 2003 San Simeon earthquake constrained by combining 1-Hz GPS, strong motion, and teleseismic data. *Geophys Res Lett* 31:L17608. <https://doi.org/10.1029/2004GL020448>
- Johansson JM et al (2002) Continuous GPS measurements of postglacial adjustment in Fennoscandia I. Geodetic results. *J Geophys Res* 107:2157. <https://doi.org/10.1029/2001JB000400>
- Johnson KM (2013) Slip rates and off-fault deformation in Southern California inferred from GPS data and models. *J Geophys Res Solid Earth* 118(10):5643–5664
- Johnson KM, Hilley G, Bürgmann R (2007) Influence of lithosphere viscosity structure on estimates of fault slip rates in the Mojave region of the San Andreas fault system. *J Geophys Res* 112: B07408. <https://doi.org/10.1029/2006JB004842>
- Johnson K, Bürgmann R, Freymueller JT (2009) Coupled afterslip and viscoelastic flow following the 2002 Denali fault, Alaska earthquake. *Geophys J Int* 176:670–682. <https://doi.org/10.1111/j.1365-246X.2008.04029.x>
- Jousset P, Mori H, Okada H (2003) Elastic models for the magma intrusion associated with the 2000 eruption of Usu volcano, Hokkaido, Japan. *J Volcanol Geotherm Res* 125:81–106. [https://doi.org/10.1016/S0377-0273\(03\)00090-8](https://doi.org/10.1016/S0377-0273(03)00090-8)
- Kanda RVS, Simons M (2010) An elastic plate model for interseismic deformation in subduction zones. *J Geophys Res* 115:B03405
- Kanda RVS, Simons M (2012) Practical implications of the geometrical sensitivity of elastic dislocation models for field geologic surveys. *Tectonophysics* 560–561:94–104
- Khan SA, Wahr J, Leuliette E, van Dam T, Larson KM, Francis O (2008) Geodetic measurements of postglacial adjustments in Greenland. *J Geophys Res* 113:B02402. <https://doi.org/10.1029/2007JB004956>
- Khan SA, Wahr J, Bevis M, Velicogna I, Kendrick E (2010) Spread of ice mass loss into Northwest Greenland observed by GRACE and GPS. *Geophys Res Lett* 37:L06501. <https://doi.org/10.1029/2010GL042460>
- Kogan MG, Vasilenko NF, Frolov DI, Freymueller JT, Steblov GM, Levin BW, Prytkov AS (2011) The mechanism of postseismic deformation triggered by the 2006–2007 great Kuril earthquakes. *Geophys Res Lett* 38:L06304. <https://doi.org/10.1029/2011GL046855>
- Kreemer C, Blewitt G, Maerten F (2006) Co- and postseismic deformation of the 28 March 2005 Nias M_w 8.7 earthquake from continuous GPS data. *Geophys Res Lett* 33:L07307. <https://doi.org/10.1029/2005GL025566>
- Kreemer C, Blewitt G, Klein EC (2014) A geodetic plate motion and global strain rate model. *Geochem Geophys Geosyst* 15:3849–3889. <https://doi.org/10.1002/2014GC005407>
- Kreemer C, Hammond WC, Blewitt G (2018) A robust estimation of the 3-D intraplate deformation of the North American plate from GPS. *J Geophys Res Solid Earth* 123:4388–4412. <https://doi.org/10.1029/2017JB015257>
- Larsen CF, Motyka RJ, Freymueller JT, Echelmeyer KA, Ivins ER (2005) Rapid viscoelastic uplift in southeast Alaska caused by post-little ice age glacial retreat. *Earth Planet Sci Lett* 237:548–560
- Larsen CF, Motyka RJ, Arendt AA, Echelmeyer KA, Geissler PE (2007) Glacier changes in Southeast Alaska and Northwest British Columbia and contribution to sea level rise. *J Geophys Res* 112: F01007. <https://doi.org/10.1029/2006JF000586>
- Larsen CF, Burgess E, Arendt AA, O'Neil S, Johnson AJ, Kienholz C (2015) Surface melt dominates Alaska glacier mass balance. *Geophys Res Lett* 42:5902–5908. <https://doi.org/10.1002/2015GL064349>
- Larson K, Freymueller J, Philipsen S (1997) Global plate velocities from the global positioning system. *J Geophys Res* 102:9961–9981
- Larson KM, Bodin P, Gombert J (2003) Using 1-Hz GPS data to measure deformations caused by the Denali fault earthquake. *Science* 300:1421–1424
- Lau HCP, Mitrovica JX, Austermann J, Crawford O, Al-Attar D, Latychev K (2016) Inferences of mantle viscosity based on ice age data sets: radial structure. *J Geophys Res Solid Earth* 121:6991–7012. <https://doi.org/10.1002/2016JB013043>
- Lemke P, Ren J, Alley RB, Allison I, Carrasco J, Flato G, Fujii Y, Kaser G, Mote P, Thomas RH, Zhang T (2007) Observations: changes in snow, ice and frozen ground. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) *Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge/New York
- Li S, Freymueller J, McCaffrey R (2016) Slow slip events and time-dependent variations in locking beneath lower Cook Inlet of the Alaska-Aleutian subduction zone. *J Geophys Res Solid Earth* 121. <https://doi.org/10.1002/2015JB012491>
- Lidberg M, Johansson JM, Scherneck H-G, Davis JL (2007) An improved and extended GPS-derived 3D velocity field of the glacial isostatic adjustment (GIA) in Fennoscandia. *J Geod* 81:213–230

- Lienkaemper JJ, Galehouse JS, Simpson RW (1997) Creep response of the Hayward fault to stress changes caused by the Loma Prieta earthquake. *Science* 276:2014–2016. <https://doi.org/10.1126/science.276.5321.2014>
- Lisowski M, Dzurisin D, Denlinger RP, Iwatsubo EY (2008) Analysis of GPS-measured deformation associated with the 2004–2006 dome-building eruption of Mount St. Helens, Washington. In: Sherrod DR, Scott WE, Stauffer PH (eds) *A volcano rekindled; the renewed eruption of Mount St. Helens, 2004–2006*, U.S. Geological Survey professional paper, vol 1750. U.S. Geological Survey, Reston, p 856. DVD-ROM [<http://pubs.usgs.gov/pp/1750/>]
- Loveless JP, Meade BJ (2010) Geodetic imaging of plate motions, slip rates, and partitioning of deformation in Japan. *J Geophys Res* 115: B02410. <https://doi.org/10.1029/2008JB006248>
- Loveless JP, Meade BJ (2011) Partitioning of localized and diffuse deformation in the Tibetan Plateau from joint inversions of geologic and geodetic observations TS AT. *Earth Planet Sci Lett* 303(1–2):11–24. <https://doi.org/10.1016/j.epsl.2010.12.014>
- Lu Z, Masterlark T, Dzurisin D (2005) Interferometric synthetic aperture radar study of Okmok volcano, Alaska, 1992–2003: Magma supply dynamics and postemplacement lava flow deformation. *J Geophys Res* 110:B02403. <https://doi.org/10.1029/2004JB003148>
- Massonnet D, Rossi M, Carmona C, Adragna F, Peltzer G, Feigl K, Rabaute T (1993) The displacement field of the landers earthquake mapped by radar interferometry. *Nature* 364:138–142
- Mattia M, Rossi M, Guglielmino F, Aloisi M, Bock Y (2004) The shallow plumbing system of Stromboli Island as imaged from 1 Hz instantaneous GPS positions. *Geophys Res Lett* 31:L24610
- Mavrommatis AP, Segall P, Uchida N, Johnson KM (2015) Long-term acceleration of aseismic slip preceding the *M*_w 9 Tohoku-oki earthquake: constraints from repeating earthquakes. *Geophys Res Lett* 42. <https://doi.org/10.1002/2015gl066069>
- McCaffrey R (2002) Crustal block rotations and plate coupling. In: Stein S, Freymueller J (eds) *Plate boundary zones*, AGU geodynamics series, vol 30. American Geophysical Union, Washington, pp 101–122
- McCaffrey R (2008) The tectonic framework of the Sumatran subduction zone. *Annu Rev Earth Planet Sci* 37:345–366
- McCaffrey R, Qamar AI, King RW, Wells R, Khazaradze G, Williams CA, Stevens CW, Vollick JJ, Zwick PC (2007) Fault locking, block rotation and crustal deformation in the Pacific Northwest. *Geophys J Int* 169:1315–1340. <https://doi.org/10.1111/j.1365-246X.2007.03371.x>
- McCaffrey R, Wallace LM, Beavan J (2008) Slow slip and frictional transition at low temperature at the Hikurangi subduction zone. *Nat Geosci* 1:316–320
- McGill SF, Wells SG, Fortner SK, Kuzma HA, McGill JD (2009) Slip rate of the western Garlock fault, at Clark Wash, near Lone Tree Canyon, Mojave Desert, California. *Geol Soc Am Bull* 121:536–554. <https://doi.org/10.1130/B26123.1>
- Meade BJ (2007) Present-day kinematics at the India-Asia collision zone. *Geology* 35:81–84. <https://doi.org/10.1130/G22924A.1>
- Meade BJ, Hager BH (2005a) Block models of crustal motion in southern California constrained by GPS measurements. *J Geophys Res* 110:B03403. <https://doi.org/10.1029/2004JB003209>
- Meade BJ, Hager BH (2005b) Spatial localization of moment deficits in southern California. *J Geophys Res* 110:B04402. <https://doi.org/10.1029/2004JB003331>
- Meade BJ, Loveless JP (2009) Predicting the geodetic signature of *M*_w > 8 slow slip events. *Geophys Res Lett* 36:L01306. <https://doi.org/10.1029/2008GL036364>
- Meade BJ, Klinger Y, Hetland E (2013) Inference of multiple earthquake-cycle relaxation timescales from irregular geodetic sampling of interseismic deformation. *Bull Seismol Soc Am* 103:2824–2835. <https://doi.org/10.1785/0120130006>
- Melbourne TI, Webb FH, Stock JM, Reigber C (2002) Rapid postseismic transients in subduction zones from continuous GPS. *J Geophys Res* 107:2241. <https://doi.org/10.1029/2001JB000555>
- Milne GA, Davis JL, Mitrovica JX, Scherneck H-G, Johansson JM, Vermeer M, Koivula H (2001) Space-Geodetic Constraints on Glacial Isostatic Adjustment in Fennoscandia. *Science* 291:2381–2385
- Milne GA, Mitrovica JX, Scherneck H-G, Davis JL, Johansson JM, Koivula H, Vermeer M (2004) Continuous GPS measurements of postglacial adjustment in Fennoscandia: 2. Modeling results. *J Geophys Res* 109:B02412. <https://doi.org/10.1029/2003JB002619>
- Minson SE, Simons M, Beck JL (2013) Bayesian inversion for finite fault earthquake source models I - theory and algorithm. *Geophys J Int* 194(3):1701–1726. <https://doi.org/10.1093/gji/ggt180>
- Minson SE, Murray JR, Langbein JO, Gombert JS (2014) Real-time inversions for finite fault slip models and rupture geometry based on high-rate GPS data. *J Geophys Res Solid Earth* 119:3201–3231. <https://doi.org/10.1002/2013JB010622>
- Mitrovica JX, Tamisiea M, Davis JL, Milne GA (2001) Recent mass balance of polar ice sheets inferred from patterns of Global Sea-level change. *Nature* 409:1026–1029
- Mitrovica JX, Gomez N, Clark PU (2009) The sea-level fingerprint of West Antarctic collapse. *Science* 323:753. <https://doi.org/10.1126/science.1166510>
- Miyagi Y, Freymueller JT, Kimata F, Sato T, Mann D (2004) Surface deformation caused by shallow magmatic activity at Okmok volcano, Alaska, detected by GPS campaigns 2000–2002. *Earth Planets Space* 56:e29–e32
- Miyazaki S, Larson KM, Choi K, Hikima K, Koketsu K, Bodin P, Haase J, Emore G, Yamagiwa A (2004a) Modeling the rupture process of the 2003 September 25 Tokachi-Oki (Hokkaido) earthquake using 1-Hz GPS data. *Geophys Res Lett* 31:L21603. <https://doi.org/10.1029/2004GL021457>
- Miyazaki S, Segall P, Fukuda J, Kato T (2004b) Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: implications for variations in fault zone frictional properties. *Geophys Res Lett* 31: L06623. <https://doi.org/10.1029/2003GL019410>
- Mogi K (1958) Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them. *Bull Earthq Res Inst, Univ Tokyo* 36:99–134
- Moreno M, Melnick D, Rosenau M, Bolte J, Klotz J, Echtler H, Baez J, Bataille K, Chen J, Bevis M, Hase H, Oncken O (2011) Heterogeneous plate locking in the South-Central Chile subduction zone: building up the next great earthquake. *Earth Planet Sci Lett* 305(3):413–424
- Neal CA, Brantley SR, Antolik L, Babb J, Burgess M, Calles K, Cappos M, Chang JC, Conway S, Desmither L, Dotray P, Elias T, Fukunaga P, Fuke S, Johanson IA, Kamibayashi K, Kauahikaua J, Lee RL, Pekalib S, Miklius A, Million W, Moniz CJ, Nadeau PA, Okubo P, Parcheta C, Patrick MP, Shiro B, Swanson DA, Tollett W, Trusdell F, Younger EF, Zoeller MH, Montgomery-Brown EK, Anderson KR, Poland MP, Ball J, Bard J, Coombs M, Dietterich HR, Kern C, Thelen WA, Cervelli PF, Orr T, Houghton BF, Gansecki C, Hazlett R, Lundgren P, Diefenbach AK, Lerner AH, Waite G, Kelly P, Clor L, Werner C, Mulliken K, Fisher G (2018) The 2018 rift eruption and summit collapse of Kilauea volcano. *Science*. <https://doi.org/10.1126/science.aav7046>
- Neri M, Casu F, Acocella V, Solaro G, Pepe S, Bernardino P, Sansosti E, Caltabiano T, Lundgren P, Lanari R (2009) Deformation and eruptions at Mt. Etna (Italy): a lesson from 15 years of observations. *Geophys Res Lett* 36:L02309. <https://doi.org/10.1029/2008GL036151>
- Nield GA, Barletta VR, Bordon A, King MA, Whitehouse PL, Clarke PJ, Domack E, Scambos TA, Berthier E (2014) Rapid bedrock uplift in the Antarctic Peninsula explained by viscoelastic response to recent ice unloading. *Earth Planet Sci Lett* 397:32–41. <https://doi.org/10.1016/j.epsl.2014.04.019>
- Nikolaïdis R, Bock Y, de Jonge P, Shearer P, Agnew D, Van Domselaar M (2001) Seismic wave observations with the global positioning system. *J Geophys Res* 106:21897–21916
- Nocquet J-M, Villegas-Lanza JC, Chlieh M, Mothes PA, Rolandone F, Jarrin P, Cisneros D, Alvarado A, Audin L, Bondoux F, Martin X,

- Font Y, Régnier M, Vallée M, Tran T, Beauval C, Mendoza JMM, Martinez W, Tavera H, Yepes H (2014) Motion of continental slivers and creeping subduction in the northern Andes. *Nat Geosci* 7:287–291. <https://doi.org/10.1038/NGEO2099>
- Norabuena E, Leffler-Griffin L, Mao AL, Dixon T, Stein S, Sacks IS, Ocola L, Ellis M (1998) Space geodetic observations of Nazca-South America convergence across the central Andes. *Science* 279:358–362
- Norabuena EO, Dixon TH, Stein S, Harrison CGA (1999) Decelerating Nazca-South America and Nazca-Pacific Plate motions. *Geophys Res Lett* 26:3405–3408
- Nyst M, Thatcher W (2004) New constraints on the active tectonic deformation of the Aegean. *J Geophys Res* 109:B11406. <https://doi.org/10.1029/2003JB002830>
- Ohta Y et al (2012) Quasi real-time fault model estimation for near-field tsunami forecasting based on RTK-GPS analysis: application to the 2011 Tohoku-Oki earthquake (M_w 9.0). *J Geophys Res* 117:B02311. <https://doi.org/10.1029/2011JB008750>
- Oskin M, Perg L, Shelef E, Strane M, Gurney E, Singer B, Zhang X (2008) Elevated shear zone loading rate during an earthquake cluster in eastern California. *Geology* 36:507–510. <https://doi.org/10.1130/G24814A.1>
- Patrick M, Orr T, Anderson K, Swanson D (2019) Eruptions in sync: improved constraints on Kilauea Volcano's hydraulic connection. *Earth Planet Sci Lett* 507:50–61. <https://doi.org/10.1016/j.epsl.2018.11.030>
- Peltier WR, Andrews JT (1976) Glacial-Isostatic Adjustment—I. The Forward Problem *Geophysical Journal International* 46(3):605–646. <https://doi.org/10.1111/j.1365-246X.1976.tb01251.x>
- Peltier WR, Argus DF, Drummond R (2015) Space geodesy constrains ice-age terminal deglaciation: the global ICE-6G_C (VM5a) model. *J Geophys Res Solid Earth* 120:450–487. <https://doi.org/10.1002/2014JB011176>
- Peltier WR, Argus DF, Drummond R (2018) Comment on “an assessment of the ICE-6G_C (VM5a) glacial isostatic adjustment model” by Purcell et al. *J Geophys Res Solid Earth* 123:2019–2018. <https://doi.org/10.1002/2016JB013844>
- Pritchard ME, Simons M, Rosen PA, Hensley S, Webb FH (2002) Co-seismic slip from the 1995 July 30 M (sub w) = 8.1 Antofagasta, Chile, earthquake as constrained by InSAR and GPS observations. *Geophys J Int* 150:362–376
- Protti M, González V, Newman AV, Dixon TH, Schwartz SY, Marshall JS, Feng L, Walter JJ, Malservisi R, Owen SE (2013) Nicoya earthquake rupture anticipated by geodetic measurement of the locked plate interface. *Nat Geosci* 7:117–121. <https://doi.org/10.1038/NGEO2038>
- Ragon T, Sladen A, Simons M (2018) Accounting for uncertain fault geometry in earthquake source inversions – I: theory and simplified application. *Geophys J Int* 214:1174–1190. <https://doi.org/10.1093/gji/ggy187>
- Reid HF (1910) Permanent displacements of the ground, in the California earthquake of April 18, 1906. In: Report of the state earthquake investigation commission. Carnegie Inst. Wash, Washington, D.C., pp 16–28
- Reid HF (1911) The elastic rebound theory of earthquakes. *Bull Dep Geol Univ Calif Publ* 6(19):413–444
- Reilinger R et al (2006) GPS constraints on continental deformation in the Africa-Arabia-Eurasia continental collision zone and implications for the dynamics of plate interactions. *J Geophys Res* 111: B05411. <https://doi.org/10.1029/2005JB004051>
- Richter A, Ivins E, Lange H, Mendoza L, Schröder L, Hormaechea JL, Casassa G, Marderwald E, Fritsche M, Perdomo R et al (2016) Crustal deformation across the Southern Patagonian Ice field observed by GNSS. *Earth Planet Sci Lett* 452:206–215
- Riel B, Simons M, Agram P, Zhan Z (2014) Detecting transient signals in geodetic time series using sparse estimation techniques. *J Geophys Res Solid Earth* 119:5140–5160. <https://doi.org/10.1002/2014JB011077>
- Robbins JW, Smith DE, Ma C (1993) Global scale tectonic plate motions measured with CDP VLBI data. In: Smith DE, Turcotte DL (eds) Contributions of space geodesy to geodynamics: crustal dynamics, AGU geophysical monograph, vol 23. American Geophysical Union, Washington, pp 21–36
- Rogers G, Dragert H (2003) Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science* 300:1942–1943. <https://doi.org/10.1126/science.1084783>
- Roy K, Peltier WR (2017) Space-geodetic and water level gauge constraints on continental uplift and tilting over North America: regional convergence of the ICE-6G_C (VM5a/VM6) models. *Geophys J Int* 210:1115–1142
- Ruhl C, Melgar D, Grapenthin R, Allen RM (2017) The value of real-time GNSS to earthquake early warning. *Geophys Res Lett* 44(16):8311–8319. <https://doi.org/10.1002/2017GL074502>
- Sasgen I, Martín-Español A, Horvath A, Klemann V, Petrie EJ, Wouters B, Horwath M, Pail R, Bamber JL, Clarke PJ, Konrad H, Wilson T, Drinkwater MR (2018) Altimetry, gravimetry, GPS and viscoelastic modeling data for the joint inversion for glacial isostatic adjustment in Antarctica (ESA STSE project REGINA). *Earth Syst Sci Data* 10:493–523. <https://doi.org/10.5194/essd-10-493-2018>
- Savage J (1983) A dislocation model of strain accumulation and release at a subduction zone. *J Geophys Res* 88:4984–4996
- Savage JC, Burford RO (1970) Accumulation of tectonic strain in California. *Bull Seismol Soc Am* 60:1877–1896
- Savage JC, Langbein J (2008) Postearthquake relaxation after the 2004 M_6 Parkfield, California, earthquake and rate-and-state friction. *J Geophys Res* 113:B10407. <https://doi.org/10.1029/2008JB005723>
- Savage J, Prescott W (1978) Asthenosphere readjustment and the earthquake cycle. *J Geophys Res* 83:3369–3376
- Scholz C (1990) The mechanics of earthquakes and faulting. Cambridge University Press, Cambridge
- Schwartz SY, Rokosky JM (2007) Slow slip events and seismic tremor at circum-pacific subduction zones. *Rev Geophys* 45:RG3004. <https://doi.org/10.1029/2006RG000208>
- Sella GF, Stein S, Dixon TH, Craymer M, James TS, Mazzotti S, Dokka RK (2007) Observation of glacial isostatic adjustment in “stable” North America with GPS. *Geophys Res Lett* 34:L02306. <https://doi.org/10.1029/2006GL027081>
- Sieh K, Jones L, Hauksson E, Hudnut K, Eberhart-Phillips D, Heaton T, Hough S, Hutton K, Kanamori H, Lilje A, Lindvall S, McGill S, Mori J, Rubin C, Spotila JA, Stock J, Thio H, Treiman J, Wernicke B, Zachariasen J (1993) Near-field investigations of the landers earthquake sequence, April to July, 1992. *Science* 260:171–176
- Smith SW, Wyss M (1968) Displacement on the San Andreas fault subsequent to the 1966 Parkfield earthquake. *Bull Seismol Soc Am* 58:1955–1973
- Steblov GM, Kogan MG, Levin BV, Vasilenko NF, Prytkov AS, Frolov DI (2008) Spatially linked asperities of the 2006–2007 great Kuril earthquakes revealed by GPS. *Geophys Res Lett* 35:L22306. <https://doi.org/10.1029/2008GL035572>
- Suito H, Freymueller JT (2009) A viscoelastic and afterslip postseismic deformation model for the 1964 Alaska earthquake. *J Geophys Res* 114:B11404. <https://doi.org/10.1029/2008JB005954>
- Sun T, Wang K, Iinuma T, Hino R, He J, Fujimoto H, Kido M, Osada Y, Miura S, Ohta Y, Hu Y (2014) Prevalence of viscoelastic relaxation after the 2011 Tohoku-oki earthquake. *Nature* 514(7520):84–87. <https://doi.org/10.1038/nature13778>
- Taira T, Bürgmann R, Nadeau RM, Dreger DS (2014) Variability of fault slip behavior along the San Andreas fault in the San Juan Bautista region. *J Geophys Res Solid Earth* 119:8827–8844. <https://doi.org/10.1002/2014JB011427>
- Tamisiea ME, Mitrovica JX, Milne GA, Davis JL (2001) Global geoid and sea level changes due to present-day ice mass fluctuations. *J Geophys Res* 106:30849–30863
- Tamisiea ME, Leuliette EW, Davis JL, Mitrovica JX (2005) Constraining hydrological and cryospheric mass flux in southeastern Alaska using

- space-based gravity measurements. *Geophys Res Lett* 32:L20501. <https://doi.org/10.1029/2005GL023961>
- Taylor M, Yin A (2009) Active structures of the Himalayan-Tibetan orogen and their relationships to earthquake distribution, contemporary strain field, and Cenozoic volcanism. *Geosphere* 5:199–214. <https://doi.org/10.1130/GES00217.1>
- Thatcher W (2003) GPS constraints on the kinematics of continental deformation. *Int Geol Rev* 45:191–212
- Thatcher W (2009) How the continents deform: the evidence from tectonic geodesy. *Annu Rev Earth Planet Sci* 17:237–262. <https://doi.org/10.1146/annurev.earth.031208.100035>
- Tong X, Smith-Konter B, Sandwell DT (2014) Is there a discrepancy between geological and geodetic slip rates along the San Andreas fault system? *J Geophys Res Solid Earth* 119:2518–2538. <https://doi.org/10.1002/2013JB010765>
- Tregoning P, Burgette R, McClusky SC, Lejeune S, Watson CS, McQueen H (2013) A decade of horizontal deformation from great earthquakes. *J Geophys Res Solid Earth* 118:2371–2381. <https://doi.org/10.1002/jgrb.50154>
- Tsang LLH, Meltzner AJ, Hill EM, Freymueller JT, Sieh K (2015) A paleogeodetic record of variable interseismic rates and megathrust coupling at Simeulue Island, Sumatra. *Geophys Res Lett* 42:10,585–10,594. <https://doi.org/10.1002/2015GL066366>
- van der Wal W, Whitehouse PL, Schrama EJO (2015) Effect of GIA models with 3D mantle viscosity on GRACE mass balance estimates for Antarctica. *Earth Planet Sci Lett* 414(0):134–143. <https://doi.org/10.1016/j.epsl.2015.01.001>
- Wallace LM, Beavan J (2010) Diverse slow slip behavior at the Hikurangi subduction margin, New Zealand. *J Geophys Res* 115: B12402. <https://doi.org/10.1029/2010JB007717>
- Wallace LM, Beavan RJ, McCaffrey R, Darby DJ (2004) Subduction zone coupling and tectonic block rotations in the North Island, New Zealand. *J Geophys Res Solid Earth* 109(B12):B12406. <https://doi.org/10.1029/2004JB003241>
- Wallace LM, Beavan RJ, McCaffrey R, Berryman KR, Denys P (2007) Balancing the plate motion budget in the South Island, New Zealand using GPS, geological and seismological data. *Geophys J Int* 168(1):332–352. <https://doi.org/10.1111/j.1365-246X.2006.03183.x>
- Wallace LM, Webb SC, Ito Y, Mochizuki K, Hino R, Henrys S, Schwartz SY, Sheehan AF (2016) Slow slip near the trench at the Hikurangi subduction zone, New Zealand. *Science* 352(6286):701–704. <https://doi.org/10.1126/science.aaf2349>
- Wallace LM, Kaneko Y, Hreinsdóttir S, Hamling I, Peng Z, Bartlow N et al (2017) Large-scale dynamic triggering of shallow slow slip enhanced by overlying sedimentary wedge. *Nat Geosci* 10(10):765–770. <https://doi.org/10.1038/ngeo3021>
- Wallace LM, Hreinsdóttir S, Ellis S, Hamling I, D’Anastasio E, Denys P (2018) Triggered slow slip and afterslip on the southern Hikurangi subduction zone following the Kaikōura earthquake. *Geophys Res Lett* 45:4710–4718. <https://doi.org/10.1002/2018GL077385>
- Wang K (2007) Elastic and viscoelastic models of crustal deformation in subduction earthquake cycles. In: Dixon TH, Moore JC (eds) *The Seismogenic zone of subduction thrust faults*. Columbia University Press, New York, pp 540–575
- Wang K, Fialko Y (2018) Observations and modeling of coseismic and postseismic deformation due to the 2015 M_w 7.8 Gorkha (Nepal) earthquake. *J Geophys Res Solid Earth* 123:761–779. <https://doi.org/10.1002/2017JB014620>
- Wang K, Hu Y, He J (2012) Deformation cycles of subduction earthquakes in a viscoelastic earth. *Nature* 484:327–332. <https://doi.org/10.1038/nature11032>
- Wang K, Sun T, Brown L, Hino R, Tomita F, Kido M, Iinuma T, Kodaira S, Fujiwara T (2018) Learning from crustal deformation associated with the M9 2011 Tohoku-oki earthquake. *Geosphere* 14:552–571. <https://doi.org/10.1130/GES01531.1>
- Wöppelmann G, Marcos M (2016) Vertical land motion as a key to understanding sea level change and variability. *Rev Geophys* 54:64–92. <https://doi.org/10.1002/2015RG000502>
- Wöppelmann G, Letetrel C, Santamaria A, Bouin M-N, Collilieux X, Altamimi Z, Williams SDP, Míguez BM (2009) Rates of sea-level change over the past century in a geocentric reference frame. *Geophys Res Lett* 36:L12607. <https://doi.org/10.1029/2009GL038720>
- Yohler R, Bartlow N, Wallace LM, Williams C (2019) Time-dependent behavior of a near-trench slow-slip event at the Hikurangi subduction zone. *Geochem Geophys Geosyst* 20:4292–4304. <https://doi.org/10.1029/2019GC008229>
- Zeng Y, Shen Z-K (2014) Fault network modeling of crustal deformation in California constrained using GPS and geologic observations. *Tectonophysics* 612–613:1–17. <https://doi.org/10.1016/j.tecto.2013.11.030>
- Zhao B, Bürgmann R, Wang D, Tan K, Du R, Zhang R (2017) Dominant controls of downdip afterslip and viscous relaxation on the post-seismic displacements following the Mw7.9 Gorkha, Nepal, earthquake. *J Geophys Res Solid Earth* 122:8376–8401. <https://doi.org/10.1002/2017JB014366>

Gravimeters

Andrew Hugill
Scintrex Ltd, Concord, ON, Canada

Synonyms

Gravity meters; Relative

Definition

An instrument which measures differences in gravity. Units: Practical units are mGal (10^{-6} g) and μ Gal (10^{-9} g). 1 Gal = 1 cm/s^2 .

Introduction

The first relative gravity measurements were made in 1672 by the French astronomer, Richer, who noted a difference in the period of a pendulum when observed in Paris and close to the equator in Cayenne, South America. For the next two centuries, relative gravity measurements, mainly for scientific purposes, were made with pendulums. In 1893, Threlfall and Pollock made the first successful spring-mass gravimeter in Sydney, Australia (Threlfall and Pollock 1900) using the newly discovered quartz fibers. From the middle of the twentieth century up until the present nearly all relative gravity measurements have been made with instruments based on a mass suspended by a quartz or metal elastic system. Instruments in which the mass is levitated in the field of a

superconducting magnet and instruments based on inertial-grade accelerometers are also used in some applications.

Because they are not intrinsically calibrated, relative gravimetry rely on measurements made by absolute gravity meters (see the ► “Gravity Measurements, Absolute” entry of this encyclopedia) for calibration.

Relative gravimeters can be classified into three types according to the mode in which they are operated, with each mode of operation having different performance requirements:

1. Survey: Measurement of relative gravity as a function of position on, or under, the surface of the earth or on the sea floor
2. Stationary: Continuous measurement of changes in gravity or low-frequency earth motion as a function of time at a fixed location on, or under, the surface of the earth or on the sea floor
3. Dynamic: measurement of relative gravity as a function of position from a moving platform such as an aircraft or a ship

Performance Requirements

Basic performance requirements for relative gravimeters are listed in Table 1. These must be met under a range of environmental conditions according to the mode of operation. Dynamic instruments have to be able to produce data while moving and subject to relatively large accelerations, whereas the defining characteristic of survey instruments is the capability to withstand acceleration and shocks during transport without sustaining changes in the reading value.

Survey instruments can be subdivided into borehole, sea-floor, and land gravimeters all with the same basic performance requirements. However, borehole and sea-floor instruments must have remote reading capabilities and be packaged to withstand high pressure. In addition, borehole instruments have size restrictions and are required to operate at higher temperatures. Survey gravimeters have a useable sensitivity of less than 5 μGal (e.g., Zumberge et al. 2008; Merlet et al. 2008). A total range of 7000 mGal covers the gravity field on

the surface of the earth. A minimum continuous operating range of around 50 mGal is needed to provide useful measurement capability.

Stationary gravimeters can be divided into those which are located in an observatory and smaller, less cumbersome portable or transportable instruments (Crossley et al. 2013). Most applications for observatory instruments (Crossley et al. 1999) are in the frequency range from dc up to approximately 7 mHz which is the maximum frequency at which the seismic normal modes are resolvable (Ingate and Berger 2004). Minimum earth noise in this band is around $3 \times 10^{-2} \mu\text{Gal}/\sqrt{\text{Hz}}$ (Ingate and Berger 2004), which represents the minimum useful instrument sensitivity. Superconducting instruments and the metal spring instruments which approach this sensitivity (Banka and Crossley 1999) have a range of a few mGal or less and are saturated by surface waves of large earthquakes (e.g., Mansinha and Kamal 1992; Agnew et al. 1986). More recently, interest has turned from using stationary gravimeters for global studies to applications related to hazard monitoring and climate change (Crossley et al. 2013), including volcano, earthquake, and aquifer monitoring. Although, in some cases requiring the same levels of performance as observatory gravimeters, the instruments used in these studies need to be portable so that they can be moved to locations where the largest signals are found.

Useable sensitivity for airborne gravity meters is limited by GPS precision as described in the *Sensor motion and orientation* section below. With a 100 s low pass filter, this limit is around 0.5 mGal (Argyle et al. 2000).

Technology

Sensor Technology

Relative gravimeters are based on one of the four core technologies listed in the top row of Table 2. Technologies have characteristics which better suit them to particular types of gravity meter as discussed below.

Fused Quartz

Fused quartz sensors perform well in survey gravimeters because quartz has high strength, remains almost perfectly elastic up to its breaking stress, and can be welded into compact structures which have exceptional mechanical integrity. Sensors are free of tares when transported and do not have to be clamped between readings, making them accurate and quick and easy to operate. The relatively high long-term drift of quartz springs is linear enough to be compensated electronically for survey applications.

Metal

Metal spring-mass sensors have very low long-term drift and low temperature coefficient, making them well suited to

Gravimeters, Table 1 Gravimeter performance requirements

Type of gravimeter	Performance			
			Range (mGal)	
	Signal bandwidth	Sensitivity (μGal)	Total	Continuous (min)
Survey	Dc	1–10	7000	50
Stationary: Observatory	Dc – 10 mHz	$3 \times 10^{-2}/\sqrt{\text{Hz}}$	7000	2
Stationary: Portable	Dc – 10 mHz	1	7000	2
Dynamic	Dc – 10 mHz	500	10,000	10,000

Gravimeters, Table 2 Gravity sensor core technology

Technology Type of gravimeter	Spring/mass		Superconducting	Inertial-grade accelerometer
	Quartz	Metal		
Survey				
Stationary				
Dynamic				

stationary measurements. However, they are more susceptible to tares in response to transportation because the elastic properties of the materials in their elastic suspension are inferior to those of quartz. Also they are assembled from a number of components fixed together, so they do not have the integrity of a welded quartz structure. These inherent disadvantages have been overcome by clamping the mass during transportation and careful handling, but at the expense of added manufacturing complexity, requirement for more skillful operators and lower productivity in the field. Historically, survey gravimeters based on metal spring-mass sensors were used extensively in survey operations because of their low long-term drift and high accuracy, especially for establishing gravity networks and other applications which require measurement of gravity at widely separated points. Absolute gravimeters are now used to make large-scale network-type measurements and survey instruments with metal spring-mass sensors are being replaced for most other applications by modern quartz instruments which are easier to manufacture and operate, and make extensive use of electronics to obtain similar or better performance.

Metal spring-mass gravimeters were first used for dynamic applications in the late 1950s (Dehlinger 1978). Since then, they have been continually refined and modified for this application and are in widespread use (e.g., Williams and MacQueen 2001; NOAA Grav-D project 2017).

Superconducting

The superconducting gravity sensor is described by one of its inventors (Goodkind 1999) as “a spring type gravimeter in which the mechanical spring is replaced by a magnetic levitation of a superconducting sphere in the field of superconducting, persistent current coils.” Because these currents are very stable, the sensor has extremely low drift which makes it well suited for stationary operation.

Inertial-Grade Accelerometers

These are devices that have been developed for inertial navigation systems so have excellent dynamic performance on moving platforms and are very compact. They have sufficient sensitivity for dynamic operations, but not for survey or

stationary operations. An example of an inertial-grade accelerometer is the Q Flex QA3000 which has a volume of approximately 1 cu. inch and is built around a chemically machined quartz suspension (Foote and Grindeland 1992). Another example is the Bell Aerospace accelerometer (Torge 1989) which has similar dimensions.

Design Principles

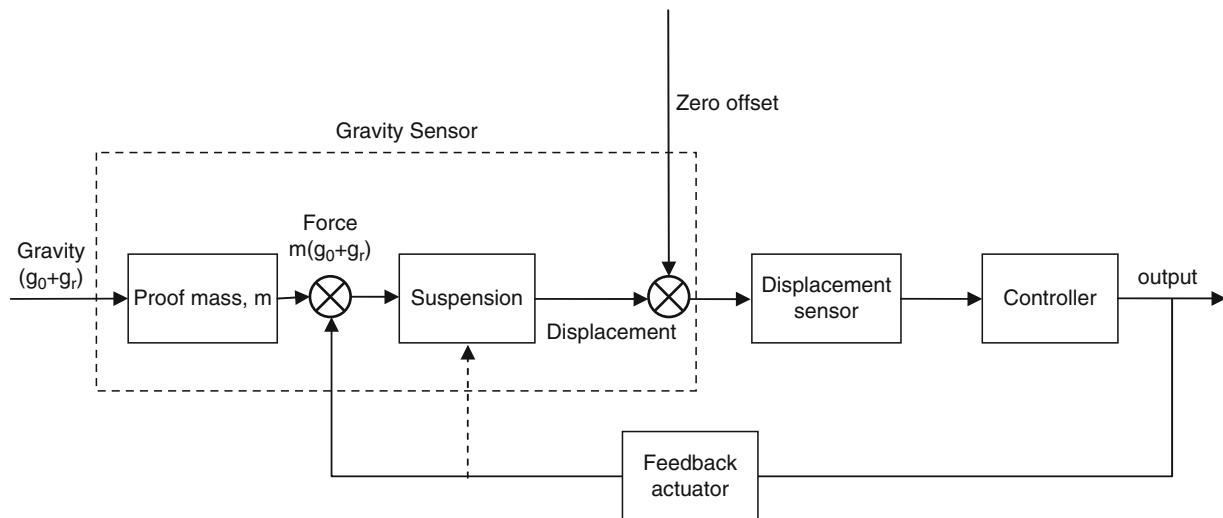
A generalized signal flow diagram for relative gravimeters is shown in Fig. 1. The proof mass is suspended so that it moves in a well-defined trajectory in response to a change in the gravitational force on it. The proof mass and suspension are termed the gravity sensor. The gravity reading is derived from the input to the calibrated feedback actuator that returns the output of the displacement sensor to the reference value. Feedback force is applied to the proof mass through an electric field or electromagnetic actuator and/or through a change to the suspension geometry.

Gravity Sensor Natural Period, Sensitivity, and Seismic Noise

The design principles of the gravity sensor determine important performance characteristics of the gravimeter that are largely independent of the technology used in construction.

The gravitational force on the mass of a spring-mass gravity sensor is balanced by the elastic force generated by the extension of a quartz or metal spring. The extended length of the spring is given by $l = \frac{mg}{k} + l_u$ where k is the spring constant and l_u is the unstressed length of the spring. A change in gravity from a value g_0 to $(g_0 + \Delta g)$ produces a change in spring length from l_0 to $(l_0 + \Delta l)$ where $\Delta l / (l_0 - l_u) = \Delta g / g_0$. As Δl is proportional to $(l_0 - l_u)$, this simple spring-mass system is termed a proportional sensor, or to use the language of seismometry, a short-period sensor.

For a practical sensor with $l_0 - l_u = 2$ cm and a change in gravity of 0.01 mGal ($1 \text{ Gal} = 1 \text{ cm/s}^2$), $\Delta g / g_0 = 10^{-8}$ and the displacement which must be detected is $\Delta l = 2 \times 10^{-10}$ m. This system responds equally to gravity and acceleration up to its natural frequency $f_n = \frac{1}{2\pi} \sqrt{\frac{k}{m}} = \frac{1}{2\pi} \sqrt{\frac{g_0}{(l_0 - l_u)}} = 3.5 \text{ Hz}$, so



Gravimeters, Fig. 1 Relative gravimeter functional diagram

unwanted signals, including microseismic and cultural noise, would need to be filtered from the output to recover the gravity signal.

It was not until the 1970s and 1980s that electronics were available to detect sub-nm displacements and effectively filter signals in a portable instrument. The Scintrex CG-3 gravity meter, introduced in 1987, and its successors the CG-5 and CG-6 are based on a proportional spring-mass system and use capacitive displacement sensing and digital filtering to achieve the required level of sensitivity and noise reduction (Hugill 1990).

Before the advent of modern electronics, many mechanical design concepts were developed (e.g., Hugill 1984) to increase sensitivity and reduce natural frequency, and thus noise response, of sensors enabling detection of gravity-induced mass displacement by eye, through a microscope. Sensors with increased mechanical sensitivity are termed “astatic” and are analogous to long-period seismometers. The most successful of these was the LaCoste suspension based on a “zero-length spring” with an unstressed length l $u = 0$ (LaCoste 1934). Worden-type and LaCoste and Romberg gravimeters, which were introduced in the 1940s and 1950s, are based on the LaCoste suspension.

The sensor in the superconducting gravimeter (Goodkind 1999) is also astaticized with its sensitivity being increased by adjusting the relative current settings in two collinear coils to shape the magnetic field gradient in the vicinity of the levitated proof mass.

Displacement Sensing and Feedback

Most modern gravimeters use capacitive displacement sensing and either electromagnetic or electrostatic feedback. Electrostatic feedback allows sensor design to be simplified as the same electrodes can be employed for sensing, using high frequency excitation, and feedback at lower frequencies

(e.g., Block and Moore 1966; Hugill 1984). Examples of instruments using capacitive displacement sensing and electrostatic feedback include the Scintrex CG-3, CG-5 and CG-6 (Hugill 1990) and the micro-g LaCoste gPhone (Micro-g LaCoste 2019).

Electromagnetic feedback has the advantage of inherent linearity and a higher force capability. Instruments with capacitive displacement sensing and electromagnetic feedback include the Bodenseewerke GSS30 (Torge 1989), the superconducting gravimeter (Goodkind 1999) and inertial-grade accelerometers. In the latter, feedback supports the full gravitational force on the mass so the spring suspension only provides constraint of the proof mass motion (e.g., Foote and Grindeland 1992), which is equivalent to having no zero offset in Fig. 1.

In most gravimeters with the LaCoste suspension, the top of the mainspring is moved to null the mass. This is done in Worden-type meters through a gearbox, screw thread, and auxiliary springs. In the LaCoste and Romberg g meter (LaCoste and Romberg 1945), the top of the spring is connected to a lever system which is moved by the operator through a calibrated gearbox and screw thread. The required displacement of the top of the spring to null a change in gravity of 0.01 mGal is 0.25 nm (Torge 1989). Electrostatic feedback systems are installed in gravimeters with metal LaCoste suspensions as an option for survey instruments to make reading easier and more accurate (e.g., Webring et al. 2004; ZLS 2019) and to enable high-sensitivity automatic reading for stationary instruments (e.g., Block and Moore 1966).

Error Sources

Error sources can be divided into three broad groups: those generated by the gravimeter or intrinsic error sources, those

resulting from the response of the instrument to external inputs, and those resulting from the effects of transportation.

Intrinsic Error Sources

Scale Factor Changes

Changes in the scale factor of gravimeters with quartz sensors can be as high as 10^{-3} in the first year after manufacture; subsequent changes are of the order of 10^{-4} – 10^{-5} per year (Zumbege et al. 2008; Meurers 2018; Carbone and Rymer 1999). Changes of around 10^{-3} have also been reported in metal spring sensors (Carbone and Rymer 1999). In contrast to the smooth drift in the scale factor of quartz sensors, the reported changes in the metal spring sensor were abrupt and separated by long periods of stability.

As the scale factor, k , relates a reading differences ($R_2 - R_1$) to the gravity change ($g_2 - g_1$) through the expression ($g_2 - g_1 = k(R_2 - R_1)$), the impact of errors depends on the gravity difference being measured and the required accuracy. Good field practice, including routine calibration and use of multiple instruments for critical measurements, prevents degradation of survey results from this error source.

Instrument Noise

Thermal noise determines the detection limit of the sensor and is given by (e.g., Richter et al. 1995)

$$a_t^2 = \frac{25k_b T}{QmT_n} \left(\frac{(\text{ms}^{-2})^2}{\text{Hz}} \right)$$

where a_t is the rms noise equivalent acceleration spectral density, k_b is Boltzman's constant, T is absolute temperature, Q is the quality factor, m is the mass of the proof mass, and T_n is the natural period. For a metal spring LaCoste and Romberg earth tide gravimeter operating at 323 K with $Q = 1$, $m = 80$ g, and $T_n = 20$ s, the thermal noise is $a_t^2 = 7 \times 10^{-20} \text{ m}^2 \text{ s}^{-4} / \text{Hz}$ (Richter et al. 1995) and for a superconducting gravity sensor operating at 4 K with $Q = 0.14$, $m = 4$ g, and $T_n = 5$ s, the thermal noise is $a_t^2 = 5 \times 10^{-19} \text{ m}^2 \text{ s}^{-4} / \text{Hz}$ and could be a factor of 3 higher, depending on the extent to which the horizontal modes couple into the vertical (Van Camp 1999).

The thermal noise limit is therefore a significant design parameter for stationary gravimeters measuring in the frequency band dc–7 mHz where earth noise power spectral density can be lower than $10^{-19} \text{ m}^2 \text{ s}^{-4} / \text{Hz}$ (Ingate and Berger 2004).

Long-Term Drift

Long-term drift is a characteristic of all relative gravimeters and imposes constraints on the way in which instruments are

used. Values range from a few $\mu\text{Gal}/\text{year}$ for instruments with superconducting sensors (e.g., Crossley et al. 2003), $5 \mu\text{Gal}/\text{day}$ or more for metal spring sensors (e.g., Crossley et al. 2013) and up to several hundred $\mu\text{Gal}/\text{day}$ or more for survey gravimeters with quartz sensors (e.g., Bonvalot et al. 1998; Merlet et al. 2008). Drift is modeled as a function of time and removed from measurements as part of data processing. Model parameters for survey instruments are derived from repeat readings on base stations at different times during a survey. Absolute gravity readings obtained at the same site are used as a reference to verify the drift model of superconducting gravimeters (e.g., Van Camp and Francis 2007).

External Noise Sources

The effect of external inputs on the sensor can be reduced to acceptable levels by a number of approaches which are described below. These include shielding the sensitive elements, compensation, and filtering and reducing the input at its source through setup and choice of location and time.

Sensor Motion and Orientation

Motion of the gravity sensor can produce a response which is only distinguishable from the gravity signal by its frequency content. If the disturbing motion is outside of the signal bandwidth, then its effects can be removed by filtering, otherwise independent means must be used to determine the sensor motion and compensate for its effects. High frequency motion and vibration can excite higher order sensor modes which can couple into the fundamental mode (e.g., Block and Moore 1970).

The gravity sensor measures the components of gravity and acceleration that are parallel to its sensitive axis. To accurately measure gravity, the sensor is vertically aligned. Vertical misalignment of the sensor results in a component of gravity being measured which varies as the cosine of the angle. Sensor tilt can be actively controlled using a levelling platform or similar mechanism, or measured and then corrected in post processing (Van Camp et al. 2017).

Horizontal acceleration can produce “cross coupling” in sensors that have the proof mass in the form of a horizontal pendulum and feedback control which doesn't keep the mass at the zero position with sufficient accuracy (e.g., Torge 1989). Horizontal velocity causes centrifugal and Coriolis accelerations which combine to produce the Eötvös-effect (Torge 1989).

The relative magnitude of each of these error sources depends on the mode in which the instrument is operated:

Airborne gravity precision and spatial resolution are ultimately limited by the recovery of aircraft vertical position with GPS (e.g., Bell et al. 1999). This limit has been estimated to be around 1 mGal rms for an 80 s filter (van Kann 2004), an

estimate which is confirmed by actual measurements of aircraft acceleration with GPS (Wei and Schwarz 1995). Performance at this level has been reported for at least two airborne gravity systems (Studinger et al. 2008). One of these, the Sander Geophysics system, which is based on inertial-grade accelerometer sensors, is unaffected by turbulence up to 0.6 m/s^2 rms (Argyle et al. 2000).

In marine gravity, the ship moves relatively slowly on an approximately equipotential surface (Torge 1989) so there is no significant vertical acceleration in the signal bandwidth and the relatively high frequency acceleration caused by wave motion is removed by low pass filtering. Horizontal position and velocity, derived from GPS measurements, are used to compensate Eötvös-effects with sufficient accuracy to enable gravity anomalies of 0.2 mGal to be detected with a 0.5 km spatial resolution (LaFehr et al. 1992).

Survey gravimeters are affected by ground motion in the form of microseismic noise, cultural and wind noise, and earthquakes. Microseismic noise peaks in the range between 3 and 8 s (e.g., Ingate and Berger 2004) and has amplitudes as high as 1 mGal pp . (Laswell et al. 2010). Distant earthquakes produce signals with peak amplitude at a period of around 20 s (Ingate and Berger 2004). Cultural and wind noise is generally above 1 Hz (e.g., Wilson 1953) and depends on local conditions. The effect of these noise sources can be reduced with low pass filtering. Astatized (long-period) sensors have naturally lower sensitivity to this noise; however, proportional (short-period) sensors with digital filtering can produce readings with an rms noise level of a few μGal or less in most conditions (Scintrex 2019; Crossley et al. 2013).

Stationary gravimeters avoid cultural and wind noise through the choice of a quiet reading site (e.g., Torge 1989). Earthquakes often generate the signal that is being measured. Filtering is used to suppress signals outside the band of interest (e.g., Agnew et al. 1976).

Temperature, Pressure, and Magnetic Field

Sensitivity of spring-mass sensors to these inputs is largely determined by the material that the gravity sensor is constructed from. For example, the elastic suspension of metal spring instruments is made from a material with an elastic modulus that has low temperature sensitivity, but relatively high magnetic field sensitivity. Fused quartz is not sensitive to changes in magnetic field, but its elastic modulus has high temperature sensitivity.

The amplitudes of temperature, pressure, and magnetic field changes that reach the gravity sensor are reduced by control, sealing, and shielding, respectively. In addition, stationary gravimeters are often placed in environments where the temperature is extremely stable (e.g., Van Camp and Francis 2007). If residual changes reaching the sensor are significant, they are compensated mechanically or

electronically. In the Scintrex CG-3, CG-5, and CG-6 quartz instruments, the temperature is controlled to the m°C level and further compensated to the $\mu^\circ\text{C}$ level based on a temperature measurement close to the elastic suspension (Hugill 1990). In the metal spring LaCoste and Romberg instruments, residual pressure changes are mechanically compensated (Ander et al. 1999).

Superconducting penetration depth determines the temperature sensitivity of superconducting gravity sensors (Goodkind 1999), and the levitation force is directly proportional to magnetic field changes. The amplitude of temperature and magnetic field changes at the sensor are reduced to acceptable levels by controlling temperature to a few $\mu^\circ\text{C}$ and surrounding the instrument with a double μ metal/Niobium magnetic shield.

Transportation Effects

Changes in readings, often called tares; changes in drift rate; and elastic relaxation can all contribute to errors as a result of transportation. However, these errors can generally be avoided through good operating procedures. Careful handling and appropriate protection from shock and vibration (e.g., Hamilton and Brule 1967) can reduce or eliminate tares. Variable drift can be recovered by frequent base station reoccupation (e.g., Webring et al. 2004) and elastic relaxation is minimized in quartz instruments by maintaining vertical orientation during transport (e.g., Zumberge et al. 2008) and by waiting a fixed time after unclamping in metal instruments (e.g., Torge 1989).

Gravimeters

Some gravimeters in use today are listed in Table 3.

Summary

Relative gravimeters are based on one of four core technologies: spring-mass systems made from quartz or metal, superconducting sensors, or inertial-grade accelerometers. Performance requirements depend on whether instruments are to be used for surveys or in stationary or dynamic modes of operation. Useable sensitivity varies from $3 \times 10^{-2} \mu\text{Gal}/\sqrt{\text{Hz}}$ for stationary instruments operated in observatories to 0.5 mGal for dynamic gravimeters operated in aircraft. The defining characteristic of survey instruments is the capability to withstand acceleration and shocks during transport without sustaining changes in the reading value. Dynamic instruments must be able to produce data while moving and subject to large accelerations.

Gravimeters, Table 3 Some relative gravimeters currently in use

Gravimeter	Type	Sensor technology	Period	Displacement sensor	Feedback actuator	Reference
LaCoste and Romberg g meter	Survey	Metal spring-mass	Long	Optical, capacitive	Mechanical (electrostatic)	Webring et al. (2004)
Scintrex CG-3/5/6	Survey	Quartz spring-mass	Short	Capacitive	Electrostatic	Hugill (1990), Scintrex (2019)
Bodenseewerke GSS30	Dynamic	Metal spring-mass	Short	Capacitive	Electromagnetic	Torge (1989)
GWR	Stationary	Superconducting	Long	Capacitive	Electromagnetic	Goodkind (1999)
ZLS Burris gravimeter	Survey/stationary	Metal spring-mass	Long	Capacitive	Electrostatic (mechanical)	ZLS (2019)
Micro-g LaCoste gPhoneX	Stationary	Metal spring-mass	Long	Capacitive	Electrostatic	Micro-g LaCoste (2019)
Sander AIRGrav	Dynamic	Inertial-grade accelerometer	Short	Capacitive	Electromagnetic	Argyle et al. (2000)
Canadian micro gravity GT gravimeters	Dynamic	Inertial-grade accelerometer	Short	Optical	Electromagnetic	Canadian Micro Gravity (2019)
ZLS dynamic gravity meter	Dynamic	Metal spring-mass	Long	Capacitive	Mechanical	ZLS (2019)
Micro-g LaCoste Air-Sea II	Dynamic	Metal spring-mass	Long	Capacitive	Mechanical	Micro-g LaCoste (2019)

Cross-References

- Earth Tides
- Geodesy, Networks, and Reference Systems
- Gravity Anomalies, Interpretation
- Gravity Data, Advanced Processing
- Gravity Data, Regional: Residual Separation
- Gravity Field of the Earth
- Gravity Field, Temporal Variations from Space Techniques
- Gravity Measurements, Absolute
- Gravity Method, Principles
- Gravity Method, Satellite
- Gravity Method, Surface
- Gravity, Data to Anomalies
- Gravity, Global Models
- Gravity, Gradiometry

Bibliography

- Agnew D, Berger J, Buland R, Farrell W, Gilbert JF (1976) International deployment of accelerometers: a network of very long period seismology. *EOS Trans Am Geophys Union* 57:181–188
- Agnew D, Berger J, Buland R, Farrell W, Gilbert JF (1986) Project IDA: a decade in review. *EOS Trans Am Geophys Union* 67(10):203–212
- Ander ME, Summers T, Gruchalla ME (1999) LaCoste and Romberg gravity meter: system analysis and instrumental errors. *Geophysics* 64:1708–1719
- Argyle M, Ferguson S, Sander L, Sander S (2000) AIRGrav results: a comparison of airborne gravity data with GSC test site data. *Lead Edge* 19:1134–1138
- Banka D, Crossley D (1999) Noise levels of superconducting gravimeters at seismic frequencies. *Geophys J Int* 139:87–97
- Bell RE, Childers VA, Arko RA (1999) Airborne gravity and precise positioning for geologic applications. *J Geophys Res* 104(B7):15281–15292
- Block B, Moore RD (1966) Measurements in the earth mode frequency range by an electrostatic sensing and feedback gravimeter. *J Geophys Res* 71:4361–4375
- Block B, Moore RD (1970) Tidal to seismic frequency investigations with a quartz accelerometer of new geometry. *J Geophys Res* 75:1493–1505
- Bonvalot S, Diamant M, Gabalda G (1998) Continuous gravity recording with Scintrex CG-3 M meters: a promising tool for monitoring active zones. *Geophys J Int* 135:470–494
- Brady JL, Hare JL, Ferguson JF, Seibert JE, Klopping FJ, Chen T, Niebauer T (2008) Results of the world's first 4D microgravity surveillance of a waterflood–Prudhoe Bay, Alaska. *SPE Reserv Eval Eng* 11(5):824–831
- Canadian Micro Gravity (2019). <http://www.canadianmicrogravity.com/>. Accessed June 2019
- Carbone D, Rymer H (1999) Calibration shifts in a LaCoste-and-Romberg gravimeter: comparison with a Scintrex CG-3 M. *Geophys Prospect* 47(1):73–83
- Crossley D et al (1999) Network of superconducting gravimeters benefits a number of disciplines. *Eos Trans AGU* 80(11):121
- Crossley D, Hinderer J, Llubes M, Florsch N (2003) Potential of ground gravity measurements to validate GRACE data. *Adv Geosci* 1:65–71
- Crossley D, Hinderer J, Riccardi U (2013) The measurement of surface gravity. *Rep Prog Phys* 76(4):046101
- Dehlinger P (1978) *Marine gravity*. Elsevier, Amsterdam
- Foote SA, Grindeland DB (1992) Model QA3000 Q-flex accelerometer high performance test results. *IEEE Aerosp Electron Syst Mag* 7(6):59–67
- Goodkind JM (1999) The superconducting gravimeter. *Rev Sci Instrum* 70(11):4131–4152
- Hamilton AC, Brule BG (1967) Vibration-induced drift in LaCoste and Romberg geodetic gravimeters. *J Geophys Res* 72(8):2187–2197
- Hugill AL (1984) The Design of a Gravimeter with Automatic Readout. Ph.D thesis. Bedford Park, South Australia, Flinders University
- Hugill A (1990) Scintrex CG-3 automated gravity meter: description and field results. *SEG Expan Abstr Soc Explor Geophys* 9:601–604
- Ingate S, Berger J (2004) Prospects for low-frequency seismometry. IRIS Broadband Seismometer Workshop, Granlibakken
- LaCoste LJB (1934) A new type long period vertical seismograph. *Physics* 5(7):178–180
- LaCoste LJB, Romberg A (1945) US Patent 2377889

- LaFehr TR, Valliant HD, MacQueen HD (1992) High-resolution marine gravity by digital control. SEG Expan Abstr Soc Explor Geophys 11:559–560
- Laswell S, Niebauer TM, Engel R., Cassidy J, Courtier N, Henton J (2010) Recent observations of increased seismic background noise using gPhone gravity meters Micro-g LaCoste 2010. <http://www.microglacoste.com/gPhoneNoise/gPhoneSeismicNoise.pdf>
- Mansinha L, Kamal (1992) A test of the superconducting gravimeter as a long-period seismometer. Phys Earth Planet Inter 71(1–2):52–60
- Melton BS (1976) The sensitivity and dynamic range of inertial seismographs. Rev Geophys Space Phys 18(2):93–116
- Merlet S, Kopaev A, Diamen M, Geneves G, Landragin A, Pereira Dos Santos F (2008) Micro-gravity investigations for the LNE watt balance project. Metrologia 45(3):265–274
- Meurers B (2018) Scintrex CG5 used for superconducting gravimeter calibration. Geodesy Geodyn 9(3):197–203
- Micro-g LaCoste (2019). <http://www.microglacoste.com/>. Accessed June 2019
- NOAA Grav-D project (2017). https://www.ngs.noaa.gov/GRAV-D/data/NGS_GRAV-D_General_Airborne_Gravity_Data_User_Manual_v2.1.pdf. Accessed June 2019
- Richter B, Wenzel H-G, Zürn W, Klotting F (1995) From Chandler wobble to free oscillations: comparison of cryogenic gravimeters and other instruments in a wide period range. Phys Earth Planet Inter 91:131–148
- Scintrex (2019). <http://www.scintrex.com/>. Accessed June 2019
- Studinger M, Bell R, Frearson N (2008) Comparison of AIRGrav and GT-1A airborne gravimeters for research applications. Geophysics 73(6):151–161
- Threlfall R, Pollock JA (1900) On a quartz thread gravity balance. Philos Trans R Soc Lond A 193:215–260
- Torge W (1989) Gravimetry. de Gruyter, Berlin
- Van Camp M (1999) Measuring seismic normal modes with the GWR C021 superconducting gravimeter. Phys Earth Planet Inter 116:81–92
- Van Camp M, Francis O (2007) Is the instrumental drift of superconducting gravimeters a linear or exponential function of time? J Geod 81:337–344
- Van Camp M, Viron O, Watlet A, Meurers B, Francis O, Caudron C (2017) Geophysics from terrestrial time-variable gravity measurements. Rev Geophys 55:938–992. <https://doi.org/10.1002/2017RG000566>
- van Kann F (2004) Requirements and general principles of airborne gravity gradiometers for mineral exploration. In: Lane RJJ (ed) Airborne gravity 2004 – abstracts from the ASEG-PESA airborne gravity 2004 workshop, Geoscience Australia Record 2004/18, pp 1–5
- Webring MW, Kucks RP, Abraham JD (2004) Gravity study of the Guernsey landfill site, Guernsey, Wyoming. U. S. Geological Survey, Reston
- Wei M, Schwarz KP (1995) Analysis of GPS derived acceleration from airborne tests. In: Proceedings on the IAG symposium on Airborne Gravity Field Determination at the IUGG XXI General Assembly, 2–14 July 1995. University of Calgary, Department of Geomatics Engineering, Alberta, pp 175–188
- Williams S, MacQueen JD (2001) Development of a versatile, commercially proven, and cost-effective airborne gravity system. Lead Edge 20(6):651–654
- Wilson CDV (1953) An analysis of the vibrations emitted by some man-made sources of microseisms. Proc R Soc London Ser A 217:188–202
- ZLS (2019). <http://zlsCorp.com/>. Accessed June 2019
- Zumberge M, Alnes H, Eiken O, Sasagawa G, Stenvold T (2008) Precision of seafloor gravity and pressure measurements for reservoir monitoring. Geophysics 73(6):133–141

Gravity Anomalies, Interpretation

Mikhail K. Kaban

Department 1: Geodesy and Remote Sensing, Sektion 1.3: Earth-System Modelling, Deutsches GeoForschungsZentrum Potsdam – GFZ, Potsdam, Germany

Definition

In geophysics *gravity anomalies* are generally defined as the difference between observed gravity field and the field of a reference model. Depending on the reference gravity model, two different types of anomaly variations are considered: *gravity anomalies* and *gravity disturbances*. The geodetic *gravity anomaly* is defined as the difference between gravity on the geoid and normal gravity on the reference ellipsoid (Blakely 1995; Heiskanen and Moritz 1967). On the other hand, the *gravity disturbance* is defined as the difference of the fields at the same point on the reference ellipsoid. It has been demonstrated that the gravity disturbances are more appropriate for geophysical purposes (e.g., Hackney and Featherstone 2003). In any case, it is necessary to take into account this difference in the interpretation. For more details see ► “Gravity, Data to Anomalies.” Here we use the general term “anomalies” for both types as traditionally accepted in geophysics.

The observed gravity anomalies reflect the effect of density variations relative to the homogeneous reference model. *Interpretation* of the gravity anomalies implies an estimation of these density heterogeneities. The density model should reproduce the observed gravity field, taking into account that the observations may be affected by measurement errors. A separate type of modeling, which may be also related to the interpretation, is an *enhancement* of the gravity anomalies. The observed field combines effects of various mass anomalies integrated over the whole Earth. Therefore, it is often difficult to select the anomaly, which relates to the object under study. The usual way to cope with the problem is to remove the effects, which mask the anomalies of interest (see also ► “Gravity Data, Regional: Residual Separation”).

Introduction

Density heterogeneity of the Earth, associated with thermal and compositional variations or with deflected boundaries separating layers of different density, is one of the main factors that control dynamic processes and deformations at both shallow and deep levels. Therefore, interpretation of the gravity anomalies or *gravity modeling* is one of the principal methods, which help to understand nature and origin of the

tectonic processes and the Earth's dynamics. On the other hand, gravity field is very sensitive to shallow heterogeneities, in particular related to mineral deposits. For this reason, gravity modeling is an important tool for both fundamental and practical purposes.

Formulation of the Problem

The relationship between the density distribution ρ and the gravity field g may be written as an operator equation:

$$A\rho = g, \quad (1)$$

where A is the integral operator. In Cartesian coordinates for a 3D body with density $\rho(x', y', z')$, this equation may be written as (Blakely 1995):

$$g(x, y, z) = \frac{\partial U}{\partial z} = -\gamma \iiint_{z' y' x'} \rho(x', y', z') \frac{(z - z')}{r^3} dx' dy' dz', \quad (2)$$

where (x, y, z) are the observation points, U is the gravitational potential, γ is the gravitational constant, and $r = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}$.

Therefore, the problem of determination of the density distribution ρ may be formulated as finding:

$$\min \left\{ \|A\rho - g_{obs}\|^2 \right\}, \quad (3)$$

where g_{obs} is the observed gravity anomaly.

Unfortunately, the problem of gravity data inversion is generally *ill posed*. The operator A has no continuous inverse A^{-1} ; thus the solution is neither unique nor stable. To overcome this principal weakness, Eq. 3 should be modified by adding a so-called regularization term Ω (Tikhonov and Arsenin 1977):

$$\min \left\{ \|A\rho - g_{obs}\|^2 + \alpha \Omega(\rho) \right\}, \quad (4)$$

where α defines strength of regularization. The regularization implies some a priori conditions imposed on the density distribution in order to limit a possible range of the solutions. If these conditions are sufficient, the problem becomes *well posed*.

There are basically two ways to obtain a unique and stable solution with the means of interpretation of the gravity anomalies. The first option implies considering simplified models, the parameters of which should be determined. An inclusion of a certain form in a homogeneous or stratified material, a thin layer with variable density, or an interface between two

layers with different density contrasts are the methods with which the number of unknowns are reduced and the problem is generalized. These models are adequate for modeling strong density contrasts, like the ocean bottom, the basement, and the crust–mantle interface, and are optionally used jointly with the concept of *isostasy*. The methods assuming a spatially restricted source are suited in the exploration of superficial bodies as ore bodies, diamond bearing Kimberlite cylinders, and salt domes. The methods using a density interface are used to define the ocean bathymetry from the gravity field or to determine the Moho, especially in remote areas where seismic studies are scarce or insufficient (e.g., Silva et al. 2006; Smith and Sandwell 1997; Van der Meijde et al. 2015). The other way to cope with the problem implies joint analysis of the gravity anomalies together with supplemental geophysical information, in particularly with seismic data that identify the basic structure of the Earth interiors. Using this method, it is possible to construct advanced models, which describe the structure of the Earth more reliably. Two basic interpretation techniques are used within this approach: forward method and inverse method (Blakely 1995) (see also ► “Gravity Method, Principles”).

Gravity Anomalies: Conditions and Enhancement

Spatial *resolution* of the gravity anomalies should be adapted to the scale and purpose of the study; it should be equal or better than the size of the smallest object of interest. *Local studies* range from meters to several tens of kilometers. They are chiefly intended to geophysical prospecting, archaeological, and construction purposes. It is also a useful tool to investigate local tectonics, for example, related to volcanic activity. If possible, gravity anomalies defined at the original observation points should be used for these investigations. It is desirable to remove a regional trend from the observed anomalies. It is clear from Eq. 2 that the gravity effect of any source rapidly decreases when we move this source to a greater depth. A useful approximation may be obtained from a spectral representation of the gravity field. A gravity effect of density variations at the depth Z having the wavelength L is reduced by the factor:

$$\exp(-2\pi Z/L) \quad (5)$$

compared to the effect of near-surface variations.

Regional studies range from tens to first thousands of kilometers. They include various tectonic applications. The required resolution of the gravity anomalies is within approximately $5'$ – $30'$. Since the original field contains sharp variation, a special attention should be paid to the data preparation. Normally, the observed values are averaged over the compartments that correspond to the grid used in the study. *Global studies* include the whole Earth and are intended to

investigate structure of the planet down to the core–mantle boundary as well as global mantle convection. The required data resolution normally ranges from 1° to 5° .

Enhancement of the gravity anomalies. A straightforward way to reduce the ambiguity of the gravity interpretation is to calculate the gravity effect of those structures, which are known from other geophysical methods, and to remove it from the observed anomalies. Then, the residual will highlight the gravity effect of “unknown” heterogeneities. *Bouguer gravity anomalies* represent the first step of the gravity “stripping.” However, this procedure may be extended using the data on the crustal structure obtained from seismic studies. Attempts to calculate residual “crust-free” gravity anomalies have been made since the first seismic profiles were measured, but a reliable 3D model could be constructed only after sufficient initial data on the crustal structure had been accumulated. One of the first maps of the residual gravity field covering a considerable part of North Eurasia was calculated by Artemjev et al. (1994).

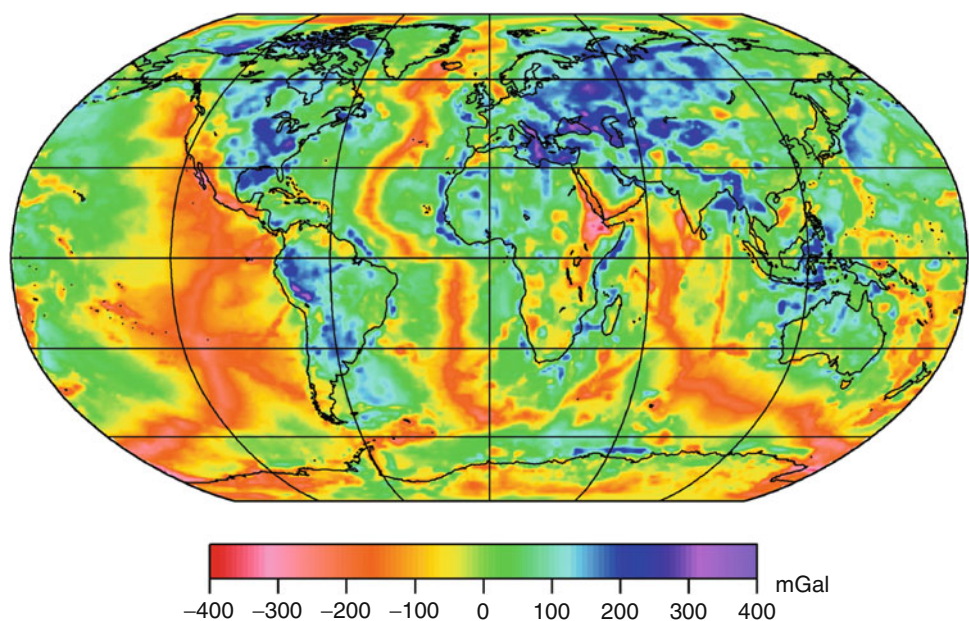
After construction of 3D models of the crust based on seismic data, this method of the gravity interpretation has been extensively used over the last decade (e.g., Kaban et al. 2003). Detailed maps of sedimentary basins and their density structure are available from numerous prospecting surveys. A number of seismic data showing crustal structure down to the crust–mantle boundary (Mohorovičić discontinuity) have also significantly increased. Based on these data, a number of 3D crustal models with regional and global coverage have been compiled (e.g., Tesauro et al. 2008; Laske et al. 2013). Density distribution within the crystalline crust is determined based on the relationships between seismic velocity and density (e.g., Christensen and Mooney

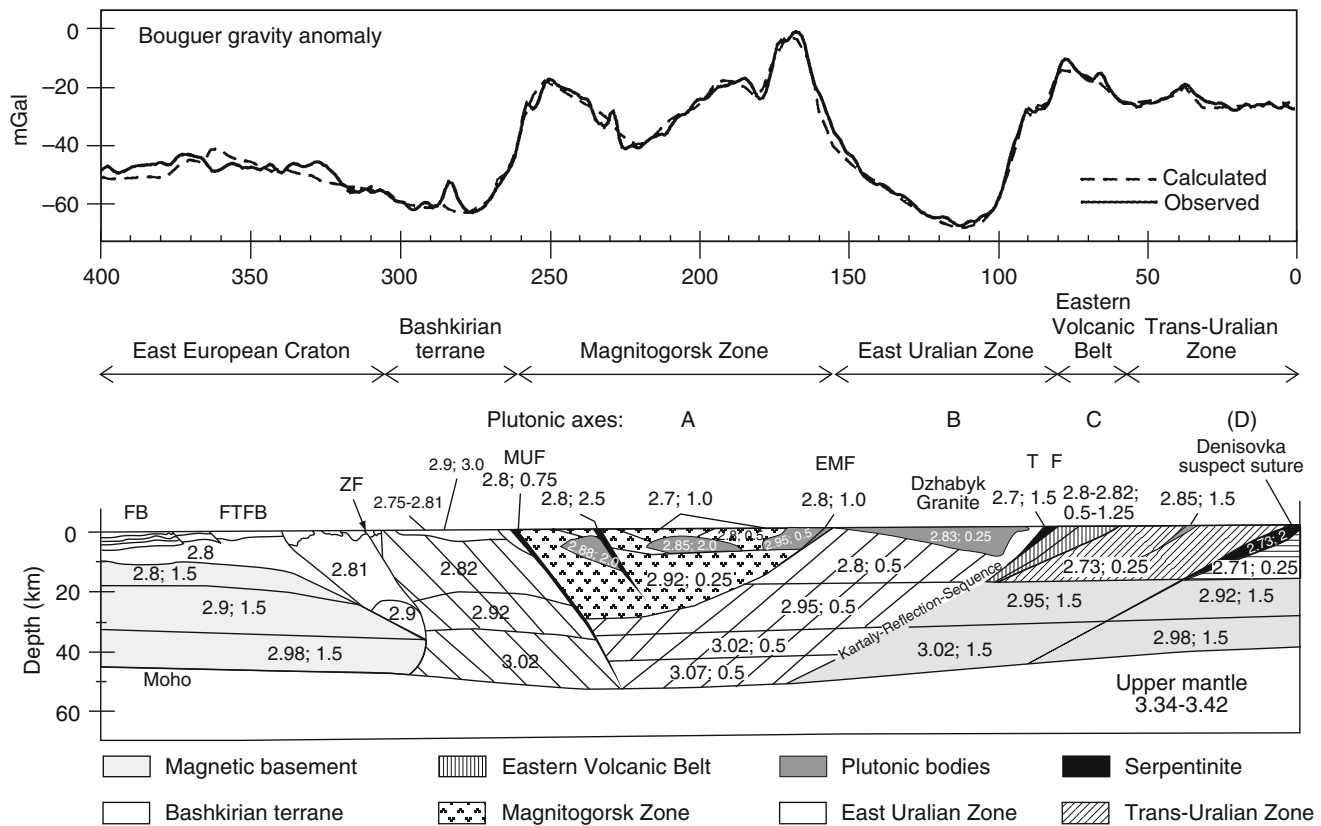
1995). Then, it is possible to calculate gravity effect of the crustal layers and to remove it from the observed field. The modern calculation methods employ integral equations similar to Eq. 2 (also in a spherical case) and can provide any desirable accuracy (e.g., IGMAS, <http://www.gravity.uni-kiel.de/igmas/>). While doing these calculations, it is important to include the regions surrounding a study area. The distance, on which this effect might be important, may be estimated from Eq. 5. For the continental regions with deep crust–mantle boundary, this distance should be at least 500 km. One of the first global maps of the residual mantle anomalies is shown in Fig. 1 (Kaban et al. 2003). This map is extensively used in studies of the mantle structure and dynamics.

Forward Method

A gravity model is constructed using an iterative approach. This type of interpretation is used when velocity models obtained from seismic profiles are used to constrain the density model. An initial model is constructed based on a priori data and modeling assumptions. The initial structure is based on seismic boundaries, and seismic velocities are converted to densities. The model’s gravity anomaly is calculated and compared with the observed field. Then, the model parameters are adjusted using an iterative process to improve the fit between the modeled and observed anomalies. At the first stage, density of the seismic blocks is modified. If the fit is not sufficient, seismic boundaries might be changed as well. One example of such interpretation along the URSEIS profile crossing the Southern Urals is shown in Fig. 2 (Kimbell et al. 2002).

Gravity Anomalies, Interpretation, Fig. 1 Residual “mantle” anomalies obtained after removing the gravity effect of the crust (relative to a horizontally homogeneous reference model) from the observed field (Kaban et al. 2003)





Gravity Anomalies, Interpretation, Fig. 2 2D gravity model for the URSEIS profile (Kimbell et al. 2002). The first number indicates density (Mg/m^3)

Inverse Method

In the inverse method, model parameters are automatically determined by solving Eq. 4. This technique is also used for joint interpretation of the gravity and seismic data. For the regional seismic profiles, it was developed by Strakhov and Romanyuk (1985), whereas for global studies based on tomography data, by Hager and O'Connell (1981). The idea of the method implies use of seismic velocities to stabilize the solution of Eq. 4. Instead of determining density variations in each grid point of the 3D model, a velocity–density relationship for selected blocks and layers should be introduced. This relationship may vary significantly; however, it is rather stable for similar compositions and physical conditions. Therefore, after selecting several blocks and layers, which are supposed to obey the same conditions, it is possible to determine a very limited number of the model parameters (velocity-to-density scaling factors for the specified blocks and layers). This provides the possibility to obtain a unique and stable solution of the inverse problem. A clear advantage of the inverse method with respect to the forward modeling is that it provides a possibility to quantitatively control sensitivity of the gravity anomalies to variations of the model parameters and their uncertainties.

The initial density model is usually produced by a simple linear conversion of the velocity disturbances into density variations. The initial scaling factors (c_j^0) are based on mineral physics and petrological data (j is the number of the selected layer or block). In the inversion they should be rescaled in a least squares adjustment. The inversion can be performed in both space and spectral (or spherical harmonic) domains. The scaling coefficients a_j (to rescale the initially adopted values) are estimated by minimizing the function (Kaban and Schwintzer 2001):

$$\chi^2 = \sum_i \cos \phi^i \left(dg_{obs}^i - \sum_j \hat{a}_j dg_j^i \right)^2 + \sum_j \alpha_j (\hat{a}_j - 1)^2, \quad (6)$$

where dg_j^i is the initial gravity effect of layer or block j at the surface grid point i characterized by the gravity anomaly dg_{obs}^i . $\cos \phi^i$ is a formal weight factor to account for the convergence of the meridians in the equal angular grid. It is not required in case of the equal space grid or spherical harmonic inversion. The last term is used to stabilize the solution. In this case it requires that the estimated scaling

factors should not be significantly different from the initial ones. α controls strength of regularization. Equation 6 leads to the normal equation system with a dimension equal to the number of the unknown scaling factors. In spherical harmonic domain, a similar equation can be derived for each set of coefficients for specific spherical harmonic degree/number, which can be solved separately under certain conditions (Hager and O'Connell 1981).

In global studies this technique is used for interpretation of both geoid and gravity anomalies. Geoid is more sensitive to the effect of deep interiors, whereas a use of the gravity anomalies provides better resolution, especially for the upper mantle (e.g., Simmons et al. 2009). An example of such study is shown in Fig. 3. The velocity-to-density scaling factor is determined for the oceanic upper mantle in the joint inversion of the residual mantle gravity anomalies and seismic tomography (Kaban and Schwintzer 2001). The obtained density–velocity ratios for the subcrustal layer (down to 75 km) and for the deep part of the oceanic upper mantle (below 225 km) are very close to the ones predicted from the mineral physics analysis. However, for the oceanic layer involving the depths 75–225 km, a significantly lower density–velocity ratio is obtained. This may be explained by a compositional layering: dry, partially depleted peridotite residuum underlain by “damp” fertile peridotite and separated by the G discontinuity at a depth of about 70 km. This result is important for geodynamic modeling. It evidences that the base of an oceanic lithosphere plate is defined by a compositional rather than thermal boundary layer or, at

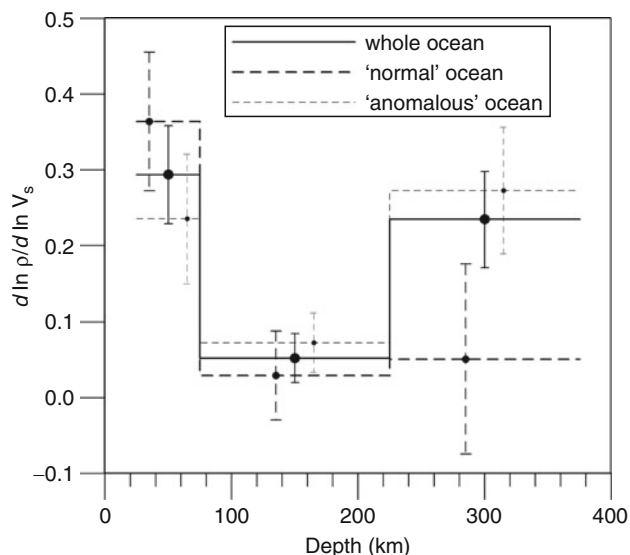
least, the base is strongly influenced by a compositional boundary.

In recent years, it has been developed a new integrative technique for construction of 3D density models of the upper mantle. In addition to the residual mantle gravity anomalies and seismic tomography, this method employs the residual topography that is the topography, which is supported by the mantle density heterogeneity. The residual gravity and topography are controlled by the same factors but in a different way (e.g., depending on depth and horizontal dimension of the heterogeneity.) This enables us to locate the position of principal density anomalies in the upper mantle (Kaban et al. 2016). One example of such inversion for the Middle East is shown in Fig. 4.

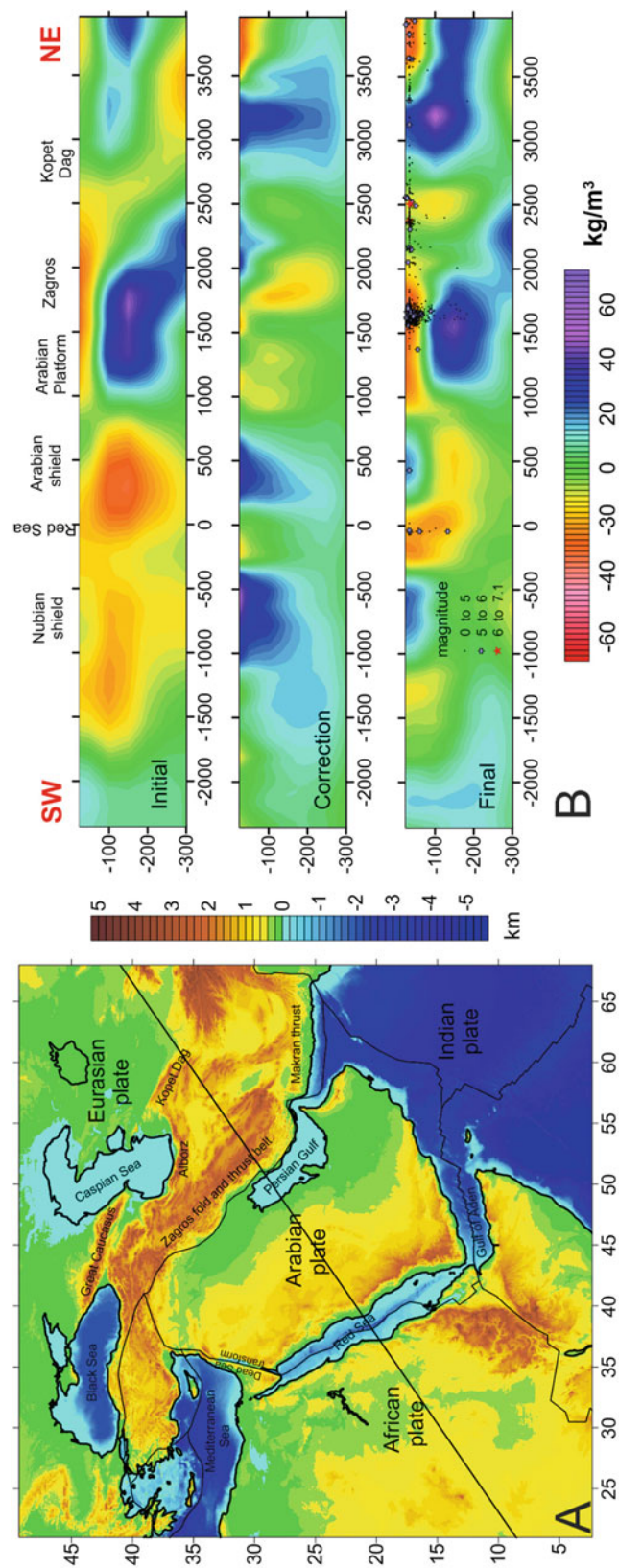
Summary

Due to new interpretation techniques and increased computer power, most of the computational problems related to the forward and inverse gravity modeling, which caused big problems even recently, are solved. For example, the direct computations using Eq. 2 may be performed for on large grids taking into account 3D density variations. Also a principal difference from the early stage period is that the interpretation of the gravity anomalies is performed now in combination with other methods. The most profitable results are obtained from the integrative interpretation of the gravity and seismic data. Further combination with geothermal data provides the possibility to separate the effects of temperature and composition variations within the Earth. The obtained density models are used in various applications. This information is essential for a geodynamic modeling that results in determination of the mantle flows, tectonic forces moving the lithosphere blocks, and stress distribution in the crust and mantle at both regional and global levels. Interpretation of the gravity anomalies is also widely used in geophysical prospecting. In some situations the seismic investigations fail to give results, for example, when a high-velocity layer masks the underlying low-velocity layers. Here the gravity and magnetic investigations are the principal means to obtain information on the underlying structure.

Two principal problems should be a key issue for the gravity modeling in future. As it was shown, reliability of the gravity modeling depends on additional data used in the interpretation. Therefore, it is necessary to bring together data from various disciplines, which might be used in construction of integrative interpretation models. Geophysicists just started approaching this ultimate goal. The second principal issue is improvement of the technique for integrative interpretation. In particular, special attention should be paid to solution of nonlinear inverse problems and to ultrahigh dimension problems.



Gravity Anomalies, Interpretation, Fig. 3 The velocity-to-density scaling coefficients adjusted for “normal” and “anomalous” oceanic areas, respectively, compared with the result for the whole oceanic area. Only the scaling coefficients and standard deviations for the upper mantle are shown, but the inversion was performed for the whole mantle (Kaban and Schwintzer 2001)



Gravity Anomalies, Interpretation, Fig. 4 3D density model for the Middle East. (a) Location of the profile. (b) Initial density model based on seismic tomography (upper profile); correction obtained in a joint inversion of the residual mantle anomalies and residual topography (middle) and final density model (bottom), Kaban et al. (2016)

Cross-References

- Gravity Data, Advanced Processing
- Gravity Data, Regional: Residual Separation
- Gravity Method, Principles
- Gravity, Data to Anomalies
- Isostasy

Bibliography

- Artemjev ME, Kaban MK, Kucherinenko VA, Demjanov GV, Taranov VA (1994) Subcrustal density inhomogeneities of Northern Eurasia as derived from the gravity data and isostatic models of the lithosphere. *Tectonophysics* 240:249–280
- Blakely RJ (1995) *Potential theory in gravity and magnetic applications*. Cambridge University Press, New York. 441 p
- Christensen NI, Mooney WD (1995) Seismic velocity structure and composition of the continental crust: a global view. *J Geophys Res* 100:9761–9788
- Hackney RI, Featherstone WE (2003) *Geophys J Int* 154:35–43
- Hager BH, O’Connell RJ (1981) A simple global model of plate dynamics and mantle convection. *J Geophys Res* 86:4843–4867
- Heiskanen W, Moritz H (1967) *Physical geodesy*. W.H. Freeman, San Francisco
- Kaban MK, Schwintzer P (2001) Oceanic upper mantle structure from experimental scaling of V_S and density at different depths. *Geophys J Int* 147:199–214
- Kaban MK, Schwintzer P, Artemieva IM, Mooney WD (2003) Density of the continental roots: compositional and thermal contributions. *Earth Planet Sci Lett* 209:53–69
- Kaban MK, El Khrepy S, Al-Arifi N, Tesauro M, Stolk W (2016) Three dimensional density model of the upper mantle in the Middle East: Interaction of diverse tectonic processes. *J Geophys Res Solid Earth* 121. <https://doi.org/10.1002/2015JB012755>
- Kimbell GS, Ayala C, Gerdes A, Kaban MK, Shapiro VA, Menshikov YP (2002) Insights into the architecture and evolution of the southern and middle Urals from gravity and magnetic data. In: Brown D et al (eds) *Mountain building in the Uralides: Pangea to the present*. Geophysical Monograph, vol 13. American Geophysical Union, Washington, DC, pp 49–65
- Laske G, Masters G, Ma Z, Pasyanos M (2013) Update on CRUST1. 0 – A 1-degree global model of Earth’s crust. In: *Geophysical Research Abstract*, vol 15. EGU General Assembly, Vienna, p 2658
- Silva JBC, Costa DCL, Barbosa VCF (2006) Gravity inversion of basement relief and estimation of density contrast variation with depth. *Geophysics* 71(5):J51–J58
- Simmons NA, Forte AM, Grand SP (2009) Joint seismic, geodynamic and mineral physical constraints on three-dimensional mantle heterogeneity; implications for the relative importance of thermal versus compositional heterogeneity. *Geophys J Int* 177:1284–1304
- Smith WHF, Sandwell DT (1997) Global seafloor topography from satellite altimetry and ship depth soundings. *Science* 277:1956–1961
- Strakhov VN, Romanyuk TV (1985) Reconstruction of the density of the Earth’s crust and upper mantle from deep seismic sounding and gravimetric data; I. *Izv Phys Solid Earth* 20(6):438–449
- Tesauro M, Kaban MK, Cloetingh S (2008) EuCRUST-07: a new reference model for the European crust. *Geophys Res Lett* 35. <https://doi.org/10.1029/2007GL032244>
- Tikhonov AN, Arsenin VY (1977) *Solutions of ill posed problems*. V. H. Winston, New York, p 258
- Van der Meijde M, Fadel IEAM, Ditmar P, Hamayun M (2015) Uncertainties in crustal thickness models for data sparse environments: a review for South America and Africa. *J Geodyn* 84:1–18

Gravity Data, Advanced Processing

Christopher J. Swain and Jonathan F. Kirby
School of Earth and Planetary Sciences, Curtin University,
Perth, WA, Australia

Synonyms

Gravity: spectral methods; Potential field transforms

Definition

Spectral analysis	Estimation and analysis of global frequency content of a signal, assumed stationary, most often using the Fast Fourier Transform (FFT).
Potential field transformations	Process of converting gravity (or magnetic) survey data into a new and physically meaningful form to facilitate its geological interpretation.
Euler and Werner deconvolution	Methods for automatically estimating depths to sources from gravity (or magnetic) survey data.
Wavelet analysis	Estimation and analysis of local frequency content of a non-stationary signal, using a wavelet transform (WT).

Introduction

The use of spectral analysis to interpret gravity anomalies goes back to the 1930s, but the modern approach in terms of the Fourier transform (FT) was developed in the 1960s and 1970s. Most of the methods are equally applicable to magnetic data, which is their more common area of use because of the huge volume of aeromagnetic data now available. Here we will not generally distinguish between the two, though the magnetics case is complicated by the fact that the field and magnetization directions are not in general the same or vertical. Blakely (1995) provides an excellent introduction to the subject, which he divides into two basic aspects: *spectral modeling* and *potential field transformations*.

In spectral modeling, the major advantage of working in the wave-number ($k = 2\pi/\lambda$) domain, after Fourier transforming the data, is that it offers simplifications and insights into the relation between anomalies and their sources. For example, the FT of the gravity anomaly over a thin horizontal layer of variable surface density $\sigma(x, y)$, at depth d , is simply the product of the FT of σ with a constant times $e^{-|k|d}$. Other models similarly separate into products of factors.

Potential field transformations can be carried out in the space domain by either numerical integration or convolution, but in the wave-number domain, the simple forms of the transformations offer considerable insight. They include upward and downward continuation, and vertical and horizontal derivatives, all of which have their uses in enhancing resolution and separating anomalies.

Both aspects were made practical by the development of the FFT, though it should be noted that there are important differences between the theoretical (continuous, infinite) FT and the (discrete, finite) FFT – see below.

The FT assumes that the statistical properties of the data do not vary spatially (i.e., the data are *stationary*), which may not be the case with real data. The wavelet transform allows spectral analysis without this assumption and therefore promises improvements over the FT.

Spectral Modeling

Many applications of spectral modeling have been devised: for example, Blakely (1995) gives expressions for the FT of the anomalies for a number of simple models. Here just two of the more commonly used methods will be described.

Parker (1972) extended the simple horizontal layer model (mentioned above) to the case of a layer bounded by irregular top and bottom surfaces. This allows the very rapid calculation (via the FFT) of the gravity anomaly over a square grid for such common cases as sedimentary basins, marine Bouguer corrections from bathymetry, and Airy isostatic corrections from topography.

The method of *statistical models* was pioneered by Spector and Grant (1970) for interpreting magnetic data. In the gravity case, the method assumes that the anomalous field results from an *ensemble* of dense vertical prisms where each of the model parameters is randomly distributed with uniform probability about a mean value. As mentioned above, the FT of the anomaly of each prism is the product of several factors, notably a depth factor ($e^{-|k|d}$), a depth extent (t) factor ($1 - e^{-|k|t}$), and a size factor (a function of the width and breadth of the prism). The method employs the radial *power spectrum* of the observed anomaly, i.e., averaged over concentric annular rings in the (k_x, k_y) plane, implying that isotropy is assumed. The theoretical power spectrum for the ensemble of prisms is the product of the *ensemble averages* of the square of each of the above factors. If the ranges of the parameters are not too large, the ensemble averages will approximate the factors for their mean values. At wavelengths much larger than the body dimensions, the size factor tends to a constant. This is equivalent to assuming a dipole, rather than prism, model. Then the mean depth and depth extents can be estimated by plotting the logarithm of the power spectrum. The depth factor gives a straight line of slope $-2d$ at medium to large k , and the depth

extent factor causes this slope to decrease as k decreases, giving rise to a maximum. Thus in principle the mean values of both d and t can be estimated. The basic problem with the method lies in knowing to what extent the assumptions are justified. If the body dimensions are not small compared to the wavelengths, then it may be possible to estimate their average values, calculate a size factor, and hence correct the power spectrum. But if parameter values are not completely random and uncorrelated, but are self-similar or fractal (for which there is some evidence) then the power spectrum will be proportional to a power of k , invalidating depth estimates. See Blakely (1995) for references.

Potential Field Transformations

Transformations of gravity and magnetic data have a long history in applied geophysics (e.g., Grant and West 1965), since their utility was realized long before they could be easily and accurately applied. Three examples of transformations that are commonly used in gravity interpretation are listed below with their uses.

- (a) Vertical derivative: largely eliminates the regional and improves resolution.
- (b) Upward continuation: very efficiently attenuates nearsurface effects; used as a standard “separation filter” (Jacobsen 1987).
- (c) Downward continuation: enhances resolution, i.e., it allows the separation of closely spaced or overlapping anomalies; approximate inversion for a uniform density layer.

These operations depend on a theorem from potential theory stating that the gravity potential caused by any 3D density distribution is identical to that of a thin layer of mass spread over one of its equipotential surfaces. This is *Green’s equivalent layer* (Blakely 1995). For example in the case of upward continuation of gravity data measured on a level surface at height z_0 to a second level surface Δz above it, the equivalent layer is taken to lie in the measurement plane. The theoretical formula for the upward continued field at a point $P(x, y, z_0 + \Delta z)$ is an integral over the complete (and infinite) plane measurement surface since each point in that plane acts as an element of mass. This integral can then be written as a convolution of the data with $\Delta z/r^3$, where r is the distance from point Q in the z_0 plane to point P . Fourier transforming this expression reveals that the FT of the upward continued data is the product of the FT of the measured data and $e^{-|k|\Delta z}$, where $k = \sqrt{(k_x^2 + k_y^2)} = 2\pi/\lambda$ (radians/m). This simple view of upward continuation as a filter with wave-number response $e^{-|k|\Delta z}$ offers physical insight that explains immediately why

the operation is effective at attenuating (exponentially) near-surface density anomalies.

The responses of some other common transforms are: downward continuation through height Δz , $e^{|k|\Delta z}$; vertical derivative, $|k|$; vertical integral (potential), $1/|k|$; x-derivative, ik_x . One transform that is particularly useful in mapping is the horizontal gradient amplitude (HGA): $\sqrt{[(\partial g/\partial x)^2 + (\partial g/\partial y)^2]}$, whose maxima occur vertical contacts. Another is the analytic signal amplitude (ASA): $\sqrt{[(\partial g/\partial x)^2 + (\partial g/\partial y)^2 + (\partial g/\partial z)^2]}$.

Using the FFT is much faster than either numerical integration or convolution, except for very small arrays of filter coefficients such as the centered difference approximation generally used for horizontal derivatives. It also implies certain requirements of the data (usually met by *preprocessing*) because the FFT differs from the continuous FT in important ways (e.g., Blakely 1995). First, it is discrete, with (1D) data spacing Δx , which means that the FFT is periodic, repeating itself every $k = 2\pi/\Delta x$. This leads to aliasing of wave numbers outside the band defined by the *Nyquist* wave number $k = \pm \pi/\Delta x$ into the FFT at smaller wave numbers, unless the data are band limited. The preprocessing can include an antialias filter to ensure this. Second, the FFT is a finite summation over a data length $N\Delta x$, rather than the infinite integral of the FT, and hence implicitly assumes that the data set is periodic, infinitely repeating itself or, equivalently, circular with the last point next to the first. To avoid first order discontinuities at the data edges (which would lead to the Gibbs effect), the preprocessing then includes extending the data in such a way as to produce a smooth transition between neighboring repetitions. With these precautions, the FFT method is capable of accurate results for most transformations, unless they require amplification of wave numbers that are not present in the data, due to it's being either too short (and therefore lacking low wave numbers – below $1/N\Delta x$) or digitized with too large an interval Δx , or otherwise inaccurate.

A modern example of an application that uses a transformation via the FFT occurs in processing data from the Falcon™ airborne gravity gradiometer system, whose sensor measures the horizontal curvature components (Dransfield and Lee 2004; Nettleton 1976). The easiest way to convert these components to the more familiar vertical gradient is to grid them and then apply a complex transformation via the FFT.

Euler and Werner Deconvolution

These are approximate inversion methods for automatically locating the depths to the tops of multiple sources. The principles of both methods are dealt with by Blakely (1995). The name is by analogy with seismic deconvolution where seismic reflection arrivals are converted into impulses

representing reflectors. The methods are similar in many ways, but Werner's method in its original form is only applicable to aeromagnetic profile data, whereas the Euler method can be applied to profiles or grids of magnetic or gravity data. However, generalizations of the methods to different types of gradient data have been made and it has been shown that Werner's method using the analytic signal is in fact a special case of Euler deconvolution (Nabighian and Hansen 2001). Werner's method will not be described in detail, but his idea was to linearise the magnetic anomaly on a profile perpendicular to a thin dipping dyke, which contains four unknowns; hence any four measurements along the profile will give four simultaneous equations which can be solved for the position and depth of the top of the dyke.

Euler Deconvolution

Euler deconvolution (Thompson 1982; Reid et al. 1990) is very similar but uses a more general equation than that of a dyke, namely Euler's equation, which can approximately model the behavior of many different simple sources. Euler's equation is satisfied by any homogeneous function $f(x, y, z)$ of degree n (one that is invariant under scaling) and can be written:

$$x \frac{\partial f}{\partial x} + y \frac{\partial f}{\partial y} + z \frac{\partial f}{\partial z} = -nf$$

For example, $f = 1/r$ satisfies this equation with $n = 1$, implying that so does the potential of a point mass. The degree n is called the *structural index*, and defines the attenuation rate of the field f due to the particular source. Examples for gravity field anomalies are (FitzGerald et al. 2004): point mass, 2; line mass, 1; finite step, 0; contact, -1 . Reid et al. (1990) showed that certain extended bodies like thin dipping sheets (e.g., irregular dykes and sills) can also satisfy Euler's equation, but this is not generally true.

In practice, the gradients are generally calculated from a grid of gravity anomaly data as described above. In order to locate the center of a body (x_0, y_0, z_0) from survey data, it is necessary to replace (x, y, z) in Euler's equation by $(x-x_0, y-y_0, z-z_0)$. Implementation of a grid Euler method follows that of a least squares version of Werner's method, where simultaneous equations are formed from a square array (or window) of grid points of between 3×3 and 10×10 equations and a LS solution found for x_0, y_0, z_0 and a background field B (i.e., f is replaced by $f-B$ on the RHS of the equation). In this case, n cannot be solved for, but must be assumed and the normal procedure is to try several values and select the one giving the most consistent results. The procedure is repeated for all points of the grid, and may generate large numbers of solutions, many of them unrealistic because the assumption of a simple source with constant n for each window is invalid. In practice, most of these solutions can be

easily eliminated by selecting only those that (a) have standard deviations for x_0 , y_0 , z_0 (obtained from the LS solution) of <5–10% of the depth and (b) fall within the window. Another useful way to restrict the number of solutions is to apply the method only close to localized anomalies as defined, for example, by areas of positive curvature of the horizontal gradient (HGA).

Extended Euler Deconvolution

For the 2D case, Mushayandebvu et al. (2001) discovered a second equation, similar to Euler's, for homogeneous functions that are also invariant under rotation. They rewrote the two equations for the particular cases of a magnetic contact and a thin dyke so that their right hand sides include terms involving the susceptibility and dip. Their *extended Euler* method finds solutions satisfying both equations that yield susceptibility and dip estimates as well as x_0 , y_0 , and z_0 . Nabighian and Hansen (2001) then showed that the additional equation could be written in terms of the *Hilbert transform* of f (e.g., Li 2006) which shows that it is just Euler's equation applied to the Hilbert transform. They further extended the new equation to 3D, in terms of generalized x- and y- Hilbert transforms (which can be represented and calculated as potential field transformations with wave-number domain responses $H_x = -ik_x/|k|$ and $H_y = -ik_y/|k|$, as shown by Nabighian 1984). It has been found that, assuming a value for the structural index and using all three equations for the anomaly data and their x- and y- Hilbert transforms gives better-constrained solutions than using just one equation for the anomaly data. Moreover, the new equations allow a number of other possibilities, including solving for the structural index (FitzGerald et al. 2004).

Tensor Deconvolution

Although gradiometer data can be used with advantage in Euler deconvolution, Mikhailov et al. (2007) show that with full tensor ("FTG") data it is possible to devise a method that dispenses with data windows and locates a source beneath each measurement point based on invariants calculated from the five independent tensor components. They suggest that since this method appears to locate the centers of bodies, whereas Euler deconvolution locates their edges, the methods should complement each other.

Wavelet Analysis, with Advantages and Limitations

Introduction

The wavelet transform (WT) was developed in the early 1980s, in an attempt to better understand the frequency content of seismic signals (e.g., Grossman and Morlet 1984). Whereas the Fourier transform (FT) provides the amplitudes

and phases of all harmonics present in the entire signal (a *global spectrum*), the WT yields these quantities at each time (or space) location of the signal (*local spectra*). For instance, the WT of Beethoven's Ninth Symphony tells one when to play C major, or when to play D minor chords. The FT, however, merely tells the musician that C major and D minor chords occur somewhere in the symphony (e.g., Antoine et al. 2004).

For this reason, the WT is superior in analyzing the harmonic constituents of *non-stationary* signals, that is, signals whose harmonics vary throughout the signal. The application to the Earth's potential fields is evident. The depth, spatial extent, and density/magnetization contrast of sources vary widely over a given region, leading to anomaly maps with clear differences in spectral characteristics from place to place. The wavelet power spectrum, or "*scalogram*," of such a map thus quantifies the relative power of each harmonic at each spatial location; in contrast, the Fourier power spectrum shows only the total harmonic composition of the map.

The WT has two distinct forms, termed "continuous" and "discrete." Unfortunately, there is no analogy between this use of the terms and their use with the FT: as mentioned, the continuous Fourier transform (CFT) is an integral equation applied to continuous analytic formulae; the discrete Fourier transform (DFT) is its specialization to discretely-sampled signals. Both CFT and DFT, however, use the same *basis functions*, the complex exponentials (sines and cosines). However, with the discrete and continuous wavelet transforms (DWT and CWT, respectively), the names "discrete" and "continuous" apply not to the signal, but to the basis functions (the wavelets). So both CWT and DWT can be applied to discretely sampled signals, and the difference between them lies (mainly) in the wavelets.

Wavelets, whether discrete or continuous, differ from the basis functions of the FT (infinitely or quasi-infinitely extending sinusoids and cosinusoids) in that they are *localized* in space (or time). That is, while they are oscillating functions (with zero mean value), their amplitude decays rapidly with distance (or time) from their center. It is this feature which localizes the wavelet spectra to a region surrounding the point of interest. Contrast this with the DFT, where the complex exponentials maintain their amplitude over the whole study area, rendering localization impossible (though see below). However, a high degree of localization in the space (or time) domain results in a poorer degree of localization in the wave-number (or frequency) domain. This property is referred to as the *uncertainty relation* of spectral analysis, by analogy with Heisenberg's uncertainty principle of quantum mechanics. In the spectral case, the more localized the position (small Δx), the higher the uncertainty in the wave-number content of the signal at that location (large Δk); and vice versa, according to:

$$\Delta x \Delta k \geq 2\pi$$

For the DFT, since the Fourier transform of a sine wave is a delta-function (i.e., a spike in the frequency domain), Δk is very (infinitesimally) small and precise, whereas Δx is very (infinitely) imprecise. For the WT, there exists a trade-off between Δx and Δk according to the *scaling* of the wavelet.

It is this scaling that allows harmonic analysis. First, the WT of a given signal is a space (or time) domain convolution of the signal with scaled versions of a mother wavelet, giving *wavelet coefficients* at that scale. When the *wavelet scale* is large, the wavelets are stretched (in space or time), and the convolution yields the low-wave-number (or frequency) components of the signal; when the scale is small the wavelets are squeezed, and the convolution returns the high-wave-number (or frequency) components. However, the uncertainty relation means that large-scale wavelet coefficients tend to be spatially (or temporally) imprecise but spectrally precise, while small-scale coefficients will have the opposite properties.

The difference between the CWT and DWT lies in the choice of scales. In the DWT, the scales are chosen so that the wavelets are orthogonal. This naturally limits the range of scales to the minimum possible without loss of information; a property which is obviously preferable for image compression (e.g., Antoine et al. 2004), and means that the signal can be reconstructed very rapidly from the wavelet coefficients. The scales in the CWT, on the other hand, are not constrained, and the wavelets are hence non-orthogonal. This, though, offers redundancy in the wavelet coefficients which can be put to good use in analysis and interpretation, though reconstruction of the signal is slower.

The CWT and Windowed Fourier Transform

There are many kinds of discrete and continuous wavelet, each having certain properties that prove useful in a variety of applications. For instance, Morlet wavelets yield global spectra that best approximate those from the Fourier transform (Kirby 2005), while Poisson wavelets have a close relationship with the Laplacian operator, lending themselves to potential field applications (e.g., Hornby et al. 1999). Furthermore, wavelets may be defined in N dimensions for analysis of N -dimensional signals; in geophysics, the most commonly used are 1D and 2D wavelets. Wavelets may also be complex or real-valued: 2D complex wavelets such as the Morlet wavelet are anisotropic, while 2D real wavelets, such as the Derivative of Gaussian wavelets, are isotropic. Torrence and Compo (1998) provide a description of several 1D continuous wavelets, while Kirby (2005) discusses various 2D continuous wavelets.

The windowed Fourier transform (WFT) is also often used to recover local spectra from a signal. In this, the basis functions are windowed complex exponentials; e.g., the Gabor transform uses a Gaussian window which slides over

the signal. In the WFT, while the frequency of the complex exponential can increase or decrease, the window width (Δx) is fixed, meaning a fixed uncertainty in frequency, by the uncertainty relation. With the CWT, however, the window width (and hence frequency uncertainty) is variable (and directly related to wavelet scale). Hence, while the WFT has its uses, it is in general inferior to the CWT for several reasons, two of which are: (1) the WFT is not as good at analyzing signals with a large bandwidth (Foufoula-Georgiou and Kumar 1994); and, (2) it is not as good as resolving closely spaced spikes or discontinuities (Addison 2002).

Applications

The advantage of being able to estimate local, as opposed to global, power or phase spectra has seen the WT applied to many disciplines, not just the Earth sciences. Indeed, a 10-s Scopus search yields over 58,000 articles with “wavelet” in the abstract, with over 4,000 published in 2009 alone. Foufoula-Georgiou and Kumar (1994), Kumar and Foufoula-Georgiou (1997), Addison (2002), and Antoine et al. (2004) discuss many. For example, Moreau et al. (1997) and Hornby et al. (1999) show the usefulness of the CWT for location and edge-detection of sources within potential field data, and subsequent inversion of the wavelet coefficients, by considering the vertical variation of the field and its horizontal derivatives. Ridsdill-Smith and Dentith (1999) also tackle derivatives, but they use the DWT which enables localized frequency filtering. In another application of the CWT, Panet et al. (2006) reveal density anomalies below Pacific island volcanoes from the new satellite gravity data. And Kirby and Swain (2009) use the CWT to estimate spatial variations in lithospheric elastic thickness via a coherency analysis in the wavelet domain. Finally, there are several types of spherical, rather than planar, wavelet transform, which lend themselves to studies in geodesy (e.g., Klees and Haagmans 2000; Keller 2004; Chambodut et al. 2005).

Summary

This contribution outlines the theory and practice of a number of methods for rapidly processing gravity data that are in the form of a grid. The term “processing” here includes transformations, spectral estimation and modeling, and “deconvolution” to produce depth estimates. These are all widely used methods that are equally applicable to grids of aeromagnetic data. We also introduce the continuous wavelet transform which has many potential applications in gravity processing, few of which appear to have been widely used as yet.

A more extensive review of the gravity method, with comprehensive reference list, can be found in Nabighian et al. (2005).

Cross-References

- [Gravity Anomalies, Interpretation](#)
- [Gravity Data, Regional: Residual Separation](#)
- [Magnetic Data Enhancements and Depth Estimation](#)

Bibliography

- Addison PS (2002) The illustrated wavelet transform handbook. Institute of Physics, Bristol
- Antoine J-P, Murenzi R, Vanderghenst P, Ali ST (2004) Two-dimensional wavelets and their relatives. Cambridge University Press, Cambridge
- Blakely RJ (1995) Potential theory in gravity and magnetic applications. Cambridge University Press, Cambridge
- Chambodut A, Panet I, Manda M, Diamant M, Holschneider M, Jamet O (2005) Wavelet frames: an alternative to spherical harmonic representation of potential fields. *Geophys J Int* 163:875–899
- Dransfield MH, Lee JB (2004) The Falcon[®] airborne gravity gradiometer survey systems. In: Lane RJL (ed) *Airborne Gravity 2004 – abstracts from the ASEG-PESA airborne gravity 2004 workshop*: Geoscience Australia Record 2004/18, pp 15–19
- FitzGerald D, Reid A, McInerney P (2004) New discrimination techniques for Euler deconvolution. *Comput Geosci* 30:461–469
- Foufoula-Georgiou E, Kumar P (eds) (1994) *Wavelets in geophysics*. Academic, San Diego
- Grant FS, West GF (1965) *Interpretation theory in applied geophysics*. McGraw Hill, New York
- Grossman A, Morlet J (1984) Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J Math Anal* 15:723–736
- Hornby P, Boschetti F, Horowitz FG (1999) Analysis of potential field data in the wavelet domain. *Geophys J Int* 137:175–196
- Jacobsen BH (1987) A case for upward continuation as a standard separation filter for potential field maps. *Geophysics* 52:1138–1148
- Keller W (2004) *Wavelets in geodesy and geodynamics*. de Gruyter, Berlin
- Kirby JF (2005) Which wavelet best reproduces the Fourier power spectrum? *Comput Geosci* 31:846–864
- Kirby JF, Swain CJ (2009) A reassessment of spectral T_e estimation in continental interiors: the case of North America. *J Geophys Res* 114: B08401. <https://doi.org/10.1029/2009JB006356>
- Klees R, Haagmans R (eds) (2000) *Wavelets in the geosciences*. Springer, Berlin
- Kumar P, Foufoula-Georgiou E (1997) Wavelet analysis for geophysical applications. *Rev Geophys* 35:385–412
- Li X (2006) Understanding 3D analytic signal amplitude. *Geophysics* 71:L13–L16. <https://doi.org/10.1190/1.2184367>
- Mikhailov V, Pajot G, Diamant M, Price A (2007) Tensor deconvolution: a method to locate equivalent sources from full tensor gravity data. *Geophysics* 72:161–169. <https://doi.org/10.1190/1.2749317>
- Moreau F, Gibert D, Holschneider M, Saracco G (1997) Wavelet analysis of potential fields. *Inverse Prob* 13:165–178
- Mushayandebvu MF, Lesur V, Reid AB, Fairhead JD (2001) Magnetic source parameters of two-dimensional structures using extended Euler deconvolution. *Geophysics* 66:814–823
- Nabighian MN (1984) Toward a three-dimensional automatic interpretation of potential field data via generalized Hilbert transforms: fundamental relations. *Geophysics* 49:780–786
- Nabighian MN et al (2005) Historical development of the gravity method in exploration. *Geophysics* 70:63ND–89ND. <https://doi.org/10.1190/1.2133785>
- Nabighian MN, Hansen RO (2001) Unification of Euler and Werner deconvolution in three dimensions via the generalised Hilbert transform. *Geophysics* 66:1805–1810
- Nettleton LL (1976) *Gravity and magnetics in oil prospecting*. McGraw-Hill, New York
- Panet I, Chambodut A, Diamant M, Holschneider M, Jamet O (2006) New insights on intraplate volcanism in French Polynesia from wavelet analysis of GRACE, CHAMP, and sea surface data. *J Geophys Res* 111:B09403. <https://doi.org/10.1029/2005JB004141>
- Parker RL (1972) The rapid calculation of potential anomalies. *Geophys J R Astron Soc* 31:447–455
- Reid AB, Allsop JM, Granser H (1990) Magnetic interpretation in three dimensions using Euler deconvolution. *Geophysics* 55:80–91
- Ridsdill-Smith TA, Dentith MC (1999) The wavelet transform in aeromagnetic processing. *Geophysics* 64:1003–1013
- Spector A, Grant FS (1970) Statistical models for interpreting aeromagnetic data. *Geophysics* 35:293–302
- Thompson DT (1982) EULDPH: a new technique for making computer-assisted depth estimates from magnetic data. *Geophysics* 47:31–37
- Torrence C, Compo GP (1998) A practical guide to wavelet analysis. *Bull Am Meteorol Soc* 77:61–78

Gravity Data, Regional-Residual Separation

Kumarendra Mallick¹, Vasanthi Anthwar¹ and Krishna Kant Sharma²

¹National Geophysical Research Institute, Hyderabad, India

²Department of Applied Geology, University of Madras, Chennai, India

Synonyms

Gravity effect of shallower and deeper bodies; Isostatic anomaly

Definition

<i>Observed Bouguer gravity</i>	Gravity field after corrections for drift, free-air, Bouguer, terrain, and latitude.
<i>Regional gravity anomaly</i>	Gravity anomaly due to deeper and broader structures.
<i>Residual gravity anomaly</i>	Gravity anomaly due to shallower bodies.
<i>Isostatic anomaly</i>	Isostasy corrected gravity anomaly.
<i>Interpolating or shape function</i>	The weighting factors for ground element covering the survey space for the computation of regional.

Introduction

The observed Bouguer gravity anomalies consist of two components – a regional component and a residual, that can be expressed by a simple relation:

$$\begin{aligned} \text{Observed Bouguer gravity field} \\ = \text{Regional gravity field} + \text{Residual gravity field} \end{aligned}$$

Commonly, the regional Bouguer gravity anomaly is the longer wavelength field due to deep sources, whereas the residual Bouguer gravity anomaly corresponds to short wavelength fields of shallower bodies. The short wavelength field dies off with depth.

However, in practice, the terminology of a regional gravity anomaly varies according to the target of the investigation and therefore does not necessarily refer to the fields generated at lower crustal depths. If the target is a mineral deposit located at the depths of 100–200 m, the basement rocks will form the regional structure. In case of hydrocarbon accumulation, the target sources are a few kilometers deep, and the regional field is generated by the rocks at the base of the sedimentary columns. The studies involving large-scale structures like subduction zones and large sedimentary basins, the regional field originates at lower crustal depths or still greater depths at Moho.

The regional and residual anomalies, therefore, are relative terms. These vary from problem to problem, and from place to place. The early workers like Griffin (1949), Elkins (1951), Nettleton (1954), Roy (1961), Steenland (1962), and Skeels (1967) viewed regional and residual from this view-point. For a regional gravity survey on a continental scale, the concept of isostasy is evoked (Jachens and Griscom 1985; Blakely 1995; Chapin 1996) to compute regional and isostatic anomalies.

In problems where the target bodies occur at shallower depths, the deeper regional field masks the response of the shallower targets and effectively acts as a noise. On the other hand, when the target body lies at greater depths, the response of shallower bodies becomes noise.

The above discussion points to the fact that it is important to separate regional and residual gravity anomalies. In effect, regional and residual gravity separation is a filtering process, and it enhances the noise-to-signal ratio. This step is essential for a precise quantitative interpretation.

It may be noted that the gravity method was first used in Geodesy. The geodesists assumed the Earth to be an ellipsoid for the processing of the Bouguer gravity data. Later on, the geophysicists used geoid as the reference surface for processing of Bouguer gravity data. Therefore, Bouguer anomaly used by geophysicists is different from that used by geodesists. Of late, there is an attempt to define gravity anomaly that

can be used by both groups (Hackney and Featherstone 2003; Hinze et al. 2005). The processing techniques are valid for both the datums. Here, we are dealing with Bouguer gravity anomaly from the view point of the geophysicists.

Regional-Residual Separation

The regional-residual separation is essentially a filtering process and is carried out broadly by two techniques: (i) graphical and (ii) analytical. The graphical method is not used these days anymore. It has only historical importance. In this method, the regional is obtained by visually smoothing the anomaly curve. The success of this approach depends on the experience of interpreter and his knowledge about the geology of the area.

Analytical Methods

The analytical techniques are developed both in space- and frequency-domain. Starting from simple four-point and nine-point averaging, Griffin's center-point-and-one-ring (1949), Agocs' least-squares (1951), Grant's statistical (1954) to polynomial fitting (Coons et al. 1967; Rao et al. 1975; El-Batroukh and Zentani 1980; Belthrao et al. 1991; Agarwal and Shivaji 1992; Leao et al. 1996), upward and downward continuations, and second derivative techniques are in space-domain. Mickus et al. (1991) has developed an algorithm using the minimum curvature technique to separate the regional and residual components.

Another approach is to build a density model of the deep-seated structures with parameters obtained by independent data and methods like seismics, aeromagnetism, fractals, spectrals, etc. and to compute their gravity effects. The field generated by these deeper sources is subtracted from the observed Bouguer gravity field to obtain the residual gravity anomalies. This approach developed at USGS employs the theory of isostasy and is frequently used to compute residual isostatic anomalies (Jachens and Griscom 1985; Blakely 1995; Lowrie 1997). A new deterministic method based on Airy-Heiskanen isostatic model had been developed by Chapin (1996). This technique assumes that the topography is fractal. The algorithm used in this technique inverts the present-day elevations to determine the depth to the base of the crust. This depth is then used as input to a 3-D gravity model that computes the gravity response (regional) of the crust. The regional term is then subtracted from the Bouguer gravity to yield the final isostatic residual.

In the frequency domain, high-pass, low-pass, and band-pass filtering techniques (Naidu 1968; Ulrich 1968; Cordell

1985; Pawlowski 1994) have been developed for regional and residual resolutions. Green's concept of an equivalent-source layer is invoked by Pawlowski (1994) to construct a data adaptive, zero-phase, Wiener band-pass filter for regional–residual separation (Pawlowski 1994), wherein the phase shift of the filtered signal is nullified by using a pair of cascaded low-pass and high-pass filters such that the phase lag caused by the low-pass filter is compensated by the phase lead introduced by the high-pass filter. The observed gravity field's Fourier power spectrum is modeled with two Green's equivalent source layers, one equivalent layer for the shallower geological sources (residual field) and another for the deeper geological features (regional field).

Possible Drawbacks in Regional Computation

In all the above-mentioned analytical techniques, the Bouguer gravity values over the anomalous zone(s) enter into the computations. Many investigators (Skeels 1967; ZurFluch 1967; Ulrich 1968; Leao et al. 1996; Ojo and Kangkolo 1997) have discovered that these techniques are not free from inaccuracies. In Griffin's (1949) center-point-and-one-ring method, the center point of the ring, which represents the regional field, moves over the entire survey space, including the anomalous zone. When the gravity values are picked up from one such anomalous zone for averaging, a small residual still remains, that is mixed with the regional anomaly. Besides this, the radius optimum choice is crucial to obtain the proper residual. Leao et al. (1996) pointed out that “any attempt to model a complex regional anomaly by a higher-order polynomial will force the polynomial to represent part of the residual field, thereby distorting the estimated regional field.” In case of the two most used techniques, both least-squares (Agocs 1951; Coons et al. 1967; El-batroukh and Zentani 1980; Leao et al. 1996) and spectral methods (Naidu 1968; Ulrich 1968; Cordell 1985; Klasner et al. 1985; Pawlowski 1994), the residual field is assumed to be zero-mean. Ulrich (1968) and Leao et al. (1996) observed that a departure from this assumption may contaminate the residual field through pseudo anomalies.

New Approach Based on Finite Element Analysis

A new space-domain operator, based on finite element concept (FEM), has recently been developed by Mallick and Sharma (1997, 1999) and applied to a variety of research problems (Sharma et al. 1999; Mallick et al. 1999; Mallick and Vasanthi 2001; Vasanthi and Mallick 2001, 2005; Kannan and Mallick 2003; Vasanthi and Sharma 2006). In addition to

these original studies, many other investigators have gained benefitted from this new technique (Kaftan et al. 2005; Agarwal and Srivastava 2010; Ekinci 2010; Ekinci and Yigitbaş 2012, 2015; Bruno et al. 2017; Kumar and Puneekar 2018).

This method is altogether different from all existing techniques in the sense that the observed gravity values over the anomalous zones in the survey space do not contribute to the regional computations. This technique is robust, well-performing, and not site-specific and overcomes most of the drawbacks encountered in the existing analytical methods.

The procedure to compute the regional field by FEM is as follows.

Given a gravity survey, a rectangular or square element is superimposed on it in order to have the maximum coverage that includes all zone(s) of interest. The ground elements for the computation of the regional field can take different sizes, shapes, and orientations. Figure 1 shows (a) four-node linear, (b) eight-node quadratic, (c) nine-node elements, corresponding to interpolating functions (or element shape functions) which are respectively linear, quadratic, and cubic. Figure 2a, b show a few higher order elements, such as (a) 12-node and (b) 16-node. Additionally, a very useful seven-node element is shown in Fig. 2c. It is an irregular element used conveniently to avoid any anomalous zone falling on any side of the ground element. In the case of irregular element, the interpolating function is specially defined in order to distribute the weights among seven nodes.

As in the finite element analysis (Cheng and Yeo 1979; Zienkiewicz 1979; Cook 1984), the field variable is assumed to be a weighted sum of discrete values at the vertices or nodal points of the element. The regional gravity field in the present case is defined in a similar manner and is expressed as:

$$g(x, y) = \sum_{i=1}^r N_i(x, y) g_i \quad (1)$$

where $N_i(x, y)$ are the shape functions, acting as weighting factors, r is the number of nodes in the element, and g_i are the observed gravity values at the nodes of the element. In the present case, $g(x, y)$ represents the regional field. The element nodes lie away from the anomalous zone(s) present in the survey space. The interpolating functions at different nodes are shown later at the end of this section.

The procedure to compute the regional field is explained with the help of an eight-node quadratic element.

Figure 3a represents an eight-node element in the x – y real space, i.e., the gravity survey space. The center is located at x_c, y_c with sides $2a$ and $2b$ are parallel to x - and y -axes, respectively.

Gravity Data, Regional-Residual Separation,

Fig. 1 Various types of elements: (a) linear, (b) quadratic, and (c) nine-node. Solid circles are nodes of the element

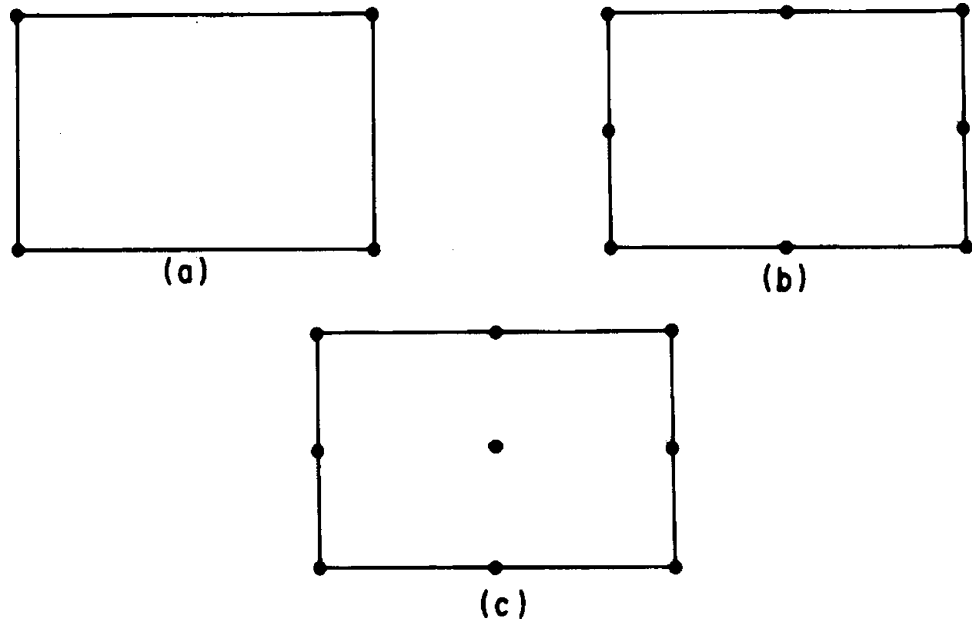
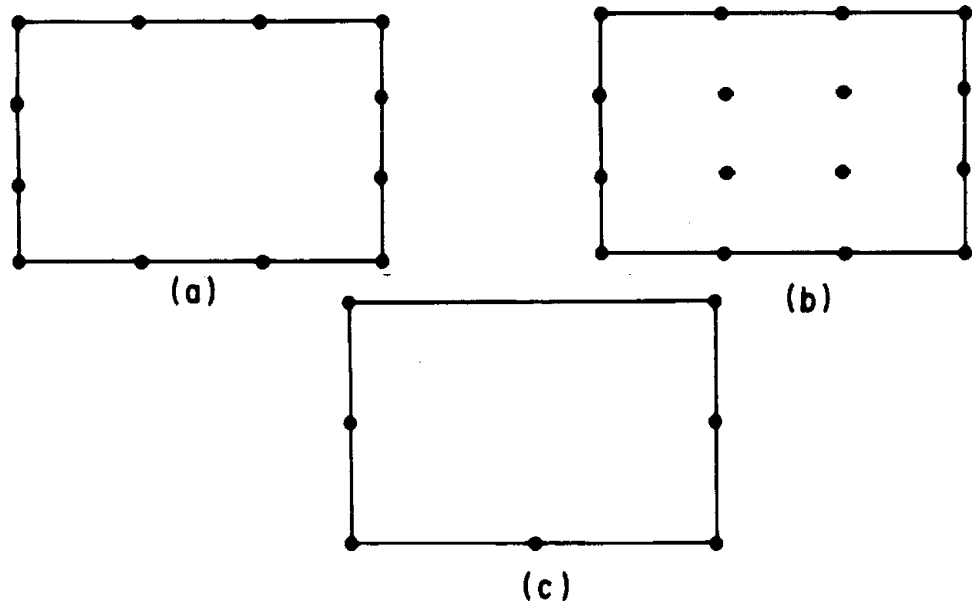
**Gravity Data, Regional-Residual Separation,**

Fig. 2 Various types of elements: (a) cubic, (b) sixteen-node, (c) seven-node. Solid circles are nodes of the element

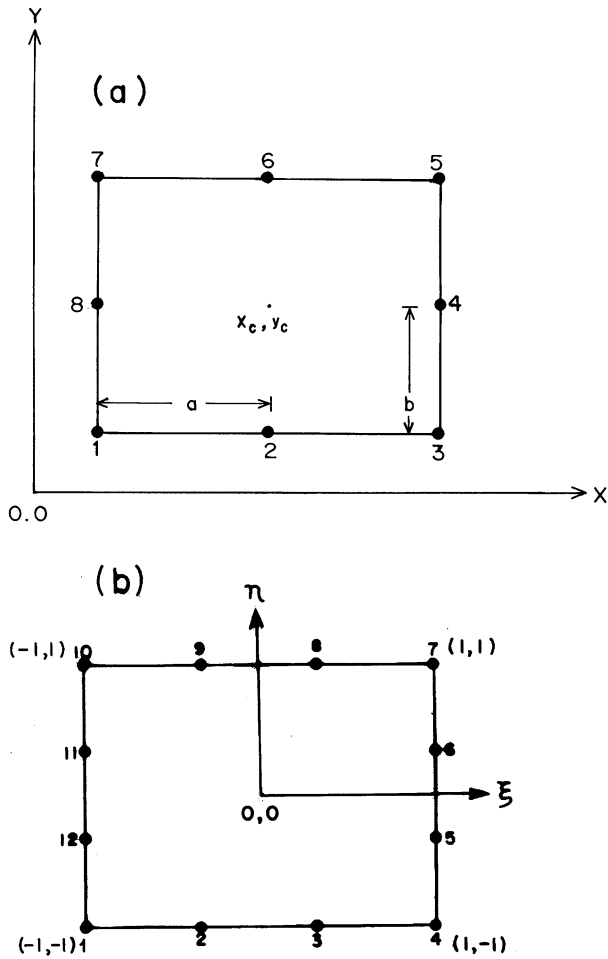


The ground element in the survey space may be of any size and may have different orientation. Therefore, there is a need to standardize the element by substituting it by a non-dimensional element. This is achieved by the following simple substitutions:

$$\xi = \frac{(x - x_c)}{a} \quad \text{and} \quad \eta = \frac{(y - y_c)}{b} \quad (2)$$

These nondimensional coordinates ξ and η vary between -1 and $+1$. In other words, the ground element of any size and orientation is replaced by ξ - η reference space as shown in Fig. 3b, the reference space is a square element. The nodes 1 through 8 in the ξ - η reference element and the nodes of the element in x - y real space have one-to-one correspondence.

This transformation (i) simplifies the computational scheme, (ii) defines the weights, without ambiguity, in terms



Gravity Data, Regional-Residual Separation, Fig. 3 (a) Eight-node element in x - y real space. (b) Eight-node element in ξ - η reference space

of shape functions, and (iii) facilitates to encompass survey space of any size and of any orientation.

With the substitutions suggested in Eq. 2, the regional gravity $g_{\text{reg}}(\xi, \eta)$ at any point (ξ, η) in the reference element can be expressed as:

$$g_{\text{reg}}(\xi, \eta) = \sum_{i=1}^r N_i(\xi, \eta) g_i, \quad (3)$$

where the weights, $N_i(\xi, \eta)$, are the shape functions of the element and r is the number of nodes. For example, in case of the 8-node quadratic element, r is equal to 8 and the shape functions are given by:

$$N_i(\xi, \eta) = \frac{(1 + \xi \xi_i)(1 + \eta \eta_i)(\xi \xi_i + \eta \eta_i - 1)}{4} \quad (4)$$

(for corner nodes, $i = 1, 3, 5, 7$)

$$N_i(\xi, \eta) = \frac{(1 - \xi^2)(1 + \eta \eta_i)}{2} \quad (5)$$

(for mid-side nodes, $i = 2, 6$)

$$N_i(\xi, \eta) = \frac{(1 - \eta^2)(1 + \xi \xi_i)}{2} \quad (6)$$

(for mid-side nodes, $i = 4, 8$)

where ξ_i and η_i are the nodal coordinates. $N_i(\xi, \eta) = 1$ at i^{th} node and zero elsewhere and $\sum N_i(\xi, \eta) = 1$. Figure 4 shows the interpolating function for corner and midpoint nodes.

The regional anomalies computed by Equation 3 in the reference element need to be translated to the real survey space. This is carried out as follows:

$$x(\xi, \eta) = \sum M_i(\xi, \eta) x_i \quad (7)$$

$$y(\xi, \eta) = \sum M_i(\xi, \eta) y_i \quad (8)$$

In case of isoparametric element used here, $M_i(\xi, \eta) = N_i(\xi, \eta)$, x_i and y_i are the nodal coordinates shown in Fig. 3a.

A similar methodology holds good for cubic (12 node) as well as irregular (missed node) elements. The appropriate shape functions need to be chosen in each case.

This technique overcomes many of the shortcomings encountered in polynomial fitting, spectral approach, and direct model-based regional computations. Some of these advantages are as follows:

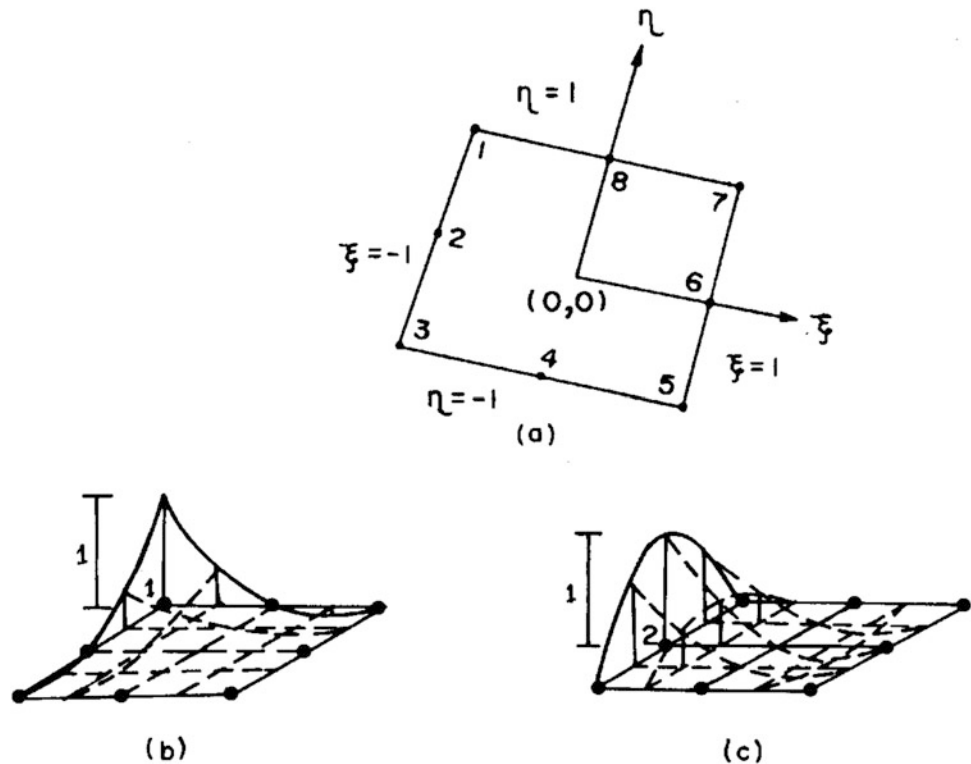
- (i) The observed gravity values only at discrete points, usually the vertices and midpoints of an element superimposing the gravity map, are required for regional computation. The anomalous zones are kept away from the nodal positions.
- (ii) For a quadrilateral ground element in x - y space of any size, shape, and any orientation, the corresponding reference element in ξ - η space is always a square. Therefore, the weighting coefficients needed for the regional computation remain unchanged. The strength of this new technique lies in this property.
- (iii) Only a second-order, or at most a third-order element, approximates the regional field, thereby eliminating the inaccuracies likely to be introduced by fitting very high-order polynomials.
- (iv) The computation is independent of a priori assumptions on the density, form, and depth of shallow or deeper structures.
- (v) The FEM approach is not site-specific, making its application universal.

Accuracy

After a brief description of the finite element concept, the next step consists of testing the accuracy of the computing

Gravity Data, Regional-Residual Separation, Fig. 4

(a) Eight-node quadratic isoparametric element, the shape function, (b) N1 (ξ, η) at a corner node, and (c) N2 (ξ, η) at a mid-side node



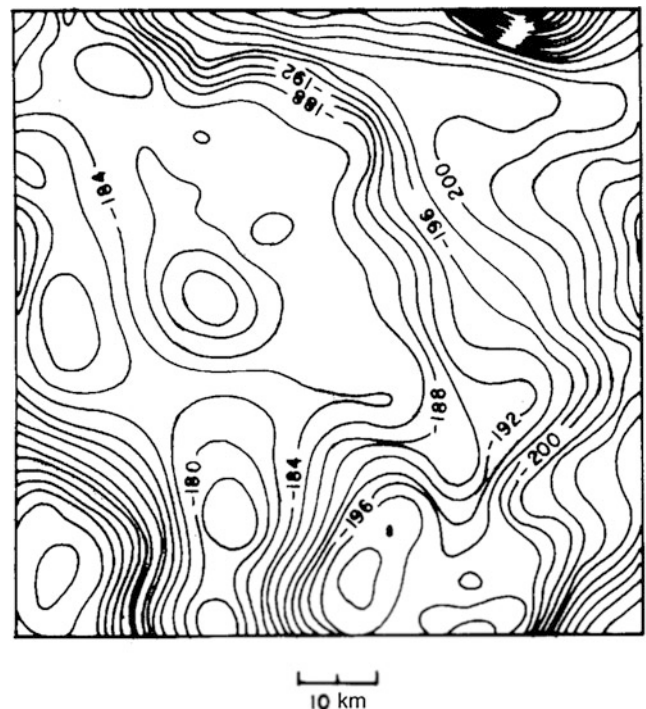
technique by comparing the results with those obtained (i) by spectral method for the Paradox Basin (Pawlowski 1994) and (ii) in the space domain for the Klamath Mountains and Cascade Range (Blakely 1995).

Paradox Basin, Utah (Frequency-Domain)

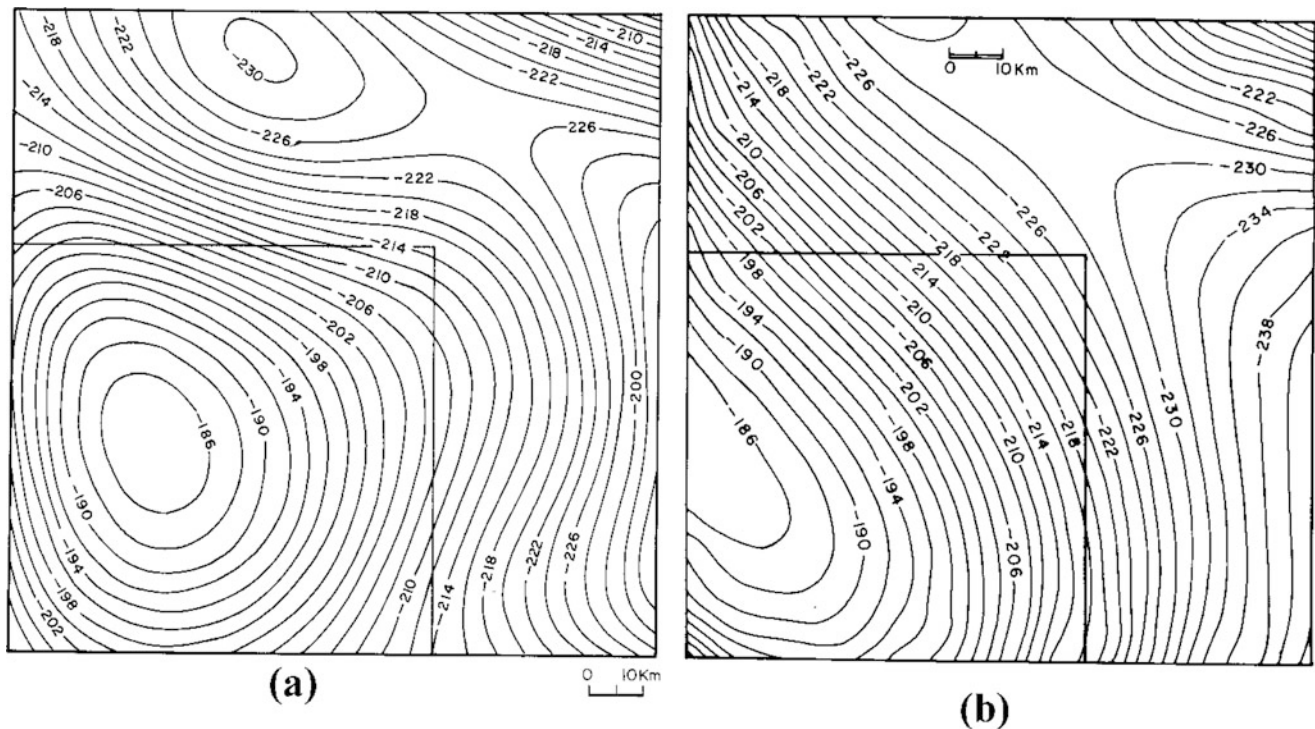
The case study of Pawlowski (1994) for the Paradox Basin, Utah, USA, is considered here because of its elegant approach using a Green's equivalent layer concept to develop a band-pass filter. Secondly, Pawlowski has compared his residual map with that obtained in an earlier study (Hildenbrand and Kucks 1983). In fact, it consists of comparing the FEM regional and residual anomalies with those derived by two independent techniques.

Figure 5 shows a 110 km $\times$ 110 km region of the Bouguer gravity map. Geologically, the Paradox Basin has a large number of NW-SE trending salt domes with hydrocarbon potential. This basin has been studied extensively by different geophysical methods, especially by the gravity method.

Based on the radially averaged Fourier spectrum of the observed gravity field, Pawlowski (1994) related the residual and regional anomalies to two equivalent layers, respectively at 5 km and 54 km depths, and has designed the band-pass filter. By applying this band-pass filter to the observed gravity data, he derived the regional and residual components.



Gravity Data, Regional-Residual Separation, Fig. 5 Bouguer gravity map of Paradox basin, Utah (Pawlowski 1994). Contour interval: 2 mGal



Gravity Data, Regional-Residual Separation, Fig. 6 (a) The regional anomaly obtained by band-pass filtering based on Green's equivalent layer (Pawłowski 1994). Two contour closures present in the map. Contour interval: 2 mGal. (b) The regional gravity anomaly

obtained by FEM approach (12 node). The regional contours trend NW-SE identical with the trends of the deep-seated salt domes. Contour interval: 2 mGal

The regional gravity field obtained by this band-pass filter method is illustrated in Fig. 6a. The regional gravity map shows two contour closures, one on north, and the more prominent one on the southwest corner. These closed contours may indicate the presence of a small portion of the shallower effects. Next, superimposing the 12-node element (nodes are shown in Fig. 2a), the regional component is computed by FEM approach. The resulting regional gravity anomaly is shown in Fig. 6b.

The new regional gravity map does not show any contour closures. The smooth regional contours show NW-SE anomalies identical with the trends of the deep-seated salt domes. In short, FEM regional reflects the deeper structures more faithfully. FEM approach therefore appears to perform better than the band-pass filter technique of Pawłowski (1994).

Residual Gravity Anomalies

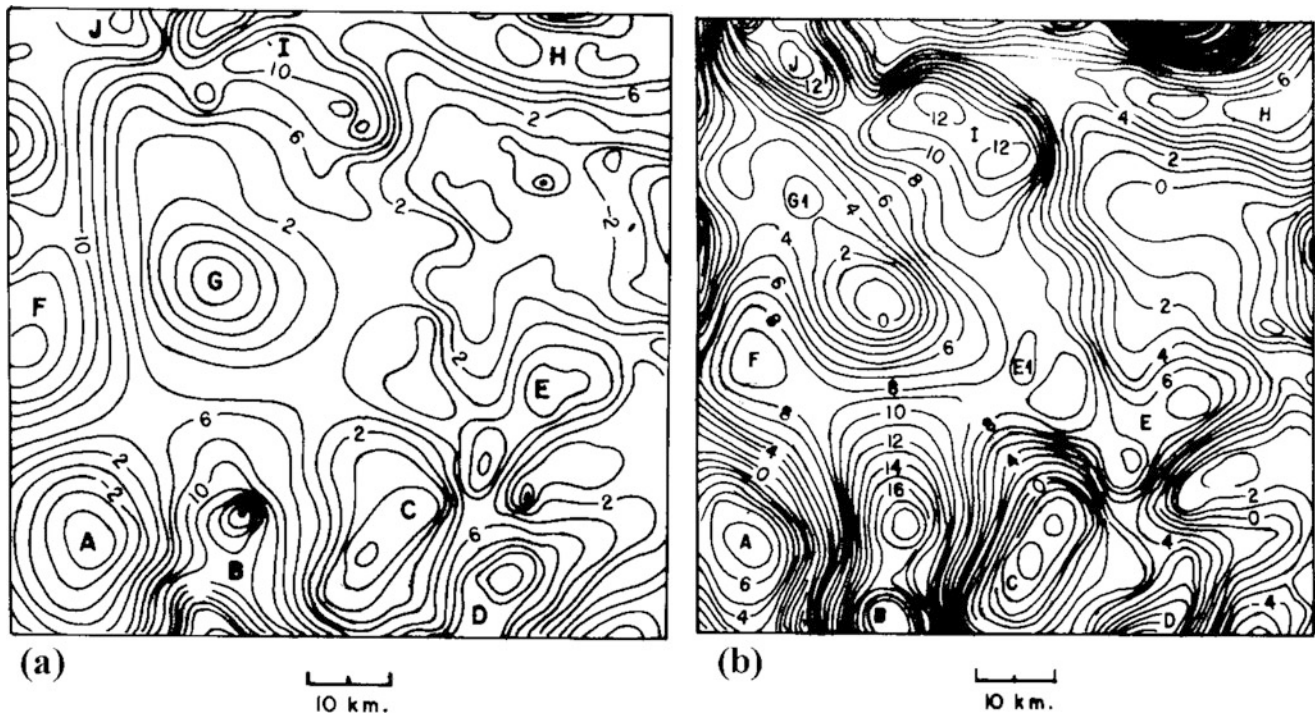
In order to get a closer look at the performance of FEM approach to compute the residual anomaly, a 75 km × 75 km window has been taken on the SW corner (see Fig. 6a, b) of the area. Figure 7a, b represent the band-passed and FEM residuals, respectively. A detailed comparison has been made by marking ten anomalous zones A to J in both maps. In each case, the FEM residual appears to be either better or comparable with the corresponding to the residual obtained by spectral technique.

Klamath Mountains and Cascade Range (Model-Based Space-Domain)

To test the accuracy in the space-domain approach, a gravity survey covering parts of Klamath Mountains and Cascade Range in north-central California lying between lat. 40° 52' 30" and 42° N, and long. 121° and 123° W has been considered. In an elegant analysis, Jachens and Griscom (1985) carried out regional-residual separation based on the theory of isostasy and provided an isostatic residual gravity map of this region. Figure 8 shows the isostatic model of Airy-Heiskanen with the following assumptions to compute the regional:

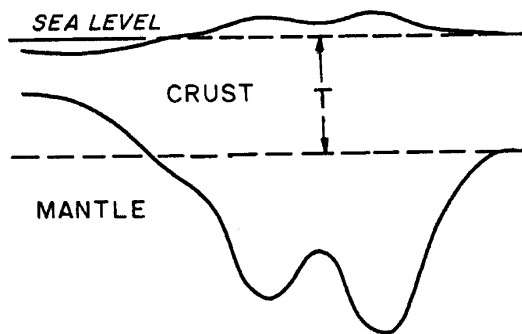
1. A three-layer crust with uniform densities
2. The density of the top layer, $\rho = 2.67 \text{ g/cm}^3$
3. The crustal thickness at sea-level, $T = 25 \text{ km}$, according to a seismic refraction survey
4. The density contrast of the lower layer, $\rho = 0.4 \text{ g/cm}^3$, close to the density contrast, $\rho = 0.385 \text{ g/cm}^3$ (Woollard 1966) based on extensive seismic studies.

In order to account for any deviation in density and crustal thickness from the above parameters, Jachens and Griscom (1985) have taken an upper bound and a lower bound of the regional gravity anomaly by considering two additional sets



Gravity Data, Regional-Residual Separation, Fig. 7 (a) The residual gravity map (Pawlowski 1994) for 75 km $\times$ 75 km survey space. The anomalous zones are shown from A to J. Zones F, I, and J are not well resolved here. Contour interval: 2 mGal. (b) The residual gravity

anomaly obtained by FEM approach (12 node) for 75 km $\times$ 75 km survey space. The anomalous zones from A to J have also well brought out here. The zones F, I, and J are well resolved and two additional zones, E1 and G1 are observed. Contour interval: 1 mGal



Gravity Data, Regional-Residual Separation, Fig. 8 Isostatic model of north-central California. (After Jachens and Griscom 1985)

of parameters: $\rho = 2.50, 2.85 \text{ g/cm}^3$; $T = 20, 30 \text{ km}$; and $\rho = 0.2, 0.6 \text{ g/cm}^3$. Figure 9 shows an east-west regional gravity profile along 41° N. The response of the standard model with $\rho = 2.67 \text{ g/cm}^3$, $T = 25 \text{ km}$, and $\rho = 0.4 \text{ g/cm}^3$ lies between upper and lower bounds. Additionally, Jachens and Griscom examined the regional anomaly for this region based on the Pratt-Hayford model response with a depth of compensation of 113.7 km. These two regional gravity anomalies did not differ appreciably from each other.

In turn, we have computed the regional of the same region by finite element approximation wherein no such assumptions

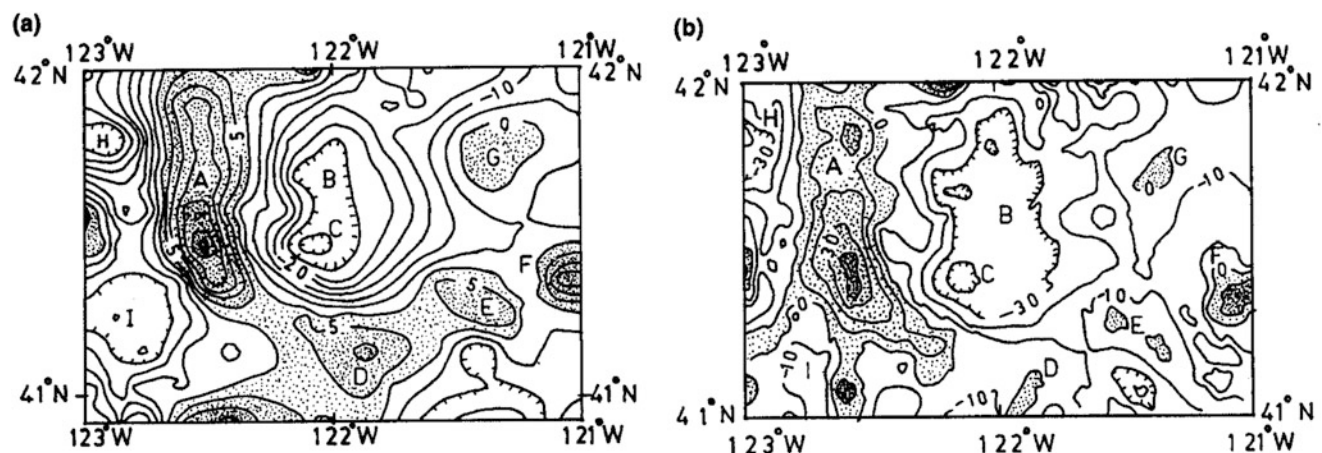
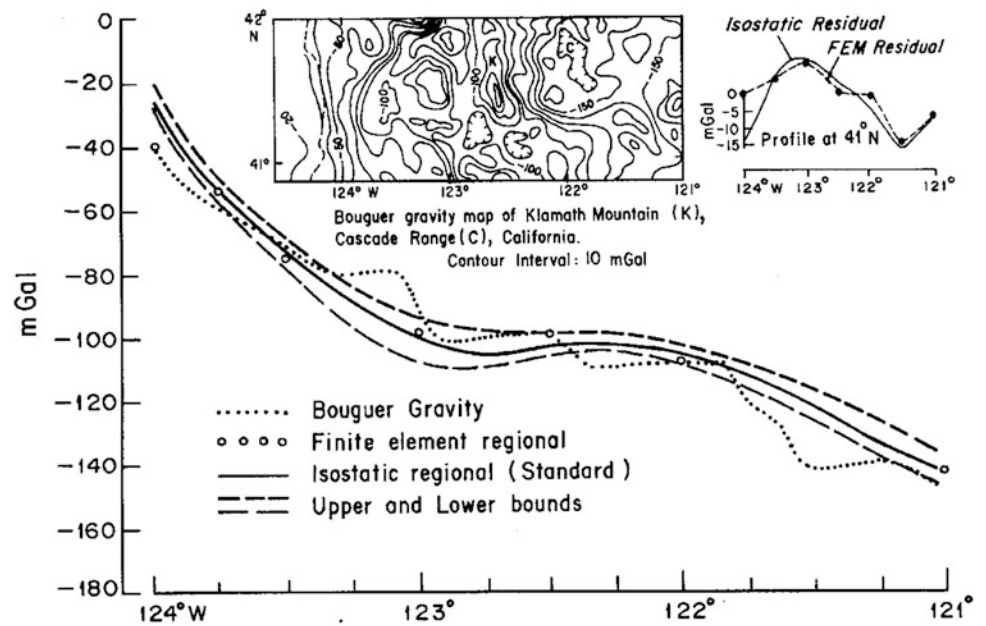
regarding *layer densities, crustal thickness, density contrast* are required. The open circles lying between the upper and lower bounds in Fig. 9 represent the FEA regional. Its agreement with the standard regional field is good. The deviations at some places between these two regionals can possibly be attributed to the hypothesis of uniform density over the entire region by Jachens and Griscom (1985).

The model-based (Blakely 1995, Fig. 7.21f) and FEA residual anomalies for Klamath Mountains and Cascade Range region are shown in Fig. 10a, b, respectively. The FEA anomalies shown from A to I appear to be better than or comparable to the corresponding model-based anomalies.

It is of interest to note that to arrive at the layer density ρ , crustal thickness (T), and density contrast ($\Delta\rho$), a considerable amount of geological information need to be collected, a large number of experiments, mostly precise (and expensive) seismic surveys, were mounted and a great deal of analysis made. In terms of time and financial input, these might be considerable. In contrast, simply starting from the Bouguer anomaly compares well with what Jachens and Griscom (1985) have very carefully computed. In many countries across the world, information about the layer density, crustal thickness, and density contrast is not available to precisely compute isostatic residual. In such areas, the FEA residual may provide a quick assessment of the gravity anomaly.

Gravity Data, Regional-Residual Separation,

Fig. 9 East-west regional gravity profile along lat. 41° N. The standard isostatic and FEA regional anomalies lie between upper and lower bound regionals of Jachens and Griscom (1985)



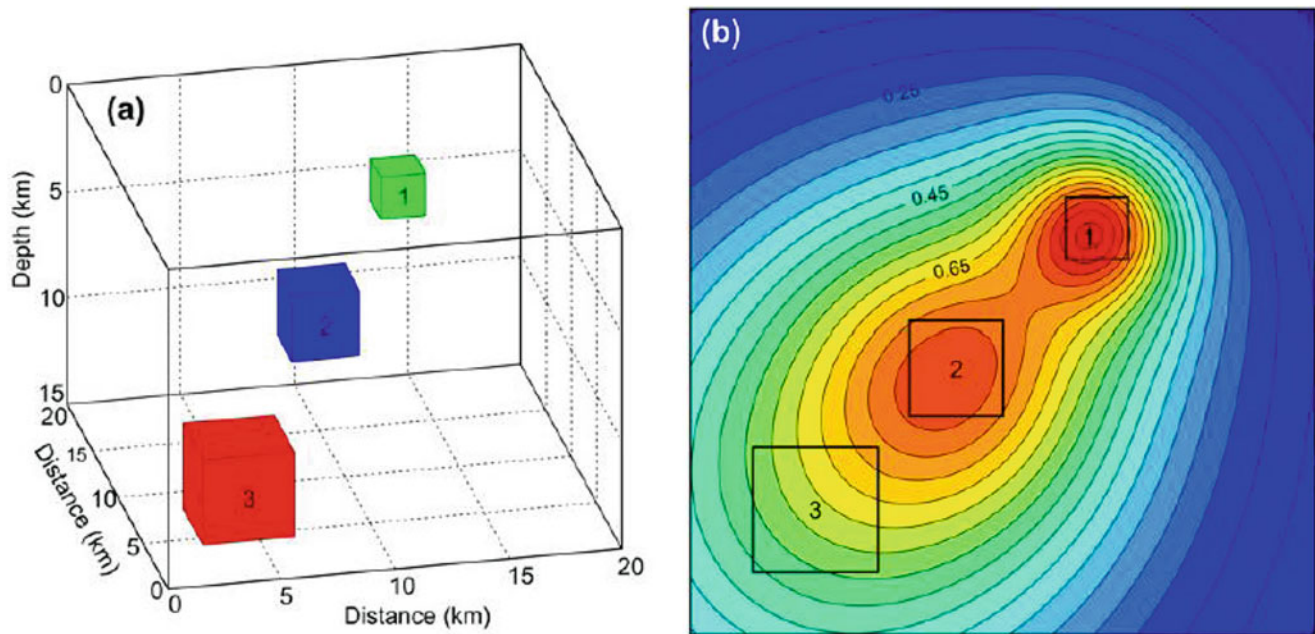
Gravity Data, Regional-Residual Separation, Fig. 10 (a) Finite element residual map of the Klamath Mountains and Cascade Range. Contour interval: 5 mGal. (b) Isostatic residual map of the Klamath

Mountains and Cascade Range. Contour interval: 10 mGal. (After Blakely 1995)

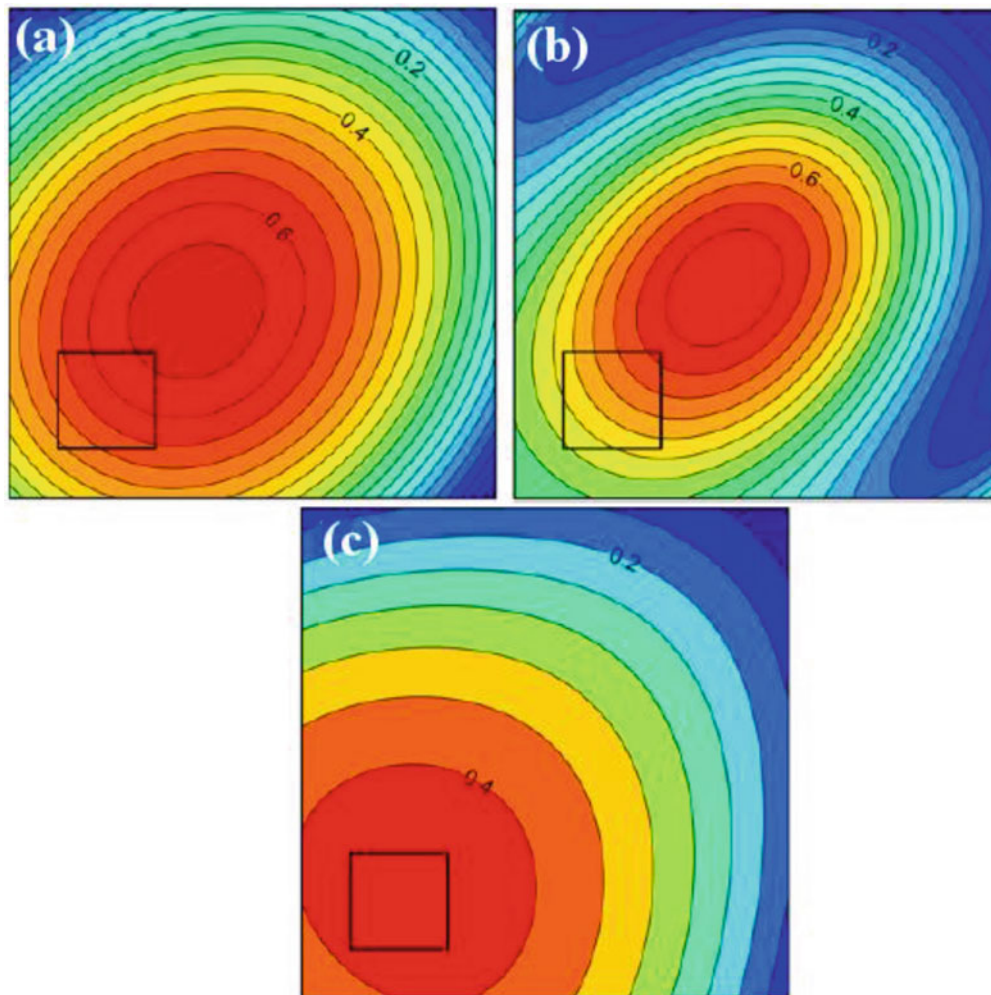
Recent Studies on FEM Applications

The success of finite element method has led many investigators to apply this technique to process the theoretical and field gravity data (Kaftan et al. 2005; Agarwal and Srivastava 2010; Ekinci 2010; Ekinci and Yiğitbaş 2012, 2015). Ekinci and Yiğitbaş (2015) have reported interesting results for synthetic simulation and field application. Figure 11 shows the gravity response of three vertical-sided prisms. First, two shallow bodies produce the residual gravity component, while the gravity response of body 3 was assumed to represent the regional effect. The regional anomalies for this model are derived by (a) second-order polynomial, (b) third-order

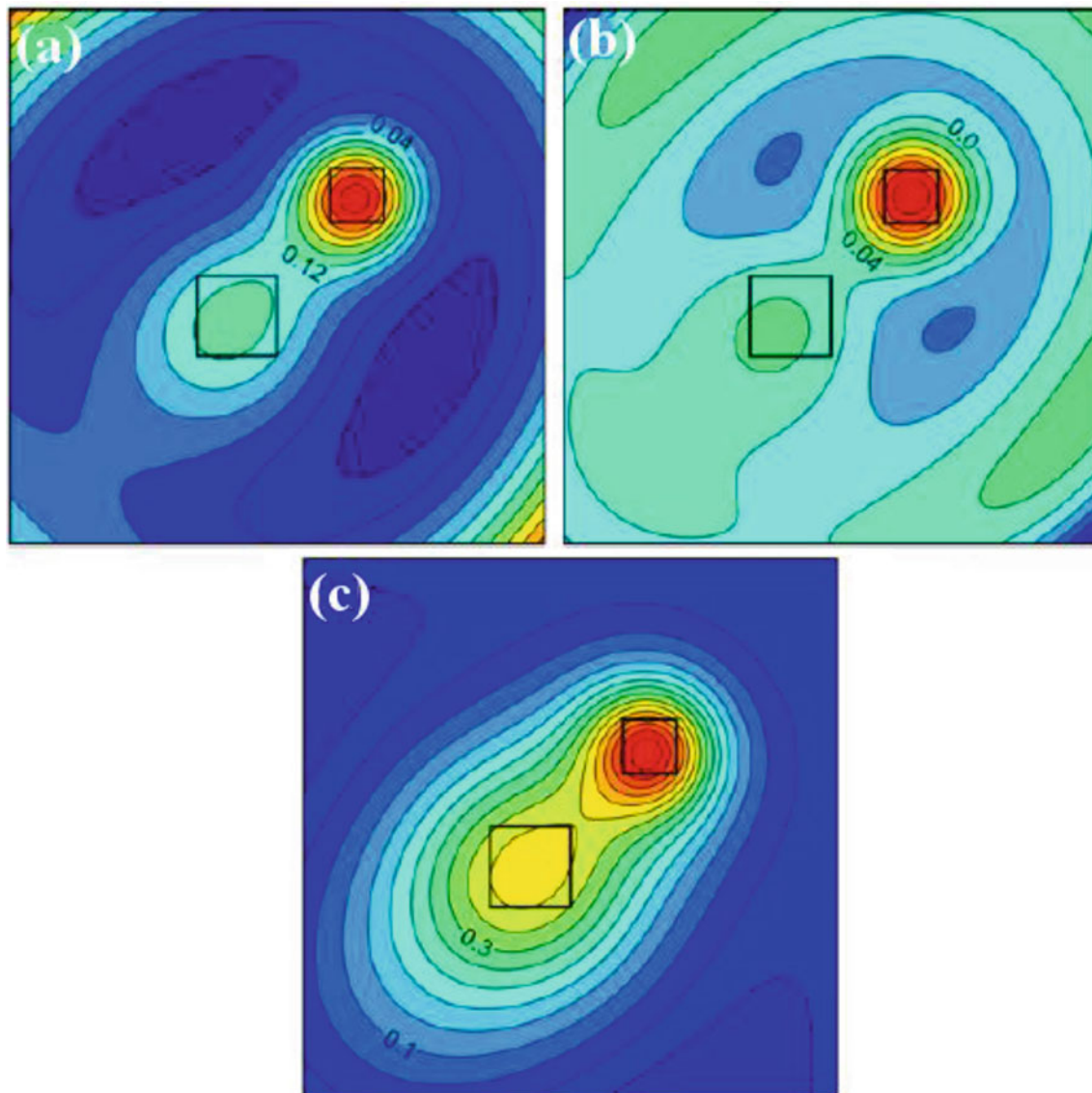
polynomial, and (c) FEM, and are shown in Fig. 12. It is seen from the figure that the gravity response of the deeper body (body 3) is interestingly enhanced by the use of FEM (Fig. 12a), whereas the maximum gravity amplitudes could not be located over body 3 by polynomial fitting procedure (Fig. 12b, c). Also, there are no pseudo anomalies present in the FEM regional, whereas they slightly contribute in second-order regional and are prominent in third-order polynomial regional. Figure 13 shows the corresponding residual gravity anomalies. The second- and third-order polynomial residuals (Fig. 13a, b) contain number of pseudo anomalies, making it unsuitable for further interpretation. In contrast, the FEM residual in Fig. 13c is far superior in terms of anomaly pattern



Gravity Data, Regional-Residual Separation, Fig. 11 Gravity response of three vertical-sided prisms (Ekinci and Yiğitbaş 2015)



Gravity Data, Regional-Residual Separation, Fig. 12 Regional gravity anomalies by (a) second-order (Ekinci and Yiğitbaş 2015), (b) third-order polynomial (Ekinci and Yiğitbaş 2015), and (c) FEM procedures (Ekinci and Yiğitbaş 2015)



Gravity Data, Regional-Residual Separation, Fig. 13 Residual gravity anomalies by (a) second-order (Ekinici and Yiğitbaş 2015), (b) third-order polynomial (Ekinici and Yiğitbaş 2015), and (c) FEM procedures (Ekinici and Yiğitbaş 2015)

and amplitude. Therefore, it can be inferred that the synthetic simulation method indicated the superiority of the FEM over the polynomial fitting procedure in computing the regional gravity component.

Summary

The separation of regional and residual components from the observed Bouguer gravity (and magnetic) anomaly has remained a subject of debate over the past six to seven decades. Starting from simple graphical to model-based estimation of regional – both in the space and frequency domains – the expected uniqueness has not been achieved. This is due to human judgment and mathematical as well as

geological assumptions. With this backdrop, a new concept based on finite element theory is described above, that performs as good as or often better than the existing techniques both in the frequency and space domain. This has been demonstrated with three examples. FEA technique is not site-specific and can be universally used.

Cross-References

- ▶ [Anthropogenic Seismicity Related to Exploitation of Georesources](#)
- ▶ [Earthquakes and Crustal Deformation](#)
- ▶ [Gravity, Data to Anomalies](#)
- ▶ [Magnetotellurics, Crustal Imaging](#)

- [Magnetovariation Studies](#)
- [Magnetic Anomalies: Interpretation](#)
- [Magnetic Anomaly Map, Global](#)

Bibliography

- Agarwal BNP, Shivaji CH (1992) Separation of regional and residual anomalies by least-squares orthogonal polynomial and relaxation techniques: a performance evaluation. *Geophys Prospect* 40:143
- Agarwal BNP, Srivatsava S (2010) A FORTRAN program to implement the method of finite elements to compute regional and residual anomalies from gravity data, *Computers & Geosciences* 36:848
- Agocs WB (1951) Least-squares residual anomaly determination. *Geophysics* 16:686
- Belthrao JF, Silva JBC, Costa JC (1991) Robust polynomial fitting method for regional gravity estimation. *Geophysics* 56:80
- Blakely RJ (1995) *Potential theory in gravity and magnetic applications*. Cambridge University Press, Cambridge
- Bruno BG, Emilson PL, Alvaro PC (2017) 3D gravimetric investigation of the Cerro do Jarau structure, Rio Grande do Sul, Brazil. *Meteorit Planet Sci* 52:565–583
- Chapin AD (1996) A deterministic approach toward isostatic gravity residuals – a case study from South America. *Geophysics* 61:1022
- Cheng YK, Yeo MF (1979) *A practical introduction to finite element analysis*. Pitman, London
- Cook RD (1984) *Concepts and applications of finite element analysis*. Wiley, New York
- Coons RL, Woollard GP, Hershey G (1967) Structural significance and analysis of Mid-Continent gravity high. *AAPG Bull* 51:2381
- Cordell L (1985) A stripping filter for potential data. In: 55th Annual international meeting, society of exploration geophysicists, extended abstracts, pp 217–218
- Ekinci YL (2010) Gravity anomaly separation using finite element method. In: 19th international geophysical congress and exhibition. Ankara, Turkey, No 59
- Ekinci YL, Yiğitbaş E (2012) A geophysical approach to the igneous rocks in the Biga Peninsula (NW Turkey) based on airborne magnetic anomalies: geological implications. *Geodin Acta* 25:267–285
- Ekinci YL, Yiğitbaş E (2015) Interpretation of gravity anomalies to delineate some structural features of Biga and Gelibolu peninsulas, and their surroundings (North-West Turkey). *Geodin Acta* 27(4):300–319
- El-batroukh SI, Zentani AS (1980) Gravity interpretation of Raguba field, Sirte Basin, Libya. *Geophysics* 45:1153
- Elkins TA (1951) The second derivative method of gravity interpretation. *Geophysics* 16:29
- Grant FS (1954) A theory for the regional correction of potential data. *Geophysics* 19:23
- Griffin WP (1949) Residual gravity in theory and practice. *Geophysics* 15:39
- Hackney RI, Featherstone WE (2003) Geodetic versus geophysical perspectives of the ‘gravity anomaly’. *Geophys J Int* 154:35
- Hildenbrand TG, Kucks RP (1983) Regional magnetic and gravity features of the Gibson Dome and surrounding region, Paradox basin, Utah. U.S. Geological Survey Open-File Report 83–359
- Hinze WJ, Aiken C, Brozena J, Coakley B, Dater D, Flanagan G, Forsberg R, Hildenbrand T, Kellar R, Kellogg J, Kucks R, Li X, Mainville A, Morin R, Pilkington M, Plouff D, Ravat D, Roman D, Jamie UF, Veronneau M, Webring M, Winster D (2005) New standards for reducing gravity data: the North American gravity database. *Geophysics* 70:J125
- Jachens RC, Griscom A (1985) An isostatic anomaly maps of California – a residual map for interpretation of anomalies from infrastructural sources. In: Hinze WJ (ed) *The utility of regional gravity and magnetic anomaly maps*. Society of Exploration Geophysicists, USA, pp 347–360
- Kaftan I, Salk M, Sari C (2005) Application of the finite element method to gravity data case study: Western Turkey. *J Geodyn* 39:431–443
- Kannan S, Mallick K (2003) Accurate regional – residual separation by finite element approach, Bouguer gravity of a Precambrian mineral prospect in northwestern Ontario. *First Break* 21:39
- Klasner JS, King ER, Jones WJ, (1985) *Geologic interpretation of gravity and magnetic data for northern Michigan and Wisconsin*. In: Hinze WJ (ed) *The utility of regional gravity and magnetic anomaly maps*. Society of Exploration Geophysicists, USA, pp 267–286
- Kumar MPR, Puneekar DV (2018) Significance of regional gravity survey in parts of Sidhi and Shadol districts. *J Indian Geophys Union* 22:478–484
- Leao JWD, Menezes PTL, Beltrao JF, Silva JBC (1996) Gravity inversion of basement relief constrained by the knowledge of depth at isolated points. *Geophysics* 61:1702
- Lowrie W (1997) *Fundamentals of geophysics*. Cambridge University Press, Cambridge
- Mallick K, Sharma KK (1997) Computation of regional gravity anomaly – a novel approach. *Proc Indian Acad Sci (Earth Planet Sci)* 106:55
- Mallick K, Sharma KK (1999) A finite element method for computation of the regional gravity anomaly. *Geophysics* 64:461
- Mallick K, Vasanthi A (2001) Delineation of subsurface extent of Perinthatta anorthosite, Kerala, India. *J Geol Soc India* 57:321
- Mallick K, Sharma KK, Rao VK (1999) A new gravity interpretation: a case study from Pahute Mesa, Nevada test site. *Curr Sci* 76:1495
- Mickus KL, Aiken CLV, Kennedy WD (1991) Regional – residual gravity anomaly separation using the minimum curvature technique. *Geophysics* 56:279
- Naidu PS (1968) Spectrum of the potential field due to randomly distributed source. *Geophysics* 33:337
- Nettleton LL (1954) Regionals, residuals and structures. *Geophysics* 19:1
- Ojo SB, Kangkolo R (1997) Shortcomings in the determination of regional fields by polynomial fitting: a simple solution. *J Appl Geophys* 36:205
- Pawlowski RS (1994) Green’s equivalent-layer concept in gravity band-pass filter design. *Geophysics* 59:69
- Rao BSR, Murthy IVR, Rao VC (1975) A successive approximation method of deriving residual gravity. *Geoexploration* 13:129
- Roy A (1961) On some properties of residual and derivatives. *J Geophys Res* 66:543
- Sharma KK, Rao VK, Mallick K (1999) Finite element gravity regional and residual anomalies and structural fabrics of northwest Ganga basin. *J Geol Soc India* 54:169
- Skeels DC (1967) What is residual gravity? *Geophysics* 32:872
- Steenland NC (1962) Gravity and aeromagnetic exploration in the Paradox basin. *Geophysics* 27:73
- Ulrich TJ (1968) Effect of wavelength filtering on the shape of the residual anomaly. *Geophysics* 33:1015
- Vasanthi A, Mallick K (2001) Occurrence patterns of kimberlite pipes based on gravity and magnetic anomalies in Wajrakarur region, AP. *J Geol Soc India* 58:251
- Vasanthi A, Mallick K (2005) Bouguer gravity anomalies and occurrence patterns of Kimberlite pipes in Narayanpet-Maddur regions, Andhra Pradesh, India. *Geophysics* 70:13
- Vasanthi A, Sharma KK (2006) Discussion article on “New standards for reducing gravity data: The North American gravity database” (W. J. Hinze et al. 2005, *Geophysics*, v. 70, J25–J32), *Geophysics*, 71: X31
- Woollard GP (1966) Regional isostatic relations in the United States. In: Steinhart JS, Smith TJ (eds) *The earth beneath the continents*. *Geophysics monograph*, American Geophys. Union, USA, vol 10, p 557
- Zienkiewicz O (1979) *The finite element method*. McGraw-Hill, London, 618p
- Zurfluh EG (1967) Application of two-dimensional linear wavelength filtering. *Geophysics* 32:1015

Gravity Field of the Earth

Christopher Jekeli

Division of Geodetic Science, School of Earth Sciences, Ohio State University, Columbus, OH, USA

Definition and Scope

The gravity field of the Earth is generally understood as the field generated by the masses of the terrestrial body. Actual measurements made on the surface include tidal components due to the sun and moon (and theoretically other planets) and the atmosphere, as well as the centrifugal acceleration due to Earth's rotation. Therefore, geodesists distinguish between terrestrial gravitation (mass attraction of the Earth) and gravity (gravitation plus centrifugal acceleration), where tidal and atmospheric effects are treated as corrections. This entry develops the basic physical concepts of terrestrial gravitation assumed to be generated solely by the subsurface mass density distribution. After an introduction to the Newtonian gravitational potential and its properties, the subsequent sections of the entry discuss the spatial variation of the field, the geophysical interpretation of the low-degree harmonics of the field, and briefly the variation of the field inside the Earth.

Introduction

Historical Notes

A study of Earth's gravitational field is a study of Earth's mass structure, its dynamical influence on satellites, and lately its mass redistribution and transport in time. It is also fundamentally and historically a geodetic study of Earth's shape, described largely (70%) by the surface of the oceans (and their conceptual continuation under the continents). Gravitation is a physical phenomenon so pervasive and incidental that humankind generally has taken it for granted with scarcely a second thought. The Greek philosopher Aristotle (384–322 BC) allowed no more than to assert that gravitation is a natural property of material things that causes them to fall (or rise, in the case of some gases), and the more material, the greater the tendency to do so. It was enough of a self-evident explanation that it was not yet to receive the scrutiny of the scientific method. Almost 2000 years later, Galileo Galilei (1564–1642) finally took up the challenge to understand gravitation through observation and scientific investigation. His experimentally derived law of falling bodies corrected the Aristotelian view and divorced the effect of gravitation from the mass of the falling object – all bodies fall with the same acceleration. This truly monumental contribution to physics was, however, only a local explanation of how bodies behaved under gravitational influence. Johannes Kepler's

(1571–1630) observations of planetary orbits pointed to other types of laws, principally an inverse-square law according to which bodies are attracted by forces that vary with the inverse square of distance. The genius of Isaac Newton (1642–1727) brought it all together in his *Philosophiae Naturalis Principia Mathematica* of 1687 with a single and simple all-embracing law that in one bold stroke explained the dynamics of the entire universe (today there is more to understanding the dynamics of the cosmos, but Newton's law remarkably holds its own). The mass of a body was again an essential aspect, not as a self-attribute as Aristotle had implied but as the source of attraction for other bodies: each material body attracts every other material body according to a very specific rule (Newton's law of gravitation). Newton regretted that he could not explain exactly why mass has this property (as one still yearns to know today within the standard models of particle and quantum theories). Even Albert Einstein (1879–1955) in developing his theory of general relativity (i.e., the theory of gravitation) could only improve on Newton's theory by incorporating and explaining action at a distance (gravitational force acts with the speed of light as a fundamental tenet of the theory). Through the principle of equivalence of inertial and gravitational mass, Einstein explained gravitation as the effect that mass has on the fabric of space-time. But what actually mediates the gravitational effect (or, attraction) still intensely occupies modern physicists and cosmologists.

Gravitation since its early scientific formulation initially belonged to the domain of astronomers, at least as far as the observable universe was concerned. Theory successfully predicted the observed perturbations of planetary orbits and even the location of previously unknown new planets (Neptune's discovery in 1846 based on calculations motivated by observed perturbations in Uranus's orbit was a major triumph for Newton's law). However, it was also discovered that gravitational acceleration varies on Earth's surface, with respect to altitude and latitude. Newton's law of gravitation again provided the backdrop for the variations observed with pendulums. An early achievement for his theory came when he successfully predicted the polar flattening in Earth's shape on the basis of hydrostatic equilibrium, which was confirmed finally (after some controversy) with geodetic measurements of long triangulated arcs in 1736–1737 by Pierre de Maupertuis and Alexis Clairaut. Gravitation thus also played a dominant role in geodesy, the science of determining the size and shape of the Earth, promulgated in large part by the father of modern geodesy, Friedrich R. Helmert (1843–1917) (Helmert 1884).

Knowledge of terrestrial gravitation through the twentieth century was crucial for geodetic determination of a well-defined reference surface (a level surface) for heights and to correct astronomic observations of latitude and longitude for the irregular direction of gravitation (Heiskanen and Veing Meinesz 1958; Molodensky et al. 1962; Heiskanen and Moritz 1967; Torge 1989; Sansò and Rummel 1997; Sansò

and Sideris 2013). Also, geophysicists realized the potential of using gravity measurements to infer subsurface structures that could trap oil and gas deposits (e.g., Heiland 1940; Nettleton 1976; Telford et al. 1990). The Earth sciences increasingly rely on accurate gravimetry and gravitational models to aid in the interpretation of other sensor data. For example, oceanographers need an accurate level surface to infer surface currents from sea surface heights measured by satellite altimetry. Geophysicists analyzing crust deformation from GPS measurements of surface point coordinates must know whether the uplift or subsidence is accompanied by a change in subsurface mass – is it an elastic or a viscoelastic response? Ice sheet mapping (e.g., in Antarctica) using airborne laser altimetry must be accompanied with other sensors, such as gravity meters, that can infer the ice-bedrock interface. Hydrologists and glaciologists can now monitor the regional seasonal and secular changes in water/ice mass as seen by gravitation-mapping satellites as they continually orbit the Earth. In our solar system, planetary probes glean significant geophysical information about their targets from the gravitational effect on their orbits as tracked from Earth.

Other perhaps more utilitarian applications also rely on gravitational information. For example, longer-term precision navigation and guidance in the absence of external references, such as GPS, rely on inertial instruments. These autonomous navigation systems sense only inertial acceleration and require gravitational compensation, since a vehicle follows a path governed by both inertial and gravitational acceleration (Einstein's equivalence principle) (Jekeli 2000).

Although not as prominent nor perhaps as prestigious in modern scientific endeavors, gravimetry at all scales, from point measurements to global surveys, continues to hold a critical, at least desirable place in the quest to answer a plethora of macroscopic inquiries into Earth's systems. For a modern view of the role that knowledge of the Earth's gravity field plays in the Earth sciences, see Plag and Pearlman (2009) and references therein.

Newton's Law of Gravitation

In its original form, Newton's law of gravitation applies only to idealized point masses. It describes the force of attraction, $\mathbf{F}$, experienced by two such solitary masses as being proportional to the product of the masses, m_1 and m_2 ; inversely proportional to the square of the distance, ℓ , between them; and directed along the line joining them, as expressed by the unit vector, $\mathbf{n}$:

$$\mathbf{F} = G \frac{m_1 m_2}{\ell^2} \mathbf{n}. \quad (1)$$

G is Newton's gravitational constant that takes care of the units between the left- and right-hand sides of the equation; it can be determined by experiment and the current accepted value is (CODATA 2018 value, NIST 2019):

$$G = (6.67430 \pm 0.00015) \times 10^{-11} \text{ m}^3/(\text{kg s}^2). \quad (2)$$

The unit vector in Eq. (1) is directed from either point mass to the other, and thus the gravitational force is attractive and applies equally to one mass as the other. Newton's law of gravitation is universal as far as we know, requiring reformulation only in Einstein's more comprehensive theory of general relativity which describes gravitation as a characteristic curvature of the space-time continuum (Newton's formulation assumes instantaneous action and differs significantly from the general relativistic concept only when very large velocities or masses are involved; or, conversely, in applications of extremely high precision).

We can ascribe a gravitational acceleration to the gravitational force, which represents the acceleration that one mass undergoes due to the gravitational attraction of the other. Specifically, from the law of gravitation, we have (for point masses) the gravitational acceleration of m_1 due to the gravitational attraction of m_2 :

$$\mathbf{g} = G \frac{m_2}{\ell^2} \mathbf{n}. \quad (3)$$

The vector $\mathbf{g}$ is independent of the mass, m_1 , of the body being accelerated (which Galileo found by experiment).

By the law of superposition, the gravitational force, or the gravitational acceleration, due to many point masses is the vector sum of the forces or accelerations generated by the individual point masses. Manipulating vectors in this way is certainly feasible, but fortunately a more appropriate concept of gravitation as a scalar field simplifies the treatment of arbitrary mass distributions.

This more modern view of gravitation (adopted by eighteenth and nineteenth century mathematicians, such as Lagrange (1736–1813), Gauss (1777–1855), and Green (1793–1841)) holds that it is a *field* having a gravitational potential, V , defined in terms of the gravitational acceleration, $\mathbf{g}$, that a test particle would undergo in the field according to the equation:

$$\nabla V = \mathbf{g}, \quad (4)$$

where ∇ is the gradient operator (a vector operator). We use here the convention that the potential is always positive. Further elucidation of gravitation as a field grew from Einstein's attempt to incorporate gravitation into his theory of special relativity where no reference frame has particular significance above all others. It was necessary to consider that gravitational force is not a real force (i.e., it is not an applied force, like friction or propulsion) – rather it is known as a kinematic force, that is, one whose action is proportional to the mass on which it acts (like the centrifugal force); see (Martin 1988). Under this precept, the geometry of space is defined intrinsically by the gravitational fields contained therein. We continue with the classical Newtonian potential,

but interpret gravitation as an acceleration different from the acceleration induced by real, applied forces. This becomes especially important when considering the measurement of gravitation (Jekeli 2007).

The gravitational potential, V , is a *scalar* function, and it is derived directly on the basis of Newton's law of gravitation. To make it completely consistent with this law and thus declare it a Newtonian potential, we must impose the following conditions of regularity at infinity:

$$\lim_{\ell \rightarrow \infty} \ell V = Gm \quad \text{and} \quad \lim_{\ell \rightarrow \infty} V = 0, \quad (5)$$

where m is the attracting mass. It easily follows from Eqs. (3, 4, and 5) that the gravitational potential at any point in space due to a point mass is:

$$V(\mathbf{x}) = \frac{Gm(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|}, \quad (6)$$

where $\ell = |\mathbf{x} - \mathbf{x}'|$ is the distance from the mass point, $\mathbf{x}'$, to the evaluation point, $\mathbf{x}$, of the potential. Note that the field of a point mass has a singularity at $\mathbf{x} = \mathbf{x}'$.

Applying the law of superposition, the gravitational potential of many point masses is the sum of the potentials of the individual points. And, for infinitely many points in a closed, bounded region with infinitesimally small masses, dm , the summation changes to an integration:

$$V(\mathbf{x}) = G \int_{\text{mass}} \frac{dm(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|}. \quad (7)$$

Changing variables, $dm(\mathbf{x}') = \rho(\mathbf{x}')dv$, where ρ represents density (mass per volume) and dv is a volume element, we have:

$$V(\mathbf{x}) = G \iiint_v \frac{\rho(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} dv. \quad (8)$$

In spherical polar coordinates, the evaluation and integration points are (θ, λ, r) and (θ', λ', r') , respectively. The volume element in this case is given by $dv = r'^2 \sin \theta' d\lambda' d\theta' dr'$. It is readily shown that if the evaluation point, $\mathbf{x}$, is inside the volume, v , then by placing the coordinate origin at $\mathbf{x}$, one has $dv = |\mathbf{x}'|^2 \sin \theta' d\lambda' d\theta' dr'$, and the apparent singularity disappears – the integral is said to be weakly singular. Indeed, the potential of a volume of mass density is continuous and differentiable everywhere.

Suppose the density distribution over the volume depends only on radial distance (from the center of mass, assumed to be at the origin): $\rho = \rho(r')$, and that the evaluation is at an exterior point. The surface bounding the masses necessarily is

a sphere (say, of radius, R) and with an appropriate choice of coordinates, one easily shows (e.g., Jekeli 2007) that the integral (8) becomes:

$$V(\theta, \lambda, r) = \frac{GM}{r}, \quad r \geq R, \quad (9)$$

where M is the total mass bounded by the sphere. To a reasonably good approximation, Earth's density structure is spherically symmetric (see section “[Low-Degree Harmonics as Density Moments](#)”), and the external gravitational potential is that of a point mass (the approximation is valid to one part in 2000).

Besides volumetric mass (density) distributions, it is of interest to consider surface distributions. Imagine an infinitesimally thin layer of mass on a surface, s , where the units of density in this case are those of mass per area. Then, analogous to Eq. (8), the potential is:

$$V(\mathbf{x}) = G \iint_s \frac{\kappa(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} ds, \quad (10)$$

where κ is the surface mass density. In this case, V is a continuous function everywhere, but its first derivatives are discontinuous at the surface.

Now, imagine two infinitesimally close density layers (double layer or layer of mass dipoles), where the units of density are those of mass per area times length. The potential in this case is given by (see Heiskanen and Moritz 1967, p. 8):

$$V(\mathbf{x}) = G \iint_s \kappa(\mathbf{x}') \frac{\partial}{\partial n} \left(\frac{1}{|\mathbf{x} - \mathbf{x}'|} \right) ds, \quad (11)$$

where $\partial/\partial n$ is the directional derivative along the perpendicular to the surface. Now, V itself is discontinuous at the surface, as are all its derivatives. In all cases, V is a Newtonian potential, having been derived from the basic formula (6) for a point mass that follows from Newton's law of gravitation, Eq. (1).

The following properties of the gravitational potential are useful and foundational for many applications. First, consider Stokes's theorem, for a vector function, $\mathbf{f}$, defined on a surface, s :

$$\iint_s (\nabla \times \mathbf{f}) \cdot \mathbf{n} ds = \oint_p \mathbf{f} \cdot d\mathbf{r}, \quad (12)$$

where p is any closed path in the surface, $\mathbf{n}$ is the unit vector perpendicular to the surface, and $d\mathbf{r}$ is a differential displacement along the path. From Eq. (4), we find:

$$\nabla \times \mathbf{g} = \mathbf{0}, \quad (13)$$

since $\nabla \times \nabla = \mathbf{0}$, and hence, applying Stokes's theorem, we find with $\mathbf{f} = m\mathbf{g}$ that:

$$w = \oint_p m\mathbf{g} \cdot d\mathbf{r} = 0. \quad (14)$$

That is, the gravitational field is conservative: the work, w , expended in moving a mass around a closed path in this field vanishes. In contrast, dissipating forces (real forces!), like friction, expend work or energy, which shows again the special nature of the gravitational force.

It can be shown (Kellogg 1953, p.156) that the second partial derivatives of a Newtonian potential, V , satisfy the following differential equation, known as Poisson's equation:

$$\nabla^2 V = -4\pi G\rho, \quad (15)$$

where $\nabla^2 = \nabla \cdot \nabla$ formally is the scalar product of two gradient operators and is called the Laplacian operator. In Cartesian coordinates, it is given by:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \quad (16)$$

Equation (15) is a local characterization of the potential field, as opposed to the global characterization given by Eq. (8). Poisson's equation holds wherever the mass density, ρ , satisfies certain conditions similar to continuity (Hölder conditions, see Kellogg 1953, pp.152–153). A special case of Eq. (15) applies for those points where the density vanishes (i.e., in free space); then Poisson's equation turns into Laplace's equation:

$$\nabla^2 V = 0. \quad (17)$$

It is easily verified that the point mass potential satisfies Eq. (17); that is,

$$\nabla^2 \left(\frac{1}{|\mathbf{x} - \mathbf{x}'|} \right) = 0, \quad \mathbf{x} \neq \mathbf{x}'. \quad (18)$$

The solutions to Laplace's Eq. (17) are known as harmonic functions (here, we also impose the conditions (5) on the solution, if it is a Newtonian potential and if the mass-free region includes infinity). Hence, every Newtonian potential is a harmonic function in free space. The converse is also true: every harmonic function can be represented as a Newtonian potential of a mass distribution.

Several structures of the Earth's crust may be approximated by forms characterized essentially by a two-

dimensional cross-section, such as sheets, slabs, dikes, and faults, that extend indefinitely in one dimension. The gravitational potential associated with such density contrasts is known as a logarithmic potential, since the potential (up to an arbitrary constant) of an infinite, straight line (that is, a differential element of the cross-section) is given by (Telford et al. 1990; Kellogg 1953):

$$\delta V(\mathbf{x}) = 2G\mu \ln \frac{1}{r}, \quad (19)$$

where r is the distance of the evaluation point, $\mathbf{x}$, from the line and μ is the constant linear mass density of the line. If the line is indefinitely extended in the y -direction, then the potential of any form with cross A in the x - z plane is:

$$V(\mathbf{x}) = 2G\mu \iint_A \ln \frac{1}{\sqrt{(x-x')^2 + (z-z')^2}} dx' dz'. \quad (20)$$

Strictly speaking, the logarithmic potential is not a Newtonian potential since it does not vanish when $r \rightarrow \infty$ (on account of the arbitrary constant); however, it is a useful approximation in many geophysical applications.

Spatial and Temporal Variation Above the Surface

This section develops two types of solutions to standard boundary-value problems when the boundary is a sphere: the spherical harmonic series and an integral with a Green's function. Attention is given to the Dirichlet boundary-value problem for the exterior region (that includes infinity). That is, the gravitational potential, V , satisfies Laplace's Eq. (17) everywhere outside the sphere, and the solution to this partial differential equation is sought based on values of V given on the sphere. Other types of boundary-value problems presume given linear combinations of zero- and first-order derivatives of the potential on the boundary. Also, other boundaries that better approximate the Earth yield more difficult solutions – the ellipsoid is relatively easy (Hotine 1969; Jekeli 1988), whereas, the Earth's surface is challenging (Molodensky et al. 1962; Moritz 1980). With the spherical boundary, one obtains a particularly straightforward spectral representation that allows analysis of the spatial variation of the field at different wavelengths.

Spherical Harmonics

For simple boundaries, Laplace's Eq. (17) is relatively easy to solve provided there is an appropriate coordinate system. For the Earth, the solutions commonly rely on approximating the boundary by a sphere of radius, R . In spherical polar coordinates, (θ, λ, r) , the Laplacian operator is given by:

$$\nabla^2 = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \lambda^2}, \quad (21)$$

and solutions to $\nabla^2 V = 0$ in the space outside a sphere of radius, R , with center at the coordinate origin are (Kellogg 1953, p. 143; Morse and Feshbach 1953, p. 1264; Hobson 1965, p. 10):

$$V(\theta, \lambda, r) = P_{nm}(\cos \theta) \sin m\lambda \frac{1}{r^{n+1}} \quad \text{and} \quad V(\theta, \lambda, r) = P_{nm}(\cos \theta) \cos m\lambda \frac{1}{r^{n+1}}, \quad (22)$$

where $P_{nm}(y)$ is the associated Legendre function of the first kind and n, m are integers such that $0 \leq m \leq n, n \geq 0$. Other solutions exist, but only Eqs. (22) are consistent with the problem of finding a real-valued Newtonian potential for the exterior space of the Earth (regular at infinity and 2π -periodic in longitude).

The general solution is a linear combination of solutions (22) for all possible integers, n and m ,

$$V(\theta, \lambda, r) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(\frac{R}{r} \right)^{n+1} v_{nm} \bar{Y}_{nm}(\theta, \lambda), \quad (23)$$

where the $\bar{Y}_{nm}$ are surface spherical harmonic functions defined as:

$$\bar{Y}_{nm}(\theta, \lambda) = \bar{P}_{n|m|}(\cos \theta) \begin{cases} \cos m\lambda, & m \geq 0 \\ \sin |m|\lambda, & m < 0 \end{cases} \quad (24)$$

and $\bar{P}_{nm}$ is a normalization of P_{nm} so that the orthogonality of the spherical harmonics is:

$$\frac{1}{4\pi} \iint_{\sigma} \bar{Y}_{nm}(\theta, \lambda) \bar{Y}_{n'm'}(\theta, \lambda) d\sigma = \begin{cases} 1, & n = n' \text{ and } m = m' \\ 0, & n \neq n' \text{ or } m \neq m' \end{cases} \quad (25)$$

where $\sigma = \{(\theta, \lambda) | 0 \leq \theta \leq \pi, 0 \leq \lambda \leq 2\pi\}$ represents the unit sphere, with $d\sigma = \sin \theta d\theta d\lambda$. For a complete mathematical treatment of spherical harmonics, see Müller (1966). The functions, $\bar{Y}_{nm}(\theta, \lambda)/r^{n+1}$, are also known as solid spherical harmonics.

The constant coefficients, v_{nm} , known as Stokes's constants, have identical units of measure. Applying the orthogonality relation, Eq. (25), to the general solution (23), these coefficients can be determined if the function, V , is known on the bounding sphere (Dirichlet boundary condition):

$$v_{nm} = \frac{1}{4\pi} \int_{\sigma} V(\theta, \lambda, R) \bar{Y}_{nm}(\theta, \lambda) d\sigma. \quad (26)$$

The spherical harmonic expansion of V , Eq. (23), with Eq. (26) for the coefficients is a solution to the Dirichlet boundary-value problem if the boundary is a sphere. The solution thus exists and is unique in the sense that these boundary values generate no other potential.

In a more formal mathematical setting, the solution (23) is an infinite linear combination of orthogonal basis functions (eigenfunctions) and the coefficients, v_{nm} , are the corresponding eigenvalues. One may also interpret the set of coefficients as the spectrum (Fourier-Legendre spectrum) of the potential on the sphere of radius, R (analogous to the Fourier spectrum of a function depending on Cartesian coordinates on the plane or line). The integers, n, m , correspond to wavenumbers and are called degree (n) and order (m), respectively.

The spherical harmonics are further classified as zonal ($m = 0$), meaning that the zeros of $\bar{Y}_{n0}$ divide the sphere into latitudinal zones; sectorial ($|m| = n$), where the zeros of $\bar{Y}_{nm}$ divide the sphere into longitudinal sectors; and tesseral ($0 < |m| < n$), where the zeros of $\bar{Y}_{nm}$ tessellate the sphere (Fig. 1).

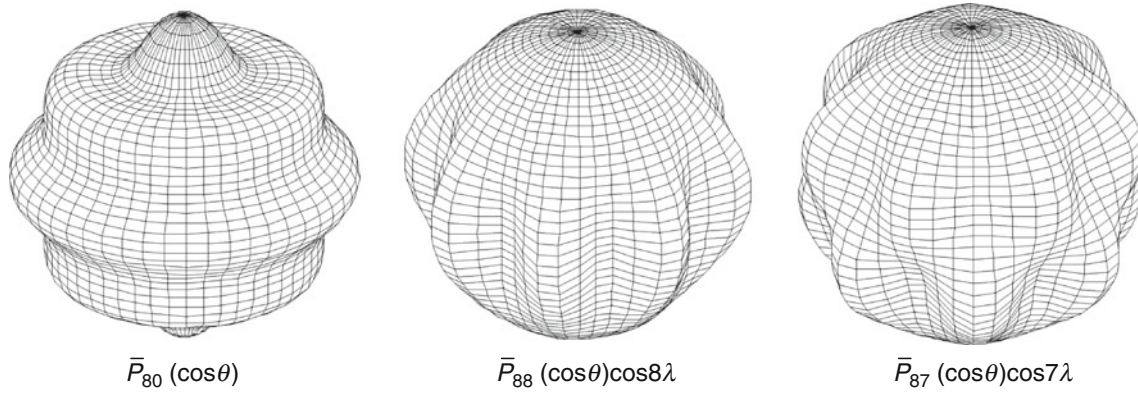
Recent global mappings of the Earth's gravitational field using satellite observations, specifically the GRACE mission (Gravity Recovery and Climate Experiment, Tapley et al. 2004), have emphasized the temporal dependence of the potential due to mass transport on and within the Earth (hydrologic flux, glacial isostatic adjustment, ice mass loss). In terms of the global spherical harmonic expansion, this implies a temporal dependence in the Stokes's coefficients:

$$V(\theta, \lambda, r; t) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(\frac{R}{r} \right)^{n+1} v_{nm}(t) \bar{Y}_{nm}(\theta, \lambda). \quad (27)$$

While the spherical harmonic series has its advantages in global representations and spectral interpretations of the field, the support region for each of its basis functions is global, which can be inefficient from the standpoint of data analysis. As more data are incorporated, or improved, each coefficient, v_{nm} , must be reestimated. Representations using local-support basis functions have recently been developed (e.g., Freedman et al. 1998) and are topics of current research. Such constructions also permit multi-resolution decompositions analogous to wavelet analysis.

Green's Functions

The Green's function representation of the field also offers a more regional characterization of the field. Although the Green's function strictly has global support, it attenuates rapidly from the evaluation point of the potential and reflects that a remote change in boundary value generally does not appreciably affect the local determination of the field.



Gravity Field of the Earth, Fig. 1 Spherical harmonics of the types zonal (left), sectorial (middle), and tesseral (right)

When the boundary is a sphere, the solution to the boundary-value problem using a Green's function is easily derived for the potential from its spherical harmonic series representation. Additional integral relationships (with appropriate Green's functions) among all the derivatives of the potential can be derived similarly (Jekeli 2007). Using the decomposition formula for the n^{th} -degree Legendre polynomial,

$$P_n(\cos \psi) = \frac{1}{2n+1} \sum_{m=-n}^n \bar{Y}_{nm}(\theta, \lambda) \bar{Y}_{nm}(\theta', \lambda'), \quad (28)$$

where

$$\cos \psi = \cos \theta \cos \theta' + \sin \theta \sin \theta' \cos(\lambda - \lambda'), \quad (29)$$

and substituting the harmonic coefficients, Eq. (26), into the series, Eq. (23), we find:

$$V(\theta, \lambda, r) = \frac{1}{4\pi} \iint_{\sigma} V(\theta', \lambda', R) U(\psi, r) d\sigma, \quad (30)$$

where with $\ell = \sqrt{r^2 + R^2 - 2rR \cos \psi}$

$$\begin{aligned} U(\psi, r) &= \sum_{n=0}^{\infty} (2n+1) \left(\frac{R}{r}\right)^{n+1} P_n(\cos \psi) \\ &= \frac{R(r^2 - R^2)}{\ell^3}. \end{aligned} \quad (31)$$

The closed formula (right side of Eq. (31)) comes from an appropriate manipulation of the Coulomb expansion (Cushing 1975, p. 155) for the reciprocal distance,

$$\frac{1}{\ell} = \frac{1}{R} \sum_{n=0}^{\infty} \left(\frac{R}{r}\right)^{n+1} P_n(\cos \psi). \quad (32)$$

Solutions (23) and (30) to the Dirichlet boundary-value problem for a spherical boundary are identical. The integral,

Eq. (30), is known as the Poisson integral and the function U is also known as Poisson's kernel.

The Disturbing Potential

For convenience, one separates Earth's gravitational potential into a reference potential (section "Normal Reference Field") and the disturbing potential, T . The disturbing potential is harmonic in free space, satisfies Poisson's integral if the boundary is a sphere, and is expandable in a series of solid spherical harmonics:

$$\begin{aligned} T(\theta, \lambda, r) &= \frac{GM}{R} \sum_{n=0}^{\infty} \\ &\times \sum_{m=-n}^n \left(\frac{R}{r}\right)^{n+1} \delta C_{nm} \bar{Y}_{nm}(\theta, \lambda) \end{aligned} \quad (33)$$

where M is the total mass (including the atmosphere) of the Earth and the δC_{nm} are unitless harmonic coefficients, being also the difference between coefficients for the total and reference gravitational potentials. The coefficient, δC_{00} , is zero under the assumption that the reference field accounts completely for the central part of the total field. Note that also these coefficients specifically refer to the sphere of radius, R .

The gravity anomaly (in spherical approximation) is defined by:

$$\Delta g(\theta, \lambda, r) = \left(-\frac{\partial}{\partial r} - \frac{2}{r}\right) T(\theta, \lambda, r), \quad (34)$$

or, using Eq. (33),

$$\begin{aligned} \Delta g(\theta, \lambda, r) &= \frac{GM}{R^2} \sum_{n=0}^{\infty} \\ &\times \sum_{m=-n}^n \left(\frac{R}{r}\right)^{n+2} (n-1) \delta C_{nm} \bar{Y}_{nm}(\theta, \lambda). \end{aligned} \quad (35)$$

The spectral relationship for $n \geq 2$ between gravity anomalies and the disturbing potential, obtained by comparing Eqs. (33 and 35), leads to a corresponding Green's function

that transforms one to the other. This function in geodesy is called more commonly Stokes's function:

$$S(\psi, r) = \sum_{n=2}^{\infty} \frac{2n+1}{n-1} \left(\frac{R}{r}\right)^{n+1} P_n(\cos \psi) \\ = 2\frac{R}{\ell} + \frac{R}{r} - 3\frac{R\ell}{r^2} - 5\frac{R^2}{r^2} \cos \psi - 3\frac{R^2}{r^2} \cos \psi \ln \frac{\ell + r - R \cos \psi}{2r} \quad (36)$$

and thus, without zero and first-degree harmonics (Hofmann-Wellenhof and Moritz 2005; Jekeli 2007):

$$T(\theta, \lambda, r) = \frac{R}{4\pi} \iint_{\sigma} \Delta g(\theta', \lambda', R) S(\psi, r) d\sigma, \quad (37)$$

which is known as the Pizzetti-Stokes integral. This equation in geodesy solves the Robin boundary-value problem for the sphere: the potential and its normal derivative are given on the boundary; see Eq. (34).

The solution is not unique since, as Eq. (35) shows, the gravity anomaly has no first-degree harmonics, which, therefore, cannot be determined from the boundary values. Conventionally, as shown in Eq. (36), the Stokes kernel also excludes the zero-degree harmonic and thus the complete solution for the disturbing potential is given by:

$$T(\theta, \lambda, r) = \frac{GM}{r} \delta C_{00} + \frac{GM}{R} \\ \times \sum_{m=-1}^1 \left(\frac{R}{r}\right)^2 \delta C_{1m} \bar{Y}_{1m}(\theta, \lambda) \\ + \frac{R}{4\pi} \iint_{\sigma} \Delta g(\theta', \lambda', R) S(\psi, r) d\sigma. \quad (38)$$

The central term, δC_{00} , is proportional to the difference in GM of the Earth and reference ellipsoid and is zero to high accuracy. The first-degree harmonic coefficients, δC_{1m} , are proportional to the center-of-mass coordinates and can also be set to zero with appropriate definition of the coordinate system (section “Low-Degree Harmonics as Density Moments”). Thus, the Pizzetti-Stokes integral (37) is the more common expression for the disturbing potential, but it embodies hidden constraints.

Spectral Contributions

The correlation function of the disturbing potential on a sphere is defined as (Moritz 1980; Jekeli 2017):

$$\phi_T(\psi) = \frac{1}{8\pi^2} \int_0^{2\pi} \int_0^{2\pi} \int_{\sigma} T(\theta, \lambda, R) T(\theta', \lambda', R) \sin \theta d\theta d\lambda d\alpha, \quad (39)$$

where the points (θ, λ, R) and (θ', λ', R) are connected by Eq. (29) and their relative azimuth, α ,

$$\tan \alpha = \frac{-\sin \theta' \sin(\lambda - \lambda')}{\sin \theta \cos \theta' - \cos \theta \sin \theta' \cos(\lambda - \lambda')}, \quad (40)$$

and where the integration is performed over all possible pairs of these points, with fixed spherical distance, ψ , $0 \leq \psi \leq \pi$.

Because of its sole dependence on ψ , ϕ_T can be expressed as an infinite series of Legendre polynomials:

$$\phi_T(\psi) = \sum_{n=0}^{\infty} (2n+1) (\Phi_T)_n P_n(\cos \psi). \quad (41)$$

where the coefficients, $(\Phi_T)_n$, constitute the Legendre transform of ϕ_T :

$$(\Phi_T)_n = \frac{1}{2} \int_0^{\pi} \phi_T(\psi) P_n(\cos \psi) \sin \psi d\psi. \quad (42)$$

Substituting Eqs. (28) and (39) into Eq. (42), and then simplifying using the orthogonality of spherical harmonics, Eq. (25), and the definition of the Fourier-Legendre spectrum, Eq. (26), we find:

$$(\Phi_T)_n = \frac{1}{2n+1} \sum_{m=-n}^n v_{nm}^2. \quad (43)$$

The quantities, $(\Phi_T)_n$, constitute the power spectral density (PSD) of T , and

$$(c_T)_n = (2n+1) (\Phi_T)_n \quad (44)$$

is known as a degree variance or variance per degree, since the total variance, from Eq. (41), is:

$$\phi_T(0) = \sum_{n=0}^{\infty} (c_T)_n. \quad (45)$$

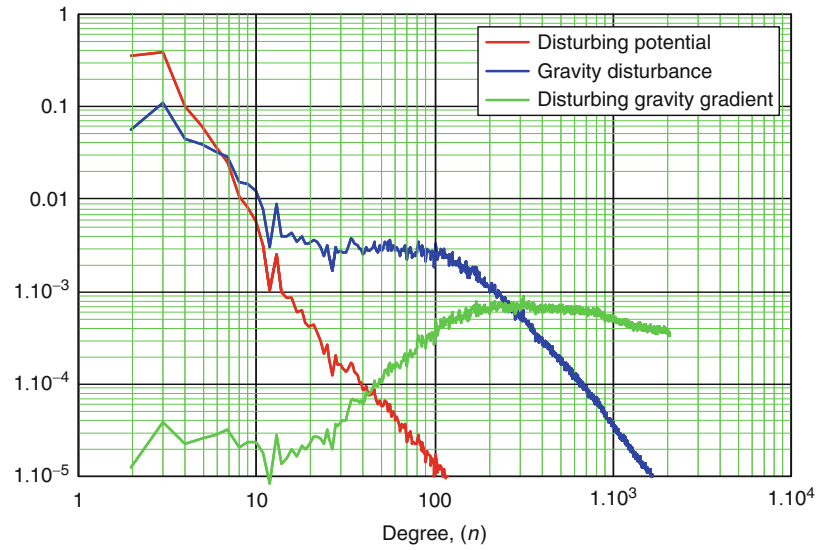
For the first and second radial derivatives of the disturbing potential, called the gravity disturbance ($\delta g = -\partial T / \partial r$) and vertical-vertical gradient disturbance ($\delta \Gamma = \partial^2 T / \partial r^2$), one has analogous PSDs that are related to the PSD of T as follows:

$$(\Phi_{\delta \Gamma})_n = \left(\frac{n+2}{R}\right)^2 (\Phi_{\delta g})_n \\ = \left(\frac{n+2}{R}\right)^2 \left(\frac{n+1}{R}\right)^2 (\Phi_T)_n. \quad (46)$$

Using one of the more recent high-degree Earth Gravitational Models, EGM2008 (Pavlis et al. 2012), Fig. 2

Gravity Field of the Earth,

Fig. 2 Degree variances of radial derivatives of the Earth's disturbing potential. All values are normalized by the respective total variance



illustrates the relative spectral content of the derivatives of the (disturbing) gravitational potential. Clearly, the higher derivatives emphasize the shorter wavelengths (or higher degrees, higher frequencies). The EGM2008 model has maximum degree 2160, corresponding to about 9.3 km resolution on the Earth's surface ($180^\circ/2160 = 5$ arcmin).

Low-Degree Harmonics

The low-degree spherical harmonics of the Earth's gravitational potential lend themselves to interpretation with respect to the most elemental distribution of the Earth's mass density, which also leads to fundamental geometric characterizations, particularly for the second-degree harmonics. Let

$$C_{nm}^{(a)} = \frac{a}{GM} \left(\frac{R}{a}\right)^{n+1} v_{nm} \quad (47)$$

be unitless coefficients that refer to a sphere of radius, a , which is the equatorial radius of the Earth ellipsoid. (Recall that coefficients, v_{nm} , Eq. (50), refer to a sphere of radius, R .) Relative to the central harmonic coefficient, $C_{00}^{(a)} = 1$, the next significant harmonic, $C_{20}^{(a)}$, is more than three orders of magnitude smaller, and the remaining harmonic coefficients are at least two to three orders of magnitude smaller than that. The attenuation of the coefficients after degree two is much more gradual (Table 1, below), indicating that the bulk of the potential can be described by an ellipsoidal field. The normal gravitational field is such a field, but it also adheres to a geodetic definition that requires the underlying ellipsoid to be an equipotential surface in the corresponding normal gravity field. The following sections examine the low-degree harmonics from these two perspectives of interpretation and reference.

Gravity Field of the Earth, Table 1 Spherical harmonic coefficients of the total gravitational potential^a

Degree, n	Order, m	$C_{nm}^{(a)}$	$C_{n,-m}^{(a)}$
2	0	-4.84170E-04	0.0
2	1	-2.39832E-10	1.42489E-09
2	2	2.43932E-06	-1.40028E-06
3	0	9.57189E-07	0.0
3	1	2.03048E-06	2.48172E-07
3	2	9.04802E-07	-6.19006E-07
3	3	7.21294E-07	1.41437E-06
4	0	5.39992E-07	0.0
4	1	-5.36167E-07	-4.73573E-07
4	2	3.50512E-07	6.62445E-07
4	3	9.90868E-07	-2.00976E-07
4	4	-1.88472E-07	3.08827E-07

^aGRACE model GGM02S (Tapley et al. 2005)

Low-Degree Harmonics as Density Moments

Returning to the general expression for the gravitational potential in terms of the Newtonian density integral (Eq. (8)) and substituting the spherical harmonic series for the reciprocal distance (Eq. (32) with (28)):

$$V(\theta, \lambda, r) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(\frac{R}{r}\right)^{n+1} \left(\frac{G}{R^{n+1}(2n+1)} \iiint_V (r')^n \bar{Y}_{nm}(\theta', \lambda') \rho(\theta', \lambda', r') dv \right) \bar{Y}_{nm}(\theta, \lambda) \quad (48)$$

yields a multipole expansion (so-called from electrostatics) of the potential. The spherical harmonic (Stokes) coefficients are multipoles of the density distribution:

$$v_{nm} = \frac{G}{R^{n+1}(2n+1)} \iiint_V (r')^n \bar{Y}_{nm}(\theta', \lambda') \rho(\theta', \lambda', r') dv. \quad (49)$$

One may also consider the n th-order moments of the mass density (from the statistics of distributions) defined by:

$$\begin{aligned} \mu_{\alpha\beta\gamma}^{(n)} &= \iiint_V (x')^\alpha (y')^\beta (z')^\gamma \rho dv, \\ n &= \alpha + \beta + \gamma. \end{aligned} \quad (50)$$

The multipoles of degree n and the moments of order n are related, though not all $(n+1)(n+2)/2$ moments of order n can be determined from the $2n+1$ multipoles of degree n , when $n \geq 2$ (clearly risking confusion, we defer to the common nomenclature of “order” for moments and “degree” for multipoles). This indeterminacy is directly connected to the inability to determine the density distribution uniquely from external measurements of the potential (Chao 2005), which is a characteristic of the classic geophysical inverse problem.

The zero-degree Stokes coefficient is coordinate invariant and is proportional to the total mass of the Earth:

$$v_{00} = \frac{G}{R} \iiint_V \rho dv = \frac{GM}{R}. \quad (51)$$

It also represents a mass monopole, and it is proportional to the zeroth moment of the density.

The first-degree harmonic coefficients (representing dipoles) are proportional to the coordinates of the center of mass, (x_{cm}, y_{cm}, z_{cm}) , which are proportional to the first-order moments of the density, as verified by recalling the definition of the first-degree spherical harmonics:

$$\begin{aligned} v_{1m} &= \frac{G}{\sqrt{3}R^2} \iiint_V \rho \begin{cases} r' \sin \theta' \sin \lambda', & m = -1 \\ r' \cos \theta', & m = 0 \\ r' \sin \theta' \cos \lambda', & m = 1 \end{cases} dv \\ &= \frac{GM}{\sqrt{3}R^2} \begin{cases} y_{cm}, & m = -1 \\ z_{cm}, & m = 0 \\ x_{cm}, & m = 1 \end{cases} \end{aligned} \quad (52)$$

Nowadays, by tracking satellites, we have access to the center of mass of the Earth (including its atmosphere) since it defines the “center” of their orbits. Ignoring the small motion of the center of mass (annual amplitude of several mm) due to the

temporal variations in the mass distribution, we may choose the coordinate origin for the gravitational potential to coincide with the center of mass, thus essentially annihilating the first-degree coefficients.

The second-order density moments likewise are related to the second-degree harmonic coefficients (quadrupoles). They also define the inertia tensor of the body. The inertia tensor is the proportionality factor in the equation that relates the angular momentum vector, $\mathbf{H}$, and the angular velocity, $\boldsymbol{\omega}$, of a body, like the Earth:

$$\mathbf{H} = \mathbf{I}\boldsymbol{\omega}, \quad (53)$$

and is given by:

$$\mathbf{I} = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix}. \quad (54)$$

It comprises the moments of inertia on the diagonal:

$$\begin{aligned} I_{xx} &= \iiint_{\text{volume}} (y'^2 + z'^2) \rho dv, \\ I_{yy} &= \iiint_{\text{volume}} (z'^2 + x'^2) \rho dv, \\ I_{zz} &= \iiint_{\text{volume}} (x'^2 + y'^2) \rho dv, \end{aligned} \quad (55)$$

and the products of inertia off the diagonal:

$$\begin{aligned} I_{xy} &= I_{yx} = - \iiint_{\text{volume}} x' y' \rho dv, \\ I_{xz} &= I_{zx} = - \iiint_{\text{volume}} x' z' \rho dv, \\ I_{yz} &= I_{zy} = - \iiint_{\text{volume}} y' z' \rho dv. \end{aligned} \quad (56)$$

Note that:

$$\begin{aligned} I_{xx} &= \mu_{020}^{(2)} + \mu_{002}^{(2)}, & I_{xy} &= -\mu_{110}^{(2)}, & I_{xz} &= -\mu_{101}^{(2)} \\ I_{yy} &= \mu_{002}^{(2)} + \mu_{200}^{(2)}, & I_{yz} &= -\mu_{011}^{(2)} \\ I_{zz} &= \mu_{200}^{(2)} + \mu_{020}^{(2)} \end{aligned} \quad (57)$$

and there as many (6) independent tensor components as second-order density moments. Using the explicit expressions for the second-degree spherical harmonics, we have from Eq. (49) with $n = 2$:

$$\begin{aligned}
v_{2,-2} &= -\frac{\sqrt{15}}{5R^3} G I_{xy}, & v_{2,-1} &= -\frac{\sqrt{15}}{5R^3} G I_{yz}, \\
v_{2,1} &= -\frac{\sqrt{15}}{5R^3} G I_{xz}, & v_{2,0} &= \frac{\sqrt{5}}{10R^3} G (I_{xx} + I_{yy} - 2I_{zz}), \\
v_{2,2} &= \frac{\sqrt{15}}{10R^3} G (I_{yy} - I_{xx})
\end{aligned} \tag{58}$$

These are also known as MacCullagh's formulas. Not all density moments (or moments of inertia) can be determined from the Stokes coefficients.

If the coordinate axes are chosen so as to diagonalize the inertia tensor (products of inertia are equal to zero), then they are known as principal axes of inertia, or also "figure" axes. For the Earth, the z -figure axis is very close to the spin axis (within several meters at the pole); both axes move with respect to each other and with respect to the Earth's crust, with combinations of various periods (daily, monthly, annual, etc.), as well as secularly in a wandering fashion. Because of these polar motions, the figure axis is not a useful coordinate axis that defines a frame fixed to the (surface of the) Earth. However, because of the proximity of the figure axis to the defined reference z -axis, the second-degree, first-order harmonic coefficients of the geopotential, and hence the products of inertia, I_{xz} and I_{yz} , are relatively small (Table 1).

The arbitrary choice of the x -axis of our Earth-fixed reference coordinate system certainly did not attempt to eliminate the product of inertia, I_{xy} (the x -axis is defined by the intersection of the Greenwich meridian with the equator, and the y -axis completes a right-handed mutually orthogonal triad). However, it is possible to determine where the x -figure axis is located by combining values of the second-degree, second-order harmonic coefficients. Let u, v, w be the axes that define a coordinate system in which the inertia tensor is diagonal, and assume that the approximation, $I_{ww} = I_{zz}$, holds. A rotation by the angle, $-\lambda_0$, about the w - (also z -) figure axis brings this ideal coordinate system back to the conventional one in which we calculate the harmonic coefficients from satellite tracking data. Tensors transform under rotation, defined by matrix, $\mathcal{R}$, according to:

$$I_{xyz} = \mathcal{R} I_{uvw} \mathcal{R}^T. \tag{59}$$

With the rotation about the w -axis given by the matrix:

$$\mathcal{R} = \begin{pmatrix} \cos \lambda_0 & -\sin \lambda_0 & 0 \\ \sin \lambda_0 & \cos \lambda_0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{60}$$

and with Eqs. (58), it is straightforward to show that:

$$\begin{aligned}
v_{2,-2} &= -\frac{\sqrt{15}G}{10R^3} (I_{uu} - I_{vv}) \sin 2\lambda_0 \\
v_{2,2} &= -\frac{\sqrt{15}G}{10R^3} (I_{uu} - I_{vv}) \cos 2\lambda_0
\end{aligned} \tag{61}$$

Hence, we have:

$$\lambda_0 = \frac{1}{2} \tan^{-1} \frac{v_{2,-2}}{v_{2,2}}, \tag{62}$$

where the quadrant is determined by the signs of the harmonic coefficients. From Table 1 and Eq. (47), we find that $\lambda_0 = -14.929^\circ$, i.e., the u -figure axis is in the mid-Atlantic between South America and Africa.

The second-degree, second-order harmonic coefficient, $v_{2,2}$, indicates the asymmetry of the Earth's mass distribution projected on the equator. Since $v_{2,2} > 0$ (for the Earth), Eqs. (61) show that $I_{vv} > I_{uu}$ and thus the equator "bulges" more in the direction of the u -figure axis; conversely, the equator is flattened in the direction of the v -figure axis. This flattening is relatively small: $1.1 \cdot 10^{-5}$.

Finally, consider the most important second-degree harmonic coefficient, the second zonal harmonic, $v_{2,0}$. Irrespective of the x -axis definition, it is proportional to the difference between the moment of inertia, I_{zz} , and the average of the equatorial moments, $(I_{xx} + I_{yy})/2$. Again, since $v_{2,0} < 0$, the Earth bulges more around the equator and is flattened at the poles. The second zonal harmonic coefficient is roughly 1000 times larger than the other second-degree coefficients and thus indicates a substantial polar flattening (owing to the Earth's early more fluid state). This flattening is approximately 0.003.

Normal Reference Field

A model for the Earth's gravitational field that is mathematically simple and yet accurate to better than one part in 63,000 (for the potential) is based on an ellipsoid of revolution rather than the sphere. The precise geodetic definition of the normal (reference) gravitational potential is that it is generated by an ellipsoidal mass that rotates with the Earth such that the gravitational potential plus the "centrifugal potential" due to the centrifugal acceleration (related as in Eq. (4)) is a constant, U_0 , on the ellipsoid. This definition sets up a Dirichlet boundary-value problem, whose solution is an analytic expression in ellipsoidal coordinates (Hofmann-Wellenhof and Moritz 2005). The boundary is defined by two parameters, the semi-major axis, a , and polar flattening, f , of the (rotationally symmetric) normal ellipsoid; the boundary values of the potential are defined by the given constant, U_0 , and the rotation rate of the Earth, ω_e (which specifies the centrifugal potential at any particular latitude). The normal ellipsoid, by definition, is centered at Earth's center of mass.

The solution for the normal potential is thus based on four fundamental constants, Earth's rotation rate, the size and shape parameters of the ellipsoid, and a gravitational scale parameter. The constant, U_0 , is related to several other gravitational quantities, such as the value of normal gravity (mass attraction plus the centrifugal acceleration) at the equator, or the total mass of the ellipsoid, or it could simply be taken as the potential value of the geoid. It can be shown (Hofmann-Wellenhof and Moritz 2005) that:

$$U_0 = \frac{GM}{E} \tan^{-1} \frac{E}{b} + \frac{1}{3} \omega_e^2 a^2, \quad (63)$$

where $b = a(1 - f)$ is the semi-minor axis, $E = \sqrt{a^2 - b^2}$ is the linear eccentricity of the ellipsoid, and M is its total mass. The mass density is not constant within the normal ellipsoid and its variation does not approximate Earth's internal structure. Rather the normal gravity potential (the sum of gravitational potential due to mass attraction and the centrifugal potential) is designed to emulate the actual external field, being constant on the ellipsoid, just as the actual gravity potential is constant on the geoid.

The normal field is always defined by adopting values of the four parameters that approximately match the corresponding actual parameters for the Earth. Before artificial Earth-orbiting satellites, the gravitational scale could best be accessed by measuring gravity (e.g., a nominal value at the equator served as the physical scale parameter). By tracking satellites, the value of GM for the Earth (total mass including the atmosphere) can be estimated to high accuracy and is still used today to define the physical scale of the normal potential. The ellipsoid parameters were traditionally determined from (partially) global geometric surveying networks projected onto a best-fitting ellipsoid. Today, the semi-major axis is well determined from a best fit of an ellipsoid to a global set of ocean heights derived from satellite altimeter data. The polar flattening, according to the section “Low-Degree Harmonics as Density Moments,” is related to the second zonal harmonic of Earth's gravitational potential, which likewise can be estimated very accurately from satellite tracking data. The second zonal harmonic coefficient, the *dynamical form factor*, thus serves as the shape parameter for the normal

ellipsoid. Table 2 lists internationally adopted normal ellipsoids over the past 100 years.

The exact analytical expression for the normal gravity potential then yields exact expressions for the normal gravity (and gravitation) and all its gradients. Nevertheless, approximate series expressions are sometimes preferred, especially to conform to a particular coordinate system. For example, the normal gravitational potential (due only to mass attraction) in spherical polar coordinates is given by:

$$V(\theta, \lambda, r) = \frac{GM}{r} \left(1 - \sum_{n=2}^{\infty} J_{2n} \left(\frac{a}{r} \right)^{2n} P_{2n}(\cos \theta) \right), \quad (64)$$

where with $e = E/a$ (the first eccentricity),

$$J_{2n} = (-1)^{n+1} \frac{3e^{2n}}{(2n+1)(2n+3)} \left(1 - n + 5n \frac{J_2}{e^2} \right), \quad (65)$$

and where

$$J_2 = \frac{1}{Ma^2} (I_{zz} - I_{xx}) = -\frac{\sqrt{5}R^3}{GMa^2} v_{2,0} = -\sqrt{5}C_{2,0}^{(a)}, \quad (66)$$

making use of Eqs. (58) and (47). From Eq. (65), it is clear that only the first few coefficients (say, up to $n = 5$) are significant for all applications, since $e^2 \approx 0.0066$ (Table 3). Also, the normal potential, comprising only even zonal harmonics, is independent of longitude and symmetric with respect to the equator.

Normal gravity (including the centrifugal acceleration) on the ellipsoid as function of *geodetic* latitude, ϕ (angle of the perpendicular to the ellipsoid with respect to the equator), is given by the Somigliana formula (Moritz 1980):

$$\gamma(\phi) = \frac{a\gamma_a \cos^2 \phi + b\gamma_b \sin^2 \phi}{\sqrt{a^2 \cos^2 \phi + b^2 \sin^2 \phi}}, \quad (67)$$

where normal gravity on the equator and at the poles is:

$$\gamma_a = \frac{GM}{ab} \left(1 - m - \frac{m}{6} \frac{e'q'_0}{q_0} \right), \quad (68)$$

Gravity Field of the Earth, Table 2 Defining parameters for normal ellipsoids of geodetic reference systems

Reference system ^a	a [m]	J_2	GM [m ³ s ⁻²]	ω_e [rad s ⁻²]
International Reference System (1930)	6,378,388	1.0920E-03 ^b	3.98633E14 ^b	7.2921151E-05
GRS67	6,378,160	1.0827E-03	3.98603E14	7.2921151467E-05
GRS80	6,378,137	1.08263E-03	3.986005E14	7.292115E-05
WGS84	6,378,137	1.08262982131E-03	3.986004418E14	7.2921151467E-05
Best values ^c	6378136.7 ±0.1 (mean-tide system)	(1.0826359 ±0.0000001)E-03 (zero-tide system)	(3.986004418 ±0.000000008)E14 (includes atmosphere)	7.292115E-05 (mean value)

^aGRS = Geodetic Reference System, WGS = World Geodetic System

^bDerived values from defining parameters, $f = 1/297$, and $\gamma_a = 9.780490$ m/s²

^cGroton (2004)

Gravity Field of the Earth, Table 3 Normal gravity harmonic coefficients corresponding to GRS80 (Moritz 2000)

n	J_{2n}
1	1.08263E-03
2	-2.37091222E-06
3	6.08347E-09
4	-1.427E-11
5	1.2E-14

$$\gamma_b = \frac{GM}{a^2} \left(1 + \frac{m}{3} \frac{e' q'_0}{q_0} \right), \quad (69)$$

with $m = \omega_e^2 a^2 b / GM$, $e' = E/b$ (the second eccentricity) and

$$\begin{aligned} q_0 &= \frac{1}{2} \left(1 + \frac{3}{e'^2} \right) \tan^{-1} e' - \frac{3}{2e'}, \\ q'_0 &= 3 \left(1 + \frac{1}{e'^2} \right) \left(1 - \frac{1}{e'} \tan^{-1} e' \right) - 1. \end{aligned} \quad (70)$$

At points along the perpendicular to the ellipsoid, the following approximate formula for normal gravity as a function also of ellipsoidal height, h , is often used:

$$\begin{aligned} \gamma(\phi, h) &= \gamma(\phi) \\ &\times \left(1 - \frac{2}{a} (1 + f + m - 2f \sin^2 \phi) h + \frac{3}{a^2} h^2 \right). \end{aligned} \quad (71)$$

which is accurate to $1.4 \times 10^{-6} \text{ m/s}^2$, or better, up to altitude, $h = 20,000 \text{ m}$, above the ellipsoid.

Earth's Internal Gravitational Field

While Earth's external gravitational field may be determined entirely from measurements and potential theory, the internal field is much less accessible to measurements and depends on models of the subsurface structure that can only be inferred from surface measurements and constraints provided by various other geophysical observations. Some have likened the study of the Earth's interior to the study of the cosmos – the investigator is incapable of performing in situ measurements and is left to use diverse (and sometimes ingenious) proxies to infer the Earth's interior makeup. Of course, there are borehole measurements and these are probing ever deeper into the Earth's crust, but they are isolated and expensive. A complete reconstruction of the Earth's (deeper) interior comes at present primarily from seismological data by analyzing the travel times of hundreds of thousands to millions of compressional (P) and shear (S) waves as they pass through Earth's body. Near-surface reconstruction also relies on seismic sounding (seismograms from intentional explosions) as well

gravimetry, magnetometry, and other geophysical sensors, including ground-penetrating radar, electromagnetic induction, and electric resistivity methods.

We may argue that knowledge of the interior gravity field is of secondary importance to an understanding of Earth's internal structure inasmuch this knowledge is derived from an Earth model that is already determined from other data. Only when gravity can be measured within the Earth (e.g., in boreholes) does the interior gravitational field offer a contributory importance to our modeling of the Earth's interior. At present, the deepest borehole in the world, at just over 12,260 m, is located in Russia on the Kola Peninsula on the Barents Sea north-east of Finland. Most holes are drilled for geological exploration and reach depths of several hundred meters.

Nevertheless, it is of interest to compute the gravitational attraction along a radius of the Earth, as modeled by a radially layered composite of mass density. From Eq. (8), it is easily shown that the gravitational potential inside a radially symmetric body (having radius, R) has the form:

$$\begin{aligned} V(r) &= \frac{4\pi G}{r} \int_0^r \rho(r') r'^2 dr' + 4\pi G \int_r^R \rho(r') r' dr', \\ 0 &\leq r \leq R. \end{aligned} \quad (72)$$

Thus, the gravitational acceleration is given by:

$$g(r) = -\frac{d}{dr} V(r) = \frac{4\pi G}{r^2} \int_0^r \rho(r') r'^2 dr'. \quad (73)$$

For the Preliminary Reference Earth Model (PREM) of 1981 (Dziewonski and Anderson 1981; Dziewonski 1989), which consists of approximately 80 layers, the radial variation of gravitation is shown in Fig. 3 together with the change in density.

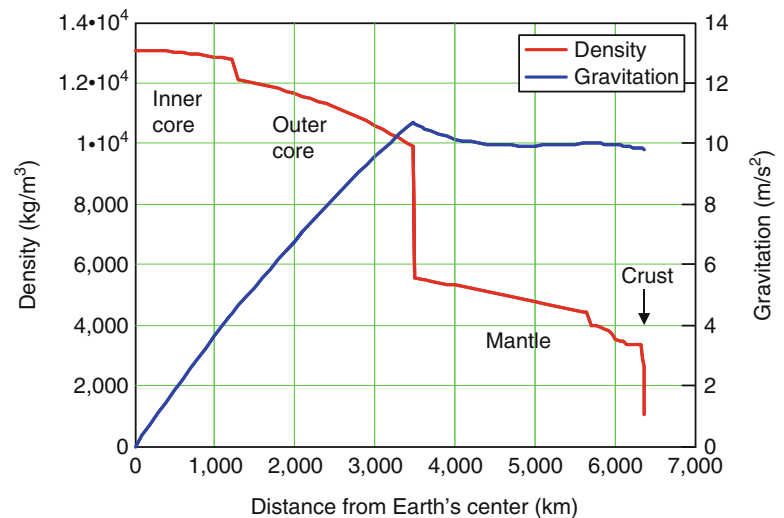
On a more local scale, we may use the attraction of an infinite horizontal plate of constant density, ρ_0 , as a model to approximate the value of gravity in the crust (or ocean) near the Earth's surface. The attraction of such a plate of thickness, d , anywhere above or below the plate is:

$$B = 2\pi G \rho_0 d, \quad (74)$$

which is also known as the Bouguer plate effect. Suppose that g_P is the value of gravity at a surface point, P , on the Earth. Assuming that the topography is relatively flat in its vicinity, removal of a plate of material with density, ρ_0 , and thickness, d , diminishes the gravity value at P by B . Now bring the gravity value to the level d through free space using a normal gradient approximation, which changes its value by $F \approx -(\partial\gamma/\partial h)d$ (also known as the free-air reduction). Finally,

Gravity Field of the Earth,

Fig. 3 Gravitation and mass density inside the Earth according to the radially symmetric PREM



restore the material plate, which again decreases the value of gravity, now at depth, d , by B .

The resulting gravity as a function of depth thus becomes:

$$g(d) \approx g_p - \left(\frac{\partial \gamma}{\partial h} + 4\pi G \rho_0 \right) d. \quad (75)$$

Using nominal values for the normal gradient, $\partial \gamma / \partial h \approx -0.3086$ mGal/m (where $1 \text{ mGal} = 10^{-5} \text{ m/s}^2$), and the ocean water or crust density, $\rho_w = 1027 \text{ kg/m}^3$ or $\rho_0 = 2670 \text{ kg/m}^3$, respectively, this becomes:

$$g(d) \approx g_p + 0.2225d \text{ [mGal]}, \quad (76)$$

d in units of meter below sea surface;

$$g(d) \approx g_p + 0.0848d \text{ [mGal]}, \quad (77)$$

d in units of meter below land surface.

This is also known as the Poincaré-Prey reduction and may be used to model gravity in boreholes and in the oceans, given gravity at the surface. The model of gravity within the Earth's upper crust (or ocean) can be improved with more careful modeling of the effects of variable topography and variable density (especially in the oceans).

Summary

The Earth's gravitational field, defined as the field generated by the subsurface mass density, embodies a wealth of geophysical information at global and local scales. The external global field is often described in terms of a spherical harmonic expansion, where the low-degree harmonics have a definite (though incomplete) interpretation with respect to the moments of the density distribution. The attenuation of the potential in its spherical harmonic (spatial frequency) domain

is sufficiently gradual that first- and second-order radial derivatives readily characterize the local, near-surface structure of the Earth. Whereas the external gravitational field of the Earth can be modeled in detail on the basis of surface measurements and potential theory, the deeper internal gravitational field model is derived secondarily from density distribution models obtained from other sources, such as seismological data.

Bibliography

- Chao BF (2005) On inversion for mass distribution from global (time-variable) gravity field. *J Geodyn* 29:223–230
- Cushing JT (1975) *Applied analytical mathematics for physical scientists*. Wiley, New York
- Dziewonski AM (1989) Earth structure, global. In: James DE (ed) *The encyclopedia of solid earth geophysics*. Van Nostrand Reinhold Company, New York, pp 331–358
- Dziewonski AM, Anderson DL (1981) Preliminary reference earth model. *Phys Earth Planet Inter* 25:297–356
- Freedman W, Gervens T, Schreiner M (1998) *Constructive approximation on the sphere, with applications in geomathematics*. Clarendon Press, Oxford
- Groton E (2004) Fundamental parameters and current (2004) best estimates of the parameters of common relevance to astronomy, geodesy, and geodynamics. *J Geod* 77:724. <https://doi.org/10.1007/s00190-003-0373-y>
- Heiland CA (1940) *Geophysical exploration*. Prentice-Hall, New York
- Heiskanen WA, Moritz H (1967) *Physical geodesy*. Freeman and Co., San Francisco
- Heiskanen WA, Veing Meinesz FA (1958) *The earth and its gravity field*. McGraw-Hill, New York
- Helmert FR (1884) *Die Mathematischen und Physikalischen Theorien der Höheren Geodäsie*, vol 2. B.G. Teubner, Leipzig. (reprinted in 1962 by Minerva GMBH, Frankfurt/Main)
- Hobson EW (1965) *The theory of spherical and ellipsoidal harmonics*. Chelsea Publishing Co., New York
- Hofmann-Wellenhof B, Moritz H (2005) *Physical geodesy*. Springer, Berlin
- Hotine M (1969) *Mathematical geodesy*. U.S. Department of Commerce, Washington, DC
- Jekeli C (1988) The exact transformation between ellipsoidal and spherical harmonic expansions. *Manuscr Geodaet* 14:106–113
- Jekeli C (2000) *Inertial navigation systems with geodetic applications*. Walter deGruyter, Berlin

- Jekeli C (2007) Potential theory and static gravity field of the earth. In: Schubert G (ed) *Treatise on geophysics*, vol 3. Elsevier, Oxford, pp 11–42
- Jekeli C (2017) *Spectral methods in geodesy and geophysics*. CRC Press, Boca Raton
- Kellogg OD (1953) *Foundations of potential theory*. Dover, New York
- Martin JL (1988) *General relativity: a guide to its consequences for gravity and cosmology*. Wiley, New York
- Molodensky MS, Eremeev VF, Yurkina MI (1962) *Methods for the study of the external gravity field and figure of the earth*. Israel Program of Scientific Translations, Jerusalem. (Russian original, 1960)
- Moritz H (1980) *Advanced physical geodesy*. Wichmann, Karlsruhe. (reprint 2008 by School of Earth Sciences, Ohio state university, Columbus)
- Moritz H (2000) Geodetic reference system 1980. *J Geod* 74(1):128–133
- Morse PM, Feshbach H (1953) *Methods of theoretical physics*, Parts I and II. McGraw-Hill, New York
- Müller C (1966) *Spherical harmonics*. Lecture notes in mathematics. Springer, Berlin
- Nettleton LL (1976) *Gravity and magnetism in oil prospecting*. McGraw-Hill, New York
- NIST (2019). https://physics.nist.gov/cgi-bin/cuu/Value?bg|search_for=gravitational+constant
- Pavlis NK, Holmes SA, Kenyon SC, Factor JF (2012) The development and evaluation of earth gravitational model (EGM2008). *J Geophys Res* 117:B04406. <https://doi.org/10.1029/2011JB008916>
- Plag HP, Pearlman M (eds) (2009) *Global geodetic observing system, meeting the requirements of a global society on a changing planet in 2020*. Springer, Berlin
- Sansò F, Rummel R (eds) (1997) *Geodetic boundary value problems in view of the one centimeter geoid*. Lecture notes in earth sciences. Springer, Berlin, p 65
- Sansò F, Sideris MG (eds) (2013) *Geoid determination, theory and methods*. Lecture notes in the earth sciences. Springer, Berlin
- Tapley BD, Bettadpur S, Watkins M, Reigber C (2004) The gravity recovery and climate experiment, mission overview and early results. *Geophys Res Lett* 31(9). <https://doi.org/10.1029/2004GL019920>
- Tapley BD, Ries J, Bettadpur S, Chambers D, Cheng M, Condi F, Gunter B, Kang Z, Nagel P, Pastor R, Pekker T, Poole S, Wang F (2005) GGM02 – an improved earth gravity field model from GRACE. *J Geod* 79:467. <https://doi.org/10.1007/s00190-005-0480-z>
- Telford WM, Geldart LP, Sheriff RE (1990) *Applied geophysics*, 2nd edn. Cambridge University Press, Cambridge, UK
- Torge W (1989) *Gravimetry*. Walter deGruyter, Berlin

Gravity Field, Temporal Variations from Space Techniques

Anny Cazenave¹, J. L. Chen² and G. Ramillien³

¹LEGOS, Observatoire Midi Pyrénées,
Toulouse, France

²Center for Space Research, University of Texas,
Austin, TX, USA

³GET, Observatoire Midi Pyrénées,
Toulouse, France

Definition

Observing the Earth from space over the past 50 years has resulted in spectacular advances in various areas of the

geosciences. This is the case, in particular, for measuring the precise shape of the Earth and its broad-scale internal structure from detailed measurements of the planet's gravity field. Since the early 1970s, tracking the orbits of tens of satellites at different altitudes and orbit inclinations has gradually improved our knowledge of Earth's average (static) gravity field (see section on "Gravity and Satellites"). Besides the precise description of the static component, attempts have been made to detect temporal variations of the gravity field. For example, a secular decrease of Earth's flattening (the degree 2, order 0 term) was discovered by precise analysis of the Lageos satellite laser ranging observations (Yoder et al. 1983) and attributed to the Global Isostatic Adjustment (GIA), the viscoelastic response of the crust and mantle after melting of ice at end of the Last Glacial Maximum (LGM) (Peltier 2009). Seasonal variations of the low-degree terms of the Earth's gravity field expansion (up to ~degree 6) have been detected from analysis of satellite orbital perturbations (Cheng and Tapley 2004). These variations were attributed to seasonal transfer of air mass inside the atmosphere as well as annual fluctuations of snow and land water storage (e.g., Cazenave et al. 1999). However, temporal changes of only the longest-wavelength (lowest degree) spherical harmonic (also called Stokes) coefficients could be monitored from orbital perturbations. This limited effective spatial resolution of observed mass change to several thousand kilometer, because large uncertainty in higher degree coefficients prevented detection of any time-varying signal.

Since 2002, the US–German jointly sponsored Gravity Recovery and Climate Experiment (GRACE) mission has been providing a precise survey of Earth's time-variable gravity field, with unprecedented temporal and spatial sampling (Tapley et al. 2004). GRACE time-variable gravity fields provide a means of measuring temporal and spatial variations of mass redistribution within the Earth system. The GRACE mission has launched a new era in studying a series of geophysical problems ranging from deep Earth structure to tracking mass redistribution on and near the surface of the Earth. GRACE has greatly improved our understanding of mass redistribution in various compartments of the climate system (atmosphere, oceans, terrestrial water, and cryosphere) and fundamentally enriched a number of fields, including (but not limited to) the global water cycle and land hydrology, mass balance of polar ice sheets and mountain glaciers, ocean mass and global sea level change, and solid Earth geophysics.

The GRACE Mission (2002–2017)

Launched in March 2002, GRACE utilized a state-of-the-art technique to observe variation of Earth's gravity by tracking the inter-satellite range and range rate between two coplanar, low-altitude satellites via a K-band ranging system (Tapley

et al. 2004). In addition, each satellite was equipped with a SuperSTAR accelerometer, GPS receiver/antenna, star cameras, and laser retro-reflector to complete the complement of science instruments. The GRACE Science Data System used the range and range rate data, along with ancillary data, to estimate a new gravity field every month, in the form of corrections to a well-defined background gravity model used in the data processing procedure. The monthly sampling rate has been selected in order to accumulate sufficient observations to ensure spatial resolution of about 400 km. Generally, there is a trade-off in selecting the temporal sampling interval. Accumulating data over longer time intervals increases spatial resolution but decreases temporal resolution and vice versa.

Modeling Temporal Geoid Variations

The static component of the gravity field G_0 corresponds to nearly 99% of the total field, mainly due to solid Earth contributions. This term can be easily evaluated and removed by computing the temporal mean of a long enough series of GRACE monthly geoids or alternatively considering a single geoid computed with a long period of time of satellite observations. The monthly time-variable geoid $\delta G(t)$ can merely be computed as the difference between the monthly geoid $G(t)$ measured by GRACE at time t and the static mean field component:

$$\delta G(t) = G(t) - G_0 \quad (1)$$

Thus, the corresponding geoid differences (also called monthly GRACE geoids in the following) only reflect short-term geoid change associated with mass redistributions inside the Earth system.

Let $\delta C_{nm}(t)$ and $\delta S_{nm}(t)$ be the time-variable “normalized” Stokes coefficients of the spherical harmonic expansion of geoid height, where n and m are the degree and order, respectively. Thus, the time-variable geoid is classically expressed as (Heiskanen and Moritz 1967):

$$\delta G(t) = \sum_{n=1}^N \sum_{m=0}^n (\delta C_{nm}(t) \cos(m\lambda) + \delta S_{nm}(t) \sin(m\lambda)) P_{nm}(\cos\theta) \quad (2)$$

where N is the maximum degree of the harmonic decomposition, θ is the colatitude, λ is the longitude, and P_{nm} is the associated Legendre polynomial which is dimensionless.

GRACE geoids are further expressed in surface mass change (usually in equivalent water height) assuming that mass redistribution occurs in thin surface layers compared

to the Earth’s dimension (Swenson and Wahr 2002), so that:

$$\begin{pmatrix} \delta C_{nm}(t) \\ \delta S_{nm}(t) \end{pmatrix} = \frac{(1+z_n) \text{Re}^2}{(2n+1)M} \iint_S \delta q(\theta, \lambda, t) \begin{pmatrix} \cos \\ \sin \end{pmatrix} (m\lambda) P_{nm}(\cos\theta) \delta S \quad (3)$$

where $\delta q(\theta, \lambda, t)$ is a surface density of surface water/ice/air mass, expressed in terms of equivalent water thickness at time t , whose harmonic coefficients, $\delta A_{nm}(t)$ and $\delta B_{nm}(t)$, can be evaluated from the corresponding geoid anomaly coefficients through the linear filtering:

$$\begin{pmatrix} \delta C_{nm}(t) \\ \delta S_{nm}(t) \end{pmatrix} = W_n \begin{pmatrix} \delta A_{nm}(t) \\ \delta B_{nm}(t) \end{pmatrix} \quad (4)$$

W_n is an isotropic spatial filter that weights the surface density coefficients, and its analytical expression is (Swenson and Wahr 2002):

$$W_n = \frac{2\pi G \text{Re} \rho_w}{(2n+1)\gamma(\theta)} (1+z_n) \quad (5)$$

In Eqs. (3) and (5) above, z_n represents the Love numbers that enables to take into account instantaneous elastic compensation of the Earth’s surface to the load; $\gamma(\theta)$ is the normal gravity on the reference ellipsoid at colatitude of θ . G ($\sim 6.67 \cdot 10^{-11} \text{ m}^3 \text{kg}^{-1} \text{s}^{-2}$) is the gravitational constant, M is the total mass of the Earth ($\sim 597602 \cdot 10^{24} \text{ kg}$), and Re ($\sim 6378 \text{ km}$) is Earth’s mean radius. ρ_w ($\sim 1000 \text{ kgm}^{-3}$) is the water density. Equations (4) and (5) are used to compute surface density variations $\delta q(\theta, \lambda, t)$ (or mass redistribution due to air, water, and ice mass change in surface fluid envelopes) from the monthly geoid harmonic coefficients.

GRACE Data Errors

Assessing the errors of GRACE monthly solutions was a major challenge. Two categories of errors were recognized: (1) errors introduced by data processing (errors of GRACE measurements and models used to remove other geophysical signals) and (2) post-processing errors (e.g., spatial smoothing of the GRACE solutions to remove high-frequency noise and leakage errors due to signal contamination from regions outside the studied area). Among errors in category (1), correlated errors due to un-modeled effects and contamination of other geophysical signals during data processing are highly problematic. Because of the limited spatiotemporal sampling provided by the GRACE satellites, these errors are aliased into the estimated spherical harmonic coefficients, giving rise to north–south stripes evident in the raw GRACE geoid and

mass maps. To reduce this organized spatial noise (category 1 error), a decorrelation filter has been proposed (Swenson and Wahr 2006), which can help suppress GRACE striping noise. Post-processing error arises mainly from two sources. One is from the limited degree and order of Stokes coefficients GRACE can provide, i.e., the truncation of GRACE gravity Stokes coefficients, typically at degree 60 or 96. The other is from the needed Gaussian smoothing to further remove the residual noise (after the decorrelation filtering). These post-processing errors significantly affect GRACE estimates of surface mass change by reducing signal amplitude at the higher degrees and orders and result in signal contamination into surrounding regions (called “leakage”). The GRACE orbital configuration (altitude and inter-satellite distance) and the need for spatial filtering and smoothing are the main factors that limit GRACE spatial resolution, estimated to be on the order of 300–400 km.

Another problem in the data processing arises from geocenter motion, i.e., the motion of the mass center of the Earth system with respect to origin of the terrestrial reference frame. GRACE gravity solutions are defined in a center of mass reference frame. Thus the degree 1 coefficients of the spherical harmonic expansion (representing geocenter motion) are missing in GRACE solutions. Independent estimates of geocenter motion from other geodetic techniques are needed (to be added to GRACE solutions), in order to accurately quantify longest-scale mass changes in the Earth system. In addition, the degree 2, order zero term is not well resolved by GRACE and traditionally is replaced by estimates based on satellite laser ranging. The accuracy of monthly GRACE solutions is estimated to be ~2 cm of equivalent water height change over areas exceeding one million square kilometer (e.g., Wahr et al. 2006).

Figure 1 shows a map of mass trends over global continental areas from GRACE over the period April 2002 to June 2017, computed from least-squares fit to time series at every grid point. Units are in cm of equivalent water height change per year. This map shows predominantly water mass gain or loss in major river basins and ice mass loss in ice-covered regions including ice sheets, and glaciers, and mass trends due to solid Earth geophysical effects, e.g., the strong GIA signals in the Hudson Bay in Canada and Scandinavia in Northern Europe, and seismic deformations due to major earthquakes (e.g., the 2004 Sumatra–Andaman and 2011 Tohoku–Oki earthquakes).

Applications of GRACE

In this section, we present a few examples of present-day mass variations detected by GRACE over continental river basins, Greenland and Antarctic ice sheets, and over the

oceans. The capability of GRACE in measuring solid Earth mass changes (e.g., GIA and co-/post-seismic deformations due to large earthquakes) is also briefly discussed.

Hydrology

The global water cycle that exchanges water among oceans, atmosphere, and land plays a key role in the physical and chemical processes that influence Earth’s climate and its change over time. On land, terrestrial water storage (TWS) change, as a major component of the global water cycle, reflects changes in water stored in rivers and lakes, soil, snow, and groundwater reservoirs and is closely connected to accumulated precipitation (P), evapotranspiration (ET), and surface and subsurface runoff (R) within a given area or basin, through the water balance equation relating a change in TWS in unit time Δt .

$$\frac{\Delta TWS}{\Delta t} = P - ET - R \quad (6)$$

For most land areas (not covered by snow), TWS is virtually a measure of total water content in surface reservoirs (rivers and lakes), soil layers, and groundwater reservoirs and is also a good indicator of abnormal climate conditions, such as droughts and floods. Accurately quantified TWS change provides a key measure of the continental water cycle and available water resources in a given region or river basin. However, estimating TWS change from in situ networks is difficult due to the lack of groundwater and soil moisture observations, which often are simply nonexistent at basin or smaller scales.

GRACE data have been successfully applied to the study of global and basin scale TWS changes (e.g., Wahr et al. 2004). Early studies mostly focused on the seasonal cycle and showed good agreement with estimates from global hydrological models. The extended GRACE time series also provide estimates of interannual and long-term variability in TWS. For example, several studies have detected severe drought events in different regions of the world, e.g., the Amazonia drought in 2005 (Chen et al. 2009) and Australia millennium drought (Leblanc et al. 2009). Anthropogenic effects on the terrestrial water cycle, such as groundwater pumping, are also detectable by GRACE (e.g., Rodell et al. 2009). Other studies have used GRACE TWS in combination with additional information to estimate changes in vertical or horizontal fluxes, e.g., $P - ET$, ET , or R , of the right-hand side of the water balance equation (Eq. 6) (e.g., Ramillien et al. 2006).

Land Ice Loss from GRACE

GRACE satellite gravity measurements offer a unique opportunity to directly estimate land ice loss and its contribution to

the global mean sea level rise. Land ice loss includes glacier melting and mass decrease of Greenland and Antarctic ice sheets. Numerous studies have been published during the last decade to provide estimates of the mass balance of Greenland and Antarctica and indicate a clear acceleration of ice mass loss from both ice sheets (e.g., Velicogna and Wahr 2006; Chen et al. 2006). Ice sheet mass loss explains about 30% of the global mean sea level rise of the last 15 years (WCRP Global Mean Sea Level Budget Group 2018). GRACE satellite gravity measurements have also been widely used to study ice mass change of mountain glaciers (e.g., Tamisiea et al. 2005). However, the quantification of mountain glaciers ice mass change is more challenging, limited by GRACE resolution. Optimal data processing methods, such as regional mascons or forward modeling, are needed to appropriately address GRACE leakage issue.

Ocean Mass Change

Net water flux into and out of the ocean causes its mass to change. Such changes cause gravitational variations detectable by GRACE (e.g., Chambers et al. 2004). This has allowed, for the first time, direct estimates of the global ocean mass contribution to sea level change. However, GRACE is also sensing GIA effect. On average over the oceanic domain, this effect is large and of the same order of magnitude as the ocean mass change due to present-day land ice melt itself. As GIA models vary significantly from each other with respect to this effect (e.g., Peltier 2009), GRACE ocean mass rates combined with land ice melt rates can place an independent constraint on the GIA correction.

Satellite altimetry allows precise measurement of present-day global mean sea level changes (e.g., Stammer and Cazenave 2018). The two main causes of the global mean sea level rise are ocean thermal expansion and ocean mass increase. GRACE is now routinely used to estimate the latter component, allowing assessment of closure/non-closure of the global mean sea level budget, when combined with satellite altimeter sea level measurements and Argo floats ocean temperature and salinity observations (e.g., WCRP Global Sea Level Budget Group 2018).

GIA Effect

GIA (also called postglacial rebound) represents large-scale viscoelastic readjustment (mostly rebounding) of Earth's crust and mantle following the removal of ice loads at the end of the LGM (Peltier 2009). GIA induces secular changes in the gravity field observable by GRACE. The large positive mass trend centered over the Hudson Bay (Canada) shown in Fig. 1 is one of the most prominent signatures of this process. Comparison between GRACE and other measures of GIA (e.g., surface gravity change, vertical crustal motions, model

estimates of ice load extent and melting history) has confirmed that GRACE is detecting Earth's recovery following melting of the Laurentide Ice Sheet at the end of the LGM (e.g., Tamisiea et al. 2007).

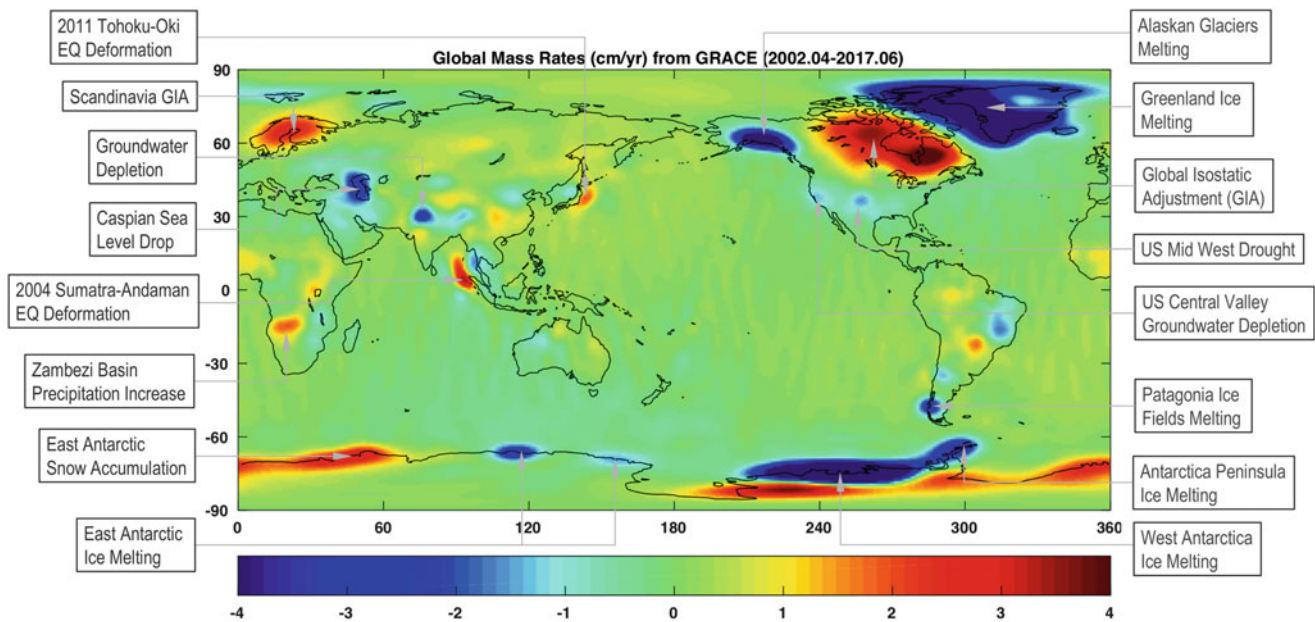
Coseismic and Post-seismic Deformations Due to Large Earthquakes

Besides GIA and the refinement of Earth's static gravity field, GRACE has contributed to the understanding of other solid Earth processes. In particular, GRACE appears able to detect the seafloor deformation, afterslip, and viscoelastic relaxation associated with large earthquakes. This is the case with the Sumatra–Andaman earthquake (magnitude of 9.3) of December 26, 2004, the largest recorded seismic event in about 40 years. GRACE successfully detected the gravity change associated with the corresponding coseismic subduction and uplift, in agreement with model predictions (e.g., Han et al. 2006). Co- and post-seismic deformations of another major earthquake, the 2011 Tohoku–Oki earthquake (magnitude of 9.0), have also been captured by GRACE (Li and Shen 2015). The seismic (mainly post-seismic) deformations of these two gigantic earthquakes have been illustrated in Fig. 1.

Summary

Since 2002, the US–German GRACE (Gravity Recovery and Climate Experiment) mission has been providing a precise survey of Earth's time-variable gravity field, with unprecedented temporal and spatial sampling. The first GRACE mission ended in 2017, but a new GRACE mission (GRACE Follow-On) was launched in May 2018, providing continuity in measuring the time-variable gravity fields, hence the spatiotemporal variations of mass redistributions within the Earth system. The GRACE missions have launched a new era in studying a series of geophysical problems ranging from deep Earth structure to tracking mass redistribution on and near the surface of the Earth. GRACE has greatly improved our understanding of mass redistribution in various compartments of the climate system (atmosphere, oceans, terrestrial water, and cryosphere) (Cazenave and Chen 2010; Rodell et al. 2018; Tapley et al. 2019). GRACE has fundamentally enriched a number of Earth science fields, including (but not limited to) the global water cycle and land hydrology, mass balance of polar ice sheets and mountain glaciers, ocean mass and global sea level change, and solid Earth geophysics.

More references on the use of GRACE for geosciences studies can be found at <https://grace.jpl.nasa.gov/publications/>.



Gravity Field, Temporal Variations from Space Techniques, Fig. 1 Global mass trends over continental areas observed by GRACE over the period April 2002 to June 2017, based on CSR RL06

gravity solutions with 300 km Gaussian smoothing applied. Some major long-term trends from climate change or solid geophysical sources are labeled on the sides

Cross-References

- [Geoid Determination, Theory and Principles](#)
- [Geoid Undulation, Interpretation](#)
- [Gravity Field of the Earth](#)
- [Remote Sensing, Applications to Geophysics](#)
- [Tracking Earth's Water in Motion from Satellite Gravity Observations](#)

Bibliography

- Cazenave A, Chen J (2010) Time-variable gravity from space and present-day mass redistribution in the Earth system. *Earth Planet Sci Lett* 298:263–274. <https://doi.org/10.1016/j.epsl.2010.07.035>
- Cazenave A, Mercier F, Bouille F, Lemoine JM (1999) Global-scale interactions between the solid Earth and its fluid envelopes at the seasonal time scale. *Earth Planet Sci Lett* 171:549–559
- Chambers DP, Wahr J, Nerem RS (2004) Preliminary observations of global ocean mass variations with GRACE. *Geophys Res Lett* 31: L13310. <https://doi.org/10.1029/2004GL020461>
- Chen JL, Wilson CR, Tapley BD (2006) Satellite gravity measurements confirm accelerated melting of Greenland ice sheet. *Science* 313:1958–1960. <https://doi.org/10.1126/science.1129007>
- Chen JL, Wilson CR, Tapley BD, Yang ZL, Niu GY (2009) 2005 drought event in the Amazon River basin as measured by GRACE and estimated by climate models. *J Geophys Res* 114:B05404. <https://doi.org/10.1029/2008JB006056>
- Cheng M, Tapley BD (2004) Variations in the Earth's oblateness during the past 28 years. *J Geophys Res* 109:B09402. <https://doi.org/10.1029/2004JB003028>
- Han S-C, Shum CK, Bevis M, Ji C, Kuo C-Y (2006) Crustal dilatation observed by GRACE after the 2004 Sumatra-Andaman earthquake. *Science* 313:658–662. <https://doi.org/10.1126/science.1128661>
- Heiskanen WA, Moritz H (1967) *Physical geodesy*. Freeman, San Francisco
- Leblanc ML, Tregoning P, Ramillien G, Tweed SO, Fakes A (2009) Basin-scale, integrated observations of the early 21st century multi-year drought in Southeast Australia. *Water Resour Res* 45:W04408. <https://doi.org/10.1029/2008WR007333>
- Li J, Shen W-B (2015) Monthly GRACE detection of coseismic gravity change associated with 2011 Tohoku-Oki earthquake using northern gradient approach. *Earth Planets Space* 67:29. <https://doi.org/10.1186/s40623-015-0188-0>
- Peltier WR (2009) Closure of the budget of global sea level rise over the GRACE era: the importance and magnitudes of the required corrections for global isostatic adjustment. *Quat Sci Rev* 28:1658–1674
- Ramillien G, Frappart F, Güntner A, Ngo-Duc T, Cazenave A, Laval K (2006) Time variations of the regional evapotranspiration rate from gravity recovery and climate experiment (GRACE) satellite gravimetry. *Water Resour Res* 42:W10403. <https://doi.org/10.1029/2005WR004331>
- Rodell M, Velicogna I, Famiglietti JS (2009) Satellite-based estimates of groundwater depletion in India. *Nature* 460. <https://doi.org/10.1038/nature08238>
- Rodell M, Famiglietti JS, Wiese DN, Reager JT, Beaulieu HK, Landerer FW, Lo M-H (2018) Emerging trends in global freshwater availability. *Nature* 557:651–659. <https://doi.org/10.1038/s41586-018-0123-1>
- Stammer D, Cazenave A (2018) *Satellite altimetry over oceans and land surfaces*. Taylor & Franc, Boca Raton
- Swenson S, Wahr J (2002) Methods for inferring regional surface mass anomalies from Gravity Recovery and Climate Experiments (GRACE) measurements of time-variable gravity. *J Geophys Res* 107:2193. <https://doi.org/10.1029/2001JB000576>

- Swenson S, Wahr J (2006) Post-processing removal of correlated errors in GRACE data. *Geophys Res Lett* 33:L8402. <https://doi.org/10.1029/2005GL025285>
- Tamisiea ME, Leuliette EW, Davis JL, Mitrovica JX (2005) Constraining hydrological and cryospheric mass flux in southeastern Alaska using space-based gravity measurements. *Geophys Res Lett* 32:L20501. <https://doi.org/10.1029/2005GL023961>
- Tamisiea ME, Mitrovica JX, Davis JL (2007) GRACE gravity data constrain ancient ice geometries and continental dynamics over Laurentia. *Science* 316:881–883. <https://doi.org/10.1126/science.1137157>
- Tapley BD, Bettadpur S, Watkins MM, Reigber C (2004) The gravity recovery and climate experiment; mission overview and early results. *Geophys Res Lett* 31(9):L09607. <https://doi.org/10.1029/2004GL019920>
- Tapley BD, Watkins MM, Flechtner F, Reigber C, Bettadpur S, Rodell M, Sasgen I, Famiglietti JS, Landerer FW, Chambers DP, Reager JT, Gardner AS, Save H, Ivins ER, Swenson SC, Boening C, Dahle C, Wiese DN, Dolslaw H, Tamisiea ME, Velicogna I (2019) Contributions of GRACE to understanding climate change. *Nature Climate Change* 9:358–369. <https://www.nature.com/nclimate>
- Velicogna I, Wahr J (2006) Measurements of time-variable gravity show mass loss in Antarctica. *Science* 311:1754–1756. <https://doi.org/10.1126/science.1123785>
- Wahr J, Swenson S, Zlotnicki V, Velicogna I (2004) Time-variable gravity from GRACE: first results. *Geophys Res Lett* 31:L11501. <https://doi.org/10.1029/2004GL019779>
- Wahr J, Swenson S, Velicogna I (2006) Accuracy of GRACE mass estimates. *Geophys Res Lett* 33:L06401
- WCRP Global Sea Level Budget Group (2018) Global sea-level budget 1993–present. *Earth Syst Sci Data* 10:1551–1590. <https://doi.org/10.5194/essd-10-1551-2018>
- Yoder CF, Williams JG, Dickey JO, Schultz BE, Eanes RJ, Tapley BD (1983) Secular variation of Earth's gravitational harmonic J2 coefficient from Lageos and nontidal acceleration of Earth rotation. *Nature* 307:57–762. <https://doi.org/10.1038/303757a0>

Gravity Field, Time Variations from Surface Measurements

Virendra M. Tiwari¹ and Jacques Hinderer²

¹Gravity and Magnetic Studies Group, CSIR-National Geophysical Research Institute (CSIR), Hyderabad, India

²Institut de Physique du Globe de Strasbourg, CNRS, Université de Strasbourg, Strasbourg, France

Synonyms

Earth's dynamic gravity field; Temporal variations of the Earth's gravity

Definition

Temporal gravity changes Variation of Earth's gravity field with time due to various tidal, oceanic, atmospheric, hydrological, glaciological, volcanic, seismic, and tectonic processes.

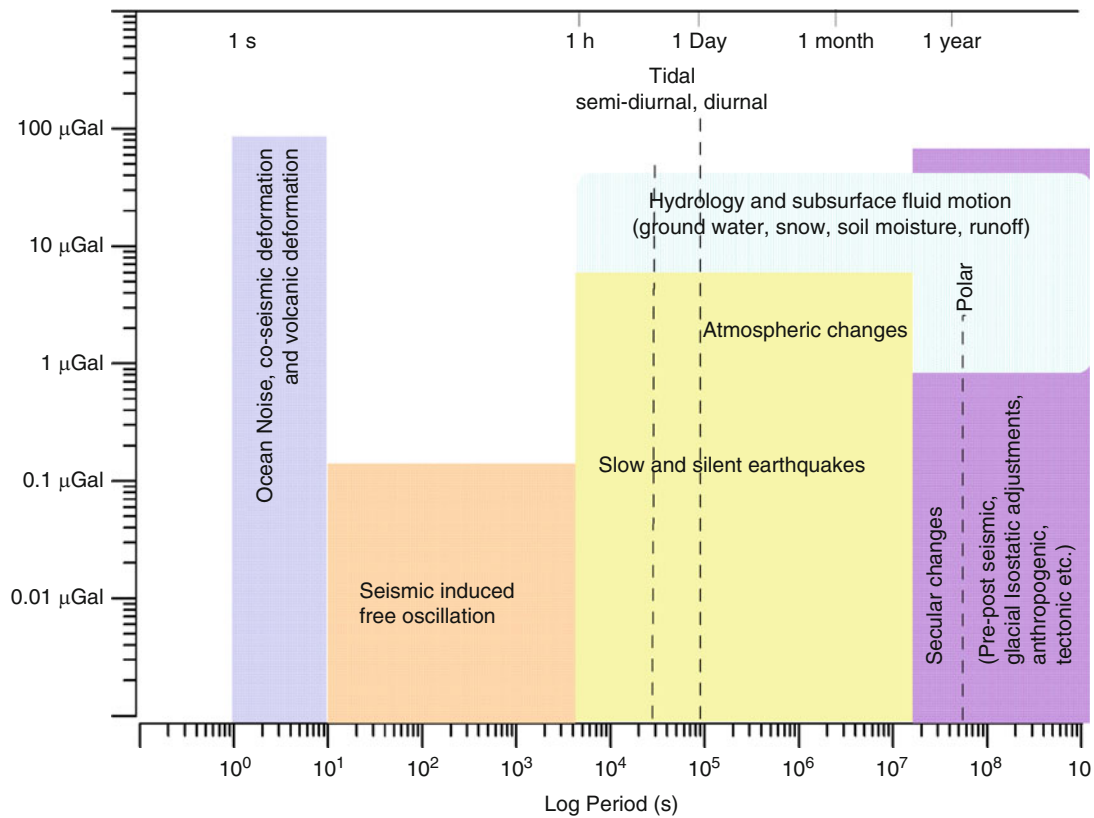
Tidal gravity changes The periodic variation of gravity field due to the attraction of Sun and Moon.

Non-tidal gravity changes Gravity changes unrelated to the attraction of Sun and Moon.

Introduction

The gravity field of the Earth is an attraction force exerted by the Earth on any unit mass normal to the equipotential surface. It is a function of mass distribution and its spatial location, inside, on, and above the Earth's surface, following the Newton's law of gravitation. Continuous changes in the Earth system and its fluid envelopes related to the solid and ocean tides, oceanic circulation, atmospheric changes, hydrological changes, volcanic, seismic and tectonic activity, Earth's polar motion, Earth's inner and fluid core motions redistribute the mass and produce variations in the Earth's gravitational field on a large range of spatial and temporal scales (Fig. 1). The variation in the gravity field on timescales of thousands to millions of years caused by redistribution of mass due to the processes that shape the Earth's surface is termed “static gravity field” as this change does not occur over human timescale.

Temporal variations of Earth's gravity field are determined by measurements of the gravitational field at the same location with time. Surface measurements of time-varying gravity field started in the twentieth century using spring type relative gravity meters that were mainly employed to study tidal related gravity changes. Gravity changes can now be precisely recorded by terrestrial measurements using absolute gravity meters (AG), superconducting gravity meter (SG), relative gravity meter, or a combination of them (AG is the most precise instrument available that can measure the absolute gravity field with an accuracy of 1–3 μGal ($1 \mu\text{Gal} = 10^{-8} \text{ m s}^{-2} = 10 \text{ nm s}^{-2}$)) (Niebauer et al. 1995). SG can measure changes in the Earth's gravity field with 0.1 μGal accuracy (Hinderer et al. 2015a). A strong effort to monitor the time-varying gravity field globally has been made with the launch of the GRACE satellites in 2002 (Wahr 2007) and more recently by GRACE-Follow On mission in 2018 after the end of the GRACE mission (2002–2017). Precise measurements of the changes in the Earth's gravity field either by satellite (NRC 1997; Wahr 2007) or from surface measurements (Hinderer and Crossley 2000; Crossley et al. 2013) thus offer the possibilities to constrain the underlying causes of dynamic changes of mass distribution in the Earth system processes. Both measuring methods have some advantages and drawbacks. For example, localized gravity changes (i.e., effect due to injection of water in oil field) cannot be detected in satellite measurements, and similarly gravity changes occurring over a large region, particularly over remote areas, cannot be mapped using surface measurements. An attempt is



Gravity Field, Time Variations from Surface Measurements, Fig. 1 Schematic diagram showing temporal gravity changes that can be studied from surface gravity measurements ($1 \mu\text{Gal} = 10^{-8} \text{ m s}^{-2} = 10 \text{ nm s}^{-2}$)

made to intercompare the satellite-derived gravity fields with surface gravity measurements observed with SG despite the fact that the ground and space gravity measurements are not sensitive to the same effects. Radial displacement is important for ground observations but has no effect on satellites, and ground gravity measurement is a point measurement while satellite measurements are integrated response over a large region (typically a few hundreds of kilometers; Wahr 2007). Nevertheless, first results of intercomparison of gravity changes observed by the European SG network and GRACE satellite data reported a fair agreement in the time-varying gravity field originated due to hydrological changes (Crossley et al. 2012; Hinderer et al. 2006).

Concepts and Applications

Primarily there are two categories of time-varying gravity signals: one that varies periodically and another one that is nonperiodic. Periodic gravity changes such as tides can be separated from the observations based on their periodicity but nonperiodic changes that arise from combinations of different contributing sources are difficult to recognize. However, they are identified using other complementary information. There are about 30 SGs in the world, which are continuously measuring gravity changes and form the international network

IGETS (International Geodynamics and Earth Tide Service) (Voigt et al. 2016) formerly called GGP (Crossley et al. 1999). IGETS stations are sparsely distributed worldwide with a very few stations in the southern hemisphere and in the equatorial band. These data are shared within the network for various applications and are also used for comparison with satellite-derived gravity changes that allows upscaling of terrestrial data and downscaling of satellite data (Hinderer et al. 2015a) for broader understanding of causes of gravity changes. AG measurements are also made at these locations for calibrations of SGs and for cross validation of secular changes. Both instruments (AG and SG) are essentially laboratory instruments; however, AG can be transported but SGs are mostly observatory instruments dedicated to continuous measurements over long durations. In the hybrid gravity approach (Hinderer et al. 2016), a combination of different surface gravity measurements is used to estimate spatiotemporal changes in Earth's gravity field. Details of surface gravity measurements, their accuracy, and potential application, as suggested by Crossley and Hinderer (2005) with minor modifications, is summarized and presented in Table 1.

The following sections present a brief description of the main sources of gravity changes that are recorded by surface gravity measurements and used to study the sources responsible for these changes. An example of different contributions (solid tides, polar motion and length of the day, atmospheric

Gravity Field, Time Variations from Surface Measurements, Table 1 Types of surface gravity measurements and their potential applications in studying time variable gravity field

	Absolute gravity meter (AG: Micro-g Solutions, FG 5)	Superconducting gravity meter (SG: GWR type)	Relative gravity meters (RG: Scintrex, CG3M, CG5M)
Precision	$\leq 1 \mu\text{Gal}$	$0.001 \mu\text{Gal}$	$1 \mu\text{Gal}$
Accuracy	$1\text{--}3 \mu\text{Gal}$	$0.1 \mu\text{Gal}$	$10 \mu\text{Gal}$
Drift	0	$0.01\text{--}0.05 \mu\text{Gal/day}$ (variable)	$3\text{--}10 \mu\text{Gal/day}$ (variable)
Setup time and operation	AG is a transportable instrument that operates in a controlled temperature environment. It can be set up in 1–3 h. One to three days measurements are desirable to get best results	Setup of SG requires a few days and some weeks to stabilize. It is housed at one place and needs calibration and drift monitoring by AG for better accuracy	RG is a field instrument that can be set up in a few minutes. A network of stations is used for measurement loops. Requires repeated measurements at a reference station every few hours
Dominant noise	Microseismic and man-made noise	Environmental effects	Environmental effects, microseismic noise, vehicle shaking, drift
Possible applications	Tidal, hydrological changes, pre-co-post seismic changes, glacial isostatic adjustment, tectonics, oil, gas, and geothermal reservoir monitoring, volcanic mass changes	Tidal, polar motion, hydrological changes, atmospheric and ocean loading's core processes, pre-co-post seismic changes, free oscillation of the glacial isostatic adjustments, volcanic mass changes	Tidal, hydrological changes, co-seismic changes for large quakes, volcanic mass changes

loading, tidal and nontidal ocean loading, soil moisture, and snow loading) to the time-varying gravity in Strasbourg is presented in Fig. 2. The amplitude of the solid tides and the ocean tidal loading would be ten times larger compared to plotted amplitudes in the Fig. 2, where only long-period tides are considered.

Tides

Tidal gravity changes are caused by the attraction of other planetary bodies like the Sun and Moon. The attraction of these planetary bodies deforms the Earth's solid and fluid surface and produces changes in the gravity field. These changes are periodic, mainly semi-diurnal and diurnal, but also with long-period components (fortnightly, monthly, half-yearly, and yearly as shown in Fig. 2). The semi-diurnal and diurnal tides can be very large (with peak to peak value of $\sim \pm 150 \mu\text{Gal}$) according to the station location and usually are the largest contribution in any surface gravity measurements. They can be computed accurately and therefore can be removed from observed gravity changes to study nontidal gravity changes (Wenzel 1996). The precise tidal gravity changes are measured with SGs, AGs, and other surface gravimeters (Hinderer et al. 2015a) on land and are recorded in satellite altimetry data over the oceanic region (Wahr 2007). Tidal gravity changes are the elastic response of the Earth to the tidal forces and hence, modeling of tidal gravity changes provide information on the internal structure and properties of the Earth as its elastic transfer function varies according to the internal distribution of density, compressibility, and rigidity parameters. The very long observational duration of some SGs of the IGETS international network exceeding for instance 25 years for the Strasbourg station

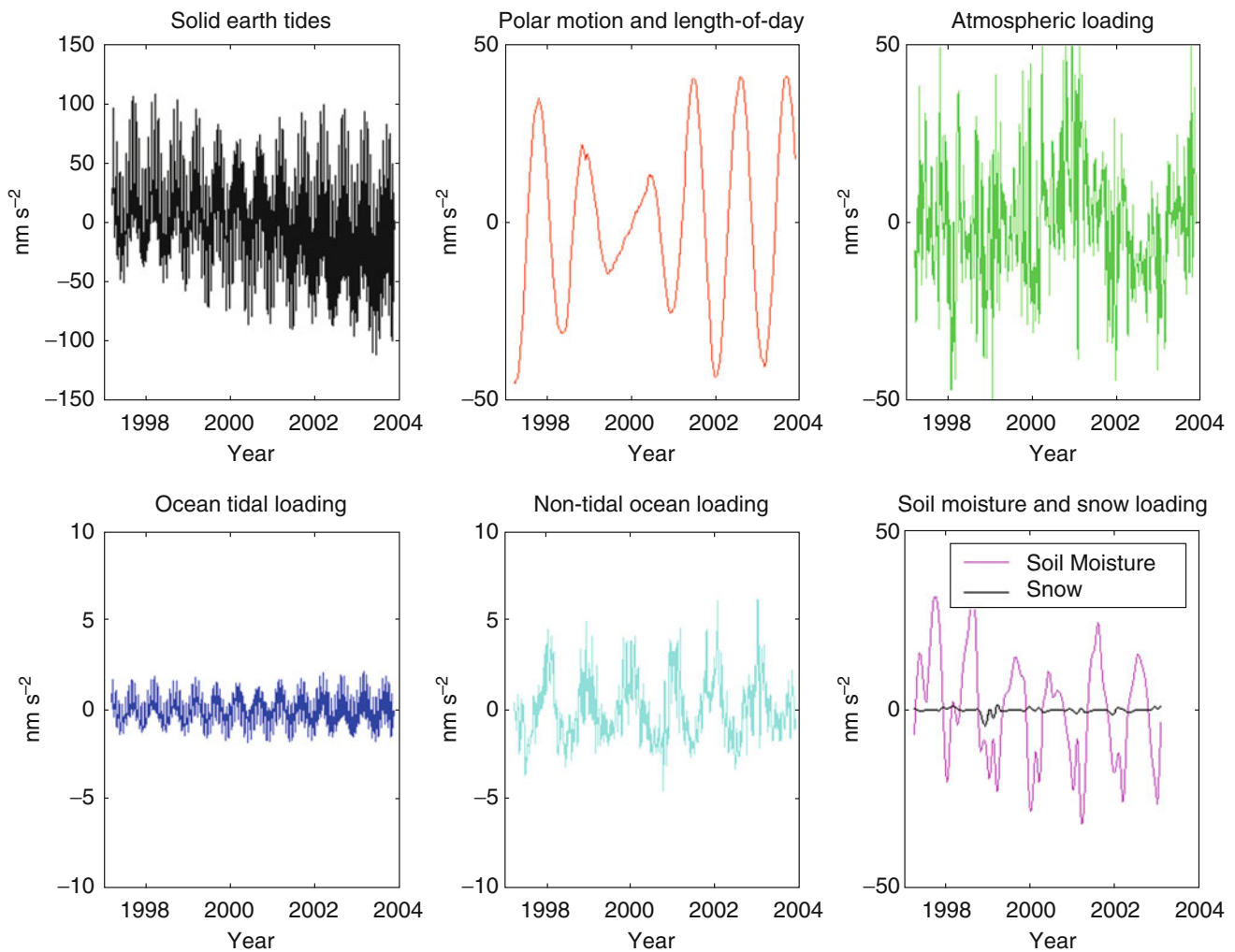
leads to the detection of tidal waves of very small amplitude that were previously hidden in noise for shorter records (Ducarme 2011; Calvo et al. 2016). Tidal observations also provide independent evidence of general relativity theory of planetary movements (Warburton and Goodkind 1976).

Sea Level Variation

Non-steric redistribution of ocean mass causes non-negligible changes in the Earth's gravity field. These are commonly recorded in satellite altimetry data. Movement of ocean waters also affects gravity changes inland; this may be quantified using results from satellite altimetry data (Fukuda and Sato 1997). It is reported that gravity changes caused by sea level change deducted from satellite altimetry data are of the order of few μGal near coastal regions that can be observed in SG data (Virtanen and Mäkinen 2003; Kroner et al. 2009).

Atmospheric Pressure

Atmospheric pressure changes alter surface gravity values because of the redistribution of mass in the atmosphere (attraction force) and the surface loading (elastic deformation) due to the change in the surface mass (Kroner and Jentzsch 1999; Boy et al. 2002; Hinderer et al. 2014). The induced gravity changes are in the order of $\pm 5 \mu\text{Gal}$ (Fig. 2) and can reach up to $10\text{--}15 \mu\text{Gal}$ for large pressure fronts ($30\text{--}50 \text{ hPa}$). They are often estimated using a simple atmospheric admittance function between gravity and atmospheric pressure, that is, $0.3 \mu\text{Gal/hPa}$ (Crossley et al. 1995). These effects are generally considered as noise in gravity data and corrected using local barometric pressure data. More sophisticated



Gravity Field, Time Variations from Surface Measurements, Fig. 2 Various contributions to time-varying gravity in Strasbourg, France (Courtesy: J.-P. Boy; $1 \mu\text{Gal} = 10 \text{ nm s}^{-2}$). Note that the

amplitude of the diurnal and semi-diurnal solid tides and the ocean tidal loading would be ten times larger than plotted amplitudes, which are only long-period tides

models of surface gravity loading where the Newtonian attraction of atmospheric masses is computed based on the 3D spatial density distribution (as derived from three-dimensional weather models) are also available (Klügel and Wziontek 2009).

Polar Motion

The direction of the Earth's axis of rotation is not steady in space but rather varies with time with two dominant periods (1 year and 435 days); this motion is generally known as polar motion or oscillation of the rotation pole. The 435-day component is referred to as Chandler wobble by astronomers. This signal can be detected by various techniques (spectral analysis, least squares adjustment) in surface gravity using continuous and stable SG observations (Hinderer et al. 2006). The

related signals are of several μGal and change according to the station coordinates (Ziegler et al. 2016) (Fig. 2).

Free Oscillations of the Earth

Large quakes are known to excite the free oscillations of the Earth (Hinderer et al. 2015a; Arora et al. 2008), which lead to surface gravity changes in the seismic frequency band. These changes accompanying the eigenmodes occur at periods between a few hundreds of seconds and 54 min, which is the longest period of the free elastic oscillations of the Earth, and can be well investigated using SG records (Rosat et al. 2004; Rosat and Hinderer 2018). Besides, free oscillations unrelated to quakes but rather due to atmosphere–ocean–solid Earth interactions (termed “hum”) were first reported from the SG measurements at Syowa, Antarctica (Nawa et al. 1998).

More recently SG data have also been analyzed for the detection of gravitational waves (Coughlin and Harms 2014) or dark matter (Horowitz and Widmer-Schmidrig 2020).

Glacial Isostatic Adjustment

On long timescales the Earth behaves like a viscoelastic material and deforms when subjected to a load. Major changes in climatic conditions of the Earth, from its birth to the present, caused large-scale changes in strengthening and weakening of the ice sheets of the cold regions of the Earth's surface. The response of the Earth to the glaciations and deglaciations is known as Glacial Isostatic Adjustment (GIA). Land-based gravity measurements have provided useful constraints to GIA, for instance, in the North American region (Lambert et al. 2001), Greenland (van Dam et al. 2000), in Svalbard (Sato et al. 2006; Mémin et al. 2014), and in Alaska (Sato et al. 2012). One of the important properties of deep Earth that GIA can provide is information about the viscosity of the Earth materials.

Volcanic Mass Changes

The growth of the volcanic dome and fluxes of volcanic masses cause local gravity changes. Volcanic mass redistributions produce gravity change in the order of tens to hundreds of μGal and can be observed using any kind of surface measurements. Surface gravity measurements along with other geodetic measurements for surface deformation like GPS or InSAR (Interferometric Synthetic Aperture Radar) are prospective tools for the estimate of mass changes during the volcanic processes (Jousset et al. 2000; Carbone et al. 2017).

Crustal Deformation Associated with Earthquakes

Sudden displacements of faults during a quake cause redistribution of mass as well as change in the elevation, both leading to local gravity changes. An example of gravity changes before and after Haicheng quake, 1975, and the Tangshan quake, 1976, is reported by Chen et al. (1979). Later on there have been several studies that document co-seismic gravity changes in surface measurements with SGs (Imanishi et al. 2004) and also in GRACE satellite observations (Han et al. 2006; Tiwari et al. 2008; de Linage et al. 2009).

In the case of the large subduction Tohoku-Oki earthquake in 2011 of magnitude 9.0, GRACE data have allowed to reveal large-scale gravity changes due to aseismic mass fluxes of tectonic origin (Panet et al. 2018).

During the same earthquake there is also a transient gravity signal with a small amplitude (fraction of μGal) that has been

detected by an SG before the arrival of seismic waves (Montagner et al. 2016) opening new perspectives in earthquake seismology and early-warning systems.

Gravity records of silent and slow quakes, for example, from Cascadian subduction zone showing the concordance of gravity changes of about 10 μGal with episodic slip and seismic tremor activity (several times a year) is one of the interesting observations in temporal gravity changes (Lambert et al. 2006).

Hydrology

Variations in land water storage (surface water, ground water, soil moisture, and snow) induce variations in the gravity field recorded on the surface (Hinderer et al. 2015a). The contributions of soil moisture and snow at Strasbourg, France, are plotted in Fig. 2. The amplitude of gravity changes associated with seasonal hydrological changes are larger than secular changes (Tiwari et al. 2006; Hinderer et al. 2015) and sometimes spread over a large area that can be mapped only through satellite measurements (Tiwari et al. 2009) because large-scale variations are difficult to be evaluated using ground-based observations due to the lack of adequate sampling. Hydrological fluxes at different time and spatial scales can now be mapped by monitoring temporal gravity changes, which holds promise to provide closing water cycle on local and regional scale (Tapley et al. 2019). Studies using the European regional network of SGs demonstrated that surface measurements of the seasonal component of continental hydrology can help in validating the satellite missions like GRACE (Hinderer et al. 2006; Crossley et al. 2012; Weise et al. 2012) and also suggest the complementary nature of satellite and ground-based observations (Tiwari et al. 2014). In the recent years there have been many experiments using SG measurements in various catchments in order to monitor water storage changes at the local scale (from meters to hundreds of meters). Some of these catchments are at mid-latitudes like in Europe (Creutzfeldt et al. 2010; Fores et al. 2016; Naujoks et al. 2010) or close to the equator in West Africa (Hinderer et al. 2012; Hector et al. 2015). Recently differential gravity studies using two simultaneously measuring SGs help even further to investigate hydrological signals at the local scale (Kennedy et al. 2014).

Oil, Gas and Geothermal Reservoir Monitoring

Time lapse gravity measurements are regularly used for monitoring subsurface density changes due to the fluid motions of the injected water and gas in the reservoir for enhancing recovery as well as for sequestration of CO_2 gases (Hare et al. 1999; Alnes et al. 2008; Sugihara et al. 2013; Jacob et al. 2015). The amplitude of gravity changes is quite

appreciable in the region where injections are large and reservoir depth is small. Such changes in gravity field occur over injection time period and are generally in the order of few tens of μGal to hundreds of μGal depending on the volume of injection and depth of injection. An estimate of other sources of the gravity changes (e.g., water table, atmospheric changes) in the area of measurements is required to model injected mass precisely. Surface gravity is also used as a monitoring tool for geothermal reservoirs and helps in better understanding long-term sustainability of the resource (Sofyan et al. 2011; Oka et al. 2012; Hinderer et al. 2015b; Portier et al. 2018).

Summary

Measurements and modeling of time-varying gravity changes allow a better understanding of various phenomena, as discussed in the earlier sections. Isolating the individual contributions from the aggregate effect in the temporal series of gravity changes has been a major challenge over a long time. Nevertheless, as number of observations and complementary information of temporal gravity changes are increasing, different components of time series of gravity changes can be better separated. Above mentioned examples and discussion suggest that time-lapse or continuous surface gravity measurements can be utilized for a wide variety of geophysical processes that deform the Earth's surface and move mass inside, on, and above the Earth's surface causing the gravity field to vary with time. Surface measurements are also vital as they could provide a test, or validation, for satellite-derived temporal gravity changes.

Cross-References

- Gravimeters
- Gravity Field of the Earth
- Gravity Field, Temporal Variations from Space Techniques
- Gravity Measurements, Absolute
- Gravity Method, Surface

Bibliography

- Alnes H, Eiken O, Stenvold T (2008) Monitoring gas production and CO_2 injection at the Sleipner field using time-lapse gravimetry. *Geophysics* 73:WA155
- Arora BR, Kamal, Kumar A, Rawat G, Kumar N, Choubey VM (2008) First observations of free oscillations from Indian superconducting gravimeter in Himalaya. *Curr Sci* 95:1611–1617
- Boy J-P, Gegout P, Hinderer J (2002) Reduction of surface gravity data from global atmospheric pressure loading. *Geophys J Int* 149:534–545
- Calvo M, Rosat S, Hinderer J (2016) Tidal spectroscopy from a long record of superconducting gravimeters in Strasbourg (France). In:

- International Association of Geodesy symposia. Springer, Berlin. https://doi.org/10.1007/1345_2016_223
- Carbone D, Poland MP, Diamant M, Greco F (2017) The added value of time-variable microgravimetry to the understanding of how volcanoes work. *Earth Sci Rev* 169:146–179. <https://doi.org/10.1016/j.earscirev.2017.04.014>
- Cazenave A et al (2011) Gravity field of the, temporal variations from satellite. In: Gupta HK (ed) *Encyclopedia of solid Earth geophysics*. Springer, Dordrecht, p 96
- Chen YT, Gu HD, Lu ZX (1979) Variations of gravity before and after the Haicheng quake, 1975, and the Tangshan quake, 1976. *Phys Earth Planet Inter* 18:330–338
- Coughlin M, Harms J (2014) Constraining the gravitational wave energy density of the Universe using Earth's ring. *Phys Rev D* 90(4). <https://doi.org/10.1103/physrevd.90.042005>
- Creutzfeldt B, Güntner A, Vorogushyn S, Merz B (2010) The benefits of gravimeter observations for modelling water storage changes at the field scale. *Hydrol Earth Syst Sci* 14(9):1715–1730. <https://doi.org/10.5194/hess-14-1715-2010>
- Crossley D, Hinderer J (2005) Using SG arrays for hydrology in comparison with GRACE satellite data, with extension to seismic and volcanic hazards. *Korean J Remote Sens* 21:31–49
- Crossley D, Jensen OG, Hinderer J (1995) Effective barometric admittance and gravity residuals. *Phys Earth Planet Inter* 90:221–241
- Crossley D, Hinderer J, Casula G, Francis O, Hsu H-T, Imanishi Y, Jentzsch G, Kaarianen J, Merriam J, Meurers B, Neumeyer J, Richter B, Shibuya K, Sato T, van Dam T (1999) Network of superconducting gravimeters benefits a number of disciplines. *EOS Trans AGU* 80(1):121, 125–126
- Crossley D, de Linage C, Hinderer J, Boy J-P, Famiglietti J (2012) A comparison of the gravity field over Central Europe from superconducting gravimeters, GRACE and global hydrological models, using EOF analysis. *Geophys J Int* 189:877–897. <https://doi.org/10.1111/j.1365-246X.2012.05404.x>
- Crossley D, Hinderer J, Riccardi U (2013) The measurement of surface gravity. *Rep Prog Phys* 76. <https://doi.org/10.1088/0034-4885/76/4/046101>
- de Linage C et al (2009) Separation of coseismic and postseismic gravity changes for the 2004 Sumatra-Andaman earthquake from 4.6 yr of GRACE observations and modelling of the coseismic change by normal-modes summation. *Geophys J Int* 176:695–714
- Ducarme B (2011) Determination of the main Lunar waves generated by the third degree tidal potential and validity of the corresponding body tides models. *J Geod*. <https://doi.org/10.1007/s00190-011-0492-9>
- Fores B, Champollion C, Le Moigne N, Bayer R, Chéry J (2016) Assessing the precision of the iGrav superconducting gravimeter for hydrological models and karstic hydrological process identification. *Geophys J Int* 208(1):269–280. <https://doi.org/10.1093/gji/ggw396>
- Fukuda Y, Sato T (1997) Gravity effects of sea level variation at the superconducting gravimeter sites, estimated from ERS-1 and Topex-Poseidon altimeter data. In: Segawa H et al (eds) *Gravity, geoid, and marine geodesy, JAG symposia*, vol 117. Springer, Berlin, pp 107–114
- Han SC, Shum CK, Bevis M, Kuo CY (2006) Crustal dilatation observed by GRACE after the 2004 Sumatra-Andaman quake. *Science* 313:658–662. <https://doi.org/10.1126/science.1128661>
- Hare JL, Ferguson JF, Carlos LV, Aikenand Brady JL (1999) The 4-D microgravity method for water flood surveillance: a model study for the Prudhoe Bay reservoir, Alaska. *Geophysics* 64:78–87
- Hector B, Séguis L, Hinderer J, Cohard J-M, Wubda M, Descloîtres M, Benarrosh N, Boy J-P (2015) Water storage changes as a marker for baseflow generation process in a tropical humid basement catchment: insights from hybrid gravimetry. *Water Resour Res* 51. <https://doi.org/10.1002/2014WR015773>
- Hinderer J, Crossley D (2000) Time variations in gravity and inferences on the Earth's structure and dynamics. *Surv Geophys* 21:1–45
- Hinderer J, Andersen O, Lemoine F, Crossley D, Boy J-P (2006) Seasonal changes in the European gravity field from GRACE: a

- comparison with superconducting gravimeters and hydrology model predictions. *J Geodyn* 41:59–68
- Hinderer J et al (2012) Land water storage changes from ground and space geodesy: first results from the GHYRAF (Gravity and Hydrology in Africa) experiment. *Pure Appl Geophys* 169(8):1391–1410. <https://doi.org/10.1007/s00024-011-0417-9>
- Hinderer J, Hector B, Boy J-P, Riccardi U, Rosat S, Calvo M, Littel F (2014) A search for atmospheric effects on gravity at different time and space scales. *J Geodyn* 80:50–57. <https://doi.org/10.1016/j.jog.2014.02.001>
- Hinderer J, Crossley D, Warburton R (2015a) Superconducting gravimetry. In: Herring T (ed), Schubert G (ed in chief) *Treatise on geophysics*, vol 3 Geodesy, 2nd edn. Elsevier, Amsterdam, pp 66–122
- Hinderer J, Calvo M, Abdelfettah Y, Hector B, Riccardi U, Ferhat G, Bernard J-D (2015b) Monitoring of a geothermal reservoir by hybrid gravimetry; feasibility study applied to the Soultz-sous-Forêts and Rittershoffen sites in the Rhine graben. *Geotherm Energy* 3:16. <https://doi.org/10.1186/s40517-015-0035-3>
- Hinderer J, Hector B, Mémin A, Calvo M (2016) Hybrid gravimetry as a tool to monitor surface and underground mass changes. In: *International Association of Geodesy symposia*. Springer, Berlin. https://doi.org/10.1007/1345_2016_253
- Horowitz CJ, Widmer-Schmid R (2020) Gravimeter search for compact dark matter objects moving in the earth. *Phys Rev Lett* 124(5). <https://doi.org/10.1103/physrevlett.124.051102>
- Imanishi Y, Sato Y, Higashi T, Sun W, Okubo S (2004) A network of superconducting gravimeters detects submicrogal coseismic gravity changes. *Science* 306:476–478
- Jacob T, Rohmer J, Manceau J-C (2015) Using surface and borehole time-lapse gravity to monitor CO₂ in saline aquifers: a numerical feasibility study. *Greenhouse Gases Sci Technol* 6(1):34–54. <https://doi.org/10.1002/ggh.1532>
- Jousset P, Dwipa S, Beauducel F, Duquesnoy T, Diamant M (2000) Temporal gravity at Merapi during the 1993–1995 crisis: an insight into the dynamical behaviour of volcanoes. *J Volcanol Geotherm Res* 100:289–320
- Kennedy J, Ferré TPA, Güntner A, Abe M, Creutzfeldt B (2014) Direct measurement of subsurface mass change using the variable baseline gravity gradient method. *Geophys Res Lett* 41(8):2827–2834. <https://doi.org/10.1002/2014gl059673>
- Klügel T, Wziontek H (2009) Correcting gravimeters and tiltmeters for atmospheric mass attraction using operational weather models. *J Geodyn* 48(3–5):204–210
- Kroner C, Jentzsch G (1999) Comparison of different barometric pressure reductions for gravity data and resulting consequences. *Phys Earth Planet Inter* 115:205–218
- Kroner C, Thomas M, Dobslaw H, Abe M, Weise A (2009) Seasonal effects of non-tidal oceanic mass shifts in observations with superconducting gravimeters. *J Geodyn* 48(3):354–359
- Lambert A, Courtier N, Sasagawa GS, Klopping F, Winster D, James TS, Liord JO (2001) New constraints on Laurentide postglacial rebound from absolute gravity measurements. *Geophys Res Lett* 28:2109–2112
- Lambert A, Courtier N, James TS (2006) Long-term monitoring by absolute gravimetry: tides to postglacial rebound. *J Geodyn* 41:307–317
- Mémin A, Spada G, Boy J-P, Rogister Y, Hinderer J (2014) Decadal geodetic variations in Ny-Alesund (Svalbard): role of past and present ice-mass changes. *Geophys J Int* 198(1):285–297. <https://doi.org/10.1093/gji/ggu134>
- Montagner J-P, Juhel K, Barsuglia M, Ampuero J-P, Chassande-Mottin E, Harms J, Whiting B, Bernard P, Clévéde E, Lognonné P (2016) Prompt gravity signal due to the 2011 Tohoku-Oki earthquake. *Nat Commun* 7:13349. <https://doi.org/10.1038/ncomms13349>
- Naujoks M, Kroner C, Weise A, Jahr T, Krause P, Eisner S (2010) Evaluating local hydrological modelling by temporal gravity observations and a gravimetric three-dimensional model. *Geophys J Int*. <https://doi.org/10.1111/j.1365-246x.2010.04615.x>
- Nawa K, Suda N, Fukao Y, Sato T, Aoyama Y, Shibuya K (1998) Incessant excitation of the Earth's free oscillations. *Earth Planets Space* 50:3–8
- Niebauer TM, Sasagawa GS, Faller JE, Hilt R, Klopping F (1995) A new generation of absolute gravimeters. *Metrologia* 32:159–180
- NRC (1997) In: Dickney JO et al (eds) *Satellite gravity and the geosphere: National Research Council Report*. National Academy Press, Washington, DC, p 112
- Oka D, Fujimitsu Y, Nishijima J, Fukuda Y, Taniguchi M (2012) Evaluation of geothermal reservoir mass change from the gravity change at the Takigami geothermal area, Oita prefecture, Japan. In: *Proceedings of the thirty-seventh workshop on geothermal reservoir engineering*. Stanford University, Stanford, 30 Jan–1 Feb 2012, SGP-TR-194
- Panet I, Bonvalot S, Narteau C, Rémy D, Lemoine JM (2018) Migrating pattern of deformation prior to the Tohoku-Oki earthquake revealed by GRACE data. *Nat Geosci* 11(5):367–373. <https://doi.org/10.1038/s41561-018-0099-3>
- Portier N, Hinderer J, Riccardi U, Ferhat G, Calvo M, Abdelfettah Y, Bernard J-D (2018) Hybrid gravimetry monitoring of Soultz-sous-Forêts and Rittershoffen geothermal sites (Alsace, France). *Geothermics* 76:201–219. <https://doi.org/10.1016/j.geothermics.2018.07.008>
- Rosat S, Hinderer J (2018) Limits of detection of gravimetric signals on Earth. *Sci Rep* 8(1). <https://doi.org/10.1038/s41598-018-33717-z>
- Rosat S, Hinderer J, Crossley D, Boy J-P (2004) Performance of superconducting gravimeters from long-period seismology to tides. *J Geodyn* 38:461–476
- Sato T, Okuno J, Hinderer J, MacMillan DS, Plag H-P, Francis O, Falk R, Fukuda Y (2006) A geophysical interpretation of the secular displacement and gravity rates observed at Ny-Alesund, Svalbard in the Arctic – effects of the post-glacial rebound and present-day ice melting. *Geophys J Int* 165:729–743
- Sato T, Miura S, Sun W, Sugano T, Freymueller JT, Larsen CF, ... Motyka RJ (2012) Gravity and uplift rates observed in southeast Alaska and their comparison with GIA model predictions. *J Geophys Res Solid Earth* 117(B1). <https://doi.org/10.1029/2011jb008485>
- Sofyan Y, Kamah Y, Nishijima J, Fujimitsu Y, Ehara S, Fukuda Y, Taniguchi M (2011) Mass variation in outcome to high production activity in Kamojang geothermal field, Indonesia: a reservoir monitoring with relative and absolute gravimetry. *Earth Planets Space* 63(11):1157–1167
- Sugihara M, Nawa K, Nishi Y, Ishido T, Soma N (2013) Continuous gravity monitoring for CO₂ geo-sequestration. *Energy Procedia* 37:4302–4309
- Tapley BD et al (2019) Contributions of GRACE to understanding climate change. *Nat Clim Chang* 5(5):358–369
- Tiwari VM, Singh B, Vyaghreswara Rao MBS, Mishra DC (2006) Absolute gravity measurements in India and Antarctica. *Curr Sci* 91:686–689
- Tiwari VM, Wahr JM, Gross R, Swenson S (2008) GRACE-observed gravity changes in areas of large quakes. *GRACE science meeting*, San Francisco, 11–12 Dec 2008
- Tiwari VM, Wahr JM, Swenson S (2009) Dwindling groundwater resources in northern Indian region, from satellite gravity observations. *Geophys Res Lett* 36:L18401. <https://doi.org/10.1029/2009GL039401>
- Tiwari VM, Srinivas N, Singh B (2014) Hydrological changes and vertical crustal deformation in South India: inference from GRACE, GPS and absolute gravity data. *Phys Earth Planet Inter* 231:74–80
- Van Camp M, de Viron O, Watlet A, Meurers B, Francis O, Caudron C (2017) Geophysics from terrestrial time-variable gravity

- measurements. *Rev Geophys* 55:938–992. <https://doi.org/10.1002/2017RG000566>
- van Dam TM, Larson K, Wahr JM, Gross S, Francis O (2000) Using GPS and gravity to infer ice mass changes in Greenland. *EOS Trans AGU* 81(37):421, 426–427
- Virtanen H, Mäkinen J (2003) The effect of the Baltic Sea level on gravity at the Metsähovi station. *J Geodyn* 35(4–5):553–565. [https://doi.org/10.1016/s0264-3707\(03\)00014-0](https://doi.org/10.1016/s0264-3707(03)00014-0)
- Voigt C, Förste C, Wziontek H, Crossley D, Meurers B, Pálinská V, Hinderer J, Boy J-P, Barriot J-P, Sun H (2016) Report on the data base of the International Geodynamics and Earth Tide Service (IGETS), scientific technical report (STR) – data; 16/08. GFZ German Research Centre for Geosciences, Potsdam. <https://doi.org/10.2312/GFZ.b103-16087>
- Wahr JM (2007) Variable gravity from satellite. In: Schubert G (ed in chief) *Treatise on geophysics*, vol 3. Elsevier, Amsterdam, pp 213–238
- Warburton RJ, Goodkind JM (1976) Search for evidence of a preferred reference frame. *Astrophys J* 208(Pt 1):881–886
- Weise A, Kroner C, Abe M et al (2012) Tackling mass redistribution phenomena by time-dependent GRACE and terrestrial gravity observations. *J Geodyn* 59–60:82–91
- Wenzel HG (1996) The nanogal software: tide data processing package ETERNA 3.3. *Bull Inf Marées Terr* 135:9425–9439
- Ziegler Y, Hinderer J, Rogister Y, Rosat S (2016) Estimation of the gravimetric pole tide using a stacking of long series of GGP superconducting gravimeters. *Geophys J Int* 205:77–88. <https://doi.org/10.1093/gji/ggw007>

Gravity Measurements, Absolute

Mark A. Zumberge

Institute of Geophysics and Planetary Physics, Scripps
Institution of Oceanography, University of California, San
Diego, La Jolla, CA, USA

Synonyms

Absolute acceleration due to gravity; Absolute *g*

Definition

Gravity	The mutually attractive force between any two masses.
Acceleration due to gravity, <i>g</i>	The value of the acceleration of a mass due to the force of gravity observed in a reference frame rotating with Earth.
Absolute	Measured with calibrated instruments traceable to metrologic standards. For the case of gravity, this requires length and time standards.
Absolute gravity meter or absolute gravimeter	An instrument used to measure gravity absolutely, traceable to time and length standards.

Introduction

The value of the acceleration due to gravity on Earth's surface, *g*, depends on position and time. A worldwide variability of roughly 0.5% from 9.8 m/s² is caused by Earth rotation, variations in density, and variations in geometry caused both by tidal fluctuations and tectonic deformation. Studying how *g* varies can reveal important facts about Earth as a whole and about a long list of local phenomena ranging from volcanic activity to evolution of oil and gas reservoirs.

Often scientists are interested only in changes in gravity with either position (gravity maps) or with time (time-lapse gravity). Many of these studies can be completed with simple, relative gravity meters which commonly consist of an electromechanical spring-mass sensor (Torge 1989). Relative gravity measurements (i.e., the difference in *g* between two sites) are more common because the interesting variations in *g* are small and its absolute value is not important for most geophysical applications. Almost all relative gravity meters undergo drift, however, making it difficult to separate real temporal gravity changes from drift in the sensor. Drift can be calculated and accounted for in cases where it is feasible to make redundant observations at multiple sites and assume gravity values at reference sites are constant. In a number of important cases, however, such drift compensating efforts are inadequate and absolute measurements are needed – they are inherently free from drift because of their reliance on length and time standards.

A note on units: practitioners in the field of gravity measurement use the unit *Gal* (after Galileo) which is defined as 1 cm/s². The nominal value of *g* is around 980 Gal. More common usage is the mGal (=10^{−3} Gal) and the μGal (=10^{−6} Gal or about 10^{−9} g). A precision of a few μGal is available from the best relative gravity meters (following a significant drift compensation effort within the measurement process) and 1 or 2 μGal with the best absolute gravity meters. Absolute determinations are characterized by both precision (the smallest observable signal or change) and accuracy (level of uncertainty or the discrepancy between the measured value and the unknown true value).

History

Late in the seventeenth century, it was discovered that a pendulum would swing with a slightly different period if it were moved to a different latitude. This effect, caused by gravity's dependence on latitude (primarily because of Earth rotation, causing both a centrifugal term and an elliptical shape to the planet as a whole), was used to make some of the earliest estimates of the deviation of Earth's figure from a sphere (Lenzen and Multhaus 1965). Two kinds of pendulum gravity meters, the reversible pendulum and the invariable

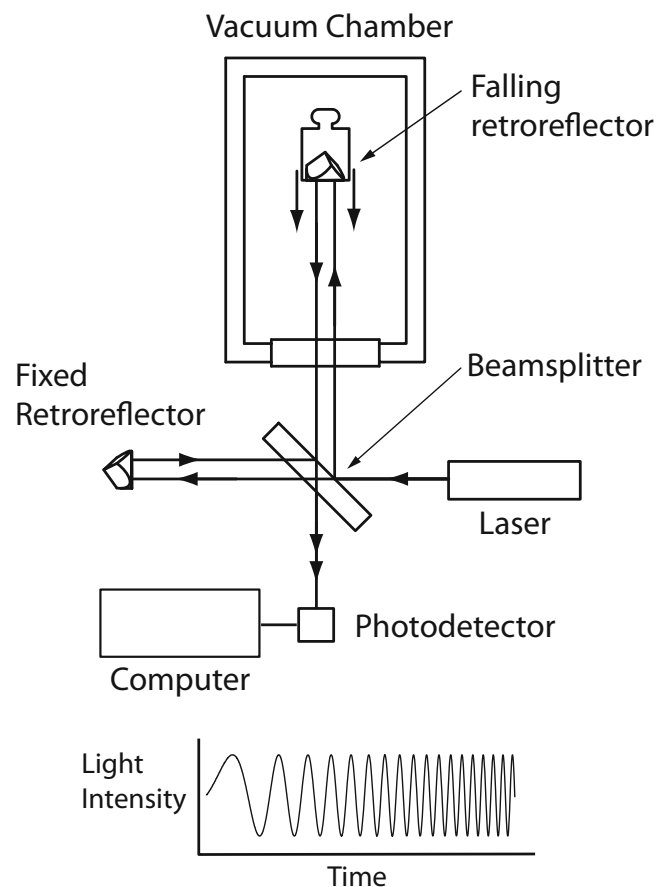
pendulum, provide a good example of the difference between absolute and relative sensors. A reversible pendulum, used extensively by Henry Kater in the nineteenth century, exploits the curious fact that two pivot points on an asymmetric pendulum which are found to yield equal oscillation periods are separated by the length of a simple pendulum (a point mass suspended from a massless rod) having that same period. This allows an absolute calculation of gravity based on the measurement of the pendulum period (referenced to a time standard) and the distance between the two oscillation points (referenced to a length standard) which have a common period, and is therefore an absolute measurement. Invariable pendulums, on the other hand, were simply swung around a single point and changes in gravity were observed as changes in the period. Pendulum measurements were made extensively by geodesists to refine the figure of the Earth and for mineral exploration up until the early twentieth century when electronic methods became adequately advanced to measure accurately the time needed for a mass in a vacuum chamber to fall a known distance. Cook (1965) and Faller (1963) pioneered some of the early free-fall measurement techniques. Towards the end of the twentieth century, the advent of the laser allowed for great improvements in absolute gravity measurement (Hammond and Faller 1967; Alasia et al. 1982; Zumbege et al. 1982; Niebauer et al. 1995).

Now, in the beginning of the twenty-first century, atomic methods are becoming viable. The most commonly used method today is to determine the acceleration of a mass freely falling in a vacuum by means of laser interferometry. Generally this has been done with macroscopic objects; however, this may be gradually replaced with a method using an ensemble of ultra-cold atoms, whose acceleration as they fall freely in a vacuum chamber is determined with matter-wave interferometry.

Principles of Modern Measurements

Macroscopic Falling Test Mass

Figure 1 is a schematic diagram of the mainstay of absolute gravity measurements since the mid-to-late twentieth century. A mechanical system inside a high vacuum chamber effects the release of a test mass. A retro-reflecting prism mounted to the test mass returns laser light entering and exiting the chamber through a window in the bottom. A beamsplitter (or half-silvered mirror) combines this reflected light with that similarly reflected from a fixed retro-reflector. The two reflected beams interfere with one another, evermore rapidly producing successive light and dark fringes as the falling mass accelerates. Two cycles of light-to-dark interference fringes are completed for each advance in the fall of a distance equaling the laser's wavelength. A photodetector senses the passage of interference fringes and generates an electrical



Gravity Measurements, Absolute, Fig. 1 A schematic diagram of a free-fall absolute gravity meter. A mechanical system (not shown) inside the vacuum chamber releases and catches the falling object. An interferometer determines the position of the test mass as it falls; the interference fringe signal is an increasing frequency sine wave

signal from which the local value of gravity (the rate at which the falling mass accelerates) can be calculated through appropriate timing analysis performed by a computer. Typically, the mass falls a distance of 10–20 cm and a thousand or more drops over a period of a day provide a statistical uncertainty of a few μGal . Because the measurement is based on calibrated length and time standards – the wavelength of the laser light and the frequency of the timing oscillator – the measurement is absolute.

Microscopic Falling Test Masses

A new and promising method for the absolute determination of gravity makes use of cold atom interferometry (Kasevich 2002). Toward the end of the twentieth century, physicists learned to use lasers to trap and manipulate clouds of atoms. Atoms can now be confined in an evacuated magneto-optical trap, cooled to very low temperatures with laser cooling, and then dropped or even launched vertically (again, with lasers) to create an “atomic fountain.” Exploiting the fact that all matter can be described as waves, atomic physicists have

learned to interfere atomic wave functions with one another and, in so doing, perform a number of highly precise measurements using them. In particular, because the phase of an atom's wave function depends on gravity, interfering atoms that have traversed separate paths through a gravitational field can yield an absolute measure of gravity. The first experiments performed by Kasevitch and Chu (1992) and followed by others (Peters et al. 2001; Lempore et al. 2008) required a laboratory full of complex optical, vacuum, and electronic equipment. Recently, however, the technology has advanced to a state where portability has been achieved (Merlet et al. 2010; Fang et al. 2016). A commercial version of a cold atom interferometric absolute gravity meter is currently available from Muquans (Talence, France).

Instrumentation

The mechanical details of conventional free-fall apparatuses vary. In some, the test mass's fall is initiated by a small elevator inside the vacuum chamber, which also surrounds the mass as it falls and acts as a shield against minute non-gravitational forces. In others, the mechanical system resides at the bottom of the vacuum chamber and launches the test mass upward, providing a mass in free flight, the acceleration of which can be measured during both its rise and fall. Another important variation is the use of a spring suspension system to isolate the reference reflector from vibrations.

Currently the most widely used absolute gravity meter is the model FG5, manufactured by Micro-g LaCoste (Lafayette, Colorado, USA) and shown in Fig. 2. With laser interferometry, it tracks the descent of a 300 g mass during a 15 cm drop. The portable system weighs 320 kg and can produce an absolute gravity value accurate to 2 μGal after an hour of data collection. Significant care has been taken by its developers to ensure that systematic errors – those whose signs do not vary and cannot be reduced by averaging – are less than the quoted accuracy. Such errors include forces on the falling mass other than gravity (e.g., residual air-drag and magnetic or electrostatic forces) and imperfections in the optical and timing systems. Periodically, many absolute gravity meters make measurements simultaneously in one of the world's standards lab for an intercomparison to test the assessments of the instruments' accuracies (Sasagawa et al. 1995).

The cold-atom interferometry method that is on the steep portion of the learning curve is schematically represented in Fig. 3. In this embodiment of the technology, developed by Muquans, a single laser cools and traps the atoms, then probes them with matter-wave interferometry during their free fall. A laser, tuned to a hyperfine transition in the falling atoms' electronic spectrum, is ramped in frequency to match the increasing Doppler shift experienced by the falling atoms as they accelerate under gravity. The rate that the probing laser's

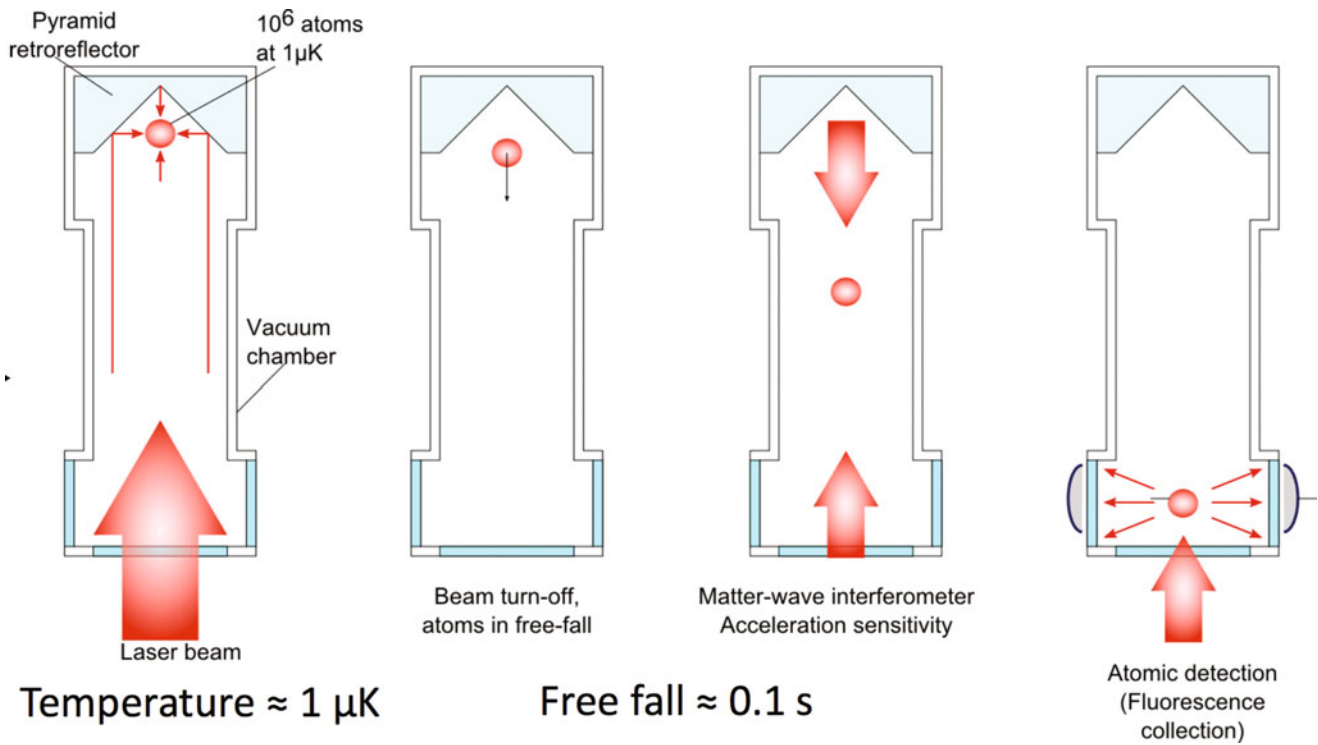


Gravity Measurements, Absolute, Fig. 2 The micro-g LaCoste FG5 is a commonly used portable absolute gravity meter

frequency must be changed to exactly match the Doppler shift is controlled by the local value of gravity. Cold atom interferometry provides an extremely sensitive way to determine this rate of frequency shift. The system is smaller in size (100 kg) and similar in cost to a macroscopic free-fall apparatus, and it has no moving parts and can collect data at a rate of two samples per second. Figure 4 is a photograph of the Muquans device.

Applications to Geophysics

Near Earth's surface, the rate of change of gravity with altitude is about 3 μGal per cm. Therefore a detection of a gravity change of a few μGal yields information on cm-level height changes. Observing gravity changes together with height information obtained by other means can provide important information for modeling the causes of surface deformation. An increase in density of 0.1 g/cm^3 occurring throughout a 25 cm thick layer below the observer (such as might be caused by water saturation) increases gravity by 1 μGal . These changes can occur very slowly which makes it desirable to use absolute measurements to detect such changes. Examples of long time series of absolute gravity used to study geophysical phenomena include observations of postglacial rebound (Lambert et al. 2001), water flooding in a natural gas reservoir (Hare et al. 1999), vertical land movement near tide-gauges (Teferle et al. 2006), and observations of coseismic offsets from an earthquake (Tanaka et al. 2001).



Gravity Measurements, Absolute, Fig. 3 In a cold-atom interferometer gravity meter, a cloud of atoms falls freely while being probed with a tunable laser (after Muquans)

Gravity Measurements, Absolute, Fig. 4 A commercially available quantum gravity meter made by Muquans (Talence, France)



A very important geophysical application of absolute gravity measurement is the establishment of calibration lines for relative gravity meters (Sasagawa et al. 1989). Absolute g measurements are also necessary in certain metrology applications, including piston-cylinder pressure calibrations and in defining the standard kilogram (Robinson and Schlamminger 2016).

Summary

Absolute gravity measurements are much less common than those made with relative gravity meters. However, because of their freedom from drift, made possible through the use of modern atomic length and time standards, absolute g meters are extremely important for monitoring long term changes associated with crustal deformation and density change and for calibrating relative meters. New advances in atomic physics may result in an absolute gravity meter which is similar in size to typical relative gravity meters, advancing the use of absolute gravity measurements in geophysics.

Bibliography

- Alasia F, Cannizzo L, Cerutti G, Marson I (1982) Absolute gravity acceleration measurements: experiences with a transportable gravimeter. *Metrologia* 18:221
- Cook AH (1965) The absolute determination of the acceleration due to gravity. *Metrologia* 1:84
- Faller JE (1963) An absolute interferometric determination of the acceleration of gravity. PhD thesis, Princeton University, Princeton
- Fang B, Dutta I, Gillot P, Savoie D, Lautier J, Cheng B, ... Landragin A (2016) Metrology with atom interferometry: inertial sensors from laboratory to field applications. *J Phys: Conf Ser* 723(1):012049. IOP Publishing
- Hammond JA, Faller JE (1967) Laser-interferometer system for the determination of the acceleration of gravity. *IEEE J Quantum Electron QE-3*:597
- Hare JL, Ferguson JF, Aiken CLV, Brady JL (1999) The 4-D microgravity method for waterflood surveillance: a model study for the Prudhoe Bay reservoir, Alaska. *Geophysics* 64:78
- Kasevich MA (2002) Coherence with atoms. *Science* 298:1363
- Kasevitch M, Chu S (1992) Measurement of the gravitational acceleration of an atom with a light-pulse atom interferometer. *Appl Phys B* 54:321
- Lambert A, Courtier N, Sasagawa GS, Klopping F, Winester D, James TS, Liard JO (2001) New constraints on Laurentide postglacial rebound from absolute gravity measurements. *Geophys Res Lett* 28:2109
- Lamporesi G, Bertoldi A, Cacciapuoti L, Prevedelli M, Tino GM (2008) Determination of the Newtonian gravitational constant using atom interferometry. *Phys Rev Lett* 100:050801
- Lenzen VF, Multhaus RP (1965) Development of gravity pendulums in the 19th century. *U.S. Natl Museum Bull* 240:301
- Merlet S, Bodart Q, Malossi N, Landragin A, Pereira Dos Santos F, Gitlein O, Timmen L (2010) Comparison between two mobile absolute gravimeters: optical versus atomic interferometers. *Metrologia* 47:L9–L11
- Niebauer T, Sasagawa G, Faller J, Hilt R, Klopping F (1995) A new generation of absolute gravimeters. *Metrologia* 32:159
- Peters A, Chung KY, Chu S (2001) High-precision gravity measurements using atom interferometry. *Metrologia* 38:25
- Robinson IA, Schlamminger S (2016) The watt or kibble balance: a technique for implementing the new SI definition of the unit of mass. *Metrologia* 53(5):A46
- Sasagawa G, Zumberge M, Stevenson J, Lautzenhiser T, Wirtz J, Ander M (1989) The 1987 southeastern Alaska gravity calibration range: absolute and relative gravity measurements. *J Geophys Res* 94:7661
- Sasagawa G, Klopping F, van Dam T (1995) Analysis of data from the 1994 international comparison of absolute gravimeters with a single computational protocol. *Metrologia* 32:192
- Tanaka Y, Okubo S, Machida M, Kimura I, Kosuge T (2001) First detection of absolute gravity change caused by earthquake. *Geophys Res Lett* 28:2979
- Teferle FN, Bingley RM, Williams SDP, Baker TF, Dodson AH (2006) Using continuous GPS and absolute gravity to separate vertical land movements and changes in sea-level at tide-gauges in the UK. *Philos Trans R Soc* 364:917
- Torge W (1989) *Gravimetry*. de Gruyter, Berlin
- Zumberge MA, Rinker RL, Faller JE (1982) A portable apparatus for absolute measurements of the Earth's gravity. *Metrologia* 18(3):145

Gravity Method, Airborne

Uwe Meyer

Sub-Department Geophysical Reconnaissance – Resources and Near Surface Processes, Federal Agency for Geosciences and Mineral Resources, Hannover, Germany

Synonyms

Aerogravity; Airborne gravimetry

Definition

Airborne	Anything carried by a fixed wing aircraft, a helicopter, an airship, or an unmanned aircraft.
Aero	Anything related to the atmosphere above ground.
Gravity	In this context nonrelativistic mass attraction as defined by Newton.
Gravimetry	In this context the measurement of the gravity field variation by mass attraction.

Introduction

The law of gravity is one the most fundamental in natural sciences. Instruments for gravity measurements have been developed first for use on land surface, later for use on ships. The instruments used for maritime surveying needed a special platform to compensate the movements of the vessel that add additional components to the main value of the

measurement: the acceleration of a well-defined mass due the gravity field of the Earth. The same accounts for airborne applications of gravity meters: The aircraft's variations in roll, pitch, and yaw angles and height changes must be either compensated for or measured for later reductions and corrections. Therefore, most instruments used in the early stages of airborne gravity meter development were basically upgraded marine gravity meters (LaCoste 1967). Due to the problem of the fast aircraft movements and therefore the even larger accelerations implied, only after the practical implementation of a global satellite-based positioning system (as GPS) non-inertial navigation became precise enough to enable satisfactory corrections and reductions (Bruton et al. 1999). At present, three different categories of airborne gravity instruments evolved: scalar gravity meters, vector gravity meters, and gradient or full tensor meters. All instruments are used to measure the regional variation of the Earth's gravity field over often inaccessible areas in a quick fashion.

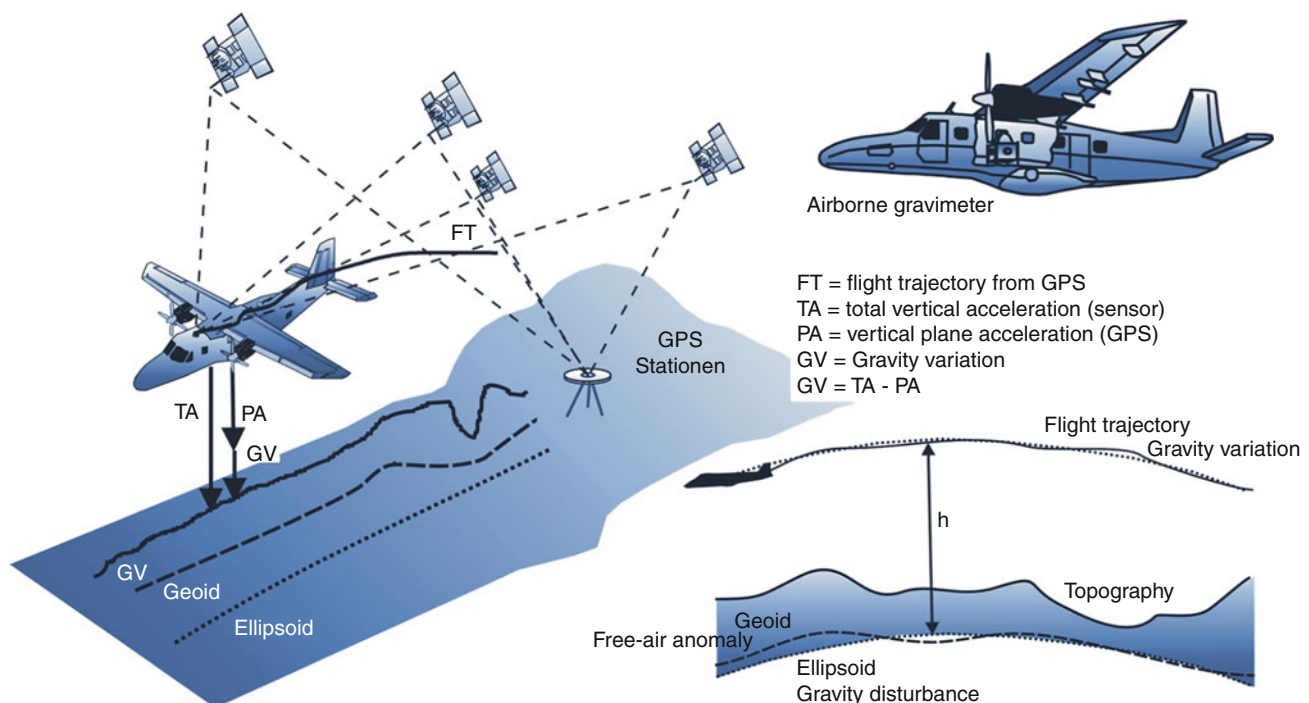
Principle Idea of Airborne Gravity

In the still most common airborne gravity applications, scalar gravity meters are used. Such meters usually measure only the relative vertical component of the Earth's gravity field using a mass coupled to a spring, a principle that was adopted from gravity land-based surveying (Fig. 1). In airborne gravity, much larger accelerations than the desired signal are introduced by attitude and height changes of the aircraft, carrying

the gravity meter. Beyond these inertial accelerations, vibrations occur that are usually damped by suitable mechanical or electromechanical means. The residual effect of the vibrations in the aircraft's cabin is usually dealt with as a noise signal. Therefore, we focus on the gravitational and inertial accelerations only. The instrument and the navigation equipment are placed in a rigid aircraft that defines the body frame. Within this body frame, we primarily measure the accelerations in the axes of the aircraft: forward, down, and sideward in a right-hand system. The gravity sensor is fixed on a platform that compensates some of the aircrafts forward and sideward accelerations that are coupled to attitude variations but not the vertical acceleration components. The main aim of the platform is to keep the gravity sensor in horizontal level in order to keep the measurement range and sensitivity of the gravity meter high. Additional navigation equipment as global positioning system antennas and inertial measurement units monitor the movements of the aircraft's body frame in the geodetic frame. Consequently, lever arms between the navigation measurement units and the gravity sensor must be corrected for in the body frame and the resulting accelerations and force must be transferred to the geodetic reference frame for later gravity mapping.

Attitude Compensation Platforms

Most platforms designed to carry a gravity sensor on a moving vehicle are two axes gimbal constructions to compensate



Gravity Method, Airborne, Fig. 1 Principle of airborne gravity

coupled roll, pitch, and yaw variations (LaCoste et al. 1967). The attitude information is commonly acquired by a gyro for short wavelengths and an accelerometer for wavelengths up to some minutes for each directional component. Some platforms compensate heading variations as well. Such platforms are Schuler-tuned. The platforms are designed to compensate attitude changes almost instantaneously in the dynamic range of the gravity sensor. Therefore, special filters are applied that define the platform period, which is usually in the order of some minutes.

Scalar Gravity Meters

Most meters in use for airborne gravity surveying are based on a test mass coupled to a spring or on torsion balance. Some new developments utilize superconductive gravity sensing systems. Such instruments measure the relative changes of the main vertical component of the Earth's gravity field. Beyond the residual platform errors, an independent non-inertial vertical acceleration determination is needed to reduce for the vertical accelerations of the body frame. This is possible only since global satellite-based positioning system navigation data is available. Scalar systems generally have a very low drift rate and therefore can be used for long wavelength gravity mapping as required for geodesy. The shortest detectable wavelengths are dependent on flight velocity, the dynamic range of the gravity sensor, and the platform period. Generally, some few mGal ($1 \text{ mGal} = 10^{-5} \text{ m/s}^2$) are achieved in accuracy related to some few kilometers wavelength resolution at an average speed of about 160 km/h above ground.

Vector Gravity Meters

Vector gravity meters are in most cases adapted inertial measurement units that measure the three vertical components of the Earth's gravity field in the body frame of the aircraft (Jekeli 2001). Analogue to most attitude-compensating platform designs, the signal is gathered from a combination of directional gyro and accelerometer measurements. In this configuration, global satellite-based positioning system navigation data must be applied for all directional acceleration reductions. The drift rates of such instruments are usually higher than those of scalar meters. At present, only very few systems are available, that are not only able to measure the significant vertical component but also the horizontal components in sufficient accuracy and resolution. Most vector gravity meters are not installed on an attitude-compensating platform but are placed on dampers only. The resolution and accuracy for the vertical component of available vector systems is better than those of scalar spring tension or torsion balance meters.

Gravity Gradient Meters

In order to enhance the short wavelengths resolution and the measurement accuracy significantly, gravity gradient meters were developed (Heiland 1940). The main idea is to eliminate the impact of the aircrafts attitudes and accelerations in the measurements by measuring the directional gradients directly. To achieve this, diagonally arranged pairs of highly accurate accelerometers are placed on a slowly rotating disk, one for each direction. Since the two accelerometers measure in opposite directions, the largest effects of the body frame accelerations are ruled out by the acceleration difference between the two accelerometers and mainly the directional gravity gradient remains. Such gradients are not stable enough for large-scale integration and therefore are only suitable for short wavelength detections in local to small-scale regional areas. This sensing system needs an exceptionally accurate attitude compensation platform. If all directional gradient components are measured, this system is called a full tensor gradient gravity meter. Available resolutions are in the order of 0.5–1 Eotvoes ($1 \text{ Eotvoes (E)} = 0.1 \text{ mGal/km} = 10^{-9} \text{ s}^{-2}$).

Processing, Reductions, and Corrections

The original data set is the sum of all measured accelerations in one or more components within the body frame of the aircraft. In order to get the gravity variation on flight level, the directional components of the body frame accelerations in a chosen geodetic frame have to be reduced for. As a next step, an Eotvoes correction has to be applied to compensate for the gravitational impact due to varying centrifugal forces (Harlan 1968). Next, a platform error correction has to be applied, usually a Harrison correction for scalar meters.

The gravity disturbance (gravity variation on ellipsoidal level) is determined by the normal gravity reduction and the free-air gravity reduction between observation level and ellipsoid.

In order to achieve the gravity anomaly, the same is done with reference to the geoid instead of the ellipsoid.

Applications and Aircraft

Airborne gravity can either be used to determine the regional geoid more precisely or for geophysical reconnaissance and prospection. For geodetic purposes rather high altitude straight and level fixed wing aircraft flights utilizing scalar gravity meters are most common (Forsberg et al. 1996). In geophysics, two aspects occur: large to mid-scale reconnaissance surveys to get an overview on tectonic or geological settings and small-scale prospection in search for mineral deposits or similar tasks (Veryaskin and McRae 2008). For reconnaissance, fixed wing aircraft with long-range capabilities are preferred. In prospection, high-resolution vector or gradient gravity meters are

used on small fixed wing aircraft, helicopters, or airships. First attempts to use unmanned airborne vehicles (UAV) are under way. New instrumental developments try to utilize modern superconductive meters, absolute gravity meters, or devices based on cold atom interferometry.

Summary

Airborne gravity has become a practical and accurate method to measure the variations of the Earth's gravitational field over regional areas. In geodesy, fixed wing aircraft equipped with scalar gravity meters are preferred. In geophysics, the same configuration is often used for reconnaissance surveys. For aerogeophysical prospection, low- and slow-flying aircraft are favored combined with vector or gradient gravity meters.

Cross-References

- [Gravity Measurements, Absolute](#)
- [Gravity Method, Principles](#)
- [Gravity, Gradiometry](#)

Bibliography

- Bruton AM, Glennie CL, Schwarz KP (1999) Differentiation for high precession GPS-velocity and acceleration determination. *GPS Solutions* 2(4):7–21
- Forsberg R, Hehl K, Bastos L, Gidskehaug A, Meyer U (1996) Development of a geoid mapping system for coastal oceanography (AGMASCO). In: *Gravity, Geoid and Marine Geodesy*, IAG symposium series, vol 117. Springer, Berlin, pp 163–170
- Harlan RB (1968) Eotvos corrections for airborne gravimetry. *J Geophys Res* 73(14):4675–4679
- Heiland CA (1940) *Geophysical exploration*. Prentice-Hall, New York
- Jekeli CH (2001) *Inertial navigation systems with geodetic applications*. Walter de Gruyter, Berlin
- LaCoste LJB (1967) Measurement of gravity at sea and in the air. *Rev Geophys* 5(4):477–526
- LaCoste LJB, Clarkson N, Hamilton G (1967) LaCoste & Romberg stabilized platform shipboard gravity meter. *Geophysics* 32:99–109
- Veryaskin A, McRae W (2008) On combined gravity gradient components modelling for applied geophysics. *J Geophys Eng* 5:348

Gravity Method, Principles

Hans-Jürgen Götze
Institute of Geosciences, Working group Satellite- and Aerogeophysics, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

Synonyms

Gravimetry methods; Gravity and figure of the Earth

Definition

Gravity methods focus on the determination of the gravitational field of the Earth's body and its anomalies and shape (determination of the figure of the Earth). The figure of the planet is essentially characterized by two forces: attraction and centrifugal force. Gravitation is a general property of all matter and is independent of the chemical composition of mass particles. The mathematical description of the gravitational field goes back to Sir I. Newton. The article describes the formulation of the law, the ambiguity in the interpretation, as well as the problems with the inversion of measured gravitational fields.

Introduction

By gravimetry (Latin “gravis”) methods are identified, which can be used to measure the gravity field of the Earth. The determination of this potential field is of greater importance for Geodesy, Geophysics, and Geotechnics. In particular in Geophysics and Geodynamics, gravity is used to investigate the Earth's interior and its dynamic processes. Furthermore precise balances and other technical equipment need gravity data for their calibration.

The main objective of gravity/gravimetry is the determination of the Earth's gravity field and of other celestial bodies (e.g., the Earth's moon). The gravity field and its gravity gradients are measured as a function of position and time on, below, and above the Earth's surface, e.g., by gravity meters on ships, in boreholes, or on board of aircrafts and satellites: CHAMP, GRACE, GOCE, and GRACE-FO mission (refer among many others to Dahle et al. 2014).

In the seventeenth and eighteenth century, scientific interest and first measurements have been caused by the understanding and development of the Physics of deformable and rigid bodies. At that time determinations of position (distances) and time became accurate enough to enable gravity measurements. Later, after scientists discovered the position dependency of gravity, it has been used to determine the figure of the Earth – then until today. Developments of the gravity method were always correlated with the interaction of technical possibilities and scientific objectives in Physics (Geophysics) and land surveying (later then Geodesy) because geodetic surveying and geophysical exploration became a strong influence on growing societies (e.g., Torge 1989). Today the gravity field and its variations – measured by satellites with very high and unexpected accuracy – are closely related with the distribution and transport of masses in the Earth's system: the ocean circulation, changes of ground-water level and solid Earth moisture, melting of continental ice sheets, river runoff, changes of sea level, and mass flows in the Earth lithosphere and mantle, all of this causes transport and redistribution of masses. Using innovative and extremely

precise gravity, sensor systems dedicate gravity field to observe the implication of these processes. Even in the investigation of other celestial bodies, gravity gains in importance: already in the 1970s, the “Lunar surface gravity meter” was deployed at the Apollo 17 ALSEP site (Talwani 2003). Later NASA’s “Gravity Recovery and Interior Laboratory (GRAIL) Mission” measured the moon’s gravity field in unprecedented detail (e.g., Zuber et al. 2013; Wimsberger et al. 2018). Gravity field models of other celestial bodies (Mars and Venus) are available at the web page of the “International Centre for Global Earth Models – ICGEM” (<http://icgem.gfz-potsdam.de/home>) for download and visualization.

In the following chapters, all values of parameters of the Earth’s gravity field are related to the World Geodetic System 1984 (WGS 84).

Potential Function

It is accepted that the force $\mathbf{F}$ of a mass point m , which is influenced by a timely constant gravity acceleration $\mathbf{g}$, can be expressed by

$$\mathbf{F} = m\mathbf{g}.$$

To move this mass along a path C with path elements $d\mathbf{r}$ from a space point P_1 to a second point P_2 , an energy U_{12} has to be applied:

$$U_{12} = - \int_{P_1}^{P_2} \mathbf{F} \cdot d\mathbf{r} = -m \int_{P_1}^{P_2} \mathbf{g} \cdot d\mathbf{r} \quad (1)$$

There is no loss of energy, and we call U_{12} independent from its path and only dependent from start and end point P_1 respective P_2 . Equation (1) shows that only the difference of potential energy at points P_1 and P_2 can be calculated. The force $\mathbf{F}$ can be derived from the potential energy as

$$\mathbf{F} = -\text{grad } U = -\nabla U \quad (2)$$

$\mathbf{F}$ is a potential field. Equation (2) holds even if we add a constant U_0 . This constant will be selected in a way that potential energy disappears in the infinite. In this case $U(P)$ defines the energy in the field of gravity acceleration which has to be applied to m to move it from point P to the infinite:

$$U(P) = - \int_P^{\infty} \mathbf{F} \cdot d\mathbf{r}.$$

In Geophysics and Geodesy, the potential energy U is replaced by the potential (potential function):

$$V = -U/m. \quad (3)$$

This is the negative potential energy per mass unit and with (2) one has

$$\mathbf{g} = \nabla V.$$

Therefore, a gravity field can be described in entire space by a single function V . The potential function fulfills the Poisson differential equation:

$$\Delta V = -4\pi f \rho \quad (4)$$

with $f = (6.67430 \pm 0.00015) \cdot 10^{-11} \text{ m}^3/\text{kg s}^2$ (gravitational constant; see below)

ρ = mass density

The Poisson equation forms the base of potential theory and is an elliptic differential equation. If space is mass-free ($\rho = 0$), the Laplace equation holds in Cartesian coordinates:

$$\Delta V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0 \quad (5)$$

It became the most universal differential equation in the physical sciences because of the wide range of phenomena that is described by it. There exists a multitude of solutions of this differential equation which are not always potential functions. For both differential equations, hold the following definition:

A space function and its second derivatives which are continuous and bounded in a region where defined are called harmonic functions if they fulfill the LAPLACE and – in the interior of a mass filled space – the POISSON differential equation.

Newton’s Law and Newtonian Potential

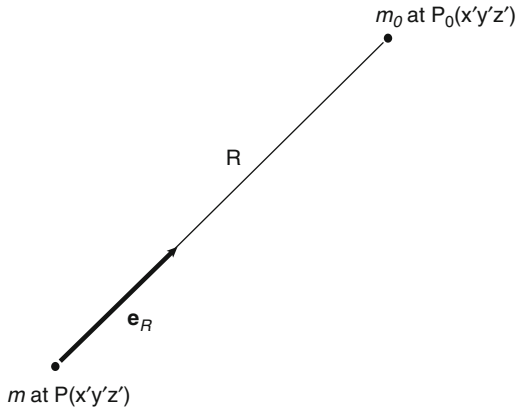
In 1687 Sir Isaac Newton published his famous treatise *Philosophiae Naturalis Principia Mathematica* which contains Newton’s law of gravitational attraction:

The magnitude of a gravitational force between two masses is proportional to each mass and inversely proportional to the square of their distance.

In a Cartesian coordinate system, a mass m_0 is centered at point $P_0(x'y'z')$ and at point $P(xyz)$ a mass m . Then a force acts on mass m_0 (Fig. 1):

$$\mathbf{F}(P) = -f \frac{mm_0}{R^2} \mathbf{e}_R \quad (6)$$

with



Gravity Method, Principles, Fig. 1 Masses m_0 and m experience a gravitational force which follows Newton's law: it is proportional to m_0 and m and $1/r^2$. By convention the unit vector $\mathbf{e}_R$ is directed from the gravitational source point (P) to the "observation" point P_0 which is located at the mass m_0

$\mathbf{R} = \overrightarrow{P_0P}$ (vector which points from P_0 to P)

$\mathbf{e}_R = \mathbf{R}/R$ (unit vector) of $\mathbf{R}$

$R = [(x - x')^2 + (y - y')^2 + (z - z')^2]^{1/2}$ (norm of vector $\mathbf{R}$)

The gravity acceleration $\mathbf{g}$ which is caused by mass m is

$$\mathbf{g}(P) = \frac{\mathbf{F}(P)}{m_0} = f \frac{m}{R^2} \mathbf{e}_R \quad (7)$$

In Cartesian coordinates one can write

$$\mathbf{g}(P) = -\frac{f m}{R^3} [(x - x')\mathbf{i} + (y - y')\mathbf{j} + (z - z')\mathbf{k}] \quad (8)$$

$\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are components of the unit vector in a coordinate system.

At the Earth's surface, gravity acceleration is approximately $g = 9.81 \text{ m/s}^2$. The potential V is a scalar and an additive quantity. Therefore, the gravity potential of n point masses can be written as

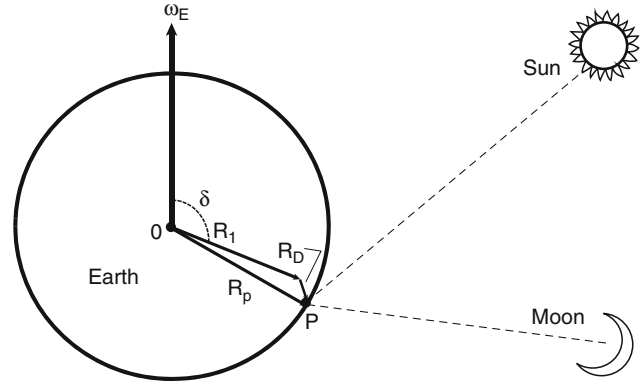
$$V(P) = f \sum_{i=1}^n \frac{m_i}{R_i}, \text{ with } R_i \text{ (spaces of the } n \text{ points } P)$$

At the point masses, $V(P)$ is not defined due to $R_i = 0$. However, except for these singular points, the Laplace equation holds.

For a mass continuum, we obtain from Eq. (6) the Earth's gravitation by Newton's law of gravitation:

$$\mathbf{b}(\mathbf{r}) = f \iiint_{\text{Earth}} \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|^3} dm$$

where $\mathbf{r}'$ and $\mathbf{r}$ are the geocentric position vectors of a mass element dm and the attracted point P (unit mass = 1) and



Gravity Method, Principles, Fig. 2 To illustrate the calculation of gravity effects in a point P at the Earth's surface, which depend on the attraction of celestial masses (sun and moon) and the centrifugal acceleration (δ = pol distance, ω_E = angular velocity vector)

$|\mathbf{r}' - \mathbf{r}| = R$. The mass element dm can be expressed by the volume density $\rho = \rho(\mathbf{r}')$ and the volume element $dm = \rho dv$:

$$\mathbf{b} = \text{grad } V$$

For the Earth's gravitational potential with v = volume of the Earth, we have

$$V(\mathbf{r}) = f \iiint_{\text{Earth}} \frac{\rho(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|} dv \quad (9)$$

Besides the attraction of the Earth at points P, generally all cosmologic masses contribute to this attraction. However, in most cases only the influence of sun and moon is considered. The entire gravity attraction can be obtained by

$$\nabla V(P) = \mathbf{a}_E(P) + \mathbf{a}_C(P) \quad (10)$$

$\mathbf{a}_E$ (attraction of Earth)

$\mathbf{a}_C$ (attraction of cosmological masses)

$\mathbf{a}_E$ causes time-dependent deformations of Earth by tidal effects, and each mass point P is dislocated by a very small value R_D . Each mass point is influenced by the rotation of Earth, and its angular velocity is given by $\omega_E = 7.292115 \cdot 10^{-5} \text{ rad/s}$.

In an Earth-centered coordinate system, the position vector $\mathbf{R}_P$ which points to point P is given by (refer also to Fig. 2)

$$\mathbf{R}_P = \mathbf{R}_1 + \mathbf{R}_D$$

Then, the gravity vector $\mathbf{g}(P)$ can be calculated by $\mathbf{g}(P) = \nabla V(P) - \ddot{\mathbf{R}}_P$. It holds

$$\dot{\mathbf{R}}_P = \dot{\mathbf{R}}_D + \dot{\mathbf{R}}_1 = \dot{\mathbf{R}}_D + (\omega_E \times \mathbf{R}_1)$$

and

$$\ddot{\mathbf{R}}_P = \ddot{\mathbf{R}}_D + (\omega_E \times (\omega_E \times \mathbf{R}_1))$$

disregarding the small changes of rotational velocity $\dot{\omega}_E$. The gravity vector can be written as:

$$\mathbf{g}(P) = \mathbf{a}_E(P) + \mathbf{a}_C(P) - \ddot{\mathbf{R}}_D + (\omega_E \times (\omega_E \times \mathbf{R}_1)) \quad (11)$$

The first term represents the attraction of Earth, the latter the centrifugal acceleration which is caused by Earth rotation.

The middle terms $\mathbf{a}_C(P) - \ddot{\mathbf{R}}_D$ display the time-dependent part of Earth tides. For precise gravity field determinations, it must be observed and eliminated from the observations. The tidal correction which is caused by sun and moon never exceeds $0.3 \times 10^{-5} \text{ m/s}^2$.

For the centrifugal acceleration $\mathbf{a}_z$, we can write

$$\begin{aligned} \mathbf{a}_z &= (\omega_E \times \mathbf{R}_1) \times \omega_E = \omega_E^2 \mathbf{R}_1 - (\omega_E^2 \cdot \mathbf{R}_1) \omega_E \\ &= \omega_E^2 \cdot \mathbf{R}_\perp \end{aligned} \quad (12)$$

with

$\mathbf{R}_\perp$ = to the Earth's rotation axis perpendicularly oriented-component of $\mathbf{R}_1$.

Because $\mathbf{R}_\perp = 1/2 \nabla R_\perp^2$, also $\mathbf{a}_z$ is a gradient field:

$$\mathbf{a}_z = \nabla Z \text{ and } Z = 1/2 \omega_E^2 R_\perp^2$$

Z is called the potential of the centrifugal acceleration, and one obtains for the complete potential of the Earth with (9):

$$W = V + Z \text{ and } \mathbf{g} = \nabla W. \quad (13)$$

Anomalies

From Eq. (13) the normal gravity field (γ_0) can be deduced from a "normal Earth body" (among others, e.g., Heiskanen

and Moritz 1967; Torge 1989; Blakely 1995). Geodesists and Geophysicists agreed that the "normal gravity" is defined at an equipotential surface which is part of a spheroidal surface that bounds a rotating, uniformly dense Earth. Due to the gravity (attraction) and rotation forces, such a spheroid has a very comparable shape to an ellipse of rotation. Consequently, it is called a reference ellipsoid: in the WGS84 geodetic system, the flattening of the ellipsoid is $f = 1/298.257223563$. Readers wishing additional information are referred to publications by Li and Götze (2001) and Hackney and Featherstone (2003).

Local variations in altitudes and densities of crust and mantle and geological formations cause variations in the Earth's gravitational field, known as gravitational anomalies. Some of these anomalies can be very extensive and cause even bulges in sea level. The study of these anomalies forms the basis of the gravity discipline in Geophysics (see *Gravity, Data to Anomalies*). The gravity field which forms these anomalies is measured with extremely sensitive gravimeters and today detected also with gradiometers (Götze and Pail 2018; Difrancesco 2007). The effect of topography and other known factors (see below) is subtracted from the observations. Such techniques are used by Geophysicists to explore oil and mineral deposits and global structures. Denser rocks (often containing mineral ores) cause higher than normal local gravitational fields on the Earth's surface; less dense sedimentary rocks cause negative anomalies. The separation of gravity anomalies which are caused by local density variations of rocks and geological formations implicates corrections of the observed gravity. It was suggested by Blakely (1995) to describe these corrections as contributions to the field which was measured:

Observed gravity = attraction of the reference ellipsoid (normal gravity)
 +effect of heights above geoid (FREE – AIR correction),
 +effect of topographic masses above geoid (BOUGUER and TOPOGRAPHIC correction),
 +effect of time – dependent variations (TIDAL correction),
 +effect of moving platforms (EOTVOS corrections),
 +effect of masses that support loads (ISOSTATIC corrections),
 +effect of both known and unknown crust/mantle density variations (GEOLOGY).

It is the goal of each interpretation to isolate the effect of geological bodies from all other terms. We have to consider that the magnitude of this effect is relatively small if compared with the value of normal gravity:

Normal gravity approximately is 9.81 m/s^2 and

Gravity effects of "Geology" approximately are $10^{-5} \text{ m/s}^2 - 10^{-3} \text{ m/s}^2$

This is less than 0.01% of the observed gravity field magnitude. In Geophysics one distinguishes the standard anomalies: free air, Bouguer and Isostatic, explained in the entry *Gravity Data to Anomalies* in this encyclopedia.

The interpretation of the physical meaning of Bouguer anomalies is an old problem associated with various and often different assumptions and simplifications introduced

for the calculation of height and mass correction terms. It is also related to the purpose the anomaly is used for in Geodesy or Geophysics. However, one has to emphasize that Bouguer anomalies do not represent the gravity effect of anomalous density at a horizontal reference level. A comprehensive, modern presentation of all problems related to the normal gravity correction has been summarized by Pašteka et al. (2017) in order to eliminate any misunderstandings.

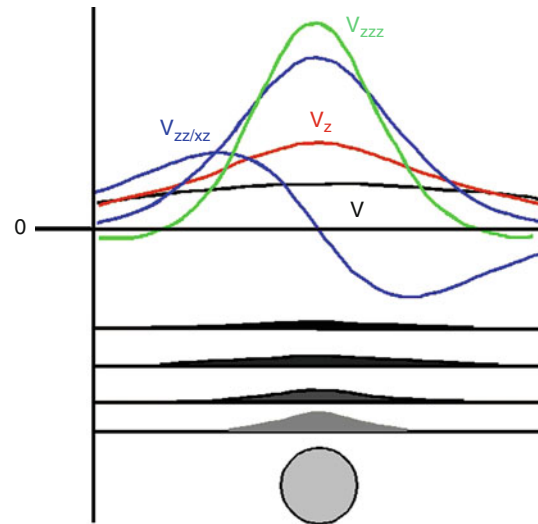
Ambiguity and Principle of Equivalence

The interpretation of gravity anomalies is not unique, and boundary conditions are always required, mainly from other geophysical disciplines and/or geology. It can be easily shown by modeling that for a given anomaly and a given density contrast, a wide range of possible interpretations can be made: at various depths, based upon different geometrical shapes. Nor does the method of interpretation by the gravity field gradients allow us to make a unique interpretation or to distinguish deep from shallow anomalies as has been claimed: it has been shown that we cannot escape the ambiguity by using second derivative quantities (gradients) or curvature and that, in fact, gravity and its derivatives are related by a corollary of Green's theorem. This theorem provides an analytical proof of ambiguity not only for the case of gravity data but also for magnetic data as well (e.g., Skeels 1947).

The Inversion Problem

The ambiguity of the inverse problem of gravimetry originates from potential theory. It is related with the computation of density functions from the observed gravity field. Density functions yield information of location, shape, and densities of causing masses. The computation leads to an integral equation, for which no unique solution exists. For a gravity anomaly generated by a particular mass distribution, infinitely many equivalent disturbing masses can be constructed at depths above the maximum depth of a disturbing body (Fig. 3). However, the total mass of the disturbing body (surface integral of the gravity anomaly, the lateral position of its mass center, and the corresponding coating on the surface of the Earth (special case of an equivalent disturbing mass) can be determined uniquely. If certain conditions are met, limiting values of maximal depth, density (difference), thickness, and lateral extension of the disturbing body can be found. The necessary additional independent information (constraints) is taken from other geophysical methods, geological, and/or petrological investigations.

In practical solutions of inversion problems, the *direct problem* of gravity appears, i.e., the computation of gravitational effects caused by particular mass contributions given in terms of location, shape, and density. In this case a unique



Gravity Method, Principles, Fig. 3 Gravitational potential (V), vertical gravity component, and derivatives of gravity over a homogeneous sphere with a depth of 5 km, a radius of 1 km, a density contrast of 1 kg/m^3 , and the equivalent masses between surface and maximum depth (after Torge 1989)

solution is always possible by application of the law of gravitation, Eq. (9).

Units and Gravitational Constant

Gravity is measured in ms^{-2} which is the standard unit in the “Système International d’Unités” (SI units are used in Physics since 1972). The six independent components of the gravity gradients are measured in s^{-2} . For many applications of gravity field, the following units are useful:

$$\begin{aligned} 1 \text{ nm s}^{-2} &= 10^{-9} \text{ m s}^{-2} \text{ or} \\ 1 \mu\text{m s}^{-2} &= 10^{-6} \text{ m s}^{-2}. \end{aligned}$$

Also the older cgm – units are still used in Geophysics and Geodesy:

$$\begin{aligned} 1 \text{ Gal} &= 1 \text{ cm s}^{-2} \\ 1 \text{ mGal} &= 10^{-5} \text{ m s}^{-2} \text{ or} \\ 1 \mu\text{Gal} &= 10^{-8} \text{ m s}^{-2}. \end{aligned}$$

“Gal” was (is) used in honor of the famous Renaissance Physicist Galileo Galilei (1564–1642). Although these older cgm units are no longer admitted, in Applied Geophysics they are still in use, together with “gravity units” ($\text{g.u.} = 1 \mu\text{m}^{-2}$).

Measurements and refinements of the gravitational constant have progressed since the time of Newton. However, up to now, laboratory experiments and physical research have not rendered a significant or generally accepted proof of $\dot{f} \neq 0$. The actual, recently published value (CODATA 2019) of the

constant is $f = 6.674\,30 \times 10^{-11} \text{ m}^3/\text{kg s}^2$, with a standard uncertainty of $\pm 0.00015 \times 10^{-11} \text{ m}^3/\text{kg s}^2$ and a relative standard uncertainty of 0.2×10^{-5} by the Committee on Data of the International Council for Science.

Summary

This contribution describes the principles of the gravity method which was developed by Sir Isaac Newton in the seventeenth century and holds until today. The theory bases on the definition of a potential function which replaces potential energy. A space function and its second derivatives which are continuous and bounded in a region where defined are called harmonic functions if they fulfill the Laplace and – in the interior of a mass filled space – the Poisson differential equation. A normal gravity field can be deduced from a “normal Earth body,” and Geodesists and Geophysicists agreed that a “normal gravity” is defined at an equipotential surface which is part of a spheroidal surface that bounds a rotating, uniformly dense Earth. Due to the gravity (attraction) and rotation forces, such a spheroid has a very comparable shape to an ellipse of rotation. Consequently, it is called a reference ellipsoid. Local variations in altitudes and densities of the Earth’s crust and mantle and geological formations cause variations in the Earth’s gravitational field, known as gravitational anomalies. In Geophysics elevation is used in all the corrections including the height correction and the complete Bouguer correction. In addition one should correct observed gravity for the geoid shape. The gravity effects due to the geoid undulations (N) are called the “indirect effects.” Any gravity field interpretation is effected by the ambiguity of the inverse problem of gravimetry which originates from potential theory; it is related with the computation of density functions from the observed gravity field.

Cross-References

- Geodesy, Figure of the Earth
- Gravity Anomalies, Interpretation
- Gravity, Data to Anomalies
- Gravity Field of the Earth
- Gravity Method, Surface
- Gravity Method, Satellite
- Gravity Modeling, Theory and Computation
- Gravity, Data to Anomalies
- Inverse Theory, Global Optimization
- Isostasy

Bibliography

Blakely R (1995) Potential theory in gravity and magnetic applications. Cambridge University Press, Cambridge

- CODATA (2019). <https://physics.nist.gov/cgi-bin/cuu/Value?bg>. Accessed 18 June 2019
- Dahle C, Flechtner F, König R, Michalak G, Neumayer KH, Gruber C, König D (2014) GFZ RL05: an improved time-series of monthly GRACE gravity field solutions. In: Flechtner F, Sneeuw N, Schuh W-D (eds) Observation of the system earth from space – CHAMP, GRACE, GOCE and future missions. GEOTECHNOLOGIEN science report, Advanced technologies in earth sciences, vol 20. Springer, Berlin, pp 29–39. <https://doi.org/10.1007/978-3-642-32135-1>
- Difrancesco D (2007) Advances and challenges in the development and deployment of gravity gradiometer systems. EGM 2007 International workshop, innovation in EM, gravity and magnetic methods: a new perspective for exploration, Capri, Italy, 15–18 April 2007
- Fixler JB, Foster GT, McGuirk JM, Kasevich MA (2003) Atom interferometer measurement of the Newtonian constant of gravity. Science 315(5808):74. <https://doi.org/10.1126/science.1135459>
- Götze HJ, Pail R (2018) Insights from recent gravity satellite missions in the density structure of continental margins – with focus on the passive margins of the South Atlantic, Gondwana Research. Focus Rev 53:285–308
- Hackney RI, Featherstone WE (2003) Geodetic versus geophysical perspectives of the ‘gravity anomaly’. Geophys J Int 154:35
- Heiskanen WA, Moritz H (1967) Physical geodesy. Freeman, San Francisco
- Li X, Götze HJ (2001) Ellipsoid, geoid, gravity, geodesy and geophysics – a tutorial. Geophysics 66(6):1660
- Pašteka R, Mikuška J, Meurers B (2017) Understanding the Bouguer anomaly: a gravimetry puzzle. Elsevier, Amsterdam
- Schmidt S, Götze HJ (2002) Gravity data processing. In: Flüh ER, Kopp H, Schreckenberger B (eds) Cruise Report SO161–1 & 4, subduction processes off Chile. Research Center for Marine Geosciences at Christian-Albrechts-Universität-Kiel, GEOMAR Report 102
- Skeels DC (1947) Ambiguity in gravity interpretation. Geophysics 12:43. <https://doi.org/10.1190/1.1437295>
- Talwani M (2003) The Apollo 17 gravity measurements on the moon. Lead Edge 22(8):786. <https://doi.org/10.1190/1.1605083>
- Torge W (1989) Gravimetry. de Gruyter Berlin, New York
- Vening-Meinesz FA (1931) Une nouvelle méthode pour la réduction isostasique régionale de l’intensité de la pesanteur. Bull Géod 29:33–51
- Wirsberger H, Krauss S, and Mayer-Gürr, T (2018) First independent Graz Lunar Gravity Field Model derived from GRAIL. Submitted to Icarus 2018_244
- Zuber MT, Smith DE, Watkins MM, Asmar SW, Konopliv AS, Lemoine FG, Melosh HJ, Neumann GA, Phillips RJ, Solomon SC, Wieczorek MA, Williams JG, Goossens SJ, Kruizinga G, Mazarico E, Park RS, Yuan DN (2013) Gravity field of the moon from the gravity recovery and interior laboratory (GRAIL) Mission. Science 339:668–671. <https://doi.org/10.1126/science.1231507>

Gravity Method, Satellite

G. Balmino

Centre National d’Etudes Spatiales, Toulouse, France

Definition

The determination of the gravitational potential of the Earth by space methods consists in computing the spherical harmonic coefficients of a truncated representation of the

external potential, from observations of artificial satellite orbits close enough to the Earth's surface, from ground stations or from other satellites (generally of higher altitudes), and also from special measurements made onboard a satellite.

Introduction

The knowledge of the gravity field of the Earth and of an associated reference surface of the altitudes (the geoid) is necessary for geodesy, for improving theories of the physics of our planet interior and surface, including various mass transfer mechanisms, and for modeling the ocean circulation in absolute. This knowledge comes from several observing techniques; on the global scale, it mostly benefited from the observation of artificial satellite orbits and has been improved over decades. The current approach, based on the accumulation of information from a large number of satellites, one being observed at a time, reached its limits around year 2000. Dramatic improvements now come from newly implemented space techniques already proposed in the 1970s and rendered possible thanks to technological progress; they were exemplified in the 2000 and 2010 decades by the CHAMP, GRACE then GRACE-FO, and GOCE missions which are realizations of the satellite-to-satellite (SST) tracking principles (in the high-low and low-low modes), of satellite gradiometry, together with the systematic use of micro-accelerometry to measure the surface forces which are a source of uncertainty in gravity field mapping from satellites.

Geodesy, Gravity Field, and Scientific Applications: Why?

Geodesy is a science which major goal is to define and study the shape of the Earth, its deformations (from relative movements of points on its surface), and its gravitational field (or its gravity field if one includes the rotational effects). It is an old discipline which acquired its fame in the eighteenth century when French scientists went measuring arcs of meridians, in Lapland and Peru, in order to determine whether the Earth was flattened at the poles or at the equator. Two hundred years later, the use of artificial satellites yielded a global and continuous view of our planet, and geodesy, too, benefited from this revolution. It became a science at the crossroads of others; for instance it is used daily for navigating satellites with high precision for oceanography, geophysics, and a variety of societal applications. In the mid-1960s first global models of the Earth's gravity field from satellites initiated a revolution in geophysics from space; about 40 years later we are witnessing a new era which sees the deciphering of our planet's gravity with unprecedented resolution and accuracy.

For a better understanding of both surface and deep phenomena which are related to the interior structure of the Earth and its temporal evolution, for studying the dynamics of the oceans and their interaction with meteorological and climate changes, for modeling the ice caps-oceans-continent relationships and for predicting the long-term evolution of the mean sea level, and also for unifying vertical reference systems and for the precise determination of satellite orbits in space, it is mandatory to know our planet gravity field to ever-increasing resolution and precision – including its temporal variations induced by the aforementioned phenomena. Details are provided by Jekeli (*this encyclopedia*); we limit ourselves to qualitative arguments on the dual role of the gravity field in Earth sciences.

On the one hand, by comparing the real field with the field of an idealized body (e.g., an ellipsoid in hydrostatic equilibrium), one defines gravity anomalies which characterize deviations from a state of internal equilibrium (i.e., nonradial density variations), which constitutes one method of sounding our planet's interior. The magnitude of gravity anomalies is in the range $-1,000/+1,000$ mgals ($1 \text{ mgal} = 10^{-5} \text{ ms}^{-2}$). As a matter of fact, this is one of four ways of “looking” inside, the other three being seismology, magnetic field studies, and the analysis of the Earth's rotation. However the gravity analysis approach is unique in providing direct information on the density field (although this is integrated and cannot be univocally inverted). When one strips the Earth from its topographic blocks (with assumed or known density), one is left with residual gravity anomalies which magnitude is not much different from the magnitude of the original ones; this exemplifies the phenomenon of isostasy, which is the concept of mechanisms of support of the topographic masses. Therefore the residual field mirrors mass excesses or deficits in depth which inform us on the lateral structure of the lithosphere and upper mantle. A significant part of this information is to-day provided by seismology and by tomographic analysis, but still suffers from uncertainties due to hypothesis on which such inversions are based. The combination of both types of data (seismic velocities and gravity anomalies) is a powerful tool to get a better picture of the interior and to make progress in the understanding of several phenomena, for instance, the accumulation of stresses and triggering of earthquakes.

On the other hand, a surface which is intimately related to gravity, the geoid, is used as the reference for defining and measuring the altitudes on the continents but also under the oceans or over the ice caps. The geoid is a particular equipotential surface of the gravity potential (the sum of the gravitational and centrifugal potentials) which may be viewed, in oceanic areas, as the surface of an ocean at rest. The Earth's rotation being sufficiently well-known and its effects being smooth at the surface (they essentially decrease with increasing latitudes), the irregularities of the geoid (in the ± 100 m range, as measured with respect to an ellipsoid of revolution

which approximates the Earth's shape) characterize the density field variations in a way similar to the gravity anomalies. In addition, on the topographic surface, water naturally flows along the geoid (or other equipotential) slope; over the oceans, seawater circulates (under the forcings of winds and density-salinity fluctuations), and these movements (different from the vertical displacements of water due to tides) and associated amounts of transported water (and heat, chemicals, nutrients, etc.) can be quantified with respect to the geoid: these are the ocean currents. One distinguishes between the mean and the variable parts of the circulation: the latter can nowadays be monitored by satellite altimetry, whereas the former (which is needed in the modeling of the climate by quantifying the long-term absolute transports of heat – especially in the shallow parts of the oceans) requires the precise knowledge of the geoid; the difference between the mean sea surface and the geoid is the ocean dynamic topography; the mean sea surface itself can be deduced from long series of altimeter measurements, but it is demanded that the geoid be independently determined.

For both solid Earth physics and oceanography, the requirements to-day are (at least) for a 100 km resolution geoid (and gravity field) with a cumulated error – i.e., up to that resolution, of 1 cm (and 1 mgal) or better. Clearly the classical domains of geodesy (unification of vertical reference systems, inertial navigation, accurate satellite orbits for monitoring the Earth's deformations, the sea surface, the Earth's kinematics, etc.) will greatly benefit from such knowledge.

In addition to this, there is a growing interest, in relationship with environmental studies and prediction capabilities, in measuring the time variations of the gravity field (see Cazenave, *this encyclopedia*). Some of them (e.g., the oceanic and solid tides) are deduced from the measurements or modeling of phenomena which are responsible for them, with satisfactory precision and resolution – or close to be satisfactory. On the contrary others can only be determined from direct observations, from which the responsible phenomena are then modeled. These are the postglacial rebound of the lithosphere (therefore of the planet surface); the variations of the ice caps, of the sea level, and of the large continental aquifers; and the distribution of precipitations (rain, snow), the soil moisture, and the evapotranspiration, that is, the water cycle in general.

Classical Ways of Measuring the Earth's Gravity Field

Up to a few years ago, three types of information have been used to describe or model, at local, regional, or/and global scales, the Earth's gravitational potential and various functionals:

- Surface measurements of gravity, on land, at sea (with ships and submarines), and by plane (more recently).

This is the oldest approach, much improved along its history. The measurements are often reduced to gravity anomalies and then gridded and made available at a variable resolution which depends on the area, on the actual density of observations, and also on restrictions still exerted by many countries for protecting national economic (sometimes military) interests. The resolution of such grids varies from a few to hundreds of kilometers, and the precision ranges from about 1 to 10 milligals – the accuracy may be worse due to systematic errors in many data sets (such as the old marine surveys). The data are archived at world level, mainly at the National Geospatial-Intelligence Agency (NGA) in the USA, and at the Bureau Gravimétrique International (BGI) in France. Airborne measurements are slowly expanding, but the precision at ground level (after downward continuation) is still insufficient. More information on gravimetry is provided in cross-references.

- The classical analysis of satellite orbit perturbations: since the 1960s a few groups in the world have determined global gravity field models from the inversion (via the law of the dynamics) of satellite orbit perturbations – which reflect the forces acting on a satellite. This has also been the current and very successful way of determining the gravity field of other solar system bodies (Nerem 1995). Most used Earth's satellites were indeed never designed for such purpose: the space geodesists simply took benefit of the more or less accurate measurements made from ground stations on these satellites for orbit knowledge. Various techniques were exploited, essentially the Doppler effect on the signal transmitted from a stable oscillator onboard a spacecraft to ground stations – or the reverse (case of the DORIS system), and roundtrip ground to satellite distances measured by laser systems – and optical determination of the satellite positions on the star background was also used at the beginning. The main limitations of these models come from the altitude of the observed satellites (due to the quasi-exponential decay of the perturbations with altitude), from the observation coverage (limited by the ground station networks and pass duration), and from uncertainties in modeling the surface forces (atmospheric drag, solar radiation pressure, and Earth albedo). If the data coverage has improved for some satellites thanks to the development of denser station networks (e.g., the DORIS system), or by tracking satellites from “above,” that is, by the GPS (Global Positioning System) constellation since the 1980s and presently by most GNSS (Global Navigation Satellite System) spacecraft, the other limitations remain and prevent from making significant progress. Solutions derived from satellite only data up to year ~2000 covered spatial scales from 500 to 40,000 km, but they were reliable (say at the centimeter level on the geoid knowledge) only down to a resolution of about 1000–1500 km. That is why they have

been – and are still combined with surface gravimetric measurements and altimetric data (see below) which bring higher resolution if not precision – depending on the area; such combined models globally reach 50 km resolution. The same strategy was applied with a much better satellite solution based on GRACE mission data and yielded the EGM 2008 model which has, in principle, a 10 km resolution though it represents the field with such details solely over areas with enough surface data. A detailed account of global gravity modeling history, techniques, and satellites used so far is given by Pavlis (*this encyclopedia*). In Sect. 5 we give an overview of the mathematical methods used to retrieve gravity field information from satellite dynamics.

- Satellite altimetry over the oceans: this is a space technique by which one measures (by radar or laser) the distance between a spacecraft and the sea surface, provided that the spacecraft position is known with sufficient precision it provides the sea surface which time average is close to the geoid but differs from it, the difference being the ocean dynamic topography (up to about 1 m in areas of strong currents such as the Gulf Stream, the Arctic Circum-polar Current, the Kuro-Shio). For several applications in marine geodesy and geophysics, it is sufficient to correct for the mean circulation with the present (approximate) models, though coastal currents, uncertainties in tidal models, and larger error measurements close to the shore limit the usage of the information to the offshore and open ocean areas. Such restriction being understood, satellite altimetry data have had an important contribution to global Earth's gravity models in increasing their resolution over the oceans – but without providing the true geoid.

New Satellite Missions

Basic principles were found years ago; technological advances have allowed for their implementation.

To get rid of the limitations of past systems, any new satellite approach should satisfy as many of the following criteria as possible, quasi-polar orbit (for the coverage), 3-D uninterrupted measurements (for recovering the gravity signals with an isotropic error), and mean altitude as low as possible (for higher resolution), counteract the effects of signal attenuation with altitude by a differencing measurement method, and isolate the gravitational signals by measuring and/or compensating the surface forces (which involves ultra-sensitive accelerometers). The missions launched during the 2000–2020 decades satisfy more or less these criteria. Working at low altitude is the constraint which has received most attention and all missions fly at or below 450 km (compared to the altitudes of satellites used so far: 800 km or larger). Surface forces are measured by micro-accelerometers (and even compensated in the case of the gradiometric mission). The other

criteria have been considered differently according to the type of mission and the used technology.

The first dedicated satellite mission, launched in 2000, was CHAMP (Challenging Mini-Satellite Payload for Geophysical Research and Application) (Reigber et al. 1996); its two main objectives were the mapping of the magnetic field and of the gravity field of the Earth. It was developed by Germany, with French contributions (magnetometers and a triaxial electrostatic micro-accelerometer) and the provision of a GPS receiver by the USA; this is one realization of the SST approach in the high-low mode (with a lower segment altitude of 450 km, decreasing to about 300 km toward the end of mission in 2010), which resulted in the first significant leap in gravity modeling precision.

CHAMP appears as the precursor of the second gravity mission GRACE (Gravity Recovery and Climate Experiment), which is a USA mission (JPL 1998) conducted in cooperation with Germany and launched in March 2002. It consisted in flying two satellites separated by about 200 km from each other, on the same mean polar orbit; the satellites are each tracked by the GPS (like CHAMP) and besides measure their relative distance and velocity with a micro-metric accuracy – this observable (of the low-low SST type) may be viewed as an extremely fine measurement of the gravitational potential difference at the satellites. Each spacecraft was equipped with a micro-accelerometer (of a class superior to the CHAMP one). The mission exceeded its designed lifetime (like CHAMP): designed for a 5-year duration, it operated until October 2017. Besides improving our knowledge of the mean gravity field (by more than one order of magnitude), it brought a wealth of information on the seasonal and inter-annual time variations of the gravity field, especially those related to the hydrological cycle and ice cap volume changes. Due to the success of this mission, right from its first years, it was so obvious that monitoring the Earth's gravity time changes was becoming an essential part of the Earth's observing system and that a follow on mission was studied and decided: GRACE-FO, a partnership between NASA and the German Research Centre for Geosciences (GFZ), was launched in May 2018. GRACE-FO uses the same two-way microwave-ranging link as GRACE, which produces so called K-band measurements, and each spacecraft houses a micro-accelerometer. In addition, GRACE-FO employs laser ranging interferometry (LRI) as a technological experiment in preparation for future satellites, an instrument which improves the precision of the separation distance measurement by a factor of 20. Unfortunately the science phase of the mission was delayed until January 2019 due to an anomaly in a system component, and the mission is still plagued by a dysfunction of one of the accelerometers, which may affect the final precision which was envisaged.

The third type mission, which was selected in November 1999 as the first core mission of the Earth Explorer program of

ESA (European Space Agency), was GOCE (Gravity Field and Steady-State Ocean Circulation Explorer) (ESA 1999) launched in 2009. For the first time, a gradiometer was flown onboard a satellite on a very low altitude orbit (around 250 km) of 96.5 degree inclination, with drag-free control. It operated until October 2013 when he reentered the Earth's atmosphere. The gradiometer instrument consisted in six ultra-sensitive micro-accelerometers of a new class (capable of measuring accelerations of $5.10^{-13} \text{ ms}^{-2}$) and delivered gravity gradients (second derivatives of the gravitational potential) in spacecraft axes with an accuracy of a few milli-Eötvös ($1 \text{ E} = 10^{-9} \text{ s}^{-2}$) per $\text{Hz}^{1/2}$ in the 0.005–0.1 Hz measurement bandwidth. The satellite was also tracked by the GPS (and by ground laser stations, actually like CHAMP, GRACE, and GRACE-FO for control and safety), and the combination of these trajectory observations and of the gradiometer measurements has produced a global gravity model with a resolution of about 100 km with a total uncertainty of ~ 2 cm on the geoid.

Overview of Satellite Dynamics Methods

In the following we describe the main mathematical concepts and tools which are in common use in satellite geodesy, that is:

1. The representation of the gravitational potential
2. The solution of the dynamical problem, with two sub-parts:
 - The equations of motion and their solution
 - The equations, derived from the preceding ones, necessary for obtaining the trajectory and physical parameters at stake (i.e., for the inverse problem)
3. The setting up of the observation equations and their linearization in the inverse problem
4. The solution of the inverse problem

Some of the described material may be redundant with Pavlis (ibid).

The Representation of the Gravitational Potential

For most global studies, the potential of the Earth's attracted masses is usually modeled by a spherical harmonic series, written at each point P as:

$$U = \frac{GM}{r} + \sum_{l>0} \frac{GM}{r} \left(\frac{R}{r}\right)^l \sum_{m=-l}^{+l} K_{lm} P_{lm}(\sin \phi) e^{im\lambda} \quad (1)$$

with G , gravitational constant; M , mass of the Earth; K_{lm} , dimensionless harmonic coefficient of degree l and order m , and r , ϕ , λ , the spherical coordinates of the point P in a reference system fixed with respect to the Earth; the P_{lm} 's are the Legendre functions of the first kind (polynomials when $m = 0$) and, for computational stability reason, are

usually normalized (like the K_{lm} 's) – which we will assume in the following.

Alternatively, the sum of all terms of the same degree l can be written with real (normalized) coefficients and functions, in the form:

$$U_l = \frac{GM}{r} \left(\frac{R}{r}\right)^l \sum_{m=0}^{+l} \times (C_{lm} \cos m\lambda + S_{lm} \sin m\lambda) P_{lm}(\sin \phi) \quad (2)$$

The C_{l0} coefficients are called zonals, and the C_{lm} , S_{lm} ($m > 0$) are the tesseral harmonics with the sectorials ($l = m$) being a special sub-class. They may vary with time.

Practically the potential series is truncated at maximum degree L , i.e., $|m| \leq l \leq L$, and therefore $\dim \{K_{lm}\} = (L+1)^2$. L depends on the type of satellite orbits and data sensitivity.

When the K_{lm} are known, it is straightforward to compute any geodetic functional F , which is linear in these parameters:

$$F(U) = \sum_l f_l U_l \quad (3)$$

For instance:

- For the gravity anomaly field: $F = \Delta g$ and $f_l = (l-1)/r_E^2$
- For the geoid height: $F = N$ and $f_l = 1/\gamma(r_E, \phi)$

where r_E is the radius vector and $\gamma(r_E, \phi)$ is the normal gravity at the evaluation point on some adopted reference ellipsoid (E) approximating the shape of the geoid.

Variances and covariances can also be propagated to these functions from the covariance matrix (see 5.4) of the spherical harmonic coefficients (in particular the variances δC_{lm}^2 and δS_{lm}^2). Useful indicators of the precision of a model are the mean errors per degree on some given geodetic functional – usually the geoid and/or the gravity anomaly field, computed as:

$$\langle \delta F \rangle_l = f_l \left[\sum_{m=0}^l (\delta C_{lm}^2 + \delta S_{lm}^2) \right]^{1/2} \quad (4)$$

The Solution of the Dynamical Problem

The equations of motion are most often considered in rectangular coordinates with respect to a given reference system and written in a specific time scale. Although approximate analytical theories exist for describing the evolution of an orbit subject to disturbing forces (they are still in use for mission analysis and first-order scientific investigations), numerical integration is systematically used for higher precision. Forces acting on a satellite are not only of gravitational origin, and their complete modeling involves parameters, such as the K_{lm} coefficients, which determination requires to integrate other

systems of equations associated with those providing the trajectory.

Reference System and Time Scale

One makes choice of a time scale, such as the Geocentric Coordinate Time or the Terrestrial Time and of reference systems: one related to the Earth's surface (the International Terrestrial Reference Frame, or ITRF) and one fixed in space (the International Celestial Reference Frame, or ICRF) as defined by the International Astronomical Union and the International Association of Geodesy (IERS 2010). The K_{lm} coefficients are expressed in ITRF, and the equations of motion of a satellite are usually written in a system (EME2000) defined by the Earth mean equator and mean equinox of epoch J2000 (January 1, 2000 at 12 h) and fixed in ICRF. The transformation from ITRF to EME2000 involves all phenomena related to the rotational motion of Earth in space, that is, precession, nutation, sidereal time, and polar motion – of which parameters may be part of the unknowns in the inverse problem.

Equations of Motion

The spacecraft motion obeys the differential equations, in EME2000:

$$\ddot{\vec{r}} = \bar{\nabla}U + \bar{\nabla}U^* + \bar{\Gamma}_{SP} + \bar{\Gamma}_D + \bar{\Gamma}_{th} \quad (5)$$

with $\vec{r}(t_0) = \vec{r}_0$; $\dot{\vec{r}}(t_0) = \dot{\vec{r}}_0$ (initial conditions at beginning of a piece of trajectory – an arc).

In the above, U is the Earth's gravitational potential, U^* is the so-called third body gravitational potential, Γ_{SP} is the acceleration vector due to solar pressures (direct from the Sun, indirect, i.e., reflected by the Earth – known as the albedo effect) and to infrared pressure (radiated by the Earth considered as a black body), Γ_D is the acceleration due to drag from the atmosphere, and Γ_{th} designates the residual acceleration due to poorly known effects (e.g., thruster accelerations in the case of an orbit controlled spacecraft). U^* includes the Sun, the Moon, and planets.

For computing the surface accelerations Γ_{SP} and Γ_D , a so-called box-and-wing (macro) spacecraft model (with a few planar faces) is used (together with ad hoc physical coefficients), and one has also to use an attitude model. Alternatively, Γ_{SP} and Γ_D may be measured by very sensitive micro-accelerometers such as in the CHAMP, GRACE, GRACE-FO, and GOCE satellites. The Γ_{th} effects may be modeled via time series or as window (rectangular) functions at some specific epochs (for instance, the times of attitude maneuvers).

Integration of the Equations of Motion

This is the direct dynamical problem. Numerical integration is systematically used. Robust multi-step integrators, of the Adams, Adams-Moulton, or Cowell type, in normal form or

in summed form (which reduces the accumulation of round-off errors), are preferred. Such integrators have a starting phase and then a running phase with a predictor-corrector iterative scheme. One big advantage of such class of methods is that one can easily speed up the algorithm, by sometimes a dramatic factor in performing a pseudo-correction by which the full force function (right-hand side member of the dynamical equations) is replaced (after the first loop of the corrector part) by an approximation (e.g., the central term of the gravity field plus a few other important terms).

Variational Equations and Their Integration

The inverse dynamical problem consists in determining the physical models parameters (or a subset of them) from the observation of the satellite trajectory – by any means. It also implies the determination of the equations of motion initial conditions in a first step or simultaneously (usually by iteration).

As it is known, and as we describe it in the next subsection, this requires to obtain at each time t the partial (first) derivatives of the satellite state vector $(\vec{r}, \dot{\vec{r}})$ with respect to the components of the initial conditions vector (which we call $\bar{Z}$) and similarly with respect to the vector of all searched for physical parameters (which we call $\bar{P}$). This is achieved via the variational equations.

We rewrite the equations of motion as follows:

$$\begin{aligned} \ddot{\vec{r}} &= \bar{\gamma}(\vec{r}, \dot{\vec{r}}, t, \bar{P}) \\ \text{with } (\vec{r}, \dot{\vec{r}})_{t_0} &= \bar{Z} \end{aligned} \quad (6)$$

$\vec{r}$ and $\dot{\vec{r}}$ are functions of t , t_0 , $\bar{Z}$, and $\bar{P}$. Let w be any Z_j or P_k . Then:

$$\frac{\partial}{\partial w} \ddot{\vec{r}} = \frac{d^2}{dt^2} \left(\frac{\partial \vec{r}}{\partial w} \right) = \frac{\partial \bar{\gamma}}{\partial \vec{r}} \left(\frac{\partial \vec{r}}{\partial w} \right) + \frac{\partial \bar{\gamma}}{\partial \dot{\vec{r}}} \left(\frac{\partial \dot{\vec{r}}}{\partial w} \right) + \frac{\partial \bar{\gamma}}{\partial w} \quad (7)$$

We note that $\partial \dot{\vec{r}} / \partial w = d(\partial \vec{r} / \partial w) / dt$ and that the last term is present only if $w = P_k$.

Therefore any vector $\bar{g} = \partial \bar{\gamma} / \partial w$ is solution of the system of so-called variational equations:

$$\ddot{\bar{g}} = \frac{\partial \bar{\gamma}}{\partial \vec{r}} \bar{g} + \frac{\partial \bar{\gamma}}{\partial \dot{\vec{r}}} \dot{\bar{g}} + \left\{ \begin{array}{c} 0 \\ \partial \bar{\gamma} / \partial w \end{array} \right\}_{w=P_k}^{w=Z_j} \quad (8)$$

with

$$\begin{aligned} g_i(t_0) &= \dot{g}_i(t_0) = 0 \quad \forall i \quad \text{if } w = P_k, \\ &= \delta_{ij} \quad \text{if } w = Z_k, \end{aligned}$$

since this is nothing but the Jacobian $\left[\frac{D(r_i, \dot{r}_i)}{D(r_i(t_0), \dot{r}_i(t_0))} \right]_{t_0}$.

We have one such system for each parameter to be estimated, which could be very costly to integrate, but we note that:

$\partial \bar{\gamma} / \partial (\vec{r}, \dot{\vec{r}})$ is common to all systems ($\forall w$) and can be evaluated once (at convergence of the predictor-corrector

loop), at each stepsize in the course of the integration of the main system, these variational equations are linear (with coefficients function of time); therefore they can be economically integrated by quadrature or, more efficiently, by using solely the predictor part of the integrator (another approach uses the matrizant of the associated homogeneous system, followed by quadratures).

The Observation Equations for the Inverse Problem

The recovery of the gravity field model parameters is an inverse problem which consists, here, in deriving the gravity coefficients (K_{lm}) from a set of observation equations.

General Form of the Equations

Each measurement q of a functional f of the satellite position (or velocity) is of the type:

$$q = f(\bar{r}, \dot{\bar{r}}, t, \bar{Q}) \quad (9)$$

or

$$q = f(\bar{r}_1, \bar{r}_2, t_1, t_2, \bar{Q}) \quad (10)$$

Direct measurements of a gravity field functional (such as gravity gradients along one or several spacecraft axis) are of the form:

$$q = f(\bar{r}, \dot{\bar{r}}, t, \bar{P}) \quad (11)$$

In the above, $\bar{r}, \dot{\bar{r}}, \bar{r}_1, \bar{r}_2$ are usually functions of $\bar{Z}$ and $\bar{P}$, respectively, the initial condition (state) vector (often comprising other “fudge” parameters – e.g., empirical multiplying factors for the surface forces) and the dynamical parameter vector (of which the gravity field parameters is a subset). $\bar{Q}$ is a vector of geometric quantities (for instance, coordinates of observing stations) and/or other physical parameters (e.g., propagation delays in a medium).

The observation equations are usually nonlinear with respect to the unknown parameters. The classical approach is to linearize them around starting values:

$$\begin{aligned} \Delta q &= q^{\text{obs}} - q^{\text{comp}} \\ &= \frac{\partial q^{\text{comp}}}{\partial \bar{Z}} \Delta \bar{Z} + \frac{\partial q^{\text{comp}}}{\partial \bar{P}} \Delta \bar{P} + \frac{\partial q^{\text{comp}}}{\partial \bar{Q}} \Delta \bar{Q} \end{aligned} \quad (12)$$

with

$$\Delta \bar{Z} = \bar{Z} - \bar{Z}_0, \quad \Delta \bar{P} = \bar{P} - \bar{P}_0, \quad \Delta \bar{Q} = \bar{Q} - \bar{Q}_0$$

Partial derivatives are computed via the variational equations of the equations of motion of the satellite(s) when q is a distance, velocity... between an observing reference point

(a tracking station on Earth, another spacecraft – like in the Earth gravity mapping missions CHAMP, GRACE, GRACE-FO and GOCE), or are directly evaluated if q can be expressed as a simple function of the physical phenomena at stake at the satellite location, such as a gravity gradient in the case of a gradiometer sensing the gravitational potential (case of the GOCE mission).

More precisely, we have:

$$\begin{aligned} \frac{\partial q^{\text{comp}}}{\partial \bar{Q}} &= \frac{\partial f}{\partial \bar{Q}} \\ \frac{\partial q^{\text{comp}}}{\partial (\bar{Z}, \bar{P})} &= \frac{\partial f}{\partial \bar{r}} \frac{\partial \bar{r}}{\partial (\bar{Z}, \bar{P})} + \frac{\partial f}{\partial \dot{\bar{r}}} \frac{\partial \dot{\bar{r}}}{\partial (\bar{Z}, \bar{P})} \end{aligned} \quad (13)$$

(or $\partial f / \partial \bar{r}_1, \partial f / \partial \bar{r}_2$ if the observable involves two spacecraft with both orbits being adjusted simultaneously like in GRACE). In current observation systems, $\partial f / \partial \dot{\bar{r}}$ is not present: the only case would be a Doppler-type measurement, but this is always treated as a difference of ranges (between the beginning and the end of a given count time).

The quantities $\partial f / \partial \bar{Q}, \partial f / \partial \bar{r}$ are in principle easy to compute at the observation epoch although in practice one has to take account of the signal propagation duration, of small motions at the observing stations (solid and fluid tides loading effects, tectonic movements), of relativistic clock effects, etc.

Finally $\partial \bar{r} / \partial (\bar{Z}, \bar{P})$ is obtained by integration of the variational equations as described in the previous sub-section.

Example: Range Measurement Generic Observation Equation

All tracking observations of a satellite orbit involve one or several range values. Consequently, it is logical to construct a generic observation equation for a single range measurement between two points S_1, S_2 , being understood that the final equation for a real measurement is a combination of generic equations. Besides this allows an easy generalization to the case when the observing station is replaced by another spacecraft (a high altitude navigation satellite – e.g., from the GPS constellation or a co-orbiting satellite like in the GRACE mission). Therefore let us consider the case where S_2 is the satellite of interest and S_1 can be a station or another spacecraft.

Firstly propagation times must be taken into account by an iterative scheme which allows to refer the quantities to the true epochs (t_1 for S_1 and t_2 for S_2). Various corrections being properly applied the range $S_1 S_2$ then reduces to the geometric quantity:

$$r_{12} = \|\bar{r}_2 - \bar{r}_1\|$$

The observation equation is therefore:

$$r_{12}^{\text{obs}} - r_{12}^{\text{comp}} \approx dr_{12} = \hat{u}_{12} \cdot (d\bar{r}_2 - d\bar{r}_1) \quad (14)$$

with $\hat{u}_{12} = \bar{r}_{12} / r_{12}$ and

$$d\bar{r}_{n,i} = \sum_j \frac{\partial \bar{r}_{n,i}}{\partial (PInt_{n,i}^j)} \Delta(PInt_{n,i}^j) + \sum_k \frac{\partial \bar{r}_{n,i}}{\partial (PExt^k)} \Delta(PExt^k)$$

($n = 1, 2$).

The quantities with *comp* superscript are derived from the numerical integration of the equations of motion of the satellite (s), and the Jacobian matrices $\partial\{\dots\}/\partial\{P\dots\}$ are obtained from the integration of the variational equations; the i subscript refers to the i .th arc in a strategy where the ensemble of the processed data is split into several batches, and we have distinguished between internal (or inner) parameters (proper to the i .th arc) and external parameters (common to all arcs). For each batch of data, the inner parameters are solely solved for in an iterative process (because of significant nonlinearities); at the end all parameters are considered and the equations are solved in ensemble. The whole procedure may then be iterated.

The Solution of the Inverse Problem

The observation equations have been linearized, which yield the system:

$$Ax = b \quad (15)$$

where x is the vector of unknowns (of dimension N), A is a $M * N$ matrix (M is the number of observations and usually $M \gg N$), and b is the residual vector (of the Δq s). A weight matrix $\Pi(M * M)$ is associated to the observations; it is very often reduced to its diagonal part for practical reasons.

In some instances, the measurement bandwidth may be rather limited, and it is then necessary to filter the observation equations as precisely as possible. This has been the case of the GRACE K-band radar measurements and of the GOCE gradiometer data, for which very specific algorithms (due to the size of the problem) have been developed (Schuh 2003; Siemes 2008). This consists in applying a filtering matrix to both sides of the observation equation system. Another approach is to filter the spherical harmonic coefficients solution afterward (Kusche 2007) or to apply space domain filters to geodetic functions (Jekeli 1981; Han et al. 2005). Such procedures have to be applied to time variable gravity models (monthly or decadal solutions) derived from the GRACE and GRACE-FO missions.

These equations can be solved by least squares, usually by forming the normal equation system:

$$(A^T \Pi A)x = A^T \Pi b \quad (16)$$

from which x is computed; $N = A^T \Pi A$ is a $(N * N)$ matrix.

A regularizing procedure may be used when the observability and/or sufficient decorrelation of all parameters are not

guaranteed; in such a case N is replaced by $N + R$, where R is the regularizing matrix. This is usually necessary for the inner unknowns (initial orbital elements, empirical parameters, etc. which may not be fully determinable). For the physical parameters, R may be:

- Derived from an approximate physical law (e.g., Kaula's law for the gravitational harmonic coefficients – usually applied above a certain degree or for particular classes of coefficients such as the zonals)
- Obtained from a priori physical information

The covariance matrix of x is finally computed as:

$$\text{cov}(x) = \sigma_0^2 (N + R)^{-1} \quad (17)$$

where σ_0 is the variance of unit weight estimated from the system and residual vector.

Other methods (orthogonal decomposition – e.g., using QR factorization) may be used for computational accuracy reasons.

The computational cost of the formation of the normal system and of its solution is:

- Proportional to MN^2 for forming the normal equations or for an orthogonal decomposition
- Proportional to N^3 for obtaining the solution and the covariance matrix.

The first step is obviously the most demanding, and this aspect has always driven the strategy of research teams working on the determination of global gravity field models.

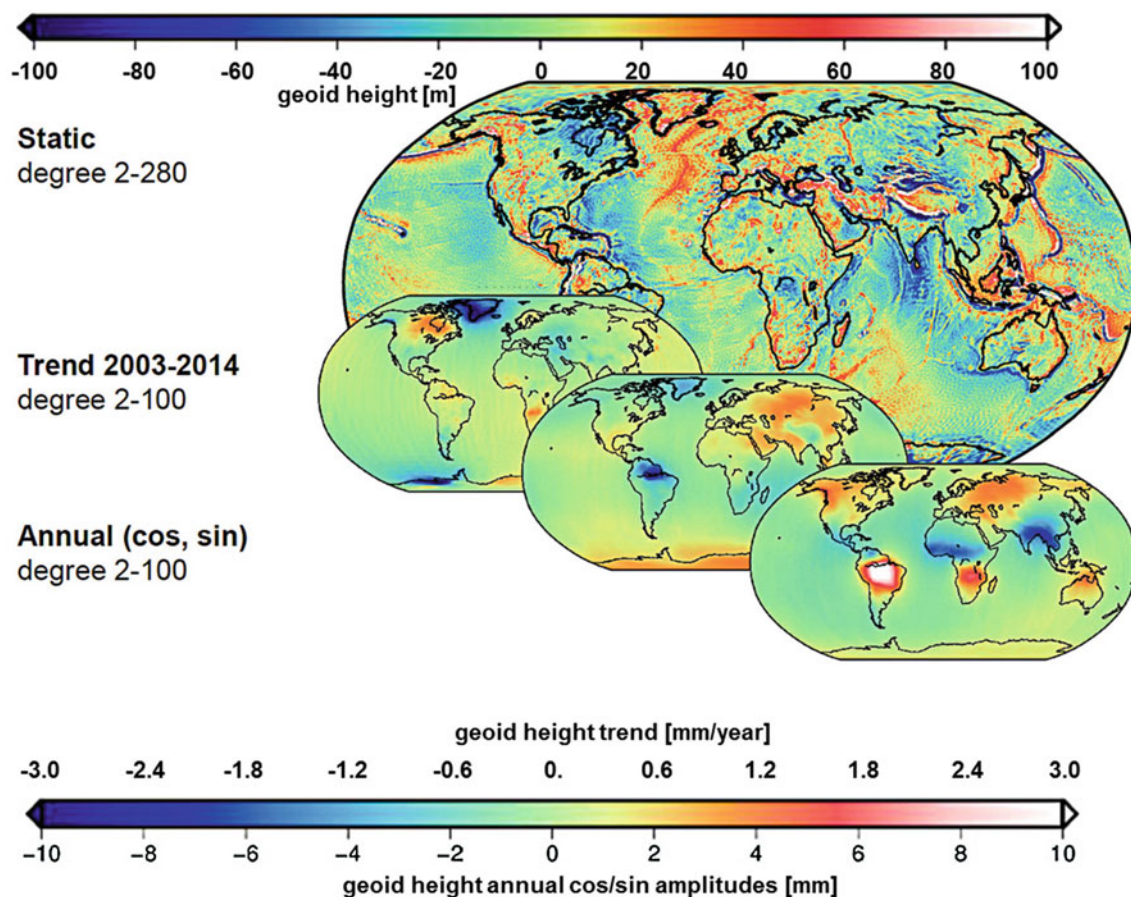
Progress Made: A Short Summary

Early in the Space Age, several research groups started working on the determination of the gravity field of the Earth. One of the first global models published was the Standard Earth I, by an American team at the Smithsonian Astrophysical Observatory (Lundquist and Veis 1966); it was a solution derived solely from analytical theory and satellite tracking (optical) data and complete to degree and order 8 (plus a few so-called resonant harmonics of higher degree and order). It was quickly followed by a combined solution (to degree-order 16) which was based on a little more satellite data and on mean surface gravity values (Gaposchkin and Lambeck 1970); this model was enriched some years later (Gaposchkin 1973). Meanwhile similar work (but based on numerical integration approach) had started at the NASA Goddard Space Flight Centre (GSFC) and in Europe where two groups (at CNES, the French Space Centre and at the Technical University in Munich) joined their efforts. This resulted in

two series of models: the GEM models on the American side (Lerch et al. 1972a, b) and the GRIM models on the European side (Balmino et al. 1976a, b). For several years these two series of models were improved: the GSFC team joined efforts with a group from the University of Texas at Austin, and the German side of the European alliance was transferred at the DGFI (Deutsches Geodatisches Forschungsinstitut) then at the GFZ (GeoForschungsZentrum). All teams involved in this domain produced many global models, each milestone being marked in general by a new satellite solution, followed by a combined model i.e., with the added (higher spatial frequency) contribution of surface gravimetry and ocean altimetry (see Pavlis, *ibid.*).

At the end of the twentieth century, models were limited (mostly in the precision of knowledge of their long and medium wavelength components) by the weaknesses of the satellite solutions; for instance the cumulated geoid error at degree 50 (400 km resolution) was still over 1 m. The revolution came with the advent of the new dedicated gravity missions: CHAMP quickly provided a gain of a factor 5 to 10 (depending on the resolution) to the models derived from it (and from other good past satellite data); GRACE allowed the

study of time variations (see Cazenave et al., this *encyclopedia*) and yielded mean (static) models which quality now reaches a centimeter error at degree 120; GOCE reached the expectation and complemented the GRACE-derived mean models in the higher spatial frequencies (up to degree 300). Due to the difficulty and importance of this type of research (which outputs require extremely thorough validation – by cross-verifications), many more groups have engaged in the processing of the new missions data, such as the Institute of Theoretical Geodesy at the University of Bonn which produced in 2010 the best gravity model from GRACE (Mayer-Gürr et al. 2010). Since then several institutes, especially in Europe, joined efforts to compute on a regular basis solutions from the new mission data, for both the mean Earth's gravity field and the time varying part. Figure 1 is an example of such results (Mayer-Gürr et al. 2015), obtained by the GOCO consortium composed of teams from the Technical Universities in Munich and Bonn (Germany), Graz (Austria), Bern (Switzerland), and the Space Research Institute in Austria. Other institutions, such as NASA-JPL (Jet Propulsion Lab., USA), the Centre for Space Research (Austin Univ., USA), GFZ (Germany), and GRGS (Groupe de Recherche de



Gravity Method, Satellite, Fig. 1 The geoid heights and their main variations from the GOCO05s model. This is a pure satellite solution which combined data from GOCE, GRACE, and all geodetic satellites tracked by laser telemetry (Lageos 1 and 2, Starlette, Stella, AJISAI, etc.)

Gravity Method, Satellite, Table 1 Some landmarks of global gravity field modeling

Model	Year	Degree	Data	Reference
IGGT-R1C	2018	240	S(GOCE, GRACE)	Lu et al. (2018)
NULP-02S	2016	250		Marchenko et al. (2016)
EIGEN6-S4	2016	300	S(GOCE, GRACE, Lageos)	Förste and Bruinsma (2016)
ITG-Grace2010s	2010	180	S(GRACE)	Mayer-Gürr et al. (2010)
GGM05G	2015	240	S(GOCE, GRACE)	Bettadpur et al. (2015)
GOCO05s	2015	280	S(GOCE, GRACE, SWARM, Lageos & other laser satellites)	Mayer-Gürr et al. (2015)
GO-CONS-GCF-2-DIR4	2013	260	S(GOCE, GRACE, Lageos)	Bruinsma et al. (2013)
GGM03S	2008	180	S(GRACE)	Tapley et al. (2007)
ITG-Grace02s	2006	170	S(GRACE)	Mayer-Gürr et al. (2006)
EIGEN-GRACE02S	2004	150	S(GRACE)	Reigber et al. (2005)
EIGEN-GRACE01S	2003	140	S(GRACE)	Reigber et al. (2003c)
EIGEN-2	2003	140	S(CHAMP)	Reigber et al. (2003b)
EIGEN-1	2002	119	S(CHAMP)	Reigber et al. (2003a)
EIGEN-1S	2002	119	GRIM5,S+ CHAMP	Reigber et al. (2002)
GRIM5S1	1999	99	S	Biancale et al. (2000)
GRIM4S4	1995	70	S	Schwintzer et al. (1997)
JGM2S	1994	60	S	Nerem et al. (1994)
JGM1S	1993	60	S	Nerem et al. (1994)
GEMT1	1987	36	S	Marsh et al. (1988)
GEM9	1977	20	S	Lerch et al. (1979)
GEM7	1976	16	S	Wagner et al. (1976)
GRIM1	1975	10	S	Balmino et al. (1976a)
GEM5	1974	12	S	Lerch et al. (1974)
GEM3	1972	12	S	Lerch et al. (1972b)
GEM1	1972	12	S	Lerch et al. (1972a)
SE1	1966	8	S	Lundquist and Veis (1966)

Geodesie Spatiale, from the French Space Centre), continue to produce gravity models in fairly great number.

Global Earth's gravity field models computed since the beginning of the Space Age are now archived at GFZ, Germany, in the ICGEM service <http://icgem.gfz-potsdam.de/>, from which documentation, spherical harmonics sets of coefficients, and gridded values of geodetic functions can be retrieved. As an example, Table 1 lists some of the main satellite only solutions published since 1966.

Conclusion and Vision of the Future of Gravity Mapping from Satellite

New progress at global scale can only come from a space approach for this only can satisfy the scientific needs (homogeneous data, full coverage, higher resolution and precision) within reasonable time and cost.

From the results already obtained, which demonstrate that the new gravity missions have met the expected outcome, one can say that several of the objectives previously discussed have been reached. However, and like in other domains, new

requirements are emerging, encouraged by the findings of these missions: higher resolution and precision models and therefore measurements are to be needed.

If we restrict ourselves to the approach by satellite-to-satellite tracking or satellite gravity gradiometry, which are used in the new satellite missions of this decade, we may foresee improvements:

- On the precision of deciphering the time variations of the gravity field at shorter spatial scale, by exploiting further the co-orbiting satellites case like in GRACE; this will be done by performing inter-satellite range measurements by laser interferometry, a technique which is presently operated in GRACE-FO,
- on the resolution of the static field (mapping of finer lateral variations of gravity) by a satellite-borne gradiometer of enhanced sensitivity, such as a cold atom interferometry gradiometer (Douch et al. 2018).

Future progress may also be possible by revisiting the differential approaches, for instance, by differencing (or double differencing) relative velocity measurements, or

by flying two or more satellites in specific geometries (gravimetric “wheel,” pendulum configuration, etc.), or by flying simultaneously several GRACE-type constellations in order to gain on the temporal resolution.

But there comes a critical aspect, GRACE, and now GRACE-FO results have shown the great potential of such missions for environmental studies and call for the need to continue monitoring, without gap, the Earth’s gravity field variations (much like we continue monitoring the mean sea level by altimetry). GRACE lasted much longer than expected, but the following mission GRACE-FO was not yet ready to be launched in time in order to avoid a gap in the gravity model series. The missing months (it could have been years) were filled thanks to the availability of the SWARM satellite mission (several low altitude spacecraft tracked by the GPS and devoted to measuring the Earth’s magnetic field), though with a drastic reduction in the resolution of the recovered gravity variations (Meyer et al. 2019). The main issue is, and will be at each time a new gravity satellite system is flown, to decide on the follow-on mission as quickly as possible so as to avoid any discontinuity in the data series.

Cross-References

- Gravimeters
- Gravity Anomalies, Interpretation
- Gravity Data, Advanced Processing
- Gravity Data, Regional: Residual Separation
- Gravity Field of the Earth
- Gravity Field, Temporal Variations from Space Techniques
- Gravity Measurements, Absolute
- Gravity Method, Airborne
- Gravity Method, Principles
- Gravity Method, Surface
- Gravity, Gradiometry

Bibliography

- Balmino G, Reigber C, Moynot B (1976a) A geopotential model determined from recent satellite observation campaigns (GRIM1). *Manusc Geodaet* 1:41–69. Berlin
- Balmino G, Reigber C, Moynot B (1976b) The GRIM2 Earth Gravity Field. *Deutsche Geodätische Kommission, Reihe A, Heft Nr. 86*, München
- Bettadpur S, Ries J, Eanes R, Nagel P, Pie N, Poole S, Richter T, Save H (2015) Evaluation of the GGM05 mean earth gravity models. *Geophysical research abstracts*, vol 17, EGU2015-4153, Vienna
- Biancale R, Balmino G, Lemoine J-M, Marty J-C, Moynot B, Barlier F, Exertier P, Laurain O, Gegout P, Schwintzer P, Reigber C, Bode A, Gruber T, König R, Massmann F-H, Raimondo JC, Schmidt R, Zhu SY (2000) A new global Earth’s gravity field model from satellite orbit perturbations: GRIM5-S1. *Geophys Res Lett* 27(22):3611–3614
- Bruinsma SL, Forste C, Abrikosov O, Marty JC, Rio MH, Mulet S, Bonvalot S (2013) The new ESA satellite-only gravity field model via the direct approach. *Geophys Res Lett* 40(14):3607–3612. <https://doi.org/10.1002/grl.50716>
- Douch D, Wu H, Schubert C, Müller J, Pereira Dos Santos F (2018) Simulation-based evaluation of a cold atom interferometry gradiometer concept for gravity field recovery. *Adv Space Res* 61(5):1307–1323
- ESA (European Space Agency) (1999) Gravity field and steady-state ocean circulation explorer mission (GOCE). *ESA-SP 1233(1)*:1–217
- Forste C, Bruinsma SL (2016) EIGEN-6S4: A time-variable satellite-only gravity field model to d/o 300 based on LAGEOS, GRACE and GOCE data from the collaboration of GFZ Potsdam and GRGS Toulouse. <https://doi.org/10.5880/icgem.2016.004>
- Gaposchkin EM (1973) Smithsonian standard earth (III). Special report no. 353, Smithsonian Astrophysical Observatory, Cambridge, MA
- Gaposchkin EM, Lambeck K (1970) 1969 Smithsonian standard earth (II). Special report no. 315, Smithsonian Astrophysical Observatory, Cambridge, MA
- Han S-C, Shum CK, Jekeli C, Kuo C-Y, Wilson CR, Seo K-W (2005) Non-isotropic filtering of GRACE temporal gravity for geophysical signal enhancement. *Geophys J Int* 163:18–25
- IERS (International Earth Rotation Service) (2010) IERS conventions (2010). IERS technical note 36, Frankfurt
- Jekeli C (1981) Alternative methods to smooth the Earth’s gravity field. Report no. 327, Department of Geodetic Science, Ohio State University
- JPL (Jet propulsion Laboratory) (1998) GRACE: gravity recovery and climate experiment. Science and Mission requirements document, revision A, JPLD-15928, NASA’s Earth System Science Pathfinder Program, pp 1–84
- Kusche J (2007) Approximate decorrelation and non-isotropic smoothing of time-variable GRACE-type gravity field models. *J Geod* 81:733–749
- Lerch FJ, Wagner CA, Smith DE, Sandson ML, Brownd JE, Richardson JA (1972a) Gravitational field models for the earth (GEM1&2). Report X55372146, Goddard Space Flight Center, Greenbelt
- Lerch FJ, Wagner CA, Putney ML, Sandson ML, Brownd JE, Richardson JA, Taylor WA (1972b) Gravitational field models GEM3 and 4. Report X59272476, Goddard Space Flight Center, Greenbelt
- Lerch FJ, Wagner CA, Richardson JA, Brownd JE (1974) Goddard earth models (5 and 6). Report X92174145, Goddard Space Flight Center, Greenbelt
- Lerch FJ, Klosko SM, Laubscher RE, Wagner CA (1979) Gravity model improvement using Geos3 (GEM9 and 10). *J Geophys Res* 84(B8):3897–3916
- Lu B, Luo Z, Zhong B, Zhou H, Flechtner F, Förste C, Barthelmes F, Zhou R (2018) The gravity field model IGGT_R1 based on the second invariant of the GOCE gravitational gradient tensor. *J Geod* 92(5):561–572
- Lundquist CA, Veis G (1966) Geodetic parameters for a 1966 Smithsonian institution standard earth. Special report no. 200, Smithsonian Astrophysical Observatory, Cambridge, MA
- Marchenko AN, Marchenko AD, Lopushansky AN (2016) Gravity field models derived from the second degree radial derivatives of the GOCE Mission: a case study. *Ann Geophys* 59(6):649–659
- Marsh JG, Lerch FJ, Putney BH, Christodoulidis DC, Smith DE, Felsentreger TL, Sanchez BV, Klosko SM, Pavlis EC, Martin TV, Robbins JW, Williamson RG, Colombo OL, Rowlands DD, Eddy WF, Chandler NL, Rachlin KE, Patel GB, Bhati S, Chinn DS (1988) A new gravitational model for the earth from satellite tracking data: GEMT1. *J Geophys Res* 93(B6):6169–6215
- Mayer-Gürr, and 24 co-authors (2015) The new combined satellite only model GOCO05s. Abstract. EGU General Assembly, Vienna
- Mayer-Gürr T, Eicker A, Ilk KH (2006) ITG-GRACE02s: a GRACE gravity field derived from short arcs of the satellite’s orbit. In:

- Proceedings of the first symposium of international gravity field service, Istanbul
- Mayer-Gürr T, Kurtenbach E, Eicker A (2010) ITG-Grace2010 gravity field model. <http://www.igg.uni-bonn.de/apmg/index.php?id=itg-grace2010>
- Meyer U, Sosnica K, Arnold D, Dahle C, Thaller D, Dach R, Jäggi A (2019) SLR, GRACE and SWARM gravity field determination and combination. *Remote Sens* 11:956. <https://doi.org/10.3390/rs11080956>
- Nerem RS (1995) Terrestrial and planetary gravity fields. *Rev Geophys* 33(suppl):469–476
- Nerem RS, Lerch FJ, Marshall JA, Pavlis EC, Putney BH, Tapley BD, Eanes RJ, Ries JC, Schutz BE, Shum CK, Watkins MM, Klosko SM, Chan JC, Luthcke SB, Patel GB, Pavlis NK, Williamson RG, Rapp RH, Biancale R, Nouel F (1994) Gravity model developments for Topex/Poseidon: joint gravity models 1 and 2. *J Geophys Res* 99(C12):24421–24447
- Reigber C, Balmino G, Moynot B, Müller H (1983) The GRIM3 earth gravity field model. *Manuscr Geodaet* 8:93–138. Stuttgart
- Reigber C, Bock R, Forste C, Grunwaldt L, Jakowski N, Lühr H, Schwintzer P, Tilgner C (1996) CHAMP Phase B executive summary. G.F.Z., STR96/13, pp 1–37
- Reigber C, Balmino G, Schwintzer P, Biancale R, Bode A, Lemoine J-M, König R, Loyer S, Neumayer H, Marty J-C, Barthelmes F, Perosanz F, Zhu SY (2002) A high quality global gravity field model from CHAMP GPS tracking data and Accelerometry (EIGEN-1S). *Geophys Res Lett* 29(14). <https://doi.org/10.1029/2002GL015064>
- Reigber C, Balmino G, Schwintzer P, Biancale R, Bode A, Lemoine J-M, König R, Loyer S, Neumayer H, Marty J-C, Barthelmes F, Perosanz F, Zhu SY (2003a) Global gravity field recovery using solely GPS tracking and accelerometer data from CHAMP. *Space Sci Rev* 29:55–66
- Reigber C, Schwintzer P, Neumayer K-H, Barthelmes F, König R, Förste C, Balmino G, Biancale R, Lemoine J-M, Loyer S, Bruinsma S, Perosanz F, Fayard T (2003b) The CHAMP-only earth gravity field model EIGEN-2. *Adv Space Res* 31(8):1883–1888. [https://doi.org/10.1016/S0273-1177\(03\)00162-5](https://doi.org/10.1016/S0273-1177(03)00162-5)
- Reigber C, Schmidt R, Flechtner F, König R, Meyer U, Neumayer KH, Schwintzer P, Zhu SY (2003c) First EIGEN gravity field model based on GRACE mission data only. http://op.gfz-potsdam.de/grace/results/grav/g001_eigen-grace01s.html
- Reigber C, Schmidt R, Flechtner F, König R, Meyer U, Neumayer KH, Schwintzer P, Zhu SY (2005) An earth gravity field model complete to degree and order 150 from GRACE: EIGEN-GRACE02S. *J Geodyn* 39:1–10
- Schuh W-D (2003) The processing of band-limited measurements – filtering techniques in the least squares context and in the presence of data gaps. In: Beutler G, Drinkwater MR, Rummel R, von Steiger R (Hrsg) *Earth gravity field from space – from sensors to earth sciences, workshop, Bern, 11–15 Mar 2002*. Space science reviews. ISSI, pp 67–78
- Schwintzer P, Reigber C, Bode A, Kang Z, Zhu SY, Massmann FH, Raimondo JC, Biancale R, Balmino G, Lemoine JM, Moynot B, Marty JC, Barlier F, Boudon Y (1997) Long wavelength global gravity field models: GRIM4S4, GRIM4C4. *J Geod* 71(4):189–208
- Siemes C (2008) Digital filtering algorithm for decorrelation within large least squares problems. PhD Dissertation, Institut für Geodäsie und Geoinformation der Universität Bonn
- Tapley B, Ries J, Bettadpur S, Chambers D, Cheng M, Condi F, Poole S (2007) The GGM03 mean earth gravity model from GRACE. *Eos Trans AGU* 88(52). Fall Meet. Suppl., Abstract G42A-03
- Wagner CA, Lerch FJ, Brownd JE, Richardson JE (1976) Improvement in the geopotential derived from satellite and surface data – GEM 7 and GEM 8. Report X9217620, Greenbelt

Gravity Method, Surface

Dinesh Chandra Mishra¹ and Virendra M. Tiwari²

¹Hyderabad, India

²Gravity and Magnetic Studies Group, CSIR-National Geophysical Research Institute (CSIR), Hyderabad, India

Definition

Gravity method	A geophysical technique based on measuring the variations in the Earth's gravity field at specific locations.
Gravity stations	Locations where gravity measurements are made.
Base gravity stations	Locations where absolute value of the Earth's gravity field is known.
Primary base	Gravity observation stations, which are made on stable platform and absolute value of the gravity field is known to a greater accuracy.
Secondary bases	Temporary bases established for a particular survey.
Closed loop	Bases that are connected in a loop for base loop correction.
Drift correction	Correction applied for nontidal variation of gravity, which is mainly dependent on the behavior of gravity meters with time.
EOTVOS correction	Correction applied to marine gravity data for the movement of ship.
	mGal; 1 mGal = 10^{-5} m/s ²
	μGal; 1 μGal = 10^{-8} m/s ²

Introduction

Gravity method in geophysics is employed to determine subsurface density distribution by measuring the variation of the Earth's gravity field. Measurements of gravity field are usually made on the Earth or ocean surface. The scale of investigations (accuracy and spatial distance of gravity measurements) depends on the type of geological problems being investigated. These vary from a few μGal to mGal of measurements accuracy, which is a few millionth of average Earth's gravity field (9.8 m/s^2) and station intervals of a few meters to a few kilometers. Based on applications, gravity survey can be broadly classified into the following three groups (Mishra 2011):

1. Regional gravity surveys: Station spacing is a few kilometers and survey accuracy is of about 1 mGal. It is generally used for tectonic and geodynamic studies.
2. Reconnaissance gravity surveys: Station spacing is a kilometer or so and survey accuracy is of 0.1–0.5 mGal. Such surveys are normally planned for delineating promising zones for hydrocarbon and mineral exploration and tectonic and lithological mapping.
3. Detailed gravity surveys: Mainly done to demarcate local geological structures favorable for mineral deposits, engineering applications, and groundwater studies. Survey accuracy is in the order of ~ 0.01 mGal and station interval is in the order of a few meters.

Measuring Equipment

Gravity Measurements

There are primarily three methods of measuring Earth's gravity field:

- (a) A method in which period of pendulum is used to find absolute value of gravity
- (b) Method based on measuring velocity of free-falling body in the vacuum chamber that also provides absolute value of gravity
- (c) Sensitive spring balance, in which a test mass is suspended through a spring and change in the spring length offers variation in the Earth's gravity

The spring balances are relative instruments; however, they are main source of gravity data recording. They measure the differences in the gravity field between two locations and one of them is a base station. The Worden gravimeter, the Scintrex, and the La Coste Romberg gravimeter are extensively employed gravimeters across the globe. Scintrex gravity meters are now only commercially available gravity meters, which have least count of $1 \mu\text{Gal}$. Similar kinds of gravity meters are also used on the ship; however, they are specially designed to take care of vertical and horizontal accelerations with gyroscopic adjustments. A detailed account of gravity measurements is described in several publications (Marson 2012; Niebauer 2015; Van Camp et al. 2017). These measurements are mainly of two kinds: absolute gravity and relative gravity measurements. Absolute measurements provide Earth's gravity values directly at any location whereas relative gravity observations offer gravity variations only.

Elevation and Position Measurements

Gravity field of the Earth increases as we move toward poles due to flattening and centrifugal force. The factor that affects the Earth's gravity field most is the distance from center of the

Earth (R). Therefore, the elevation of stations and its position should be precisely known for gravity surveys. There are several ways to know the position and elevation, but nowadays global positioning system (GPS) provides a best method of determining position and elevation by utilizing information from several orbiting satellites. Sometimes, when measurements are made in shadow regions, particularly in the forests, the GPS navigations are not very good. In such cases, theodolite, navigation based on maps, and elevation using pressure measurement devices are often used as per the requirements of survey accuracy. GPS measurements are improved by using them in differential GPS (DGPS) mode. The accuracy of a position obtained using DGPS mode is a few centimeters in horizontal and tens of centimeter in vertical. Correctness of determined height is dependent on the knowledge of geoid height in the survey area. DGPS navigation systems that provide position, speed, and heading information of ship are common now for shipborne gravity measurements.

Land Gravity Survey

Gravity survey often refers to the measurement of the Earth's gravity field along with the position and elevation of the observation point. Elevation effect is approximately $0.2\text{--}0.3 \text{ mGal/m}$ implying that an error of 1 m in elevation of the stations will affect the gravity measurement by about 0.2 mGal which is quite significant, especially when the accuracy of gravimeters is $0.01\text{--}0.001 \text{ mGal}$. Therefore, ample efforts are made to record elevation of the stations of the desired accuracy for a particular survey. In general, gravity measurements are made over a plain area a few meters away from depression, ponds, mound, etc., which can also affect the gravity readings.

Preprocessing of Field Data

Gravimeters are highly sensitive and they are affected by the changes caused due to the mechanical and environmental changes such as jerks, pressure, and temperature. Gravimeters measure variations in the gravity field, thus the measurements using them are required to tie with the base stations where the absolute values of the gravity field are known. Therefore, measurements need to be corrected for factors influencing measurements to achieve the needed accuracy. There are mainly two corrections, namely (1) drift correction and (2) base loop correction, which are applied during course of the field work.

Drift Correction

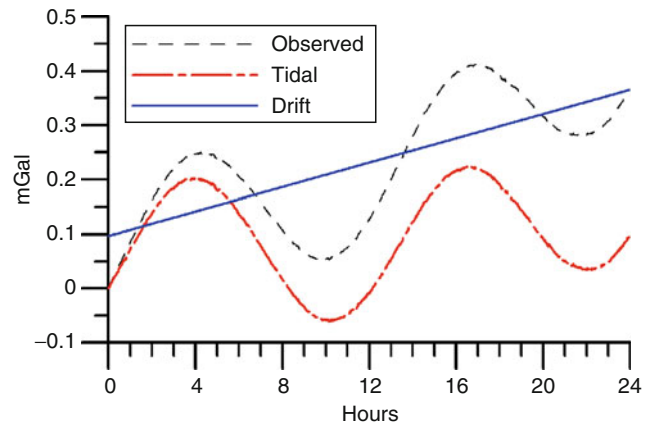
When the repeat observations are made by a gravimeter at the same location, it is found to be varying with time. The variation in the gravity field with time is caused due to several geophysical phenomena (Tiwari and Hinderer 2011). In a

gravity survey, time-varying gravity field is primarily attributed to the tidal effect and nontidal changes, caused due to the mechanical behavior of the gravimeter and contributions from atmospheric or hydrological effects during few hours of repeat observations. The peak to peak tidal effect is about 300 μGal during new moon day and can be precisely calculated for specific locations using standard available software such as ETERNA (Schüller 2015). However, precise estimate of tidal effects that include both the solid Earth and ocean loading contributions is not straightforward (Van Camp et al. 2016).

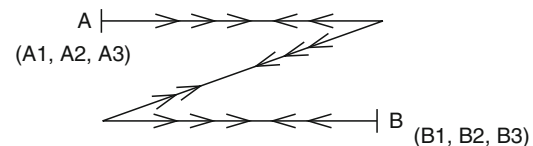
The nontidal variation is referred as drift of the gravimeter (Fig. 1). This is normally $\sim 0.0\text{--}0.03$ mGal/h and is corrected after removing tidal effects computed for that specific location. The drift correction is estimated by making repeat observations at the interval of 1–2 h and the difference in the two nontidal gravity values, known as drift of the instrument, is distributed proportionally to the other observations made during that period. Figure 1 shows the variation of the gravity field with time, tidal effect, and drift at Hyderabad. Repeat observations at a fixed location are also important to understand the condition of gravimeters. If drift of the gravimeter is more than a permissible limit of <0.05 mGal/h, the instrument needs to be thoroughly checked for scale adjustments, battery, etc. or survey needs to be redone.

Base Loop Correction

Gravity base stations are the points where the absolute value of the gravity field is known to an accuracy of 0.01 mGal (10 μGal) and are used to obtain the absolute value of the gravity field at the stations in the field by measuring its variations with respect to base stations. There are primary base stations in the country usually at fixed locations such as railway stations, airports, etc. In any gravity survey, a network of secondary bases is established encompassing survey block, which is connected to any of the known primary base. To achieve the accuracy, a particular procedure of measurements is followed. The bases are connected by looping three times between two bases A and B as A-B-A-B-A-B (A and B being two stations, Fig. 2) and occupying them usually within an hour to ensure that the instrument drift is linear and minimum. This measurement procedure is referred to as double tie of bases. This kind of looping provides three observations at every base, which are corrected for the drift of the gravimeter based on consecutive five readings. The consistent value from the three observations or the average of close-by readings if they are not consistent is chosen for every base station. In case of large survey area such as in a regional survey or hydrocarbon exploration, bases are first established in the region that are subsequently used to tie the gravity stations for their absolute values and simultaneously used for drift correction so as to avoid reoccupying same



Gravity Method, Surface, Fig. 1 Observations of gravity changes at Hyderabad for a day. Tidal and drift of the instrument are plotted separately



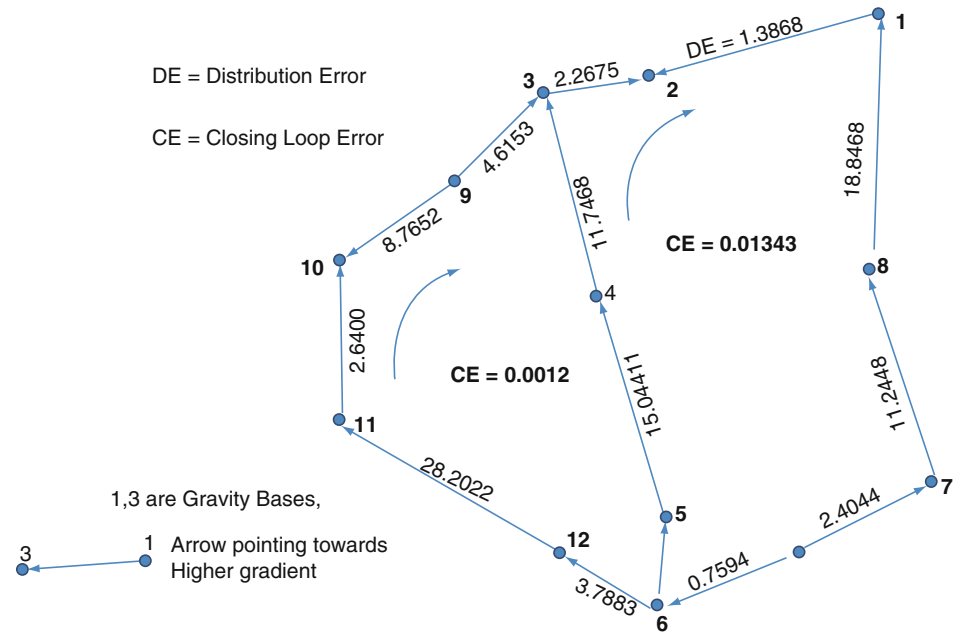
Gravity Method, Surface, Fig. 2 A and B are two bases that are occupied five times starting from A to B and going back and forth that provide three diurnal corrected readings at each station which are used to obtain correct reading at base stations

base after 2 h. In these cases, the survey starts from a base station and closed at some other base station after every 1–2 h. The difference in the gravity field is adjusted with respect to their absolute values distributed to the station occupied during that period.

The bases are established on permanent structures for easy identification during the survey and afterward. Figure 3 shows a typical example of a network of bases, which shows the differences in the gravity values between two bases along the arms with arrows showing toward higher values. These differences along the arms of a network (loop) are added or subtracted as the case may be depending on the direction of the arrows to obtain the difference at closing base of the loop. Ideally, it should be zero, but in all practical surveys a difference is found, which is known as closed loop error (CE). This difference is divided by number of arms that gives the distribution error (DE). The arrow inside the loop points in the direction of the positive loop error. In this case, station number 1 is a primary base that is used to establish secondary bases in the survey block for ease of recording the gravity readings in the survey block. Initial checks for loop closure errors are done in the field itself in order to repeat the erroneous bases in case the loop error is found to be more than the permissible error. The permissible error for the closed loop is $0.02 \sqrt{n}$ mGal, where n is number of arms in a loop. Sazhina

Gravity Method, Surface,

Fig. 3 Closed loops of gravity base stations for loop correction. Station numbers are marked as filled circle with arrows indicating toward the increasing values. The closing loop error (CE) is the departure of vector algebraic sum of the each arm (differences in the observed gravity field at adjoining bases) from zero. Observed gravity difference between two adjoining stations (values of arm) is written on the corresponding arm and CE is also written for each loop. Distribution error (DE) is CE divided by number of arms in the loop, which is added to each loop to arrive final values



and Grushiny (1971) have provided detailed description of these corrections under “Adjustment of Gravity Networks” similar to geodetic networks. Having established the requisite number of base stations in an area, the gravity observations at various stations are tied to these bases within a maximum time limit of 1–2 h for drift correction during that period. In regional surveys where survey is conducted along roads at station spacing of 3–5 km, it may not be possible to make loops. However, as the desired accuracy of the gravity data in this survey is limited to 1 mGal, bases can be tied along the roads at about 20–30 km interval as A-B-A-B which provides two drift-corrected readings at every base. The average of the two readings is taken as the base value. In case of mineral exploration, if the area is small and isolated far away from any known base, an arbitrary base is chosen near the camp with any arbitrary value, for example, 0, 100, 1,000, etc., and all the readings are reduced to this arbitrary base value. This provides the gravity anomalies in the survey block with reference to the arbitrary base. In case it is desired, this arbitrary base can be tied to standard bases in the nearby region even after the survey and observed gravity values at stations are updated with reference to the absolute value of the gravity field at the base station accounting for the arbitrary value adopted previously for the survey.

There are several other corrections like latitude, terrain, elevation corrections, etc., that are applied to the observed gravity field at a station to obtain Bouguer anomaly or simply gravity anomaly at a station (Lafehr 1991; Gotze and Li 1996). These corrections are dealt under data reduction or preparation of gravity anomaly maps.

Marine Gravity Survey

The fundamental difference of marine gravity survey from the land survey is measurements on moving platform, which distorts the measurement accuracy. However, if the speed and accelerations are measured, the corresponding corrections might be applied to the gravity data. Nevertheless, the measurement accuracy on the moving platform like ship will not be as accurate as measurements made on ground. The standard accuracy of marine gravity survey is an order of magnitude less than ground survey. The sea state poses another difficulty in the marine survey; it is apparently calm in the deep water and more predictable than the shallow sea or near the coast. Thus, marine gravity measurements are generally corrected for effects of moving platform, Earth’s tide, drift, and adjusted for crossover tracks as briefly explained below.

Moving Platform Correction

The gravity field measurements are distorted due to movement of ship, as well as due to horizontal and vertical accelerations, which can be corrected and eliminated by using supplementary data. They are classified into three categories (Dehlinger 1978).

- *EOTVOS corrections*: When a ship moves, particularly in the east west direction, it experiences centripetal accelerations about the Earth’s axis that modifies the acceleration due to gravity. This can be precisely corrected if the speed and heading of ship is known.

- *Corrections for horizontal and vertical accelerations:* These corrections are applied to gravity measurements for any horizontal and vertical motion of the platform. In the modern type of gravimeter, they are adjusted by designing stabilized platform or gimbal-suspended systems. They are also removed using GPS data and then averaging of several gravity values.
- *Corrections for cross coupling accelerations:* This correction is applied to a beam type of gravity measuring instruments when it is rigidly attached to stabilized platform. The error occurs when horizontal and vertical components of ship accelerations have equal periods and have a phase difference.

Earth Tide Correction

As discussed above, Earth and ocean are subject to the gravitational attraction of Sun and Moon. Over the land, major part of tidal effect is perturbation of solid Earth; however, over ocean the perturbation of outer water surface also has pronounced effect. Therefore, tidal correction applied to ship-borne gravity data should have a model of both solid Earth and ocean (Van Camp et al. 2017).

Drift Correction

All the marine gravity surveys start with known absolute gravity value at the port of sailing and closed to the known gravity station at the end of survey. Since the time of reoccupation of known gravity bases (reference stations) are generally weeks to month, the drift computed are not very accurate. Nevertheless, all the spring associated gravity meters also have long time drift that are generally taken into account while closing the marine survey though they are often small.

Network or Crossover Adjustments

The problem of offset in the marine gravity survey is quite known. It arises due to several reasons (Wessel and Watt 1988). One of the main reasons of offset is mechanical tare or jar of the spring type of gravity meters. Cross over corrections in the same leg of the marine survey is similar to close loop correction discussed for land survey. In case of marine survey over a block, a cross cutting profile is recorded that is used for instrumental check, drift, and data accuracy.

Design of Survey

Designing of gravity survey is a vital component to optimize the resources according to the geological problems. Following are the details.

Regional Surveys for Geodynamic Studies

As the name indicates, these surveys are basically regional in nature and are conducted in large regions for geodynamic studies especially for crustal and upper mantle structures. These surveys are carried out along roads and tracks at a station spacing of 3–5 km, which delineates long wavelength anomalies related to large-scale structures and density inhomogeneities. Anomalies recorded in these surveys are of the order of tens of mGal (10–100 mGal) spread over a few hundred km (~100–1,000 km). The accuracy of such surveys is limited to ± 1 –2 mGal that can be obtained by controlling elevation of stations using DGPS with accuracy better than 1 m in elevation.

Hydrocarbon Exploration

There are two aspects of gravity surveys for hydrocarbon exploration on land, namely (a) reconnaissance gravity survey and (b) detailed/integrated gravity surveys (Nettleton 1976).

Reconnaissance Gravity Survey

Gravity surveys for exploration of hydrocarbons are basically conducted as reconnaissance surveys to delineate basement structures and structures in sedimentary basins to delimit the areas for seismic surveys. It is generally carried out in a large area covering a specific basin along roads and tracks at a station spacing of approximately 2.0 km. It is planned in such a way that network of stations at about 2.0 km spacing are available in the entire region. The order of gravity anomalies expected from such structures are of a few mGal in amplitude, and therefore, the desired accuracy of such surveys are ± 0.1 mGal. Station elevation is therefore obtained by geodetic leveling or DGPS providing elevation to an accuracy of a few cm (<10 cm). To check the accuracy of recorded gravity value, 5–10% of gravity stations are repeated on different days and preferably by different observers. In ocean, surveys are conducted along profiles 4–5 km apart over a specific block and bathymetry is recorded with echo sounder.

Detailed/Integrated Gravity Survey

While conducting seismic surveys for exploration of hydrocarbons, it is a general practice to follow the same profiles by gravity surveys at closer stations spacing of 0.5–1 km. These surveys are basically conducted for an integrated approach to supplement the information derived from seismic surveys and delineate structures within the sediment. The desired accuracy of such surveys are ± 0.01 mGal (10 μ Gal), and therefore, station location and their elevation to accuracy of a few centimeter (<5 cm) are usually obtained from geodetic surveys or dual frequency DGPS. In case of shallow gas occurrences, such surveys can be conducted over a grid with both profile and station spacing of 10–100 m as the case may be. At

the places where reflection seismic data is poor, gravity survey is conducted to supplement the seismic.

Surveys for Minerals, Groundwater, and Engineering Sites

The application of gravity survey for mineral and groundwater exploration primarily lies in delineating surface/shallow structures such as faults, fractures, etc., which may be mineralized (Parasnis 1997). Same is the case with engineering sites where shallow structures such as faults, fractures, etc., are delineated to avoid those sections for large scale constructions such as dams, power plants, etc. The survey procedures for application to these studies are therefore, similar in logistics. They fall under two categories, namely, reconnaissance and detailed/integrated surveys.

Reconnaissance Survey

This kind of survey is carried out to limit the area of interest from a larger area. It is conducted along profiles separated by a few hundred meters a part (100–500 m) at stations spacing of a few tens of meters (10–100 m). If the area is large and inaccessible to lay down specific profiles, it is conducted along available roads and motorable tracks at a station spacing of a few tens of meters (10–100 m). Due to desired accuracy of gravity data, elevation of stations in these surveys are obtained by geodetic leveling or DGPS, providing elevation to an accuracy of a few centimeters (<10 cm). If the area is not large and accessible, a base line is laid out parallel to the general strike of the geological formation in the center of the block and profiles are laid out perpendicular to it at desired intervals. Stations are marked along the profiles at regular intervals by geodetic survey or using DGPS. Based on these reconnaissance surveys, promising zones and geologically significant structures such as faults are delineated which are followed up by detailed gravity and other geophysical surveys such as electrical, electromagnetic, etc. In case of surveys for engineering sites, same reconnaissance procedure is followed to delimit the area as described above for mineral exploration.

Detailed/Integrated Gravity Survey

Detailed gravity survey for the mineral exploration is basically limited in areas of a few square meters and are carried out in grids with profiles and station spacing of a few meters (1–10 m) or may be even less depending on the size of the area and the target. A baseline is laid out parallel to the general strike of the geological formations, and profiles are laid perpendicular to it at almost same spacing of 1–10 m as per requirement of particular survey. Stations along the profiles are marked by the geodetic survey at the same interval as profiles forming a grid. In certain cases, the observed gravity anomaly may be less than 0.1 mGal, which require survey accuracy of about 0.01 mGal (10 μ Gal).

Summary

Gravity method, surface includes both onshore and offshore measurements of the gravity field of the Earth for geodynamic studies and resource exploration. Accuracy of the gravity survey required for various purposes and precautions to be observed in this regard are discussed. Geodynamic studies to an accuracy of 1–2 mGal can be carried out using normal gravity survey but resource exploration usually requires high-resolution gravity survey to an accuracy of $\sim 10 \mu$ Gal using microGal gravimeters. In such cases, base loop correction in ground surveys and correction for crossover points in marine surveys are essential and are emphasized.

Cross-References

- Earth Tides
- GPS, Data Acquisition, and Analysis
- Gravity Field of the Earth
- Gravity Field, Temporal Variations from Space Techniques
- Gravity Field, Time Variations from Surface Measurements
- Gravity Method, Principles
- Gravity, Data to Anomalies

Bibliography

- Dehlinger P (1978) Marine gravity. Elsevier, Amsterdam
- Gotze HJ, Li X (1996) Topography and geoid effects on gravity anomalies in mountainous areas as inferred from the gravity field of the Central Andes. *Phys Chem Earth* 21:295–297
- Lafehr TR (1991) Standardization in gravity reduction. *Geophysics* 56:1170–1178
- Marson I (2012) A short walk along the gravimeters path. *Int J Geophys*. ID 687813, 9 pages. <https://doi.org/10.1155/2012/687813>
- Mishra DC (2011) Gravity and magnetic methods for geological studies principles, integrated exploration and plate tectonics. BS Publications, Hyderabad. 938 pp. ISBN: 978-93-81075-27-2
- Nettleton LL (1976) Gravity and magnetics in oil exploration. McGraw-Hill, New York
- Niebauer T (2015) Gravimetric methods – absolute and relative gravity meter: instruments concepts and implementation. In: *Treatise on geophysics*. Elsevier, Amsterdam, pp 37–57
- Parasnis DS (1997) Principles of applied geophysics. Chapman & Hall, London
- Sazhina N, Grushiny N (1971) Gravity prospecting. Mir Publishers, Moscow
- Schüller K (2015) Theoretical basis for earth tide analysis with the new ETERNA34-ANA-V4. 0 program. *Bull Inf Marées Terr* 149(12): 024–012
- Tiwari VM, Hinderer J (2011) Time variation of earth's gravity field: surface measurements. In: Gupta H (ed) *Encyclopedia of solid earth geophysics*. Springer, Berlin
- Van Camp M, de Viron O, Avouac J-P (2016) Separating climate-induced mass transfers and instrumental effects from tectonic signal in repeated absolute gravity measurements: uncertainty of absolute gravity data. *Geophys Res Lett* 43:4313–4320. <https://doi.org/10.1002/2016GL068648>

- Van Camp M, de Viron O, Watlet A, Meurers B, Francis O, Caudron C (2017) Geophysics from terrestrial time-variable gravity measurements. *Rev Geophys* 55:938–992. <https://doi.org/10.1002/2017RG000566>
- Wessel P, Watt AB (1988) On the accuracy of marine gravity measurements. *J Geophys Res* 94:7685–7729

Gravity Modeling, Theory and Computation

Jean-Pierre Barriot and Lydie Sichoix
Observatoire Géodésique de Tahiti, University of French
Polynesia, Faaa-Tahiti, French Polynesia

Definition

Gravity modeling Explaining the variations of the external gravity field of a body in terms of its internal density variations. Applications range from oil prospecting to the deep structure of the Earth.

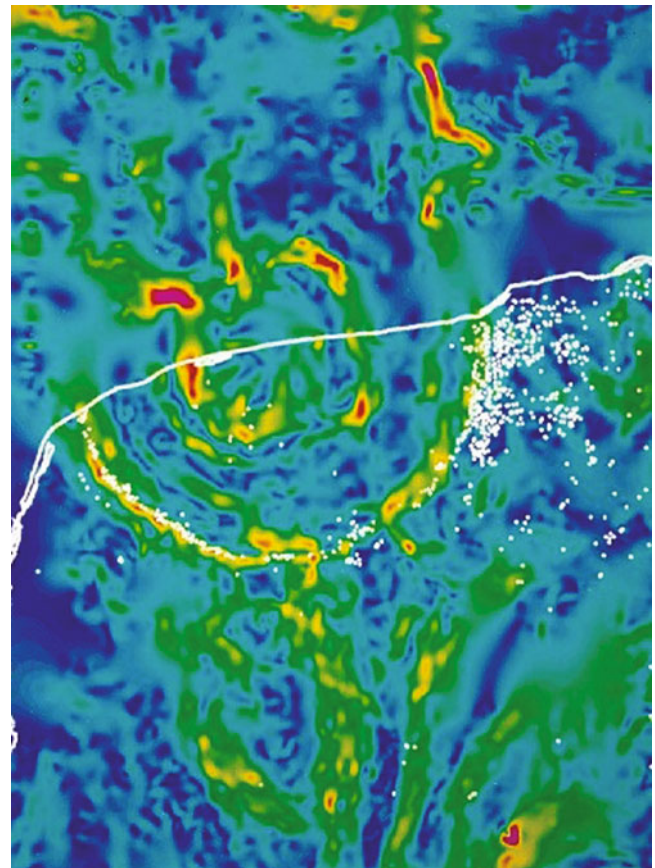
Inverse problem In Gravity modeling, the forward problem is to compute the external gravity field of a body knowing its internal density distribution. The inverse problem is the opposite, and is much more difficult to solve for.

The Gravity Modeling from a Theoretical Point of View

Gravity is an elusive quantity to deal with, but a powerful one. This is mainly because unlike electromagnetic forces, gravity cannot be shielded as there are no negative masses. We refer to LaFehr (1980) and Milsom (1996) for an introduction to the gravimetric techniques (see also ► “Gravity Method, Principles”), and Zidarov (1990) for the subtleties of gravimetric inversion. Good monographs are available on the theory and computational aspects of gravity modeling. The one from Blakely (1996) is particularly exhaustive (see also ► “Gravity Anomalies, Interpretation”). Here we give just a short introduction, with emphasis on the modeling directly done on the data, and not on its spectral transforms.

A landmark success of the application of gravimetric interpretation is the identification of the Chicxulub crater buried underneath the Yucatan Peninsula (Hildebrand et al. 1991, Fig. 1, see ► “Impact Craters on Earth”).

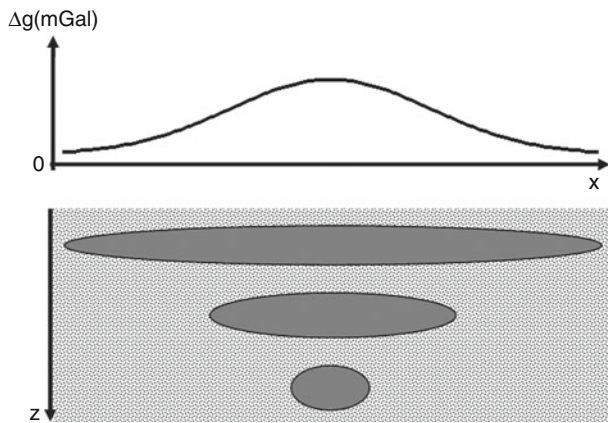
The reasons why gravity is so elusive are profound and lie deep in potential theory. Let us start with a simple example. The radial gravity outside a homogeneous sphere of mass M and radius r is given by GM/R^2 , where G is the Newtonian constant and R the distance from the center of the sphere. But



Gravity Modeling, Theory and Computation, Fig. 1 The Chicxulub structure (diameter 180 km) is widely believed to be the remnant of the impact that killed the dinosaurs 65 millions of years ago. This structure was identified on a gravity map of the coast, with no other visible evidence. The impacting asteroid that formed the crater was at least 10 km in diameter. (Image source: Shuttle radar topography mission, Chicxulub Crater Gravity Map with Cenotes Shown as White Dots with the Yucatan Coastline Being the White Line)

r does not intervene in the computation, as long as $R > r$. This means that two homogeneous spheres of radius r and r' have the same external field, as long as they have the same mass. This non-uniqueness has been studied in detail and theory shows that a 3D repartition of density that causes a gravity anomaly can be split in two additive parts: one that is harmonic, that is, that satisfies the Laplace equation, and another that is non-harmonic. Only the harmonic part of the anomalous density can be unambiguously reconstructed from the anomalous gravity signal (Moritz 1989; Vanicek and Kleusberg 1985; Ballani et al. 1993). The non-harmonic part of the density belongs to the so-called null-space of the model. In terms of geophysics, a typical example of non-uniqueness is shown in Fig. 2.

Far more useful is to note that the ambiguity in gravity modeling can be reduced, or even reduced to zero if we know the support, or mathematically, the spatial extension of the density anomaly. For example, if we know by other means the



Gravity Modeling, Theory and Computation, Fig. 2 Each of these underground bodies (gray) creates the same gravity anomaly. If one supposes that the body is reduced to a point mass, one can infer from this assumption the maximum possible depth for the causative body. (Image source: Dubois and Diamant 2005, redrawn)

radius r of the homogeneous sphere, the density ρ is automatically fixed, with $g = GM/R^2$ by $M = 4/3 \times \pi \times \rho \times r^3$.

Chao (2005) has extended this reasoning to the modeling of gravity in terms of spherical harmonics (see ► [“Spherical Harmonic Analysis Applied to Potential Fields”](#)). If we consider a 2D spherical skin covering the Earth, the inverse solution of the surface density distribution is unique. The latter applies quite readily to the inversion of time-variable gravity signals originating from within the crust (such as those observed by the GRACE/GOCE space missions, see ► [“Gravity Method, Satellite”](#)).

Another possibility is to suppose that the density known a priori. In that case, the radius of the sphere instantly follows. This seems trivial, but it is not. It has been demonstrated that the gravity field outside a body can be totally explained by a combination of point masses, linear masses, and sheet masses that lie deep inside the body (Zhdanov 1988). But here also, this result is of little help for the geophysicist. Moreover, the inverse gravity problem is unstable: small variations in the data (errors) can cause unbounded variations in the solution. Technically, the inverse gravity problem is ill-posed in Hadamard’s sense (1902, see also Sabatier 2000 and Menke 1989, and ► [“Inverse Theory, Global Optimization”](#)).

The core fact that permits a useful application of gravity anomalies in geophysics is that these anomalies, even contaminated by noise, reflect *real* structures, that is, structures that *we know* have a reasonable density contrast with the surroundings, and/or exhibit a density that varies smoothly inside the expected structure – except at the edges or along faults – and/or are located inside a given volume. For example, we know from borehole measurements that an iron ore is located between two depths: one at $-1,000$ m and the other one at $-2,000$ m. To have bounds on the possible densities

and/or the locations reduce tremendously the world of possibilities. In short, we are then looking for *plausible* structures.

With such cautions and restrictions, gravity modeling is a powerful tool if correctly used. The best use of gravity measurements is to team them up with other measurements with different null-spaces, like magnetic surveys or even seismic imaging. For example, the velocity of seismic waves increases with density, and the anomalous delays of propagation are linked to the anomalous velocity field (and so to the density through an equation of state) by an averaging integral along the ray known as the Radon transform (Trojan and Hayakawa 2002). A joint seismic/gravity inversion is a powerful tool and probably the best way to remove non-uniqueness (Strykowski 1998, see also ► [“Seismic Tomography”](#)).

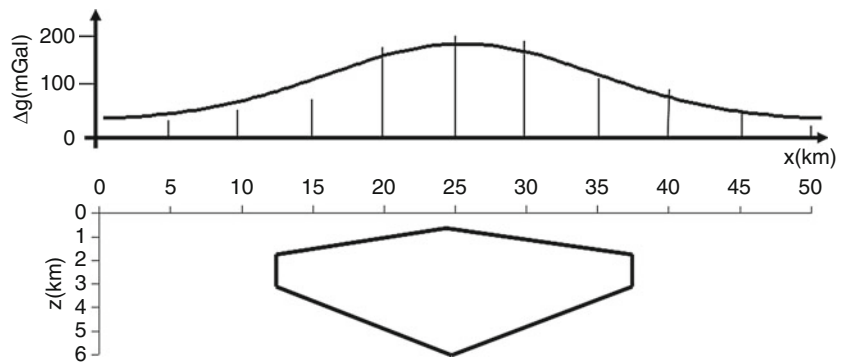
But this is hardly done in practice. The driver for field measurements is cost. A gravity survey, even on remote areas, is cheap, and often done as a precursor to more expensive and accurate methods, like seismic tomography. Gravity modeling thus allows to identify promising structures on its own. The game is tricky, as the observed gravity is of course affected by noise, but also by data processing (see ► [“Gravity, Data to Anomalies”](#)). Data processing consists for example of correcting the gravity data for drifts, referring to some “constant” reference in altitude (usually the geoid), and removing the gravity signal of the topography (Bouguer correction, see ► [“Gravimeters”](#)). Besides, the gravity signal is by nature nonlocal, and sums up the signature of other distant bodies or interfaces, like the Mohorovic discontinuity. These distant signatures must be guessed and withdrawn from the gravity signal *before* any local geological interpretation is undertaken (see ► [“Gravity Data, Regional: Residual Separation”](#)).

There are two ways to do this. One is to use direct modeling with feedback. Technically, an operator draws a 2D *educated* guess of the plausible structure (see Fig. 3) on a tablet or other graphic device, and the computer program derives the anomalous gravity model caused by the structure and draws it against the real signal. The operator then modifies slightly the guessed structure in order to obtain a better match with the real signal. For the geophysicist, “anomalous gravity” is synonymous of “Bouguer anomaly,” that is to say a measurement of gravity corrected for the variation of altitude and for the topography-induced variations. Other definitions of “anomalous gravity” are possible. For example, it is theoretically possible to work indifferently with geoid undulations or Bouguer anomalies. In practice, the modeling of geoid undulations is reserved for very-large-scale structures in isostatic equilibrium (see ► [“Isostasy”](#)), while Bouguer anomalies usually refer to small- and medium-scale structures (see ► [“Gravity, Data to Anomalies”](#) and ► [“Geoid Undulation, Interpretation”](#)).

This process can be automated with the computing code adjusting step by step the shape of the structure to exactly

Gravity Modeling, Theory and Computation, Fig. 3

The geological body is approximated by a polygonal structure of infinite length perpendicular to the paper sheet and of constant density contrast with the surroundings. The game is to move the summits of the polygon in order to match (solid curve) the observed gravity (needles)

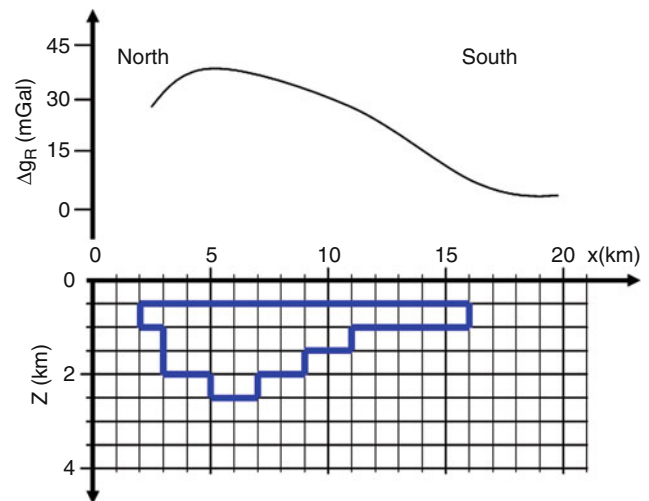


match the observed gravity anomaly. With some more complexity in terms of calculus, 3D modeling can be done in almost the same manner, with polyhedra replacing polygons. But there is a huge difference at another level. A human operator has always trouble to visualize and modify 3D solid models, even with the assistance of powerful 3D graphic representations. Such a tedious tool is used with parsimony. Efficient formula that replaces volume integrals by integrals along vertices have been developed by several investigators, from Talwani et al. (1959), to Tsoulis and Petrovi (2001) to implement polygon/polyhedron calculus.

For 3D modeling, a better approach is to discretize the volume where we are seeking for the geological structure in tiny adjacent cells of simple fixed shape (e.g., cubes) but of unknown densities. It is important to note the difference with the previous approach: by looking at the densities, we obtain a mathematically linear problem, in the sense that the effect of each cell sums up additively in the gravity model. To the contrary, altering the skeleton of a polyhedron of constant density (even only a single corner point) has a non-additive effect on the computed gravity. Technically, we are dealing with a non-linear model. Non-linear problems are intrinsically much more difficult to solve for.

The Linear Inverse Gravimetric Problem

Let us focus on the linear gravimetric inverse problem, illustrated in Fig. 4. We have to define how the discrepancies between measurements and computed gravity from the model are accounted for. The usual way is to measure this discrepancy as a Euclidean distance, by summing up the square of the differences per data point between measurements and computed gravity, and then taking the square root of the sum. This is called the L2 norm of discrepancies. Another way is to sum up the absolute values of the discrepancies. This is called the L1 norm of discrepancies. Specialized computer languages (not specialized libraries) able to



Gravity Modeling, Theory and Computation, Fig. 4 The linear inverse gravimetric problem. The subsurface is divided in small squares (2D, or pixels) or small cubes (3D, or voxels), and we are looking for the anomalous densities for each square or cube. (Image source: Blakely 1996, redrawn)

directly tackle L2 norm problems and Bayesian inferences have recently being commercialized (Davidson-Pilon 2016). Approaches where the number of parameters in the inversion process is not kept fixed (Transdimensional Bayesian Algorithms with Monte-Carlo framework) are also gaining momentum (Izquierdo et al. 2020). See also in this Encyclopedia ► “Inverse Theory, Monte Carlo Method.”

The set of all unknown densities x over the mesh and the set g of the gravity measurements on the ground are linked through the matrix relation $Ax = g$. In the frame of the L2 norm, solving for the ill-posed inverse linear gravimetric problem, consists of finding a set x of densities that verifies at the best this relation, and fulfilling at the same time another relation of the form $Bx = k$, where k is another physical or mathematical quantity. For example k can be magnetic measurements and B a design matrix linking densities to magnetic measurements. If we add in such way “sufficient” physical information, the solution of this augmented problem becomes

unique. Very often, additional physical information is lacking, and uniqueness is imposed by setting $B = I$ and $k = 0$, where I is the identity matrix. This is equivalent to looking for a set x of densities that realizes a compromise between $Ax = g$ and $x = 0$, that is, an x that explains the gravity signal and is not too large (i.e., close to zero). This is known in the literature as a Tikonov–Arsenine (1976) regularized inverse or Marquardt–Levenberg inverse (Levenberg 1944; Marquardt 1963). Technically, the solution depends also on the way that the hidden Euclidean spaces used to set up the problem are defined. They are not necessarily orthonormal spaces. The angular separation of their axes and unit lengths can carry additional information about the physical problem (called in the literature as “a priori weights”). Tarantola (2005) gave a complete survey of the probabilistic setting for the Tikonov–Arsenine formula, with the weights seen as the inverses of a priori covariance matrices. The covariance linked to I is especially important as it controls to which extent the density contrasts found are “reasonable.” A density contrast of 0.2 is sound; a density contrast of 20 is not, even if the computed gravity matches more closely the observed signal!

The simplest form of the Tikonov–Arsenine inverse formula is of the form:

$$x^*(p) = \text{inv}(\text{trans}(A)A + pI)\text{trans}(A)g \quad (1)$$

where $x^*(p)$ is the solution of the inverse problem (here densities over the mesh), A is the design matrix defined above, p is a scalar parameter, I the identity matrix, and g the set of gravity measurements. In the Tarantola–Valette approach, p is the ratio between the variance of the noise error and the “variance” (the size) of the expected solution with respect to zero. The scalar p is the “smoothing” parameter. The geophysicist chooses “by hand” p in order to have both “reasonable” L2 residuals, defined as $\text{trans}(Ax^* - g)(Ax^* - g)$, and a reasonable L2 size (densities) of x , defined as $\text{trans}(x^*)x^*$.

This solution x^* minimizes, with a unique minimum, the following functional (or cost function) $Q(x)$ with respect to x :

$$Q(x) = \text{trans}(Ax - g)(Ax - g) + p \text{trans}(x)x \quad (2)$$

Without doubt, the celebrated Tikonov–Arsenine formula is the workhorse of gravity inversion.

Another approach, also very popular, but un-implementable for large problems, is to look at the so-called singular value decomposition of A , a generalization of the eigenvalue decomposition of square matrices (Campbell and Meyer 1979, see ► “Inverse Theory, Singular Value Decomposition”). The use of parallel computing and especially GPU (Graphical Processing Units) is a recent way to alleviate the computational burden (Akimova et al. 2018).

Looking at a solution of the inverse problem in terms of the L1 norm seems just a little adjustment with respect to the L2 norm. However, it is a totally different approach. Let us imagine that we have a set of n gravity measurements and a mesh of m cells, assuming that n is larger than m . Technically, the problem is overdetermined, with a matrix A that has many more lines than columns. To solve for this system in the L1 sense consists of solving with respect to x each subset of m lines of A (each subsystem is now square with a trivial inverse), and looking at the solution that gives the best L1 residuals with respect to the complete matrix A . There are $n!/m!(n-m)!$ subsystems. For 1,000 measurements and 50 cells, a computer crushing numbers from the dawn of the universe will have trouble to handle this inversion. Thankfully, Dantzig and colleagues (see Branham 1990) found a way to carry out the L1 inversion with on-the-shelf computers, called the simplex algorithm (linear programming, see ► “Inverse Theory, Linear”). This type of inversion of gravity data has also to be supplemented with a priori (external) constraints to obtain a geologically “sound” solution, for example, by imposing that the inertia moments of the repartition of densities are minima (Cuer and Bayer 1980), or by looking at extrema solutions, for example, the “ideal” body of Parker (1975, see also Goodacre 2006), that is, the set of densities characterizing the geological structure that gives a minimum density contrast with the surroundings.

This type of inverse modeling has been revived recently by the use of neural networks. Neural networks are nothing new, as they were introduced in 1958 (see Widrow and Lehr 1990, for a review of the early developments). The main idea, for the linear case, is first to have a fixed discretization of the subsurface, like the subsurface mesh shown in Fig. 4. Secondly, families (indexed by an integer i) of pixels or voxels densities are generated over this mesh, following a given strategy (for example approximating an ore deposit). Each family can be arranged as a vector “signal” X_i and generates a corresponding Y_i gravity signal (as the profile A in Fig. 4). It then can be demonstrated that the best forward linear link $Y = M X$ between the two families, ordered as matrices in consecutive columns $X = \{X_i, \text{ for all } i\}$ and $Y = \{Y_i, \text{ for all } i\}$ is given by $M = Y X^+$ where the $+$ denotes the pseudo-inverse (or Moore–Penrose inverse) of a matrix (Campbell and Meyer 1979). The best inverse linear link $X = N Y$ is symmetrically given by $N = X Y^+$ (see an application in Barriot et al. 1999, Annex B). This a supervised learning procedure in the modern jargon of artificial intelligence, and the inverse operator N can be used to obtain a best guess of an unknown x from a given measured y by $x = N y$. We can say that N represents an a priori knowledge about the expected subsurface repartition of densities, as it is built from a family of acceptable, from a geophysicist point of view, repartitions of subsurface densities. We refer to Shimelevich et al. (2019) for a recent application in gravimetry and geoelectrics modeling, and to

Valentine and Sambridge (2020a, b) for an overview of inverse gravimetric problems in the light of machine learning.

The Nonlinear Inverse Gravimetric Problem

In this section we come back to the polygon/polyhedron modeling of gravity anomalies, where we assume that the density of the body is uniform.

Mathematically, the problem expressed as $g = f(x)$, where now x is the set of coordinates of the corner points, and is nonlinear. We want to find a set of coordinates that explain “at the best” the gravity signal, that is, $g(\text{observed}) - f(x)$ is kept to a minimum, for example, with respect to the L2 norm.

We then have to minimize a cost function of the form:

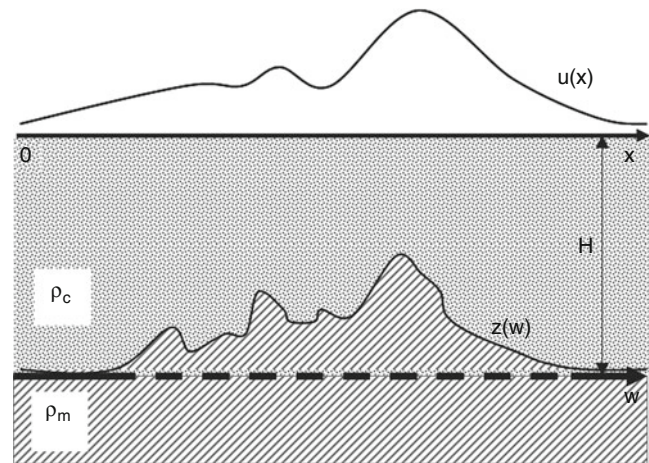
$$Q(x) = \text{trans}(g - f(x))(g - f(x)) \quad (3)$$

There is an important difference with the linear case. We can have several well-separated, inherent minima to this functional (Fig. 5). There are two ways to solve for this inverse problem. The first one is to linearize the problem, that is, to look at a solution close to a known estimate x_0 , that is, by writing that $x = x_0 + dx$, with a dx bounded in some way.

A typical illustration in gravimetry is to look at the shape of an unknown interface between two strata of different and constant densities (see Fig. 6). This is an extension of the polygonal model where some points are pushed to infinity and has been discussed by Tarantola and Valette (1982) in their milestone paper, an extension of the pioneering work of Cordell and Henderson (1968). This is often used to refine a model obtained by the modeler from a priori geological evidence, but can be hardly used without any a priori knowledge of the solution. The Backus and Gilbert approach (1968), very popular during the 1970s, can also be seen as a nonlinear problem.

The other approach, more radical, consists in trying to explore the whole set of possible minima of the nonlinear cost-functional. A popular approach to do so is to mimic the biological process of evolution. The genetic algorithm of Holland (1975) can be summarized in four basic steps: coding, selection, crossover, and mutation.

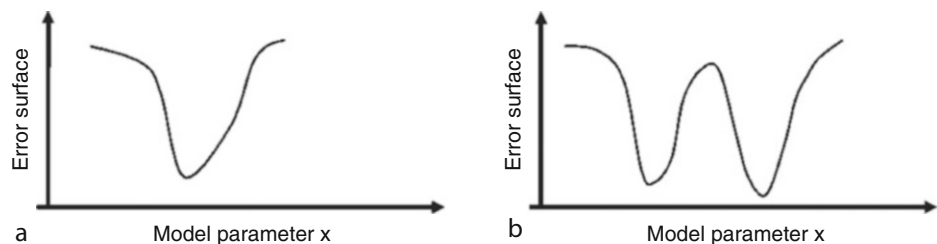
Common to any genetic algorithm is the digitization of an a priori and sufficiently populated list of model parameters using a binary code scheme (for example, we know that the density we are looking for is bounded by 2.5 and 2.8). Assume that the desired resolution (discretization) is 0.01. Then we can make a one-to-one correspondence between a binary word of 5 bits (a “chromosome”) and this discretization. A subset of models is then selected from the model population by using some global fitness criterion, for example, the ratio of each model fitness (evaluated in the L2 sense with respect to the observed gravity and using the one-to-one correspondence) to the sum of all fitness criteria for all models. We then take the models the ratio of which is over a given threshold. For example, for a population of L models, we choose a subset of $L/2$ of “best fit” models. We then mate the $L/2$ models by pairs, each pair bearing two off-springs by recombination of their chromosomes by processes called crossover and mutation that mimic the meiosis of real chromosomes, giving birth to a complete new generation of L models that replaces the old one. Generally, the average fitness increases between generations, since only the best models from each generation are selected for breeding. Other algorithms exist to locate the “best” minimum, like



Gravity Modeling, Theory and Computation, Fig. 6 Modeling an interface between two homogeneous separate strata. The problem is a special case of the polygon/polyhedron inverse problem. (Image source: Tarantola and Valette 1982, redrawn)

Gravity Modeling, Theory and Computation, Fig. 5

(a) The cost function of the linear inverse problem presents only one minimum. (b) The cost function of the nonlinear inverse problem may present several minima, indicating a non-unique solution. (Image source: Dimri 1992, redrawn)



the tunneling algorithm (Levy and Montalvo 1985) or the particle swarm optimization (Monteiro Santos 2010). Neural networks can also be applied to the non-linear case, but it is too early to say whether this is a fashion effect in gravimetric modeling or will lead to real practical tools. We refer to LeCun et al. (2015) for an in-depth review of the latest advances in neural networks and deep learning. Powerful no-cost computer libraries like TensorFlow for deep learning applications are now in wide use (Abadi et al. 2015). Another kinds of supervised learning algorithms are the so-called Support Vector Machines (SVM) classifiers. An application of SVM to gravimetric modeling can be found in Wu et al. (2007). Unsupervised learning as also being attempted (Levassor et al. 2009). See also in this Encyclopedia ► “Inverse Theory, Artificial Neural Networks” and ► “Inverse Theory, Monte Carlo Method”.

Spectral Approaches

Spectral approaches rely on the expression of the gravity signal along a basis of given functions: very often a Fourier decomposition in the planar case or spherical harmonics on a global basis. New basis are gaining a wide acceptance, like wavelets, that are like “sine” or “cosine” waves but with a limited spatial support (Chui 1992). Large classes of inverse gravimetric problems can then be recast as deconvolution problems (Dimri 1992), with all the artillery available from Fourier analysis (Dean 1958; Blakely 1996). Nevertheless, this approach suffers from all the drawbacks of the Fourier analysis, especially the need to smooth the discontinuities of the gravity signal at the edges of the zone of interest (this is of course not the case for spherical harmonics). Such approaches have been used in planetary geophysics to estimate the global and local average depth of the Moho interface on Mars (Zuber et al. 2000) using spherical harmonics modeling of the gravity field and of the topography of the planet, without any other surface or subsurface information. Let us also mention the use of wavelets that permits a multiscale regularized deconvolution of the density variations in the Earth’s crust (Volker 2005).

Summary

Gravity modeling is a cheap tool in geophysical prospecting. It can also be a very potent tool if used with some precaution. This always implies the introduction, both in forward and inverse modeling, of some a priori knowledge of the shape of structures that are guessed, or some bounds about their dimensions or about acceptable density contrasts.

Cross-References

- [Geoid Undulation, Interpretation](#)
- [Gravimeters](#)
- [Gravity Anomalies, Interpretation](#)
- [Gravity Data, Regional: Residual Separation](#)
- [Gravity Method, Principles](#)
- [Gravity Method, Satellite](#)
- [Gravity, Data to Anomalies](#)
- [Impact Craters on Earth](#)
- [Inverse Theory, Artificial Neural Networks](#)
- [Inverse Theory, Global Optimization](#)
- [Inverse Theory, Linear](#)
- [Inverse Theory, Singular Value Decomposition](#)
- [Isostasy](#)
- [Seismic Ray Theory](#)
- [Seismic Tomography](#)
- [Spherical Harmonic Analysis Applied to Potential Fields](#)

Bibliography

- Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jozefowicz R, Jia Y, Kaiser L, Kudlur M, Levenberg J, Mané D, Schuster M, Monga R, Moore S, Murray D, Olah C, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: large-scale machine learning on heterogeneous systems. Software available from [tensorflow.org](https://www.tensorflow.org)
- Akimova E, Misilov V, Sultanov M (2018) Solving the structural inverse gravimetry problem in the case of multilayered medium using GPU. In: Proceedings of the Ural-PDC 2018 workshop on parallel, distributed, and cloud computing for young scientists, Yekaterinburg
- Backus GE, Gilbert JF (1968) The resolving power of gross earth data. *Geophys J R Astron Soc* 16:169–205
- Ballani L, Engelis J, Grafarend G (1993) Global base functions for the mass density in the interior of a massive body (Earth). *Manuscr Geodaet* 18(2):99–114
- Barriot J-P, Kofman W, Herique A, Leblanc S, Portal A (1999) A two dimensional simulation of the CONSERT experiment (radio tomography of comet Wirtanen). *Adv Space Res* 24(9):1127–1138
- Blakely RJ (1996) Potential theory in gravity and magnetic applications. Cambridge University Press, Cambridge. ISBN 0-521-57547-8
- Branham RL Jr (1990) Scientific data analysis: an introduction to over-determined systems. Springer, New York. ISBN 0-387-97201-3
- Campbell SL, Meyer CD Jr (1979) Generalized inverses of linear transformations, Surveys and reference works in mathematics, vol 4. Pitman, Boston. ISBN 0-273-08422-4
- Chao BF (2005) On inversion for mass distribution from global (time-variable) gravity field. *J Geodyn* 39(3):223–230
- Chui CK (1992) An introduction to wavelets. Academic, San Diego
- Cordell L, Henderson RG (1968) Iterative three-dimensional solution of gravity anomaly data using a digital computer. *Geophysics* 33:596–601
- Cuer M, Bayer R (1980) Fortran routines for linear inverse problems. *Geophysics* 45:1706–1719

- Davidson-Pilon C (2016) Bayesian methods for hackers: probabilistic programming and Bayesian inference. Addison-Wesley, New York. ISBN-13 978-0133902839
- Dean WC (1958) Frequency analysis for gravity and magnetic interpretation. *Geophysics* 23:97–127
- Dimri V (1992) Deconvolution and inverse theory: application to geophysical problems, methods in geochemistry and geophysics. Elsevier, Amsterdam. ISBN 0-444-89493-4
- Dubois J, Diamant M (2005) Géophysique – Cours et exercices corrigés, 3ème édition. Dunod, Paris. ISBN 9782100491858
- Goodacre AK (2006) Estimation of the minimum density contrast of a homogeneous body as an aid to the interpretation of gravity anomalies. *Geophys Prospect* 28(3):408–414
- Hadamard J (1902) Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton Univ Bull* 13:49–52
- Hildebrand AR, Penfield GT, Glen T, Kring DA, Pilkington M, Zanutto AC, Jacobsen SB, Boynton WV (1991) Chicxulub Crater: a possible Cretaceous/Tertiary boundary impact crater on the Yucatan Peninsula, Mexico. *Geology* 19(9):867–871
- Holland JH (1975) Adaptation in natural and artificial systems. Michigan University Press, Ann Arbor
- Izquierdo K, Lekić V, Montési LGJ (2020) A Bayesian approach to infer interior mass anomalies from the gravity data of celestial bodies. *Geophys J Int* 220(3):1687–1699. <https://doi.org/10.1093/gji/ggz544>
- LaFehr TR (1980) Gravity method. *Geophysics* 45(11):1634–1639. ISSN 0016-8033
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444. <https://doi.org/10.1038/nature14539>
- Levassor F, Foudil Bey N, Caumon G (2009) Kohonen's neural networks: application to determine geological models from gravimetric datasets. In: Proceedings of 29th Gocad meeting, Nancy
- Levenberg K (1944) A method for the solution of certain problems in least squares. *Q Appl Math* 2:164–168
- Levy AV, Montalvo A (1985) The tunneling algorithm for the global minimization of functions. *SIAM J Sci Stat Comput* 6(1):15–29
- Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. *SIAM J Appl Math* 11:431–441
- Menke W (1989) Geophysical data analysis: discrete inverse theory, International geophysics series, vol 45. Academic, San Diego. ISBN 0-12-490921-3
- Milson J (1996) Field geophysics. Wiley, Chichester. ISBN 0-471-96634-7
- Monteiro Santos FA (2010) Inversion of self-potential of idealized bodies anomalies using particle swarm optimization. *Comput Geosci* 36(9):1185–1190
- Moritz H (1989) A set of continuous density distributions within a sphere compatible with a given external gravitational potential. *Gerlands Beitr Geophysik* 98:185–192
- Parker RL (1975) The theory of ideal bodies for gravity interpretation. *Geophys J R Astron Soc* 42:315–334
- Sabatier PC (2000) Past and future of inverse problems. *J Math Phys* 41(6):4082–4124
- Shimelevich MI, Osborne EA, Osborne IE, Rodionov EA, Lyakhovets DA (2019) Formalized inversion of geophysical data using neural network technologies with application to the tasks of geoelectrics and gravimetry. In: Proceedings of the 15th conference in engineering and mining geophysics, Apr 2019, vol 2019. European Association of Geoscientists & Engineers, pp 1–5. <https://doi.org/10.3997/2214-4609.201901717>
- Strykowski G (1998) Some technical details concerning a new method of gravimetric-seismic inversion. In: Proceedings of the XXIII General Assembly of the European Geophysical Society, Nice. Physics and chemistry of the earth, vol 24. pp 207–214
- Talwani M, Worzel JL, Landisman M (1959) Rapid gravity computations for two-dimensional bodies with application to the Mendocino submarine fracture zone. *J Geophys Res* 64:49–61
- Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. SIAM, Philadelphia. ISBN 978-0-898715-72-9
- Tarantola A, Valette B (1982) Generalized nonlinear inverse problems solved using the least squares criterion. *Rev Geophys Space Phys* 20(2):219–232
- Tikhonov A, Arsenine V (1976) Méthodes de Résolution des Problèmes Mal Posés. Editions Mir, Moscou
- Troyan VN, Hayakawa M (2002) Inverse geophysical problems. Terrapub Editor, Tokyo. ISBN 4-88704-131-4
- Tsoulis D, Petrovi S (2001) Short Note On the singularities of the gravity field of a homogeneous polyhedral body. *Geophysics* 66(2):535–539
- Valentine AP, Sambridge M (2020a) Gaussian process models – I. A framework for probabilistic continuous inverse theory. *Geophys J Int* 220(3):1632–1647. <https://doi.org/10.1093/gji/ggz520>
- Valentine AP, Sambridge M (2020b) Gaussian process models – II. Lessons for discrete inversion. *Geophys J Int* 220(3):1648–1656. <https://doi.org/10.1093/gji/ggz521>
- Vanicek P, Kleusberg A (1985) What an external gravitational potential can really tell us about mass distribution. *B Geofis Teor Appl XXCII* (108):243–250
- Volker M (2005) Regularized wavelet-based multiresolution recovery of the harmonic mass density distribution from data of the Earth's gravitational field at satellite height. *Inverse Probl* 21:997–1025. <https://doi.org/10.1088/0266-5611/21/3/013>
- Widrow B, Lehr M (1990) 30 years of adaptive neural networks: perceptron, Madaline, and backpropagation. *Proc IEEE* 78(9):1415–1442
- Wu S-Y, Wang Y-J, Li L, Hu J-S, Feng G-Z, Yan H-J (2007) The application of support vector machine in the joint inversion of gravimetric and seismic data. *Prog Geophys* 22(5):1611–1616
- Zhdanov MS (1988) Integral transform in geophysics. Springer, Berlin. ISBN 3-540-17759-0
- Zidarov D (1990) Inverse gravimetric problems in geoprospecting and geodesy. Elsevier, Amsterdam. ISBN 0-444-98777-0
- Zuber MT, Solomon SC, Phillips RJ, Smith DE, Tyler GL, Aharonson O, Balmino G, Banerdt WB, Head JW, Johnson CL, Lemoine FG, McGovern PJ, Neumann GA, Rowlands DD, Zhong S (2000) Internal structure and early thermal evolution of Mars from Mars global surveyor topography and gravity. *Science* 287(5459):1788–1793

Gravity, Data to Anomalies

Ron Hackney

Petroleum and Marine Division, Geoscience Australia,
Canberra, ACT, Australia

Definition

Gravity anomaly. The difference between gravity measured at a point and a model value at that point that is based on the normal gravity of a reference ellipsoid, corrected for the gravity effects of elevation above the reference ellipsoid and the mass of rock between the point and the reference ellipsoid.

Introduction

The study of anomalous gravity has its roots in geodesy where it is used to determine the shape of the Earth (see ► “Geodesy, Physical” and ► “Geodesy, Figure of the Earth”). Gravity anomalies have also proved extremely useful in the interpretation of subsurface geological structure at various scales. Like many geophysical techniques, resource exploration has been the greatest driver of the use of gravity data (e.g., Nabighian et al. 2005). Gravity anomalies are often useful in the early stages of an exploration program as they provide insight into the form of low-density sediment accumulations (basins) or the location of high-density ore deposits. At larger scales, gravity anomalies provide insight into crustal structure and, most recently, time variations in the long-wavelength gravity field measured using satellites (see ► “Gravity Method, Satellite”) provide insight into changes in polar ice mass and ground-water fluctuations.

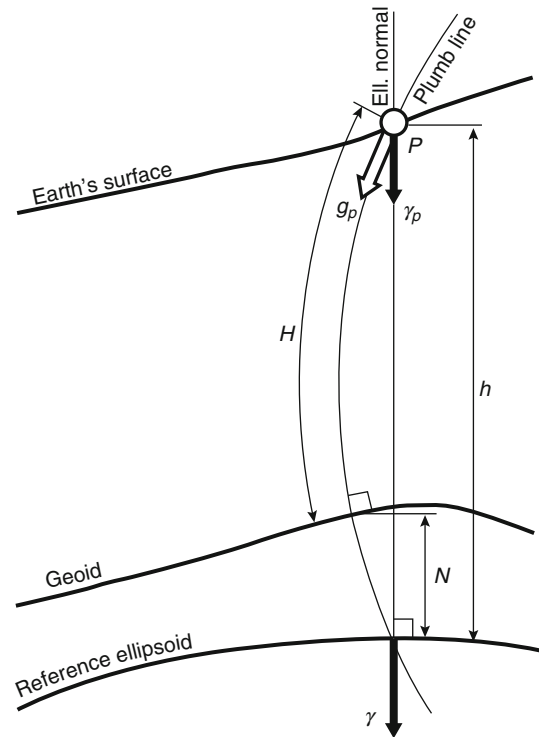
Before gravity data are useful in resource exploration or crustal studies, non-geological contributions to the gravity field that obscure the anomalies of interest need to be accounted for and effectively removed. This chapter explains the process of converting gravity measurements to gravity anomalies for geophysical applications. It is important to stress the geophysical context as there are often subtle differences in the way geodesists and geophysicists think of and use the various gravity-related quantities (cf. Li and Götze 2001; Hackney and Featherstone 2003; Hofmann-Wellenhof and Moritz 2006).

Key Parameters

Figure 1 shows the key parameters required for computing anomalous gravity. The relevant gravity quantities for geophysical applications are:

- Gravity measured at the point P , g_P
- Normal gravity on the surface of the reference ellipsoid, γ
- A model or predicted gravity value at the measurement point P , γ_P , derived from γ

The elevation of the point P is also required to account for the decrease in gravity with increasing elevation and the gravity effect of the mass of rock between the measurement point and the reference ellipsoid. Two height quantities need to be considered (Fig. 1). The height of the point P above the geoid, H , is measured along a plumb line, a curved line that is perpendicular to equipotential surfaces (the geoid is the equipotential surface that most closely approximates mean sea



Gravity, Data to Anomalies, Fig. 1 Diagram illustrating the key parameters required for computing geophysical gravity anomalies

level – see ► “Geoid Determination, Theory and Principles”). The ellipsoidal height, h , is measured along an ellipsoidal normal, the line through P that is perpendicular to the surface of the reference ellipsoid. When computed algebraically (i.e., disregarding the vector nature of gravity), the difference between h and H gives the geoid–ellipsoid separation, N , where

$$N = h - H. \quad (1)$$

Observed Gravity

The measured value of gravity at a point P is the value typically determined from relative gravity measurements that are converted to an observed value by tying to an international network of reference gravity stations, currently the International Gravity Standardization Network of 1971 (IGSN71; Morelli et al. 1974). Country-specific datums may also be used (e.g., Tracey et al. 2007), but these are also generally tied to IGSN71. The relative measurements may be made on land or at sea (see ► “Gravity Method, Surface”) or even in the air (see ► “Gravity Method, Airborne”). The relative measurements will have been corrected for instrument drift and Earth tide variations (see ► “Gravity Method, Surface”).

Anomalous Gravity in Geophysics

An anomaly is the difference between something that we measure at a particular place and what we expect from a reference model at that place. In seismology, measured travel times are compared to those expected from simplified models of the whole Earth (see ► “Seismic Tomography”). The resulting travel-time anomalies tell us about regions through which seismic waves travel faster or slower than through the simplified representation of reality. This in turn tells us something about the temperature or composition of the Earth’s interior. In geophysical applications, gravity anomalies are best thought of in a similar way: Measurements of the gravity field at the Earth’s surface are compared to the gravity expected from a reference model of the Earth. The differences in gravity between our measurements and the reference model tell us something about lateral mass-density variations within the Earth.

In this context, a geophysicist generally seeks to remove large-scale, non-geological gravity effects that mask the gravity anomalies that can be related to subsurface density variations (e.g., related to ore deposits, sedimentary basins, or crustal structure). The large-scale effects include Earth mass and rotation and, depending on the application, maybe some or all of the crust and mantle. Most of the large-scale effect is accounted for by normal gravity, the acceleration due to gravity on the surface of a reference ellipsoid, an ellipsoid of rotation that best fits the shape of the Earth (represented by the geoid – see ► “Geodesy, Figure of the Earth”).

If a model value is required at a point that does not lie on the surface of the reference ellipsoid, adjustments must be made to normal gravity to account for gravity effects related to (a) the difference in height between the measurement point and the surface of the reference ellipsoid (because gravity decreases with distance from the Earth proportionally to $1/r^2$) and (b) the body of rock (or water) lying between the measurement point and the reference ellipsoid. Other regional-scale gravity effects that arise from crustal structure might also be removed by a process of regional–residual field separation (see ► “Gravity Data, Regional: Residual Separation”) or filtering (see ► “Gravity Data, Advanced Processing”).

In the geophysical context, anomalous gravity is calculated by subtracting a predicted or model gravity value for a point P from the measured (observed) gravity at that point. This anomalous gravity quantity is technically referred to as a gravity disturbance (cf. Hackney and Featherstone 2003), but the use of gravity “anomaly” is and will likely remain entrenched in the geophysical vernacular. In a geophysical sense, a gravity anomaly can be expressed as

Gravity anomaly = (Measured value) – (Model value), or mathematically as:

$$\Delta g_P = g_P - \gamma_P \quad (2)$$

where

- Δg_P is the anomalous gravity
- g_P is measured gravity at the point P
- γ_P is a predicted or model value of gravity at the measurement point P (our best estimate of the Earth’s gravity field at P that considers Earth’s mass and rotation and the elevation, h , of the point P with respect to the reference ellipsoid)

As illustrated in Fig. 2, the predicted gravity value, γ_P is computed by applying corrections to normal gravity that account for the height of the measurement point above the reference ellipsoid (Free air correction) and the mass of rock that lies between the measurement point and the reference ellipsoid (Bouguer and terrain corrections). Mathematically, the computation of γ_P can be expressed as:

$$\gamma_P = \gamma - \delta g_F + \delta g_B - \delta g_T \quad (3)$$

where

- γ is normal gravity on the reference ellipsoid
- δg_F is the Free air correction
- δg_B the Bouguer correction
- δg_T the terrain correction

The application of these corrections is described in the following section.

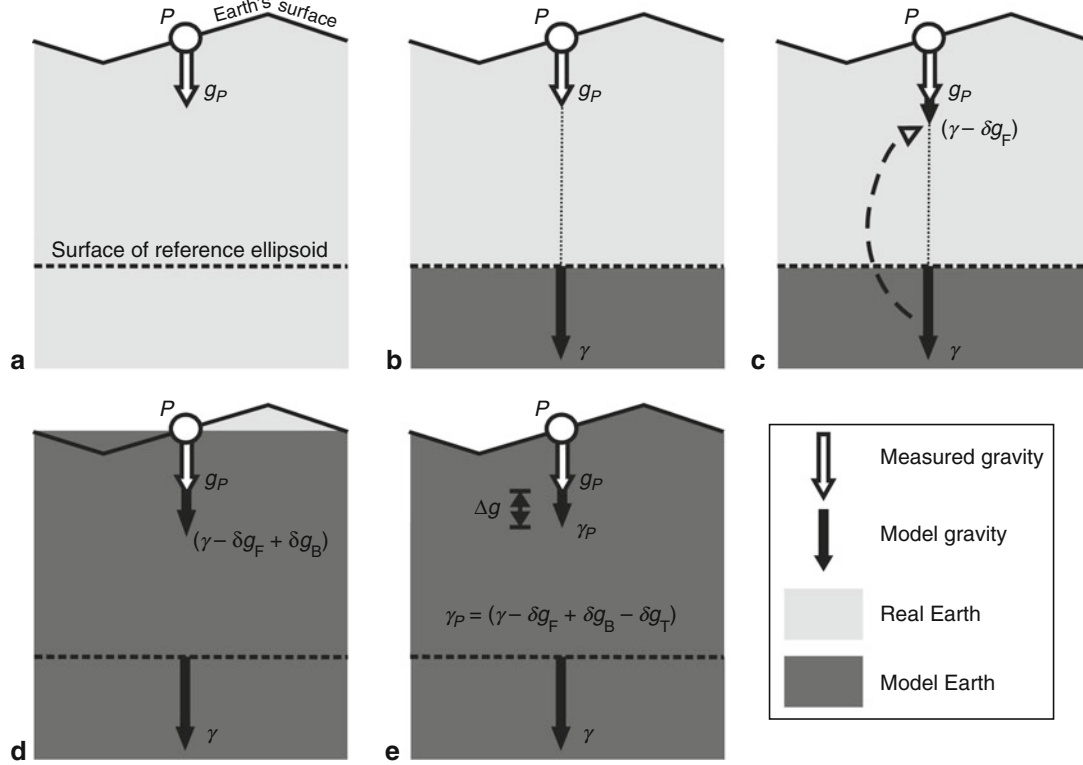
Computing Model Gravity at the Measurement Point

A Note on Units

The standard SI unit of acceleration due to gravity is $\text{m}\cdot\text{s}^{-2}$. However, because gravity anomalies are generally about three to seven orders of magnitude smaller than total gravity acceleration, the use of derivative units is more practical. Historically, the mGal was the standard derivative unit ($1 \text{ mGal} = 10^{-5} \text{ m}\cdot\text{s}^{-2}$). The mGal is still in common use, but gravity anomalies and related quantities are often also given in $\mu\text{m}\cdot\text{s}^{-2}$ ($10^{-6} \text{ m}\cdot\text{s}^{-2}$). In this description of the anomaly computation procedure, physical constants are given in standard SI units. This means that unless otherwise stated, the equations presented give quantities in $\text{m}\cdot\text{s}^{-2}$.

Normal Gravity (Fig. 2b)

Normal gravity, γ , represents the gravity effect of a reference ellipsoid whose geometry is defined to best represent the



Gravity, Data to Anomalies, Fig. 2 Illustration of the process of computing anomalous gravity in geophysics. (a) Observed gravity g_P reflects mass below the Earth's surface (light gray shading). (b) Normal gravity, γ , considers only mass below the reference ellipsoid (dark gray shading). (c) The free-air correction, δg_F , accounts for the elevation difference between the point P and the surface of the reference ellipsoid,

but not the rock mass in that space. (d) The Bouguer correction, δg_B , adds the gravity effect of rock mass between P and the surface of the reference ellipsoid, but assumes that rock mass has a smooth upper surface. (e) The terrain correction, δg_T , corrects for the smooth-topped Bouguer plate or cap

shape of the Earth. Because an ellipsoid is a mathematically defined body, computing its gravity effect is straightforward. A closed-form expression for computing normal gravity on the surface of a *geocentric* reference ellipsoid is given by the Somigliana–Pizetti formula (e.g., Hofmann-Wellenhof and Moritz 2006; see also ► “International Gravity Formula”):

$$\gamma = \gamma_a \frac{1 + k \sin^2 \phi}{\sqrt{1 - e^2 \sin^2 \phi}}, \quad (4)$$

where γ_a is normal gravity at the equator of the reference ellipsoid, ϕ is the *geocentric* latitude of the measurement point, and k and e^2 are quantities derived from the parameters describing the form of the reference ellipsoid (Table 1).

Equation 4 is exact and has now been adopted in many countries as a standard when computing gravity anomalies for geophysical applications (e.g., Hinze et al. 2005; Tracey et al. 2007). In the past, simpler expressions were used for computing normal gravity (cf. Li and Götze 2001). These approximate formulae were used largely for computational convenience, but given the measurement accuracy achievable today, their continued use is difficult to justify.

Free Air Correction (Fig. 2c)

Equation 4 gives normal gravity on the surface of the reference ellipsoid. However, the measurement point P is almost never located exactly on the surface of the reference ellipsoid. This means that the model value at a measurement point above the reference ellipsoid will be less than γ . The decrease in gravity above the reference ellipsoid is accounted for by the Free air correction. This correction essentially uses the rate of change in gravity above the surface of the Earth to correct for the decrease in gravity with elevation.

Historically, the value of the Free air correction was approximated as $0.3086 \times 10^{-5} \text{ m}\cdot\text{s}^{-2}/\text{m}$ (i.e., 0.3086 mGal/m), the gravity gradient above a spherical Earth. A more rigorous expression for the Free air correction, δg_F , can be derived from the gradient of normal gravity above the surface of the reference ellipsoid. Normal gravity at a height h above the ellipsoid can be computed directly from (Hofmann-Wellenhof and Moritz 2006):

$$\gamma_h = \gamma \left[1 - \frac{2}{a} (1 + f + m - 2f \sin^2 \phi) h + \frac{3}{a^2} h^2 \right], \quad (5)$$

Gravity, Data to Anomalies, Table 1 Parameters defining the GRS80 geocentric reference ellipsoid (Moritz 2000), the reference ellipsoid currently recommended by the International Union of Geodesy and

Geophysics as representing the size, shape, and gravity field of the Earth with an accuracy sufficient for most geodetic, geophysical, astronomical, and hydrographic applications

Parameter	Name	Equation	GRS80 value
a	Semi-major axis (equatorial radius)		6,378,137 m
GM	Geocentric gravitational constant		$3,986,005 \times 10^8 \text{ m}^3 \cdot \text{s}^{-2}$
J_2	Dynamic form factor		$108,263 \times 10^{-8}$
ω	Angular velocity		$7,292,115 \times 10^{-11} \text{ rad} \cdot \text{s}^{-1}$
b	Semi-minor axis (polar radius)		6,356,752.3141 m
γ_a	Equatorial normal gravity		$9.780\,326\,7715 \text{ m} \cdot \text{s}^{-2}$
γ_b	Polar normal gravity		$9.832\,186\,3685 \text{ m} \cdot \text{s}^{-2}$
k	Normal gravity constant	$(b\gamma_b/a\gamma_a) - 1$	0.001 931 851 353
e^2	First numerical eccentricity	$(a^2 - b^2)/a^2$	0.006 694 380 022 90
f	Ellipsoidal flattening	$(a - b)/a$	0.003 352 810 681 18
m	Geodetic parameter	$\omega^2 a^2 b/GM$	0.003 449 786 003 08
R_e	Arithmetic mean earth radius	$(2a + b)/3$	6,371,008.7714 m

Note that the parameters a , GM , J_2 , and ω are exact defining parameters – the other parameters are derived geometric and physical constants

or the magnitude of the Free air correction, δg_F , to be subtracted from γ can be computed using,

$$\delta g_F = \frac{2\gamma_a}{a} \left(1 + f + m + \left(-3f + \frac{5}{2}m \right) \sin^2 \phi \right) h - \frac{3\gamma_a}{a^2} h^2. \quad (6)$$

The parameters f (ellipsoidal flattening) and m (geodetic parameter) are defined in Table 1.

Bouguer Correction (Fig. 2d)

Once normal gravity is corrected for the vertical separation between the reference ellipsoid and the measurement point, a further adjustment needs to be made for the body of rock between that point and the reference ellipsoid. The gravity effect of this rock is not represented in normal gravity and is not accounted for by the Free air correction. This correction, the Bouguer correction, δg_B , was historically computed by representing the body of rock between the measurement point and surface of the reference ellipsoid as a horizontal slab extending to infinity. The infinite-slab Bouguer correction is given by:

$$\delta g_B = 2\pi G \rho t \quad (7)$$

where ρ is the assumed density of the slab, t is the thickness of the slab, and G is the universal gravitational constant ($6.67428 \times 10^{-11} \text{ m}^3 \cdot \text{kg}^{-1} \cdot \text{s}^{-2}$, the 2006 value recommended by the Committee on Data for Science and Technology, International Council of Science). In most cases, the slab thickness is equivalent to the height of the measurement point above the reference ellipsoid, h . The density of the slab generally takes on a standardized value of $2,670 \text{ kg} \cdot \text{m}^{-3}$, a value that closely represents the average density of crustal rocks (Hinze 2003). In some cases, a density that is more representative of local

geology may be used to minimize the correlation between the Bouguer anomaly and topography. However, using standardized values facilitates the comparison of different datasets and avoids incorporating geological interpretation into the anomaly computation procedure (LaFehr 1991a).

More sophisticated representations of the rock mass below the measurement point take into account the curvature of the Earth's surface. A commonly used representation is the spherical cap, the gravity effect of which is given by (LaFehr 1991b):

$$\delta g_B^{\text{cap}} = 2\pi G \rho [\mu t - \lambda(R_E + t)] \quad (8)$$

where R_E is the mean radius of the reference ellipsoid (Table 1), t the thickness of the cap (generally equivalent to the height of the measurement point above the reference ellipsoid, h), and μ and λ are dimensionless coefficients related to R_E and the angle subtended by the spherical cap at the center of the Earth. The spherical cap is generally taken to extend to a distance of 166.735 km from the measurement point, the distance originally selected by Bullard (1936) to minimize the difference between the gravity effect of the spherical cap and the infinite slab for a range of elevations (cf. LaFehr 1991b). Equation 8 is valid for surveys made on land, but Argast et al. (2009) present an extended form of the equation that is valid for gravity measurements made in the air and at sea.

Terrain Correction (Fig. 2e)

Regardless of whether the Bouguer correction is computed from an infinite slab or from a spherical cap, it is computed for a body with a flat or curved top that does not match topography. Topographic variations can have a significant influence on measured gravity, particularly in areas of high relief (e.g., mountain ranges or near the continental slope). For this reason, it is often beneficial to compute a terrain correction that effectively imparts an irregular surface onto the top of the Bouguer plate or cap.

Various methods exist for computing the terrain correction (cf. Nowell 1999; Hackney and Featherstone 2003), but the methods generally involve dividing the region around the measurement point into a series of rectangular prisms or concentric ring segments. A digital elevation model is then used to determine the average elevation for each prism relative to the height of the measurement point. Accuracy can be improved by using prisms with sloping tops. The size of the prisms is generally increased away from the measurement point as the influence of distant prisms is less than for nearer prisms. To further simplify computations, some methods approximate the more distant prisms as a line mass and fast-Fourier transform techniques have also been developed.

Often the different methods for computing terrain corrections are combined. For example, Hinze et al. (2005) describe the procedure adopted for the North American gravity database that uses segmented rings and estimated topography within 100 m of the measurements point, segmented rings and high-resolution digital elevation data from 100 to 895 m, and vertical prisms and digital elevation data from 895 m to 166.7 km (including corrections for Earth curvature beyond 14 km).

Historically, the most commonly used terrain correction procedure was based on computing the gravity effect of a number of concentric rings divided into individual segments (Hammer 1939). The gravity effect of each prism can be computed from:

$$\delta g_{\text{prism}} = 2\pi G \frac{\rho}{n} \left(r_2 - r_1 + \sqrt{r_1^2 + z^2} - \sqrt{r_2^2 + z^2} \right) \quad (9)$$

where n is the number of prism segments in each concentric ring (zone), r_2 is the outer radius of the concentric ring, r_1 the inner radius, and z is the elevation difference between the measurement point and the mean elevation of the ring segment. The other parameters are the same as in Eqs. 6 and 7. The density, ρ , used in Eq. 9 should be the same as that used for the Bouguer correction.

The Hammer (1939) formulation includes concentric rings out to a distance of 22 km, but ideally terrain corrections should be computed out to 166.735 km ($\sim 1.5^\circ$), the full radius of the standard spherical cap. When terrain corrections are computed to this distance, Earth curvature must also be considered. However, the influence of terrain corrections beyond about 10–20 km from the station is often insignificant and sometimes ignored.

Anomaly Computation

Once the Free air, Bouguer, and terrain corrections are applied to normal gravity, the gravity anomaly at the point P can be computed from Eqs. 2 and 3, i.e.,

$$\begin{aligned} \Delta g &= g_P - \gamma_P \\ &= g_P - (\gamma - \delta g_F + \delta g_B - \delta g_T) \\ &= g_P - (\gamma_h + \delta g_B - \delta g_T) \end{aligned} \quad (10)$$

In Eq. 10, the term in brackets represents the predicted or model value of gravity at the measurement point. The Free air correction is subtracted because normal gravity is reduced above the surface of the ellipsoid (Fig. 2c), while the Bouguer correction is added because the extra gravity effect related to the body of rock between P and the reference ellipsoid is not included in γ (Fig. 2d). The terrain correction must be subtracted from γ (Fig. 2e) to remove the effect of the non-existent slab/cap mass assumed to fill valleys and to reduce the model valley of gravity for topography higher than the measurement point (mass above the measurement point reduces gravity).

Once the model value of gravity is subtracted from the measured value of gravity at P , the remaining quantity is the anomalous component of the gravity field that cannot be explained by large-scale gravity effects related to the Earth's shape and rotation, changes in gravity with elevation, and the bulk influence of rock below the measurement point. The resulting anomalous gravity reflects subsurface density variations and can be interpreted geologically (see ► “Gravity Anomalies, Interpretation” and ► “Gravity Modeling, Theory and Computation”).

Other Factors

Atmospheric Correction

The parameters that define the GRS80 reference ellipsoid (Table 1) rely on satellite-based measurements, meaning that normal gravity includes the mass of the atmosphere. The application of a correction to normal gravity for the atmospheric mass is usually insignificant in comparison to errors associated with terrain corrections, meaning that its application is not essential for many geophysical applications (e.g., Featherstone and Dentith 1997). Nevertheless, the application of an atmospheric correction is included in new standards for gravity anomaly computation (e.g., Hinze et al. 2005; Tracey et al. 2007). If it is to be applied, the value to be subtracted from normal gravity can be computed using (Wenzel 1985).

$$\delta g_A = 0.874 - 9.9 \times 10^{-5}h + 3.56 \times 10^{-9}h^2. \quad (11)$$

This equation has an accuracy of about $0.01 \times 10^5 \text{ m}\cdot\text{s}^{-2}$ to 10 km elevation.

Further Isolating Anomalies of Interest

The Bouguer anomaly computed using Eq. 10 is dominated by the gravity effect of crustal-thickness variations that

more-or-less isostatically compensate for topography (thick crust under mountains, thin crust under ocean basins – see ► “Isostasy”). While the Bouguer anomaly is useful for constraining crustal-scale studies, the gravity effect of smaller-scale density variations within the upper crust (e.g., ore deposits or sedimentary basins) is often obscured by the long-wavelength gravity signature related to crustal-thickness variations.

The long-wavelength component of the Bouguer anomaly can be removed using regional–residual field separation techniques (see ► “Gravity Data, Regional: Residual Separation”). Alternatively, some form of isostatic compensation can be assumed (see ► “Isostasy”) and the gravity effect of the Moho that compensates for topography can be computed and subsequently subtracted.

Distant Relief

Computation of the Bouguer correction and associated terrain corrections are generally made out to a standard distance of 166.735 km. In certain circumstances, the gravity effect of topography beyond this distance can be significant (e.g., surveys in areas of high topography and continental- and global-scale studies: LaFehr 1991a; Mikuška et al. 2006). Users of gravity data should be aware of these distant relief effects and consider them if necessary.

Efforts to include the effects of whole-Earth topography are emerging. For example, Kuhn et al. (2009) present a grid of terrain corrections for the Australian continent that considers topography globally and Tenzer et al. (2009) computed gravity anomalies for the entire Earth that incorporate corrections for global topography.

The Geophysical “Indirect Effect”

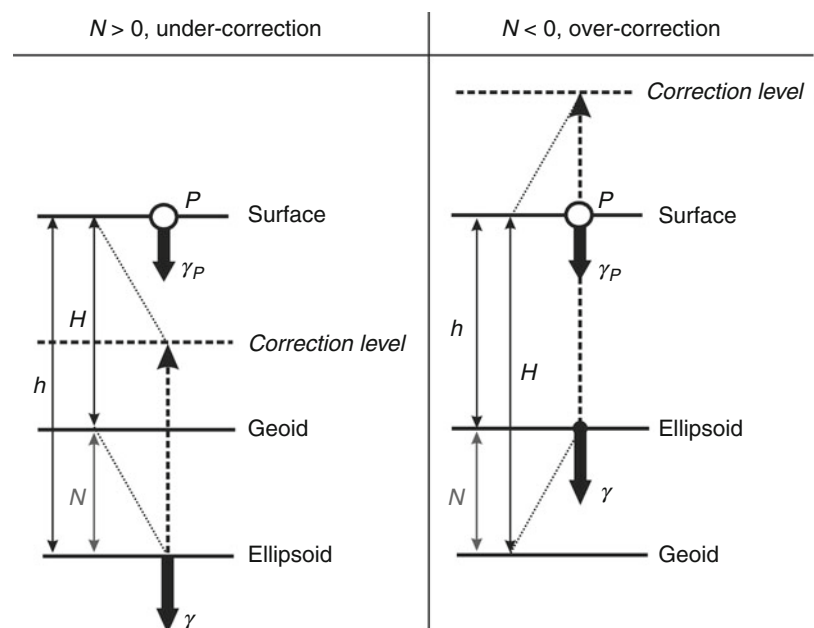
The formulation for anomalous gravity described here (Eq. 2) is the most logical way in which to think of gravity anomalies for geophysical applications. However, the process of anomaly computation is often described as a reduction of g_P to the sea level datum. While this process is useful in geodesy, it is not strictly correct in geophysical applications (Hackney and Featherstone 2003).

The misconception that gravity anomaly computation for geophysical purposes involves reducing gravity measurements to the geoid (or sea level) is partly related to the measurement of elevation. Before the advent of GPS (see ► “GPS, Data Acquisition, and Analysis”), elevations were normally determined by leveling over long traverses from reference tide gauges at the coastline (see ► “Geodesy, Ground Positioning, and Levelling”). Leveling thus provided heights relative to sea level, or the geoid. However, normal gravity is derived from a mathematically defined body, the reference ellipsoid, which does not coincide exactly with the geoid.

The reference ellipsoid is generally chosen to best fit the geoid either regionally or globally, but the ellipsoid never perfectly matches the geoid. This means that if heights relative to sea level are used to compute anomalous gravity, then the corrections applied will either be an under- or over-correction, depending on whether the ellipsoid is above or below the geoid (Fig. 3). The magnitude of this under- or overcorrection is referred to as the *geophysical indirect effect* (e.g., Hackney and Featherstone 2003). Globally, the geoid–ellipsoid separation is about ± 100 m at most, meaning that the geophysical indirect effect varies by about $\pm 30 \times 10^{-5} \text{ m}\cdot\text{s}^{-2}$. This variation can induce long-

Gravity, Data to Anomalies,

Fig. 3 Illustration of the under- or over-correction of model predicted gravity that results when the Free air and Bouguer corrections are computed using heights relative to mean sea level (geoid) rather than relative to the reference ellipsoid. N is defined by Eq. 1



wavelength errors in gravity data, especially in areas of high geoid gradient. For more regional surveys, the effect is generally minor and, in many cases, essentially constant. For regional surveys, the geophysical indirect effect is a bias that will often be accounted for during regional–residual field separation.

With the advent of GPS, elevations tied directly to the WGS84 ellipsoid became readily available. Because the differences in the dimensions of the WGS84 and GRS80 ellipsoids are minor, anomalous gravity can be computed directly from the heights measured by GPS. However, care is required in order to ensure that consistent heights are used for both older and newer datasets. Updated procedures adopted in some countries recognize the need to use heights tied to the reference ellipsoid (e.g., Hinze et al. 2005; Tracey et al. 2007).

Gravity Anomalies for Geodetic Purposes

This entry focuses on the geophysical approach to computation of gravity anomalies, but it is worth outlining the ways in which gravity anomalies are computed for geodetic purposes. In geodesy (see ► “Geodesy, Physical” and ► “Geoid Determination, Computational Methods”), gravity anomalies are used to determine the shape of the Earth using either a “classical” or “modern” approach (e.g., Hofmann-Wellenhof and Moritz 2006). In the classical approach, the geoid is determined using gravity anomalies computed from the difference between surface gravity measurements reduced (downward) to the geoid and normal gravity on the reference ellipsoid. The modern approach leads to the computation of a surface referred to as the telluroid and requires gravity anomalies computed from the difference between surface gravity and normal gravity corrected (upward) to the telluroid. The separation between the telluroid and the surface (the “height anomaly”) can then be used to determine a geoid-proxy referred to as the quasigeoid.

Computation Example: The Central Andes

The Central Andes in South America are an ideal region to illustrate the process of computing anomalous gravity for geophysical purposes. This is because the high topography and associated thick crust lead to large corrections and high-amplitude gravity anomalies. Observed gravity and the different types of corrections, model values, and anomalies for the Central Andes are illustrated in Fig. 4. The gridded data shown in Fig. 4 are based on a dataset that contains more than 4,000 surface gravity measurements (Schmidt and Götze 2006) (Fig. 4a). Figure 4b shows the observed (measured) values of gravity, g_P , that are tied to reference stations linked to the IGSN71 gravity datum. Observed

gravity values are large (9.773 – $9.796 \text{ m}\cdot\text{s}^{-2}$) and the highest parts of the Central Andes are associated with the lowest measured gravity. This is due mostly to the pronounced decrease in gravity with elevation, but the mass deficit associated with the thick, low density crust (with respect to the underlying mantle) also contributes to the decreased gravity in the mountainous areas.

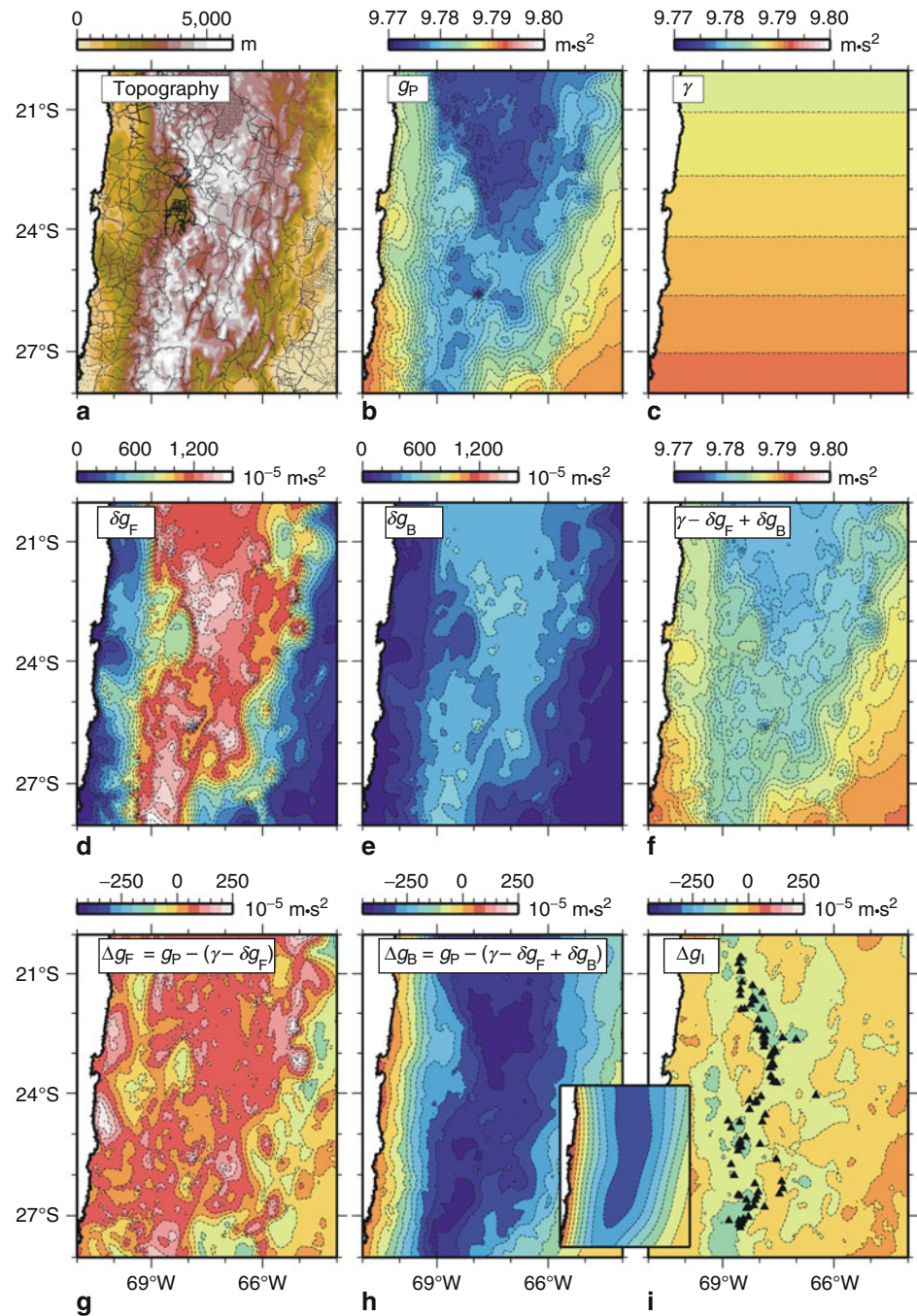
Normal gravity on the GRS80 reference ellipsoid (Eq. 4 and Fig. 4c) shows a north-to-south increasing trend that reflects the pole-ward increase in gravity related to the shape and rotation of the reference ellipsoid. To derive a model value directly at each of the data points shown in Fig. 4a, Free air (Eq. 6) and Bouguer (Eq. 7 or 8) corrections are applied to normal gravity. The magnitude of the Free air correction, δg_F , is shown in Fig. 4d – the correction is largest where topography is highest. The Free air correction is subtracted from normal gravity to account for the decrease in normal gravity with elevation above the reference ellipsoid. The Free air anomaly results when the height-corrected normal gravity is subtracted from observed gravity (Fig. 4g). Compared to observed gravity, the amplitude range of the Free air anomaly is significantly reduced ($\pm 300 \times 10^{-5} \text{ m}\cdot\text{s}^{-2}$).

A map of Free air anomalies (Fig. 4g) is difficult to interpret geologically because it retains a strong correlation with topography. This correlation is modified when the gravity effect of the rock between the surface and the reference ellipsoid is accounted for using the Bouguer correction (Fig. 4e). The Bouguer correction is added to normal gravity (because normal gravity does not include the gravity effect of the rocks) and, together with the addition of the Free air correction, results in a model value, γ_P , that resembles the observed gravity (Fig. 4f). When this model value is subtracted from the observed gravity, the Bouguer anomaly results (Fig. 4h). The Bouguer anomaly of the Central Andes is large and negative (minimum values reach about $-450 \times 10^{-5} \text{ m}\cdot\text{s}^{-2}$), which reflects the thick crustal root beneath the Central Andes.

An isostatic residual gravity map is shown for the Central Andes in Fig. 4i. The regional field (inset between Fig. 4h, i) is computed by assuming regional isostatic compensation of topography using an effective elastic thickness of 20 km (see ► “Isostasy”). Crustal thickness is computed for this model and then the gravity effect of the isostatic Moho is estimated and subtracted from the Bouguer anomaly. The Central Andean isostatic residual field has a vastly reduced amplitude (-150 to $+60 \times 10^{-5} \text{ m}\cdot\text{s}^{-2}$) and bears a closer resemblance to surface geology, particularly the location of active volcanoes. The coincidence between active volcanoes and negative isostatic residual anomalies suggests the presence of low-density material within the crust or that the assumption of regional isostatic compensation is not applicable to the volcanic chain.

Gravity, Data to Anomalies,

Fig. 4 Illustration of the various components of anomalous gravity for the Central Andes in South America. (a) Topography from the GTOPO30 global digital elevation model – points show the location of gravity measurements (Schmidt and Götze 2006) where the various corrections are computed and gridded using a cell size of one arc minute; (b) observed gravity; (c) normal gravity on the surface of the GRS80 reference ellipsoid computed using Eq. 4 – note the latitudinal variation; (d) the magnitude of the Free air correction that is subtracted from normal gravity to account for the reduction in normal gravity above the reference ellipsoid; (e) the Bouguer correction that represents the gravity effect of the mass of rock between the surface and the reference ellipsoid; (f) model value for gravity at the surface that is subtracted from the observed gravity shown in b to produce the Bouguer anomaly; (g) the Free air anomaly; (h) the Bouguer anomaly (map in b minus map in f); (i) isostatic residual anomaly, δg_I , computed by subtracting from the Bouguer anomaly an isostatic regional field (inset between h and i) computed by assuming regional isostatic compensation of Andean topography with an elastic thickness of 20 km. The triangles in (i) show active volcanoes (Trumbull et al. 2006). (Note that the Bouguer anomaly in (h) does not incorporate a terrain correction, but when plotted at this scale, the effect of the terrain correction is not evident)

**Summary**

From a geophysical perspective, a gravity anomaly is best thought of as the difference between measured gravity and a model value, at the same point, that accounts for large-scale, non-geological effects that mask anomalies related to subsurface geology. The model gravity value is derived by correcting normal gravity for the effects of height above the reference ellipsoid (Free air correction) and the

mass of rock between the reference ellipsoid and the measurement point (Bouguer correction). Additional corrections (terrain and atmospheric corrections, indirect effect) improve on the relatively straightforward Free air and Bouguer corrections. Since the earliest use of gravity data, our ability to compute the corrections to normal gravity has improved to the point where simplistic application of the corrections is giving way to procedures that consider global influences.

Depending on the application, long-wavelength field components remaining in the Bouguer anomaly may also be removed (isostatic correction or regional–residual field separation). Gravity anomalies can then be used to constrain geological interpretations of relatively shallow crustal features (e.g., basins, ore bodies, or even archeological sites), or be used to examine crustal structure and isostatic state.

Bibliography

- Argast D, Bacchin M, Tracey R (2009) An extension of the closed-form solution for the gravity curvature (Bullard B) correction in the marine and airborne cases. ASEG Extended Abstracts 2009(1):6. <https://doi.org/10.1071/ASEG2009ab129>
- Bullard EC (1936) Gravity measurements in East Africa. *Philos Trans R Soc Lond* 235:486–497
- Featherstone WE, Dentith MC (1997) A geodetic approach to gravity data reduction. *Comput Geosci* 23(10):1063–1070
- Hackney RI, Featherstone WE (2003) Geodetic versus geophysical perspectives of the “gravity anomaly”. *Geophys J Int* 154:35–43. See also: Erratum, 2003. *Geophysical Journal International*, 154: 596; Corrigendum, 2006. *Geophysical Journal International*, 167: 585
- Hammer S (1939) Terrain corrections for gravimeter stations. *Geophysics* 4:184–194
- Hinze WJ (2003) Bouguer reduction density, why 2.67? *Geophysics* 68(5):1559–1560
- Hinze WJ, Aiken C, Brozena J, Coakley B, Dater D, Flanagan G, Forsberg R, Hildenbrand T, Keller GR, Kellogg J, Kucks R, Li X, Mainville A, Morin R, Pilkington M, Plouff D, Ravat D, Roman D, Urrutia-Fucugauchi J, Véronneau M, Webring M, Winchester D (2005) New standards for reducing gravity data: the North American gravity database. *Geophysics* 70(4):J25–J32. <https://doi.org/10.1190/1.1988183>
- Hofmann-Wellenhof B, Moritz H (2006) *Physical geodesy*. Springer, Vienna
- Kuhn M, Featherstone WE, Kirby JF (2009) Complete spherical Bouguer anomalies over Australia. *Aust J Earth Sci* 56(2):213–223. <https://doi.org/10.1080/08120090802547041>
- LaFehr TR (1991a) Standardization in gravity reduction. *Geophysics* 56(8):1170–1178
- LaFehr TR (1991b) An exact solution for the gravity curvature (Bullard B) correction. *Geophysics* 56(8):1179–1184
- Li X, Götze H-J (2001) Ellipsoid, geoid, gravity, geodesy, and geophysics. *Geophysics* 66(6):1660–1668
- Mikuška J, Pašteka R, Marušiak I (2006) Estimation of distant relief effect in gravimetry. *Geophysics* 71(6):J59–J69. <https://doi.org/10.1190/1.2338333>
- Morelli C, Ganter C, Hankasalo T, McConnell RK, Tanner JB, Szabo B, Uotila U, Whalen CT (1974) The international gravity standardization net 1971. International Association of Geodesy. Special publication number, 4. International Association of Geodesy, Paris
- Moritz H (2000) Geodetic reference system 1980. *J Geod* 74(1):128–162. <https://doi.org/10.1007/s001900050278>
- Nabighian MN, Ander ME, Grauch VJS, Hansen RO, LaFehr TR, Li Y, Pearson WC, Peirce JW, Phillips JD, Rude ME (2005) Historical development of the gravity method in exploration. *Geophysics* 70(6):63ND–89ND. <https://doi.org/10.1190/1.2133785>
- Nowell DAG (1999) Gravity terrain corrections – an overview. *J Appl Geophys* 42:117–134
- Schmidt S, Götze H-J (2006) Bouguer and isostatic maps of the Central Andes. In: Oncken O, Chong G, Franz G, Giese P, Götze H-J, Ramos VA, Strecker MR, Wigger P (eds) *The Andes: active subduction orogeny*. Springer, Berlin/Heidelberg. *Frontiers in Earth Sci*, vol 1, pp 559–562
- Tenzer R, Hamayun K, Vajda P (2009) Global maps of the CRUST 2.0 crustal components stripped gravity disturbances. *J Geophys Res* 114:B05408. <https://doi.org/10.1029/2008JB006016>
- Tracey R, Bacchin M, Wynne P (2007) AAGD07: a new absolute gravity datum for Australian gravity and new standards for the Australian National Gravity Database. ASEG Extended Abstracts 2007(1):1–3. <https://doi.org/10.1071/ASEG2007ab149>
- Trumbull RB, Riller U, Oncken O, Scheuber E, Munier K, Hongn F (2006) The time – space distribution of Cenozoic volcanism in the Central Andes: a new data compilation and some tectonic implications. In: Oncken O, Chong G, Franz G, Giese P, Götze H-J, Ramos VA, Strecker MR, Wigger P (eds) *The Andes: active subduction orogeny*. Springer, Berlin/Heidelberg. *Frontiers in Earth Sci*, vol 1, pp 29–43
- Wenzel H (1985) *Hochauflösende Kugelfunktionsmodelle für des Gravitationspotential der Erde [1]*. Wissenschaftliche arbeiten der Fachrichtung Vermessungswesen der Universität, Hannover, p 137

Gravity, Global Models

Nikolaos K. Pavlis

Geodesy and Geophysics Basic and Applied Research,
National Geospatial-Intelligence Agency (NGA), Reston,
VA, USA

Synonyms

Geopotential models; Global gravitational models; Potential coefficient models

Definition

A *Global Gravitational Model* (GGM) is a mathematical approximation to the external gravitational potential of an attracting body.

Gravitation is the term used to describe the potential generated by the masses of an attracting body like the Earth. *Gravity* is the term used for the potential of gravitation plus the centrifugal potential due to the rotation of the attracting body.

Global Gravitational Models

Introduction

A GGM is a mathematical approximation to the external gravitational potential of an attracting body. We will focus here on the case where the attracting body is the Earth, although many of the concepts that we discuss apply equally

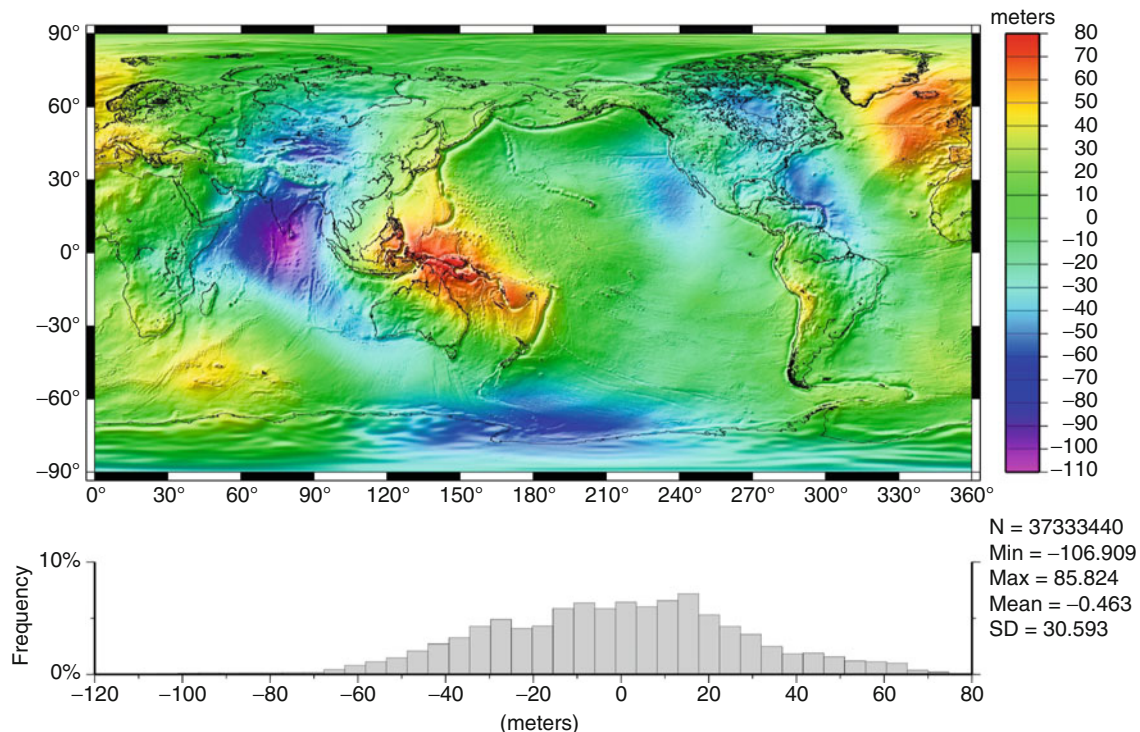
well to other planets and celestial bodies. In its most complete and general form, a GGM consists of a set of numerical values for certain parameters, the statistics of the errors associated with these values (as described, e.g., by their error variance-covariance matrix), and a collection of mathematical expressions and algorithms that allow a user to perform:

1. *Synthesis*, that is, computation of the numerical values of quantities related to the gravitational potential (*functionals* of the gravitational field), given the position of the evaluation point.
2. *Error Propagation*, that is, computation of the expected errors of the computed functionals, as implied by the propagation of the errors of the parameters defining the GGM.

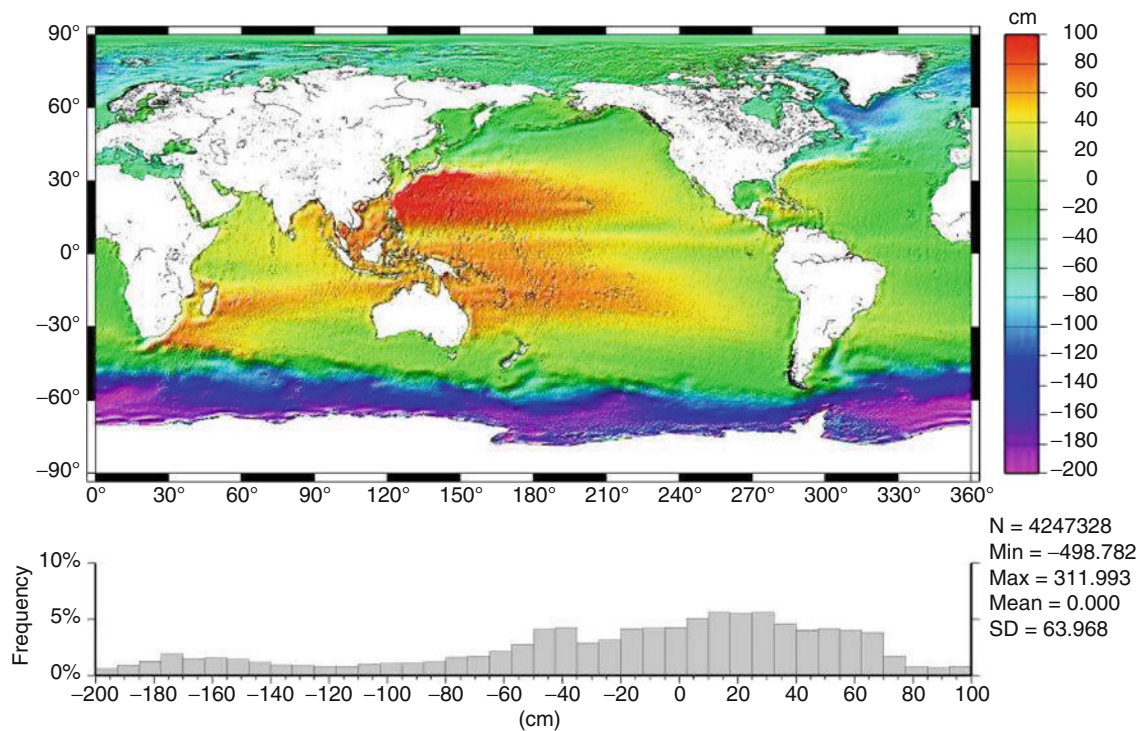
A GGM must be able to support such computations at arbitrary points, located on or above the Earth's surface, in a fashion that is both rigorous and efficient. In addition, a GGM should fulfill certain conditions stemming from the underlying physics. Namely, it should represent a scalar function of position that is harmonic outside the attracting masses and vanishes at infinity as the reciprocal of the distance between attracted point and attracting mass element. Moreover, the GGM should permit the computation of any functional of the field in a way that guarantees *self-consistency*. This

means that the model should preserve the relationships (differential or integral) between the various functionals. A GGM has numerous uses, both operational and scientific (see also [Tscherning 1983]), including:

1. Orbit determination applications necessary for space surveillance (the detection, tracking, and orbit prediction of Earth-orbiting objects).
2. Inertial navigation applications for trajectory determination of airplanes and missiles.
3. *Geoid* determination. The geoid is one of the infinitely many *level* (or *equipotential*) surfaces of the gravity field of the Earth, that is, the gravity potential is constant at every point on the geoid. The geoid is the specific equipotential surface that would coincide exactly with the mean ocean surface of the Earth, if the oceans were in equilibrium, at rest, and extended through the continents (such as with very narrow canals). Figure 1 shows the geoid surface that is implied by the EGM2008 GGM (Pavlis et al. 2008). Geoid determination allows one to transform a *geometric* (ellipsoidal) height to an *orthometric* one, that is, to an elevation referenced to an equipotential surface, as those elevations that appear on topographic maps. This application has attracted great interest in recent years, because GPS positioning and geoid determination offer the possibility of determining



Gravity, Global Models, Fig. 1 Geoid undulations (in meters), implied by the EGM2008 Global Gravitational Model, to degree 2,160, with respect to the WGS 84 reference ellipsoid



Gravity, Global Models, Fig. 2 Dynamic Ocean Topography (DOT) (in centimeters), obtained by subtracting the EGM2008 geoid undulations from a Mean Sea Surface grid computed at the Danish National

Space Center (DNSC). Notice the low associated with the Antarctic Circumpolar Current (ACC) and the highs associated with the Pacific Ocean gyres

orthometric heights and height differences without the need for the expensive and laborious spirit leveling (Schwarz et al. 1987).

4. Oceanographic applications that require the estimation of the Dynamic Ocean Topography (DOT), an estimate of which is shown in Fig. 2, and of its slopes. These quantities are directly related to ocean circulation. This application puts very stringent accuracy and resolution requirements on GGMs (Ganachaud et al. 1997).
5. An accurate and high-resolution GGM may be used to provide a reference surface – a global, high-resolution geoid surface – suitable for the realization of a global vertical datum (Rapp and Balasubramania 1992).
6. Geophysical prospecting applications where, in combination with other information (e.g., seismic data), a GGM may provide important constraints that aid the determination of the subsurface density distribution.
7. Comparisons of the decay of the spectrum of the gravitational potential to the decay of the spectrum of the Earth's topography (Rapp 1989).

These and other applications represent integral parts of various civilian and military activities. Each of these applications has (in general) different accuracy and resolution requirements, as far as the supporting GGM is concerned. For example, due to the attenuation of the gravitational field

with increasing altitude, a relatively low-resolution GGM (e.g., a *Spherical Harmonic Analysis Applied to Potential Fields* to maximum degree and order 70 or 90) is currently adequate for the precise orbit determination of most Earth-orbiting satellites. In contrast, accurate determination of the slopes of the equipotential surface (deflections of the vertical) demands a GGM of much higher resolution.

Geodesists have at various times developed “special purpose” GGMs that optimize performance for a particular application (e.g., orbit determination of a particular satellite, or geoid undulation computation over a specific geographic region). Although such “tailored” models have found some uses in the past, the ultimate goal has always been the development of a *single*, general purpose, GGM that addresses the different and diverse applications in an optimal manner, without over-performing in one application at the expense of its performance in others.

The development of a high-resolution GGM is a task that involves the optimal combination of a variety of data (satellite, land, marine, airborne). This is because a single data type with both global coverage and with uniformly high accuracy and high *spectral* sensitivity does not (yet) exist. The aforementioned data are of complementary character (in terms of spectral sensitivity and/or of geographic coverage), so that their optimal combination enables a GGM to satisfy the variety of applications described before.

Comprehensive GGM solutions that combine different types of data may involve the simultaneous estimation of the parameters that describe the gravitational potential, along with parameters that describe the dynamic ocean topography, solid Earth and ocean tides, Earth orientation and tracking station position parameters, as well as a plethora of “nuisance” parameters necessary to model completely the content of certain data types (e.g., biases and delays associated with certain satellite tracking data). The result of a successful GGM development effort is a model that can be used as a *standard* for numerous applications, over a substantial period of time.

Local and Regional Gravimetric Models

The accuracy and the resolving power of the data that were used in its development dictate the accuracy and resolution of a GGM. Geopolitical and/or proprietary data issues many times prevent the individual or the team developing a GGM from having access to *all* the existing data. However, over some regions, data of higher accuracy and/or resolving power (e.g., geographically dense sets of gravity and elevation data) may be available to some individual(s) or may become available *after* a GGM has been developed. These data may be used in combination with the existing GGM to improve the accuracy and/or resolution of the determination of *one or more specific functionals of the field*, over the region where the detailed data became available. This local or regional “densification” can produce a specific local or regional gravimetric product or model.

Such densification has been among the favorite geodetic activities over many decades now, and represents the geodesist’s way of creating a *multi-resolution* gravitational model resembling a “quilt.” In effect, patches of fine detail (the *Local Gravimetric Models* – LGMs) are sewn on top of a more or less homogeneous piece of fabric (the reference GGM). The GGM is the foundation upon which any regional or local gravimetric approximation study and application is built. Geodesists do not necessarily have to reevaluate the reference GGM every time a new set of data becomes available locally. Such reevaluation is mostly warranted if new and improved satellite data become available, spanning a sufficiently long time period, and/or if new terrestrial data (of higher accuracy and/or resolution) become available over areas with substantial geographic extent.

Signal Representation and Data Characteristics

Although geodesists have variously considered and studied the representation of the gravitational potential using *point masses* (Sünkel 1981, 1983), *finite element* methods (Meissl 1981; Baker 1988), and *splines* (Sünkel 1984; Jekeli 2005), these approaches have seen only limited application in the representation of (especially) the “static” (i.e., the time-averaged) gravitational field of the Earth. *Spherical harmonic*

functions have prevailed as the standard form used for the representation of the gravitational potential globally, from the very early days of global determinations to the present. Indeed, the set of coefficients of a *Spherical Harmonic Analysis Applied to Potential Fields* of the gravitational potential has become pretty much synonymous with a GGM.

Nerem, Jekeli, and Kaula (1995) provide a review of the gravity field determination process, including a review of global-scale modeling. Bouman (1997) provides a comprehensive “survey” of global gravity field models starting from the 1970s. Rapp (1998) provides a review of the geopotential modeling developments of the twentieth century, which includes an extensive list of references.

The Earth’s external gravitational potential, V , at a point P defined by its geocentric distance (r_P), geocentric colatitude (θ_P) (defined as 90° minus the geocentric latitude), and longitude (λ_P), can be expressed as:

$$V(r_P, \theta_P, \lambda_P) = \frac{GM}{r_P} \left[1 + \sum_{n=2}^{\infty} \left(\frac{a}{r_P} \right)^n \sum_{m=-n}^n C_{nm} Y_{nm}(\theta_P, \lambda_P) \right]. \quad (1)$$

GM is the geocentric gravitational constant (the product of the universal gravitational constant, G , times the mass of the Earth including its atmosphere, M) and a is a scaling factor associated with the fully normalized, unitless, spherical harmonic coefficients C_{nm} (a is usually numerically equal to the equatorial radius of an adopted mean-Earth ellipsoid). The absence of first-degree terms in Eq. 1 is a consequence of placing the origin of the coordinate system at the Earth’s center of mass. The *Spherical Harmonic Analysis Applied to Potential Fields* are defined as (Heiskanen and Moritz 1967, Sect. 1–14):

$$Y_{nm}(\theta_P, \lambda_P) = \bar{P}_{n|m|}(\cos \theta_P) \cdot \begin{cases} \cos m\lambda_P & \text{if } m \geq 0 \\ \sin |m| \lambda_P & \text{if } m \leq 0 \end{cases}. \quad (2)$$

$\bar{P}_{n|m|}(\cos \theta_P)$ is the fully-normalized associated legendre function of the first kind, of degree n and order $|m|$. The normalization of the surface spherical harmonic functions Y_{nm} is such that:

$$\frac{1}{4\pi} \int_{\sigma} Y_{nm}^2 d\sigma = 1, \quad (3)$$

where the integral is taken over the unit sphere σ , whose area element is $d\sigma$. In practice, the degree summation in Eq. 1 is truncated to some finite degree N , which defines (approximately) the resolution of the GGM. A useful rule of

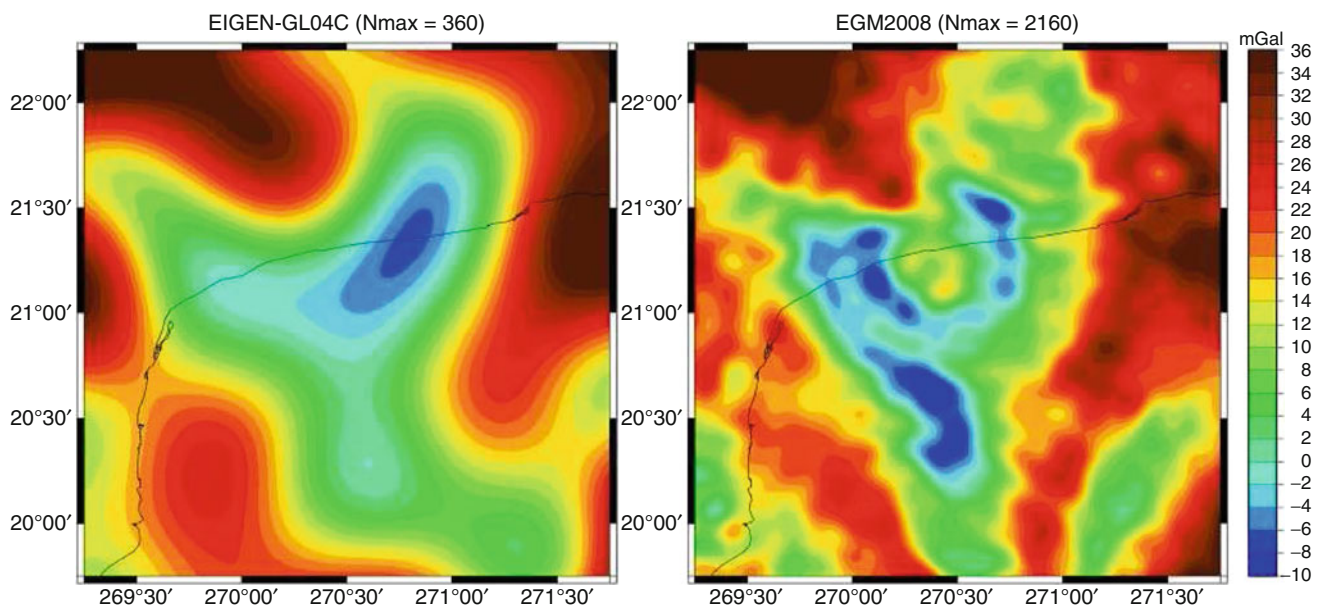
thumb is that the half-wavelength linear resolution ($\lambda/2$), on the Equator, expressed in kilometers, is related to N by:

$$\lambda/2 = \frac{180^\circ}{N} \cdot 111 \text{ (km)}. \quad (4)$$

Figure 3 shows a representative example of the fine detail (high resolution) that can be captured with an expansion to degree 2,160, compared to the lower-resolution information that is retrievable using an expansion to degree 360. A spherical harmonic expansion complete to degree and order N involves $(N + 1)^2$ spherical harmonic coefficients, a number that grows rapidly as the maximum degree, and therefore the resolution of the model, increases. For example, an expansion to degree 360, as the EGM96 model involves $O(130 \text{ k})$ coefficients, while EGM2008, which extends to degree 2,160, involves approximately 4.7 million coefficients. The availability of efficient numerical algorithms for the estimation and for the subsequent use of these spherical harmonic expansions is obviously of critical importance. The goal of global high-resolution gravitational modeling is to estimate, as accurately as possible, the coefficients C_{nm} , to as high degree as possible, through the optimal combination of gravitational information from a variety of data sources. Of equal importance is the estimation of reliable error estimates for the C_{nm} values. The estimated C_{nm} values can then be used to compute functional(s) of the field (e.g., Geoid *undulations*, *gravity anomalies*, deflections of the vertical) while their associated errors (and error correlation when available) can

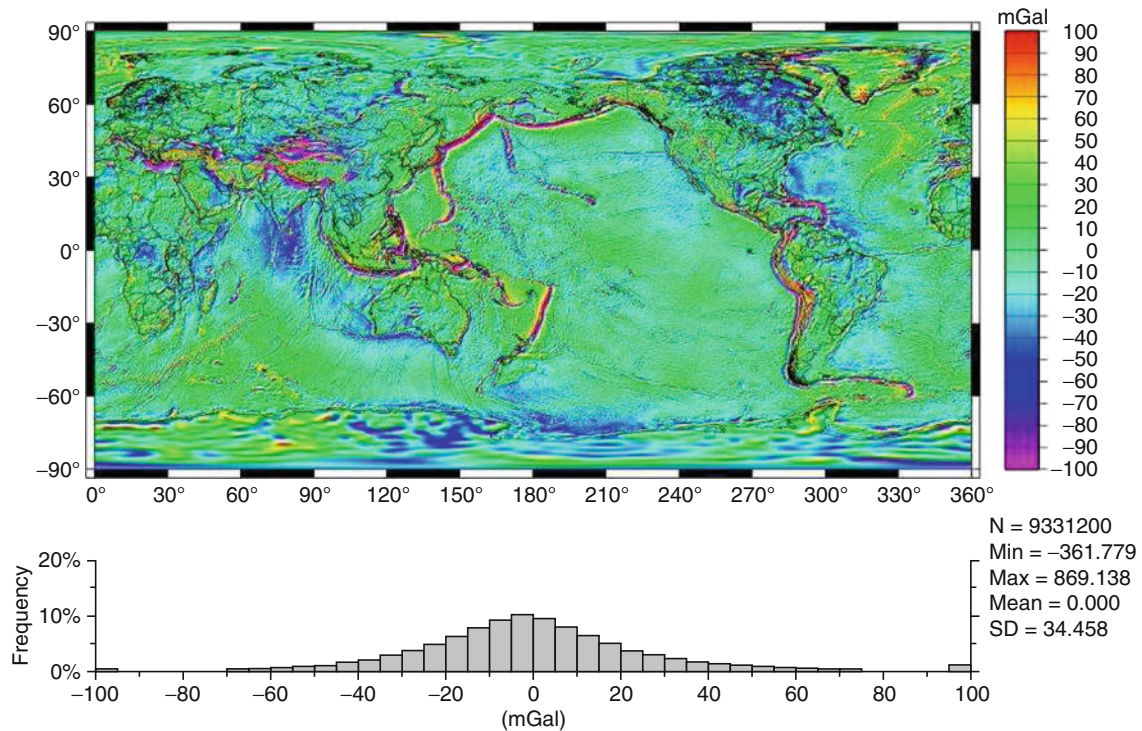
be propagated to yield the errors of the derived functional(s). Before the dawn of the new millennium and the availability of data from the satellite missions CHAMP, GRACE, and GOCE, four kinds of gravitational information were commonly available for the development of high-degree combination gravitational models like EGM96 (Lemoine et al. 1998):

1. Gravitational information obtained from the analysis of satellite orbit perturbations that are deduced from tracking data. This is of critical importance for the accurate determination of the low degree part of the model. *Satellite-only* models have progressed from solutions to degree 4 in the early 1960s, to models complete to degree 70 or 90 available by the end of the twentieth century. These advances were made through the availability of ever more accurate tracking data acquired over a continuously expanding constellation of Earth orbiters. Tracking data from approximately 40 satellites have been used in the development of the satellite-only solution supporting EGM96 (denoted EGM96S) (Lemoine et al. 1998). These data include optical, radio Doppler and radio interferometric observations, Satellite Laser Ranging (SLR), Doppler Orbit determination and Radiopositioning Integrated on Satellite (DORIS), and Satellite-to-Satellite Tracking (SST) data from the Global Positioning System (GPS) and Tracking and Data Relay Satellite System (TDRSS) constellations to lower Earth orbiters. Despite these advances, these tracking data types are incapable of resolving the fine



Gravity, Global Models, Fig. 3 Free-air gravity anomalies (in mGal, where $1 \text{ mGal} = 10^{-5} \text{ m/s}^2$) implied by two GGMs of different maximum degrees (different resolving powers), over the Yucatán Peninsula. EIGEN-GL04C extends to degree 360, while EGM2008 extends to

degree 2,160. The Chicxulub impact crater “ring” is clearly visible in the EGM2008 gravity anomalies, but cannot be discerned in the EIGEN-GL04C gravity anomalies due to the lower resolution of this model



Gravity, Global Models, Fig. 4 Free-air gravity anomalies (in mGal) implied by EGM2008 to degree 2,160, averaged over $5' \times 5'$ cells on the surface of the Earth. These values reveal numerous geophysical features, such as oceanic trenches, ridges, subduction and fracture zones, and

seamount chains. The lack of detailed data over Antarctica is also clearly evident from the smoothness of the values over this region (for the development of EGM2008, only GRACE satellite information was available over Antarctica)

structure of the field due to the attenuation of the gravitational signal with altitude. Moreover, the available satellites do not sample uniformly the range of orbital inclinations and altitudes, which is a necessary condition for the de-correlation of the harmonic coefficients estimated from satellite tracking data only. This causes strong correlation especially among coefficients of higher degrees within satellite-only solutions, and necessitates the use of a priori constraints in the development of satellite-only models (Lerch et al. 1979).

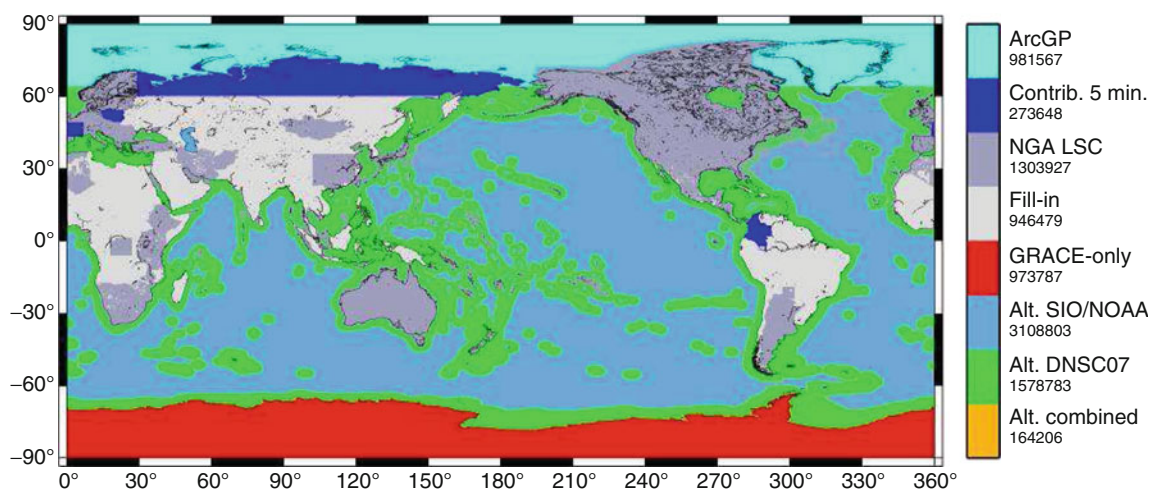
2. Surface (land, marine, and airborne) gravimetric data, which are (essentially) measurements of the magnitude of the acceleration of gravity. These data are in principle capable of resolving both long and short wavelength features of the gravity field. This, however, requires uniform global coverage with dense gravity data of uniformly high accuracy. The best available data sets ca. 1996 (Kenyon and Pavlis 1996) represent information derived from over 4,000 sources of detailed gravity data collected over several decades. The accuracy and density of point data vary substantially with geographic region, with extended regions (e.g., Antarctica) being practically void of gravity measurements. *Gravity Anomalies, Interpretation* values are the remainders after subtraction from the gravity measurements of the gravity acceleration

implied by an ellipsoidal reference (normal) field (see also Fig. 4). These gravity anomalies are susceptible to various systematic errors (Heck 1990). These errors, in conjunction with the nonuniformity of coverage, degrade the long wavelength integrity of the gravitational information that can be extracted from surface gravimetry. Nevertheless, surface and airborne gravimetry presently provide the only data that can resolve short wavelength gravity features, especially over land areas. In addition, shipborne gravity measurements may aid the separation of the geoid from the DOT signal when used in combination with satellite altimetry. The processing of surface gravity data also requires accurate and high-resolution information of the Earth's topography. This information is necessary both for the preprocessing of the gravity data and for the computation of correction terms that are necessary in order to determine the geoid surface over the continents, from the estimated GGM coefficients (Rapp 1997). It may also be used to supplement low-resolution gravitational information that may be available over certain geographic regions, with high-resolution information implied by the gravitational effects of the Earth's topography. This approach was used effectively for the development of EGM2008, as it is discussed by Pavlis et al. (2007).

3. Satellite altimeter data have enabled an unsurpassed mapping of the field over the oceans, in terms of both accuracy and resolution. TOPEX/Poseidon (T/P) (Fu et al. 1994) was the first satellite altimetry mission that routinely provided measurements of the Sea Surface Height (SSH), which were not significantly contaminated by radial orbit error (RMS radial orbit error at the ± 2 cm level). However, altimetric measurements are confined over the ocean areas bounded by the satellite's inclination, and furthermore provide a mapping of the sum of the geoid undulation (the height of the geoid surface with respect to the surface of an ellipsoid of revolution) plus the DOT. These aspects weaken somewhat the contribution of altimeter data in the determination of the long wavelength gravitational field and necessitate the appropriate modeling and estimation of the DOT when altimeter SSH data are used in combination solutions. There is, however, another way of incorporating altimeter data into a high-degree GGM, which is discussed next.
4. The combination of altimeter data from multiple missions (spanning several years), some of which have produced very closely spaced ground tracks, has provided a dense sampling of most of the ocean's surface. These data, in the form of SSH and/or SSH slopes, can be used to estimate ocean-wide sets of gravity anomalies, at very fine resolution (e.g., $2' \times 2'$ and even $1' \times 1'$). Spatial averages of these values can be "merged" with corresponding land and airborne gravity data and gravity information obtained from models of the potential generated by the topography and its *isostatic* compensation (Pavlis and Rapp 1990) to produce a complete global equiangular grid of gravity anomaly values. Figure 5 shows the identification of the data sources, which were used to compile the "merged"

$5' \times 5'$ free-air gravity anomaly data file that supported the development of EGM2008. The geometry of such a grid, in conjunction with the symmetry properties of the Legendre and the trigonometric functions, allows the applicability of very efficient harmonic analysis and *synthesis* methods (Rizos 1979; Colombo 1981a). These methods have revolutionized the development and use of very high-degree spherical harmonic expansions. These approaches also allow efficient combination with *satellite-only* information, as was done, for example, by Rapp and Pavlis (1990). However, incorporation of altimeter data into a GGM in this fashion requires some a priori knowledge of the DOT, so that the altimetry-derived gravity anomaly values are estimated from appropriately corrected SSH.

Satellite tracking, altimetric, and surface gravimetric data are of complimentary character both in a spectral as well as in a geographic sense. Their combination enables the determination of the gravitational field, over a wider band of its spectrum, with improved accuracy, than can be obtained by using any of the three data types alone. The particular means of combining these data, in order to develop a high-degree GGM, constitutes a solution strategy. A critical consideration in the design of a solution strategy is the treatment of altimeter data (Rapp 1993), that is, if these data will be incorporated as in (3) or as in (4) above. OSU91A (Rapp et al. 1991) and EGM96 (Lemoine et al. 1998) represent the result of implementing a particular solution strategy, whereby altimeter data were used as in (3) for the determination of the low degree part of these models (maximum degree 50 and 70, respectively), and as in (4) for the higher degree part. The OSU89A/B models (Rapp and Pavlis 1990) and the



Gravity, Global Models, Fig. 5 The major sources of free-air gravity anomaly data used to compile the "merged" file that supported the development of EGM2008. ArcGP stands for "Arctic Gravity Project," SIO/NOAA for "Scripps Institute of Oceanography/National Oceanic

and Atmospheric Administration," which provided the major portion of the altimetry-derived gravity anomalies. Fill-in values were computed by augmenting low-resolution gravitational information, by the high-resolution gravitational information implied by the topography

EGM2008 model (Pavlis et al. 2008) incorporated satellite altimeter data as in (4) above. Certain characteristics of the above data types that are particularly important for their effective combination are discussed next.

- *Information Content.* The observables within the above four categories contain information not only about the gravitational field, but also about numerous other effects. Some of these effects are of interest in their own right (e.g., the DOT information contained within altimetric SSH), while others represent, at least as far as gravitational modeling is concerned, systematic noise (e.g., the non-conservative forces acting on a satellite). In either case, effective incorporation of a particular data type into the combination solution requires precise modeling and optimal estimation of all the effects and signals contained within the observable. Otherwise, the estimated gravitational model can be severely corrupted by the mis-modeled (or unmodeled) systematic effects.
- *Spectral Sensitivity Overlap.* The development of a GGM through a *least squares adjustment* procedure combining different data types is meaningful, provided that the data used in the adjustment share some common degree of sensitivity to the gravitational signal over a certain portion of its spectrum (a range of harmonic degrees). Otherwise, there is little “adjustment” being performed to data representing disjoint spectral bands. For example, pre-GRACE satellite-only models have a narrow spectral sensitivity overlap with models recovered from surface gravity data alone. This complicated considerably the problem of optimal combination of these two data types. On the other hand, this also means that setting up and inverting extremely large linear systems corresponding to very high degree models may not be necessary, if a single data type (e.g., a complete global equiangular grid of gravity anomalies) uniquely determines the higher-degree portion of such a GGM.
- *Relative Weighting.* The optimal estimation of a GGM depends critically on the optimality of the relative weights assigned to the different data types. Considering the numerous sources of data that may be involved, this may

require the solution to a very large *component of variance* estimation problem, complicated further by the fact that the extraction of gravitational information from satellite orbit observations is a strongly nonlinear problem. Although approximate solutions to this relative weight estimation problem have been used with considerable success (Lerch 1991), many times the experience and intuition of the model developer(s) guide the selection of appropriate data weights more than anything else.

The Recent Gravity-Mapping Satellite Missions

The satellite data used for the development of all GGMs published by the end of the twentieth century represent tracking of “targets of opportunity,” that is, of spacecraft designed and equipped with instrumentation for applications other than the mapping of the gravitational field from space. As a result of three satellite missions, this situation has changed dramatically during the last decade. These three missions are CHAMP (Reigber et al. 1996), GRACE (GRACE 1998), and GOCE (ESA 1999). Table 1 summarizes the main characteristics of these missions.

A comprehensive discussion regarding the concepts involved in these three mission scenarios can be found in (ESA 1999, Sect. 2.3). Mapping of the gravitational field from space requires missions that adhere as much as possible to the following fundamental design constraints:

- Uninterrupted tracking in three spatial dimensions.
- Measurement or compensation of the effects of non-gravitational forces.
- Orbital altitude as low as possible, to enhance sensitivity to the gravitational signal, and inclination as high as possible, to permit (near) global coverage.
- Counteraction of the field’s attenuation at altitude through the measurement of derivatives of the potential.

All three missions have in common the high–low Satellite-to-Satellite Tracking component (SST-*hl*) from the GPS (and GLONASS in the case of GOCE) constellation, and the measurement of nongravitational forces by the onboard accelerometers. These data permit highly accurate orbit

Gravity, Global Models, Table 1 Main characteristics of three gravity-mapping satellite missions

Mission	Status	Orbit	Mission objective	Instrumentation, tracking, and comments
CHAMP	Launched on 7/15/2000 active	Alt. = 450 km $e \approx 0.004$ $i = 87^\circ$	Gravity and magnetic fields Atmospheric limb Sounding Ionosphere sounding	3-axis STAR accelerometer GPS and SLR altitude will decay from 450 km (BOL) to 300 km (EOL)
GRACE	Launched on 3/17/2002 active	Alt. = 485 km $e \approx 0.001$ $i = 89^\circ$	Gravity field and its temporal variation	3-axis accelerometers (1 per s/c) GPS and SLR K-band inter-satellite ranging between the 2 s/c
GOCE	Launched on 3/17/2009 active	Alt. ≈ 250 km $i = 96.7^\circ$ Sun-synchronous	Gravity field (Especially static)	Six 3-axis accelerometers forming the gradiometer GPS/GLONASS and SLR

determination for all three missions, and in addition may enhance the gravitational field determination at very long wavelengths (very low degrees). In addition to that, GRACE involves the continuous measurement of the range between two identical satellites that “chase each other,” which constitutes a low-low SST formation (SST-II). GOCE’s accelerometer array on the other hand provides the measurements necessary to determine the gravitational *tensor* (i.e., the 3×3 matrix of second-order spatial derivatives of the gravitational potential) at altitude. GOCE is unique in the sense that it will provide boundary data at altitude covering the entire Earth, *except* for two polar cap areas of $\sim 6.7^\circ$ radius (due to the satellite’s inclination). The data acquired by each of these three missions result in different sensitivities to the gravitational spectrum. Simulation studies examining the geopotential recovery attainable from these (and other) mission scenarios have been reported, for example, by Sneeuw and Ilk (1997). When considering the data from these satellite missions two main questions arise:

1. What is the optimal way of analyzing their data?
2. What is the optimal way of combining their data with existing data, for example, from surface gravimetry and from satellite altimetry, in order to develop high-degree combination gravitational models?

1. *Data Analysis.* In the case of CHAMP, the gravitational information is extracted from the analysis of the perturbations of a low Earth orbiter, in a fashion similar to other preexisting satellite missions. However, CHAMP’s low orbit, in conjunction with the accelerometer data and with the availability of nearly global tracking data coverage, enabled *for the first time* the determination of an accurate long wavelength global gravitational model from a *single* satellite’s data. Indicative of this “new state of affairs” was the fact that a very preliminary solution (complete to degree and order 91) was already developed based on a *single* month’s worth of CHAMP data *only* and was presented during the 2001 meeting of the International Association of Geodesy (IAG) by Reigber et al. (2001). Although significantly better models that include GRACE data have by now surpassed considerably this preliminary solution, it served as a good example of the improvements that were to follow.

Compared to CHAMP, GRACE added the SST-II component, which permitted higher-resolution gravitational information to be extracted from the analysis of the orbital perturbation *differences* along the line of sight of the two low orbiting satellites. One can use traditional orbit perturbation analysis methods to process the GRACE data and derive a GGM, for example, in spherical harmonics. GGM01S (Tapley et al. 2004) and GGM02S (Tapley et al. 2005) were estimated

following such a procedure. This analysis scenario, albeit costly, is within current computational capabilities for models extending to degree and order 180 or so. Geodesists have also considered alternative analysis methods for GRACE-type missions (e.g., Wolff 1969; Colombo 1981b; Jekeli 1999a; Rowlands et al. 2002). Such methods provide higher computational efficiency at the cost of introducing certain approximations. Setting aside for a moment the details of optimal GRACE data processing, it is important to recognize here the quantum leap that has been accomplished with the GRACE mission. Approximately 14 months of GRACE data *alone* have been used to develop GGM02S, whose cumulative global Root Mean Square (RMS) geoid undulation error to degree 70 is less than ± 1 cm (Tapley et al. 2005, Fig. 2). By comparison, the cumulative geoid undulation global RMS error to degree 70 for EGM96, which required the combination of data from tens of satellites, along with surface gravity and satellite altimetry, was ± 19 cm (Lemoine et al. 1998, Table 10.3.2-1).

As far as GOCE is concerned, numerous investigations of the various aspects of its data analysis and of the development of a GGM from them can be found in (ESA 2000). One particular issue that receives increased attention relates to the polar gaps and their impact on analysis schemes that exploit regularity and completeness in the data coverage.

2. *Data Combination.* While the existing satellite gravity-mapping missions (CHAMP and especially GRACE) have already delivered (or promise to deliver as in the case of GOCE) quantum leaps in the accuracy and resolution of the *satellite-only* gravitational models, there is still a need to combine that information with terrestrial gravity and satellite altimetry data in an optimal fashion. This is required so that a “seamless” extension of the gravitational spectrum can be achieved, taking advantage of the rich high frequency content of the surface and altimetric data. The higher resolution of GRACE-based and GOCE-based *satellite-only* models will significantly increase the spectral overlap with surface gravity and altimetry. This will enable for the first time the estimation of high-resolution models of the DOT as well as the significant reduction of systematic errors present in surface gravity data (see also [Pavlis 2000]).

Beyond the Sensitivity of Satellite Data

Due to the attenuation of the gravitational signal with increasing altitude, there is obviously a limit to the gravitational information that can be extracted from spaceborne sensors. This implies a limit to the resolution (maximum degree) of satellite-only models. Spectral estimation indicates that for

GRACE, the signal-to-noise ratio becomes one around degree 170, while for GOCE the corresponding degree is not expected to exceed 300. Observations made on the Earth's surface (or on airplanes flying at low altitudes) can extend the resolution of gravitational models considerably. Surface and airborne data like *gravity anomalies*, *gravity disturbances*, etc. are therefore capable of supporting the development of much higher resolution gravitational models than those developed based on satellite data only. One way of developing such high-degree models involves first the formation of a regular grid of spatially averaged values of some functional of the field that completely covers the globe. These values represent data derived on the basis of other primary observables, using techniques like *Least Squares Collocation* (LSC) (Moritz 1980). Surface gravity anomalies are a suitable choice for such a functional, both from a spectral sensitivity and from an availability viewpoint. Very high-degree and order spherical harmonic models are then developed from the analysis of such global anomaly grids, using efficient harmonic analysis techniques like those put forward by Colombo (1981a). Wenzel (1998, 1999) reported such expansions complete to degree and order 1800. Pavlis et al. (2005) reported the development and evaluation of the PGM2004A preliminary gravitational model, which combined GRACE information with a global set of $5' \times 5'$ gravity anomalies, and extended to maximum degree and order 2,160. The final product of that modeling effort was EGM2008 which also extends to degree 2,160 (Pavlis et al. 2008). EGM2008 was completed and released to the public in April 2008.

State-of-the-Art Global Gravitational Modeling

We briefly outline the main aspects of the development of two global gravitational models, representative of the state-of-the-art, given the data that were available, at the time of each model's development: EGM96 (Lemoine et al. 1998), which represents the state-of-the-art before the availability of data from the CHAMP and GRACE missions, and EGM2008 (Pavlis et al. 2008), which represents currently (March 2010) the model with the highest resolution and accuracy, prior to the anticipated availability of data from GOCE. A comparison of the approaches that were followed for the development of these two models reveals the critical changes in model development, which were brought about by the availability of GRACE data on one hand, and of high quality $5' \times 5'$ area-mean gravity anomalies (from the combination of terrestrial and altimetry-derived data sources) on the other.

In theory, the estimation of a GGM complete to some (arbitrary) high degree and order, from the combination of various (heterogeneous) data, could be carried out as follows:

- (a) Form separate normal equations from each individual data type, to a maximum degree and order that corresponds to the resolution of the available data and their sensitivity to the gravitational signal.
- (b) Treat satellite altimeter data as “direct tracking” observations, that is, ranges from the spacecraft to the ocean surface whose upper endpoint senses (through the orbit dynamics) attenuated gravitational signals (static and time varying), while their lower end point senses the combined effects of geoid undulation, DOT as well as tides and other time varying effects, without any attenuation. In this manner, altimeter data contribute to the estimation of the satellite's orbit, as well as the estimation of the DOT and of the potential coefficients.
- (c) Combine the various normal equations (with appropriate relative weights) and invert the resulting system, to estimate the combination solution to its high degree, along with its *full* error covariance matrix.

Such an “ideal” approach would permit the most rigorous modeling of the observables and would allow the greatest flexibility in terms of data weighting. A combination solution to degree 360, if performed as outlined above, would require the formation of full (symmetric) normal equation matrices (from satellite altimetry and surface gravimetry) for approximately 130,000 parameters (considering *only* the gravitational potential coefficients). For maximum degree 2,160, there would be approximately 4.7 million such parameters involved. Such computational tasks are beyond our present computational capabilities. Therefore, at present, one may choose between, or combine, two main solution strategies to attack the problem:

- *Solution Strategy (A)* Apply the “ideal” estimation strategy outlined above, to obtain a combination solution for the lower degree part of the field, up to a maximum degree that is computationally manageable. Apart from reasons of computational feasibility, this maximum degree should enable the appropriate modeling of the gravitational signal contained in the currently available satellite tracking data. Furthermore, since the DOT signal is of long wavelength nature, the benefits of “direct” altimetry are almost entirely retained here. To avoid aliasing effects however, the contribution to the altimetry and surface gravity data from the coefficients beyond the solved-for degree has to be filtered out of the data prior to the normal equation formation. This may be done using a preexisting high-degree solution (Pavlis 1988; Denker and Rapp 1990). Herein, we will refer to this type of solution as the *comprehensive* low-degree combination model. The obvious shortcomings of this approach are the relatively low maximum attainable degree (approximately 250 at present) and its computational demands. Some models developed using this approach (or similar ones) include JGM-1 and JGM-2 (Nerem et al. 1994) and JGM-3 (Tapley et al. 1996), the

part of EGM96 up to degree and order 70 (see Lemoine et al. 1998, Chap. 7), and EIGEN-GL04C (Förste et al. 2008).

- *Solution Strategy (B)* Consider that one is willing to make the following two approximations:
 1. The orbits of the altimeter satellites, whose data are included in the combination solution, are perfectly known (at least radially). This approximation is justifiable if one is working with altimeter satellites supported by T/P-class precise orbit determination. Moreover, after the availability of GRACE-based gravitational models for precise orbit determination of altimeter satellites, errors arising from gravitational model inaccuracies do not dominate the orbit error budget of these satellites. Errors due to, for example, mis-modeling of nonconservative forces acting on the spacecraft are likely to be more significant at present. In this regard, to allow the orbits of altimeter satellites to contribute (through their dynamics) to the determination of gravitational parameters within a combination solution may not be a desirable approach presently, because the effects of orbit errors of non-gravitational origin could corrupt the solved-for gravitational parameters.
 2. The DOT is known a priori, for example, from an Ocean Circulation Model (OCM) or from a previous low-degree comprehensive combination solution.

The implication of (1) is that satellite altimetry does not have to be treated as “direct” tracking anymore, which simplifies the problem considerably, since now orbit dynamics are not involved in the altimeter data processing. One is left with a “surface” problem, where the geoid (N) and the DOT (ζ) signals have to be separated, given the “observed” SSH (h), which is their sum. If in addition, the approximation (2) is introduced, then altimetry contributes to the combination solution “observed” geoid heights (N) over (parts of) the ocean.

In addition to the above two approximations, a key issue here is that altimetric information may also be provided in a *gridded* form. This is possible through the use of a Mean Sea Surface (MSS), obtained from multiple altimetric missions. The success of T/P has significantly improved the accuracy of MSS data sets (especially at long wavelengths). This is accomplished by adjusting the SSH data from other altimetric missions (e.g., ERS-1, ERS-2, GEOSAT, GEOSAT-Follow-On), to the surface defined by T/P, using crossover minimization techniques. Such MSS data sets have been developed by, for example, Yi (1995), Kim et al. (1995), Anzenhofer et al. (1996), Wang (2001), and more recently by Andersen and Knudsen (2009). One may also have available in *gridded* form, altimetry-derived gravity anomaly values. Such values have been estimated using various techniques, on an ocean-wide basis by, for example, Rapp and

Basic (1992), Andersen and Knudsen (1998), Andersen et al. (2010), Trimmer and Manning (1996) and Sandwell and Smith (1997, 2009) among others.

The two simplifying approximations discussed above and particularly the availability of altimetric information in gridded form (especially in the form of gravity anomalies) make applicable an alternative class of high-degree combination solution techniques. These techniques combine the satellite-only information, with potential coefficient information obtained from the analysis of *complete*, regular grids of functional(s) of the disturbing potential (e.g., N , Δg), and are based on the highly efficient harmonic analysis algorithms originally studied and put forward by Colombo (1981a). These algorithms exploit the regularity of the data grids and the symmetry properties of Legendre and trigonometric (sine/cosine) functions. Inherent in these algorithms is the use of Fast Fourier Transform (FFT) techniques to process data arrays residing over latitude bands that are symmetric with respect to the equator, in a highly efficient manner. Estimators of this type are the (simple) Numerical Quadrature (NQ), the Block-Diagonal (BD) least-squares adjustment, and the Optimal Estimation (OE) technique. Models developed using the NQ approach include OSU86E/F (Rapp and Cruz 1986a) and OSU89A/B (Rapp and Pavlis 1990). BD techniques of varying sophistication have been used to develop GPM2 (Wenzel 1985), DGF192A (Gruber and Bosch 1992), GFZ95A (Gruber et al. 1996), and EGM2008 (Pavlis et al. 2008). OE was used to develop the OSU86C/D models (Rapp and Cruz 1986b).

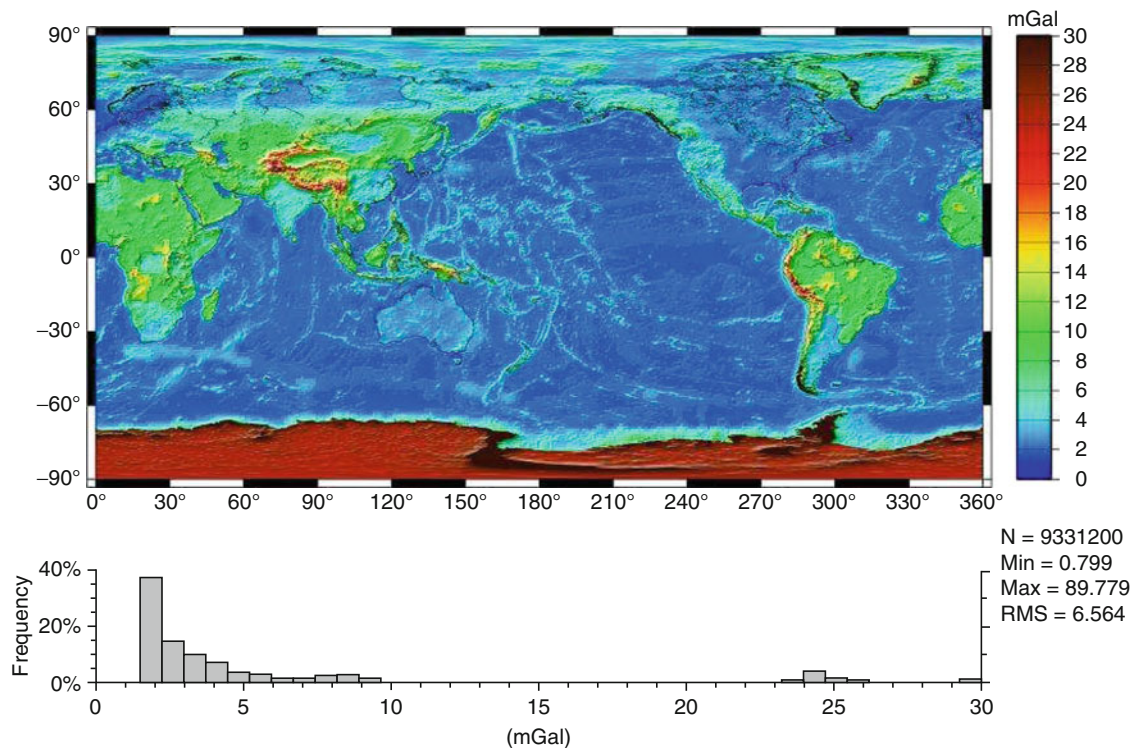
The two solution strategies (A) and (B) discussed above have their respective advantages and disadvantages. Due to computational limitations, the development of a GGM to spherical harmonic degree 360 or higher, currently requires the implementation of either a combination of the two strategies (as it was done for EGM96 and EIGEN-GL04C), or the implementation of strategy (B), as it was done for OSU89A/B and EGM2008. EGM96 (Lemoine et al. 1998) employed a comprehensive solution to degree 70, augmented by a BD solution from degree $n = 71$ to degree $n = 359$, while the $n = 360$ coefficients (which cannot be estimated using the BD technique) were obtained from an NQ model. The main reason for the choice of this solution strategy was the fact that the satellite-only component of EGM96 (denoted EGM96S), was accompanied by a variance-covariance matrix that was fully occupied. This was due to the fact that the heterogeneous tracking data from the (approximately) 40 satellites that were used to derive EGM96S to degree and order 70, were incapable of de-correlating adequately the spherical harmonic coefficients up to this degree and order. Therefore, in order to preserve the integrity of the least-squares adjustment used to combine EGM96S with the surface gravity and altimetry data, one had to consider the satellite-only normal equations in their complete (fully

occupied) form. Any block-diagonal approximation of these normal equations would result in estimation errors that could not be tolerated. This situation changed dramatically with GRACE, where the corresponding normal equation matrix, due to the global coverage and uniform GRACE data accuracy could be safely approximated by a block-diagonal matrix, without significant loss of accuracy. In addition, after the availability of satellite-only models from GRACE, there is really no need to incorporate altimeter data into the combination solution in the form of “direct” tracking. Instead, a preliminary model from GRACE and an MSS can be used to estimate a preliminary model of the DOT. This could then be used to correct the SSH of the MSS and estimate an ocean-wide set of altimetry-derived gravity anomalies. These can be “merged” with terrestrial and airborne data to form a complete global grid (see also Fig. 5). The gravitational information implied by these gridded data could then be combined (in a least-squares adjustment) with the satellite-only model from GRACE, to derive the combination solution, up to the high degree (2,160), in a single step. The entire process may be iterated, using the high-degree combination solution to derive the next estimate of the DOT, and so on. This is essentially the approach that was used to develop EGM2008 (Pavlis et al. 2008). Two iterations of the estimation of the altimetry-derived gravity anomalies were performed, which was

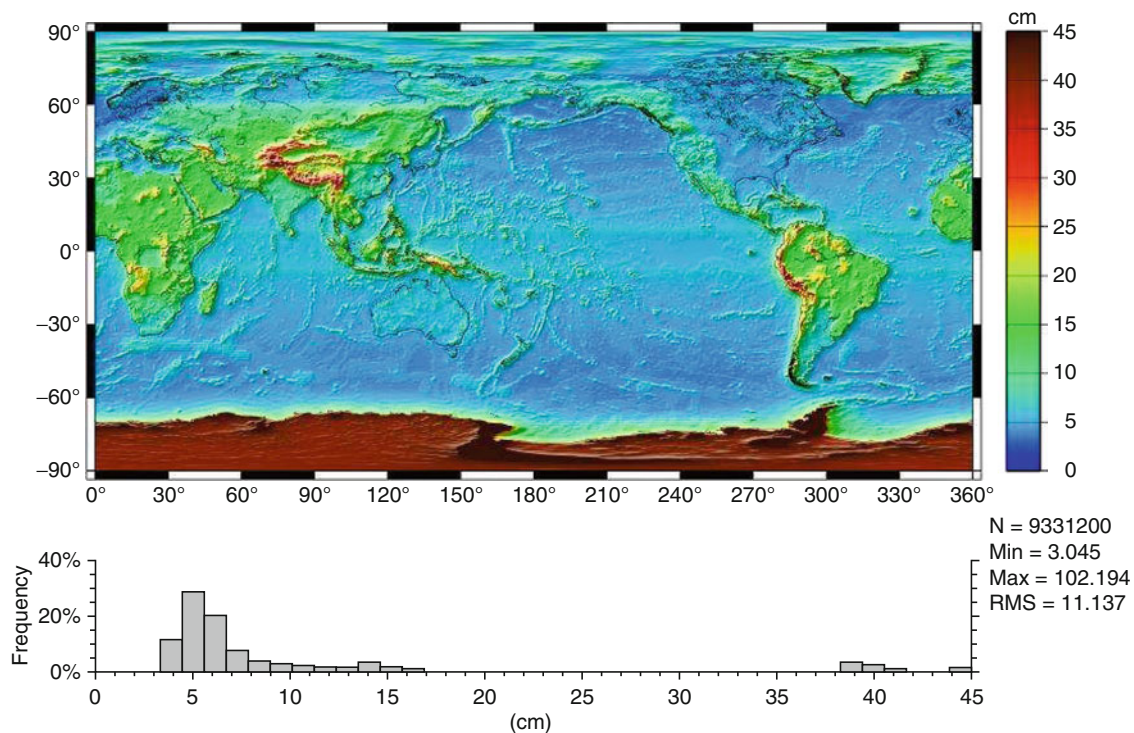
sufficient for the process to converge. Despite its iterative nature, this approach permits the development of very high-degree combination solutions in an efficient manner, and moreover in a single adjustment step, thereby avoiding the “piecewise” nature of models like EGM96.

Model Evaluation and Accuracy Assessment

Of central importance to the development of a GGM are the evaluation of its performance, and the estimation of reliable error measures to quantify its accuracy. The former requires the comparison of model-derived quantities to corresponding values obtained *independently*. Several different data types, withheld from a model’s development, are used for this purpose. These independent data have different spectral sensitivities and/or occupy different geographic regions. Tests that are usually employed here include satellite orbit determination and comparisons with tracking data, comparisons with geoid undulations obtained from GPS and leveling data, comparisons with deflections of the vertical obtained from astrogeodetic techniques (Jekeli 1999b), comparisons employing altimetry and general ocean circulation models, etc. A useful practice, introduced during the development of EGM96, and used also during the development of EGM2008, is to invite a voluntary evaluation working group, independent of a model’s developers, that evaluates and provides



Gravity, Global Models, Fig. 6 Propagated error estimates (in mGal) of the free-air gravity anomalies computed using the EGM2008 model. Notice the large error estimates over Antarctica and over mountainous areas (Himalaya, Andes), due to the lack of accurate and detailed data



Gravity, Global Models, Fig. 7 Propagated error estimates (in centimeters) of the geoid undulations computed using the EGM2008 model. The larger values around the equator are due to the lower density of GRACE data at low latitudes

feedback to the model's developers regarding candidate preliminary solutions, as well as the final outcome from a modeling effort, in a manner as objective as possible. These groups work under the auspices of the International Association of Geodesy (IAG) and upon completion of a certain evaluation effort they report their findings in IAG-sponsored publications, which can be accessed freely by the public. In the case of EGM2008, the results from such an evaluation by 25 different international investigating teams are reported in *Newton's Bulletin No. 4*, which is jointly published by the Bureau Gravimétrique International (BGI) and the International Geoid Service (IGeS).

The estimation of reliable errors that should accompany a model's derived quantities is also very important. These errors are geographically dependent, reflecting the different quality of the data used to develop the model, as a function of geographic area. For models like EGM2008, which involve a very large number of parameters (approximately 4.7 million), one cannot perform such error propagation using variance-covariance matrices and linear algebra methods, due to the excessively large sizes of the matrices involved. Instead, Pavlis and Saleh (2005) introduced and implemented an alternative approach that employs global convolutions with band-limited kernels. This technique is capable of providing very reliable error estimates, with geographic specificity, in a very efficient manner. Figures 6 and 7 illustrate geographically the propagated errors in free-air

gravity anomalies and in geoid undulations, implied by EGM2008 to degree 2,160.

Summary

Accurate knowledge of the gravitational potential of the Earth, on a global scale and at high resolution, is a fundamental prerequisite for various geodetic, geophysical, and oceanographic investigations and applications. Over the last 50 years, continuing improvements and refinements to the basic theory have been paralleled by the availability of more accurate and complete data and by dramatic improvements in the computational resources available for numerical modeling studies. These advances have brought the state-of-the-art from the early spherical harmonic models of degree 8 (Zhongolovich 1952) to solutions like EGM2008 that extends to degree 2,160. The availability of satellite missions like CHAMP, GRACE, and GOCE, dedicated to the mapping of the gravitational field from space, is enabling the determination of the long wavelength features of the field with unprecedented accuracy. Nevertheless, the determination of the fine details of the field still relies on the availability of terrestrial and airborne data, of high accuracy and sufficient geographic coverage and density. In this regard, several areas of the Earth, most notably Antarctica, await more comprehensive gravimetric surveying.

Cross-References

- [Geoid](#)
- [Geoid Determination, Computational Methods](#)
- [Geoid Determination, Theory and Principles](#)
- [Geoid Undulation, Interpretation](#)
- [Gravity Anomalies, Interpretation](#)
- [Gravity Field of the Earth](#)
- [Gravity Field, Temporal Variations from Space Techniques](#)
- [Gravity Field, Time Variations from Surface Measurements](#)
- [Gravity Measurements, Absolute](#)
- [Gravity Method, Airborne](#)
- [Gravity Method, Principles](#)
- [Gravity Method, Satellite](#)
- [Gravity Method, Surface](#)
- [Gravity Modeling, Theory and Computation](#)
- [Spherical Harmonic Analysis Applied to Potential Fields](#)

Bibliography

- Andersen OB, Knudsen P (1998) Global marine gravity field from the ERS-1 and GEOSAT geodetic mission altimetry. *J Geophys Res* 103(C4):8129–8137
- Andersen OB, Knudsen P (2009) DNSC08 mean sea surface and mean dynamic topography models. *J Geophys Res* 114(C11):C11001
- Andersen OB, Knudsen P, Berry PAM (2010) The DNSC08GRA global marine gravity field from double retracked satellite altimetry. *J Geod* 84(3):191–199. <https://doi.org/10.1007/s00190-009-0355-9>
- Anzenhofer M, Gruber T, Rentsch M (1996) Global high resolution mean sea surface based on ERS-1 35- and 168-day cycles and TOPEX data. In: Rapp RH, Cazenave AA, Nerem RS (eds) *Global gravity field and its temporal variations*. IAG Symposia, vol 116. Springer, Berlin/Heidelberg
- Baker EM (1988) A finite element model of the Earth's anomalous gravitational potential. Rep. 391, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Bouman J (1997) A survey of global gravity models. Delft Institute for Earth-Oriented Space Research. Rep. No. 97.1, Delft Institute of Technology, Delft
- Colombo OL (1981a) Numerical methods for harmonic analysis on the sphere. Rep. 310, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Colombo OL (1981b) Global geopotential modelling from satellite-to-satellite tracking. Rep. 317, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Denker H, Rapp RH (1990) Geodetic and oceanographic results from the analysis of 1 year of geosat data. *J Geophys Res* 95(C8):13151–13168
- ESA (2000) From Eötvös to mGal – Final report, ESA/ESTEC Contract No. 13392/98/NL/GD, Graz
- ESA SP-1233 (1) (1999) The four candidate Earth explorer core missions – gravity field and steady-state ocean circulation mission. ESA Publications Division, ESTEC, Noordwijk
- Förste C, Schmidt R, Stubenvoll R, Flechtner F, Meyer U, König R, Neumayer H, Biancale R, Lemoine J-M, Bruinsma S, Loyer S, Barthelmes F, Esselborn S (2008) The GeoForschungsZentrum Potsdam/Groupe de Recherche de Géodésie Spatiale satellite-only and combined gravity field models: EIGEN-GL04S1 and EIGEN-GL04C. *J Geod* 82(6):331–346. <https://doi.org/10.1007/s00190-007-0183-8>
- Fu L-L, Christensen E, Yamarone C Jr, Lefebvre M, Ménard Y, Dorner M, Escudier P (1994) TOPEX/POSEIDON: mission overview. *J Geophys Res* 99(C12):24369–24381
- Ganachaud A, Wunsch C, Kim M-C, Tapley B (1997) Combination of TOPEX/POSEIDON data with a hydrographic inversion for determination of the oceanic general circulation and its relation to geoid accuracy. *Geophys J Int* 128:708–722
- GRACE (1998) Gravity Recovery and Climate Experiment: science and mission requirements document, revision A, JPLD-15928, NASA's Earth System Science Pathfinder Program
- Gruber T, Bosch W (1992) A new 360 gravity field model, Presented at the XVII General Assembly of the European Geophysical Society. Edinburgh
- Gruber T, Anzenhofer M, Rentsch M (1996) The 1995 GFZ high resolution gravity model. In: Rapp RH, Cazenave AA, Nerem RS (eds) *Global gravity field and its temporal variations*. IAG Symposia, vol 116. Springer, Berlin
- Heck B (1990) An evaluation of some systematic error sources affecting terrestrial gravity anomalies. *Bull Géodésique* 64:88–108
- Heiskanen WA, Moritz H (1967) *Physical geodesy*. WH Freeman, San Francisco
- Jekeli C (1999a) The determination of gravitational potential differences from satellite-to-satellite tracking. *Celestial Mech Dyn Astr* 75(2):85–100
- Jekeli C (1999b) An analysis of vertical deflections derived from high-degree spherical harmonic models. *J Geod* 73(1):10–22. <https://doi.org/10.1007/s001900050213>
- Jekeli C (2005) Spline representations of functions on a sphere for geopotential modeling. Rep. 475, Department of Civil and Environmental Engineering and Geodetic Science, Ohio State University, Columbus
- Kenyon SC, Pavlis NK (1996) The development of a global surface gravity data base to be used in the joint DMA/GSFC geopotential model. In: Rapp RH, Cazenave AA, Nerem RS (eds) *Global gravity field and its temporal variations*. IAG Symposia, vol 116. Springer, Berlin
- Kim M-C, Tapley BD, Shum C-K, Ries JC (1995) Center for Space Research mean sea surface model, Presented at the TOPEX/POSEIDON Working Team Meeting, Pasadena, California
- Lemoine FG, Kenyon SC, Factor JK, Trimmer RG, Pavlis NK, Chinn DS, Cox CM, Klosko SM, Luthcke SB, Torrence MH, Wang YM, Williamson RG, Pavlis EC, Rapp RH, Olson TR (1998) The development of the joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) geopotential model EGM96, NASA technical publication, TP-1998-206861, p 575
- Lerch FJ (1991) Optimum data weighting and error calibration for estimation of gravitational parameters. *Bull Géodésique* 65:44–52
- Lerch FJ, Klosko SM, Laubscher RE, Wagner CA (1979) Gravity model improvement using GEOS-3 (GEM9 and 10). *J Geophys Res* 84(B8):3897–3916
- Meissl P (1981) The use of finite elements in physical geodesy. Rep. 313, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Moritz H (1980) *Advanced physical geodesy*. Herbert Wichmann Verlag, Karlsruhe. Federal Republic of Germany
- Nerem RS, Lerch FJ, Marshall JA, Pavlis EC, Putney BH, Tapley BD, Eanes RJ, Ries JC, Schutz BE, Shum CK, Watkins MM, Klosko SM, Chan JC, Luthcke SB, Patel GB, Pavlis NK, Williamson RG, Rapp RH, Biancale R, Nouel F (1994) Gravity model development for TOPEX/POSEIDON: joint gravity models 1 and 2. *J Geophys Res* 99(C12):24421–24447
- Nerem RS, Jekeli C, Kaula WM (1995) Gravity field determination and characteristics: retrospective and prospective. *J Geophys Res* 100(B8):15053–15074. <https://doi.org/10.1029/94JB03257>
- Pavlis NK (1988) Modeling and estimation of a low degree geopotential model from terrestrial gravity data. Rep. 386, Department of Geodetic Science and Survey, Ohio State University, Columbus

- Pavlis NK (2000) On the modeling of long wavelength systematic errors in surface gravimetric data. In: Schwarz K-P (ed) *Geodesy beyond 2000 – the challenges of the first decade*. IAG Symposia, vol 121. Springer, Berlin
- Pavlis NK, Rapp RH (1990) The development of an isostatic gravitational model to degree 360 and its use in global gravity modelling. *Geophys J Int* 100:369–378
- Pavlis NK, Saleh J (2005) Error propagation with geographic specificity for very high degree geopotential models. In: Jekeli C, Bastos L, Fernandes J (eds) *Gravity, geoid and space missions*. IAG Symposia, vol 129. Springer, Berlin
- Pavlis NK, Holmes SA, Kenyon SC, Schmidt D, Trimmer R (2005) A preliminary gravitational model to degree 2160. In: Jekeli C, Bastos L, Fernandes J (eds) *Gravity, geoid and space missions*. IAG Symposia, vol 129. Springer, Berlin
- Pavlis NK, Factor JK, Holmes SA (2007) Terrain-related gravimetric quantities computed for the next EGM. In: *Gravity field of the Earth, Proceedings of the 1st international symposium of the International Gravity Field Service (IGFS)*, Harita Dergisi, Special Issue 18, Istanbul, pp 318–323
- Pavlis NK, Holmes SA, Kenyon SC, Factor JK (2008) An Earth gravitational model to degree 2160: EGM2008, Presented at the 2008 General Assembly of the European Geosciences Union, Vienna, April 13–18
- Rapp RH (1989) The decay of the spectrum of the gravitational potential and the topography of the Earth. *Geophys J Int* 99:449–455
- Rapp RH (1993) Use of altimeter data in estimating global gravity models. In: Rummel R, Sansò F (eds) *Satellite altimetry in geodesy and oceanography*. Springer, Berlin, pp 373–417
- Rapp RH (1997) Use of potential coefficient models for geoid undulation determinations using a spherical harmonic representation of the height anomaly/geoid undulation difference. *J Geod* 71:282–289
- Rapp RH (1998) Past and future developments in geopotential modeling. In: Forsberg R, Feissel M, Dietrich R (eds) *Geodesy on the move – gravity, geoid, geodynamics and Antarctica*. IAG Symposia, vol 119. Springer, Berlin
- Rapp RH, Balasubramania N (1992) A conceptual formulation of a world height system. Rep. 421, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Rapp RH, Basic T (1992) Oceanwide gravity anomalies from GEOS-3, SEASAT and GEOSAT altimeter data. *Geophys Res Lett* 19(19):1979–1982
- Rapp RH, Cruz JY (1986a) Spherical harmonic expansions of the Earth's gravitational potential to degree 360 using 30' mean anomalies. Rep. 376, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Rapp RH, Cruz JY (1986b) The representation of the Earth's gravitational potential in a spherical harmonic expansion to degree 250. Rep. 372, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Rapp RH, Pavlis NK (1990) The development and analysis of geopotential coefficient models to spherical harmonic degree 360. *J Geophys Res* 95(B13):21885–21911
- Rapp RH, Wang YM, Pavlis NK (1991) The Ohio State 1991 geopotential and sea surface topography harmonic coefficient models. Rep. 410, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Reiger C, Bock R, Förste C, Grunwaldt L, Jakowski N, Lühr H, Schwintzer P, Tilgner C (1996) CHAMP Phase-B executive summary, GFZ, STR96/13
- Reiger C, et al (2001) Global gravity field recovery with CHAMP. Presented at the 2001 IAG Scientific Assembly, Budapest, September 2–7
- Rizos C (1979) An efficient computer technique for the evaluation of geopotential from spherical harmonic models. *Aust J Geodesy, Photogramm Cartogr* 31:161–170
- Rowlands DD, Ray RD, Chinn DS, Lemoine FG (2002) Short-arc analysis of intersatellite tracking data in a gravity mapping mission. *J Geod* 76(6–7):307–316. <https://doi.org/10.1007/s00190-002-0255-8>
- Sandwell DT, Smith WHF (1997) Marine gravity anomaly from Geosat and ERS 1 satellite altimetry. *J Geophys Res* 102(B5):10039–10054
- Sandwell DT, Smith WHF (2009) Global marine gravity from retracked Geosat and ERS-1 altimetry: ridge segmentation versus spreading rate. *J Geophys Res* 114:B01411. <https://doi.org/10.1029/2008JB006008>
- Schwarz K-P, Sideris MG, Forsberg R (1987) Orthometric heights without leveling. *J Surv Eng* 113(1):28–40
- Sneeuw N, Ilk KH (1997) The status of spaceborne gravity field mission concepts: a comparative simulation study. In: Segawa J, Fujimoto H, Okubo S (eds) *Gravity, geoid and marine geodesy*. IAG Symposia, vol 117. Springer, Berlin
- Sünkel H (1981) Point mass models and the anomalous gravitational field. Rep. 328, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Sünkel H (1983) The generation of a mass point model from surface gravity data. Rep. 353, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Sünkel H (1984) Splines: their equivalence to collocation. Rep. 357, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Tapley BD, Watkins MM, Ries JC, Davis GW, Eanes RJ, Poole SR, Rim HJ, Schutz BE, Shum CK, Nerem RS, Lerch FJ, Marshall JA, Klosko SM, Pavlis NK, Williamson RG (1996) The joint gravity model–3. *J Geophys Res* 101(B12):28029–28049
- Tapley BD, Bettadpur S, Watkins M, Reiger C (2004) The gravity recovery and climate experiment: mission overview and early results. *Geophys Res Lett* 31:L09607. <https://doi.org/10.1029/2004GL019920>
- Tapley B, Ries J, Bettadpur S, Chambers D, Cheng M, Condi F, Gunter B, Kang Z, Nagel P, Pastor R, Pekker T, Poole S, Wang F (2005) GGM02 – an improved earth gravity field model from GRACE. *J Geod* 79(8):467–478. <https://doi.org/10.1007/s00190-005-0480-z>
- Trimmer RG, Manning DM (1996) The altimetry derived gravity anomalies to be used in computing the joint DMA/ NASA earth gravity model. In: Rapp RH, Cazenave AA, Nerem RS (eds) *Global gravity field and its temporal variations*. IAG Symposia, vol 116. Springer, Berlin
- Tscherning CC (1983) The role of high-degree spherical harmonic expansions in solving geodetic problems. In: *Proceedings of the international association of geodesy symposia, IUGG XVII General Assembly*, vol 1, Department of Geodesy Science and Survey, Ohio State University, Columbus, pp 431–441
- Wang YM (2001) GSFC00 mean sea surface, gravity anomaly, and vertical gravity gradient from satellite altimeter data. *J Geophys Res* 106(C12):31167–31174
- Wenzel H -G (1985) *Hochauflösende Kugelfunktionsmodelle für das Gravitationspotential der Erde*, Wiss. Arb. 137, Fachrichtung Vermess. der Univ. Hannover, Hannover, Federal Republic of Germany
- Wenzel G (1998) Ultra high degree geopotential models GPM98A, B and C to degree 1800. <http://www.gik.uni-karlsruhe.de/~wenzel/gpm98abc/gpm98abc.htm>
- Wenzel G (1999) Schwerefeldmodellierung durch ultrahochauflösende Kugelfunktionsmodelle. *Z Vermessungswesen* 124(5):144–154
- Wolff M (1969) Direct measurements of the Earth's gravitational potential using a satellite pair. *J Geophys Res* 74(22):5295–5300
- Yi Y (1995) Determination of gridded mean sea surface from TOPEX, ERS-1 and GEOSAT altimeter data. Rep. 434, Department of Geodetic Science and Survey, Ohio State University, Columbus
- Zhongolovich ID (1952) The external gravitational field of the Earth and the fundamental constants related to it. *Acad Sci Publ Inst Teor Astron Leningrad*

Gravity, Gradiometry

Christopher Jekeli

Division of Geodetic Science, School of Earth Sciences, Ohio State University, Columbus, OH, USA

Definition and Scope

Gravity gradiometry encompasses all manners of measuring the spatial gradients of the gravity vector field. In a broader sense, it includes the techniques for developing and applying gradient models, from both analytic and numerical perspectives, in order to validate gradient sensors, to study geologic bodies, and to test physical theories. Gravity gradiometry has a long and varied history of static (ground-based), shipborne, airborne, and spaceborne measurement systems. This entry includes the mathematical foundations and properties of gradients starting with Newton's law of gravitation. A description of straightforward modeling is followed by a rudimentary measurement error analysis; and the entry concludes with a comprehensive account of measurement systems, as well as some sample applications.

Introduction

Gravity gradiometry is over 120 years old, and it offers a rich perspective of the Earth's external gravitational field as generated by its crustal mass density structure. Like the measurement of gravity, gradiometry is used to make both qualitative and quantitative interpretations of the Earth's near subsurface interior. The spatial gradients of the Earth's gravity acceleration vector form a tensor of nine elements, three components for each of the three gravity vector components. They emphasize the fine structure and the short-wavelength characteristics of the field. They are largest over sharp subsurface geologic contrasts, such as faults, the edges of salt domes and igneous intrusions, and the flanks of synclines and anticlines. As the second derivatives of the gravity potential, the gravity gradients describe the curvature of the potential, and values of the differential curvature help to indicate directions of principal axes of source bodies. Gravity gradiometry also plays a significant role in understanding the fundamental nature of gravitation from a cosmological viewpoint.

While popular in the early part of the twentieth century to aid in the exploration for oil and gas, static gravity gradiometry has largely been replaced by scalar gravimetry (measurements of the magnitude of the gravity vector), both in the static mode and on airborne platforms, due to its greater efficiency and yet comparable accuracy. Gradiometry has resurfaced recently because for moving-base systems (e.g., aircraft and satellites), it is arguably the preferred type of

measurement since it does not require an independent system that determines the kinematic accelerations of the vehicle with equal precision, as does the gravimeter due to the principle of equivalence between inertial and gravitational mass.

Either the static or moving-base measurement of gradients requires exquisitely sensitive instrumentation, which over these last 120 years has improved less in accuracy than in resolution (integration time) for the surface and airborne systems. Typical airborne systems today advertise the same precision that the static instruments had at the beginning of the twentieth century. However, they have a much higher temporal resolution (of the order of a few seconds) compared to the many hours required for the early static instruments. On the other hand, the recent launch of a satellite-borne gradiometer represents a major advance in global gradiometry with an accuracy improvement of one to two orders of magnitude, but at the expense of sensing the field in the relatively far removed (from the sources), but dynamically quiet environment of space. Because of the relatively great expense of current gradiometer technology, to date only a few terrestrial (airborne and shipborne) systems are in operation.

The following sections cover the basic theory of gravitational gradients, the history of the sensor technology from the first torsion balance to modern-day gradiometers, the forward modeling of gravitational gradients, and applications in geodesy, geophysics, and physics.

Mathematical Foundations

Gravitational Gradients

Newtonian gravitational field theory adequately serves both to develop the necessary field equations, their solutions, and the derivation and use of gravitational gradients. No attempt is made or needed for present purposes to discuss the gradients in terms of Einstein's theory of general relativity. One may start with Newton's law of gravitation that posits the force between two point masses as proportional to each of the masses, inversely proportional to the square of the distance between them, and directed along the connecting line. An equivalent concept is the Newtonian gravitational field potential that quantifies the work (energy) needed to bring a unit test mass from a position, $\mathbf{x}$, to infinity for a given attracting point mass, m , located at $\mathbf{x}'$,

$$V(\mathbf{x}) = \frac{Gm}{|\mathbf{x} - \mathbf{x}'|}, \quad (1)$$

where G is Newton's gravitational constant and the geophysical convention of positive potential is used. By the law of superposition, the potential is generalized by summation to a discrete set of mass points and, ultimately, using the concept of mass density, $\rho = dm/dv$, to (piecewise) continuous distributions of mass,

$$V(\mathbf{x}) = G \int_v \frac{\rho(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} dv, \quad (2)$$

where v is the region in space occupied by the mass density. The potential of a volume distribution is a differentiable (hence also continuous) function for all points inside and outside v and on its boundary.

The spatial gradient of the gravitational potential is the gravitational acceleration (gravitation) experienced by the test mass at $\mathbf{x}$,

$$\mathbf{g}(\mathbf{x}) = \nabla V(\mathbf{x}). \quad (3)$$

The spatial gradient of the components of the gravitation vector constitutes the elements of the gravitational gradient tensor,

$$\mathbf{\Gamma}(\mathbf{x}) = \nabla \nabla^T V(\mathbf{x}) = \nabla \mathbf{g}^T(\mathbf{x}), \quad (4)$$

where “T” denotes the transpose of a vector (or matrix). In Cartesian coordinates, $\mathbf{x} = (x_1, x_2, x_3)^T$, the gravitational gradient tensor is given by

$$\begin{aligned} \mathbf{\Gamma} &= \begin{pmatrix} \frac{\partial^2 V}{\partial x_1^2} & \frac{\partial^2 V}{\partial x_1 \partial x_2} & \frac{\partial^2 V}{\partial x_1 \partial x_3} \\ \frac{\partial^2 V}{\partial x_2 \partial x_1} & \frac{\partial^2 V}{\partial x_2^2} & \frac{\partial^2 V}{\partial x_2 \partial x_3} \\ \frac{\partial^2 V}{\partial x_3 \partial x_1} & \frac{\partial^2 V}{\partial x_3 \partial x_2} & \frac{\partial^2 V}{\partial x_3^2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_2}{\partial x_1} & \frac{\partial g_3}{\partial x_1} \\ \frac{\partial g_1}{\partial x_2} & \frac{\partial g_2}{\partial x_2} & \frac{\partial g_3}{\partial x_2} \\ \frac{\partial g_1}{\partial x_3} & \frac{\partial g_2}{\partial x_3} & \frac{\partial g_3}{\partial x_3} \end{pmatrix} = \begin{pmatrix} \Gamma_{1,1} & \Gamma_{1,2} & \Gamma_{1,3} \\ \Gamma_{2,1} & \Gamma_{2,2} & \Gamma_{2,3} \\ \Gamma_{3,1} & \Gamma_{3,2} & \Gamma_{3,3} \end{pmatrix}. \end{aligned} \quad (5)$$

The tensor is symmetric and exists where the potential is twice differentiable.

Poisson’s field equation characterizes a Newtonian gravitational field through its second-order gradients, specifically by the trace of the gravitational gradient tensor,

$$\nabla \cdot \nabla V(\mathbf{x}) = \nabla^2 V(\mathbf{x}) = -4\pi G \rho(\mathbf{x}), \quad (6)$$

where the Laplacian operator in Cartesian coordinates is

$$\nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}. \quad (7)$$

Equation (6) shows that the gradient does not exist at discontinuities of the mass density, such as at the exact interface between the solid Earth and atmosphere. By its symmetry

and Poisson’s equation, the gravitational gradient tensor has only five independent elements for points where the mass density distribution is known. For later reference, the potential in free space satisfies Laplace’s equation,

$$\nabla^2 V(\mathbf{x}) = 0. \quad (8)$$

If a local coordinate system has its three-axis aligned with the vertical, positive downward, and the one-axis points northward, then the horizontal rate of change of the vertical component of gravitation, g_3 , is called the horizontal gradient, with magnitude and direction of greatest variation (from north), respectively, given by

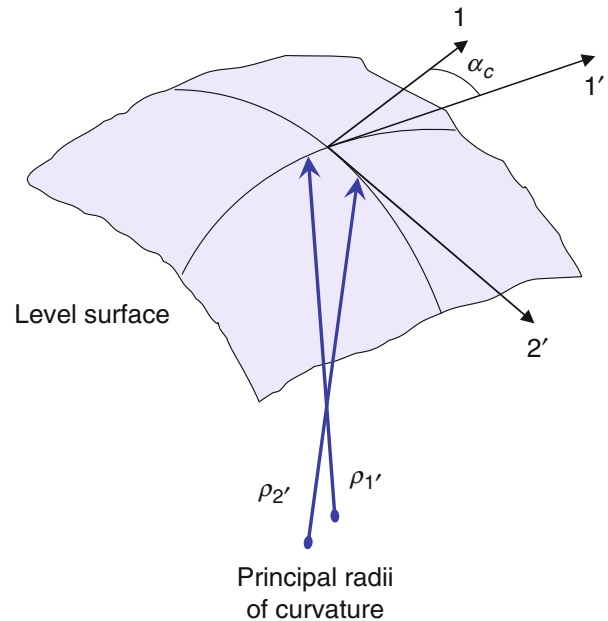
$$\Gamma_H = \sqrt{\Gamma_{1,3}^2 + \Gamma_{2,3}^2}, \quad \alpha_H = \tan^{-1} \frac{\Gamma_{2,3}}{\Gamma_{1,3}}. \quad (9)$$

Gravitational gradients also describe the curvature of a surface of constant gravitational potential (a level surface) (Fig. 1). Let $\rho_{1'}$ and $\rho_{2'}$ be the two principal radii of curvature at a point on the surface, whose reciprocals are, respectively, the minimum and maximum curvatures along corresponding (orthogonal) axes, $x_{1'}$ and $x_{2'}$. It can be shown (Heiland 1940; Nettleton 1976) that the differential curvature, $\Delta\kappa = g_3(1/\rho_{2'} - 1/\rho_{1'})$, is given by

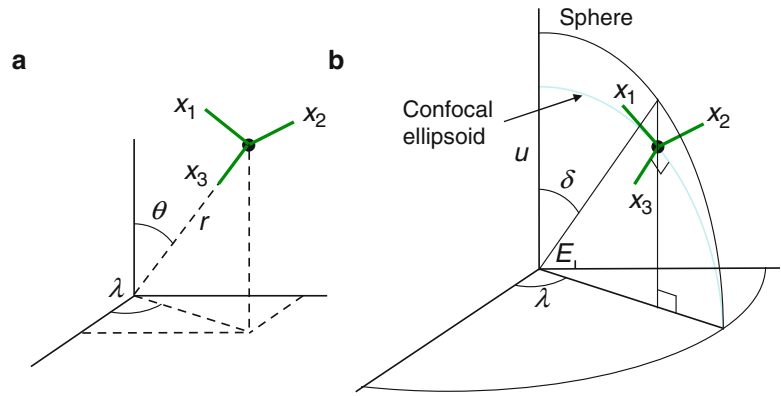
$$\Delta\kappa = \sqrt{(2\Gamma_{1,2})^2 + (\Gamma_{2,2} - \Gamma_{1,1})^2}. \quad (10)$$

The direction of minimum curvature is

$$\alpha_c = \frac{1}{2} \tan^{-1} \frac{-2\Gamma_{1,2}}{\Gamma_{2,2} - \Gamma_{1,1}}. \quad (11)$$



Gravity, Gradiometry, Fig. 1 Principal radii of curvature at a point on an equipotential surface

Gravity, Gradiometry,**Fig. 2** Local Cartesian coordinates relative to (a) spherical polar coordinates and (b) ellipsoidal (spheroidal) coordinates

If $a_c = 0$, then the local coordinate axes, x_1 and x_2 , coincide with the directions of minimum and maximum curvature; and $\Gamma_{1,2} = 0$. In general, we also have

$$\begin{aligned}\Gamma_{2,2} - \Gamma_{1,1} &= g_3(1/\rho_{2'} - 1/\rho_{1'}) \cos 2\alpha_c, \\ \Gamma_{1,2} &= -\frac{1}{2}g_3(1/\rho_{2'} - 1/\rho_{1'}) \sin 2\alpha_c.\end{aligned}\quad (12)$$

Other Coordinate Systems

For a local Cartesian coordinate system embedded in a spherical polar coordinate system as shown in Fig. 2a, the first- and second-order derivatives are related to corresponding derivatives with respect to the spherical coordinates, as follows,

$$\frac{\partial}{\partial x_1} = -\frac{1}{r} \frac{\partial}{\partial \theta}, \quad \frac{\partial}{\partial x_2} = \frac{1}{r \sin \theta} \frac{\partial}{\partial \lambda}, \quad \frac{\partial}{\partial x_3} = -\frac{\partial}{\partial r}; \quad (13)$$

$$\begin{aligned}\frac{\partial^2}{\partial x_1^2} &= \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2} \\ \frac{\partial^2}{\partial x_1 \partial x_2} &= \frac{\cot \theta}{r^2 \sin \theta} \frac{\partial}{\partial \lambda} - \frac{1}{r^2 \sin \theta} \frac{\partial^2}{\partial \theta \partial \lambda} \\ \frac{\partial^2}{\partial x_1 \partial x_3} &= -\frac{1}{r^2} \frac{\partial}{\partial \theta} + \frac{1}{r} \frac{\partial^2}{\partial \theta \partial r} \\ \frac{\partial^2}{\partial x_2^2} &= \frac{\cot \theta}{r^2} \frac{\partial}{\partial \theta} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \lambda^2} \\ \frac{\partial^2}{\partial x_2 \partial x_3} &= \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \lambda} - \frac{1}{r \sin \theta} \frac{\partial^2}{\partial \lambda \partial r} \\ \frac{\partial^2}{\partial x_3^2} &= \frac{\partial^2}{\partial r^2}\end{aligned}\quad (14)$$

with corresponding Laplacian operator,

$$\begin{aligned}\nabla^2 &= \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} \\ &= \frac{\partial^2}{\partial r^2} + \frac{2}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial}{\partial \theta} + \frac{1}{r^2 \sin^2 \theta} \\ &\quad \times \frac{\partial^2}{\partial \lambda^2}.\end{aligned}\quad (15)$$

Usually the Earth is approximated by an oblate rotational ellipsoid (spheroid) generated by rotating an ellipse with linear eccentricity, E , about its semi-minor axis, identified as the polar axis. Corresponding ellipsoidal (spheroidal) coordinates, (u, δ, λ) , are shown in Fig. 2b, where surfaces of constant u are confocal ellipsoids and surfaces of constant δ are hyperboloids of revolution of one sheet. As for spherical coordinates, surfaces of constant longitude, λ , are half-planes containing the polar axis. The local first- and second-order derivatives with respect to Cartesian coordinates (where the local x_3 -axis is orthogonal to the confocal ellipsoid that passes through the origin of the local system) are related to the corresponding derivatives with respect to the ellipsoidal coordinates, as follows,

$$\begin{aligned}\frac{\partial}{\partial x_1} &= -\frac{1}{w} \frac{\partial}{\partial \delta}, \quad \frac{\partial}{\partial x_2} = \frac{1}{v \sin \delta} \frac{\partial}{\partial \lambda}, \quad \frac{\partial}{\partial x_3} \\ &= -\frac{w}{v} \frac{\partial}{\partial u};\end{aligned}\quad (16)$$

$$\begin{aligned}\frac{\partial^2}{\partial x_1^2} &= \frac{1}{w^2} \left(\frac{uv^2}{w^2} \frac{\partial}{\partial u} + \frac{E^2 \sin 2\delta}{2w^2} \frac{\partial}{\partial \delta} + \frac{\partial^2}{\partial \delta^2} \right) \\ \frac{\partial^2}{\partial x_1 \partial x_2} &= \frac{1}{wv \sin \delta} \left(\cot \delta \frac{\partial}{\partial \lambda} - \frac{\partial^2}{\partial \delta \partial \lambda} \right) \\ \frac{\partial^2}{\partial x_1 \partial x_3} &= \frac{v}{w^2} \left(-\frac{u}{w^2} \frac{\partial}{\partial \delta} + \frac{E^2 \sin 2\delta}{2w^2} \frac{\partial}{\partial u} + \frac{\partial^2}{\partial \delta \partial r} \right) \\ \frac{\partial^2}{\partial x_2^2} &= \frac{\cot \delta}{w^2} \frac{\partial}{\partial \delta} + \frac{u}{w^2} \frac{\partial}{\partial u} + \frac{1}{v^2 \sin^2 \delta} \frac{\partial^2}{\partial \lambda^2} \\ \frac{\partial^2}{\partial x_2 \partial x_3} &= \frac{1}{w \sin \delta} \left(\frac{u}{v^2} \frac{\partial}{\partial \lambda} - \frac{\partial^2}{\partial \lambda \partial u} \right) \\ \frac{\partial^2}{\partial x_3^2} &= \frac{1}{w^2} \left(-\frac{E^2 \sin 2\delta}{2w^2} \frac{\partial}{\partial \delta} - \frac{uE^2 \sin^2 \delta}{w^2} \frac{\partial}{\partial u} + v^2 \frac{\partial^2}{\partial u^2} \right)\end{aligned}\quad (17)$$

where

$$v^2 = u^2 + E^2, \quad w^2 = u^2 + E^2 \cos^2 \delta. \quad (18)$$

The Laplacian operator is

$$\nabla^2 = \frac{1}{w^2} \times \left(v^2 \frac{\partial^2}{\partial u^2} + 2u \frac{\partial}{\partial u} + \frac{\partial^2}{\partial \delta^2} + \cot \delta \frac{\partial}{\partial \delta} + \frac{w^2}{v^2 \sin^2 \delta} \frac{\partial^2}{\partial \lambda^2} \right). \quad (19)$$

Invariants

Being a second-order tensor, the gradients change under a rotation of Cartesian coordinates from system s to system t according to $\mathbf{\Gamma}^t = \mathbf{C}_s^t \mathbf{\Gamma}^s \mathbf{C}_t^s$, where $\mathbf{C}_s^t$ is the matrix of appropriate direction cosines. For any arbitrary Cartesian system, s , there exists a rotation that diagonalizes the tensor, $\mathbf{\Gamma}^s$,

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \mathbf{\Gamma}^t = \mathbf{C}_s^t \mathbf{\Gamma}^s \mathbf{C}_t^s. \quad (20)$$

The diagonal elements, λ_j , are the eigenvalues of $\mathbf{\Gamma}^s$, or any rotation thereof and are called invariants of the tensor with respect to coordinate rotations. They are real numbers because the tensor is real and symmetric. All combinations of eigenvalues are also invariants.

Given the tensor, $\mathbf{\Gamma}$, in some coordinate system, the eigenvalues are roots to the following characteristic equation ($\det$ = determinant),

$$\begin{aligned} \det(\lambda \mathbf{I} - \mathbf{\Gamma}) &= (\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3) \\ &= \lambda^3 - I_0 \lambda^2 + I_1 \lambda - I_2 = 0, \end{aligned} \quad (21)$$

from which it is easy to show that, with the gradient elements as designated in Eq. (5),

$$I_0 = \lambda_1 + \lambda_2 + \lambda_3 = \Gamma_{1,1} + \Gamma_{2,2} + \Gamma_{3,3}, \quad (22)$$

$$\begin{aligned} I_1 &= \lambda_1 \lambda_2 + \lambda_2 \lambda_3 + \lambda_3 \lambda_1 \\ &= \Gamma_{1,1} \Gamma_{2,2} + \Gamma_{2,2} \Gamma_{3,3} + \Gamma_{1,1} \Gamma_{3,3} - \Gamma_{1,2}^2 - \Gamma_{2,3}^2 \\ &\quad - \Gamma_{3,1}^2, \end{aligned} \quad (23)$$

$$I_2 = \lambda_1 \lambda_2 \lambda_3 = \det \mathbf{\Gamma}; \quad (24)$$

and, therefore, these quantities are invariants with respect to rotation. For example, Poisson's Eq. (6) implies $I_0 = -4\pi G\rho$.

From Eqs. (22) and (23), we have $I_0^2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 + 2I_1$. If $I_0 = 0$ (e.g., above the Earth's surface, neglecting the atmosphere), then $I_1 < 0$. Pedersen and Rasmussen (1990) show that

$$0 \leq I \equiv -\left(\frac{I_2}{2}\right)^2 / \left(\frac{I_1}{3}\right)^3 \leq 1; \quad (25)$$

and call I an indicator of two-dimensionality. $I = 1$ implies that the source of the gradients is like a monopole and $I = 0$ implies that the generated field is invariant in some direction, which means that the source is two-dimensional and can be described by a cross section.

Finally, we note that both Γ_H (Eq. 9) and $\Delta\kappa$ (Eq. 10) are horizontal invariants, that is, independent of the azimuthal orientation of the local frame.

Models and Modeling

As is quantified later in this section, the gravitational gradients, being second-order derivatives of the potential, reflect primarily the short-wavelength, local features of the Earth's crust. For this reason the forward modeling of gradients due to sources has largely concentrated on formulating the gradient signals of polyhedra of constant density, which by suitable assemblage can simulate the local crust, specifically the local topography. Given a regular distribution of elevation data for a particular region, the topography may be modeled as a collection of right rectangular prisms whose vertical dimension is defined by the elevation of a data point and whose base has dimensions equal to the horizontal resolution of the data. Other options to model the topography include triangular prisms (using a triangulation of the elevation data) and any of several numerical integration methods applied to (from Eq. 2)

$$\Gamma_{j,k}(\mathbf{x}) = G \int_V \rho(\mathbf{x}') \frac{\partial^2}{\partial x_j \partial x_k} \frac{1}{|\mathbf{x} - \mathbf{x}'|} dv. \quad (26)$$

The purposes of modeling gravitational gradients range from validation of gradiometer accuracy to introducing constraints on the geophysical inverse problem. As new gradiometer systems are developed, particularly airborne systems, determining their accuracy is aided by flying over regions with well-known topographic elevations and density contrasts. Comparing the measurements to gradients, forward-modeled from the topography, calibrates the noise level and enables tuning of appropriate data filters. Conversely, detecting subsurface geologic features, such as kimberlites (for diamond prospecting), from gradiometric surveys is facilitated by knowing what to look for in terms of a particular gradient signature. Thus, even in the inversion of gradiometric data to infer sources, forward modeling of structures plays an important role.

The gradients generated by constant-density, right-rectangular prisms follow from Eq. (26). If the prism has dimensions bounded by Cartesian coordinates, $a_1 \leq x'_1 \leq a_2$, $b_1 \leq x'_2 \leq b_2$, $c_1 \leq x'_3 \leq c_2$, the gradients for points outside the prism are given by

$$\Gamma_{jk}(\mathbf{x}) = -G\rho \ln(x_\ell - x'_\ell + r) \Big|_{x'_1=a_1}^{a_2} \Big|_{x'_2=b_1}^{b_2} \Big|_{x'_3=c_1}^{c_2}, \quad (27)$$

$$\Gamma_{jj}(\mathbf{x}) = G\rho \tan^{-1} \frac{(x_k - x'_k)(x_\ell - x'_\ell)}{(x_j - x'_j)r} \Big|_{x'_1=a_1}^{a_2} \Big|_{x'_2=b_1}^{b_2} \Big|_{x'_3=c_1}^{c_2}, \quad (28)$$

where $r^2 = (x_1 - x'_1)^2 + (x_2 - x'_2)^2 + (x_3 - x'_3)^2$ and where (j, k, ℓ) is a cyclic permutation of $(1, 2, 3)$. Using these formulas to model the gradients associated with the local topography is analogous to a numerical integration of Eq. (26) using the rectangular rule.

To take advantage of various other numerical integration techniques based on a discretization of the topographic surface, $h(x'_1, x'_2)$, over an area, A , one first integrates Eq. (26) with respect to the vertical, x'_3 ,

$$\begin{aligned} \Gamma_{jk}(\mathbf{x}) &= G\rho \iint_A T_{jk}(\mathbf{x}; x'_1, x'_2) dA \\ &\approx G\rho \sum_{i_1} \sum_{i_2} T_{jk}(\mathbf{x}; (x_1)_{i_1}, (x_2)_{i_2}) \Delta A_{i_1, i_2}, \end{aligned} \quad (29)$$

where the functions,

$$T_{jk}(\mathbf{x}; x'_1, x'_2) = \int_0^{h(x'_1, x'_2)} \frac{\partial^2}{\partial x_j \partial x_k} \left(\frac{1}{r} \right) dx'_3, \quad (30)$$

have well-defined analytic expressions (Jekeli and Zhu 2006). The discretization indicated by Eq. (29) then may take any number of forms.

For example, it may also be desirable to express $T_{j,k}$ as a Taylor series in h , in which case Eq. (29) becomes a series of two-dimensional convolutions on the plane,

$$\Gamma_{jk}(\mathbf{x}) = G\rho \iint_A \left(\sum_n \frac{1}{n!} \frac{\partial^n}{\partial h^n} T_{jk}|_{h=0} h^n \right) dx'_1 dx'_2, \quad (31)$$

where it is noted that the derivatives of $T_{j,k}$ are functions of $x_1 - x'_1$ and $x_2 - x'_2$ (besides x_3 and h). Discretized, such convolutions can be evaluated rapidly using fast algorithms for the discrete Fourier transform (Forsberg 1985), provided the variable, x_3 , remains constant (which is a limitation of the method). Similar Fourier transform methods have also been developed by (Parker 1972), but found to be less suitable (Jekeli and Zhu 2006) due to the use of the *analytic* Fourier transforms of the kernel functions, $\partial^n T_{j,k} / \partial h^n$. While the

linear term in the Taylor series may suffice in some cases, high-order expansions are required when evaluating the gradients near Earth's surface. Nonlinear effects are significant even for the problem of inverting gradients on the ocean surface to infer seafloor topography at depths of several kilometers (Yang et al. 2018).

Other fast methods transform the volume integral (Eq. 26) to surface and line integrals along the theory of Gauss (divergence theorem, etc.); Petrovic (1996) and Tsoulis and Petrovic (2001) have developed corresponding formulas for arbitrary polyhedra. In all these cases, the underlying coordinate system is Cartesian. Forward modeling of gravitational gradients in spherical coordinates that account for the convergence of coordinate lines of a data grid, as well as the convergence of radial (height) lines, has not received much attention (however, see Smith et al. 2001), since the gradients are influenced primarily by very local topography.

Spectral models of the gravitational gradients can be derived directly from corresponding models of the gravitational potential in local regions, as well as for the entire Earth; such global models exist in abundance as compiled, for example, by the International Center for Global Earth Models. The latter global models invariably represent the gravitational potential, V , in a series of solid spherical harmonic functions, $(R/r)^{n+1} \bar{Y}_{n,m}(\theta, \lambda)$, whose projections onto a sphere of radius, $r = R$, comprise the basis functions for the Fourier-Legendre spectrum, $\{(GM/R)C_{n,m}\}$, of the potential on that sphere. Extended into free space, $r \geq R$, the potential satisfies Laplace's Eq. (8) and is thus modeled according to

$$V(r, \theta, \lambda) = \frac{GM}{R} \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(\frac{R}{r} \right)^{n+1} C_{n,m} \bar{Y}_{n,m}(\theta, \lambda), \quad (32)$$

assuming no masses exist outside the sphere of radius, R , and total mass, M . The (surface) spherical harmonic functions are defined by

$$\bar{Y}_{n,m}(\theta, \lambda) = \bar{P}_{n,|m|}(\cos \theta) \begin{cases} \cos m\lambda, & m \geq 0 \\ \sin |m|\lambda, & m < 0 \end{cases} \quad (33)$$

where the functions, $\bar{P}_{n,m}$, are associated Legendre functions of the first kind, fully normalized so that

$$\begin{aligned} &\frac{1}{4\pi} \iint_{\Omega} \bar{Y}_{n',m'}(\theta, \lambda) \bar{Y}_{n,m}(\theta, \lambda) d\Omega \\ &= \begin{cases} 1, & n' = n \text{ and } m' = m \\ 0, & n' \neq n \text{ or } m' \neq m \end{cases} \end{aligned} \quad (34)$$

and where Ω represents the unit sphere; see also (Jekeli 2017) for further details.

Practically, a spherical harmonic model is given with a limited spectrum up to maximum degree, $n_{\max}$. The gradients, $\Gamma_{j,k}$, corresponding to this model in a local coordinate system, as shown in Fig. 2a, can then be determined using Eq. (14). However, the spectrum with respect to the surface spherical harmonics does not exist for all the gradients. Specifically, any of the derivatives in latitude destroys the orthogonality of the basis (eigen-) functions on the sphere. It is easy to see, however, that the second radial derivative possesses a Fourier-Legendre spectrum.

Similarly, for a given ellipsoidal harmonic model with coefficients, $C_{n,m}^e$ (Jekeli 1988a),

$$V(u, \delta, \lambda) = \frac{GM}{R} \sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{\bar{Q}_{n,|m|}(iu/E)}{\bar{Q}_{n,|m|}(ib/E)} C_{n,m}^e \bar{Y}_{n,m}(\delta, \lambda), \quad (35)$$

where $\bar{Q}_{n,m}$ is the fully normalized associated Legendre function of the second kind, the gradients in a local coordinate system are obtained using the derivatives in Eq. (17). Again, derivatives of the potential with respect to the ellipsoidal coordinate, δ , do not have a Fourier-Legendre spectrum with respect to the eigenfunctions, $\bar{Y}_{n,m}(\delta, \lambda)$. Although the second derivative with respect to the ellipsoidal coordinate, u , could be interpreted to have a Fourier-Legendre spectrum on a confocal ellipsoid with respect to the functions, $\bar{Y}_{n,m}(\delta, \lambda)$, it is not directly related to the local gradient, $\partial^2 V / \partial x_3^2$.

A representation of the potential and its derivatives using locally supported basis functions on the sphere has found particular application for gravitational gradients due to their inherently local spectral content (Eicker et al. 2014). In particular, the geopotential models constructed with radial basis functions (RBFs) (Freeden et al. 1998; Schmidt et al. 2007) have the form,

$$V(r, \theta, \lambda) = \sum_{k=1}^N c_k B_k(r, \theta, \lambda), \quad (36)$$

where the k^{th} RBF is

$$B_k(r, \theta, \lambda) = \sum_{n=0}^{n'} (2n+1) \left(\frac{R}{r}\right)^{n+1} B_n P_n(\cos \psi_k), \quad (37)$$

with spectral tuning coefficients, B_n , and central angle, ψ_k (the “radial” distance), from the central node, (θ_k, λ_k) , of its “local” support,

$$\cos \psi_k = \cos \theta \cos \theta_k + \sin \theta \sin \theta_k \cos(\lambda - \lambda_k). \quad (38)$$

In general, these RBFs still have global support, but with appropriate definition of their spectrum, they can be defined to

vanish outside local regions (Schreiner 1997). This leads then to spherical spline and wavelet representations of the gravitational potential on and external to the sphere. In each case the local gravitational gradients (i.e., in a local Cartesian coordinate system) are determined according to (14).

Usually, a local Cartesian coordinate system suffices to model the Fourier spectrum of the gradients. In this case, the spectra of all gradients may be derived from the spectrum of the potential. Starting with the representation of the potential in terms of its Fourier spectrum, $v(f_1, f_2)$, on a plane ($x_3 = 0$) and including the harmonic extension into free space ($-x_3 > 0$ in the north-east-down system), we have

$$V(x_1, x_2, x_3) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} v(f_1, f_2) e^{-2\pi f |x_3|} e^{i2\pi(f_1 x_1 + f_2 x_2)} df_1 df_2, \quad (39)$$

where $f = \sqrt{f_1^2 + f_2^2}$ and f_1, f_2 are spatial, cyclical frequencies. The Fourier spectrum, $\mathcal{G}$, of each gravitational gradient on the $|x_3|$ -plane is obtained by applying the appropriate derivative,

$$\begin{aligned} \mathcal{G}_{1,1}(f_1, f_2; x_3) &= -4\pi^2 f_1^2 v(f_1, f_2) e^{-2\pi f |x_3|} \\ \mathcal{G}_{1,2}(f_1, f_2; x_3) &= -4\pi^2 f_1 f_2 v(f_1, f_2) e^{-2\pi f |x_3|} \\ \mathcal{G}_{1,3}(f_1, f_2; x_3) &= -i4\pi^2 f_1 v(f_1, f_2) e^{-2\pi f |x_3|} \\ \mathcal{G}_{2,2}(f_1, f_2; x_3) &= -4\pi^2 f_2^2 v(f_1, f_2) e^{-2\pi f |x_3|} \\ \mathcal{G}_{2,3}(f_1, f_2; x_3) &= -i4\pi^2 f_2 v(f_1, f_2) e^{-2\pi f |x_3|} \\ \mathcal{G}_{3,3}(f_1, f_2; x_3) &= 4\pi^2 f^2 v(f_1, f_2) e^{-2\pi f |x_3|}. \end{aligned} \quad (40)$$

These relationships can be used to determine any gradient from any other gradient or from any derivative of the potential. For example, the gradient, Γ_{13} , at altitude, x_3 , is the inverse Fourier transform of $f \times 2 \times \pi \times f_1 \times \gamma(f_1, f_2) e^{-2\pi f |x_3|}$, where $\gamma(f_1, f_2) = -2\pi f v(f_1, f_2)$ is the Fourier transform of the vertical derivative of V (the gravitational acceleration) at ground level. In such transformations, due care must be exercised when calculating the higher frequencies of the gradients if the original data have noise and are limited in resolution due to finite data spacing.

Reference Field and Power Spectral Density

The first-order approximation of the Earth's gravity potential is the “normal potential,” which includes the centrifugal component due to Earth's rotation and whose defining property is that it is constant on a co-rotating ellipsoidal model for the Earth. It may be expressed by the spherical harmonic expansion,

$$U(r, \theta, \lambda) = \frac{GM}{r} \left(1 - \sum_{n=1}^{\infty} J_{2n} \left(\frac{a}{r} \right)^{2n} P_{2n}(\cos \theta) \right) + \frac{1}{2} \omega_e^2 r^2 \sin^2 \theta, \quad (41)$$

where a is the semimajor axis of the ellipsoid, ω_e is Earth's rotation rate, and the coefficients are given by

$$J_{2n} = (-1)^{n+1} \frac{3e^{2n}}{(2n+1)(2n+3)} \left(1 - n + 5n \frac{J_2}{e^2} \right), \quad (42)$$

where e is the first eccentricity of the ellipsoid, and J_2 is the dynamical form factor related to the difference in principal moments of inertia of the Earth, or also the polar flattening of the ellipsoid (Heiskanen and Moritz 1967; Hofmann-Wellenhof and Moritz 2005). The series (41) is usually truncated at $n = 5$ with entirely sufficient accuracy. For values of the defining parameters, GM , a , J_2 , ω_e , of the current and past reference fields, see chapter ► “Gravity Field of the Earth.”

The normal gravity potential, U , does not depend on longitude due to the rotational symmetry of the ellipsoid; and, in view of Eq. (14), only the local gravity gradients, $\Gamma_{1,1}$, $\Gamma_{1,3}$, $\Gamma_{2,2}$, and $\Gamma_{3,3}$, include a normal component. A standard approximation for the normal gradient $\Gamma_{3,3}^{(U)}$ is $\Gamma_{3,3}^{(U)} \approx 3086 \times 10^{-9} \text{ s}^{-2}$, based on just the first term in Eq. (41). The common unit of measure for gravity gradients is the Eötvös ($1 \text{ E} = 10^{-9} \text{ s}^{-2}$), in honor of the Hungarian geophysicist, Loránd Eötvös, who invented the torsion balance (section ► “Gravity, Gradiometry”).

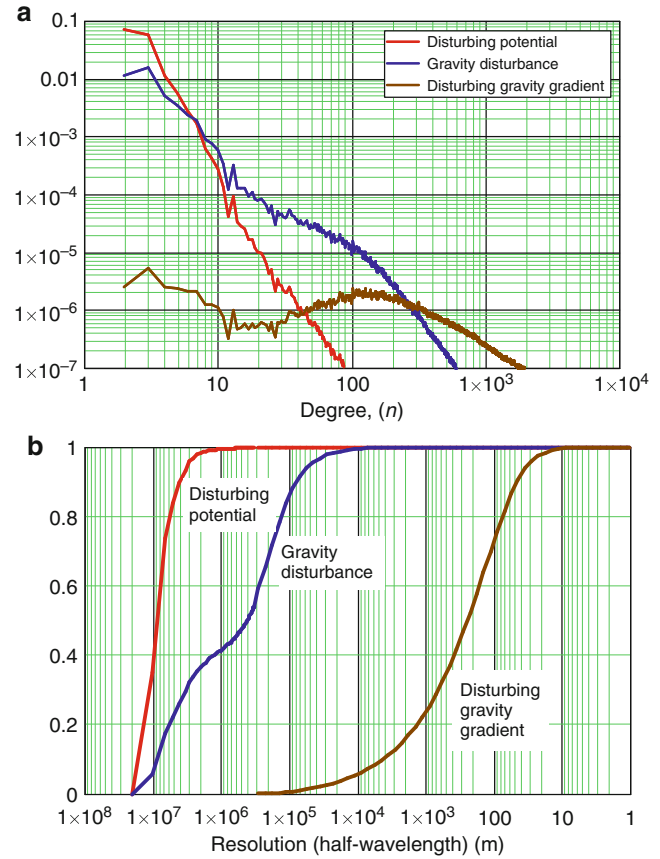
The residual potential, actual minus normal potentials, also called the disturbing potential, then gives rise to gradient disturbances, which may in some instances attain an order of magnitude similar to the reference part, especially in very mountainous areas (e.g., Hein 1977).

Globally, the power spectral density (PSD) of the potential, defined on the sphere of radius, R , is given in terms of its Fourier-Legendre spectrum by (Jekeli 2017)

$$\Phi_n = \frac{1}{2n+1} \left(\frac{GM}{R} \right)^2 \sum_{m=-n}^n C_{n,m}^2. \quad (43)$$

Corresponding radial derivatives have PSDs scaled by fixed powers of the degree, meaning that they emphasize the higher degrees or shorter wavelengths of the field. Figure 3a shows the PSDs of the disturbing potential and its first two radial derivatives (relative to the normal field) according to the Earth Gravitational Model 2008 (EGM2008, Pavlis et al. 2012).

Also shown (Fig. 3b) is the cumulative contribution of the disturbing field and its radial derivatives up to a given degree, n , or resolution, where spatial resolution (half-wavelength) in kilometers is given by $\Delta s = 6371\pi/n$, 6371 km being the mean Earth radius. While about 99.3% of the disturbing

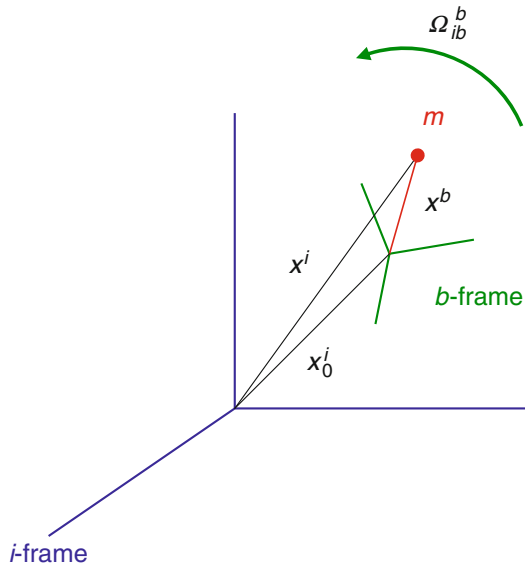


Gravity, Gradiometry, Fig. 3 (a) Power spectral densities of radial derivatives of the Earth's disturbing potential normalized by the respective total variance. (b) Cumulative relative power of each radial derivative as function of spatial resolution of the field

potential is defined by the field with wavelengths longer than 2000 km, the gravity disturbance (first radial derivative) comprises only about 41.4% up to these wavelengths, and the gradient disturbance barely has any (relative) power at wavelengths longer than 200 km. Here, it is noted that the cumulative contribution of the gradient disturbance is highly dependent on the local field and is only an illustrative example. As another example, Jordan (1978) provided a PSD model wherein 99.995% of the power of the gradient disturbance is due to the density variations in the upper 150 m of the crust, while this regime contributes 89% to the gravity disturbance and only 7% to the disturbing potential.

Measurement Error Analysis

All currently operational gravity gradiometers measure “gradients” by differencing, in one form or another, the inertial accelerations sensed by linear or angular accelerometers. In the limit of infinitesimal separation, the model with linear accelerometers can be used to derive the theoretical equation for gravity gradiometers, which also offers a general interpretation of measured gradients on a moving platform. We start



Gravity, Gradiometry, Fig. 4 The geometry for the lever-arm effect on the acceleration of m in the b -frame rotating with respect to the i -frame

with the lever-arm equation (Jekeli 2000) for the output of an accelerometer (Fig. 4),

$$\mathbf{a}^b = \mathbf{a}_0^b + (\mathbf{g}_0^b - \mathbf{g}^b) + \dot{\Omega}_{ib}^b \mathbf{x}^b + \Omega_{ib}^b \Omega_{ib}^b \mathbf{x}^b, \quad (44)$$

where $\mathbf{a}^b$ is the inertial acceleration of a test mass, m , in a frame (the b -frame) that rotates with respect to the inertial frame (i -frame) with angular rate given by the vector, $\boldsymbol{\omega}_{ib}^b = (\omega_1 \ \omega_2 \ \omega_3)^T$, expressed in terms of a cross product as

$$[\boldsymbol{\omega}_{ib}^b \times] = \Omega_{ib}^b = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}. \quad (45)$$

The acceleration of the b -frame in the i -frame is $\mathbf{a}_0^i$, and the location of the test mass (the lever-arm) in the b -frame is $\mathbf{x}^b$. The gravitational acceleration of the b -frame (in b -frame coordinates) is $\mathbf{g}_0^b$ and that of the test mass is $\mathbf{g}^b$. The last two terms of Eq. (44) represent, respectively, Coriolis and centrifugal accelerations due to the frame rotation.

The spatial derivative of Eq. (44), analogous to differencing the inertial accelerations of two test masses in the b -frame, yields

$$\frac{\partial \mathbf{a}^b}{\partial \mathbf{x}^b} = -\frac{\partial \mathbf{g}^b}{\partial \mathbf{x}^b} + \dot{\Omega}_{ib}^b + \Omega_{ib}^b \Omega_{ib}^b, \quad (46)$$

showing that the measured acceleration gradient is the combination of the gravitational gradient and angular accelerations. Thus, a gravity gradiometer must either be stabilized against rotations with respect to the inertial frame or some

Gravity, Gradiometry, Table 1 Roughly commensurate (per row) sensor errors in mobile gravity gradiometry. Emphasized entries represent typical calibrated error levels for high-accuracy components

Gradiometer white noise	Orientation error	Gyro bias	Gyro white noise
30 E/ $\sqrt{\text{Hz}}$	20°	0.5°/h	3°/h/ $\sqrt{\text{Hz}}$
10 E/ $\sqrt{\text{Hz}}$	6°	0.15°/h	1°/h/ $\sqrt{\text{Hz}}$
1 E/ $\sqrt{\text{Hz}}$	0.6°	0.015°/h	0.1°/h/ $\sqrt{\text{Hz}}$
0.1 E/ $\sqrt{\text{Hz}}$	0.06°	0.0015°/h	0.01°/h/ $\sqrt{\text{Hz}}$
0.01 E/ $\sqrt{\text{Hz}}$	0.006°	0.00015°/h	0.001°/h/ $\sqrt{\text{Hz}}$

other realizable frame, or these rotations must be measured independently with gyroscopes. An analysis of the required accuracy of these measurements (or, equivalently, of the stabilization) was conducted by Jekeli (2006) and is summarized in Table 1. Each row of this table corresponds approximately to commensurate sensor errors (gradiometer and gyroscope). For typical airborne gradiometer systems with noise levels in the range 1 – 10 E/ $\sqrt{\text{Hz}}$, conventional high-accuracy gyroscopes are sufficient to separate the gravitational gradients from the non-gravitational components. It is also evident that the sensitivity to position error (registration error) is not strong in view of the relatively small spatial gradient of the gravitational gradients, of the order of 1.45×10^{-3} E/m for the main field.

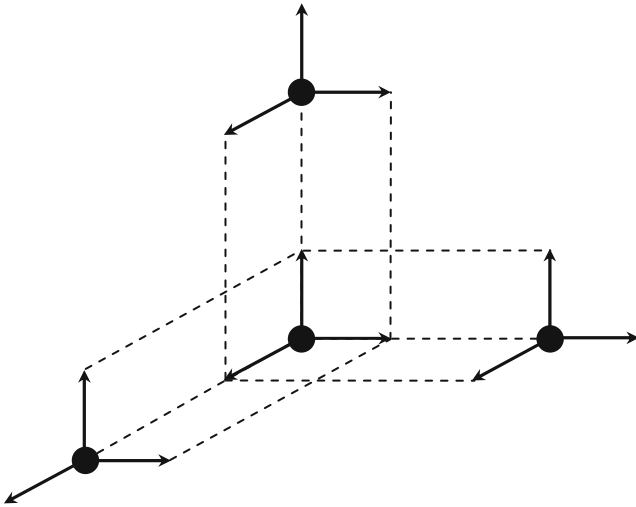
Due to the symmetry of the gravitational tensor, $\Gamma^b = (\partial \mathbf{g}^b / \partial \mathbf{x}^b)$, and the anti-symmetry of Ω_{ib}^b , the measurement of the full matrix, $(\partial \mathbf{a}_m^b / \partial \mathbf{x}^b)$, permits some simplifications, such as

$$\Gamma^b = -\frac{1}{2} \left(\frac{\partial \mathbf{a}^b}{\partial \mathbf{x}^b} + \left(\frac{\partial \mathbf{a}^b}{\partial \mathbf{x}^b} \right)^T \right) + \Omega_{ib}^b \Omega_{ib}^b. \quad (47)$$

The next section offers further details on the measurement of gravity gradients.

Gradiometry

Whereas the gravity meter, or any other accelerometer, senses the inertial reaction of a single test mass to an applied force, the (gravitational) gradiometer senses the difference of reactions from two test masses. In general terms, one may think of a gradiometer as the juxtaposition of two accelerometers and the sensed gradient as the ratio of the acceleration difference to the baseline length between them. As such the gradiometer measures only an approximation of the gradient at a point, which is adequate for sufficiently small baselines. Measuring the entire tensor of gradients (a “full-tensor gradiometer”) requires a minimum of 12 accelerometers as shown in Fig. 5, although, because some gradients share an accelerometer in this configuration, the errors between them may be



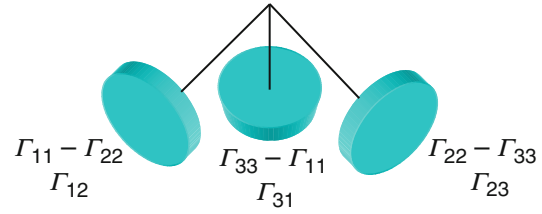
Gravity, Gradiometry, Fig. 5 Conceptual minimal configuration for a full-tensor gradiometer

correlated. Completely independent measurements of all nine tensor elements, therefore, require 18 accelerometers. On the other hand, taking advantage of the tensor symmetry and Laplace's field Eq. (8) brings the minimum to five accelerometer pairs.

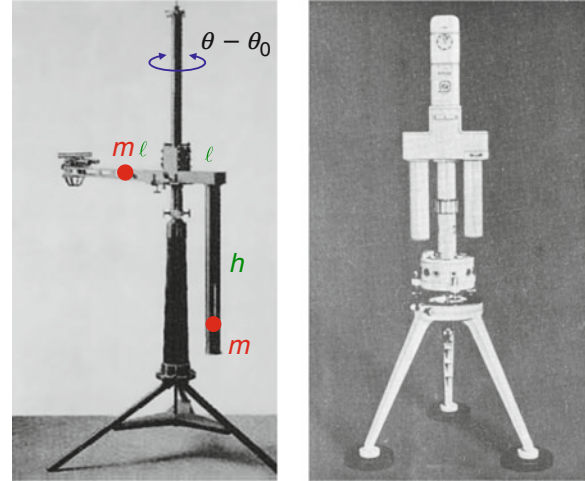
Most gradiometers actually are designed to sense components of the differential curvature of the field (Eq. 12), and thus they rely on Laplace's equation to determine the individual diagonal components. Specifically, imagine that a gradiometer senses $\Gamma_{2,2} - \Gamma_{1,1}$ and $\Gamma_{1,2}$ in the instrument frame. Combining three such devices with mutually orthogonal orientation within the instrument system (Fig. 6) thus yields in addition to $\Gamma_{1,2}$, $\Gamma_{2,3}$, $\Gamma_{3,1}$ the components $\Gamma_{2,2} - \Gamma_{1,1}$, $\Gamma_{3,3} - \Gamma_{2,2}$, and $\Gamma_{1,1} - \Gamma_{3,3}$. The latter always sum to zero (except for measurement errors) and cannot by themselves separate the diagonal elements of the tensor. Nevertheless, it is common to refer to this as a full-tensor gradiometer (FTG), even though effectively only five of the nine tensor elements are sensed independently.

The first instrument specifically designed to measure the gravity gradient was the torsion balance, invented by the Hungarian geophysicist, Loránd Eötvös (1848–1919), at the end of the nineteenth century (One makes the distinction in geodesy between “gravitation” and “gravity,” where the latter refers to measurements on the Earth's surface and is the vector sum of acceleration due to mass attraction (gravitation) and the centrifugal acceleration associated with Earth's rotation). Prolific in many endeavors of exploration geophysics, one of his passionate goals was to design an instrument that could sense all gravitational gradients (he left the vertical-vertical gradient to others, since it would not be possible for his apparatus) (Szabó 1998).

Eötvös adapted the concept of Coulomb's torsion balance (designed to sense electrostatic forces) to measure gravitational



Gravity, Gradiometry, Fig. 6 Mutually orthogonal gradiometer sensors in the “umbrella configuration”



Gravity, Gradiometry, Fig. 7 Eötvös's original torsion balance (left, from Szabó 1998, reprinted with permission, Eötvös Loránd Geophysical Institute of Hungary) and the more modern version manufactured by the Eötvös Loránd Geophysical Institute of Hungary, Budapest in the 1950s (right, from Badekas 1967)

gradients. After several initial prototype instruments, including the curvature variometer, which was sensitive to the differential curvature components (Eq. 12), and the horizontal variometer, which measured the horizontal gradients (Γ_{13} , Γ_{23}), the final product combined these into one instrument, shown in Fig. 7. Two equal masses, m , are attached to the ends of a beam of length, 2ℓ , which is suspended by a platinum-iridium (or, tungsten) wire that may twist about the vertical axis in response to the curvature of the gravitational field. If both masses were at the same height ($h = 0$, in Fig. 7), the instrument would be the curvature variometer. With the vertical offset of one mass, the instrument is also sensitive to the horizontal gradients.

Assuming the beam has negligible mass, the torque on the beam is

$$\mathbf{L} = \mathbf{r}_A \times m\mathbf{g}_A + \mathbf{r}_B \times m\mathbf{g}_B, \quad (48)$$

where $\mathbf{r}_A = (-x, -y, 0)$ and $\mathbf{r}_B = (x, y, h)$ are the position vectors of the masses in a north-east-down coordinate system with origin at the instrument center, and $\mathbf{g}_A$, $\mathbf{g}_B$ are the gravity

vectors at these points. The torque component about the vertical is

$$L_3 = mx(g_2(x, y, h) - g_2(-x, -y, 0)) + my(g_1(-x, -y, 0) - g_1(x, y, h)). \quad (49)$$

To first-order approximation, the gravity components are

$$g_1 = \Gamma_{1,1}x + \Gamma_{1,2}y + \Gamma_{1,3}z, \quad (50)$$

$$g_2 = \Gamma_{2,1}x + \Gamma_{2,2}y + \Gamma_{2,3}z, \quad (51)$$

which yields

$$L_3 = 2m(x^2 - y^2)\Gamma_{1,2} + 2mxy(\Gamma_{2,2} - \Gamma_{1,1}) + mh(x\Gamma_{2,3} - y\Gamma_{1,3}). \quad (52)$$

Here, the $\Gamma_{j,k}$ are gravity gradients as opposed to the gravitational gradients of the previous sections.

We may introduce the azimuth, α , of the instrument with respect to north, such that $x = \ell \cos \alpha$ and $y = \ell \sin \alpha$. Then using the angular form of Hooke's law, $L_3 = -\tau(\theta_\alpha - \theta_0)$, Eq. (52) becomes

$$-\tau(\theta_\alpha - \theta_0) = m\ell^2((\Gamma_{2,2} - \Gamma_{1,1}) \sin 2\alpha + 2\Gamma_{1,2} \cos 2\alpha) + m\ell h(\Gamma_{2,3} \cos \alpha - \Gamma_{1,3} \sin \alpha), \quad (53)$$

where τ is the torsion coefficient of the wire; θ_0 is the torsion-free, zero-direction of the beam; and θ_α is the angle of the beam relative to the zero-direction when the instrument (casing) is placed at azimuth, α . Measurements at several azimuths are required to determine all unknown quantities: $\theta_0, \Gamma_{2,2} - \Gamma_{1,1}, \Gamma_{1,2}, \Gamma_{1,3}, \Gamma_{2,3}$.

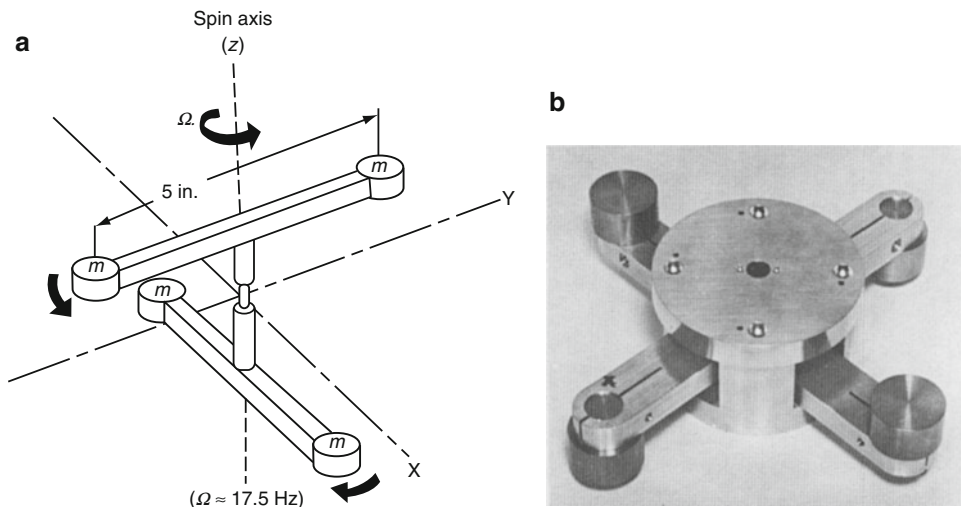
Eötvös designed several other versions, such as the double torsion balance (two opposing beams) that he used to study the equivalence of inertial and gravitational mass (which Einstein postulated as intrinsically true for his general theory of relativity; see section ► “Magnetic Gradiometry”). Though quite laborious and extremely sensitive to nearby masses in the field, the torsion balance (several versions built by various manufacturers) was widely used for geophysical prospecting of oil and gas in the early twentieth century both in Europe and in the USA. Ultimately, it gave way to the more efficient and equally accurate land gravity meter (Nettleton 1976; Chapin 1998).

Gravity gradiometry did not regain significant attention until the dawn of the space age since only gradiometers on planetary probes can make in situ measurements of the gravitational field. Early developments of moving-base gradiometers mimicked the torsional sensing apparatus, similar, in principle, to Eötvös's horizontal variometer, but in a smaller package (thus sensing only $\Gamma_{2,2} - \Gamma_{1,1}$ and $\Gamma_{1,2}$ in a local frame; Eq. (53) with $h = 0$). In addition, by rotating the device about the axis perpendicular to the mass baseline, the sensed gradients are modulated by twice the rotation rate due to the symmetry of the mass-baseline configuration. In this way some of the errors (such as accelerometer scale mismatches and misalignments) that modulate at once the rotation rate, Ω , can be separated in the frequency domain from the gradient signal that is modulated at twice this rate. That is, demodulation of the output signal at frequency, 2Ω , effectively filters these errors.

One of the earliest modern gradiometers was developed by Hughes Aircraft Research Laboratories for a lunar orbiter mission, the Hughes Rotating Gravity Gradiometer (RGG) (Forward 1981). Figure 8 shows a schematic of the device that senses the relative torques on the two arms that are coupled to the base by torsion pivots. Two dumbbells are required to

Gravity, Gradiometry,

Fig. 8 Concept (from Heller 1977) and breadboard model (from Forward 1981) of the Hughes Aircraft Rotating Gravity Gradiometer



distinguish the gravitational gradients from torques on the case in a dynamic environment. Furthermore, the rotation rate of the device is designed to be half its mechanical resonant frequency so as to amplify sensed gradients. A demonstration model of the RGG (which was never put into operation) had a diameter of 14 cm and an accuracy of 1 E for a 10 s integration time (Forward 1981).

The Charles Stark Draper Laboratory of Cambridge, Massachusetts, using its leading expertise in gyroscope technology, also devised several gradiometer concepts for moving-base applications. The “floated gradiometer,” pictured in Fig. 9a, was derived from the floated gyroscope whose buoyant/viscous/magnetic support system is virtually insensitive to high levels of acceleration and vibration. In this case, the device, like Eötvös’s torsion balance, had two anomalous masses at different heights (skewed mass configuration). If, for example, $\alpha = -90^\circ$ is substituted in Eq. (53) (maintaining the north-east-down coordinate system and indicating that the first axis of the device points west, toward $-x_2$), then the measurement is the applied restoring moment, M , in response to a deflection due to the gradients,

$$M = \tau(\theta_x - \theta_0) = 2m\ell^2\Gamma_{1,2} - m\ell h\Gamma_{1,3}, \quad (54)$$

where ℓ is the radius of the device.

On a stabilized platform with axes constrained to the local north-east-down directions, a cluster of three such devices (Fig. 9b) yields measured moments,

$$\begin{aligned} M_1 &= 2m\ell^2\Gamma_{1,2} - m\ell h\Gamma_{1,3} \\ M_2 &= 2m\ell^2\Gamma_{1,2} + m\ell h\Gamma_{1,3} \\ M_3 &= -2m\ell^2\Gamma_{1,2} + m\ell h\Gamma_{2,3} \end{aligned} \quad (55)$$

which can be combined to yield the horizontal gradients, $\Gamma_{1,3}$ and $\Gamma_{2,3}$. Derived from the differences of corresponding torques (as shown), these measured gradients are insensitive

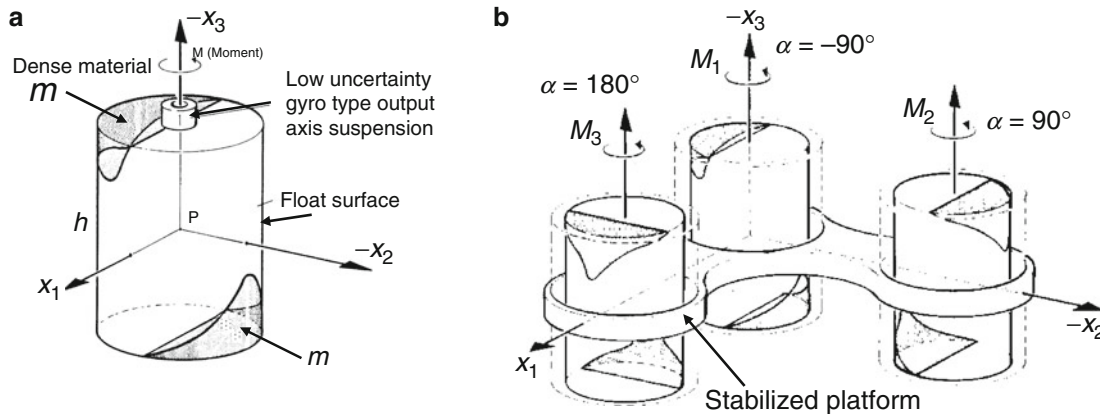
to angular accelerations about the vertical axis, the direction about which a north stabilized platform is least accurate. Later designs used electrostatically suspended spheres, with anomalous masses, that in an appropriate triad configuration would constitute a full-tensor gradiometer (with Laplace’s constraint) and be less susceptible to platform jitter effects (Wells 1984). Like the Hughes Aircraft instruments, the Draper Lab gradiometers did not reach the production stage but did demonstrate very good performance in the laboratory.

Another concept that utilized device rotation to separate the gradient signal from noise sources in the frequency domain was the gravity gradiometer instrument (GGI) of Bell Aerospace. The instrument consisted of pairs of accelerometers mounted on a rotating disk as shown in Fig. 10 (left). The input axes are perpendicular to the radius of the disk, and, thus, in analogy to the Hughes Aircraft RGG, their measurements may be thought of as proportional to torques about the spin axis. Substituting $\alpha = \Omega t$ and $\alpha = \Omega t + \pi/2$ for the respective pairs of accelerometers and $M = mr(a_1 + a_2)$ and $M = mr(a_3 + a_4)$ for the respective torques into Eq. (53) (with $h = 0$), the measurement equation, therefore, is given by

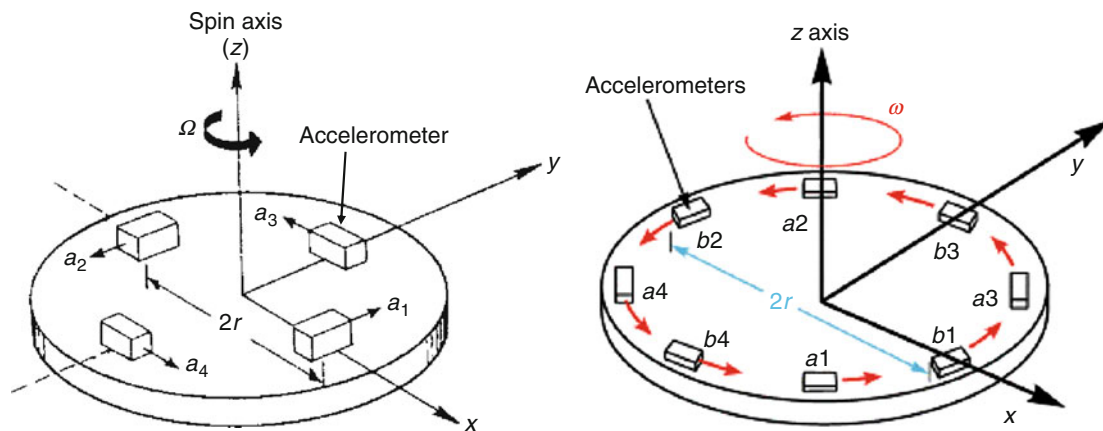
$$\begin{aligned} a_1(t) + a_2(t) - (a_3(t) + a_4(t)) \\ = 2r(\Gamma_{2,2} - \Gamma_{1,1}) \sin 2\Omega t + 4r\Gamma_{1,2} \cos 2\Omega t, \end{aligned} \quad (56)$$

where Ω is the spin rate of the disk and t is time. The factor of 2 on the right side comes from the dual pairs of accelerometers. This instrument has been used for military and commercial applications, and today’s version (acquired by Lockheed Martin, Fig. 10, right) has four pairs of accelerometers per disk, thus, in essence, further doubling the signal amplitude in the output (Schweitzer et al. 2000). However, three GGIs are still needed (e.g., in the configuration of Fig. 6) to obtain all elements of the gradient tensor.

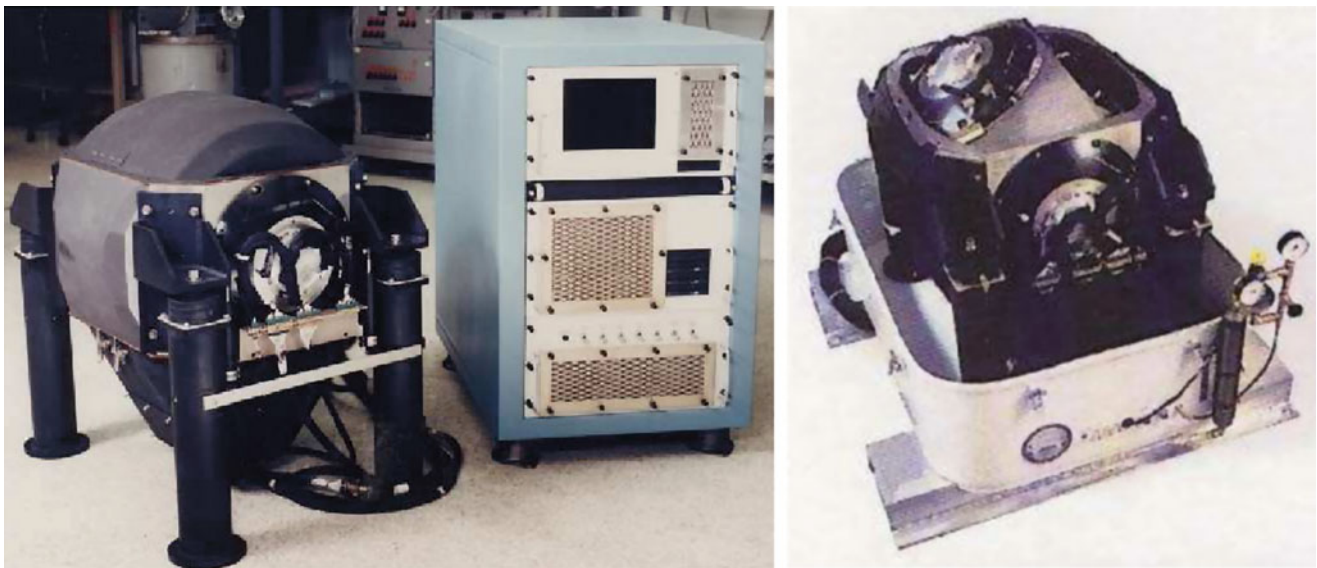
The Bell Aerospace (Lockheed Martin) gradiometer is currently operated for various geophysical and geodetic



Gravity, Gradiometry, Fig. 9 Floated gradiometer geometry of the Charles Stark Draper Laboratory; (a) single unit, (b) cluster of three units on a stabilized platform (both diagrams after Trageser 1970, reprinted with permission, Draper Laboratory)



Gravity, Gradiometry, Fig. 10 Left, Gravity Gradiometer Instrument (GGI), Bell Aerospace; right, current version, Lockheed Martin



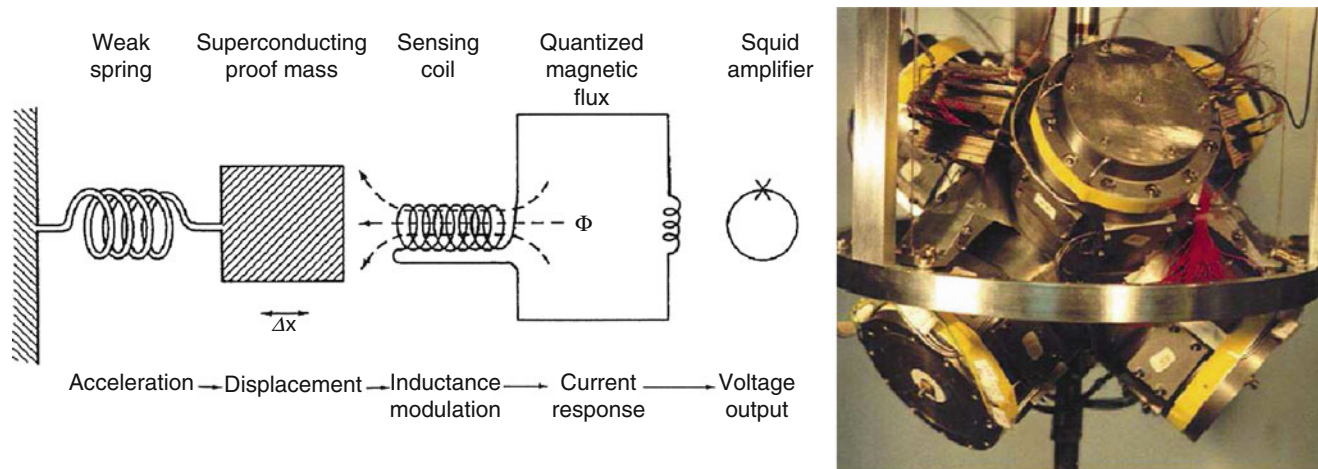
Gravity, Gradiometry, Fig. 11 Bell Aerospace (Lockheed Martin) GGI used on current terrestrial gradiometer systems. Left, the Fugro Falcon single-axis system (previously BHP Billiton, from van Leeuwen 2000, reprinted with permission, Society of Exploration Geophysics);

right, the full-tensor 3 GGI system of Bell Geospace (Murphy 2004; reproduced with permission, Bell Geospace). Both systems are shown on inertially stabilized platforms

applications by CGG (Compagnie Générale de Géophysique, using improved versions of the erstwhile BHP Falcon single-disk system, Fig. 11) and by Bell Geospace (full-tensor gradiometer, with the original four-accelerometer disks). These systems used primarily in the airborne mode for high efficiency have also been deployed on ocean vessels and ground vehicles.

Satellite in situ gravitational field measurements are limited fundamentally to gradiometry. Aside from small atmospheric drag forces and solar radiation (and Earth albedo) pressure, an Earth-orbiting satellite is in free fall. Accelerometers (at the center of mass) on a satellite sense only the non-gravitational forces, but a gradiometer, as on any moving

platform, senses the gravitational gradient (Eq. 46). Satellite gradiometry has a long, interesting, and ultimately successful history. Besides the Hughes RGG (which did not fly on a satellite), a major effort was undertaken by NASA to put a superconducting gravity gradiometer (SGG) into low Earth orbit in the 1980s. The SGG was developed by the University of Maryland (Moody et al. 1986; Paik et al. 1988) as an outgrowth of the technology for highly sensitive gravity wave detectors. Based on accelerometers that use the magnetic flux exclusion principle of superconducting bodies, the SGG was designed primarily to sense the in-line gradients. Referring to Fig. 12 (left), as the gap between the superconducting proof mass and the sensing coil changes due to

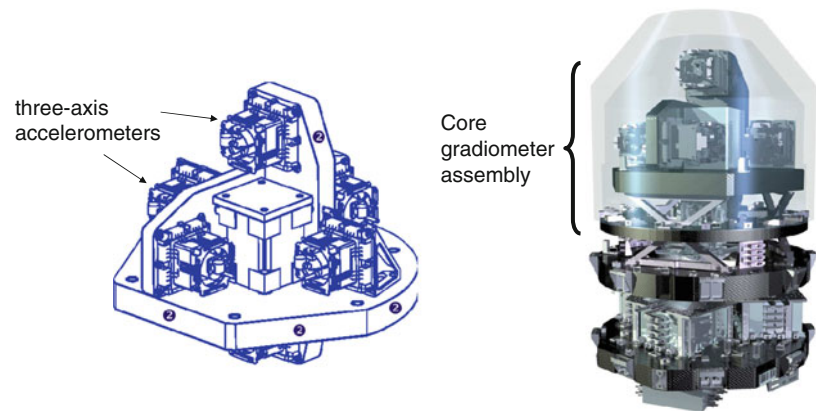


Gravity, Gradiometry, Fig. 12 Left, the accelerometer concept of the superconducting gravity gradiometer of the University of Maryland (reprinted with permission from Moody et al. (1986). Copyright

(1986), American Institute of Physics). Right, a three-axis system that senses the three in-line gradients (Paik 2004; reprinted with permission of author)

Gravity, Gradiometry,

Fig. 13 GOCE core gradiometer assembly (left) and total instrument (right)



acceleration, the magnetic flux is forced out of the sensing coil and into the input coil of a superconducting quantum interference device (SQUID), which is able to sense extremely small changes in the flux. The system was not flown on a satellite but did contribute to physics experiments involving the search for non-Newtonian gravitational effects (see section ► “Magnetic Gradiometry in Archaeo-geophysics”).

In March 2009, the first satellite-borne gravity gradiometer was launched by the European Space Agency (ESA) into low Earth orbit (altitude = 255 km) for a mission to map the Earth’s gravitational field globally to a resolution of 100 km, or better (Drinkwater et al. 2003). The *Gravity field and steady-state Ocean Circulation Explorer* (GOCE) gradiometer is a true full-tensor device (Fig. 13), based on three pairs of three-axis accelerometers, separated pair-wise by 0.5 m, thus in principle measuring all nine elements of the tensor, $\partial^2 a^b / \partial x^b$ (Eq. 46). Similar to airborne gravity gradiometry, the dynamical orbit of the satellite need not be determined with exquisite accuracy, and tracking was performed by the (nevertheless very precise) Global Positioning System (GPS). More

important is the attitude stabilization (see Eq. 47), and this was controlled by an active ion-propulsion drag compensation system, as well as onboard star trackers and magnetic torquing (Sechi et al. 2006), yielding a drag-free stabilized gradiometer platform. The originally planned 20-month mission lasted for about 55 months until in November 2013 the orbital decay could no longer be controlled resulting in the satellite’s demise.

Each accelerometer of the GOCE gradiometer has one weakly prelaunch calibrated axis, due to the earthly 1-g laboratory environment. As a consequence, the in-line components, $\partial a_1 / \partial x_1$, $\partial a_2 / \partial x_2$, $\partial a_3 / \partial x_3$, and the cross components, $\partial a_1 / \partial x_3$, $\partial a_3 / \partial x_1$, are the most accurate. The latter allow an accurate determination of the angular velocity, ω_2 , about the normal to the orbital plane, which is greatest in magnitude (Earth-pointing stabilization), since from Eq. (46) and the symmetry of the gravity gradient tensor, we have

$$\frac{\partial a_1}{\partial x_3} - \frac{\partial a_3}{\partial x_1} = 2\dot{\omega}_2. \quad (57)$$

As a consequence and since ω_1 and ω_3 are comparatively small, the most accurate gravity gradients obtained by the GOCE gradiometer are the in-line gradients and the cross gradient, $\Gamma_{1,3}$ (Rummel et al. 2011).

Other types of gradiometers may be found discussed in the literature and are at various stages in their development. For example, a cryogenic gradiometer based on the same principle as the Hughes Aircraft RGG is being constructed and tested by the Canadian company, Gedex (Tryggvason et al. 2004). Stanford University has investigated a gradiometer using dual cold-atom interferometer accelerometers (McGuirk et al. 2002). And, MEMS (microelectromechanical systems) technology is being applied to the manufacture of miniature gradiometers (Flokstra et al. 2009). The reader is referred also to the review by Difrancesco (2007).

Designing and using a gravity gradiometer system requires consideration of several significant sources of error. Gradiometers on a moving platform must detect a very small differential gravitational signal buried in a large-amplitude acceleration environment. An accuracy of 1 E over an accelerometer baseline of 0.1 m implies better than $7 \times 10^{-11} \text{ m/s}^2$ accuracy in the sensed accelerations, a resolving power that is challenging for most accelerometers. Rotation of the gradient sensor has been a key development to separate signal from noise in the frequency domain. Cryogenics is another solution, as is the dynamically quiet environment of a space vehicle.

Essential error sources for accelerometer-based gradiometers include the scale factor and bias stability of the accelerometers, as well as the degree to which their scale factors are matched and their sensitive axes are aligned. These all determine how well the common linear acceleration is rejected in the difference that ideally leaves only the gravitational gradient (with proper stabilization). Often, specialized electronic or mechanical devices or procedures (rotation and induced shaking) are used to eliminate or calibrate these erroneous effects. A particularly important calibration is the self-gradient of the vehicle that carries the gradiometer. Since the platform is stabilized (e.g., in the local level frame), the rotation of the vehicle about the platform creates a varying gradient field due to the changing attitude of the vehicle masses. This field must be calibrated, often using elaborate procedures that place the vehicle in varying attitudes while the gravitational field is constant (i.e., the vehicle is stationary) – see (Jekeli 1988b).

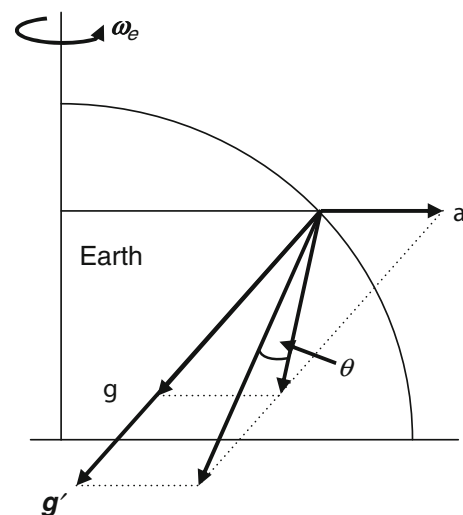
In summary, the development of instruments that measure gravitational gradients spans over 100 years; and, despite a variety of system advancements, mostly to make gradiometry mobile, many modern sensor concepts have strong analogies to the original Eötvös curvature variometer – they measure differential curvature parameters. In-line differential accelerometers are the only systems that have been or are in operation for scientific purposes, both on Earth and in orbit; but other technologies are also in development for future systems. The 1 E

accuracy of the Eötvös torsion balance rivals or still bests the terrestrial mobile systems of today, but the latter have improved the integration time vastly from many hours to tens of seconds, or better. In principle, gradiometry relies primarily on the technology of inertial sensors, and the theoretical limit in measurement precision in many cases has not yet been reached at the operational level. It is clear that gravity gradiometry is the holy grail of precision gravimetry on moving platforms – the future will see many more ideas and prototype sensors added to the existing repertoire of instruments.

Applications

One of the more interesting applications of the torsion balance was Eötvös's experimental validation of the equivalence of inertial and gravitational mass, assumed by Einstein in his theory of general relativity. Inertial mass is the proportionality factor between velocity and linear momentum (or, between acceleration and *applied* force), while gravitational mass is the proportionality factor between gravitational acceleration and gravitational force. Conceptually, the acceleration caused by the gravitational attraction of a mass body need not be indistinguishable from the acceleration caused by a force, such as propulsion, friction, or similar physical action. Yet the assumption of their equivalence led Einstein to his very successful theory.

Eötvös experimented already in the 1890s with the question of dependence of gravitational attraction on the composition of mass (continuing with Galileo's experiments that showed no such dependence). The idea of his experiment is illustrated in Fig. 14 showing that different types of test



Gravity, Gradiometry, Fig. 14 Eötvös's experiment to determine if different types of mass experience different gravitational accelerations (g versus g') that could be measured as a deflection, θ , with respect to the constant centrifugal acceleration, a

masses on the beam of the torsion balance might yield different gravitational accelerations that in combination with the unchanging centrifugal acceleration due to Earth's rotation would yield a measurable deflection. Eötvös et al. (1922) found no deviation at the 5 part-per-billion level. If indeed, such a deflection is detected, it could also be interpreted as evidence of a non-constant gravitational constant or an additional non-Newtonian force (Fischbach et al. 1986; Fischbach and Talmadge 1992).

Another experimental test of Newton's law of gravitation involving gravitational gradients considers the sum of in-line gradients, which for a Newtonian field must sum to zero (Fig. 15) in a vacuum (Laplace's Eq. 8). Moody and Paik (1993) used a swinging pendulum in the laboratory to induce modulated in-line gradients that were measured with their superconducting gradiometer (section ► "Gravity, Gradiometry"). The sum of the measured gradients was zero with a statistical variation ($\sigma = 3 \times 10^{-4}$ E) that placed further limits on any local departure from Newton's law of gravitation.

The principal utility of the torsion balance and later the gravity gradiometer was and is in geophysical exploration. The Eötvös torsion balance (and its successors) played a significant part, with over a hundred in operation worldwide, in the exploration for oil and gas reservoirs in the early twentieth century, 1918–1940. The achievable accuracy was as good as 1–3 E, but the measurements were arduous, requiring up to 6 h per station. Thus, the torsion balance quickly lost ground when the more portable and robust, as well as sufficiently accurate spring-type gravimeters were developed in the 1930s and 1940s.

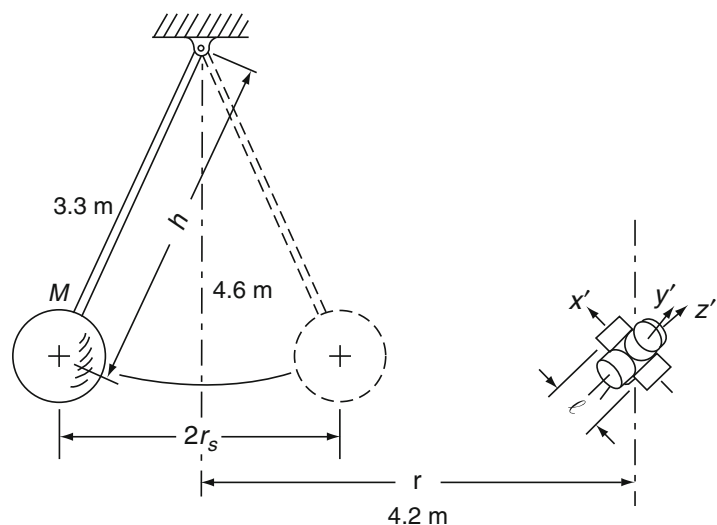
Gravity gradiometry in modern times is conducted almost exclusively with mobile systems, in order to take advantage of the increased resolution capability and avoiding the difficulty

of separating the kinematic and inertial accelerations, required in mobile gravimetry systems (Jekeli 2000). Submarine gradiometers are used to compensate inertial navigation systems for the effects of gravity and shipborne gradiometers have been used for marine gravity surveys. A significant investment has been made by a number of commercial exploration companies to use airborne gradiometers to map the gravitational field at high resolution in search for minerals, such as diamonds found in kimberlite pipes (Liu et al. 2001).

The first satellite-borne gradiometer mission to map the Earth's global gravitational field, GOCE, has yielded a number of high-resolution (100 km) global models derived solely from satellites and expressed as series of spherical harmonics. Since the gravitational gradients emphasize the shorter spatial wavelengths, most of these global models are obtained by combining the GOCE data with other satellite-tracking data, e.g., from GPS (Pail et al. 2011), or often the data from other satellite missions, specifically GRACE (Tapley et al. 2004), as well as previous long-wavelength models derived from ground-based tracking of a multitude of satellites. Furthermore, terrestrial gravity data serve to complement the satellite data for so-called combination models (e.g., Fecher et al. 2017). On the other hand, after more than 3 years of GOCE data, a model based exclusively on these data was also developed (Brockmann et al. 2014). Besides the global models, the GOCE mission has opened a new chapter on regional gravity modeling (e.g., Eicker et al. 2014; Lieb et al. 2015; Wu et al. 2017), as well as subsurface geophysics (e.g., Ebbing et al. 2013; Bouman et al. 2016; Eshagh and Pitoňák 2019). These detailed models of the global and regional field will continue to advance the Earth sciences in many fields, from oceanography to regional tectonics and the dynamics of the upper mantle and crust, and to polar ice sheets and glaciology.

Gravity, Gradiometry,

Fig. 15 Experimental setup to test Laplace's equation by measuring the sum of in-line gradients modulated by the changing gravitational field induced by a swinging pendulum (from Paik 2004; reprinted with permission of author)



Summary

Gravity gradiometry is a highly specialized approach to mapping the finer details of a gravity field. Although gradient sensors have existed for well over a century, the most efficient applications have come more recently from their deployment on moving-base systems, such as on ships, but more prominently on aircraft and space vehicles. From a theoretical viewpoint, measuring the gradients in these modes is more advantageous than measuring the field using conventional gravity meters since they are insensitive, in principle, to the linear accelerations of the vehicle, and sensing gradients implies detecting the shorter-wavelength features of the field. Despite these benefits, in practice, developing sensors of functional precision needed to discern these relatively small quantities requires exquisite isolation of vibrations and scale factor calibration. For this reason, to date there are but a few systems in operation.

Bibliography

- Badekas J (1967) The horizontal gradients of gravity in southwest Ohio. Report 89, Department of Geodetic Science, Ohio State University, Columbus
- Bouman J, Ebbing J, Fuchs M, Sebera J, Lieg V, Szwillus W, Haagmans R, Novak P (2016) Satellite gravity gradient grids for geophysics. *Sci Report* 6:21050
- Brockmann JM, Zehentner N, Hock E, Pail R, Loth I, Mayer-Gürr T, Schuh WD (2014) EGM_TIM_RL05: An independent geoid with centimeter accuracy purely based on the GOCE mission. *Geophys Res Lett* 41(22):8089–8099. <https://doi.org/10.1002/2014gl061904>
- Chapin D (1998) Gravity instruments; past, present, future. *Lead Edge* 17:100–112
- Difrancesco D (2007) Advances and challenges in the development and deployment of gravity gradiometer systems. Proceedings of the EGM 2007 International Workshop Innovation in EM, Grav and Mag Methods: a new Perspective for Exploration, Capri, 15–18 Apr 2007. <http://www2.ogs.trieste.it/eageseg/egm2007/INDEX.pdf>
- Drinkwater MR, Floberghagen R, Haagmans R, Muzi D, Popescu A (2003) GOCE: ESA's first earth explorer core mission. In: Beutler GB, Drinkwater M, Rummel R, von Steiger R (eds) *Earth gravity field from space – from sensors to earth sciences*, Space sciences series of ISSI, vol 18. Kluwer, Dordrecht, pp 419–432
- Ebbing J, Bouman J, Fuchs M, Lieb V, Haagmans R, Meekes S, Abdul Fattah R (2013) Advancements in satellite gravity gradient data for crustal studies. *Lead Edge* 32:900–906. <https://doi.org/10.1190/le32080908.1>
- Eicker A, Schall A, Kusche J (2014) Regional gravity modelling from spaceborne data: case studies with GOCE. *Geophys J Int* 196:1431–1440
- Eötvös L, Pekár D, Fekete E (1922) Beiträge zum Gesetze der Proportionalität von Trägheit und Gravität. *Ann. Physik, Series IV*, 68, 11. English translation in: Szabó Z (ed) (1998): *Three Fundamental Papers of Loránd Eötvös*. Loránd Eötvös Geophysical Institute of Hungary, Budapest
- Eshagh M, Pitoňák M (2019) Elastic thickness determination from on-orbit GOCE data and CRUST1.0. *Pure Appl Geophys* 176:685–696
- Fecher T, Pail R, Gruber T (2017) GOCO05c: a new combined gravity field model based on full normal equations and regionally varying weighting. *Surv Geophys* 38(3):571–590. <https://doi.org/10.1007/s10712-016-9406-y>
- Fischbach E, Talmadge C (1992) Six years of the fifth force. *Nature* 356(6366):207–215
- Fischbach E, Sudarsky D, Szafer A, Talmadge C, Aronson SH (1986) Reanalysis of the Eötvös experiment. *Phys Rev Lett* 56(1):3–6
- Flokstra J, Cuperus R, Wiegink RJ, van Essen MC (2009) A MEMS-based gravity gradiometer for future planetary missions. *Cryogenics* 49(11):665–668
- Forsberg R (1985) Gravity field terrain effect computations by FFT. *Bull Geod* 59:342–260
- Forward RL (1981) Gravitational science from the polar orbiting lunar observatory. Presented at the Gravity Gradiometer Workshop, NASA Headquarters, Washington, DC, 14–15 Jan 1981
- Freedon W, Gervens T, Schreiner M (1998) *Constructive approximation on the sphere*. Clarendon Press, Oxford University Press, New York
- Heiland CA (1940) *Geophysical exploration*. Prentice-Hall, New York
- Hein G (1977) Die Berechnung von anomalen vertikalen Schweregradienten hoher Präzision. *Deutsche Geodätische Kommission, Reihe C*, no 234, München
- Heiskanen WA, Moritz H (1967) *Physical geodesy*. Freeman and Co., San Francisco
- Heller WG (1977) Error models for prototype moving-base gravity gradiometers. Report AFGL-TR-77-0131, by The Analytic Sciences Corporation for Air Force Geophysics Laboratory, Hanscom AFB
- Hofmann-Wellenhof B, Moritz H (2005) *Physical geodesy*. Springer, Wien
- Jekeli C (1988a) The exact transformation between ellipsoidal and spherical harmonic expansions. *Manuscripta Geodaetica* 14:106–113
- Jekeli C (1988b) The gravity gradiometer survey system. *EOS, Trans Am Geophys U* 69(8):105, 116–117
- Jekeli C (2000) *Inertial navigation systems with geodetic applications*. Walter de Gruyter, Berlin
- Jekeli C (2006) Airborne gradiometry error analysis. *Surv Geophys* 27(2):257–275
- Jekeli C (2017) *Spectral methods in geodesy and geophysics*. CRC Press, Boca Raton
- Jekeli C, Zhu L (2006) Comparison of methods to model the gravitational gradients from topographic data bases. *Geophys J Int* 166:999–1014. <https://doi.org/10.1111/j.1365-246X.2006.03063.x>
- Jordan SK (1978) Statistical model for gravity, topography, and density contrasts in the Earth. *J Geophys Res* 83(B4):186–1824
- Lieb V, Bouman J, Dettmering D, Fuchs M, Schmidt M (2015) Combination of GOCE gravity gradients in regional gravity field modelling using radial basis functions. In VIII Hotine-Marussi Symposium on Mathematical Geodesy. *Int Assoc Geod Symp* 142:101–108. https://doi.org/10.1007/1345_2015_71
- Liu G, Diorio P, Stone P, Lockhart G, Christensen A, Fitton N, Dransfield M (2001) Detecting kimberlite pipes at Ekati with airborne gravity gradiometry. 15th conference, Australian Society of Exploration Geophysicists, Extended Abstracts
- McGuirk JM, Foster GT, Fixler JB, Snadden MJ, Kasevich MA (2002) Sensitive absolute-gravity gradiometry using atom interferometry. *Phys Rev A* 65:033608
- Moody MV, Paik HJ (1993) Gauss's law test of gravity at short range. *Phys Rev Lett* 70(9):1195–1198
- Moody MV, Chan HA, Paik HJ (1986) Superconducting gravity gradiometer for space and terrestrial applications. *J Appl Phys* 60(12):4308–4315
- Murphy CA (2004) The Air-FTG airborne gravity gradiometer system. In: Lane RJL (ed) *Airborne gravity 2004 – abstracts from the ASEG-PESA airborne gravity 2004 workshop: geoscience Australia record 2004/18*, pp 7–14

- Nettleton LL (1976) Gravity and magnetics in oil prospecting. McGraw-Hill, New York
- Paik HJ (2004) Probing extra dimensions using a superconducting gravity gradiometer. Presented at Caltech-JPL Association for Gravitational Wave Research Seminar, 13 Apr 2004, <http://cajagwr.caltech.edu/scripts/seminars.html>
- Paik HJ, Leung JS, Morgan SH, Parker J (1988) Global gravity survey by an orbiting gravity gradiometer. EOS, Trans Am Geophys U 69(48):1601–1611
- Pail R, Bruinsma S, Migliaccio F, Förste C, Goiginger H, Schuh WD, Hoeck E, Reguzzoni M, Brockmann JM, Abrikosov O, Veicherts M, Fecher T, Mayrhofer R, Krasbutter I, Sansò F, Tscherning CC (2011) First GOCE gravity field models derived by three different approaches. J Geodesy 85:819–843. <https://doi.org/10.1007/s00190-011-0467-x>
- Parker RL (1972) The rapid calculation of potential anomalies. Geophys J Roy Astro Soc 31:447–455
- Pavlis NK, Holmes SA, Kenyon SC, Factor JF (2012) The development and evaluation of Earth Gravitational Model (EGM2008). J Geophys Res 117:B04406. <https://doi.org/10.1029/2011JB008916>
- Pedersen LB, Rasmussen TM (1990) The gradient tensor of potential field anomalies – some implications on data collection and data processing of maps. Geophysics 55(12):1558–1566
- Petrovic S (1996) Determination of the potential of homogeneous polyhedral bodies using line integrals. J Geodesy 71:44–52
- Rummel R, Yi W, Stummer C (2011) GOCE gravitational gradiometry. J Geodesy 85:777–790. <https://doi.org/10.1007/s00190-011-0500-0>
- Schmidt M, Fengler M, Mayer-Gürr T, Eicker A, Kusche J, Sanchez L, Han SC (2007) Regional gravity modelling in terms of spherical base functions. J Geodesy 81:17–38
- Schreiner M (1997) Locally supported kernels for spherical spline interpolation. J Approx Theory 89:172–194
- Schweitzer M, Feldman WK, König WF, DiFrancesco DJ, Sieracki DL, San Giovanni CP (2000) System and process for optimizing gravity gradiometer measurements. US Patent no 6125698. US Patent and Trademark Office
- Sechi G, André G, Andreis D, Saponara M (2006) Magnetic attitude control of the GOCE satellite. Proceedings of the 6th International ESA Conference on Guidance, Navigation and Control Systems, Loutraki, 17–20 Oct 2005 (ESA SP-606, January 2006)
- Smith D, Robertson D, Milbert D (2001) Gravitational attraction of local crustal masses in spherical coordinates. J Geodesy 74:783–795
- Szabó Z (ed) (1998) Three Fundamental Papers of Loránd Eötvös. Loránd Eötvös Geophysical Institute of Hungary, Budapest
- Tapley BD, Bettadpur S, Watkins M, Reigber C (2004) The gravity recovery and climate experiment, mission overview and early results. Geophys Res Lett 31(9). <https://doi.org/10.1029/2004GL019920>
- Trageser MB (1970) A gradiometer system for gravity anomaly surveying. Report R-588, Charles Stark Draper Laboratory, Massachusetts Institute of Technology, Cambridge, MA
- Tryggvason B, Main B, French B (2004) A high resolution airborne gravimeter and airborne gravity gradiometer. In: Lane RJJ (ed) Airborne gravity 2004 – abstracts from the ASEG-PESA airborne gravity 2004 workshop: geoscience Australia record 2004/18, pp 41–47
- Tsoulis D, Petrovic S (2001) On the singularities of the gravity field of a homogeneous polyhedral body. Geophysics 66(2):535–539
- van Leeuwen EH (2000) BHP develops airborne gravity gradiometer for mineral exploration. Lead Edge 19(12):1296–1297
- Wells WC (ed) (1984) Spaceborne gravity gradiometers. Proceedings of the Spaceborne Gravity Gradiometer Workshop, NASA Goddard Space Flight Center, Greenbelt, 28 Feb–2 Mar 1983, NASA Conference Publication 2305
- Wu Y, Hao Z, Zhong B, Luo Z (2017) Regional gravity field recovery using the GOCE gravity gradient tensor and heterogeneous gravimetry and altimetry data. J Geophys Res 122. <https://doi.org/10.1002/2017JB014196>
- Yang J, Jekeli C, Liu L (2018) Seafloor topography estimation from gravity gradients using simulated annealing. J Geophys Res 123:6958–6975. <https://doi.org/10.1029/2018JB015883>

Great Earthquakes

Roger M. W. Musson
British Geological Survey, Edinburgh, UK
School of Geosciences, University of Edinburgh,
Edinburgh, UK

Definition

Any earthquake with magnitude over 8 Mw. Alternatively, any earthquake famous for one reason or another

Introduction

Under the technical definition, one can find some very obscure “great” earthquakes – a magnitude 8 earthquake in the Southern Ocean is likely to go wholly unremarked by the world’s media. Also, given the inherent imprecision in magnitude scales, it is possible for an earthquake with a magnitude of 8.0 to be successively promoted to and demoted from the canon of greatness with each successive revision.

If one took a sample of educated laymen and asked them to name any earthquakes they could think of, one could be sure to get a few of the most recent high-profile calamities; further back the results might vary from country to country. In the central USA, probably many are aware of the 1811 New Madrid earthquakes; in central Europe, the name of the 1356 Basel earthquake might come up; and so on. But probably not many could name correctly the largest earthquake ever known, the 1960 Valdivia, Chile, earthquake, which reached 9.6 Mw in magnitude, a value many seismologists suspect to be a physical upper limit of earthquake size. This devastating earthquake also produced a huge tsunami, which, on the Chilean coast, pitched boats up to 3 km inland (see Box-Out: The Ten Largest Earthquakes).

The title “largest earthquake ever” gives the 1960 Valdivia event an indisputable claim to greatness by any reckoning. Other earthquakes are famous for a variety of reasons: their place in history, the lessons learned from them, their significance in the development of seismology, or their sheer excessive destructiveness. The earthquakes of 1356 and 1811 mentioned above are both important for the same two reasons – they demonstrate that significant earthquakes can still

occur well away from plate boundaries and also that the historical record can reveal potential threats in places not much remarked for seismicity today.

The 1556 Shaanxi earthquake bears the record for being the most deadly earthquake on record, with about 830,000 dead. The figure of one million dead in one earthquake has not yet been reached (contrary to some sources, the 20 May 1202 earthquake in the Levant did not kill this many; Ambraseys and Melville 1988), but, with several current megacities dangerously exposed, it is feared that it is only a matter of time (see Box-Out: The Ten Deadliest Earthquakes).

There are probably four earthquakes that stand out as the most unquestionably great earthquakes of history. First was the 1755 Lisbon earthquake. Not only was this (with its accompanying tsunami) one of the most deadly and devastating earthquakes in European history (not forgetting its impact on North Africa), it sent shock waves throughout the whole of Western culture. Second, the 1906 San Francisco earthquake so ingrained itself into history that the very name of San Francisco conjures earthquakes to mind, to which must be added the important role this event played in the development of seismological theory. Third, the 2004 Sumatra earthquake, staggering in its immensity, changed both the popular and professional understanding of earthquakes. Lastly and more recently still, the 2011 Tohoku earthquake ended up being responsible for the worst nuclear accident since Chernobyl.

1 November 1755: The Advance of Rationalism

Of all great earthquakes, the one with the most lasting impact is certainly the Lisbon earthquake of 1755. Not only did it destroy what was then one of the most important cities in the world, the capital of the Portuguese empire; it shook up the very development of Western philosophy. Furthermore, it continues to be controversial among seismologists to this very day.

The first of November is the Feast of All Souls, one of the most solemn occasions in the calendar of the Catholic Church. It was on the morning of this day that a massive earthquake struck somewhere off the coast of Portugal. Damage occurred over a wide area across Portugal and western Spain; significant damage occurred even in Morocco. The shaking was perceptible throughout the Iberian Peninsula and into France. Even beyond France, long-period waves caused the agitation of water in ponds and lakes at great distances – as far away as Scotland (Loch Ness was disturbed) and Scandinavia.

The worst damage, however, occurred in Lisbon. The churches, many of which collapsed, were packed with worshippers when the quake struck, and casualties were heavy. The sea then mysteriously withdrew, leaving fishes floundering in the mud. Many ventured out from curiosity or to collect

some of the fish. Shortly afterward the tsunami roared in and surged up into the semi-ruined city, collapsing many buildings that had already been damaged by the earthquake.

All Europe was shocked, and not just by the scale of the destruction. It was axiomatic in Christian thinking that all prodigious events such as earthquakes were of divine origin. The milder ones were sent as a warning of God's displeasure; the more severe ones were a punishment to send mankind back on the path of righteousness. Furthermore, this was true whether or not it was appreciated that such events had natural causes, as many printed works of the previous century explicitly argued: God works through, and controls, the natural world. Whether or not earthquakes were due to natural causes (the dominant explanation at the time was the Aristotelian view that earthquakes were due to subterranean winds), the timing and location were divinely ordained (Walsham 1999).

But the injustice of the destruction of Lisbon strained this train of thought to breaking point. Not only did the earthquake occur on one of the holiest days of the Christian year – the godly citizens who were marking the occasion by attending mass were killed, whereas sinners who chose to spend the morning in playhouses or brothels by and large escaped. If this was God's judgment, it was a very strange one. The earthquake was seized upon by Voltaire as a means of attacking the optimistic school of philosophy characterized by Leibniz (Kendrick 1956). In Lisbon, a decree was issued that anyone found preaching that the earthquake was sent by God should be put to death.

For modern study of the 1755 earthquake, there are several sorts of information that can be used to try and understand which fault produced it. Firstly, there is the distribution of earthquake effects; for most historical earthquakes, analysis of the distribution of intensity is the predominant means of location, though this is markedly less easy for offshore events. Secondly, for tsunamigenic earthquakes like 1755, the time and run-up of the tsunami at different locations can be inverted to estimate the parameters of the engendering fault rupture. Thirdly, some use can be made of information from aftershocks. Since these are generally much smaller than the main event, they are usually felt only at short distances. Lastly, in the case of a major earthquake, there should be some geological trace of the structure that produced it.

In terms of available geological structures, there are two main candidates, and a possible third. The first, and most obvious, is the Gorringe Bank, a segment of the offshore plate boundary between Europe and Africa. This was identified as the source of the 1755 earthquake by Johnston (1996); it was also the location of a large earthquake in 1969, and the intensity distribution of the 1969 event resembled to some extent a scaled-down version of the pattern observed from the 1755 event (Levret 1991).

The second candidate is a more recent discovery – a fault off the coast of Portugal trending NNE-SSW showing signs of

recent activity (Baptista et al. 1998). This was named the Marquês de Pombal Thrust in honor of the man who oversaw the reconstruction of Lisbon. This fault provides a good fit to analyses of the tsunami arrivals on the Portuguese coast (Baptista et al. 2003). However, it doesn't seem very likely that an earthquake on this fault would have caused damage in Morocco.

The third is a controversial suggestion by Gutscher et al. (2002) and Gutscher (2004) that the causative feature was a N-S aligned subduction zone in the Gulf of Cadiz, dipping eastward under Gibraltar. Apart from the fact that the very existence of such a feature is disputed, if this was the source, it is hard to understand why Lisbon was destroyed in place of Cadiz. The earthquake is well described in Gibraltar itself, and damage was insignificant.

There are still more problems though. Given that the magnitude of this great earthquake must have been around 8.7 (Johnston 1996), none of the faults described seem long enough to host such a great earthquake. Recent suggestions to account for this propose multiple near-simultaneous earthquakes: Baptista et al. (2003) and Baptista and Miranda (2005) propose a composite source involving the Marquês de Pombal Thrust and a fault segment on the southern flank of the Guadalquivir Bank, while Vilanova et al. (2003) have the second earthquake on a fault in the Lower Tagus Valley, close to Lisbon. The issue is still unresolved and controversial; the source of this great earthquake is still an enigma.

18 April 1906: The Archetype of Earthquakes

The San Francisco earthquake of 18 April 1906 is of all earthquakes, probably the most iconic in popular culture, the one earthquake that practically any layman can mention. As Davison (1927) remarked, "if the Californian earthquake were the only shock known to mankind, the attention paid to it could hardly have been more exclusive." Not only was it a major disaster in a romantic locale, it even had a celebrity cast – Caruso being in the city at the time. Thanks to the telegraph, the news spread quickly around the world, much faster than had been the case in 1755.

As with the Lisbon earthquake, fire was a contributor to the damage; but this time, it was the major contributor. Earthquakes are great fire starters, upsetting stoves and breaking gas pipes. But they also hinder firefighting efforts. As the fire gripped the city, much of which consisted of wooden buildings, it was found that the water supply system was also damaged by the quake. Unable to use hosepipes, the city authorities resorted to trying to control the fire with a system of firebreaks. Unfortunately, this was carried out somewhat ineptly. The black powder that was used often simply set fire

to the building being blasted; and where it didn't, it blew houses into clouds of matchwood that easily ignited – the firebreaks being often created too close to the advancing flames. Furthermore, no attempt was made to check what the buildings marked for demolition actually contained. One such building, belonging to the Viavi pharmaceuticals company, contained a huge store of spirit alcohol. The resulting fireball so exacerbated the fire that probably another 50 blocks in the northern part of the city were destroyed as a result (Fradkin 2006).

The final death toll is a matter of dispute; initial figures of around 700 are clearly underestimates, partly for propaganda purposes (reporters were encouraged to call it the Great San Francisco Fire and play down mention of the earthquake). Current estimates are about 3000 dead (Stover and Coffman 1993).

From a seismological point of view, the 1906 earthquake is rather more straightforward than the 1755 event. The fault it occurred on is not just known; it is perhaps the most famous fault in the world: the San Andreas, forming the plate boundary between the North American and Pacific plates. This is a right-lateral strike-slip fault; the fault rupture extended probably around 430 km including the offshore extension, with slip of over 6 m in places (Lawson 1908–1910; Segall and Liskowski 1990; Stover and Coffman 1993). Since strike-slip faults are limited in width to the thickness of the crust, this gives a magnitude of only around 7.8 Mw. The felt area was much less than that of the Lisbon earthquake, covering only California and extending partially into Nevada and Oregon (attenuation of seismic shaking is higher in California than in Iberia).

The dislocation of fence posts across the fault trace made it very obvious that the earthquake was due to fault displacement. It was studying the 1906 rupture that led Reid (1910) to develop the theory of elastic rebound, which still underpins basic seismological theory today – that a fault gradually deforms due to the buildup of crustal stresses, until a locking asperity breaks, releasing energy as the fault violently snaps. The connection between earthquakes and faulting was well-known at the time; Hoernes (1878) held that the majority of earthquakes were tectonic in origin, due to displacement along fractures related to mountain building. As early as 1855, Lyell identified repeated earthquake-related faulting in New Zealand (Lyell 1856), but it is not clear that he understood that the fault produces the earthquake and not vice versa. The earliest statement that dislocation on a fault actually causes earthquake shaking seems to be Mallet (1851), but he was ahead of his time.

However, Reid's fundamental understanding began a new era in seismology, opening the door to the interpretation of seismograms in terms of fault mechanics (Wald et al. 1993; Agnew 2002).

26 December 2004: Cataclysm

Although most of the world's subduction zones are around the Pacific, there are two in the Indian Ocean: on the Makran coast of Iran and along the south coast of Sumatra (extending to Burma in the north and bending round to Java in the east). While the historical seismicity of Sumatra (including a major event in 1833) was by no means unknown (Newcomb and McCann 1987), it was not generally considered high on the list of potential localities for catastrophe before the earthquake of Boxing Day 2004. The epicenter was off the Sumatran coast at the northern end of the island, near Banda Aceh, and the rupture propagated almost entirely northward, for a staggering distance of 1600 km (Nalbant et al. 2005), reaching up to the Andaman Islands. The northern part of the rupture zone, where the plate boundary is extremely oblique to the direction of convergence, was almost north-south in orientation. The resulting tsunami, instead of heading off into the southern Indian Ocean as occurred in 1833, headed due east and west, hitting crowded tourist beaches in Thailand about 15 minutes later (shortly after 8 a.m. local time) and India and Sri Lanka two hours after that.

The effect was literally a cataclysm (Gr: down washing), as the waves swept inland up to about the 30 m mark in some places, causing immense destruction. About 230,000 lost their lives, of which 167,000 were in Indonesia. It tends to be the case that most high-fatality earthquakes kill local people in substandard housing; this event was exceptional in killing large numbers of foreign tourists, which accounts in part for the great impact of the event on public consciousness in the West. (For a while afterward, journalists would enquire of seismologists “will there be a tsunami?” even in cases for earthquakes in central Russia.)

The result was a sudden impetus toward installing tsunami warning systems in different parts of the world, though perhaps not always with the understanding that technical solutions are of limited use without the practical means to evacuate people at risk in a timely fashion.

Given that no earthquake anything like as large had occurred anywhere in the world since the 1960s, in the following days, the analysis of the earthquake tested the methods of modern seismology in new ways. The shock waves circled the globe, being distinctly recognizable even on the third pass, and even altered the rotation of the planet itself, the length of day dropping by up to $-5 \mu\text{s}$ (Bizouard 2005). The pole itself shifted by up to 2 mas (milliarcseconds) toward 90° E (Bizouard 2005; 1 mas equals an equatorial shift at the Earth's surface of 3.1 cm). The difficulty for analysts was caused by the excessive duration of the rupture; the unidirectional propagation of the rupture took about 10 minutes to reach its terminus. This complicated the seismic records as waves propagating from different parts of the rupture at

different times overwrote each other. The immediate impact was underestimation of the earthquake magnitude – immediate press releases giving values of 8.5 or 8.7. Initial attempts to model the slip distribution could only manage the southernmost part of the rupture (Ji 2004). The first complete slip models were made by Ammon et al. (2005) from a combination of body waves, intermediate-period waves and long-period waves, and Vigny et al. (2005) using an inversion of GPS data (the two models disagree extensively). The final magnitude assessment is 9.2 Mw.

Previously, it had been thought that such a large earthquake was impossible on this part of the Sunda thrust system, on tectonic grounds, and due to parts of the system having already broken in previous large earthquakes. The fact that the 2004 earthquake ruptured through segments that had already recently ruptured has caused seismologists to reassess the risk of great thrust earthquakes in other areas (e.g., Suárez and Albini 2009).

It was speculated at the time that the movement due to the earthquake would increase the stress on the Great Sumatran Fault, a strike-slip fault running the length of the island, and bring this close to failure (Nalbant et al. 2005). As of the time of writing, the expected event on this fault has not yet occurred. Instead, a remarkable cascade of events occurred on segments of the thrust zone south of the 2004 rupture, including further great earthquakes on 25 March 2005 and 12 September 2007 with magnitudes ≥ 8.5 Mw. Current estimates (e.g., Megawati and Pan 2009) are that there is still potential slip in this part of the system to make further great earthquakes on the coast of Sumatra a continuing threat.

11 March 2011: Prepared or Not Prepared?

Just over 6 years after the Great Sumatra earthquake, another event with magnitude over 9 occurred, this time off the east coast of Japan. The earthquake occurred during the working day (14:46 JST), with an epicenter offshore from Tohoku region and a rupture extending some 400 km, spreading bilaterally NNE and SSW of the epicenter, unlike the mostly unidirectional rupturing in 2004. Inevitably, this great offshore subduction earthquake (dubbed by some a “mega-thrust”) produced a major tsunami, and it is the tsunami that had the greatest impact. The death toll from the earthquake was a few hundreds, which for a magnitude 9 earthquake striking a heavily populated country has to be considered a remarkable achievement for Japanese earthquake engineering. But the tsunami killed about 20,000. What went wrong?

One factor was that some Japanese remembered previous tsunamis that had originated on the American side of the Pacific and which were therefore much attenuated by the time they reached Japan. This fostered the idea that tsunamis

were not especially dangerous. Footage from video cameras showed people idling on the harbor side while klaxons were still sounding the alert. There is even footage of one Japanese man laboriously lugging a heavy television set up a hill while the approaching tsunami is already visible behind him. The video record of this tsunami is extensive and often astonishing. It includes video made from a dashboard camera in a car that was picked up and tossed around by the tsunami. Amazingly, both driver and camera survived the experience.

More serious was the fact that Japanese authorities had severely underestimated the hazard. Before 2011, researchers had considered the subduction system east of Japan to be well-understood; that the plate interface regularly broke in a characteristic fashion, in which earthquakes of 7.0–7.5 Mw would rupture a single segment, and that there was therefore no risk of several segments failing together in a single devastating event (Hino 2015). Thus the areas marked out as safe refuges in the event of a tsunami were anything but safe. This was despite evidence that earthquakes in the distant past had caused great tsunamis, notably the Jogan earthquake of 869 (Abe et al. 1990; Sawai et al. 2008).

To compound matters, the tsunami hit the Fukushima Daiichi nuclear plant. Again, there had been precautions only against a much smaller tsunami, and the defenses were directly overrun. The earthquake took out the main power supply, and the emergency generators were in the basement, which was immediately flooded. With no power at all on site, the reactor cooling was inoperable, and a meltdown was the result – the worst nuclear disaster since Chernobyl. The area around the site is still an exclusion zone, and cleaning it up will take decades.

Estimates of the total economic loss from the Tohoku earthquake and tsunami range from \$250 to 350 billion, making it probably the single most expensive natural disaster ever. It raises difficult questions about preparedness. Evidently the affected settlements could have been better prepared; but fishing communities have to live on the coast. And is protecting a town with a 15 m high sea wall a realistic option?

Box-Out: The Ten Deadliest Earthquakes

A number of lists have appeared over the years purporting to give the ten earthquakes with the highest number of fatalities. Many of these are prone to easily avoidable errors, such as perpetuating fake earthquakes like the notorious 1737 Calcutta event (see Bilham 1994 for the refutation of this) or the 893 Ardebil (Iran) earthquake. However, compiling an accurate list is difficult, since for even relatively modern earthquakes, death tolls are often inaccurate, given the sheer difficulty of making a body count in the

aftermath of a major disaster (as witness the difficulty of estimating the death toll in the 2010 Haiti earthquake). The problem is greatly multiplied for historical earthquakes, where sources are vague, unspecific, or exaggerated, and in most cases, there was no census to begin with.

The earthquake generally agreed to be the deadliest of all time is the 1556 Shaanxi earthquake, with 830,000 deaths. This great earthquake ($M \sim 8$) affected an area where many people were living in loess caves, which collapsed, burying the inhabitants.

The second deadliest is probably the 1976 Tangshan disaster. The death toll from this event has been a matter of controversy; the earthquake occurred during the Cultural Revolution in China, when news about disasters was generally suppressed. Shortly after the earthquake, a suspiciously accurate figure of 655,237 deaths appeared in the Hong Kong media (South China Morning Post, 5 Jan 1977). The official death toll later released was 242,419 (Chen 1988), though Marza (2004) suggests the true figure may have been over a million.

There are a number of earthquakes with death tolls in the region of 200,000–300,000 that are hard to discriminate between. Willis (1928) attributes 260,000 deaths to the Syrian earthquake of 115, but a comprehensive review of the sources by Ambraseys (2009) shows that while the death toll must have been very high, the total figure is unknown. There are several candidates for third place. The 1920 Ningxia earthquake killed 235,502 according to Utsu (2002) and 160,000 according to Tiedemann (1992). The official death toll of the 26 December 2004 (9.2 Mw) Sumatran earthquake is 227,898, mostly from the tsunami. The 1139 earthquake in Ganja, Azerbaijan, is considered by Ambraseys and Adams (1989) to have killed 230,000, a figure rejected by Guidoboni and Comastri (2005) as exaggerated (considering the total population of the affected area). Finally, the 12 January 2010 Haiti earthquake, at the time of writing, is believed to have killed around 220,000–230,000.

Events with death tolls around 200,000 include another Chinese earthquake, in Shanxi province in 1303, with a death toll around 200,000 (Fu et al. 2005), and the 856 Qumis earthquake in Iran, believed to have killed around 200,000 (Ambraseys and Melville 1982).

Next highest is undoubtedly the 1923 Kanto (Japan) earthquake with 143,000. The only other earthquake (leaving aside events with very doubtful high death

(continued)

tolls) with attributed deaths over 100,000 is the 1138 Aleppo earthquake, but this is another case where, according to Ambraseys (2009), the actual figure is quite unknown (Ambraseys 1962 gives 130,000).

If one removed 1138 and 1139 from the list as too doubtful, the next two events would be the Sichuan earthquake of 2008 with 87,587 and the 1908 Messina earthquake in Southern Italy (82,000).

This gives us the following list (all magnitudes are moment magnitude, Mw):

1.	23 January 1556	~8	Shaanxi (830,000)
2.	27 July 1976	7.5	Tangshan (240,000)
3.	16 December 1920	7.8	Ningxia (235,000)
4.	26 December 2004	9.2	Sumatra (230,000)
5.	12 January 2010	7.0	Haiti (230,000)
6.	17 September 1303	~8	Shanxi (200,000)
7.	22 December 856	~7	Qumis (200,000)
8.	25 September 1139	~7	Azerbaijan (high)
9.	1 September 1923	7.9	Kanto (143,000)
10.	15 October 1138	~7	Aleppo (130,000)

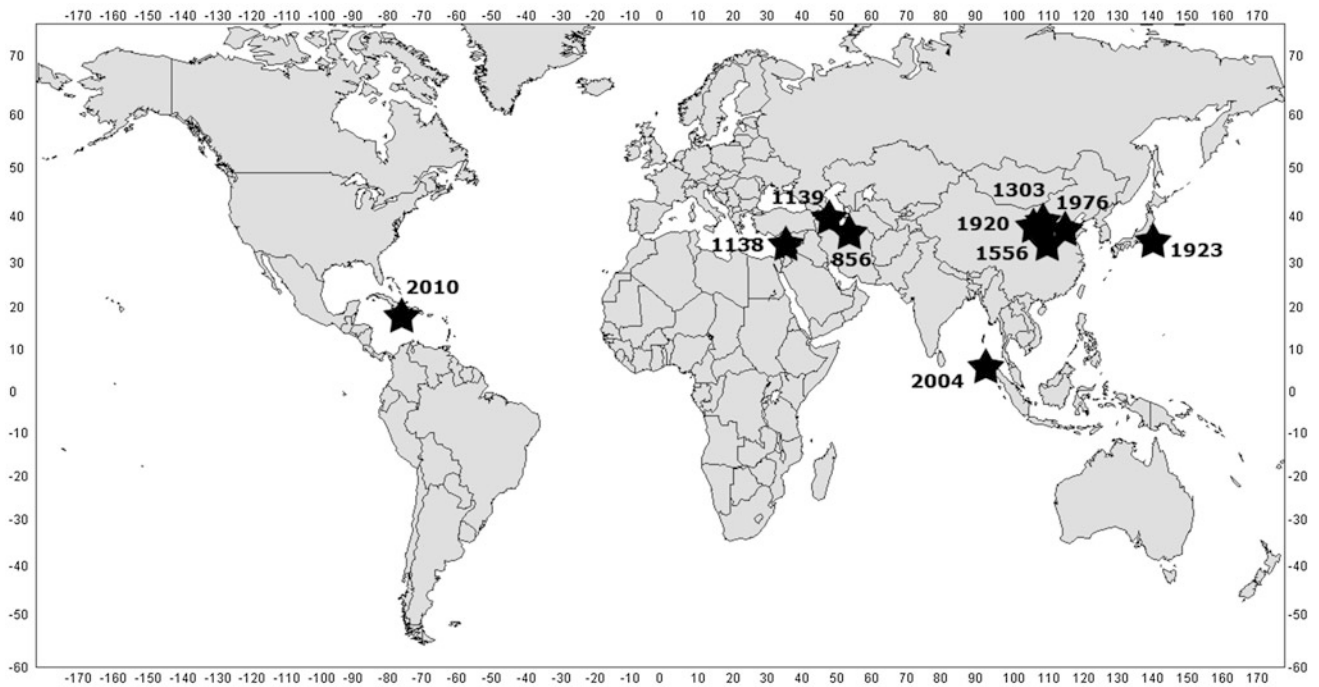
Numbers of fatalities are rounded to the nearest thousand; the precise ranking is obviously conjectural, as the figures are really only indicative of general loss levels. The distribution is shown in Fig. 1

Box-Out: The Ten Largest Earthquakes

Listing the ten largest earthquakes presents different problems. For the period since 1900, it is possible, at least in theory, to provide indisputable magnitude values from instrumental data, although figures are not precise for the very largest events. For the pre-instrumental era, the situation is much harder. Early attempts to convert historical data into estimates of magnitude tended to rely on maximum intensity; however, this rapidly saturates in large earthquakes. More sophisticated methods based on intensity data fields are available (e.g., Sibol et al. 1987), but these are unlikely to be accurate for very large events (the lack of calibration is one problem).

Earthquakes around the 9.0 Mw mark are necessarily subduction events – only large subduction thrust planes can provide large enough rupture areas to accommodate the vast energies required. Estimation of magnitude from non-instrumental data in such cases cannot rely on intensity distribution, which is liable to be incomplete, but really needs to be based on estimating rupture area and displacement. It was possible, for instance, on the morning of 26 December 2004, to see at once that the great Sumatran earthquake was at least 9 Mw just by looking at the huge extent of

(continued)



Great Earthquakes, Fig. 1 Map of the ten deadliest earthquakes in history

the aftershock distribution, which stretched for over 1000 km.

Since 1900, one can at least be certain that the catalogue for very large earthquakes is complete. Taking the Centennial Catalogue of Engdahl and Villaseñor 2002 as a basis, one can list the ten largest earthquakes since 1900 as follows (all magnitudes are moment magnitude, M_w):

1.	22 May 1960	9.6	Chile
2=.	28 March 1964	9.2	Alaska
2=.	26 December 2004	9.2	Sumatra
4.	9 March 1957	9.1	Andreanof Islands
5=.	11 March 2011	9.0	Tohoku
5=.	4 November 1952	9.0	Kamchatka
7=.	31 January 1906	8.8	Colombia-Ecuador
7=.	27 February 2010	8.8	Chile
9=.	11 November 1922	8.7	Chile
9=.	28 March 2005	8.7	Sumatra

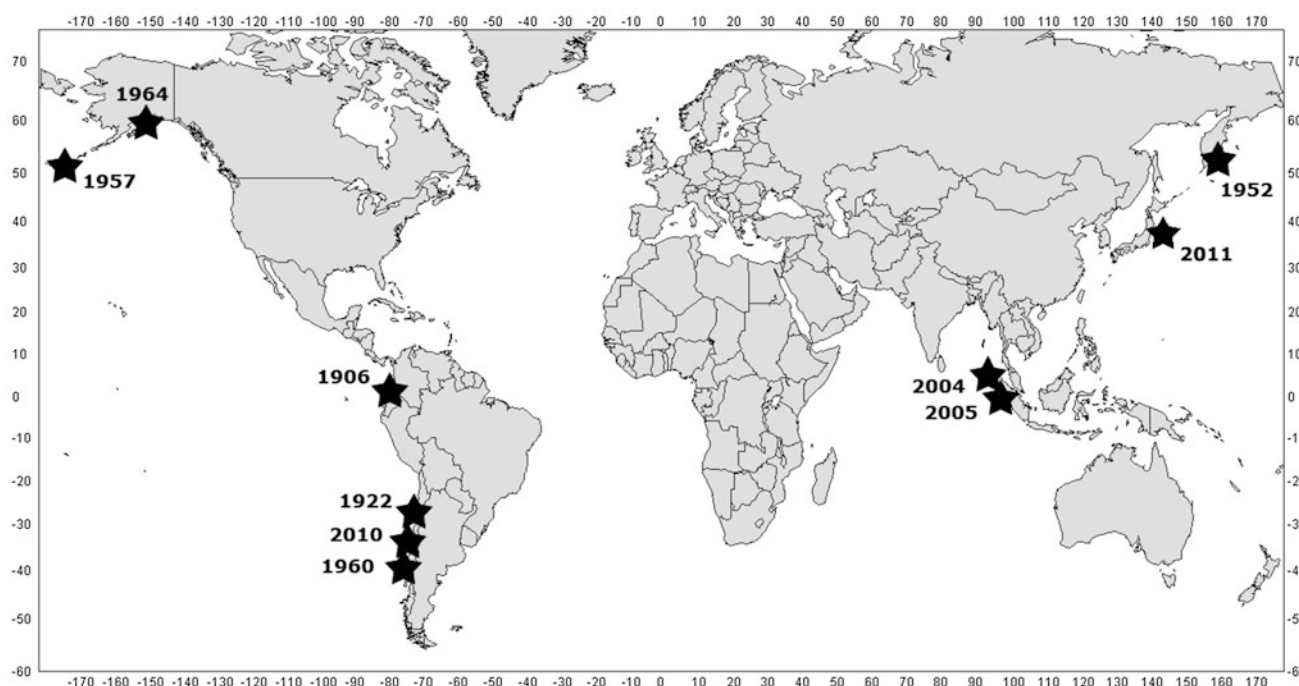
The 2005 Sumatra earthquake, while not strictly an aftershock of the 26 December 2004 event, was certainly triggered by it. It was a very strange experience to be telling journalists that it was “only” 8.7 and small in comparison to the Boxing Day earthquake. It is somewhat remarkable that all the magnitude 9s in the

twentieth century, accounting for most of the seismic energy release of the whole century, occurred in the space of just 12 years, from 1952 to 1964. The years 2004–2011 appear to making a similar spike in the energy release record.

The distribution is shown in Fig. 2, which shows a very different pattern from Fig. 1. The 2004 Sumatra earthquake is the only one to appear on both lists. This should emphasize that the deadliness of an earthquake generally owes more to its location with respect to population centers, than to its size.

Summary

Two lists of earthquakes are presented here, the ten largest known earthquakes in historical times and the ten deadliest in terms of reported fatalities. The lists have marked differences, since a very large earthquake in a remote area will not kill many and a moderate earthquake in a heavily populated area may kill many, especially where building standards are low. Four case studies are presented of great earthquakes that have left a particular legacy.



Great Earthquakes, Fig. 2 Map of the ten largest earthquakes since 1900

Cross-References

- [Characteristic Earthquakes and Seismic Gaps](#)
- [Earthquake, Aftershocks](#)
- [Earthquake, Magnitude](#)
- [Seismicity, Subduction Zone](#)
- [Seismic Seiches](#)
- [Subduction Zones](#)
- [Tsunami](#)

Acknowledgments The author would like to thank the late Professor N.N. Ambraseys and Dr. D.J. Kerridge for helpful reviews of the text.

Bibliography

- Abe H, Sugeno Y, Chigema A (1990) Estimation of the height of the Sanriku Jogan 11 earthquake-tsunami (A.D. 869) in the Sendai Plain. *J Seismol Soc Jpn* 43:513–525
- Agnew DC (2002) History of seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*. Academic Press, San Diego, pp 3–12
- Ambraseys NN (1962) A note on the chronology of Willis' list of earthquakes in Palestine and Syria. *Bull Seismol Soc Am* 52:77–80
- Ambraseys NN (2009) *Earthquakes in the mediterranean and middle East*. Cambridge University Press, Cambridge
- Ambraseys NN, Adams RD (1989) Long-term seismicity of Northern Armenia confirmed by the historical record. *Eos* 70:152–154
- Ambraseys NN, Melville CP (1982) *A history of persian earthquakes*. Cambridge University Press, Cambridge
- Ambraseys NN, Melville CP (1988) An analysis of the Eastern mediterranean earthquake of 20 May 1202. In: Lee WHK, Meyers H, Shimazaki K (eds) *Historical seismograms and earthquakes of the World*. Academic Press, San Diego, pp 181–200
- Ammon CJ et al (2005) Rupture process of the 2004 Sumatra-Andaman earthquake. *Science* 308:1133–1139
- Baptista MA, Miranda JM (2005) Evaluation of the 1755 earthquake source using tsunami modelling. In: *International conference of the 250th anniversary of the 1755 Lisbon Earthquake*. PRO, Lisbon, pp 574–577
- Baptista MA, Miranda PMA, Miranda JM, Mendes-Victor L (1998) Constraints on the source of the 1755 Lisbon tsunami inferred from numerical modelling of historical data. *J Geodyn* 25(2):159–174
- Baptista MA, Miranda JM, Chierici F, Zitellini N (2003) New study of the 1755 earthquake source based on multi-channel seismic survey data and tsunami modelling. *Nat Hazards Earth Syst Sci* 3:333–340
- Bilham R (1994) The 1737 Calcutta earthquake and cyclone evaluated. *Bull Seismol Soc Am* 84(5):1650–1657
- Bizouard C (2005) Influence of the earthquakes on the polar motion with emphasis on the Sumatra event, *Journées Systèmes de Référence Spatio-Temporels*. Proceedings, Warsaw, pp. 229–232
- Chen Y (1988) *The great Tangshan earthquake of 1976: an anatomy of disaster*. Pergamon, New York
- Davison C (1927) *The founders of seismology*. Cambridge University Press, Cambridge
- Engdahl ER, Villasenor A (2002) Global seismicity: 1900–1999. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*. Academic Press, San Diego, pp 665–690
- Fradkin PL (2006) *The great earthquake and firestorms of 1906*. University of California Press, Berkeley
- Fu Z-X, Liu G-P, Shao H-C, Ding X (2005) An overview on shallow strong earthquake activity and earthquake live losses of Chinese mainland in the century from 1901 to 2001. *Acta Seismol Sin* 18(4):392–401
- Guidoboni E, Comastri A (2005) *Catalogue of earthquakes and tsunamis in the mediterranean area from the 11th to the 15th century*. INGV, Bologna
- Gutscher M-A (2004) What caused the great Lisbon earthquake? *Science* 305:1247–1248
- Gutscher M-A et al (2002) Evidence for active subduction beneath Gibraltar. *Geology* 30:1071–1074
- Hino R (2015) An overview of the Mw 9, 11 March 2011, Tohoku earthquake. *SummBull Int Seismolo Cent* 48:100–132
- Hoernes R (1878) *Erdbeben Studien*. Jahrbuch der Kaiserlich-Königlichen Geologischen Reichsanstalt 28:387–448
- Ji C (2004) Preliminary result of the 04/12/26 (Mw 9.0), off W coast of northern Sumatra earthquake. <http://www.gps.caltech.edu/~jichen/Earthquake/2004/aceh/aceh.html>
- Johnston AC (1996) Seismic moment assessment of earthquakes in stable continental regions. III. New Madrid 1811–1812, Charleston 1886, and Lisbon 1755. *Geophys J Int* 126:314–344
- Kendrick TD (1956) *The Lisbon earthquake*. Methuen, London
- Lawson AC (ed) (1908–1910) *The California earthquake of April 18, 1906: report of the state earthquake investigation commission, vol 2*. Carnegie Institution of Washington, Washington DC
- Levet A (1991) The effects of the November 1, 1755 Lisbon earthquake in Morocco. *Tectonophysics* 193:83–94
- Lyell C (1856) Sur les effets du tremblement de terre du 23 Janvier, 1855, à la Nouvell Zélande. *Bulletin de la Société Géologique de France* 13: 661–667
- Mallet R (1851) First report on the facts of earthquake phenomena. *Rep Br Assoc Advanc Sci* 20:1–89
- Marza V (2004) On the death toll of the 1999 Izmit (Turkey) major earthquake, European seismological commission. PRO, Potsdam
- Megawati K, Pan TC (2009) Regional seismic hazard posed by the Mentawai Segment of the Sumatran Megathrust. *Bull Seismol Soc Am* 99:566–584
- Nalbant SS, Steacy S, Sieh K, Natawidjaja DH, McCloskey J (2005) Earthquake risk on the Sunda Trench. *Nature* 435:756–757
- Newcomb KR, McCann WR, 1987. Seismic history and seismotectonics of the Sunda Arc. *Journal of Geophysical Research* 92(1):421–439.
- Reid HF (1910) The mechanics of the earthquake. In: Lawson AC (ed) *The California earthquake of April 18, 1906: report of the state earthquake investigation commission*. Carnegie Institution of Washington, Washington DC
- Sawai Y, Fujii Y, Fujiwara O, Kamataki T, Komatsubara J, Okamura Y, Satake K, Shishikura M (2008) Marine incursions of the past 1500 years and evidence of tsunamis at Suijin-numa, a coastal lake facing the Japan Trench. *The Holocene* 18:517–528
- Segall P, Liskowski M (1990) Comparison of surface displacements in the 1906 San Francisco and 1989 Loma Prieta earthquakes. *Science* 250:1241–1244
- Sibol MS, Bollinger GA, Birch JB (1987) Estimations of magnitudes in central and eastern North America using intensity and felt area. *Bull Seismol Soc Am* 77:1635–1654
- Stover CW, Coffman JL (1993) *Seismicity of the United States, 1568–1989 (Revised)*. United States Government Printing Office, Washington
- Suárez G, Albin P (2009) Evidence for great tsunamigenic earthquakes (M 8.6) along the Mexican subduction zone. *Bull Seismol Soc Am* 99:892–896
- Tiedemann H (1992) *Earthquakes and volcanic eruptions: a handbook on risk assessment*. Swiss Re, Zurich, 951 pp.
- Utsu T (2002) A list of deadly earthquakes in the World: 1500–2000. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds)

- International handbook of earthquake and engineering seismology. Academic Press, San Diego, pp 691–718
- Vigny C et al (2005) Insight into the 2004 Sumatra-Andaman earthquake from GPS measurements in southeast Asia. *Nature* 436:201–206
- Vilanova SP, Nunes CF, Fonseca JFBD (2003) Lisbon 1755: A case of triggered onshore rupture? *Bull Seismol Soc Am* 93(5):2056–2068
- Wald DJ, Kanamori H, Helmberger D, Heaton TH (1993) Source study of the 1906 San Francisco earthquake. *Bull Seismol Soc Am* 83:981–1019
- Walsham A (1999) *Providence in early modern England*. Oxford University Press, Oxford
- Willis B (1928) Earthquakes in the Holy Land. *Bull Seismol Soc Am* 18:73–103

Heat and Ground Water Flow

Vincenzo Pasquale, Massimo Verdoya and Paolo Chiozzi
Department of Earth, Environment and Life Sciences,
University of Genoa, Genoa, Italy

Definition

Aquifer	Permeable underground that can store or transmit water.
Heat advection	Heat transfer by water motion due to pressure gradient.
Heat convection	Heat transfer by water motion due to density gradient.

Energy Equation for Ground Water Flow

The basic theory of heat and ground water flow is similar to other diffusion theories. The energy equation governing the mechanism of heat transfer in addition to water flow is expressed as

$$\rho c \frac{\partial T}{\partial t} = k \nabla^2 T - \rho_w c_w (u \cdot \nabla T) \quad (1)$$

where c , ρ , and k are the water-rock matrix specific heat, density, and thermal conductivity, respectively, t is the time, $\rho_w c_w$ is the volumetric heat capacity of water, and ∇^2 is the Laplace operator. The term $u \cdot \nabla T$ is the scalar product of the Darcy velocity u , or volumetric flow rate of water per unit area, and the temperature gradient ∇T . Equation 1 neglects heat sources and considers thermal conductivity as isotropic and temperature-independent. Moreover, it assumes that rock and water are in thermal equilibrium (see e.g., Turcotte and Schubert 2002; Pasquale et al. 2017).

The water displacement is linearly proportional to the driving force given by pressure and inversely proportional to the resistance due to water viscosity and type of interstices through which water flows. Darcy's law describes ground water flow by stating that the Darcy velocity u is connected with water pressure p in the form

$$u = -\frac{\kappa_p}{\eta_w} \nabla(p + \rho_w g z) = -\frac{\kappa_p \rho_w g}{\eta_w} \nabla H_h = -\lambda_c \nabla H_h \quad (2)$$

where κ_p is the permeability, η_w and ρ_w are the water dynamic viscosity and density, respectively, $H_h = z + p/(\rho_w g)$ is the hydraulic head, z is the elevation above a standard datum, g is the acceleration due to gravity, $\lambda_c = \kappa_p \rho_w g / \eta_w$ is the hydraulic conductivity. When ground water flow is caused only by an externally applied pressure gradient, the heat transfer mechanism is referred to as advection. If the flow is driven by buoyancy effect (i.e., variation in density caused by temperature gradient), the mechanism is called thermal convection.

In case of steady heat and water flows and only vertical heat transport (the horizontal temperature gradient is negligible), substitution of Eq. 2 into Eq. 1 gives

$$\frac{d^2 T}{dz^2} = \frac{\rho_w c_w \kappa_p}{k \eta_w} \left[\frac{d}{dz} (p + \rho_w g z) + g z \frac{d\rho_w}{dz} \right] \frac{dT}{dz} \quad (3)$$

where ρ_w is a function of depth z . By introducing in this equation the hydraulic head and the expansion coefficient of water $\alpha_w = (d\rho_w/dT)/\rho_w$, one obtains

$$\frac{d^2 T}{dz^2} = \frac{\rho_w^2 c_w g \kappa_p}{k \eta_w} \left(\frac{dH_h}{dz} + \alpha_w z \frac{dT}{dz} \right) \frac{dT}{dz} \quad (4)$$

This one-dimensional equation contains terms expressing the conductive heat transfer, heat advection, and thermal convection in a uniform, isotropic, homogenous aquifer.

Heat Advection

In case of advection the buoyancy force is negligible, and Eq. 4 becomes

$$\frac{d^2T}{dz^2} - \frac{c_w \rho_w u_z}{k} \frac{dT}{dz} = 0 \quad (5)$$

with u_z the Darcy velocity in the vertical direction (positive downwards). The solution of Eq. 5 is

$$T = T_1 + (T_2 - T_1) \frac{\exp(\beta_z z/h) - 1}{\exp \beta_z - 1} \quad (6)$$

where $\beta_z = c_w \rho_w u_z h/k$ is a dimensionless parameter, which may be positive or negative depending on whether u_z is downward or upward, and T_1 and T_2 are the temperatures at the top and at the bottom of the vertical distance h within the aquifer (Fig. 1). Equation 6 is valid for steady-state thermal conditions and for a flow rate assumed constant and so slow that thermal equilibrium between water and rock matrix is maintained.

More generally, because surface topographic relief usually exists across an aquifer, heat and water flow, particularly in a sloping semi-confining aquifer, is neither purely horizontal nor purely vertical. In this case, Lu and Ge (1996) demonstrated that the solution for a two-dimensional flow is an extension of Eq. 5. Assuming a linear variation of temperature along the horizontal direction, in the left-hand member of the equation we must add the term $-\Gamma_o(\beta_o/h)$ that accounts for the constant horizontal flow of heat and water. Therefore, we have

$$\frac{d^2T}{dz^2} - \frac{\beta_z}{h} \frac{dT}{dz} - \frac{\beta_o}{h} \Gamma_o = 0 \quad (7)$$

where $\beta_o = c_w \rho_w u_o h/k$, and Γ_o and u_o are the horizontal components of the temperature gradient and the Darcy velocity, respectively. The gradient Γ_o is taken positive (or negative) when the heat flow is leftward (or rightward). The solution of Eq. 7 is

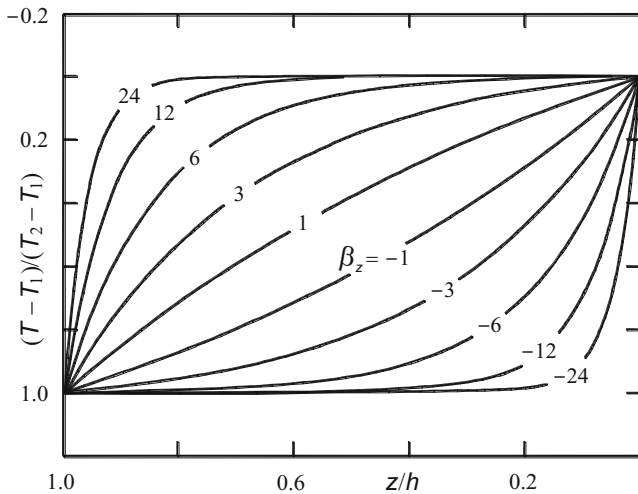
$$T = T_1 + (T_2 - T_1) \left[\frac{\exp(\beta_z z/h) - 1}{\exp \beta_z - 1} + \frac{\Gamma_o \beta_o}{\Gamma_z \beta_z} \left(\frac{\exp(\beta_z z/h) - 1}{\exp \beta_z - 1} - \frac{z}{h} \right) \right] \quad (8)$$

where Γ_z is the vertical temperature gradient. In the absence of horizontal heat or water flow ($\Gamma_o = 0$ or $u_o = 0$), Eq. 8 reduces to Eq. 6. Ratio $(T - T_1)/(T_2 - T_1)$ versus z/h for $(\Gamma_o \beta_o)/(\Gamma_z \beta_z)$ and different β_z is shown in Fig. 2.

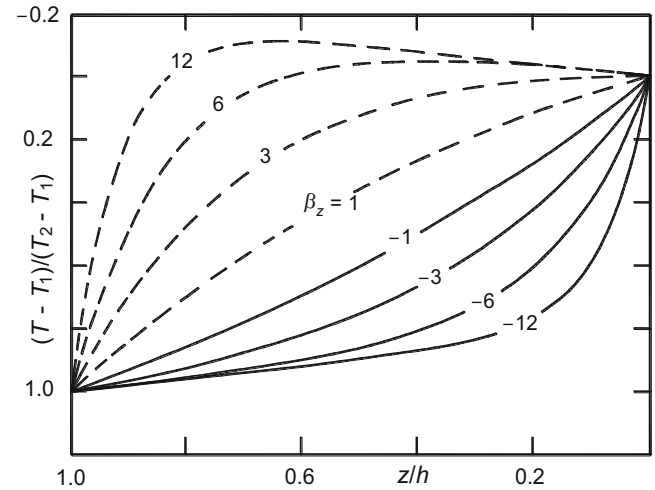
Reiter (2001) suggested that the analysis of the thermal effect of ground water can be practicable by comparing the temperature gradient Γ_z with temperature and depth. Integrating Eq. 7 once, we obtain

$$\Gamma_z = \frac{\beta_z T}{h} + \frac{\beta_o \Gamma_o z}{h} + C \quad (9)$$

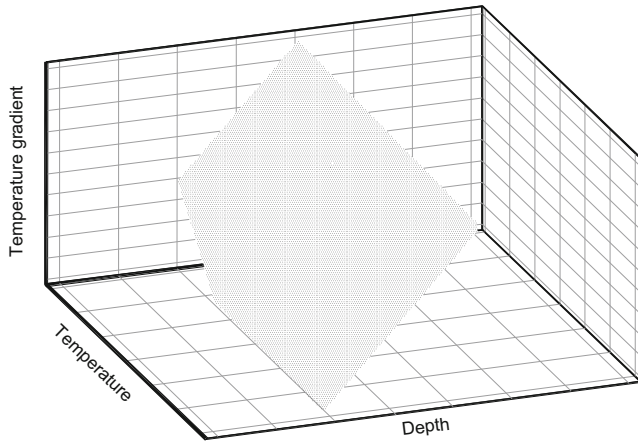
where C is the integration constant. Equation 9 can be seen as a plane whose slopes in relation with the axes of temperature and depth contain information on the vertical and horizontal components of the Darcy velocity (Fig. 3). When one component of water flow is dominant, Eq. 9 reduces to a straight line. Then, the temperature gradient is only a function of depth for $u_z = 0$, while it is a function of temperature for negligible horizontal flow.



Heat and Ground Water Flow, Fig. 1 Ratio $(T - T_1)/(T_2 - T_1)$ versus z/h for different values of β_z (water downward) and $-\beta_z$ (upward water)



Heat and Ground Water Flow, Fig. 2 Ratio $(T - T_1)/(T_2 - T_1)$ versus z/h for $(\Gamma_o \beta_o)/(\Gamma_z \beta_z) = 0.2$ (hatched curve), $(\Gamma_o \beta_o)/(\Gamma_z \beta_z) = -0.2$ (solid curve) and different β_z



Heat and Ground Water Flow, Fig. 3 The plane containing the temperature gradient versus depth and temperature (Eq. 9)

The quantities β_z and β_o measure the relative efficiency of an aquifer for the simultaneous heat transport by both water flow and pure conduction. They appear to be analogues of the thermal Péclet number, which quantifies the potential for advection to perturb the temperature-depth distribution. The vertical Péclet number Pe_z can be expressed as the ratio of the advected heat flow q_{ad} and the conducted heat flow q_c over a characteristic length L

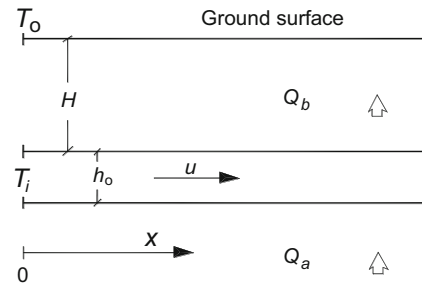
$$Pe_z = \frac{q_{ad}}{q_c} = \frac{\rho_w c_w |u_z| (T_2 - T_1)}{k (T_2 - T_1)/L} = \frac{\rho_w c_w |u_z| L}{k} \quad (10)$$

When $Pe_z \gg 1$ vertical advection dominates, while the conductive component prevails for $Pe_z \ll 1$ (Clauser and Villinger 1990). The length L can be chosen in many ways and is often selected depending on the scale of the flow system. Normally, it is assumed to correspond to the thickness of the aquifer. By substituting u_o for u_z into Eq. 10, it is possible to derive also the horizontal Péclet number.

The foregoing analytical solutions for heat and water flow can be applied to borehole temperature data, recorded by means of high precision equipment. The analysis procedure, which consists in matching temperature and temperature gradient data with theoretical curves obtained from Eqs. 6, 8, and 9, leads to the estimation of the Darcy velocity (see for details, Verdoya et al. 2008; Pasquale et al. 2010).

Hydrothermal Systems

There are special analytical solutions that provide estimates of the water temperature along flow paths like those occurring in an advection-dominated hydrothermal systems. An analysis of the heat transported by the water from the recharge to the discharge area can be carried out by assuming geometry and hydraulic properties of the hydrothermal system.



Heat and Ground Water Flow, Fig. 4 Sketch of a horizontal aquifer

Figure 4 sketches a horizontal water path of thickness h_o , at depth H below the surface, in which water moves at a Darcy velocity u . The water temperature T varies only along the horizontal x -axis. In a section dx of the aquifer, during time $dt = dx/u$, the heat per unit width that enters is $Q_a = q dx^2/u$, whereas the heat flowing out is $Q_b = k (T - T_o) dx^2/(H u)$, where q is geothermal flow beneath the aquifer, T_o the temperature at the surface and k the water-rock matrix thermal conductivity, assumed to be uniform with depth. The gain of heat $Q = (Q_a - Q_b)$ is

$$Q = \left(\frac{q}{u} - k \frac{T - T_o}{H u} \right) dx^2 \quad (11)$$

and the corresponding increase of water temperature is $dT = Q/(\rho_w c_w h_o \sigma dx)$, where σ is the effective porosity. Thus, along the direction x , we obtain

$$\frac{dT}{dx} = \left(q - k \frac{T - T_o}{H} \right) \frac{1}{\rho_w c_w \sigma h_o u} \quad (12)$$

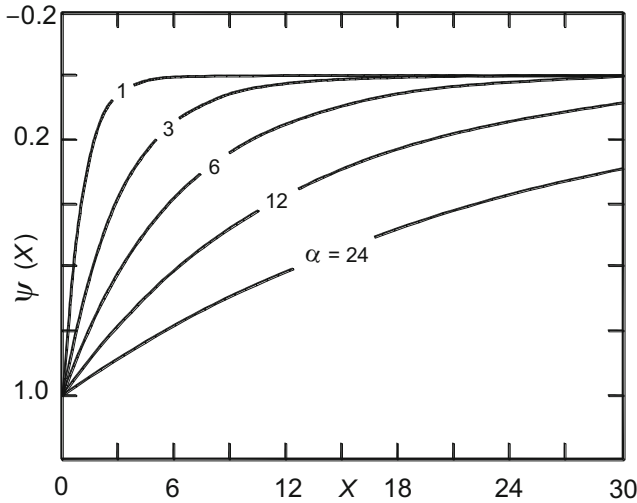
At $x = 0$, if the temperature $T = T_i$, the normalized temperature ψ as a function of X is given by (Schoeller 1962; Pasquale et al. 2011)

$$\psi(X) = \exp(-X/\alpha) \quad (13)$$

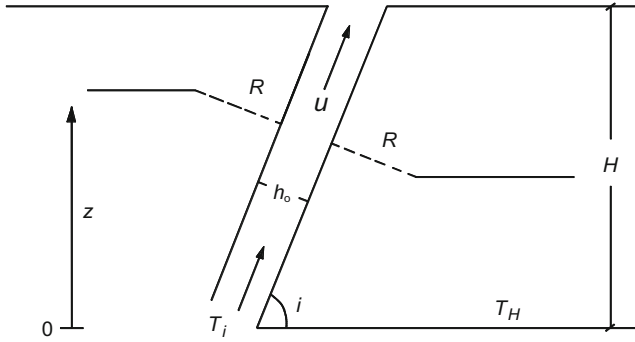
where $\psi(X) = (T(x) - T_H)/(T_i - T_H)$, $X = x/H$ and T_H is the background temperature at depth H given by $T_o + q H/k$ (Fig. 5). The parameter $\alpha = c_w \rho_w u h_o \sigma/k$ is dimensionless. This approach assumes that water and the aquifer inner boundaries have the same temperature because of the large thermal resistance of the rocks, the slow water flow, and the large contact surface.

In a recharge area, the ground water may seep through a descending flow path. The ground water temperature along the sloping flow path is (Schoeller 1962; Pasquale et al. 2011)

$$\phi(x) = \frac{\Gamma_z x}{\cot i + \alpha} \quad (14)$$



Heat and Ground Water Flow, Fig. 5 $\psi(X)$ for different values of α along a horizontal aquifer



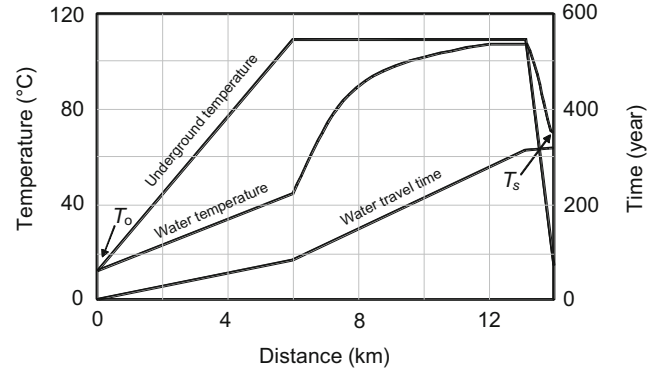
Heat and Ground Water Flow, Fig. 6 Scheme of an ascending aquifer

where $\varphi(x) = T(x) - T_i$ is the temperature of water along the horizontal x axis, T_i is the temperature of water at the beginning of the flow path, Γ_z is the background vertical temperature gradient, and i is the slope of the aquifer.

For ground water flow ascending to the surface, like, e.g., in a discharge area (Fig. 6), the water temperature as a function of z is given by

$$\vartheta(z) = -\Gamma_z z + (T_i - T_H) \exp\left(-\frac{2z}{R \alpha \sin i}\right) + \frac{\Gamma_z R \alpha \sin i}{2} \left[1 - \exp\left(-\frac{2z}{R \alpha \sin i}\right)\right] \quad (15)$$

where $\vartheta(z) = T(z) - T_H$ and T_H is the unperturbed temperature at depth $H(z = 0)$. The term $R = h_o e^5$ represents the thickness of the medium surrounding the aquifer and perturbed by the heat exchange. If we assume the discharge area as a single subvertical cylindrical conduit, by replacing h_o with the radius r of the conduit, the temperature is given by



Heat and Ground Water Flow, Fig. 7 Subsurface temperature, water temperature and water travel time in the hydrothermal system of Acqui Terme (NW Italy) (after Pasquale et al. 2011). $T_o = 12.5$ °C and $T_s = 70.0$ °C are the water temperature values in the recharge area and at the hot spring, respectively

$$Y(z) = -\Gamma_z z + \frac{\Gamma_z \sin i}{A} \left[1 - \exp\left(-\frac{A z}{\sin i}\right)\right] \quad (16)$$

where $Y(z) = T(z) - T_H$ is the water temperature as a function of depth and $A = 2 k/[c_w \rho_w u r^2 \ln(R/r)]$.

An example of the temperature and the water travel time estimated by means of Eqs. 13, 14, 15, and 16 is shown in Fig. 7. From the recharge area of the hydrothermal system, located at about 14 km from the hot spring (70 °C), meteoric water heats up due to the background geothermal flow while descending and horizontally flowing within the fractured basement. After the maximum temperature (110 °C) is reached (at a depth of 2800 m), water rapidly rises to the spring, through a sub-vertical fault. This scheme of deep ground water flow is consistent with geochemical data, and with borehole temperature logs. The water residence time within the hydrothermal system is about 300 years.

Heat Convection

In high-temperature hydrothermal systems, typically occurring in volcanic areas, convection is the dominating heat-transfer mechanism. Convection can occur also in deep carbonate reservoirs hosting low-temperature thermal water. A convenient way to analyze the occurrence of heat convection in an aquifer heated from below is the dimensionless Rayleigh number, R_a , defined as

$$R_a = \frac{g \alpha_w \rho_w c_w \kappa_p L \Delta T}{v_w k} \quad (17)$$

where v_w is kinematic viscosity of the saturating water, L is the length over which convection occurs, ΔT is the temperature difference between the top and bottom of this length interval. In the temperature range expected within an aquifer,

v_w can be evaluated with the following expressions (Holzbecher 1998)

$$v_w = 10^{-3} [1 + 0.015512 (T - 20)]^{-1.572} \quad 40^\circ \text{C} < T < 100^\circ \text{C} \quad (18)$$

$$v_w = 0.2414 \rho_w \times 10^{\frac{247.8}{T+133.15}} \times 10^{-4} \quad 100^\circ \text{C} < T < 300^\circ \text{C} \quad (19)$$

where water density ρ_w varies with temperature T as

$$\rho_w = 1000 \left[1 - \frac{(T - 3.9863^2) (T + 288.9414)}{508929.2 (T + 68.12963)} \right] \quad (20)$$

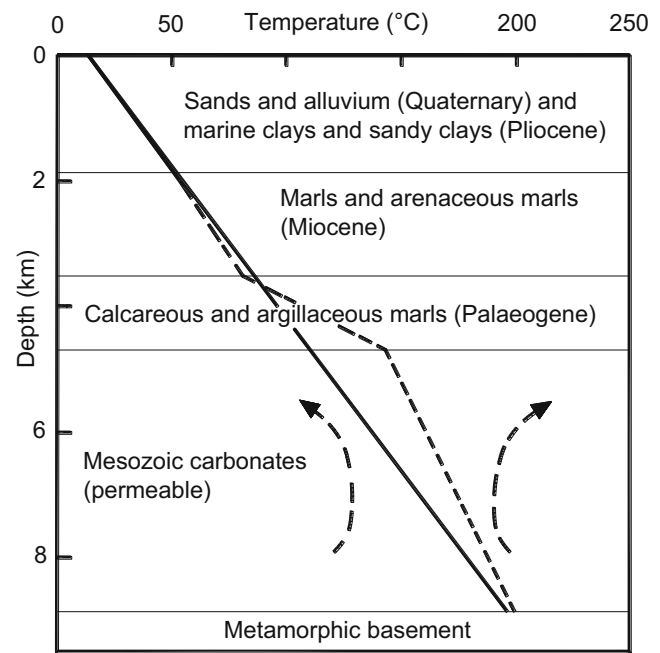
Therefore, convection occurs under appropriate thermal and permeability conditions. In practical terms, the Rayleigh number helps to differentiate between purely conductive systems, where isothermal constant density models are applicable, and low-grade hydrothermal systems, where nonisothermal models must be used. When the Rayleigh number exceeds the critical value of $4\pi^2$, convection starts and groundwater may form convective cells.

By analyzing thermal records from hydrocarbon exploration wells, Pasquale et al. (2013) provided evidence of thermal convection in the deep carbonate reservoir of the eastern Po Plain (Italy). Figure 8 shows the temperature-depth distribution as inferred through inversion of thermal data together with the temperature simply extrapolated from the observed thermal gradient. In the uppermost sedimentary layers, the inferred temperature matches the observed temperature, whereas within the Palaeogene layer and in the Mesozoic reservoir there are significant differences. As the thermal conductivity variation is small within the sedimentary sequence and the hydrogeological characteristics (including litho-stratigraphic sequences and structural setting) hardly permit advection, thermal convection can account for these differences. Convection could cause a decrease of temperature gradient in the reservoir and a significant increase in the overlying Palaeogene low-permeability unit.

The potential for thermal convection to take place within the reservoir can be quantified with Eq. 17. By assuming that $R_a = 4\pi^2$, $g = 9.8 \text{ m s}^{-2}$, $\rho_w c_w = 4032 \text{ kJ m}^{-3} \text{ K}^{-1}$, $k = 3.1 \text{ W m}^{-1} \text{ K}^{-1}$, $\alpha_w = 0.82 \times 10^{-3} \text{ K}^{-1}$, $v_w = 2.60 \times 10^{-7} \text{ m}^2 \text{ s}^{-1}$, and that all the parameters are independent from pressure, Eq. 17 becomes

$$\left(\frac{dT}{dz} \right)_{\min} = \frac{9.82 \times 10^{-10}}{\kappa_p L^2} \quad (21)$$

where the length L of convection is the thickness of the reservoir. Equation 21 provides the minimum temperature gradient



Heat and Ground Water Flow, Fig. 8 Temperature versus depth as inferred from inversion of temperature recorded in petroleum wells (broken line) and by extrapolating the observed temperature gradient (continuous line). Arrows indicate likely thermal convection in the carbonate reservoir

required for convection. For typical values of temperature gradient (14 mK m^{-1}) and thickness (5 km) of the Mesozoic carbonate reservoir, the inferred permeability, k_p , is larger than $3 \times 10^{-15} \text{ m}^2$. This value is consistent with the minimum range of $5 \times 10^{-17} - 10^{-15} \text{ m}^2$ required for thermal convection to occur in porous rocks (Manning and Ingebritsen 1999).

Conclusions

Although modeling of heat transfer in aquifers of complex structure with particular boundary conditions is often addressed by means numerical methods (see e.g., Haenel et al. 1988; Bartels et al. 2003), in mid-low permeability aquifers of simple geometry, the analytical methods presented here can quantitatively describe the heat transport coupled with single-phase water flow.

Cross-References

- Deep Scientific Drilling
- Geophysical Well Logging
- Geothermal Heat Pumps
- Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
- Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

Bibliography

- Bartels J, Kühn M, Clauser C (2003) Numerical simulation of reactive flow using SHEMAT. In: Clauser C (ed) Numerical simulation of reactive flow in hot aquifers – SHEMAT and Processing SHEMAT. Springer, Heidelberg, pp 5–74
- Clauser C, Villing H (1990) Analysis of conductive and convective heat transfer in a sedimentary basin, demonstrated for the Rheingraben. *Geophys J Int* 100:393–414
- Haenel R, Rybach L, Stegena L (1988) Fundamentals of geothermics. In: Haenel R, Rybach L, Stegena L (eds) Handbook of terrestrial heat-flow density determination. Kluwer Academic, Dordrecht, pp 9–57
- Holzbecher EO (1998) Modeling density-driven flow in porous media. Springer, Berlin
- Lu N, Ge S (1996) Effect of horizontal heat and fluid flow on the vertical temperature distribution in a semiconfining layer. *Water Resour Res* 32:1449–1453
- Manning CE, Ingebritsen SE (1999) Permeability of the continental crust: implications of geothermal data and metamorphic systems. *Rev Geophys* 37:127–150
- Pasquale V, Verdoya M, Chiozzi P (2010) Darcy velocity and Péclet number analysis from underground thermal data. *Boll Geofis Teor Appl* 51:361–371
- Pasquale V, Verdoya M, Chiozzi P (2011) Groundwater flow analysis using different geothermal constraints: the case study of Acqui Terme area, northwestern Italy. *J Volcanol Geoth Res* 199:38–46
- Pasquale V, Chiozzi P, Verdoya M (2013) Evidence for thermal convection in the deep carbonate aquifer of the eastern sector of the Po Plain, Italy. *Tectonophysics* 594:1–12
- Pasquale V, Verdoya M, Chiozzi P (2017) Geothermics: heat flow in the lithosphere, 2nd edn. Springer, Cham
- Reiter M (2001) Using precision temperature logs to estimate horizontal and vertical groundwater flow components. *Water Resour Res* 37:663–674
- Schoeller H (1962) Les eaux souterraines. Masson et Cie, Paris
- Turcotte DL, Schubert G (2002) Geodynamics: application of continuum physics to geological problems, 2nd edn. Cambridge University Press, New York
- Verdoya M, Pasquale V, Chiozzi P (2008) Inferring hydro-geothermal parameters from advectively perturbed thermal logs. *Int J Earth Sci* 97:333–344

Heat Flow Determinations, Continental

John H. Sass (Deceased)¹ and Graeme Beardsmore^{2,3}

²University of Melbourne, Parkville, VIC, Australia

³Hot Dry Rocks Pty Ltd, Windsor, VIC, Australia

Definition

Continental heat flow is the flux density of heat being conducted to the surface of continents. It is expressed as mW m^{-2} . Heat flow is calculated as the product of temperature gradient (mK m^{-1} or $^{\circ}\text{C per km}$ – a scale appropriate to the dimensions of geological formations) and thermal conductivity ($\text{W m}^{-1} \text{K}^{-1}$).

Introduction

Terrestrial heat flow, Q , is estimated from the product of temperature gradient (Γ) and thermal conductivity (K), according to:

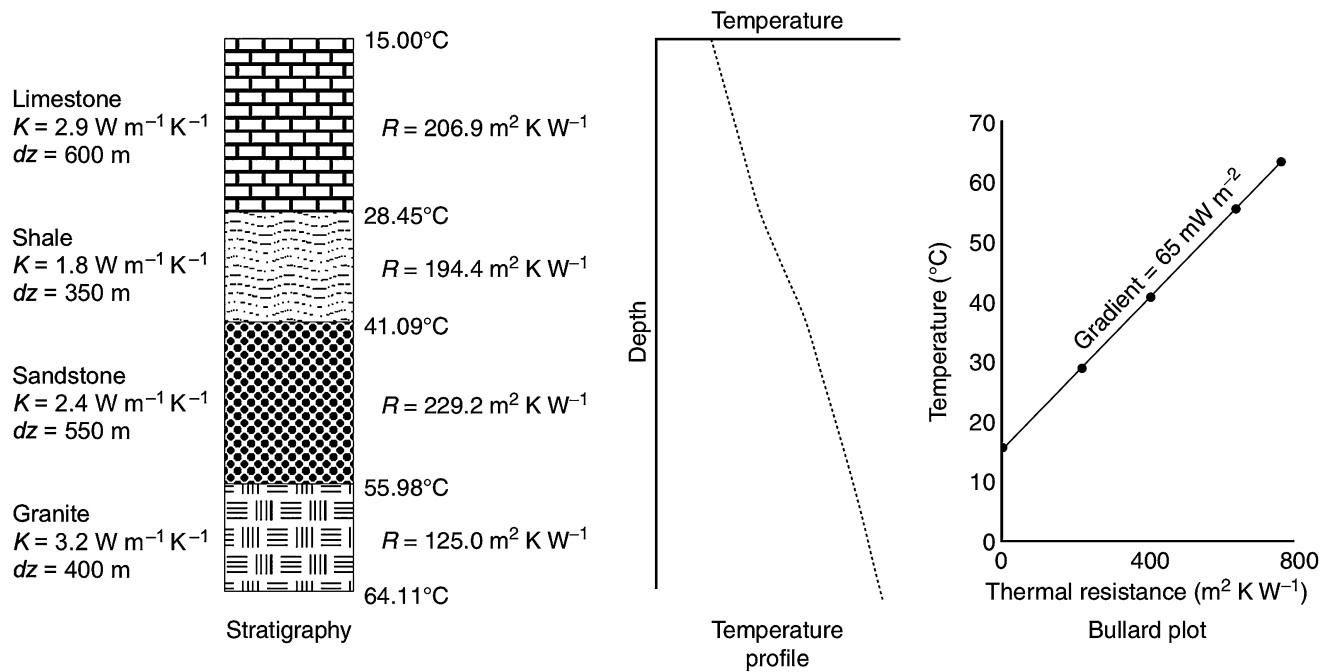
$$Q = K\Gamma \quad (1)$$

In practice, Γ is derived from point measurements of temperature at two or more discrete depths. K can be determined in situ, but it is acceptable and often more practical to measure it on samples brought to the surface, either in the field or in a laboratory.

Beardsmore and Cull (2001) reviewed several options for solving Eq. 1 using discrete data sets. These usually involve combining the harmonic mean thermal conductivity with the average temperature gradient over discrete intervals of the temperature profiles. Where there is significant thermal conductivity stratification, the “Bullard plot” is a useful method for calculating average heat flow (see Fig. 1). In this method, temperature is plotted against integrated thermal resistance, R , where $R = dz/K$, dz being the thickness of the interval with mean conductivity K . If heat flow is constant with depth, the relation is linear and the slope of the line is the heat flow.

Uncertainty is inherent in heat flow determinations:

- The average temperature gradient and the average thermal conductivity must be determined over the same depth interval for a precise determination of heat flow.
- Temperature gradient estimates become more precise by averaging over greater depth intervals, but thermal conductivity estimates become less precise.
- It is rarely practical to measure thermal conductivity continuously over an entire depth interval, so measured samples must be assumed to also represent unsampled sections.
- The transient thermal disturbance due to seasonal surface temperature cycles may extend to a depth of several tens of meters, so stable results can only be obtained at greater depths.
- The diffusion of historical step changes in surface temperature (e.g., from land clearance or the retreat of glaciation) may affect heat flow determinations down to hundreds of meters or more.
- Continental heat flow might not be parallel to vertical boreholes because heat flow is affected by significant lateral and temporal variations in surface temperature, especially when compared to determinations of heat flow beneath the oceans.
- Heat flow on the continents is also more likely than oceanic heat flow to be distorted by heat refraction by near-surface variations in geological structures and topography.



Heat Flow Determinations, Continental, Fig. 1 A hypothetical rock sequence with thermal properties and unit thicknesses as indicated. The thermal resistance, R , of each unit is equal to its thickness, dz , divided by its thermal conductivity, K . A temperature–depth profile through the

sequence is nonlinear, but a Bullard plot of temperature versus integrated thermal resistance is linear with slope equal to the heat flow. Heat flow in this example is 65 mW m^{-2}

Derived heat flow values may vary depending on the depth and sampling strategy of the survey, so careful thought should precede the design of any survey aimed at quantifying regional continental heat flow.

Temperature

The earliest measurements of temperature beneath the earth's surface were obtained from mines and tunnels. As exploratory drilling for petroleum and minerals became widespread, temperature profiles obtained in boreholes became the most common means of determining temperature gradients within the earth. Electronic thermometers deployed on a wire line are the most common tool for measuring temperatures within the earth, although the use of fiber optic systems is increasing.

The accuracy and precision of temperature gradient estimates depend on the accuracy and precision of temperature and depth measurements. Best practice calls for a precision of $\sim 0.001 \text{ K}$ at $\sim 1 \text{ m}$ intervals over several hundred meters. Temperature sensor types appropriate for borehole measurements include the following:

- **Thermistors:** A thermistor is a bead of sintered metallic oxide, commonly germanium. The electrical resistance across a thermistor decreases strongly with increasing

temperature (typically -4% per K between 0°C and 100°C). Simple electrical resistance measurement devices such as off-balance Wheatstone bridges and $4\frac{1}{2}$ digit digital multimeters (DMMs) are sufficient for 1 mK temperature precision (McGee and Thomas 1988). Thermistor probes for use in “normal” continental borehole environments usually have a resistance of between 3 and 100 k Ω at 25°C and provide precise estimates of temperature between about 0°C and 80°C , above which their temperature sensitivity drops off because of low resistance. Each thermistor has a unique resistance–temperature response so must be individually calibrated.

- **Platinum “resistance temperature detectors” (RTDs):** The key component of platinum RTDs is a length of pure platinum wire. The electrical resistance of platinum wire increases predictably with increasing temperature, although the response is less sensitive than that of thermistors. Depending upon their construction and configuration, platinum RTDs can operate successfully up to 500°C (see Tew and Strouse 2001, for principles and standards). 1 mK temperature precision can be obtained with RTDs if $6\frac{1}{2}$ digit DMMs are employed for readouts. RTD probes for continental borehole environments typically have a resistance of 1 or 2 k Ω at 20°C . One advantage over thermistors is that RTD sensors are interchangeable without the need for individual calibration.

- Diodes: Effective temperature sensors can be constructed from ordinary semiconductor diodes. Such sensors have linear voltage or current responses with temperature, are of relatively small size, operate over a limited temperature range (typically -40°C to $+120^{\circ}\text{C}$), are relatively cheap, and can achieve equivalent precision and accuracy to thermistors and RTDs if individually calibrated. Diode temperature sensors tend to be less robust (electrically and physically) than other sensor types, which tends to rule them out of most borehole applications. For stationary applications, however, commercial semiconductor sensors are available that provide an output signal as a voltage, a current, a resistance, or as digital data.
- Fiber optic “distributed temperature sensors” (DTS): Temperature measurement using fiber optic cables employs the phenomenon of “Raman scattering” of laser light. About one in ten million of the photons directed along the cable with a constant wavelength scatter in such a way that their wavelength either increases or decreases. The intensity of the scattered band with shorter wavelength is temperature dependent, while the intensity of the longer wavelength band is independent of temperature. The local temperature of the cable is derived from the ratio of the shorter to longer wavelength intensities. The location along the cable is inferred from the time elapsed between the incident laser pulse and the detection of the scattered signals. Temperature resolution increases relative to the number of repeated pulses over which each measurement is averaged. Commercially available instruments can achieve temperature accuracy of $\sim 1^{\circ}\text{C}$, temperature resolution of $\sim 0.01^{\circ}\text{C}$, and depth accuracy and resolution of $\sim 1\text{ m}$ (Selker et al. 2006). Instantaneous records of temperature distribution along continuous lengths of cable up to 30,000 m can be obtained within a fraction of a minute. Förster et al. (1997) presented a comparison between fiber-optic and conventional temperature logs.
- Down-hole oscillator with dual-conductor logging cable: Resistance-based elements can be incorporated into an oscillator circuit to create a frequency–temperature dependence. With a power source and amplifier included with the oscillator circuit in the probe, the frequency of the oscillating output current can be monitored in real time from the surface via the dual conductors, unaffected by leakage resistance between conductors. Doig et al. (1961) reported precision of 0.05°C .
- “Memory” tools: Memory, or “computer,” tools are fully self-contained probes including power source, memory, clock, and temperature sensor. Some also include onboard pressure gauges as a proxy for a depth sensor. Memory tools are initialized at the surface and deployed downhole on a single wire slickline. They record time–temperature data at preprogrammed intervals and store the data in onboard memory. The data are downloaded from the tool after it is recovered to the surface (Suto et al. 2008). Time–depth data are required to convert the time–temperature data into a depth–temperature log. Time–depth data can be recorded at the winch or wellhead during deployment or can be interpreted from time–pressure data if the probe is fitted with a pressure gauge.
- Fiber optic DTS: DTS systems include a length of fiber optic cable weighted on the end, a laser source, and a spectrometer to detect and quantify the frequency and intensity of back-scattered light. Henniges et al. (2005) described the deployment of fiber optic cable systems for both synoptic temperature logs and long-term monitoring of temperature.

The relative uncertainty in temperature gradient is the sum of the relative uncertainties in temperature differential and depth interval. These uncertainties decrease with increasing depth interval. In most cases, temperature gradient determined over an interval of several tens of meters provides relative uncertainty better than 5%.

Borehole Temperature Systems include the following types:

- Four-conductor logging cable from the sensor probe to the surface: Typically, one pair of conductors forms a closed loop while the other pair connects through the sensor probe. This provides for accurate measurement of the resistance of the temperature sensor in the probe by quantifying the contribution of the cable to the total resistance. It requires that the leakage resistance between conductors be effectively infinite, a specification that can be met by manufacturers of geophysical logging cables. The relatively high resistance of thermistors with respect to typical logging cable makes them the most suitable sensors for these systems.

Thermal Conductivity

The International Society for Rock Mechanics and Rock Engineering recommends three different methods for measuring thermal conductivity on samples at atmospheric pressure (Popov et al. 2016): (a) steady state divided bar, (b) transient point-, line-, or plane-source, and (c) optical scanning.

Lodge (1878) originally introduced the concept of the divided bar. The device is essentially a vertical “stack” within which the temperature drop across a standard substance is compared with that across a specimen of unknown thermal conductivity. By design, it measures uniaxial conductivity, so specimens can be prepared to measure conductivity along any

axis of interest. Benfield (1939) and Birch and Clark (1940) described early divided bars at Cambridge and Harvard universities, respectively, with constant heat sources at the top of their stacks and isothermal sinks at their bases. Evolution to a constant temperature difference regulated by water baths in the 1950s (Beck 1957) resulted in more rapid measurements. In the late 1960s, the divided bar was adapted to the measurement of the thermal conductivity of rock fragments, by far the most common source of material (Sass et al. 1971). Substitution of Peltier devices for the constant temperature water baths provided hitherto unattainable portability, economy, and speed of measurement from 2010 (Antriasian 2010; Fig. 2).

With transient methods, heat is applied to a rock at a known rate by an element approximating an ideal point, line, or surface. The rock warms in response to the heater at a rate dependent on its thermal conductivity. Thermal conductivity can be deduced by analyzing the temperature–time curve of the heater element or a nearby point. In general, more conductive rocks draw heat away from the element at a greater rate and result in a slower increase in temperature.

Transient methods can be applied in a permanent laboratory, in mobile field laboratories, or under certain conditions, within boreholes. Jaeger (1959) described the mathematical

framework for measuring thermal conductivity with linear and cylindrical heat sources embedded in materials approximating infinite solids, such as needle-like probes inserted into rocks. Vacquier (1985) extended the method to linear sources applied to the flat surface of an infinite half-space. Compared to the divided bar, transient source methods present difficulties in characterizing uniaxial (e.g., vertical) conductivity because they conduct heat radially, rather than linearly, in all directions away from the source. In homogenous sequences, conductivity is the same in all directions, but the values may be significantly different in heterogeneous sequences. Horai (1971) described the adaption of line-source technique to measure the thermal conductivity of rock fragments. Grubbe et al. (1983) showed that if large enough specimens are available, the principal axial values of conductivity can be derived by making several measurements with the source in different orientations. Sass et al. (1984) described the United States Geological Survey's (USGS) versions of line-source techniques in detail. Transient source units available commercially include QTM-710/700 (KEM undated), TLS-100 (Thermtest undated), and C-Therm TCi (C-Therm Technology 2018).

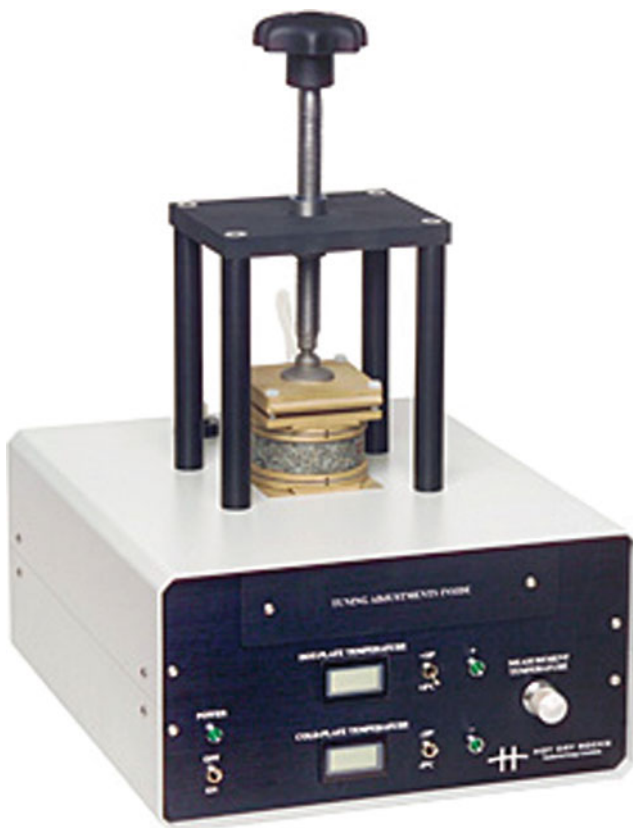
In the 1980s, Popov et al. (1983) described a rapid, non-destructive method for measuring the thermal conductivity of rocks along continuous lines using an optical scanning technique. Rock faces are painted black and aligned with known standard materials along a track. A unit containing a laser heat source and infrared temperature sensors moves beneath the specimens at a constant speed, focusing the laser onto the painted surfaces. The radiometer measures the rise in temperature of the specimens at a constant time and distance after the heating due to the applied laser pulse. The temperature rise is related to the thermal conductivity of the specimen.

A three-way comparison between a divided bar operated by the USGS, a line-source half-space apparatus operated by GeoForschungsZentrum (GFZ) in Germany, and an optical scanner operated by Popov in Russia showed close agreement over a variety of crystalline rock types representing a wide range of thermal conductivity values (Popov et al. 1999).

In Situ Thermal Conductivity Measurement in Boreholes

The in situ measurement of thermal conductivity in vertical boreholes generally presents great experimental and logistical difficulties. In anisotropic rocks, it also measures conductivity radially away from the borehole so that independent measurements of anisotropy on core are required to calculate vertical conductivity. Conductivity was measured over a ~2 m interval of the KTB pilot core hole using an in situ cylindrical heat source (Burkhardt et al. 1990) as a demonstration of the feasibility of the method. The organization Posiva Oy in Finland has done much to develop the technique since the late 1990s.

During the boom in geothermal exploration in the mid- to late 1970s, the importance of determining heat flow as well as



Heat Flow Determinations, Continental, Fig. 2 A commercial divided bar utilizing Peltier devices as heat source/sink. (Photo courtesy of Hot Dry Rocks Pty Ltd, Australia)

temperature gradients became apparent. This was coupled with concerns over the security of exploration wells in competitive lease areas. To address these issues, the Geothermal Studies Project of the US Geological Survey developed a method for determining temperature gradient, thermal conductivity, and thus, heat flow in situ over discrete intervals during the drilling process (Sass et al. 1981).

Details of the technology can be found in Sass et al. (1979) with the caveat that the electronics and computer hardware described in that reference have since been superseded.

More recently, Freifeld et al. (2008) demonstrated thermal conductivity profiling of entire boreholes by coupling a down-hole line-source heating loop with a fiber optic DTS system.

Sampling

Thermal conductivity can rarely be practically measured over the full interval of interest. Certain sampling protocols must be devised and followed to estimate the conductivity over significant depth intervals. This may involve:

- Characterizing the conductivity of individual geological formations where these are largely homogenous
- Selecting samples from evenly spaced intervals as a “random” sampling strategy
- Measuring mixtures of chips from sequential depth intervals

Many studies have shown that the standard deviation of conductivity measurements on samples from seemingly homogenous formations can be as high as 30% of the mean. Measurements must be obtained from sufficient samples, therefore, to constrain the mean conductivity of the formation to within an acceptable standard error. Measurements on ten samples of a given homogenous formation should yield a mean conductivity within 10% of its true value.

Conductivity is temperature dependent and values measured in the laboratory should be corrected for in situ conditions especially for very deep wells, and in geothermal areas with high temperature gradients.

When no rock samples are available for measurement of thermal conductivity, estimates can be obtained from empirical relations between conductivity and well-log parameters or by choosing a “flux plate,” a layer of more than a few meters thickness composed of a rock type whose conductivity has been characterized by measurements elsewhere. See Blackwell and Steele (1989) for a review of these methods.

Summary

Heat flowing from the surface of the continents cannot be measured directly. It is calculated as the product of the

temperature gradient (most often measured in boreholes) and the thermal conductivity of the formation within which the temperature gradient was measured. Thermal conductivity can be measured in situ, but by far the majority of measurements are made on rock samples from the borehole, or from nearby outcrops of the relevant formations. To determine heat flow with a relative uncertainty better than 10%, temperature gradient should be determined over a minimum depth interval of tens of meters, with conductivity measured on ten or more samples from within that interval.

Temperatures are usually measured by electronic transducers deployed on multiconductor electrical cables or incorporated into “memory tools” that are lowered into the borehole on steel or alloy wires. Fiber optic cables can also be inserted into boreholes to provide synoptic temperature profiles or a time series of profiles.

A variety of techniques, based on both steady-state and transient heat flow, can be used to characterize the thermal conductivity of rock specimens. Optical scanning methods provide rapid profiles of thermal conductivity estimates along extended scan lines and are valuable reconnaissance tools complementary to divided bar and point- or line-source transient techniques.

The heat flow determined for any given location and depth interval might be strongly influenced by local spatial or temporal thermal disturbances and not represent regional heat flow conditions. Heat flow is most likely to represent regional conditions if determined within boreholes that have had time to thermally re-equilibrate to natural conditions within a volume of rock with relatively uniform properties over space and time. In practice, locations that best approximate such conditions have low topographic relief, minimal surface disturbance over prior decades or centuries, low rates of erosion or deposition, laterally homogenous geology, and minimal groundwater movement.

Cross-References

- Heat Flow, Continental
- Heat Flow, Seafloor: Methods and Observations
- Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
- Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

Bibliography

Antriasian AM (2010) The portable electronic divided bar (PEDB): a tool for measuring thermal conductivity of rock samples. In: Proceedings, World Geothermal Congress 2010, Bali, Indonesia, 25–29 April 2010. 7 pp. <http://www.geothermal-energy.org/pdf/IGAstandard/WGC/2010/1331.pdf>

- Beardsmore GR, Cull JP (2001) *Crustal heat flow: a guide to measurement and modeling*. Cambridge University Press, Cambridge, p 324. <https://doi.org/10.1017/CBO9780511606021>
- Beck AE (1957) A steady-state method for the rapid measurement of the thermal conductivity of rocks. *J Sci Instrum* 34:186–189. <https://doi.org/10.1088/0950-7671/34/5/304>
- Benfield AE (1939) Terrestrial heat flow in Great Britain. *Proc R Soc Lond A Math Phys Sci* 173(955):428–450. Retrieved from <https://www.jstor.org/stable/97402>
- Birch AF, Clark H (1940) The thermal conductivity of rocks and its dependence upon temperature and composition. *Am J Sci* 238(8):529–558. Retrieved from <http://search.proquest.com.ezp.lib.unimelb.edu.au/docview/52921021?accountid=12372>
- Blackwell DD, Steele JL (1989) Thermal conductivity of sedimentary rocks: measurements and significance. In: Naeser ND, McCulloh TH (eds) *Thermal history of sedimentary basins*. Springer, New York, pp 13–36. https://doi.org/10.1007/978-1-4612-3492-0_2
- Burkhardt H, Honarmend H, Pribnow D (1990) First results of thermal conductivity measurements with a borehole tool for great depths. KTB report 90–6a. Springer, New York, pp 245–258
- C-Therm Technology (2018) Simplifying thermal conductivity (k). C-Therm Technologies Ltd, Fredericton, p 6. Retrieved from http://web.archive.org/save/https://ctherm.com/products/tci_thermal_conductivity/
- Doig R, Saull VA, Butler RA (1961) A new borehole thermometer. *J Geophys Res* 66:4263–4264. <https://doi.org/10.1029/JZ066i012p04263>
- Förster A, Schrötter J, Merriam DF, Blackwell DD (1997) Application of optical-fiber temperature logging – an example in a sedimentary environment. *Geophysics* 62:1107–1113. <https://doi.org/10.1190/1.1444211>
- Freifeld BM, Finsterle S, Onstott TC, Toole P, Pratt LM (2008) Ground surface temperature reconstructions: using in situ estimates for thermal conductivity acquired with a fiber-optic distributed thermal perturbation sensor. *Geophys Res Lett* 35:L14309. <https://doi.org/10.1029/2008GL034762>
- Grubbe K, Haenel R, Zoth G (1983) Determination of the vertical component of thermal conductivity by line source methods. *Zeitblatt Geol Paläontol (Teil) 1H(1/2)*:49–56
- Henniges J, Zimmermann G, Büttner G, Schrötter J, Erbas K, Huenges E (2005) Wireline distributed temperature measurements and permanent installations behind casing. In: *Proceedings World Geothermal Congress, Antalya, Turkey, 24–29 April 2005*, 5 pp
- Horai KI (1971) Thermal conductivity of rock-forming minerals. *J Geophys Res* 76:1278–1308. <https://doi.org/10.1029/JB076i005p01278>
- Jaeger JC (1959) The use of complete temperature-time curves for determination of thermal conductivity with particular reference to rocks. *Aust J Phys* 12:203–217. <https://doi.org/10.1071/PH590203>
- KEM (undated) Quick thermal conductivity meter: QTM 710/700. Kyoto Electronics Manufacturing Co. Ltd, Tokyo, p 8. Retrieved from <http://web.archive.org/web/20190606044600/http://www.kyoto-kem.com/en/pdf/catalog/QTM-710.pdf>
- Lodge OJ (1878) On a method of measuring the absolute thermal conductivity of crystals and other rare substances. *London Edinburgh Dublin Philos Mag J Sci* 5:110–117. Retrieved from https://zs.thulb.uni-jena.de/receive/jportal_jpvolume_00058127
- McGee, Thomas D (1988) *Principles and methods of temperature measurement*. Wiley-Interscience, New York, p 608
- Popov YA, Semenov VG, Korostelev VM, Berezin VV (1983) Non-contact evaluation of thermal conductivity of rocks with the aid of a mobile heat source. *Izvestiya Acad Sci USSR Phys Solid Earth* 19(7):563–567
- Popov YA, Pribnow DFC, Sass JH, Williams CF, Burkhardt H (1999) Characterization of rock thermal conductivity by high-resolution optical scanning. *Geothermics* 28:253–276. [https://doi.org/10.1016/S0375-6505\(99\)00007-3](https://doi.org/10.1016/S0375-6505(99)00007-3)
- Popov Y, Beardsmore G, Clauser C, Roy S (2016) ISRM suggested methods for determining thermal properties of rocks from laboratory tests at atmospheric pressure. *Rock Mech Rock Eng* 49(10):4179–4207. <https://doi.org/10.1007/s00603-016-1070-5>
- Sass JH, Lachenbruch AH, Munroe RJ (1971) Thermal conductivity of rocks from measurements on fragments and its application to heat-flow determinations. *J Geophys Res* 76:3391–3401. <https://doi.org/10.1029/JB076i014p03391>
- Sass JH, Kennelly JP Jr, Wendt WE, Moses TH Jr, Ziagos JP (1979) In situ determination of heat flow in unconsolidated sediments. U.S. Geological Survey open-file report. U.S. Geological Survey, pp 79–593, 73pp. Available in digital form from <https://pubs.er.usgs.gov/publication/ofr79593>
- Sass JH, Kennelly JP, Wendt WE, Moses TH Jr, Ziagos JP (1981) In situ determination of heat flow in unconsolidated sediments. *Geophysics* 46:76–83. <https://doi.org/10.1190/1.1441142>
- Sass JH, Kennelly JP Jr, Smith EP, Wendt WE (1984) Laboratory line-source methods for the measurement of thermal conductivity of rocks near room temperature. U.S. Geological Survey open-file report. U.S. Geological Survey, pp 84–91, 20 pp. <https://doi.org/10.3133/ofr8491>
- Selker JS, Thévenaz L, Huwald H, Mallet A, Luxemburg W, van de Giesen N, Stejskal M, Zeman J, Westhoff M, Parlange MB (2006) Distributed fiber-optic temperature sensing for hydrologic systems. *Water Resour Res* 42:W12202. <https://doi.org/10.1029/2006WR005326>
- Suto Y, Sakuma S, Takahashi H, Hatakeyama N, Henfling J (2008) Temperature memory gauge survey and estimation of formation temperature of the USDP-4 conduit hole at Unzen Volcano, Japan. *J Volcanol Geotherm Res* 175(1–2):20–27. <https://doi.org/10.1016/j.jvolgeores.2008.03.034>
- Tew WL, Strouse GF (2001) Standard reference material 1750: standard platinum resistance thermometers, 13.8033 K to 429.7485 K, NIST Special Publication 260–139. U.S. Government Printing Office, Washington, DC, p 37. Specific information on RTD devices is best obtained by entering “resistance temperature detectors” into an Internet Search Engine. One excellent contemporary source is <http://web.archive.org/save/https://www.omega.com/en-us/resources/rtd>
- Thermtest (undated) TLS-100 portable thermal conductivity meter for measurement of soil, rock, concrete, and polymers. Thermtest Thermophysical Instruments, Fredericton, p 4. Retrieved from <https://web.archive.org/web/20190606050255/https://thermtest.com/tls-100>
- Vacquier V (1985) The measurement of thermal conductivity of solids with a transient linear heat source on the plane surface of a poorly conducting body. *Earth Planet Sci Lett* 74(2–3):275–279. [https://doi.org/10.1016/0012-821X\(85\)90027-5](https://doi.org/10.1016/0012-821X(85)90027-5)

Heat Flow, Continental

Paul Morgan
Colorado Geological Survey, Denver, CO, USA

Synonyms

Heat flow; Heat flow density; Heat flux, land; Terrestrial heat flow

Definition

Heat flow	The outward flow of thermal energy by conduction from within the Earth through the solid surface of the Earth.
Continental heat flow	Heat flow from the continental crust or lithosphere, which are those portions of the plates that are not directly formed by seafloor spreading at mid-ocean ridges and are not generally subducted.

Introduction

Heat flow is a fundamental property of the Earth. The Earth is a heat engine that drives plate tectonics and intra-plate processes such as hot-spot volcanism unrelated to plate margins. Heat flow, the outward flow of thermal energy from within the Earth through the solid surface of the Earth, is a measure of the thermal energy that drives this heat engine. In addition, temperature in the Earth, which is closely related to heat flow, is a primary parameter controlling physical properties of materials within the Earth. In general, areas of high heat flow are underlain by higher temperatures, at least at shallow depth, than areas of low heat flow. We seek therefore to know the distribution of heat flow from the surface of the Earth and to understand the factors that control its variations.

The primary factor controlling the distribution of heat flow is the type of lithosphere, continental or oceanic. Continental lithosphere is generally relatively old (>500 Ma) and the dominant mechanism of heat transfer through this lithosphere is conduction. A significant portion of heat flow from this lithosphere is generated by radiogenic heat production in the continental crust. In contrast, oceanic lithosphere is young (<200 Ma), heat is convected into this lithosphere as it is formed at mid-ocean ridges and it has very little internal radiogenic heat production (see ► [“Heat Flow, Seafloor: Methods and Observations”](#)). A variety of geologic processes modify continental lithosphere and most of these processes also modify the thermal regime. In addition, the upper continental crust commonly hosts moving groundwater that redistributes heat in the upper crust. Thus, the long and complex history of continental lithosphere, significant and lateral variability in radiogenic heat generation, and redistribution of heat by groundwater combine to make continental heat flow heterogeneous.

Fourier’s Law

Under conditions of steady state, heat flow, q , is a function of thermal conductivity, K , and temperature, T , given by Fourier’s law:

$$q = -K\nabla T, \quad (1)$$

where ∇ is the gradient operator. This equation assumes that thermal conductivity is isotropic, that is, it does not change with direction. In general, spatial changes in thermal conductivity and anisotropy must be considered. The negative sign in the equation recognizes that heat flows from higher temperatures to lower temperatures and that flow of heat is in the opposite direction to the temperature gradient. On average, Earth heat flow is primarily vertical and the vertical component of heat flow, q_z , is commonly quoted:

$$q_z = -K \frac{\partial T}{\partial z}, \quad (2)$$

where z is depth. In practice, the subscript and negative sign are commonly omitted:

$$q = K \partial T / \partial z. \quad (2a)$$

The working units of q , K , and $\partial T / \partial z$ are mW m^{-2} , $\text{W m}^{-1} \text{ } ^\circ\text{C}^{-1}$ or $\text{W m}^{-1} \text{ K}^{-1}$, and $^\circ\text{C km}^{-1}$, respectively.

Under transient conditions the heat conduction equation in one dimension becomes:

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial z^2}, \quad (3)$$

where t is time and κ is thermal diffusivity, given by $\kappa = K/\rho c$, where ρ is density and c is specific heat. The general working units for t and κ are s and $\text{mm}^2 \text{ s}^{-1}$, although Ma and $\text{km}^2 \text{ Ma}^{-1}$ are also sometimes used.

If heat generation, A , is also included, the heat flow equation becomes:

$$\frac{\partial^2 T}{\partial z^2} - \frac{1}{\kappa} \frac{\partial T}{\partial t} + \frac{A}{K} = 0. \quad (4)$$

The working units for radiogenic heat generation, A , are $\mu\text{W m}^{-3}$.

Geothermal Gradient

The rate of increase in temperature with depth, $\partial T / \partial z$, is the *geothermal gradient*. This term is usually reserved for zones over which heat is transferred by conduction and the rate of increase in temperature with depth is linear (see ► [“Heat Flow Determinations, Continental”](#)), but if detailed temperature data are not available or gradients are estimated from oil-well bottom-hole temperatures, an approximate average geothermal gradient may be given. The profile of temperature plotted as a function of depth is the *geotherm*.

Geothermal gradients range from very low to near zero in fore-arc regions, the zones between subduction zones and their associated volcanic belts where heat is convected downward by the subducting lithosphere. Similarly, low gradients are found in zones of groundwater recharge. Gradients in excess of $20,000\text{ }^{\circ}\text{C km}^{-1}$ have been measured at one location in the upper few meters of the sediments beneath Yellowstone Lake inside the caldera of Yellowstone National Park, Wyoming, USA (Morgan et al. 1977), but such high gradients are buffered by the boiling point curve of water as soon as temperatures reach the boiling temperature at depth. Thus, there is a very wide range of geothermal gradients in the upper continental crust. Temperatures are ultimately controlled by the boiling point curve of water in the uppermost crust and then the solidus curves of the crust and mantle lithosphere, and temperatures at depth generally do not range as widely as shallow geothermal gradients might suggest. Thermal conductivity varies by a factor of about 5 in the main rock types that comprise the continental crust (see ► “Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat”): there is some variation in heat flow associated with changes in thermal conductivity, but the most significant variations in heat flow are accompanied by changes in geothermal gradient.

Sources of Heat

There are two sources of heat contributing to continental heat flow, heat that enters continental lithosphere from below, and heat generated within the continental lithosphere (see ► “Energy Budget of the Earth”). Although the Earth beneath the lithosphere has thermal energy that must ultimately be lost to the surface either through the lithosphere or in the process of the formation of new oceanic lithosphere at mid-oceanic ridges, there is no requirement where or when that heat must be lost, that is, there is no constant basal heat flux that flows into continental lithosphere. The thermal boundary condition at the base of the lithosphere is probably more accurately described as a constant temperature boundary, and this constant temperature may be maintained either by small-scale convection and/or lateral flow in the asthenosphere. Heat flow into the base of the lithosphere is then determined by the thermal conductivity structure of the lithosphere, the thickness of the lithosphere, heat production within the lithosphere, and its surface temperature. The resulting thermal profile is disrupted during major deformation of the lithosphere or changes in its boundary conditions.

By far the largest source of heat within the lithosphere is radiogenic heat production from the decay series of the unstable isotopes ^{232}Th , $^{235,238}\text{U}$, and ^{40}K . These isotopes all have half-lives of the same order as the age of the Earth and thus remain sufficiently abundant to be significant and continue to

decay at a rate that is significant in terms of heat production (see ► “Radiogenic Heat Production of Rocks”). Other radiogenic isotopes are important for age dating, such as ^{87}Rb , ^{147}Sm , and ^{187}Re (see ► “Absolute Age Determinations: Radiometric”), but the heat produced by the decay of these isotopes and other isotopes used for age dating is insignificant. Uranium, thorium, and potassium are incompatible elements and tend to be concentrated in silicic magmas. Hence, they are depleted in the mantle during crustal formation and generally concentrated in the crust in silicic rocks. Heat flow data show that their concentrations in large silicic plutons are relatively predictable from surface measurements (Roy et al. 1968), but their distributions are less predictable in metamorphic and sedimentary terrains (see ► “Lithosphere, Continental: Thermal Structure”). Measurements indicate that from as little as less than 20% to as much as 80% of heat flow may be derived from crustal radiogenic heat generation. Typical values of heat production for crustal rocks range from $<0.1\text{ }\mu\text{W m}^{-3}$ for basic igneous rocks to $>6\text{ }\mu\text{W m}^{-3}$ for radiogenic granites.

Other crustal heat sources include frictional heat from faults and heat from chemical reactions. Attempts to measure heat associated with active faults have generally failed to detect heat or produced inconclusive results. The primary chemical reaction that has been observed as a heat source is the oxidation of pyrite, but any exothermic metamorphic or weathering reaction is a potential heat source. However, friction and chemical heat sources are likely to be only local in extent and minor in terms of the total continental thermal budget.

Mechanisms of Heat Transport

In stable continental crust, the primary mechanism of heat transfer is conduction as described by Fourier’s law (Eqs. 1 and 2). In the mantle, the effects of radiative heat transfer may be significant as olivine has a significant component of radiative heat transfer above a temperature of about $500\text{ }^{\circ}\text{C}$. This radiative component may be included in Fourier’s law by substituting a temperature-dependent value of thermal conductivity that includes a correction for radiative heat transfer.

Even in stable regions, heat may be redistributed in the upper few kilometers by groundwater flow. In stable regions the geothermal gradient is seldom high enough to drive free convection (see below), and groundwater flow is driven by differences in elevation of the piezometric surface (water table). Heat transferred by the forced vertical flow of groundwater, q_{gw} , may be approximated by:

$$q_{gw} = \bar{v}c_w |dT|, \quad (5)$$

where $\bar{v}$ is the equivalent vertical velocity if the porosity were 100%, c_w is the specific heat of water, and $|dT|$ is the absolute temperature drop over the vertical extent of the water flow. Heat transfer will be in the direction of water flow and heat flow is decreased in areas of recharge and increased in areas of discharge. This type of convection is termed forced convection or advection.

If the geothermal gradient is sufficiently high, water flow may be driven by thermal buoyancy forces. The magnitude of the gradient required for the onset for thermal convection depends on the geometry of the heat source and the medium in which convection occurs (e.g., Turcotte and Schubert 2002, p. 393 et seq.). For a homogeneous porous layer of permeability km^2 and thickness b km heated uniformly from below, the minimum thermal gradient, $(\partial T/\partial z)_{\min}$, required for convection to occur is given by:

$$\left(\frac{\partial T}{\partial z}\right)_{\min} = \frac{4.2 \times 10^{-10}}{kb^2}. \quad (6)$$

This minimum gradient may not sustain convection and a higher gradient may be required for continuous convection. Convection driven by thermal buoyancy forces is termed free convection.

Heat will also be convected when there are net vertical movements within the lithosphere associated with deformation or changes in either the top or bottom boundaries of the lithosphere. In general, extensional deformation causes thinning of the lithosphere resulting in a net upward movement of the material in the lithosphere relative to its surface and a compression of the isotherms, or horizons of equal temperature within the lithosphere (Fig. 1a). In contrast, compressional deformation causes thickening of the lithosphere resulting in a net downward movement of the material in the lithosphere relative to its surface and an increase of the spacing of the isotherms within the lithosphere (Fig. 1b). For simplicity, deformation in Fig. 1 is shown to be 100% by pure shear and uniformly distributed throughout the lithosphere, a highly unlikely distribution of geological strain. As shear strength changes from brittle to ductile laterally and with depth throughout the lithosphere, the strain distribution is always much more complex: zones of simple shear, imbricated layers, and other complexities typically characterize lithospheric deformation. However, at a lithospheric scale, the net material movements and compression and extension of the geotherms illustrated in Fig. 1 are valid.

Stable continental lithosphere has a relatively low heat flow corresponding to a low geothermal gradient and strong lithosphere. These conditions are not favorable for deformation. Observations indicate that deformation commonly occurs in lithosphere that is still recovering from a previous thermal event or has experienced a preceding thermal event. This thermal event may be manifested in the form of

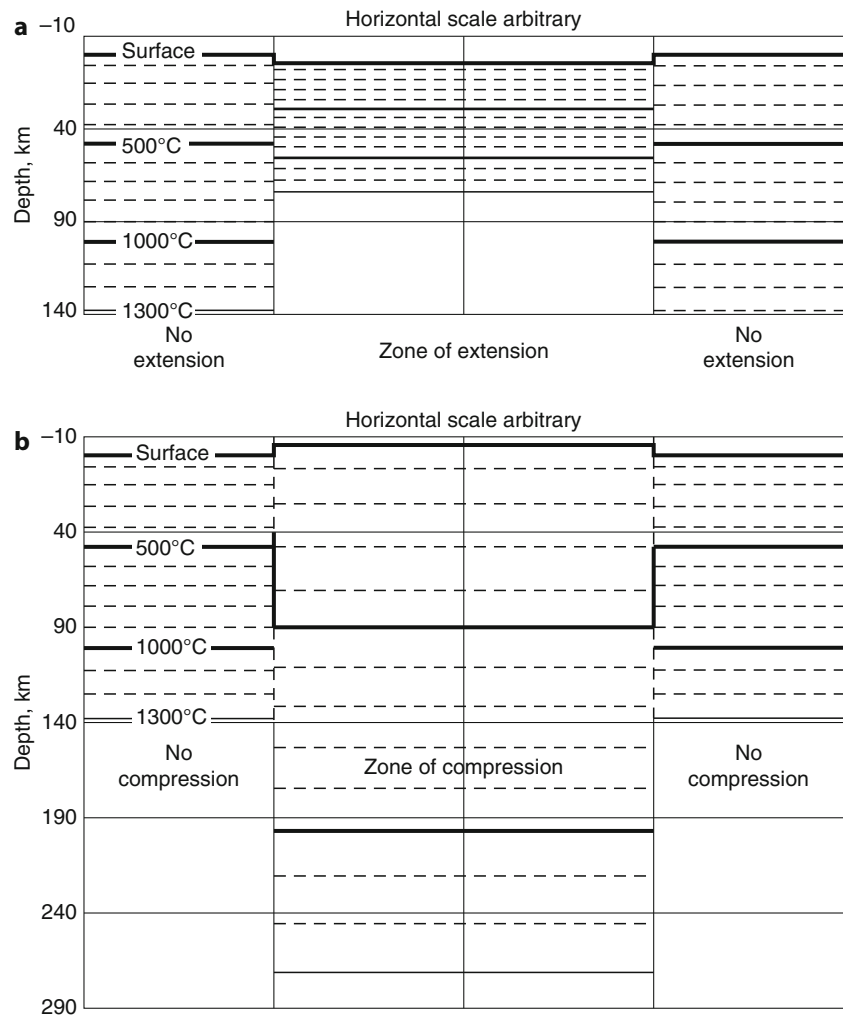
volcanism in which heat is advected in the lithosphere by ascending magmas. Heat transfer by magmas is conceptually similar to heat transfer by groundwater described by Eq. 5, if the specific heat of magma is substituted for the specific heat of water. Additional heat is released from a magma during solidification, and this heat is represented by the latent heat of solidification. Many analytical solutions for conductive modeling of magma intrusions may be found in Carslaw and Jaeger (1959) and other references, and these processes may also be modeled numerically. Magma generation may be initiated from below the lithosphere by the ascent of hot material to the base of the lithosphere (mantle plumes?) generating magmas by pressure-release melting and raising the temperature at the lower boundary of the lithosphere. Additionally magmas may be generated within the lithosphere by pressure-release melting and the melting of lower-melting temperature components within the lithosphere.

Once deformation is initiated additional magma generation will be initiated. The net upward movement of material during extension may trigger pressure-release melting, especially if the lithosphere has been preheated by pre-deformation magmatism. During compression, there is a net syn-deformational lowering of the geotherm. However, thickening of the crustal layer of radiogenic heat-producing elements will generally cause the geothermal gradient to return to a higher value than its initial value. This higher geotherm may result in melting of the thickened crust and crustal derived magmatism, the results of which are commonly observed in belts of compressional deformation. Thus, magmatic convection of heat is an important heat transfer mechanism that commonly accompanies heat convection by tectonic deformation.

Finally, there is heat convection associated with changes at the Earth's surface. In discussing deformation and magmatism, the effects of these processes on surface elevation were ignored. Cooling oceanic lithosphere as it ages away from oceanic ridges results in a relatively simple change in depth of the ocean floor (see ► [“Heat Flow, Seafloor: Methods and Observations”](#)). A similar, but more complex relation exists for the thermal and density structure of continental lithosphere (Lachenbruch and Morgan 1990; ► [“Isostasy, Thermal”](#)). Changes in surface elevation are generally accompanied by either erosion or sedimentation with the net result that the material in the lithosphere has a net movement either up or down relative to the surface, respectively. If the rates of erosion or sedimentation are known, a correction may be applied for these effects. If erosion creates three-dimensional topography, a terrain correction may also be applied. Not strictly a heat transfer mechanism, but changes in surface temperature associated with past climate change or changes in land use may also change surface heat flow and a correction may be applied for these perturbations.

Heat Flow, Continental,

Fig. 1 Diagrammatic examples of the effects of instantaneous pure shear deformation on lithospheric isotherms. Both diagrams have the same thermal conditions in undeformed lithosphere (no extension and no compression) calculated for a 35 km thick crust with a surface heat flow of 40 mW m^{-2} , a surface heat production of $2 \mu\text{W m}^{-3}$, and a crustal thermal conductivity of $2.5 \text{ W m}^{-1} \text{ K}^{-1}$. The crustal heat production is assumed to decay exponentially with depth with a depth parameter of 10 km, and the mantle is assumed to have a heat production of $0.084 \mu\text{W m}^{-3}$ and a temperature-dependent thermal conductivity appropriate for olivine. The base of the lithosphere is assumed to be the $1,300^\circ\text{C}$ isotherm, (a) shows the effect of thinning the lithosphere by a factor of 2 by extending the lithosphere by a factor of 2. (b) shows the effect of doubling the thickness of the lithosphere by compressing the lithosphere by a factor of 2



The duration of a thermal disturbance associated with thermal convection in the lithosphere depends primarily on the stable thickness of the lithosphere and its thermal diffusivity. There is no standard for defining this duration as temperatures relax asymptotically to their stable values, and the form of this return depends on the structure of the thermal disturbance, which may be highly variable and very complex in continental deformation with magmatism. However, a useful guideline for the duration of the thermal disturbance, τ , is:

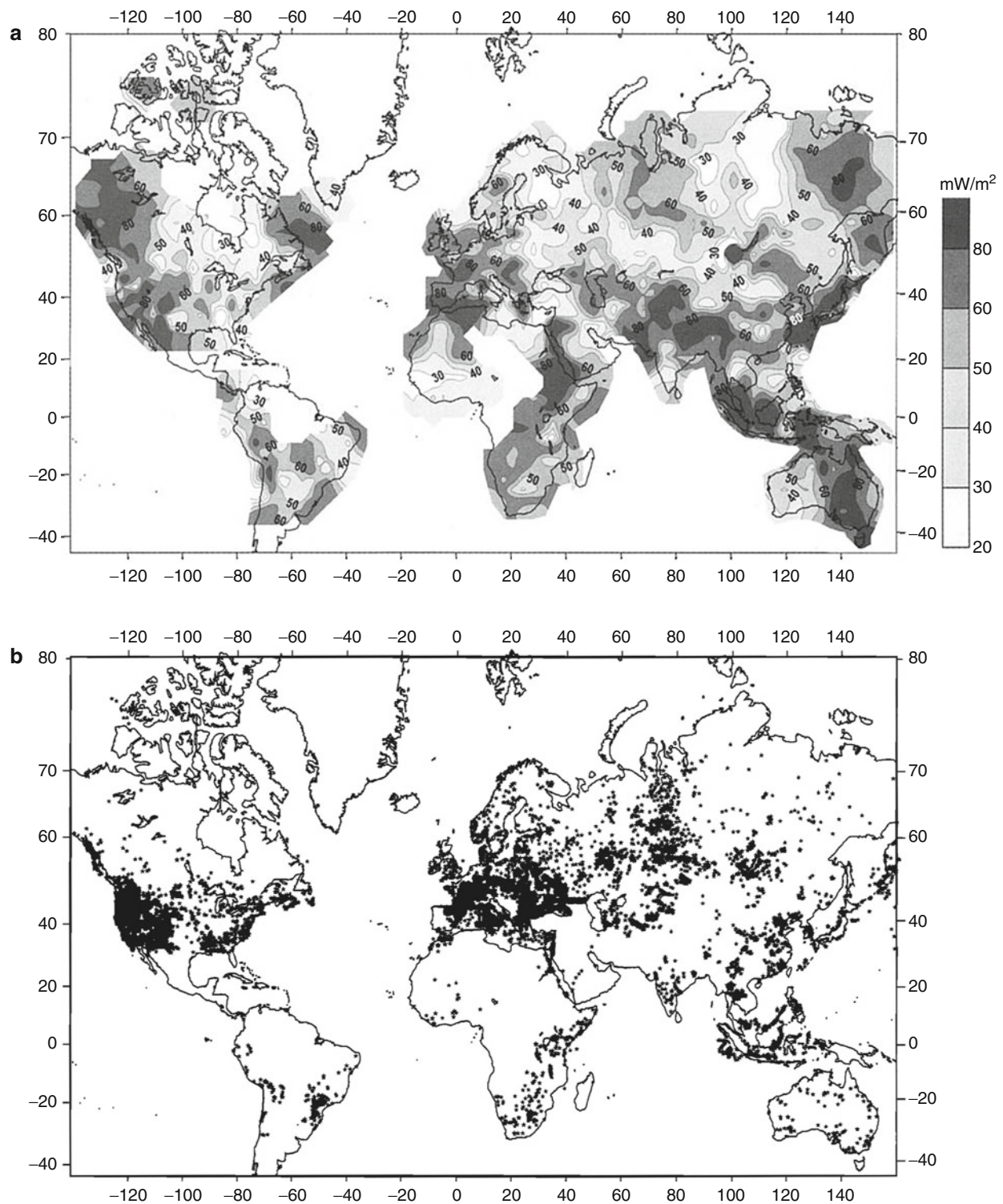
$$\tau = \frac{L^2}{4\kappa} \quad (7)$$

where L is the stable thickness of the lithosphere and κ is its thermal diffusivity. For example, using an average thermal diffusivity of $32 \text{ km}^2 \text{ Ma}^{-1}$ and a thickness of 100 km for stable oceanic lithosphere, the calculated duration of the thermal disturbance is about 78 Ma. Oceanic crust is observed to flatten at about 70–80 Ma. For continental lithosphere, if the stable thickness of this lithosphere is 100 km, the duration

of the thermal disturbance would be about 78 Ma; if the stable thickness were 150 km, the duration would be ~175 Ma; if the thickness were 200 km, the duration would be ~310 Ma; and if the stable thickness were 250 km, the duration would be ~500 Ma. These durations are from the start of stabilization: if a deformation and magmatism event were followed by significant erosion or sedimentation, the duration of the thermal stabilization would not begin until the end of the erosion or sedimentation. However, as the effects of sedimentation and erosion are relatively minor, their measurable effects would probably be dissipated before the end of the formal disturbance duration.

Global Distribution of Continental Heat Flow

The global distribution of continental heat flow is shown in Fig. 2a. The density of data is far from uniform (Fig. 2b), but there are sufficient data that some generalizations about the heat flow distribution may be made. In general, high heat



Heat Flow, Continental, Fig. 2 (a) Distribution of continental heat flow, based on the data of Pollack et al. (1993) and subsequent updates, presented with a $10^\circ \times 10^\circ$ interpolation (Artemieva and Mooney 2001,

Fig. 3). (b) Locations of continental heat flow measurements used in constructing Fig. 2a. (From Artemieva and Mooney 2001, Fig. 4)

flow corresponds with zones of active seismicity, which in turn correlate with plate margins. Where the zones of seismicity are diffuse, such as the western US and the Himalaya,

the plate margin deformation is distributed and the zones of high heat flow are extensive. Some portions of active plate margins are associated with low heat flow, such as the

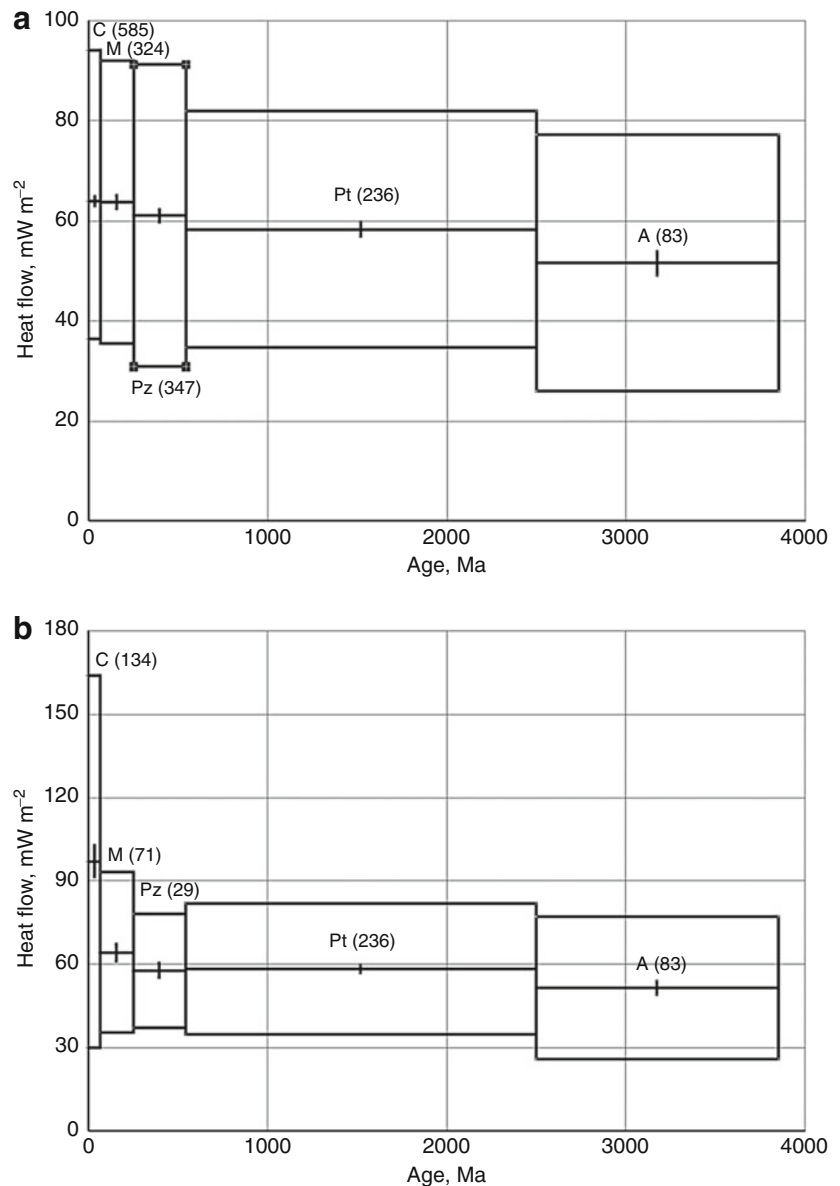
western margin of South America and the Pacific Northwest of the USA, west of the volcanic Cascade Range. These areas overlie subducting oceanic lithosphere where heat is being convected down into the mantle. Associated with these low heat flow zones is a rapid lateral transition to high heat flow in the active volcanic zones of the subduction zones. In addition, plate margins are generally associated with elevated topography, which received more precipitation than low elevations and in which heat flow is generally more variable because of redistribution of heat by forced and free groundwater convection.

The association of high and low heat flow with active plate margins is demonstrated in a plot of mean heat flow as a function of age of heat flow site (Fig. 3). Data are shown for sedimentary and metamorphic sites in Fig. 3a and igneous

sites in Fig. 3b. The data for sedimentary and metamorphic sites (Fig. 3a) show a small decrease from Cenozoic and Mesozoic to Paleozoic sites and from Paleozoic to Proterozoic sites, and a larger decrease from Proterozoic to Archean sites. This result suggests that the plate margin disturbances are relatively minor in the Phanerozoic sedimentary and metamorphic site data. In the igneous site data (Fig. 3b), however, there is a significant decrease in both the mean and standard deviation from both Cenozoic to Mesozoic and from Mesozoic to Paleozoic sites. There is almost no difference in the mean from Paleozoic to Proterozoic sites. The data are the same for Proterozoic and Archean sites as Fig. 3a. The igneous site data clearly show thermal perturbations with ages extending into Mesozoic age sites relative to older sites. This result is consistent with a continental lithospheric thickness of

Heat Flow, Continental,

Fig. 3 Continental heat flow as a function of age of data site. Widths of boxes and bars indicate age range. Heights of boxes indicate $\pm$ one standard deviation about the mean; short vertical line indicates $\pm$ one standard error about the mean. Ages of sites are grouped as C – Cenozoic, M – Mesozoic, Pz – Paleozoic, Pt – Proterozoic, and A – Archean. Numbers in parentheses are the number of 1° by 1° blocks represented in each age group. (a) shows sedimentary and metamorphic sites for Phanerozoic data. (b) shows igneous sites for Phanerozoic data. (Data re-plotted from Pollack et al. (1993))



the order of 200 km. The observation that the igneous site data show a perturbation at young ages but there is only a small signal in the sedimentary and metamorphic site data suggests that the sedimentary and metamorphic site data are dominated by sedimentary sites and that these sites are either distant from plate boundaries or high heat flow is suppressed at these sites by downward advection of heat by advection during sedimentation.

The difference between the mean heat flow at Proterozoic and Archean sites has been commented upon from previous compilations of global heat flow data (Morgan 1985) in addition to this data set (Artemieva and Mooney 2001). To explain this difference as a transient effect would require a lithosphere of the order of 500–600 km in thickness, which is probably unrealistic. However, a mean heat flow difference of 7 mW m^{-2} may be explained in terms of a difference in the contributions of the mean heat production of Proterozoic and Archean crusts to their heat flows. Assuming an average crustal thickness of 35 km, an average difference of $0.2 \text{ } \mu\text{W m}^{-3}$ would explain the difference between the mean measured heat flow at Archean and Proterozoic sites. This difference is not universally observed in all areas (Jaupart and Mareschal 1999; ► “Lithosphere, Continental: Thermal Structure”), but it is consistent with bulk observations of the differences in chemistry between the upper crust of Proterozoic and Archean terrains (McLennan et al. 2006). Archean upper crust is statistically depleted in the heat-producing elements relative to younger crust. In addition, xenolith data indicate that Archean lithosphere is statistically distinct from younger lithosphere so the upper crustal thermal observation may indicate that Archean lithosphere is statistically thermally distinct (Griffin et al. 2009).

In summary, the global distribution of continental heat flow is complex. Neglecting measurement errors (see ► “Heat Flow Determinations, Continental”), variations associated with surface temperature changes, uncorrected topographic effects, and anomalies induced by changes in surface cover, heat flow may be described as steady state or transient. At any site, near-surface heat flow may be redistributed by groundwater flow, but this is most likely at sites with high elevation and rugged topography that are likely to have experienced relatively young (Cenozoic or Mesozoic) mountain building and/or igneous activity. Transient thermal perturbations in the lithosphere associated with tectonic deformation and/or igneous activity will dissipate with time and are likely to be restricted to heat flow sites in lithosphere with Cenozoic or Mesozoic tectonic or thermal ages. To distinguish tectonic and igneous ages from sedimentary ages, the term *tectonothermal age* is commonly used to describe the age of heat flow sites. Data from sites with tectonothermal ages Paleozoic and older are generally expected to be in steady state unless erosion or sedimentation are significant.

Igneous activity always heats continental lithosphere but deformation may either raise or lower the geothermal gradient (Fig. 1). There is therefore no predictable anomaly for continental lithosphere at or near plate margins. The wide range in observed heat flow measured at young sites is in part associated with redistribution of heat by groundwater flow, but also associated with the real range in anomalies, both positive and negative, generated by deformation.

Whether the heat flow and geotherm be transient or steady state, significant variation among sites may occur associated with crustal radiogenic heat generation. Heat generation in the upper crust is very variable being generally low in mafic rocks and high in silicic rocks. Its contribution may vary from <20% to >70% of heat flow, and much of the range in heat flow represented in the standard deviations about the means of heat flow in Fig. 3 are probably associated with chemical heterogeneity in the continental crust resulting in lateral variations in the contributions of radiogenic heat production to heat flow.

Paleo-Heat Flow

The discussion above has been restricted to heat flow measured at present although there is some lag in heat flow measured at the surface as it is conducted from depth: the order of that time lag may be estimated from Eq. 7. This heat flow and associated temperatures are appropriate for most use for geophysical parameters, such as seismic wave velocities, electrical conductivities, and density calculations, but the time lag may be significant in some area where temperatures at depth in the lithosphere are changing. However, for some geological observations, including most techniques of radiometric age dating, the former temperature of a rock or mineral are preserved in a chemical or radiometric system when the rock cools as it is brought to the surface. This process may allow the temperature and sometimes the geothermal gradient of a site to be reconstructed at the time the rock was uplifted and this general topic is sometimes referred to as paleo-heat flow.

The rates of most chemical reactions increase at higher temperatures. For example, the maturation of organic material in sediments roughly follows the Arrhenius' equation, which describes the velocity of the reaction, r , as:

$$r = \lambda \exp(-E/RT_K) \quad (8)$$

where λ is the frequency factor, E is the activation energy, R is the gas constant, and T_K is temperature in Kelvin (Stegenga 1988). The combination of the frequency factor and temperature gives a time-temperature index that, together with composition of the organic matter, provides a useful predictor of hydrocarbon maturation. Temperatures in sedimentary basins

are usually estimated by the increase in coal rank of carbonaceous matter in the sediments as measured by its vitrinite reflectance (e.g., Suggate 1998), and time may be estimated by the sedimentation history of the basin. Vitrinite reflectance measures the maximum temperatures attained by the sediments, which may exceed modern temperatures.

Radiometric methods for age dating (see ► [“Absolute Age Determinations: Radiometric”](#)) do not determine rock ages but the time at which the radiometric system on which the age determination is based cools through its closure temperature, the temperature at which the age is set. Thus, radiometric ages may be used for the additional purpose of determining the thermal history of a rock, especially if some estimate of the depth at which the age was set may be determined (see below). Radiometric systems do not close at a single temperature but follow the Arrhenius’ equation; at high temperatures cooling is generally relatively rapid. Examples of approximate mineral-closure temperatures for the uranium-lead method are monazite >1,000 °C, zircon >1,000 °C, titanite 650–600 °C, apatite 500–450 °C, and rutile 450–400 °C (Flowers 2005). Examples of approximate mineral-closure temperatures for the potassium-argon system are hornblende 530 ± 40 °C, muscovite ~350 °C, and biotite 280 ± 40 °C (McDougall and Harrison 1999). Different dates from these minerals may be used to calculate rates of erosion or exhumation of terrains. However, young ages may also indicate reheating for minerals with relatively low-temperature closure temperatures.

The minerals used for some radiogenic systems have geologically very low closure temperatures. For example, the fission track technique is based on radiation damage (tracks) in different minerals caused by the spontaneous fission of ^{238}U (Garver 2008). Closure of this system is by thermal annealing of the fission tracks, which occurs at >120 °C for apatite, ~200 °C for zircon, and ~300 °C for sphene. At these low temperatures, a geothermal gradient is usually assumed and the fission track ages are commonly used to indicate recent uplift/erosion rates (uplift/erosion rate = closure temperature/[fission track age $\times$ assumed geothermal gradient]).

Finally, paleo-thermal conditions are preserved by metamorphism, and in particular in mineral systems called geothermometers. Geothermometers are typically pairs of minerals in which concentration of one element in one of the minerals has a strong temperature dependence but is essentially pressure independent; the other mineral, which must be in intimate contact with the first mineral grain, acts as a reservoir for this element and allows the first mineral to be in thermal equilibrium with the element. Geothermometers tend to equilibrate rapidly at high temperatures but re-equilibrate slowly as the temperature decreases, so if cooling and exposure at the surface is relatively rapid, the geothermometer records peak thermal conditions.

Geobarometers also exist in which the concentration of an element in the first mineral is strongly pressure (depth) dependent but insensitive to temperature. If mineral pairs representing a geothermometer and a geobarometer occur in the same outcrop, then temperature and depth may be estimated allowing calculation of the paleo-geothermal gradient and heat flow at the time of peak thermal conditions.

Studies of paleo-thermal gradients from metamorphic terrains suggest that, on average, geothermal gradients were higher than modern average geothermal gradients. However, although global heat flow was almost certainly higher in the past than at present (see ► [“Energy Budget of the Earth”](#)), these observations do not demonstrate that conclusion for the following reasons: (1) only peak or near-peak thermal conditions are preserved resulting in an overestimation of average continental heat flow; (2) peak preserved geothermal gradients are no higher than geothermal gradients measured today – modern geothermal gradients are buffered in the uppermost crust by the water boiling point curve and at greater depths by the crust solidus, and the same conditions would have applied in the earliest Earth preserved in the continental crust; and (3) if there were zones of low temperature/high pressure metamorphism (blueschist) in the Archean/Early Proterozoic, they are unlikely to be preserved because a geotherm required to maintain these metamorphic conditions requires active subduction. With the cessation of subduction, the geotherm would relax to a warmer geotherm, and blueschist metamorphism is likely to be overprinted. Similarly, low temperatures would not be recorded by geothermometers as the lithospheric temperatures increase.

Summary

Continental heat flow is a complex, but fundamental parameter that, together with oceanic heat flow, constrains the energy budget of the Earth heat engine. It is particularly complex because continents are complex and heterogeneous. It has components from beneath the lithosphere and radiogenic heat production within the lithosphere. Heat is advected into and within the lithosphere by magmatic activity, and heat is redistributed by lithospheric deformation associated with plate tectonics and at relatively shallow depths by groundwater flow. Changes in surface conditions also modify heat flow in the shallow crust. The time constant of thermal relaxation of continental lithosphere is of the order of 300 Ma for stable lithosphere 200 km in thickness. Typically heat flow data from igneous sites of Cenozoic and Early Mesozoic ages reflect transient thermal conditions; data from sites of older ages do not. However, continental heat flow remains very variable even at Precambrian sites because of lateral variability in crustal radiogenic heat production.

Cross-References

- [Absolute Age Determinations: Radiometric](#)
- [Energy Budget of the Earth](#)
- [Heat Flow, Continental](#)
- [Isostasy](#)
- [Lithosphere, Continental: Thermal Structure](#)
- [Lithosphere, Oceanic: Thermal Structure](#)
- [Radiogenic Heat Production of Rocks](#)
- [Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat](#)

Bibliography

- Artemieva IM, Mooney WD (2001) Thermal thickness and evolution of Precambrian lithosphere. *J Geophys Res* 106:16387–16414
- Carslaw HS, Jaeger JC (1959) Heat conduction in solids. Oxford University Press, Oxford, 510. pp
- Flowers RM (2005) Tempo of burial and exhumation within the deep roots of a magmatic arc, Fiordland, New Zealand. *Geology* 33:17–20
- Garver JI (2008) Fission-track dating. In: Gornitz V (ed) *Encyclopedia of paleoclimatology and ancient environments*. Springer, Dordrecht, pp 247–249
- Griffin WL, O'Reilly SY, Afonso JC, Begg GC (2009) The composition and evolution of lithospheric mantle: a reevaluation and its tectonic implications. *J Petrol* 50:1185–1204
- Jaupart C, Mareschal J-C (1999) The thermal structure and thickness of continental roots. *Lithos* 48:93–114
- Lachenbruch AH, Morgan P (1990) Continental extension, magmatism and elevation; formal relations and rules of thumb. *Tectonophysics* 174:39–62
- McDougall I, Harrison TM (1999) *Geochronology and thermochronology by the $^{40}\text{Ar}/^{39}\text{Ar}$ method*. Oxford University Press, New York
- McLennan SM, Taylor SR, Hemming SR (2006) Composition, differentiation, and evolution of continental crust: constraints from sedimentary rocks and heat flow. In: Brown M, Rusher T (eds) *Evolution and differentiation of continental crust*. Cambridge University Press, New York, pp 92–134
- Morgan P (1985) Crustal radiogenic heat production and the selective survival of ancient continental crust. *J Geophys Res* 90(Suppl): C561–C570
- Morgan P, Blackwell DD, Spafford RE, Smith RB (1977) Heat flow measurements in Yellowstone Lake and the thermal structure of the Yellowstone Caldera. *J Geophys Res* 82:3719–3732
- Pollack HN, Hurter SJ, Johnson JR (1993) Heat flow from the Earth's interior: analysis of the global data set. *Rev Geophys* 31:267–280
- Roy RF, Blackwell DD, Birch F (1968) Heat generation of plutonic rocks and continental heat flow provinces. *Earth Planet Sci Lett* 5:1–12
- Stegena L (1988) Paleogeothermics. In: Haenel R, Rybach L, Stegena L (eds) *Handbook of terrestrial heat-flow density determination*. Kluwer, Dordrecht, pp 391–419
- Suggate RP (1998) Relations between depth of burial, vitrinite reflectance and geothermal gradient. *J Pet Geol* 21:5–32
- Turcotte DL, Schubert G (2002) *Geodynamics*. Cambridge University Press, Cambridge

Heat Flow, Seafloor: Methods and Observations

Earl E. Davis¹ and Andrew T. Fisher²

¹Pacific Geoscience Centre, Geological Survey of Canada, Sidney, BC, Canada

²Department of Earth and Planetary Sciences, University of California at Santa Cruz, Santa Cruz, CA, USA

Definition

Conductive heat flux	The rate of heat transfer through a substance per unit area. The conductive heat flux, commonly defined in one dimension, is the product of the thermal gradient and thermal conductivity, with SI units of W m^{-2} . Heat flux is referred to in historical and some current oceanographic literature as “heat flow,” but in precise physical terms, heat flux equals heat flow density.
Advective heat flux	The rate of heat transfer per unit area as a consequence of material motion. The advective heat flux is commonly defined as the product of the temperature, heat capacity, and material velocity, with SI units of W m^{-2} .
Thermal conductivity	The constant of proportionality that relates the thermal gradient to the conductive heat flux, indicating the ability of a medium to transfer heat by conduction. SI units are in $\text{W m}^{-1} \text{K}^{-1}$.
Hydrothermal circulation	Large-scale pore-fluid convection driven by geothermal buoyancy, with flow being strongly influenced by permeability structure of the host formation.

History of Observations

Pioneering measurements of temperature below the seafloor (e.g., Petterson 1949; Revelle and Maxwell 1952; Bullard 1954) were made to compare the thermal state of the ocean crust with that of continents and thus to improve the knowledge of the present-day heat loss from the Earth. The early data demonstrated that gravity-driven probes and corers that penetrated just a few meters into seafloor sediments could provide meaningful measurements of geothermal gradients, and they thus established the foundation for the marine heat flow discipline. Methods for measuring the subseafloor

thermal gradient and thermal conductivity improved in subsequent years, the number and geographic distribution of determinations increased, and patterns of seafloor heat flux were gradually revealed. Initially, the average seafloor heat flux appeared to be similar to that through continents, despite the contribution of radiogenic heat produced by the continental crust that was not present in the ocean crust. Heat-flux values over mid-ocean ridges were found to be significantly higher than in the flanking basins, and locally values were often inexplicably scattered (Von Herzen and Uyeda 1963; Lee and Uyeda 1965). Higher heat flux at mid-ocean ridges was consistent with emerging ideas about seafloor spreading, although the values measured were lower than expected from early theoretical models for the formation of ocean lithosphere. For nearly two decades, regionally low and scattered seafloor heat-flux values remained unexplained, although there was early speculation about the possibility that fluids flowing through the seafloor might influence heat loss from the ocean crust (Elder 1965; Langseth et al. 1966).

By the 1970s, heat flow surveys with more closely spaced measurements, made in the context of local sediment thickness and igneous crustal structure, helped to confirm the hypothesis that hydrothermal circulation in the igneous crust caused both the scatter and the lower-than-expected values in young areas (Lister 1972). Further improvements to instrumentation and observational strategies, in particular the development of probes that could be used with great efficiency for multiple measurements during a single instrument lowering, and the practice of making measurements in the context of geologic structure, led to the use of heat flow as a tool to study the process of hydrothermal circulation itself (Williams et al. 1974). With this new knowledge, it became possible to decipher the local and regional variability of measurements in a way that could lead to a better quantification of deep-seated heat flux, the goal of the original marine heat flow studies, and to understand the hydrologic processes behind the perturbations. This led to studies across a range of spatial and temporal scales using marine heat flow and complementary measurements to solve numerous problems. A summary of the suite of tools currently in use for these studies is provided in the next section, along with a brief description of how heat-flux determinations are made. This is followed by a few examples of data from specific studies that illustrate how data are used, and a summary of some of the major conclusions that have been made through such studies.

Methods

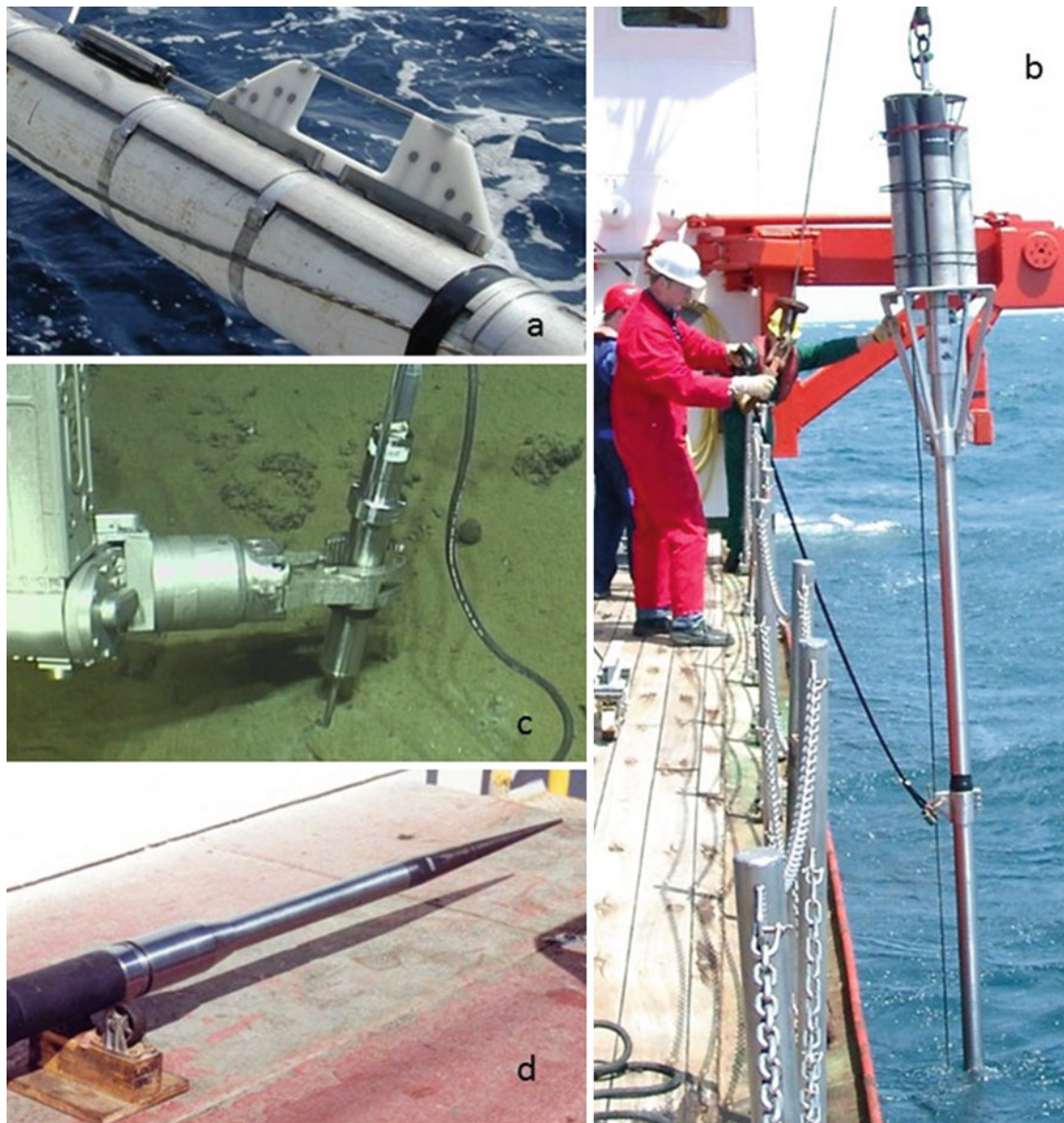
Shallow Measurements in Marine Sediments

Heat flux through the seafloor is often determined using the combination of temperatures measured with a series of

sensors mounted on the outside of gravity-driven corers (Fig. 1a) and thermal conductivities measured on the recovered sediment cores. Depths of penetration in excess of 10 m can be achieved in soft sediment, providing a valuable check on potential perturbations from bottom-water temperature variations (discussed below), although accuracy is often limited by physical disturbances caused by the coring process, by changes in the physical properties of the recovered material, and by the imperfect depth registration between the cores and the intervals between the temperature sensors (Fisher et al. 2002). Probes devoted exclusively to heat-flux measurements are typically limited to lengths of a few meters, but they have several distinct advantages. They allow thermal conductivity to be measured under in situ conditions and at depths that are co-registered with the sensors used to determine the gradient, and they allow transects of many measurements to be made efficiently during a single instrument lowering. A typical heat-flow probe (Fig. 1b) employs a heavy strength member that resists bending during repeated sediment penetration and withdrawal; this strength member supports a small-diameter, rapidly responding tube containing thermistor sensors and a linear heater element. Tools used with submersibles are shorter and use a thin sensor tube, but not a strength member (Fig. 1c). In situ temperatures are estimated by extrapolating transient decays following probe penetration, and conductivities are determined from the rate of change of temperature following steady or impulsive activation of a heater. A typical data record is shown in Fig. 2, as are the resulting determinations of temperature and thermal conductivity. Heat flux is determined as the slope of a linear regression fit of temperature vs. cumulative thermal resistance, R :

$$R = \Sigma [\Delta z / \lambda(z)],$$

where λ is thermal conductivity measured at a series of depths, z , and Δz is the depth interval assumed to be represented by each measurement. This is equivalent to multiplying the thermal gradient by the thermal conductivity if the latter does not vary with depth. Errors associated with possible bottom-water temperature variations (discussed later) are evaluated by examining systematic deviations from linearity as a function of the number of thermistors included in the fit, working progressively up towards the shallowest measurement point. If there is systematic curvature in a plot of T versus R , it may indicate rapid, vertical fluid flow through sediments with associated advective heat transport. This has been noted in some hydrothermally active areas (e.g., Stein and Fisher 2001; Wheat et al. 2004), but is otherwise rare. Complete descriptions of instruments and discussions of data



Heat Flow, Seafloor: Methods and Observations, Fig. 1 Sensors and probes for measuring temperatures and thermal conductivity in marine sediments, including (a) outrigger temperature sensors mounted to the outside of a sediment corer (core barrel is 12 cm diameter), (b) a multipenetration heat-flow probe for measuring sediment temperatures

and thermal conductivities (total length is 6 m; photo courtesy of H. Villinger), (c) a 60-cm-long probe for measuring sediment heat flux with manned submersibles and remotely operated vehicles, and (d) a high-strength probe that extends below a drill bit for bottom-hole temperature measurements (length is 1.2 m, tip diameter 1 cm)

reduction methods can be found in Lister (1979), Hyndman et al. (1979), Davis (1988), Wright and Loudon (1989), Villinger and Davis (1987), and Stein and Fisher (2001).

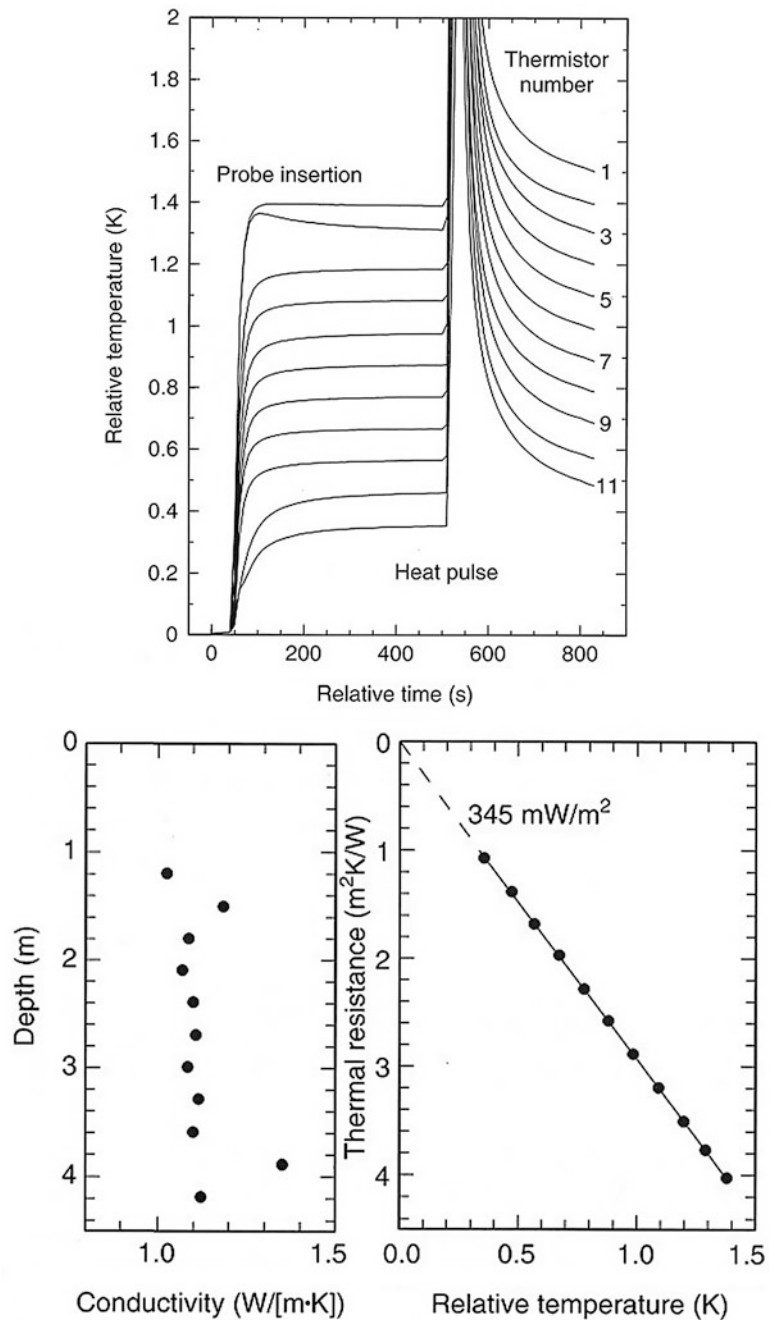
Deep Borehole Measurements

Where observations are needed in hard formations or at depths greater than can be penetrated with a gravity-driven device, drilling is required. For research objectives, this has been done primarily through the Deep Sea Drilling Project and the successive Ocean Drilling, Integrated Ocean Drilling, and International Ocean Discovery Programs. In relatively

unconsolidated sediments (typically the uppermost 50–100 m below the seafloor), hydraulically driven piston corers are deployed from the bottom of the drill string, and temperatures are measured at the tip of the core barrel (Horai and Von Herzen 1985; see Heesemann et al. 2006). At greater depths below the seafloor, high-strength probes can be pushed in with the weight of the drill string sufficiently far below the bottom of the hole (c. 1 m) to gain a measurement unperturbed by drilling (Uyeda and Horai 1980; Davis et al. 1997a) (Fig. 1d). Deeper than a few hundred meters in sediment, or at any level in crystalline rock, direct formation

Heat Flow, Seafloor: Methods and Observations,

Fig. 2 Typical data (upper panel) collected with a marine heat probe like that shown in Fig. 1b. Natural sediment temperatures are determined by extrapolating the transients following probe insertion, and thermal conductivities are determined from the rates of decay following the metered pulse of heat. In this example, high conductivities associated with two turbididic sand layers are present



measurements are not feasible; instead, temperature logs in the open hole or long-term borehole measurements are required to discriminate the natural formation thermal state from the commonly large and long-lived perturbations from drilling and subsequent fluid flow into or out of the hole. The most reliable method for determining the natural thermal state of crustal rocks has been to seal holes and install thermistor strings for long-term monitoring (Davis et al. 1992) (Fig. 3).

In a few cases, borehole and seafloor probe measurements have been made in close proximity (Fig. 4; Davis et al. 1999, 2003). Comparisons like these provide verification that

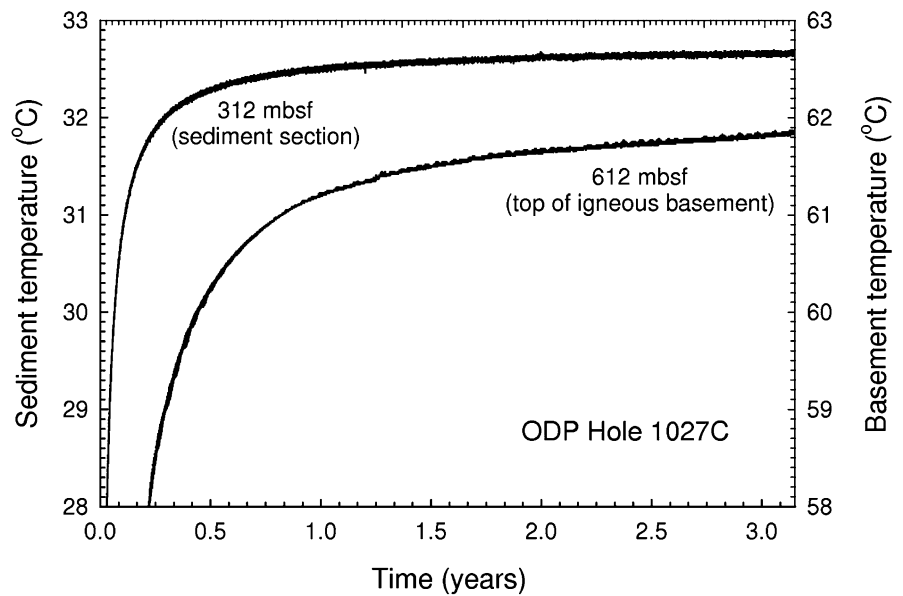
perturbations from processes like bottom-water temperature variations (discussed below), sedimentation, mass wasting, or pore-fluid advection are not present, and that shallow probe measurements give accurate values of the local geothermal heat flux.

Estimating Heat Flux From the Depth Limit of Gas Hydrate Stability

Sediments in some areas of the oceans contain sufficient quantities of gas to host subseafloor, solid gas hydrates.

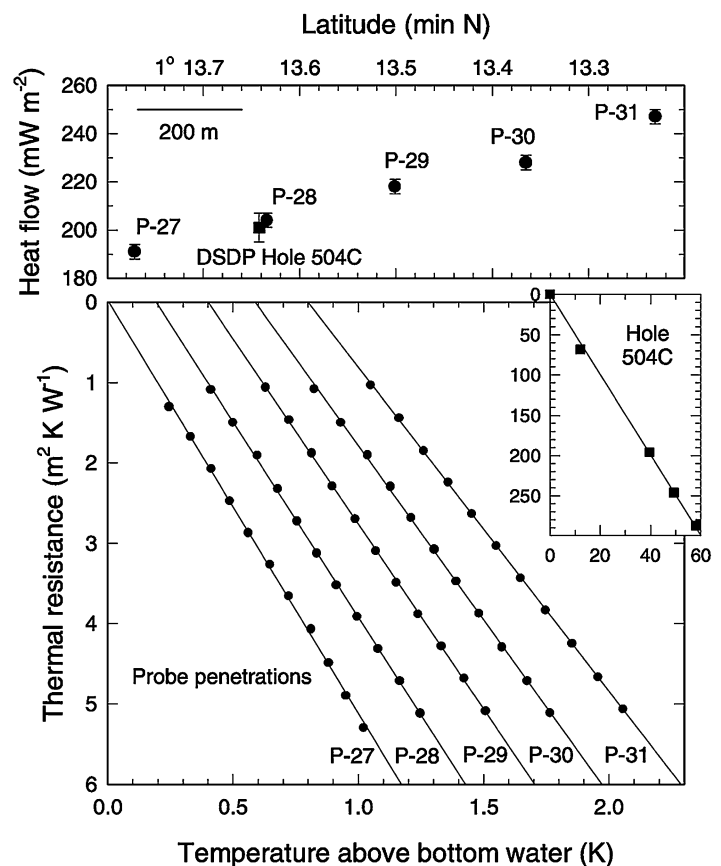
Heat Flow, Seafloor: Methods and Observations, Fig. 3

Two records from a 10-thermistor cable suspended in a borehole that penetrates a c. 600-m-thick sediment layer and into the underlying uppermost igneous oceanic crust. The longer recovery time at the deeper level reflects the cooling effects of the large volume of water that invaded the uncased and permeable igneous section during the 5 days between the time of drilling and when the hole was sealed and instrumented



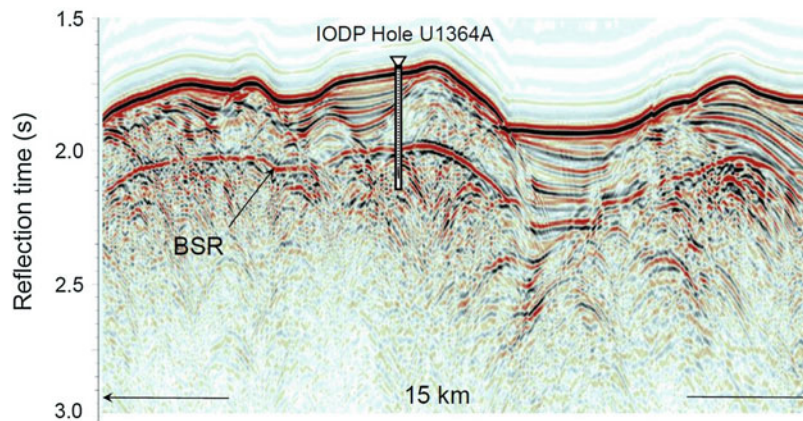
Heat Flow, Seafloor: Methods and Observations, Fig. 4

Comparison of co-located seafloor probe and borehole heat-flux observations. Variations in probe measurements along the transect adjacent to the borehole site illustrate how carefully such comparisons must be done. (Data from Davis et al. 2003)



Beyond being interesting in their own right, gas hydrates offer a means for estimating heat flux. The stability of gas hydrate is dependent on temperature and pressure, and the stability limit is commonly seen as a seismic reflector that marks the

boundary between sediment above containing gas hydrate, and sediment below containing free gas. Where heat flux is relatively constant, this reflector follows small local variations in seafloor depth and is referred to as a bottom



Heat Flow, Seafloor: Methods and Observations, Fig. 5 A methane-hydrate bottom simulating reflector in the Cascadia subduction zone accretionary prism offshore western Canada. In this example, coring, logging, and long-term temperature monitoring along a thermistor array in a co-located Integrated Ocean Drilling Program borehole

have allowed precise measurements of the depth and temperature at the BSR and of the thermal conductivity of the overlying sediments, and thus an accurate determination of the heat flux at the location of the borehole (53 mW m^{-2}) and elsewhere along this and nearby seismic profiles

simulating reflector (BSR; Fig. 5). The strongest factor governing the subseafloor depth to the limit of hydrate stability is temperature, and thus the BSR depth can be used to estimate the local heat flux. The methodology, introduced by Yamano et al. (1982) and reviewed in detail by Villinger et al. (2010), is especially valuable where bottom-water temperature variability is too large to permit accurate heat-flux determinations with shallow probes, where sediments are too hard to allow probe penetration, to interpolate between or extrapolate from sparse direct measurements, or to gain information where standard measurements are simply not available.

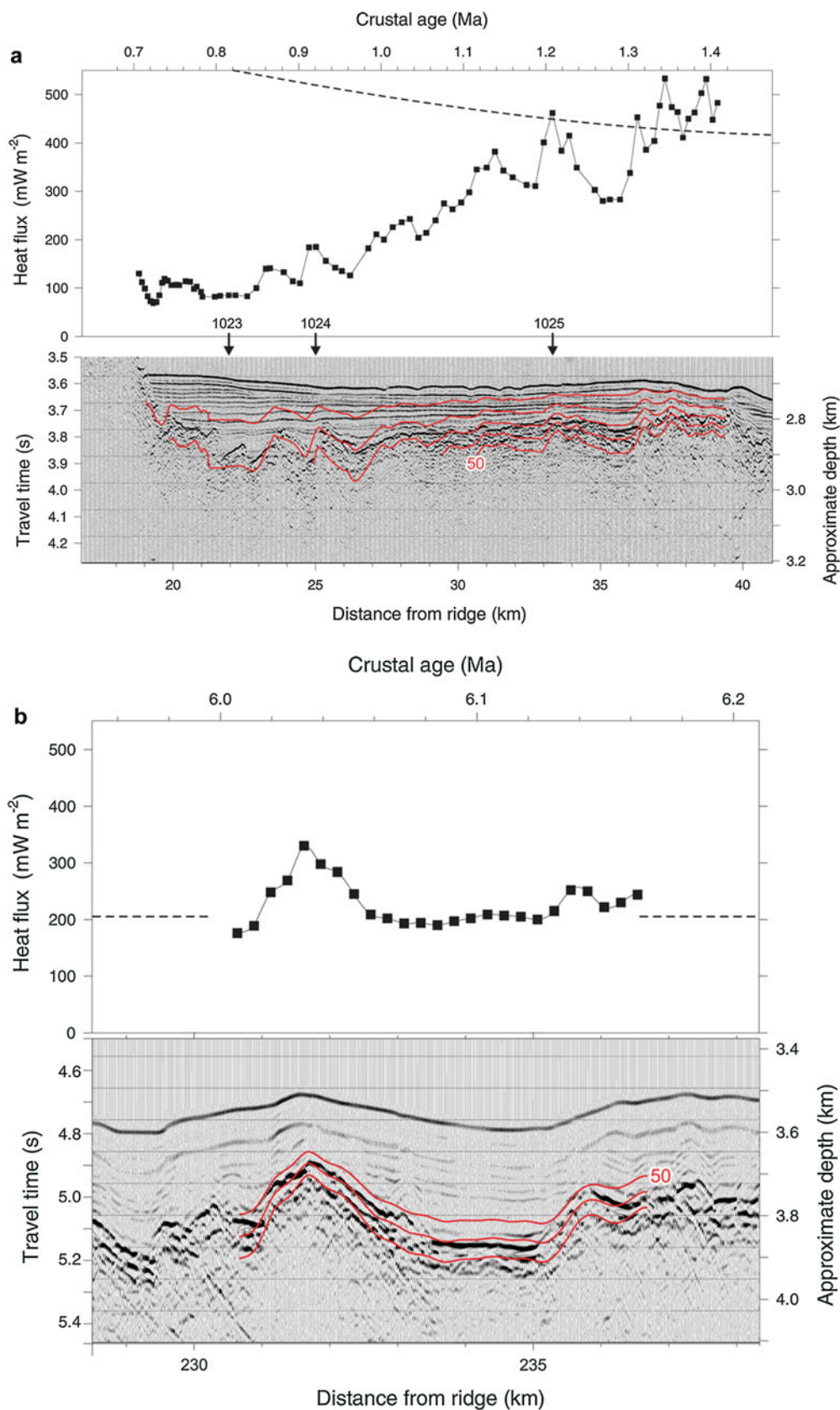
Several factors must be considered for accurate application of this method, including the temperatures at the seafloor and at the BSR hydrate stability boundary. The latter can be defined via an estimation of pressure at the boundary and the equation of state for the gas/gas-hydrate/water system or from an empirical relationship (e.g., Dickens and Qinby-Hunt 1997). It is normal to assume a seawater/methane system, although the presence of higher hydrocarbons and of porewater having higher or lower salinity than seawater both have effects (e.g., Andreassen et al. 2000; Ruppel et al. 2005). Next, the depth to the BSR must be estimated. This requires a conversion between two-way travel time (from seismic reflection data) and sediment thickness from the seafloor to the BSR, using velocities that can be estimated from seismic refraction, wide-angle reflection, or drilling data. If both the depth and temperature can be defined from drilling (e.g., Fig. 5), then much uncertainty can be eliminated and the resulting estimates of thermal gradient derived from BSRs can be extended regionally with greater confidence. To determine the heat flux, the cumulative thermal resistance from the seafloor to the depth of the BSR is required. This is commonly calculated

using the average of thermal conductivity values measured in core samples, or estimated using porosity or seismic velocity data.

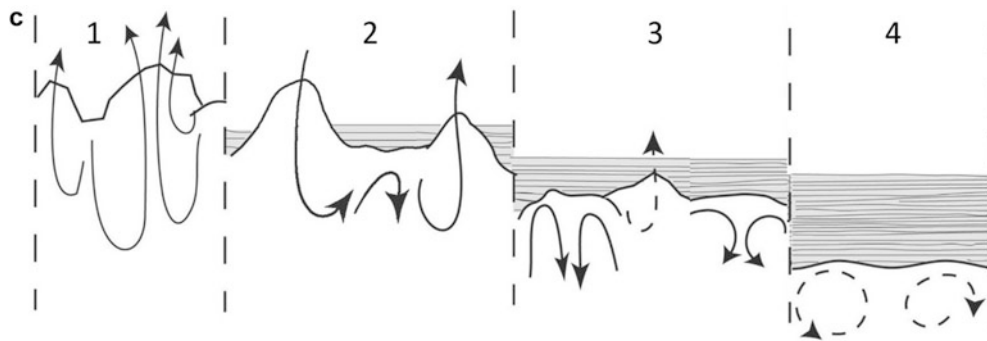
Example Studies

Heat Flow Signals from Hydrothermal Circulation

Hydrothermal circulation can both extract and redistribute lithospheric heat, making it difficult to determine the heat flux from deep in the lithosphere, although the pattern of redistribution can provide a valuable tracer of fluid flow. This has been the focus of a large number of studies (see review in Fisher and Harris 2010). Two instructive examples of the influence of hydrothermal circulation are shown in Fig. 6, where closely spaced measurements were made along transects striking perpendicular to basement structure. The first (Fig. 6a) crosses a sediment-covered area adjacent to a region of outcropping igneous crust, close to a spreading ridge axis (where measurements with conventional probes are impossible). Local variations correlated with sediment thickness variations are present, as is a systematic variation over a larger scale, with heat flux increasing with distance from the area of basement outcrop, opposite to the expected trend of decreasing heat flux with increasing seafloor age. Temperatures estimated at the top of the igneous section also increase systematically with distance from the ridge. Both trends suggest that crustal heat is laterally advected from the crust and lost to ventilated hydrothermal circulation at neighboring locations where sediment cover is absent or discontinuous. Away from areas of igneous outcrop, the hydrothermal system becomes increasingly sealed by sediments, and the measured seafloor heat flux values approach that expected from models of lithospheric cooling. Examples elsewhere suggest that the



Heat Flow, Seafloor: Methods and Observations, Fig. 6 (continued)



Heat Flow, Seafloor: Methods and Observations, Fig. 6 Transects of heat flux colocated along seismic reflection profiles on the flanks of the Juan de Fuca Ridge (a) striking away from a region of extensive basement near the left end of this profile, and the Costa Rica Rift (b) where an extensive sediment cover is present. Both show the effects of hydrothermal circulation on conductive seafloor heat flux and the crustal thermal regime. Heat flux estimated on the basis of the local lithosphere age (see text) is shown as the dashed line. Temperatures below the

seafloor extrapolated from seafloor probe measurements and observed in boreholes (Ocean Drilling Program site numbers given) are shown near the top of the igneous crust at intervals of 10 °C. The schematic diagram in (c) shows the influence of sediment burial on hydrothermal circulation and advective heat loss within and from the igneous crust, with panels 1–4 reflecting a general age progression from young, unsedimented seafloor to old, sediment-buried crust with permeability ultimately diminished by hydrothermal alteration

effects of advective heat loss from the upper igneous crust may be felt laterally as far as 50–100 km (e.g., Fisher et al. 2003; Fisher and Wheat 2010) and that advective heat transport is most effective in directions parallel to structural strike.

The second example (Fig. 6b) is in an area where sediment cover is continuous over a broad region spanning several tens of km. Locally, heat flux varies inversely with sediment thickness, much as in the case of local variations seen in Fig. 6a. Such variations are common in areas of young seafloor, but can occur across relatively old seafloor as well (Embley et al. 1983; Von Herzen 2004; Fisher and Von Herzen 2005). They are the consequence of thermally efficient, local convection in permeable igneous rocks beneath low-permeability sediment cover. If the permeability of the igneous “basement” formation is high, vigorous convective flow maintains nearly constant temperatures at the sediment/basement interface despite variations in the thermal resistance of the overlying sediment layer. In this instance, the average seafloor heat flux is close to that expected from the underlying lithosphere, suggesting that the circulation in the upper igneous crust may redistribute heat locally, but is not extracting heat advectively.

Ever since the early work of Lister (1972), the mere presence of local variability has been used as a diagnostic indicator of hydrothermal circulation in both young and old areas (e.g., Fig. 6c), but with widely spaced observations, neither of the signals exemplified in Fig. 6a, b could be resolved coherently; values were simply scattered and averages were often low. When systematic, detailed transects of observations began to be completed in context of co-located seismic data, the vigor of the convection could be inferred quantitatively from the nonconductive thermal regime (Fisher and Becker 1995; Davis et al. 1997b), and the amount of heat lost from the

crust by fluid advection could be estimated with growing confidence (e.g., Anderson and Skilbeck 1980; Stein and Stein 1994; Harris and Chapman 2004).

Two major lessons are learned from detailed observations like these for drawing conclusions about deep-seated heat flux. First, to ensure that observations do not suffer from the bias caused by advective heat extraction, it must be demonstrated that there are no exposures of permeable rock at faults or volcanic edifices within distances of at least several tens to 100 km or more. Second, large numbers of closely spaced observations must be made, ideally co-located with seismic reflection data, so that the local variability can be understood and meaningfully averaged, and the locally relevant lateral transport scale can be assessed (e.g., Sclater et al. 1976; Davis et al. 1999; Villinger et al. 2002). To develop a regional understanding, these transects should be created as part of a spatial grid, extending both across and along the structural strike of the underlying crust, to help resolve the possibility of advective heat loss from the crust in any direction.

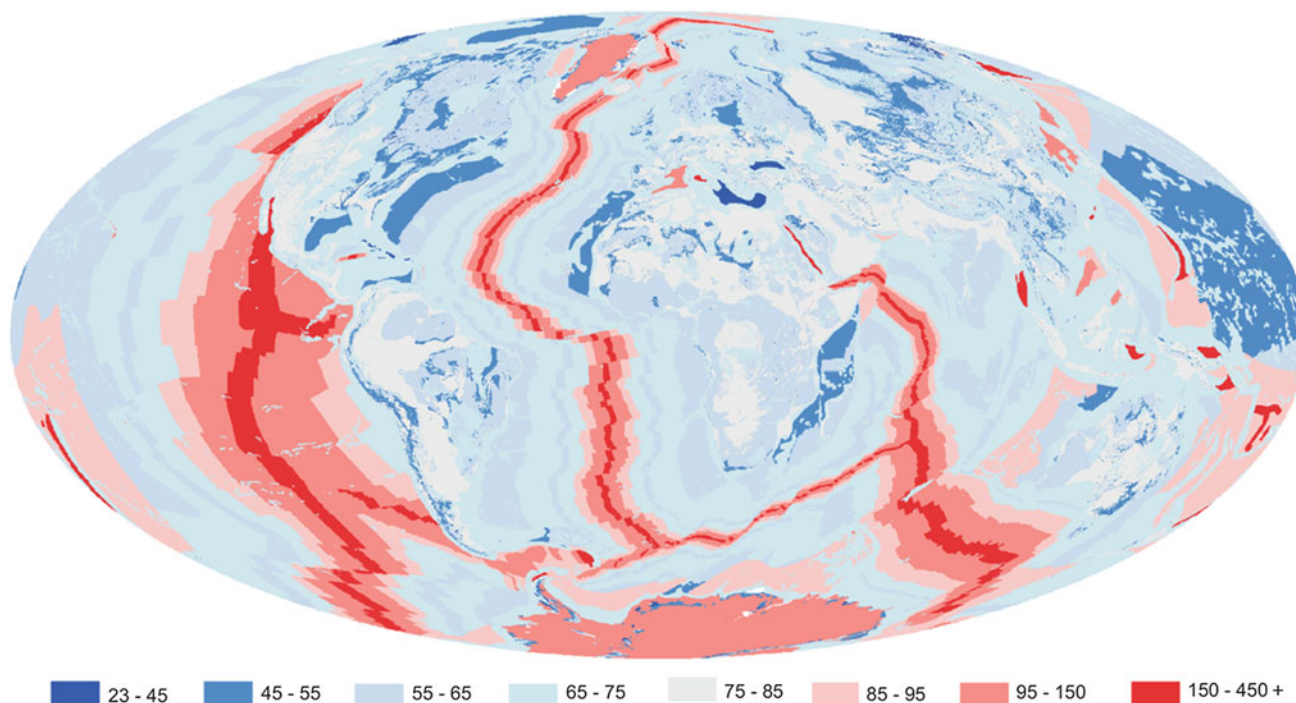
Interest in hydrothermal processes and their influences on water-rock-energy-chemical-biological interactions continues to grow. Estimates for the temperatures of circulation, the chemistry of the fluids, the volumetric rates of exchange between the crust and the ocean, and the consequent effects on crustal alteration and ocean chemistry have become reasonably well understood (e.g., Mottl and Wheat 1994; Elderfield and Schultz 1996; Fisher and Wheat 2010). Ongoing topics of research include the distribution of crustal permeability (e.g., Becker et al. 2013; Fisher et al. 2014), the percentage of rock affected by hydrothermal alteration (e.g., Coggon et al. 2014), and the influence of thermal and chemical conditions on chemosynthetic microbial populations (e.g., Orcutt et al. 2015).

Dependence of Heat Flux on Age and the Global Average

With the potentially large influence of hydrothermal circulation in mind, it is clear that a simple compilation of heat-flux data will provide a biased view of global heat loss. Except in old ocean basins (and in locations that are tens to hundreds of km from the nearest areas of exposed basement), values are likely to be highly variable and low relative to the heat loss expected from the underlying lithosphere. But by taking only those measurements that are sufficiently far from known permeable igneous outcrops, and sufficiently numerous to provide a reliable local average, a subset of data can be gathered that provides a reliable determination of deep-seated heat flux. When considered in context of lithospheric age, the results have been found to be consistent with both the characteristics of age-dependent seafloor subsidence and with simple lithospheric cooling theory (see entry on “Oceanic Lithosphere: Thermal Structure”). In young areas, heat flux is found to decline linearly with the inverse square root of age, following the simple relationship $Q = C t^{-1/2}$ (where Q is heat flux in mW m^{-2} , t is age in Ma, and C is a constant estimated between 475 and 510; see Harris and Chapman 2004). High-quality observations in older regions (>100 Ma) show heat flux to be generally uniform, in the range of $45\text{--}50 \text{ mW m}^{-2}$ (Lister et al. 1990; Hasterok et al. 2011), suggesting that the thermal structure of the lithosphere may become stabilized in

a state regulated either by the convectively supplied heat flux from the underlying asthenosphere or by the combination of a compositionally established lithospheric thickness and the relatively uniform temperature of the vigorously convecting asthenosphere.

With the relationship between heat flux and age thus defined, the problems of the unknown bias and large local variability in young areas and of the sparse distribution of measurements in large portions of the oceans can be overcome. This, together with seafloor ages defined by magnetic anomalies, allows a robust determination of the heat flux throughout the oceans (Fig. 7). Furthermore, with a seafloor area/age relationship (e.g., Parsons 1982; Wei and Sandwell 2006), the total heat flow through the oceans can be estimated. Several such estimates have been made (e.g., Williams and Von Herzen 1974; Sclater et al. 1980; Jaupart et al. 2007; Davies and Davies 2010), and all fall in a relatively narrow range centered around 32 TW (with contributions from marginal seas and hot-spot swells included). This yields an average total seafloor flux of roughly 107 mW m^{-2} , a number that has little physical significance but is considerably greater than that estimated in the early days of marine heat flow, and greater than the average though continents (c. 67 mW m^{-2}), particularly when the latter is adjusted for the contribution of continental crustal radiogenic heat (c. 33 mW m^{-2}).



Heat Flow, Seafloor: Methods and Observations, Fig. 7 Total heat flux through the oceans and continents, with factors such as advective hydrothermal heat loss through the seafloor accounted for. Color scaling is given in mW m^{-2} . The distribution of constraining observations is highly variable and sparse in many regions; values shown are based

primarily on age in the oceans (in the manner described in the text), and on age and geologic province in continental regions (from Davies and Davies 2010). The pattern of the regional variations demonstrates the important role of lithospheric creation at seafloor spreading centers in liberating Earth's internal heat

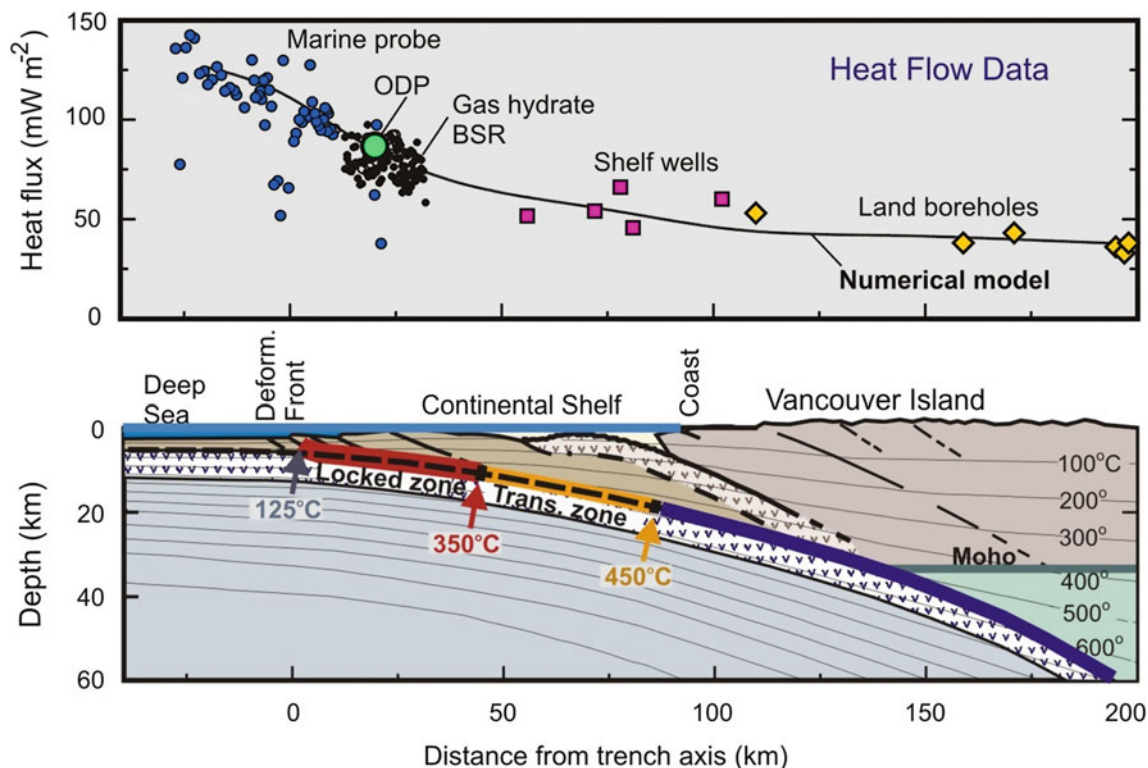
Conversely, with the total heat flow thus constrained, the heat lost advectively by hydrothermal circulation flowing through the seafloor (Fig. 6c panels 1 and 2) can be estimated from the difference between the total as determined above and the age-binned average of unfiltered observations. Such estimates of this “heat deficit” fall in the neighborhood of 10 TW (Stein and Stein 1994; Harris and Chapman 2004), meaning that over 20% of Earth’s total heat loss (32 and 14 TW from the oceans and continents, respectively) occurs by this mechanism. Most of the oceanic hydrothermal heat loss occurs from seafloor less than 8–10 Ma in age, and the deficit disappears (on average) by ~65 Ma. The actual age at which advective loss becomes insignificant is locally variable, depending primarily on the continuity of accumulating sediments that bury the igneous crust (Anderson and Skilbeck 1980; Harris and Chapman 2004).

The Signature of Subduction

Marine heat flow is used extensively to constrain deep thermal structure in studies of continental margins, marginal basins, and subduction zones. Particular attention has been given to the last in an effort to help define associated seismic hazards. A transect crossing the forearc prism of the Cascadia subduction zone illustrates one such application (Fig. 8). This transect begins with standard gravity-driven probe

measurements over the incoming plate and outermost accretionary prism, where bottom-water temperature variability is small. At the inner prism and on the continental shelf, where seafloor is shallower than 1500–2000 m, other measurement techniques are used, including borehole measurements and estimates made using the depth to a BSR which marks the limit of methane-hydrate stability. Observations like these allow the underlying thermal structure to be inferred, providing critical constraints on dehydration reactions, the rheology and frictional properties of the sediments and rocks, and the potential for seismogenic slip along the subduction thrust interface. In this example, the seafloor heat flux is variable locally, but regional values and trends are consistent with the expected thermal state created by the subduction of the oceanic plate and accretion of the thick sedimentary section.

More recently documented cases summarized by Harris et al. (2017) include the Middle America Trench, where the thermal state of the incoming Cocos plate is influenced by regionally varying hydrothermal heat loss. These variations correlate with differences in seismic processes occurring in the subduction zone itself (Newman et al. 2002; Fisher et al. 2003; Harris et al. 2010). One part of the plate is extensively cooled by hydrothermal circulation before the Cocos Plate is subducted, and earthquakes observed within the subduction wedge in this area are relatively deep (>20 km).



Heat Flow, Seafloor: Methods and Observations, Fig. 8 Structural and heat-flux transect across the Cascadia subduction forearc, with temperatures estimated from a numerical model for underthrusting and

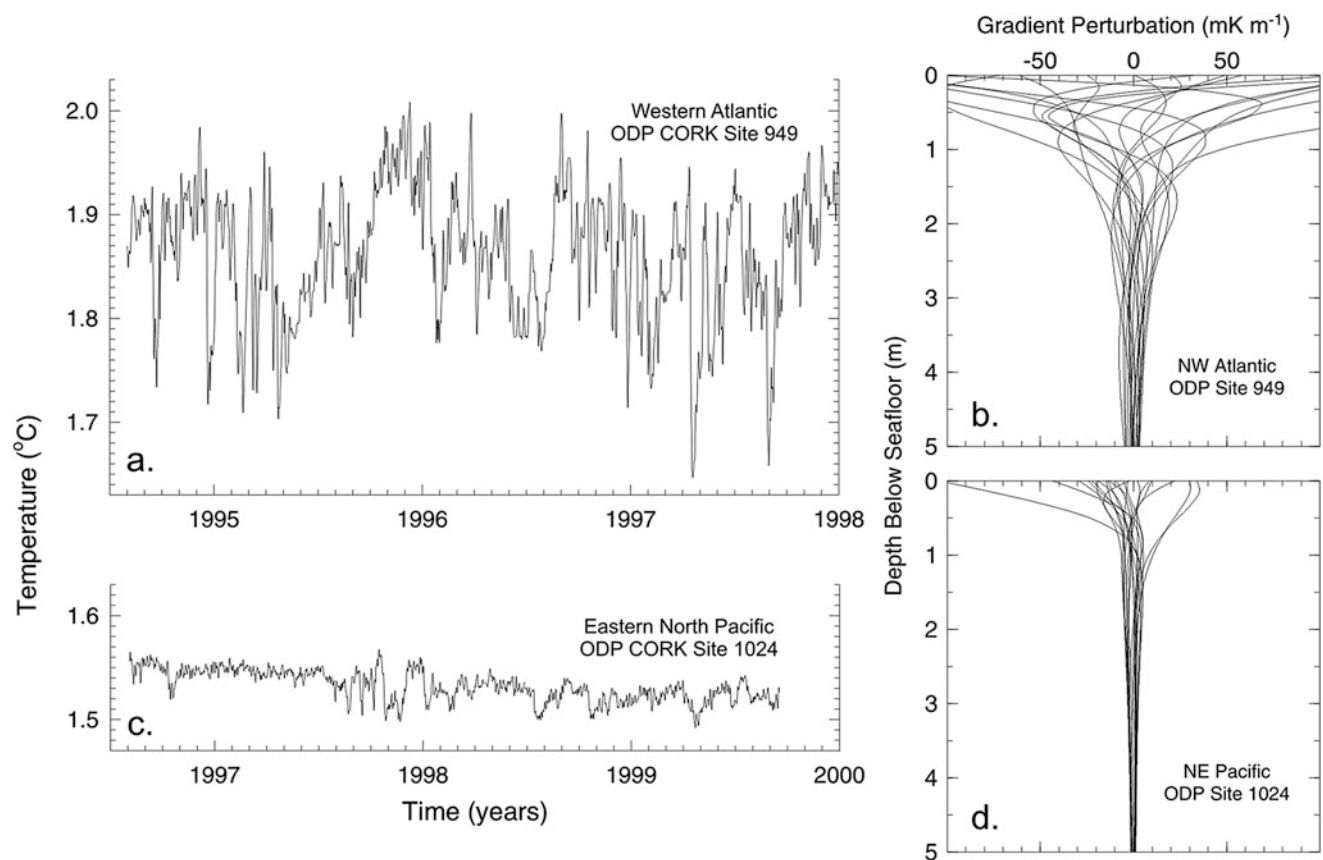
sediment thickening, constrained by the heat-flux data (following compilation of Hyndman and Wang 1993)

Earthquakes tend to be shallower (<20 km) along an adjacent segment of the subduction zone, where there is no evidence for regional advective heat extraction from the oceanic plate. One explanation for the different earthquake depths is that cooling of part of the Cocos Plate slows dewatering and deepens the transition of smectite to illite in subducting sediments (Spinelli and Saffer 2004; Lauer et al. 2017). The delay in heating associated with hydrothermal circulation in the crust prior to subduction causes a landward shift of the locked region where earthquakes are most likely to occur.

Bottom-Water Temperature Perturbations

Seafloor heat-flux measurements rely on the assumption that both long- and short-term bottom-water temperature (BWT) variations are small. This is often considered a good assumption because the volume of ocean bottom water is large and the temperature of the source in polar regions is regulated by the formation of sea ice. However, many early heat-flux measurements were made with 2–3 subsurface temperatures extending to depths of only a few meters, and it was difficult to identify nonlinearities that might indicate a recent change

in bottom water temperature. Good checks on the stability of BWTs were not possible until temperature observations began to be made in deep-sea boreholes and long records of bottom-water temperature were acquired (Hyndman 1984; Davis et al. 2003). An ideal suite of observations that would allow errors associated with bottom-water temperature variations to be quantified throughout the world's oceans – one that is broadly distributed both geographically and with ocean depth – does not yet exist, but the available data show that gradients measured a few meters below the seafloor do permit accurate determinations of heat flux in large areas of the oceans, generally where depths are greater than ~ 2000 m. One such check is illustrated in Fig. 4, where closely co-located seafloor probe and borehole observations are compared. Significant bottom-water temperature variations are ruled out by the linearity of the plots of temperature versus cumulative thermal resistance and by the agreement between the shallow probe and deep borehole determinations. This illustration also shows the importance of precise co-location when doing such a comparison, given the scale and magnitude of the local spatial variability of heat flux as defined by neighboring probe measurements.



Heat Flow, Seafloor: Methods and Observations, Fig. 9 Bottom-water temperature variations observed at 5000 m depth in the western Atlantic Ocean (a) and 2600 m depth in the Pacific Ocean (c), and perturbations of the geothermal gradient as a function of depth (b and

d) stemming from the variations, calculated at an evenly distributed suite of times during the temperature time series (data from Davis et al. 2003). These calculations show how data collected with short probes may be subject to significant errors due to changes in bottom water temperature

A more direct approach uses observations of bottom water fluctuations themselves. Perturbations attenuate with depth over a scale that is determined by the period and thermal diffusivity of the sediment, following the relationship $T(z) = T(0) \exp(-z/2\kappa\Gamma)^{1/2}$, where $T(0)$ is the fluctuation amplitude, κ is the thermal diffusivity, typically $= 4 \times 10^{-7} \text{ m}^2 \text{ s}^{-1}$, and Γ is the period of fluctuation. The effect on the thermal gradient is illustrated in Fig. 9 for two specific examples. The first, from a 5000-m-deep site in the western Atlantic, shows that oceanographic perturbations fall below the threshold of significance for heat-flux determinations by a depth of 2–3 m below the seafloor (Fig. 9a, b). The second, from a 2600-m-deep site in the eastern Pacific, shows smaller variability; there, significant perturbations would propagate less than 1 m below the seafloor (Fig. 9c, d). Generally speaking, the greatest stability is found in regions of great water depth and great distances from bottom-water sources in polar regions. A better definition of this as a potential source of error in heat-flux measurements is expected to emerge as the number of long-duration oceanographic monitoring experiments increases. In areas where bottom water temperature variations are likely to be problematic, local monitoring can allow local heat-flux measurements to be corrected (e.g., Neumann et al. 2017).

Summary

Heat flux, the product of the thermal gradient and thermal conductivity, can be measured with high accuracy in most deep-ocean settings with gravity-driven probes that penetrate a few meters into seafloor sediment. Meaningful values for deep-seated heat flux require a full understanding of the potentially large influence of hydrothermal circulation in the underlying permeable igneous rocks of the upper oceanic crust, however. This is best gained through closely spaced transects of heat-flux measurements co-located with seismic reflection profiles that show the structure of the underlying volcanic rocks. Such studies provide valuable constraints on the flow of water within the oceanic crust, the exchange of water, heat, and solutes between the crust and the oceans, the formation of hydrothermal mineral deposits, the accumulation of gas hydrates, and the development and maintenance of a sub-seafloor microbial biosphere. Where the effects of hydrothermal circulation can be accounted for or are known to be insignificant, deep-seated flux can be reliably determined and used to estimate the thermal history of sedimentary basins and the thermal regime of seismogenic fault zones; to understand seafloor spreading, the aging of oceanic lithosphere, and the origin of mid-plate swells; and to define the global heat budget which is dominated by heat loss through the seafloor.

Cross-References

- [Heat Flow, Continental](#)
- [Lithosphere, Oceanic: Thermal Structure](#)

Bibliography

- Anderson RN, Skilbeck JN (1980) Oceanic heat flow. In: Emiliani C (ed) *The Sea*, vol 7. Wiley Interscience, New York, pp 489–523
- Andreassen K, Mienert J, Bryn P, Singh SC (2000) A double gas-hydrate related bottom simulating reflector at the Norwegian continental margin. In *gas hydrates: challenges for the future*. *Ann N Y Acad Sci*, New York 912:126–135
- Becker K, Fisher AT, Tsuji T (2013) New packer experiments and borehole logs in upper oceanic crust: evidence for ridge-parallel consistency in crustal hydrogeologic properties. *Geochim Geophys Geosyst* 14(8). <https://doi.org/10.1002/ggge.20201>
- Bullard EC (1954) The flow of heat through the floor of the Atlantic Ocean. *Proc R Astron Soc London A* 222:408–429
- Coggon RM, Rehkämpera M, Atteck C, Teagle DAH, Teagle AJC, Cooper MJ (2014) Controls on thallium uptake during hydrothermal alteration of the upper ocean crust. *Geochim Cosmochim Acta* 144:25–42. <https://doi.org/10.1016/j.gca.2014.09.001>
- Davies JH, Davies DR (2010) Earth's surface heat flux. *Solid Earth* 1:5–24
- Davis EE (1988) Oceanic heat flow density. In: Haenel R, Rybach L, Stegena L (eds) *Handbook of terrestrial heat-flow density determination*. Kluwer Academic Publishers, Dordrecht, pp 223–260
- Davis EE, Becker K, Pettigrew T, Carson B (1992) CORK: a hydrologic seal and downhole observatory for deep-sea boreholes. *Proc Ocean Drill Program Initial Rep* 139:45–53
- Davis EE, Wang K, He J, Chapman DS, Villinger H, Rosenberger A (1997a) An unequivocal case for high Nusselt number hydrothermal convection in sediment-buried igneous oceanic crust. *Earth Planet Sci Lett* 146:137–150
- Davis EE, Villinger H, Macdonald RD, Meldrum RD, Grigel J (1997b) A robust rapid-response probe for measuring bottom-hole temperatures in deep-ocean boreholes. *Mar Geophys Res* 19:267–281
- Davis EE, Chapman DS, Wang K, Villinger H, Fisher AT, Robinson SW, Grigel J, Pribnow D, Stein JS, Becker K (1999) Regional heat flow variations on the sedimented Juan de Fuca Ridge eastern flank: constraints on lithospheric cooling and lateral hydrothermal heat transport. *J Geophys Res* 104:17,675–17,688
- Davis EE, Wang K, Becker K, Thomson RE, Yashayaev I (2003) Deep-ocean temperature variations and implications for errors in seafloor heat flow determinations. *J Geophys Res* 108:2034. <https://doi.org/10.1029/2001JB001695>
- Dickens GR, Qinby-Hunt MS (1997) Methane hydrate stability in pore water: a simple theoretical approach for geophysical applications. *J Geophys Res* 102:773–783
- Elder JW (1965) Physical processes in geothermal areas. In: Lee WHK (ed) *Terrestrial heat flow*. Am Geophys Union, Washington, DC, pp 211–239
- Elderfield H, Schultz A (1996) Mid-ocean ridge hydrothermal fluxes and the chemical composition of the ocean. *Annu Rev Earth Planet Sci* 24:191–224
- Embley RW, Hobart MA, Anderson RN, Abbott D (1983) Anomalous heat flow in the northwest Atlantic: a case for continued hydrothermal circulation in 80-M.y. crust. *J Geophys Res* 88:1,067–1,074
- Fisher AT, Becker K (1995) The correlation between heat flow and basement relief: observational and numerical examples and implications for upper crustal permeability. *J Geophys Res* 100:12,641–12,657

- Fisher AT, Harris RN (2010) Using seafloor heat flow as a tracer to map subseafloor fluid flow in the oceanic crust. *Geofluids* 10:142–160
- Fisher AT, Von Herzen RP (2005) Models of hydrothermal circulation within 106 Ma seafloor: constraints on the vigor of fluid circulation and crustal properties below the Madeira Abyssal Plain. *Geochim Geophys Geosyst* 6. <https://doi.org/10.1029/2005GC001013>.
- Fisher AT, Wheat CG (2010) Seamounts as conduits for massive fluid, heat, and solute fluxes on ridge flanks. *Oceanography* 23:74–87
- Fisher AT, Villinger H, Harris RN, Von Herzen RP, Pfender M, Müller M, Grevemeyer I, Kaul N, Wheat CG (2002) Comment on “Deep-penetration heat flow probes raise questions about interpretations from shorter probes,” by Géli et al., 2001. *EOS Trans Am Geophys Union* 82(29):317, 320, *Eos Trans Am Geophys Union* 83(18):196
- Fisher AT, Stein CA, Harris RN, Wang K, Silver EA, Pfender M, Hutnak M, Cherkaoui A, Bodzin R, Villinger H (2003) Abrupt thermal transition reveals hydrothermal boundary and role of seamounts within the Cocos plate. *Geophys Res Lett* 30(11):1550. <https://doi.org/10.1029/2002GL016766>.
- Fisher AT, Alt JC, Bach W (2014) Hydrogeologic properties, processes and alteration in the igneous ocean crust. In: Stein R, Blackman D, Inagaki F, Larsen H-C (eds) *Earth and life processes discovered from subseafloor environment – a decade of science achieved by the Integrated Ocean Drilling Program (IODP)*. Elsevier, Amsterdam/New York, pp 507–551
- Harris RN, Chapman DS (2004) Deep-seated oceanic heat flux, heat deficits, and hydrothermal circulation. In: Davis EE, Elderfield H (eds) *Hydrogeology of the oceanic lithosphere*. Cambridge University Press, Cambridge, pp 311–336
- Harris RN, Grevemeyer I, Ranero CR, Villinger H, Barckhausen U, Henke T, Mueller C, Neben S (2010) Thermal regime of the Costa Rican convergent margin: 1. Along-strike variations in heat flow from probe measurements and estimated from bottom-simulating reflectors. *Geochim Geophys Geosyst* 11:Q12S28. <https://doi.org/10.1029/2010GC003272>
- Harris RN, Spinelli GA, Fisher AT (2017) Hydrothermal circulation and the thermal structure of shallow subduction zones. *Geosphere* 13:1–20
- Hasterok D, Chapman DS, Davis EE (2011) Oceanic heat flow: implications for global heat loss. *Earth Planet Sci Lett* 311:386–395. <https://doi.org/10.1016/j.epsl.2011.09.044>
- Heesemann M, Villinger H, Fisher AT, Trehu AM, Witte S (2006) Testing and deployment of the new APC3 tool to determine in situ temperature while piston coring. In: Collett TS, Riedel M, Malone MJ (eds) *Proceedings of the integrated ocean drilling program expedition 311*. Integrated Ocean Drilling Program Management International, Inc., College Station
- Horai K, Von Herzen RP (1985) Measurement of heat flow on Leg 86 of the Deep Sea Drilling Project. Initial Reports, DSDP 86:759–777
- Hyndman RD (1984) Review of Deep Sea Drilling Project geothermal measurements through Leg 71. Initial Reports, DSDP 78B:813–823
- Hyndman RD, Wang K (1993) Thermal constraints on the zone of major thrust earthquake failure: the Cascadia subduction zone. *J Geophys Res* 98:2039–2060
- Hyndman RD, Davis EE, Wright JA (1979) The measurement of marine geothermal heat flow by a multipenetration probe with digital acoustic telemetry and in situ thermal conductivity. *Mar Geophys Res* 4:181–205
- Jaupart C, Labrosse S, Mareschal J-C (2007) Temperatures, heat, and energy in the mantle of the Earth. *Treatise on Geophysics* 7:253–303
- Langseth MG, LePichon X, Ewing M (1966) Crustal structure of the midocean ridges, 5, heat flow through the Atlantic Ocean floor and convection currents. *J Geophys Res* 71:5321–5355
- Lauer RM, Saffer DM, Harris RN (2017) Links between clay transformation and earthquakes along the Costa Rican subduction margin. *Geophys Res Lett* 44:7725–7732. <https://doi.org/10.1002/2017GL073744>
- Lee WHK, Uyeda S (1965) Review of heat flow data. In: Lee WHK (ed) *Terrestrial heat flow*. Geophysical monograph, vol 8. American Geophysical Union, Washington, DC, pp 87–190
- Lister CRB (1972) On the thermal balance of a mid-ocean ridge. *Geophys J R Astron Soc* 26:515–535
- Lister CRB (1979) The pulse-probe method of conductivity measurement. *Geophys J R Astron Soc* 57:451–461
- Lister CRB, Sclater JG, Davis EE, Villinger H, Nagihara S (1990) Heat flow maintained in ocean basins of great age: investigations in the north-equatorial west Pacific. *Geophys J Int* 102:603–630
- Mottl MJ, Wheat CG (1994) Hydrothermal circulation through mid-ocean ridge flanks: fluxes of heat and magnesium. *Geochim Cosmochim Acta* 58:2,225–2,237
- Neumann F, Negrete-Aranda R, Harris RN, Contreras J, Sclater JG, González-Fernández A (2017) Systematic heat flow measurements across the Wagner Basin, northern gulf of California. *Earth Planet Sci Lett* 479:340–353
- Newman AV, Schwartz SY, Gonzalez V, DeShon HR, Protti JM, Dorman LM (2002) Along-strike variability in the seismogenic zone below Nicoya Peninsula, Costa Rica. *Geophys Res Lett* 29. <https://doi.org/10.1029/2002GL015409>
- Orcutt BN, Sylvan JB, Rogers DR, Delaney J, Lee RW, Girguis PR (2015) Carbon fixation by basalt-hosted microbial communities. *Front Microbiol* 6. <https://doi.org/10.3389/fmicb.2015.00904>
- Parsons B (1982) Causes and consequences of the relation between area and age of the ocean floor. *J Geophys Res* 87:289–303
- Petterson H (1949) Exploring the bed of the ocean. *Nature* 416:468–470
- Revelle RR, Maxwell AE (1952) Heat flow through the floor of the eastern North Pacific Ocean. *Nature* 170:199–202
- Ruppel C, Dickens GR, Castellini DG, Gilhooly W, Lizarralde D (2005) Heat and salt inhibition of gas hydrate formation in the northern Gulf of Mexico. *Geophys Res Lett* 32:L04605
- Sclater JG, Crowe J, Anderson RN (1976) On the reliability of ocean heat flow averages. *J Geophys Res* 81:2,997–3,006
- Sclater JG, Jaupart C, Galson D (1980) The heat flow through oceanic and continental crust and the heat loss of the earth. *Rev Geophys* 18:269–311
- Spinelli GA, Saffer D (2004) Along-strike variations in underthrust sediment dewatering on the Nicoya margin, Costa Rica related to the updip limit of seismicity. *Geophys Res Lett* 31. <https://doi.org/10.1029/2003GL018863>
- Stein JS, Fisher AT (2001) Multiple scales of hydrothermal circulation in Middle Valley, northern Juan de Fuca Ridge: physical constraints and geologic models. *J Geophys Res* 106(B5):8563–8580
- Stein CA, Stein S (1994) Constraints on hydrothermal heat flux through the oceanic lithosphere from global heat flow. *J Geophys Res* 99:3,081–3,095
- Uyeda S, Horai K (1980) Heat flow measurements on Deep Sea Drilling Project Leg 60. In: Initial reports, DSDP, vol 60. U. S. Govt. Printing Office, Washington, DC, pp 789–800
- Villinger H, Davis EE (1987) A new reduction algorithm for marine heat flow measurements. *J Geophys Res* 92:12,846–12,856
- Villinger H, Grevemeyer I, Kaul N, Hauschild J, Pfender M (2002) Hydrothermal heat flux through aged oceanic crust: where does the heat escape? *Earth Planet Sci Lett* 202:159–170
- Villinger H, Trehu AM, Grevemeyer I (2010) Seafloor marine heat flux measurements and estimation of heat flux from seismic observations of bottom simulating reflectors. In: Riedel M, Willoughby EC, Chopra S (eds) *Geophysical characterization of gas hydrates*. Society of Exploration Geophysicists, Tulsa, pp 279–300
- Von Herzen RP (2004) Geothermal evidence for continuing hydrothermal circulation in older (> 60 M.y.) ocean crust. In: Davis EE, Elderfield H (eds) *Hydrogeology of the oceanic lithosphere*. Cambridge University Press, Cambridge, pp 414–447

- Von Herzen RP, Uyeda S (1963) Heat flow through the eastern Pacific Ocean floor. *J Geophys Res* 68:4,219–4,250
- Wei M, Sandwell D (2006) Estimates of heat flow from Cenozoic seafloor using global depth and age data. *Tectonophysics* 417:325–335
- Wheat CG, Mottl MJ, Fisher AT, Kadko D, Davis EE, Baker E (2004) Heat and fluid flow through a basaltic outcrop on a ridge flank. *Geochem Geophys Geosyst* 5(12). <https://doi.org/10.1029/2004GC000700>
- Williams DL, Von Herzen RP (1974) Heat loss from the earth: new estimate. *Geology* 2:327–330
- Williams DL, Von Herzen RP, Sclater JG, Anderson RN (1974) The Galapagos spreading center: lithospheric cooling and hydrothermal circulation. *Geophys J R Astron Soc* 38:587–608
- Wright JA, Loudon KE (1989) *Handbook of seafloor heat flow*. CRC Press, Boca Raton. 498 pp
- Yamano M, Uyeda S, Aoki Y, Shipley TH (1982) Estimates of heat flow derived from gas hydrates. *Geology* 10:339–343

Height Systems, Vertical Datums and Their Unification

Michael G. Sideris

Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada

Definition

Height or *elevation* of a point P on the Earth's surface is the distance from P to a reference surface of zero height, measured along a precisely defined path.

Different *Height Systems* result from the choice of the reference surface and the path from P to that surface.

Vertical Datum is a surface of zero height, typically defined as the equipotential (level) surface of the Earth's gravity field that passes through an accessible *fundamental point*.

Vertical Datum Unification is the process of connecting vertical datums to each other, or to one main datum, by determining the offsets between existing vertical datums.

Note: the above definitions refer primarily to *physically meaningful heights*, i.e., to heights that express, and depend on, differences of the Earth's gravity potential.

Introduction

Height differences and absolute heights can typically be obtained by spirit levelling (see ► “[Geodesy, Ground Positioning, and Levelling](#)”) or by combining Global Navigation Satellite Systems (GNSS) positioning methods with a model of the geoid (see ► “[Geoid Determination, Theory and Principles](#)”). Their accurate determination is essential for many engineering and environmental applications of high

societal impact, both globally and regionally. In such applications, height differences need to represent potential differences between points and not just geometric differences as those obtained from, e.g., GNSS. Furthermore, these heights need to refer to the same zero-height reference surface or *vertical datum*, and hence the need for determining the offsets between existing vertical datums, i.e., vertical datum unification. These offsets can be estimated if the geodetic, physical and geoid heights (from GNSS, spirit leveling and gravity information, respectively) are known for the fundamental and/or other points in each vertical datum. Besides a geoid model obtained from the solution of the geodetic Boundary Value Problem (BVP), height data from GNSS-surveyed levelling benchmarks on land, GNSS-surveyed tide gauge stations along coasts, and a mean dynamic topography model at sea can be used in the computations. In this entry, therefore, we present the most common height systems, the theoretical definition and practical realization of a vertical datum, and the unification of existing vertical datums, geared in particular towards the realization of the International Height Reference System (Ihde et al. 2017; Sánchez and Sideris 2017).

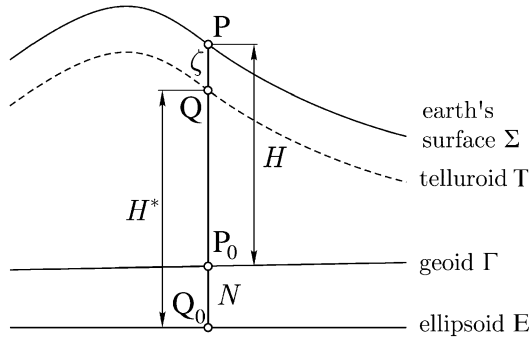
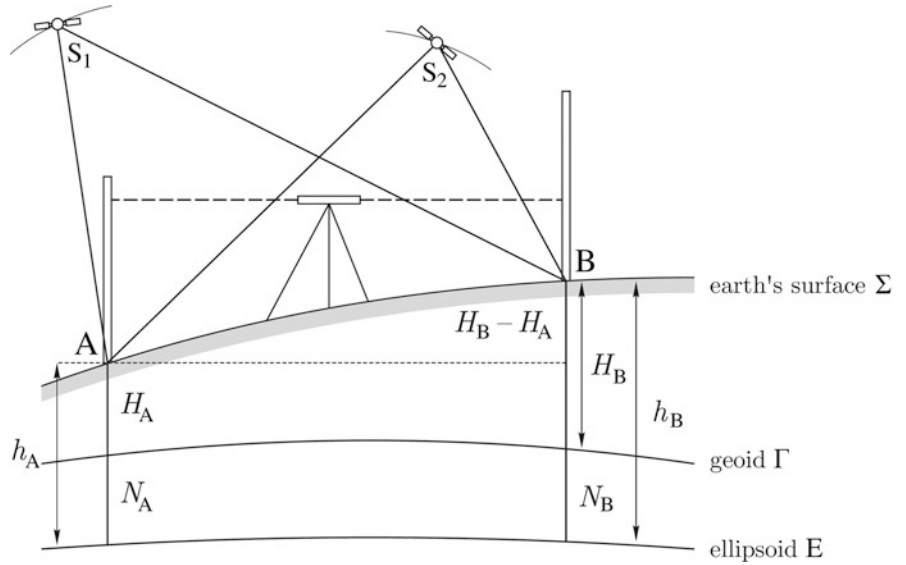
Obtaining height differences by traditional spirit leveling supplemented by gravity observations between benchmarks is a very labor-intensive, time-consuming, terrain- and weather-dependent, and costly undertaking. Height modernization in many countries now relies on height differences by GNSS positioning methods and the geoid (Véronneau and Huang 2016; see Fig. 1). Although in this way the limitations of spirit leveling are avoided, the resulting heights are *ellipsoidal* or *geometric heights*, h , above the surface of a reference ellipsoid while the former method provides *orthometric heights*, H , above the surface of the geoid (length of line PQ_0 and curved line PP_0 in Fig. 2, respectively). Only the orthometric heights have a physical meaning as they represent potential differences between benchmarks (see section “[Height Systems](#)” below), which are indispensable for environmental monitoring applications (e.g., sea level rise, coastal inundation, floods) and water resources applications (e.g., dams, drainage, irrigation). As it can be seen in Figs. 1 and 2, it is possible to convert GNSS-derived ellipsoidal heights to orthometric heights, through the use of geoid undulations, N , by use of the simple expression $H = h - N$ (see also Eq. 4 below). National height modernization efforts rely of this simple equation and allows users to obtain orthometric heights at any point by simply using a GNSS receiver together with the national geoid model.

Height Systems

We have already mentioned that the reference surface, i.e., the surface of zero height in a particular height system, is called

Height Systems, Vertical Datums and Their Unification,

Fig. 1 Traditional and modern way of height determination (h , H and N denote ellipsoidal heights, orthometric heights and geoid undulations, respectively)



Height Systems, Vertical Datums and Their Unification,

Fig. 2 Geometry between ellipsoidal ($h = PQ_0$), orthometric ($H = PP_0$), and normal heights ($H^* = QQ_0$), geoid undulations ($N = P_0Q_0$), and height anomalies ($\zeta = PQ$)

vertical datum, and heights are measured along straight or curved lines above it. For example, geodetic (or ellipsoidal) heights h derived from GNSS observations are “measured” along the normal to the surface of an ellipsoid, which is the datum in the *geodetic height system*. Because geodetic heights do not represent potential differences and therefore cannot be used for applications like the ones mentioned in the introduction, other, physically meaningful, height systems are in use. They are based on the following equation that relates differentially height to gravity g and gravity potential W (Hofmann-Wellenhof and Moritz 2006):

$$dh = -\frac{dW}{g} \quad (1)$$

Using h to denote a generic height, $\tilde{g}$ to denote the average gravity from the reference surface to a point P on the Earth's surface along the line perpendicular to the gravity

equipotential surfaces (plumb line), and integrating along the plumb line we obtain

$$\begin{aligned} C_P = W_0 - W_P &= -\int_0^{h_P} dW = \int_0^{h_P} g dh = h_P \frac{1}{h_P} \int_0^{h_P} g dh \\ &= h_P \tilde{g}_P \end{aligned} \quad (2)$$

$$h_P = \frac{C_P}{\tilde{g}_P}, \quad \tilde{g}_P = \frac{1}{h_P} \int_0^{h_P} g dh \quad (3)$$

C_P in Eq. (2) is the difference between the gravity potential W_0 of the datum (typically, the geoid) and the gravity potential W_P of any point P on the Earth's surface, called *geopotential number*. The unit for C is the geopotential unit, g.p.u., where 1 g.p.u. = 1000 Gal m. Numerically, geopotential numbers in g.p.u. are almost equal to heights in meters since, using an average g -value of 981 Gal, $C_P = \tilde{g}_P h_P \approx 0.981 h_P$.

Three different height systems can be defined based on the choice of $\tilde{g}$ in Eq. (3):

- The *dynamic height system* ($h = H^d$ when $\tilde{g} = \gamma_{45}$)
- The *orthometric height system* ($h = H$ when $\tilde{g} = \bar{g}$)
- The *normal height system* ($h = H^*$ when $\tilde{g} = \bar{\gamma}$)

where γ is the gravity of the reference ellipsoid, γ_{45} is the value of γ at geodetic latitude $\phi = 45^\circ$, and the overbar denotes average gravity along the corresponding plumb line. In Fig. 1 in Sideris, ▶ “[Geoid Determination, Theory and Principles](#)”, h , H , and H^* are the lengths of the lines PQ_0 , PP_0 , and QQ_0 , respectively. Usually, the curvature of the plumb lines and the

small deflection angles are ignored, resulting in a simpler figure, such as Fig. 2 above, which illustrates the relationship between geodetic, orthometric, and normal heights, as well as geoid undulations N and height anomalies ζ :

$$h = H + N = H^* + \zeta \quad (4)$$

N and ζ – and the corresponding surfaces of the geoid and telluroid/quasi-geoid – are obtained as the solutions to the Stokes and Molodensky geodetic BVPs, respectively; the relevant theory and methodology are given in this volume in ► “Geoid Determination, Theory and Principles” and ► “Geoid Determination, Computational Methods”. Furthermore, using Eq. (4) above and Eq. (23) in Sideris, ► “Geoid Determination, Theory and Principles”, they can be related by the following expression through the Bouguer anomaly Δg_B :

$$N - \zeta = H^* - H = \frac{\bar{g} - \bar{\gamma}}{\bar{\gamma}} H \approx \frac{\Delta g_B}{\bar{\gamma}} H \quad (5)$$

It should be noted that Fig. 2 does not depict dynamic heights H^d as they have no geometrical representation; they are basically potential differences given in units of length. Nevertheless, they are very important in applications that require direct knowledge of potential differences; as an example, dynamic heights are used in the International Great Lakes Datum of 1985 (IGLD85) for safe navigation in the Great Lakes of North America. A very good introduction to these height systems and vertical datums is given in Jekeli (2000), and a detailed treatment can be found in Sansò et al. (2019).

Although physical heights are not directly measurable, they could be derived from height differences Δn_{AB} observed by spirit leveling between two points A and B (see Fig. 1), by applying a gravity-dependent *height correction* hc_{AB} :

$$\begin{aligned} \Delta h_{AB} &= h_B - h_A = \frac{C_B - C_A}{\bar{g}} = \frac{1}{\bar{g}} \int_A^B g dn = \frac{1}{\bar{g}} \int_A^B (\tilde{g} + g - \tilde{g}) dn \\ &= \int_A^B dn + \int_A^B \frac{g - \tilde{g}}{\bar{g}} dn = \Delta n_{AB} + hc_{AB}, \quad hc_{AB} = \int_A^B \frac{g - \tilde{g}}{\bar{g}} dn \end{aligned} \quad (6)$$

Then for the three height systems we have:

- Dynamic height differences $\Delta h_{AB} = \Delta H_{AB}^d$ and height corrections $hc_{AB} = dc_{AB}$
- Orthometric height differences $\Delta h_{AB} = \Delta H_{AB}$ and height corrections $hc_{AB} = oc_{AB}$
- Normal height differences $\Delta h_{AB} = \Delta H_{AB}^*$ and height corrections $hc_{AB} = nc_{AB}$

The height differences and height corrections are given by the expressions (Hofmann-Wellenhof and Moritz 2006, Sects. 4.2–4.4)

$$\begin{aligned} \Delta h_{AB}^d &= \Delta n_{AB} + dc_{AB}, \\ dc_{AB} &= \int_A^B \frac{g - \gamma_{45}}{\gamma_{45}} dn \approx \sum_A^B \frac{g - \gamma_{45}}{\gamma_{45}} \delta n \end{aligned} \quad (7)$$

$$\begin{aligned} \Delta H_{AB} &= \Delta n_{AB} + oc_{AB}, \\ oc_{AB} &= \int_A^B \frac{g - \bar{g}}{\bar{g}} dn \approx \sum_A^B \frac{g - \bar{g}}{\bar{g}} \delta n = dc_{AB} \\ &\quad + \frac{\bar{g}_A - \gamma_{45}}{\gamma_{45}} H_A - \frac{\bar{g}_B - \gamma_{45}}{\gamma_{45}} H_B \end{aligned} \quad (8)$$

$$\begin{aligned} \Delta H_{AB}^* &= \Delta n_{AB} + nc_{AB}, \\ nc_{AB} &= \int_A^B \frac{g - \bar{\gamma}}{\bar{\gamma}} dn \approx \sum_A^B \frac{g - \bar{\gamma}}{\bar{\gamma}} \delta n = dc_{AB} \\ &\quad + \frac{\bar{\gamma}_A - \gamma_{45}}{\gamma_{45}} H_A^* - \frac{\bar{\gamma}_B - \gamma_{45}}{\gamma_{45}} H_B^* \end{aligned} \quad (9)$$

where δn are the height difference segments observed between points A and B.

For the orthometric heights, the average value of gravity along the plumb line is often approximated by use of Prey’s reduction, resulting in *Helmert orthometric heights* (or *normal orthometric heights* when the actual gravity value is unknown and is approximated by the normal gravity value):

$$\bar{g}_P = \frac{1}{H_P} \int_{P_0}^P g dH \approx g_P - \frac{1}{2} \frac{\partial g}{\partial H} H_P \approx g_P - \left(2\pi G \rho + \frac{1}{2} \frac{\partial \gamma}{\partial H} \right) H_P \quad (10)$$

For the normal heights, the average value of normal gravity along the normal plumb line can be obtained iteratively from the expression

$$\begin{aligned} \bar{\gamma}_P &= \frac{1}{H_P^*} \int_{Q_0}^Q \gamma dH^* \approx \gamma_P \left[1 - \left(1 + f + \frac{\omega^2 a}{\gamma_a} - 2f \sin^2 \phi_P \right) \frac{H_P^*}{a} \right. \\ &\quad \left. + \left(\frac{H_P^*}{a} \right)^2 \right] \end{aligned} \quad (11)$$

where ϕ_P is the geodetic latitude of point P, a, f, ω are the semimajor axis, flattening and rotational rate of the reference ellipsoid, and γ_a is the normal gravity at the equator.

Vertical Datum Definition and Realization

In the following, we will use the orthometric height system, which utilizes the geoid as its surface of zero heights, although it is also possible to work with normal heights, height anomalies, and the telluroid or the quasi-geoid. The reason for this choice is that, even in the era of GNSS positioning, most countries do not store their gravity data as gravity disturbances δg but as gravity anomalies Δg , and also because it is easier to solve the Stokes geodetic BVP rather than the Molodensy geodetic BVP; see definitions and equations in Sideris, ► “[Geoid Determination, Theory and Principles](#)”. Vertical datum unification can also be made using just geopotential numbers instead of a particular type of height system. This has the advantage that all developments can be made in terms of geopotential numbers or potential differences and then be converted into the desirable height system using the equations in the preceding section. In order to keep this entry short, however, these alternatives will not be discussed here, and the interested reader is referred to Sánchez and Sideris (2017) and Sansò et al. (2019).

Traditionally, regional and national (orthometric) vertical datums have been realized by selecting a *fundamental point* (or *origin*) $\bar{P}_0$ on the geoid surface of constant gravity potential W_0 , with $H_{\bar{P}_0} = 0, W_{\bar{P}_0} = W_0$. $\bar{P}_0$ is typically at, or connected to, a coastal tide gauge. Through levelling observations connected to $\bar{P}_0$, orthometric height differences and then orthometric heights are derived for all the points of a levelling network. This choice is due to the historical preference of measuring heights above Mean Sea Level (MSL) and the use of the geoid – the gravity field equipotential surface that best approximates the idealized MSL – as the vertical datum for orthometric heights, even though it is well known that the MSL is not an equipotential (level) surface. In order to be able to unify (connect) vertical datums referring to two different level surfaces, knowledge of their potential difference is required. Vertical datum unification by use of the MSL is in fact possible by the *ocean method* (or *ocean levelling method*), which relies on the connection of the MSL at various tide gauges by the mean dynamic topography (MDT), i.e., the height of the average sea surface above the geoid (see, e.g., Rummel and Ilk 1995; Woodworth et al. 2012). Here, we will concentrate on the *geodetic method* that uses as observables gravity measurements, orthometric heights from levelling, and ellipsoidal heights from GNSS, and determines the geoid by solving the Stokes geodetic BVP; see also Heck and Rummel (1990) and Sideris (2014).

Geodetic BVP Solutions for Single and Multiple Vertical Datums

Solution to the single (global) vertical datum geodetic BVP.

We will use here the more general version of the Stokes geodetic BVP than the one in Sideris, ► “[Geoid Determination, Theory and Principles](#)”, for an arbitrary geocentric reference ellipsoid. Namely, we assume that the normal potential of the reference ellipsoid U_0 differs from the gravity potential of the geoid W_0 , i.e.,

$$\Delta W_0 = W_0 - U_0 \quad (12)$$

and that the mass of the reference ellipsoid differs from the mass of the geoid by an amount δM . In this case, Eqs. (12) and (13) in Sideris, ► “[Geoid Determination, Theory and Principles](#)”, become (see also Hofmann-Wellenhof and Moritz 2006, Chap. 2)

$$N = \frac{k\delta M}{R\gamma} - \frac{\Delta W_0}{\gamma} + \frac{T}{\gamma} = N_0 + \frac{T}{\gamma} \quad (13)$$

$$\Delta g = -\frac{\partial T}{\partial h} + \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} (T - \Delta W_0) \quad (14)$$

where k is Newton’s gravitational constant and R is the mean radius of the Earth.

The *generalized Bruns’s formula* of Eq. (13) above can be used to estimate the geoid undulation from T , and Eq. (14) is the boundary condition relating the unknown T to the data Δg on the boundary surface of the geoid Γ .

T is obtained as the solution of the Stokes geodetic BVP, which is defined by Laplace’s equation together with the boundary condition of Eq. (14) on Γ . In spherical approximation, by using a sphere with radius $r = R$, where R is the mean radius of the Earth, we obtain

$$\frac{1}{\gamma} \frac{\partial \gamma}{\partial h} \approx \frac{1}{\gamma} \frac{\partial \gamma}{\partial r} = -\frac{2}{R} \quad (15)$$

The single (global) vertical datum geodetic BVP in spherical approximation is then defined as follows:

$$\begin{aligned} \nabla^2 T &= 0, & \text{outside } \Gamma \\ -\frac{\partial T}{\partial r} - \frac{2}{R} T &= \Delta g - \frac{2}{R} \Delta W_0, & \text{on } \Gamma \\ T &= 0, & \text{at infinity} \end{aligned} \quad (16)$$

For a computational point P, the solution for T is given (spherical approximation, spherical boundary surface) by Eq. (17) in Sideris, ► “[Geoid Determination, Theory and Principles](#)”, i.e.,

$$T_P = \frac{R}{4\pi} \iint_{\sigma} \Delta g S(\psi) d\sigma = S \Delta g \quad (17)$$

Here, instead of Eq. (18) in Sideris, ► “Geoid Determination, Theory and Principles”, we will use the spherical harmonic expansion of $S(\psi)$ as it more appropriate for the developments in the section that follows:

$$S(\psi) = \sum_{n=2}^{\infty} \frac{2n+1}{n-1} P_n(\cos \psi) \quad (18)$$

where σ is the unit sphere, ψ is the spherical distance between data and computation point, P_n is the Legendre polynomial of degree n , and $S(\psi)$ is the Stokes function.

The geoid undulation at P can finally be obtained from the generalized Bruns’s Eq. (13):

$$N_P = \frac{k\delta M}{R\gamma} - \frac{\Delta W_0}{\gamma} + \frac{1}{\gamma} S \Delta g = N_0 + \frac{1}{\gamma} S \Delta g \quad (19)$$

The evaluation of the term $\gamma^{-1} S \Delta g$ in Eq. (19) is not computed by evaluating Eq. (17) directly but by the remove-compute-restore technique presented in Sideris, ► “Geoid Determination, Computational Methods” (Eq. 1). Here, for brevity, and also because we have assumed that the gravity anomalies have already been reduced onto the boundary surface, we do not provide the equations for the contribution of the topography even though they should not be omitted in areas of rough terrain. A reader interested in the detailed procedures and formulas should consult the abovementioned entries in this encyclopedia, and Sansò and Sideris (2013, Chaps. 4 and 8).

The long wavelengths of the gravity field are first removed from the data by use of geopotential model (GM), then the residual anomalies are the input to Stokes’s integral to produce residual undulations, and then the long wavelength part of the geoid is computed from the same GM and restored. The GM-contributions to Δg and N , i.e., Δg_{GM} and N_{GM_P} ,

respectively, are computed from Eqs. (2) and (3) in Sideris, ► “Geoid Determination, Computational Methods”. The residual quantities are obtained from

$$\Delta g_{res} = \Delta g - \Delta g_{GM} \quad (20)$$

$$N_{resP} = \frac{1}{\gamma} S_{res} \Delta g_{res} \quad (21)$$

where the *residual Stokes kernel*

$$S_{res}(\psi) = S(\psi) - \sum_{n=2}^{n_{max}} \frac{2n+1}{n-1} P_n(\cos \psi) \quad (22)$$

is employed in order to match the maximum degree and order n_{max} of the GM used, and the integration in Eq. (17) is carried out only to the corresponding spherical cap ψ_o around the computation point P. The N value is then obtained as

$$N_P = N_{GM_P} + N_{resP} \quad (23)$$

Solution to the multiple vertical datums geodetic BVP.

We assume that the whole sphere σ is covered by $J+1$ non-overlapping vertical datum zones $\sigma^j, j = 0, 1, 2, \dots, J$ (Rummel and Teunissen 1988). Each of these vertical datums is defined by the level surface with potential W_0^j passing through a fundamental point $\bar{P}_0^j$ (see Fig. 3). The zone σ^0 defined by $W_0^0 = W_0$ and $\bar{P}_0^0 = \bar{P}_0$ is chosen arbitrarily as the reference datum with respect to which the geopotential differences

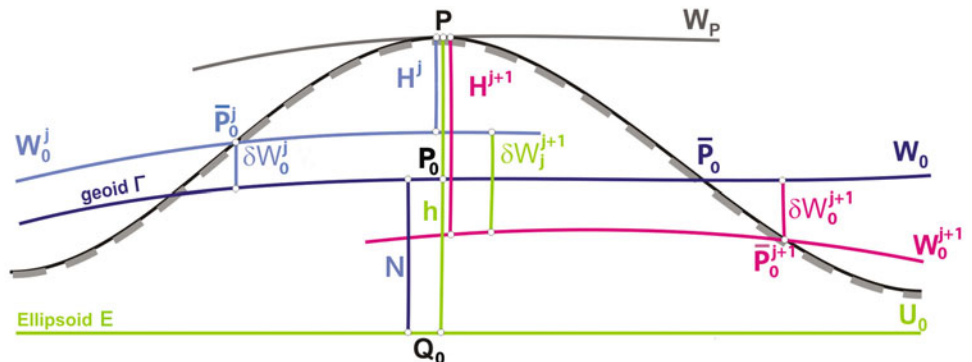
$$\delta W_0^j = W_0 - W_0^j, \quad j = 1, 2, \dots, J \quad (24)$$

and the associated vertical datum offsets will be computed. Since we have

$$\begin{aligned} \Delta W_0^j &= W_0^j - U_0 = W_0^j - U_0 + W_0 - W_0 = \Delta W_0 - \delta W_0^j, \\ j &= 1, 2, \dots, J \end{aligned} \quad (25)$$

Height Systems, Vertical Datums and Their Unification,

Fig. 3 Geometry of multi datum geodetic BVP and datum offsets ($\bar{P}_0^j, W_0^j$ and $\delta W_0^j, \delta W_0^{j+1}$ denote datum fundamental points, datum potentials, and datum offsets, respectively)



$$\Delta g^j = \Delta g - \frac{2}{R} \delta W_0^j, \quad j = 1, 2, \dots, J \quad (26)$$

where, for simplicity, we assume the same reference ellipsoid for all datums, we see that with respect to the corresponding values for the reference datum, in datum j not only the geoid undulations N^j and the orthometric heights H^j will be biased but so will the gravity anomalies, as well (Sideris 2014). Specifically, Δg^j are biased with respect to those on the reference datum by the second term in the right hand side of Eq. (26), which can be considered as the “free-air” reduction needed to reduce Δg^j from the zero-height level W_0^j to the reference level W_0 . The more general case of different datums having different reference ellipsoids can be found, in ellipsoidal approximation, in Sansò and Usai (1995) and in Sansò and Venuti (2002).

By use of the above equations, the geodetic BVP for $J + 1$ datum zones is defined by the following equations:

$$\begin{aligned} \nabla^2 T &= 0, & \text{outside the boundary} \\ -\frac{\partial T}{\partial r} - \frac{2}{R} T &= \Delta g - \frac{2}{R} \Delta W_0 - \frac{2}{R} \sum_{j=1}^J c_p^j \delta W_0^j, & \text{on the boundary} \\ T &= 0, & \text{at infinity} \end{aligned} \quad (27)$$

where c_p^j indicates whether the computation point P belongs to zone σ^j ($c_p^j = 1$) or not ($c_p^j = 0$).

The solution of the above geodetic BVP using biased gravity anomalies is

$$\begin{aligned} T_p^j &= S \Delta g^j = S \left(\Delta g - \frac{2}{R} \sum_{j=1}^J c_p^j \delta W_0^j \right) \\ &= S \Delta g - \frac{2}{R} \sum_{j=1}^J \delta W_0^j I_p^j = T_p - \delta T_p^j \end{aligned} \quad (28)$$

from which we obtain T as

$$T_p = T_p^j + \delta T_p^j = S \Delta g^j + \frac{2}{R} \sum_{j=1}^J \delta W_0^j I_p^j \quad (29)$$

where

$$I_p^j = \frac{R}{4\pi} \iint_{\sigma^j} S(\psi) d\sigma \quad (30)$$

The geoid undulation using biased gravity anomalies referring to datum j is obtained by Bruns's equation

$$\begin{aligned} N_p &= \frac{k\delta M}{R\gamma} - \frac{\Delta W_0^j}{\gamma} + \frac{T_p}{\gamma} = \frac{k\delta M}{R\gamma} - \frac{\Delta W_0}{\gamma} + \frac{\delta W_0^j}{\gamma} + \frac{1}{\gamma} S \Delta g^j \\ &\quad + \frac{1}{\gamma} \frac{2}{R} \sum_{j=1}^J \delta W_0^j I_p^j \\ &= N_0 + \delta N^j + \frac{T_p^j}{\gamma} + \frac{\delta T_p^j}{\gamma} \end{aligned} \quad (31)$$

Hence the use of biased gravity anomalies in the solution of the geodetic BVP introduces two effects in the derived geoid: the *vertical datum offset* $\delta N^j = \delta W_0^j / \gamma$ which is a *direct bias* term on N , and $\delta T_p^j / \gamma$ which is the effect of an *indirect bias* term on N . Writing this last term explicitly as

$$\delta N_p^{j,\text{ind}} = \frac{\delta T_p^j}{\gamma} = \frac{1}{\gamma} \frac{2}{R} S \delta W_0^j = \frac{2}{R} \delta N^j I_p^j + \frac{2}{R} \sum_{i=1, i \neq j}^J \delta N^i I_p^i \quad (32)$$

we can see that the geoid height of point P in the datum zone σ^j is affected by the offsets δN^j (or δW_0^j) of all J datum zones and not only by the offset of the j^{th} one.

As in the single-datum geodetic BVP solution, the evaluation of the term $N_p^j = \gamma^{-1} T_p^j$ in Eq. (31) is not computed by Eq. (17) directly as $\gamma^{-1} S \Delta g^j$ but by the remove-compute-restore technique. The residual quantities are obtained from

$$\Delta g_{\text{res}}^j = \Delta g^j - \Delta g_{\text{GM}}^j \quad (33)$$

$$N_{\text{resp}}^j = \frac{1}{\gamma} S_{\text{res}} \Delta g_{\text{res}}^j \quad (34)$$

using the residual Stokes kernel of Eq. (22) and the N_p^j value is then obtained as

$$N_p^j = N_{\text{GMp}}^j + N_{\text{resp}}^j \quad (35)$$

It is very important to note that the residual Stokes kernel should also be used in Eqs. (30) and (32) for the computation of the indirect effect, which should be evaluated as

$$\delta N_p^{j,\text{ind}} = \frac{\delta T_p^j}{\gamma} = \frac{1}{\gamma} \frac{2}{R} S_{\text{res}} \delta W_0^j = \frac{2}{R} \delta N^j I_{\text{resp}}^j + \frac{2}{R} \sum_{i=1, i \neq j}^J \delta N^i I_{\text{resp}}^i \quad (36)$$

Finally, Eq. (31) can now be evaluated as

$$N_p = N_0 + \delta N^j + N_{\text{GMp}}^j + N_{\text{resp}}^j + \delta N_p^{j,\text{ind}} \quad (37)$$

The indirect bias term $\delta N_p^{j,\text{ind}}$ vanishes only if T is derived from unbiased gravity anomalies obtained from, e.g., a satellite-only GM. This, however, creates a non-negligible omission error in the computed value of N . Fortunately, the use of the residual Stokes kernel in Eq. (36) results in an indirect bias term that can safely be omitted as it has been shown that for any GM with $n_{\text{max}} \geq 70$, this term is below 1 cm. Numerical results for both the omission error and the indirect bias term can be found in Sideris and Fotopoulos (2012), Sideris (2014), and Sánchez and Sideris (2017).

Determination of Vertical Datum Offsets

In order to be able to determine all J unknown datum offsets, namely, δW_0^j or δN^j , $j = 1, 2, \dots, J$, together with ΔW_0 or N_0 , at least one data point P per datum zone should be given with known ellipsoidal height h_P determined from GNSS positioning and orthometric height H_P^j from levelling in the datum σ^j . This can be, for example, the fundamental point/tide gauge of each datum; see discussion in Sideris (2014). We can then form the following *observation equation* for each point P using N_P^j from Eq. (37) into Eq. (4):

$$l_P = h_P - H_P^j - N_{\text{GM}_P} - N_{\text{resp}}^j = N_0 + \left(1 + \frac{2}{R} I_{\text{resp}}^j\right) \delta N^j + \frac{2}{R} \sum_{i=1, i \neq j}^J I_{\text{resp}}^i \delta N^i + v_P \quad (38)$$

where l_P denotes the “observable” and v_P denotes the stochastic residual.

When more points than the minimum required are available, the solution is obtained by least-squares adjustment. For a set of $n \geq J + 1$ observation equations, the matrix form of the functional model of Eq. (38) is

$$l = A \delta N + v \quad (39)$$

with corresponding stochastic model

$$E\{v\} = 0, E\{vv^T\} = C_l = C_h + C_H + C_{N_{\text{GM}}} + C_{N_{\text{res}}} \quad (40)$$

where $E\{\cdot\}$ is the expectation operator, A is the design matrix, and C are variance-covariance matrices. The least-squares solution of the vector of datum offsets and their variance-covariance matrix is

$$\widehat{\delta N} = (A^T C_l^{-1} A)^{-1} A^T C_l^{-1} l, \widehat{C}_{\delta N} = (A^T C_l^{-1} A)^{-1} \quad (41)$$

The above is the general least-squares solution that can be applied to the global unification of height systems. It can also be used with, besides tide gauges, observations from GNSS/levelling benchmarks and from ocean levelling, either separately for each data group or in a combined fashion adjusting together all available data groups.

Solution simplifications and practical considerations.

Since, as already mentioned, the indirect bias term can be rendered negligible for cm-level accuracy by choosing a GM with $n_{\text{max}} \geq 70$, Eq. (37) can be simplified to

$$l_P = h_P - H_P^j - N_{\text{GM}_P} - N_{\text{resp}}^j = N_0 + \delta N^j + v_P \quad (42)$$

which makes the least-squares solution much easier to implement.

Either because of lack of information or to improve the efficiency of the computations, another simplification often applied in practice without significant loss of accuracy is the use of diagonal or block-diagonal variance-covariance matrices in Eq. (41). Examples can be found in Gerlach and Fecher (2012) and Rangelova et al. (2016).

Another often-used simplification of Eq. (38) is to set N_0 or ΔW_0 equal to zero. This is equivalent to selecting a conventional value for W_0 , either by setting W_0 equal to U_0 , or by computing W_0 with data at tide gauges or in the oceans. The approach relies on the geoid being the level surface around which the sea surface heights (SSH), H_S , are near zero when averaged globally over all oceanic regions σ_s (Gauss-Listing geoid). Mathematically, we require that

$$\frac{\partial}{\partial W_0} \iint_{\sigma_s} H_S^2 d\sigma \approx \frac{\partial}{\partial W_0} \iint_{\sigma_s} \left(\frac{W_0 - W}{\gamma} \right)^2 d\sigma = 0 \quad (43)$$

which results in the following expression for W_0 :

$$W_0 = \frac{\iint_{\sigma_s} \frac{W}{\gamma^2} d\sigma}{\iint_{\sigma_s} \frac{1}{\gamma^2} d\sigma} \quad (44)$$

The SSH is typically derived from altimetry over the oceans and W_0 can be estimated from gravity data. There are many problems to be resolved in this approach, such as the spatial coverage/integration area and choice of ocean models (typically the polar regions are left out), the choice of reference epoch and long-term variations of MSL, the choice (ideally, satellite-only) and maximum degree of a GM, plus the ones mentioned in the next paragraph (see also Heck 2005). Although this approach has its proponents (see, e.g., Burša et al. 2008; Sánchez 2008), it has not been

adopted here, on the one hand because of the inherent problems just mentioned and, on the other hand, because selecting a particular W_0 is not really necessary for global height unification, unless governments decide to reference all their data to a unified global vertical datum, which seems unlikely at this point in time.

In general, when estimating vertical datum offsets, besides knowing the error characteristics of all data sets that enter into the least-squares adjustment, care must be taken so that all data are consistent in terms of observation epoch, global terrestrial reference frame (e.g., ITRF), reference ellipsoid, height type, tide system, and crustal motion and sea-level change effects, wherever present. For example, in areas with significant crustal motion such as Canada and Fennoscandia, heights of tide gauge benchmarks should be corrected using GNSS-derived vertical crustal velocities or geophysical models of crustal motion due to glacial isostatic adjustment (Rangelova et al. 2012; Peltier et al. 2015).

Summary and Outlook

We have briefly outlined how height systems and vertical datums are defined and implemented, how they are related to the Earth's gravity field, and how they can be unified by employing the solution to the multi-datum geodetic BVP. We have shown in detail how vertical datum unification can be achieved when ellipsoidal heights from GNSS or altimetry, orthometric heights from geodetic or oceanographic levelling, and geoid heights from global geopotential models and gravity anomalies are available at each fundamental point, and/or other stations, in each of the existing national or regional vertical datums. The functional and stochastic models used for the determination of the offsets between existing vertical datums by least-squares adjustment have also been discussed, together with the theoretical and practical issues that must be considered in this estimation process. Numerical results on the unification of regional and continental vertical datums have been published in the last two decades. Such examples are the articles by Featherstone and Filmer (2012) for Australia, Amos and Feathersone (2009) and Tenzer et al. (2011) for New Zealand, Sánchez and Sideris (2017) for South America, and Sideris and Fotopoulos (2012) and Rummel et al. (2014) for Europe and North America.

Although global vertical datum unification has not been attempted systematically yet, countries in Europe, North America, and Oceania have started abandoning traditional levelling networks and adopting geoid (or quasi-geoid)-based vertical datums realizable through a geoid surface and GNSS positioning. Examples of such height modernization efforts can be found at <https://www.nrcan.gc.ca/height-reference-system-modernization/9054> for Canada, and at <https://www.ngs.noaa.gov/heightmod/> for the USA. The

effort for global unification of all of these datums is already underway, coordinated by the International Association of Geodesy (IAG) and its Global Geodetic Observing System (GGOS). A strategy for the realization of the International Height Reference System (IHRF) has already been developed (Ihde et al. 2017; Sánchez and Sideris 2017), and efforts for its implementation are currently underway by the international scientific community (IAG Joint Working Group on the Implementation of the International Height Reference Frame, IHRF) and national government agencies.

Cross-References

- Geodesy, Figure of the Earth
- Geodesy, Ground Positioning, and Levelling
- Geodesy, Networks, and Reference Systems
- Geodesy, Physical
- Geoid
- Geoid Determination, Computational Methods
- Geoid Determination, Theory and Principles
- Gravity Field of the Earth
- Gravity, Global Models
- Spherical Harmonic Analysis Applied to Potential Fields

Bibliography

- Amos MJ, Feathersone WE (2009) Unification of New Zealand's local vertical datums: iterative gravimetric quasigeoid computations. *J Geod* 83:57–68
- Burša M, Kouba J, Raděj K, True SA, Vatr V, Vojtišková M (2008) Mean Earth's equipotential surface from Topex/Poseidon altimetry. *Studia Geoph et Geod* 42(4):459–466
- Featherstone WE, Filmer MS (2012) The north–south tilt in the Australian Height Datum is explained by the ocean's mean dynamic topography. *J Geophys Res* 117:C08035
- Gerlach C, Fecher F (2012) Approximations of the GOCE error variance-covariance matrix for least-squares estimation of height datum offsets. *J Geod Sci* 2(4):247–256
- Heck B (2005) Problems in the definition of vertical reference frames. In: *Proceedings of V Hotine–Marussi symposium on mathematical geodesy*. IAG symposia, vol 127. Springer, Berlin, pp 164–173
- Heck B, Rummel R (1990) Strategies for solving the vertical datum problem using terrestrial and satellite geodetic data. In: *Proceedings of sea surface topography and the geoid*. IAG symposia, vol 104. Springer, Berlin, pp 116–128
- Hofmann-Wellenhof B, Moritz H (2006) *Physical geodesy*, 2nd edn. Springer, Wien/New York
- Ihde J, Sánchez L, Barzaghi R, Drewes H, Foerste C, Gruber T, Liebsch G, Marti U, Pail R, Sideris M (2017) Definition and proposed realization of the International Height Reference System (IHRF). *Surv Geophys* 38:549–570
- Jekeli C (2000) Heights, the geopotential and vertical datums. OSU report no. 459. Department of Civil and Environmental Engineering and Geodetic Science, Ohio State University, Columbus
- Peltier WR, Argus DF, Drummond R (2015) Space geodesy constrains ice age terminal deglaciation: the global ICE-6G_C (VM5a) model. *J Geophys Res Solid Earth* 120:450–487

- Rangelova E, van der Wal W, Sideris MG (2012) How significant is the dynamic component of the north American vertical datum? *J Geod Sci* 2(4):281–289
- Rangelova E, Sideris MG, Amjadiparvar P, Hayden T (2016) Height datum unification by means of the GBVP approach using tide gauges. In: *Proceedings of VIII Hotine-Marussi symposium on mathematical geodesy*. IAG symposia, vol 142. Springer, Berlin, pp 121–130
- Rummel R, Ilk KH (1995) Height datum connection – the ocean part. *Allgemeine Vermessungsnachrichten* 8(9):321–330
- Rummel R, Teunissen P (1988) Height datum definition, height datum connection and the role of the geodetic boundary value problem. *Bull Géod* 62:477–498
- Rummel R, Gruber T, Sideris MG, Rangelova E, Woodworth P, Hughes C, Ihde J, Liebsch G, Rülke A, Schafer U (2014) STSE–GOCE+: height system unification with GOCE – summary and final results. ESA study contract report, Doc. Nr. GO-HSU-PR-0021, issue 1. www.gocplushsu.eu/ext/doc/fin-doc/GO-HSU-RP-0021_1.0_Final_Summary_Report.pdf
- Sánchez L (2008) Approach for the establishment of a global vertical reference level. In: *Proceedings of VI Hotine-Marussi symposium on theoretical and computational geodesy*. IAG symposia, vol 132. Springer, Berlin, pp 119–125
- Sánchez L, Sideris MG (2017) Vertical datum unification for the International Height reference System (IHRs). *Geophys J Int* 209:570–586
- Sansò F, Sideris MG (eds) (2013) *Geoid determination – theory and methods*. Lecture notes in Earth science, vol 110. Springer, Berlin/Heidelberg
- Sansò F, Usai S (1995) Height datum and local geodetic datum in the theory of geodetic boundary value problems. *Allgemeine Vermessungsnachrichten* 8–9:343–385
- Sansò F, Venuti G (2002) The height datum/geodetic datum problem. *Geophys J Int* 149:768–775
- Sansò F, Reguzzoni M, Barzaghi R (2019) *Geodetic Heights*. Springer Geophysics. Springer, Cham
- Sideris MG (2014) Geodetic world height system unification. In: Freedon W, Nashed MZ, Sonar T (eds) *Handbook of geomathematics*, 2nd edn. Springer, Berlin Heidelberg, pp 3067–3085
- Sideris MG, Fotopoulos G (eds) (2012) Special issue on regional and global geoid-based vertical datums. *J Geod Sci* 2(4):246–376
- Tenzer R, Vatr V, Gan L, Abdalla A, Dayoub N (2011) Combined approach for the unification of leveling networks in New Zealand. *J Geod Sci* 1(4):324–332
- Véronneau M, Huang J (2016) The Canadian geodetic vertical datum of 2013 (CGVD2013). *Geomatica* 70(1):9–19
- Woodworth PL, Hughes CW, Bingham RJ, Gruber T (2012) Towards worldwide height system unification using ocean information. *J Geod Sci* 2(4):302–318

High-Frequency Seismology

Simanchal Padhy
Seismological Observatory, CSIR-National Geophysical
Research Institute, Hyderabad, India

Synonyms

Elastic waves in heterogeneous media

Definition

By high-frequencies, we mean spatial variations in elastic properties of the medium over a distance scale much larger than the wavelength of seismic waves (i.e., heterogeneities with a scale length less than the seismic wavelength), which includes seismic wave propagation on a local to regional scale. Frequency is, therefore, an important parameter for broadly classifying seismology into three categories, each with a different scale of investigation: low-frequency (<20 mHz), high-frequency ($20\text{--}50$ mHz $< f < 1\text{--}30$ Hz), and exploration (>10 Hz) seismology. High-frequency signals are noncoherent (random), additive in terms of power/energy instead of amplitudes ignoring phase: they can be band-passed with a zero-phase filter and squared to obtain their energy content. Their waveform shapes are highly complex, where the complexity could stem from heterogeneity in source rupture process and/or complicated wave propagation effect. Source properties such as spatial variations of rupture velocity, slip velocity and stress drop, and complex wave propagation (structure) effects like reflection, refraction at interfaces separating two different media, multipath arrivals over different azimuths, focusing/defocusing of energy due to lateral changes in medium properties, attenuation (anelastic) and scattering of seismic waves could lead to the observed waveform complexity. The wave gets amplified or/and attenuated near the recording sites as they pass through unconsolidated near-surface material. Thus, accounting accurately for the improved source, wave propagation, site effects including the soil nonlinearity in ground-motion modeling offers potential benefits in more realistic region-specific seismic hazard assessment.

Introduction

The propagation of high-frequency seismic waves in a 3-D laterally heterogeneous elastic medium is a challenging problem. The complexity of resulting seismogram results from mode conversions, multipath, anisotropy, attenuation, and scattering of seismic waves and their superposition from small-scale heterogeneities encountered during wave propagation. Superposition of scattered waves results in a long-duration complex wave train following the direct wave arrival called as the coda. Given a multitude of wave propagation phenomena for their origin, a better understanding of their source of complexity is essential for detailed imaging of medium heterogeneities that requires proper identification, characterization, and origin of coda waves. The high-frequency seismic waves are very sensitive to small-scale heterogeneities in the lithosphere. These heterogeneities strongly influence seismic waves; their effects on the waveform can be studied through (1) propagation and attenuation

(decay in amplitude with distance) of direct (body and surface) and coda waves recorded at local to regional distances through the lithosphere, and (2) fluctuation in amplitude and phase of teleseismic P-wave train at ~ 1 Hz observed at a closely spaced seismic array. The teleseismic P-wave train, based on spectral analyses, is used for discriminating earthquakes from underground nuclear explosions, for early tsunami warning by fast determining magnitude, duration of events, and source rupture complexity (Wu et al. 2019).

The organization of the chapter is as follows. I present a brief introduction to the basics of seismic wave (both acoustic and elastic) propagation theory with the governing equations, standard methods of data analyses, and finally summary and outlook.

Basic Wave Propagation Principle

The fundamental theory of wave propagation primarily rests on two equations: Newton's second law (acceleration is proportional to the divergence of the stress tensor), and generalized Hooke's constitutive law stating the proportionality between stress and strain in 1-D, with an introduction of stress tensor or traction in higher dimensions (2-D and 3-D).

Acoustic Wave Propagation

For the acoustic approximation, the dominant wave type is a compressional (or longitudinal) wave with no shear contribution, governed by the following equations.

$$P = -\kappa \nabla \cdot \mathbf{u} \quad (\text{Hooke's law}), \quad (1)$$

$$\frac{\partial P}{\partial x} = -\rho(x, y, z) \frac{\partial^2 \mathbf{u}}{\partial t^2} \quad (\text{Newton's law}), \quad (2)$$

where $\mathbf{u} = (u, v, w)$ are the displacement field-components along the (x, y, z) axes, $P(x, y, z, t)$ is the spatio-temporal change in acoustic pressure wave (say, the pressure exerted by medium on face of a cube), $\kappa(x, y, z)$ is the spatially varying bulk modulus, and $\rho(x, y, z)$ is the density distribution of medium. The acoustic wave equation can be expressed as:

$$\nabla^2 P - \frac{1}{V^2} \frac{\partial^2 P}{\partial t^2} = -\frac{\rho(x, y, z)}{\kappa(x, y, z)} \ddot{f}(x, y, z, t), \quad (3)$$

$f(x, y, z, t)$ is the spatio-temporal variation of body force, double dots over it represent double time derivative, $V = \sqrt{\frac{\kappa(x, y, z)}{\rho(x, y, z)}}$ is wave speed. The solution of wave equation in 1-D is $P(x, t) = A_0 e^{i(kx - \omega t)}$, where the function $e^{i(kx - \omega t)}$ describes a right-going plane wave, k is wave number; $k^2 = (\omega/V)^2$ defines the dispersion relation linking the

temporal and spatial variables in the frequency domain. For a point source, Eq. (3) becomes:

$$\begin{aligned} \nabla^2 G(\mathbf{x}, t|\mathbf{x}', t') - \frac{1}{V^2} \frac{\partial^2 G(\mathbf{x}, t|\mathbf{x}', t')}{\partial t^2} \\ = -\delta(\mathbf{x} - \mathbf{x}') \delta(t - t'), \end{aligned} \quad (4)$$

where $\delta(\mathbf{x} - \mathbf{x}')$ is the spatially delayed Dirac delta function, with the source location at $\mathbf{x}'$, with its temporal counterpart $\delta(t - t')$ with the source excitation time t' , and t is time variable. The $G(\mathbf{x}, t|\mathbf{x}', t')$ is the response of the medium, defined between the point source $(\mathbf{x}', t')$ and the receiver $(\mathbf{x}, t)$, to the impulsive point source called as the Green's function. The solution to Eq. (4) can be expressed as:

$$G(\mathbf{x}, t|\mathbf{x}', 0) = -\frac{e^{i(k|\mathbf{x} - \mathbf{x}'| - \omega t)}}{4\pi |\mathbf{x} - \mathbf{x}'|}, \quad (5)$$

where $1/|\mathbf{x} - \mathbf{x}'| = 1/r$ is the geometrical spreading factor for body waves; the solution is similar to plane wave solution, except the wave fronts are the concentric spheres with the source at the center point $\mathbf{x}$. For a medium with intrinsic attenuation, seismic wave amplitude decays exponentially with distance, and Eq. (5) can take the form $G(\mathbf{x}, t|\mathbf{x}', 0) = \frac{e^{i(kr - \omega t)} e^{-\omega r/2VQ}}{4\pi r}$, where Q is the quality factor describing attenuation.

Elastic Wave Propagation

Wave propagation through any media having nonzero shear strength (elastic materials) more closely approximates waves propagating through the real earth, especially true while studying S- and surface waves of the full wave field. Just like for acoustic waves dealing with the compressional force and dilatational strains (i.e., volume changing), Hooke's law for the elastic medium states the proportionality between stress and strain tensors related to elastic constants.

$$\tau_{ij} = c_{ijkl} \varepsilon_{kl}, \quad (6)$$

where τ_{ij} are the components of the stress tensor (the external force applied per unit area in the i^{th} direction along the surface normal to the j^{th} direction), ε_{kl} are the components of the strain tensor, c_{ijkl} are the components of the elastic stiffness tensor of medium properties, and $i, j, k, l = 1, 2, 3$. There are 81 elastic constants corresponding to 9 different stress and 9 different strains. Symmetries in the stress (i.e., $\tau_{ij} = \tau_{ji}$) and strain tensors (i.e., $\varepsilon_{ij} = \varepsilon_{ji}$) together with conservation of energy and momentum reduce the number of independent constants from 81 to 21 (Aki and Richards 1980). For acoustic media, there is only one independent constant, that is, the bulk modulus. For elastic isotropic

media, there are only two independent constants related through (Aki and Richards 1980):

$$\tau_{ij} = \lambda \delta_{ij} \varepsilon_{kk} + 2\mu \varepsilon_{ij}, \quad (7)$$

where μ is the shear modulus, and λ is Lamé's constant.

Elastic Wave Equation

Newton's law for the elastic media can be obtained by equating the external body forces with the inertial forces as:

$$\rho \ddot{u}_i = \tau_{ij,j} + f_i, \quad (8)$$

where f_i is the i th component of the body force. After expressing stress in terms of strain tensor and other constants, the vectorial form of Eq. (8) (elastic wave equation) can be expressed as:

$$\rho \ddot{\mathbf{u}} = (\lambda + \mu) \nabla(\nabla \cdot \mathbf{u}) + \mu \nabla \cdot \nabla \mathbf{u}, \quad (9)$$

The vector wave field can be split into two potential (scalar and vector) wave fields: $\mathbf{u} = \nabla \phi + \nabla \times \psi$, where ϕ and ψ are the scalar and vector potential, respectively; this along with the constraint $\nabla \cdot \psi = 0$ can decompose Eq. (9) into two potential wave field equations mentioned below:

$$\rho \ddot{\phi} - (\lambda + 2\mu) \nabla^2 \phi = 0, \quad \text{and} \quad (10)$$

$$\rho \ddot{\psi} - \mu \nabla^2 \psi = 0, \quad (11)$$

Equations (10) and (11) above, respectively, correspond to the wave equations for scalar and vector potential wave fields, satisfied by plane P- and S-waves propagating with speeds of $V_p = \sqrt{(\lambda + 2\mu)/\rho}$, and $V_s = \sqrt{\mu/\rho}$.

Theory of Seismic Wave Attenuation and Its Effects on Seismic Waveform

Seismic attenuation is quantified by reciprocal of dimensionless quality factor Q , defined as the ratio of stored energy E to the loss of energy ΔE of a plane wave in one cycle:

$$Q = \frac{4\pi E}{\Delta E} = \frac{2\alpha V}{\omega}, \quad (12)$$

where α is attenuation coefficient, V the propagation velocity, and ω the angular frequency. Attenuation causes the wave amplitude exponentially reduced with distance, x as $e^{-\alpha x}$. The propagation of seismic waves in earth can be approximated by wave propagation in an ideal elastic solid. In reality, earth material significantly differs from elastic solid, however, because earth is absorptive, inelastic, and heterogeneous. Absorption of energy in elastic media is a linear phenomenon

(linear filter) where stress and strain are linearly related as $\sigma(\omega) = M(\omega)\varepsilon(\omega)$, where ω is angular frequency, and σ , ε , and M are stress, strain, and complex modulus, respectively. The filter, represented by M , is causal, meaning that its output σ never precedes input ε . The modulus being complex can be expressed as $M = M_1 + iM_2$ or in amplitude and phase as $M = Ae^{i\phi}$ (A is amplitude, ϕ is phase, and i is imaginary unit). Attenuation is defined as $Q^{-1} = \frac{M_2}{M_1} = \tan(\phi)$. In visco-elastic media, there is a phase shift between them; accordingly elastic modulus characterizing rock material is complex and frequency-dependent.

The velocity, density, and absorption properties of earth materials affect seismic waveforms. To accurately model the seismic wave field in real media, we need to consider seismic attenuation. Effect of attenuation can be implemented in two ways: either by using the complex velocity or by convolving dissipation operators with the synthetics computed for a purely elastic media. Several attenuation models of complex modulus or propagation velocity are available to describe wave propagation in visco-elastic media over a wide frequency range. These models describe wave propagation in terms of Q and phase velocity. Of all the models, the constant- Q , nearly constant- Q (Futterman 1962) models in seismic band are the most popular ones. The standard linear solid (SLS) is preferred in finite-difference modeling of wave propagation as stress and strain are related by differential operators, which allows computing synthetics for models with arbitrary distribution of velocity and Q values (Emmerich and Korn 1987).

In time domain, attenuation in elastic solids can empirically be defined using memory variables, where elastic moduli act as time convolution operators (Ben-Menahem and Singh 1981). The corresponding modulus in frequency domain results in complex propagation velocity giving rise to attenuation accompanied by dispersion; these phenomena are related through Hilbert transform to satisfy the causality of the propagating wavelet (Aki and Richards 1980). Depending on the complexity of the medium, several moduli can be used to model attenuation; for example, superposition of SLSs can be used to approximate constant- Q over a specified frequency range. The causal linear absorption model, where attenuation is accompanied by dispersion, can be defined in terms of frequency-dependent Q and phase velocity, V (Futterman 1962) as:

Attenuation : $Q(\omega)$

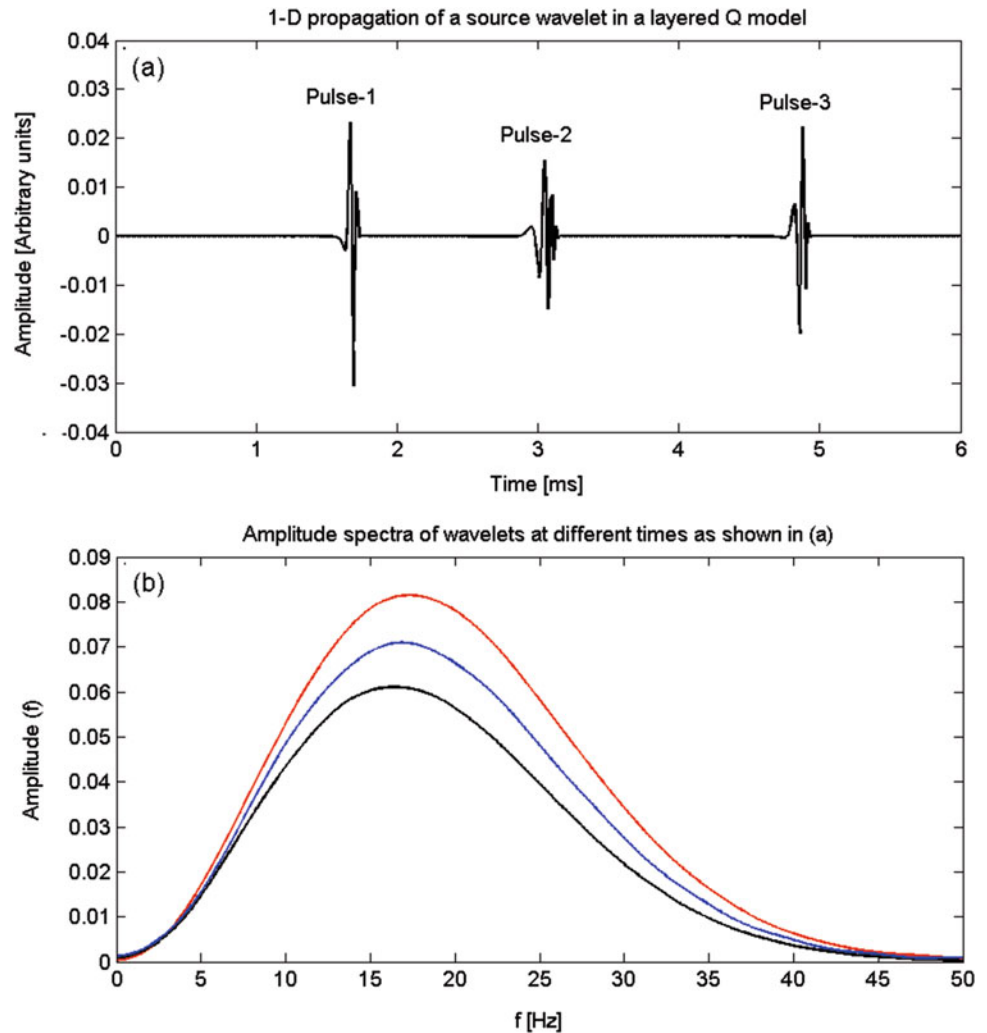
$$= Q(\omega_0) \left[1 - \frac{1}{\pi Q_0} \ln \left(\frac{\omega}{\omega_0} \right) \right], \quad \text{and} \quad (13)$$

Dispersion (frequency – dependence of phase velocity) :

$$V(\omega) = V_l \left[1 - \frac{1}{\pi Q_0} \ln \left(\frac{\omega}{\omega_0} \right) \right] - \frac{i}{2Q(\omega)}, \quad (14)$$

High-Frequency Seismology,

Fig. 1 1-D wave propagation in a layered Q structure. (a) Propagating Ricker source wavelet at different travel times in a 4-layered Q-model as an example ($Q_p = 30, 40, 50$, and 60 ; $Q_s = Q_p/2$) (b) Amplitude spectra of the source wavelet at different times, as shown in (a). Note the reduction in amplitude and loss in frequency with time because of attenuation



where ω_0 is reference frequency, $Q_0 = Q(\omega = \omega_0)$, V is complex propagation velocity, and V_l is layer velocity. This attenuation model can be incorporated into a layered earth model, for example, the plane wave reflectivity algorithm to generate seismic traces. As an example, a 4-layered attenuation structure model can generate plane waves characterized by pulses with diminution amplitude, wider spread (waveform distortion) with time (Fig. 1a); their spectra show reduced high-frequency energy of the pulse at latter time because of attenuation (Fig. 1b). Once Q of the Earth reliably estimated, deconvolution or inverse Q -filtering is performed on the seismic data to restore the high frequencies to enhance seismic resolution.

Methods of Analysis of High-frequency Seismograms

There are several techniques used for analysis of high-frequency seismograms, such as scattering and attenuation of both direct and coda waves, array or polarization analysis,

converted wave analyses, local to regional scale tomography, short-period receiver function analyses, surface wave dispersion measurements including those of ambient noise, and their combinations for imaging media with an improved resolution. Because the literature on each topic mentioned above is huge, I have limited the scope of this chapter to scattering and attenuation of seismic waves at local/regional scale.

Scattering of Seismic Waves in Heterogeneous Media

Scattering from the 3-D heterogeneities with fluctuations in velocity, density, and other elastic constants alters the amplitude and phase of seismic waves, resulting in a wave train consisting of dispersed and dissipated direct arrivals followed by scattered coda waves.

Coda Wave Analysis

Analysis of coda waves is one of the major components of high-frequency seismology for characterizing the small-scale medium heterogeneities. They are more sensitive to changes

in property of medium than the direct arrivals, sampling the medium over a larger volume. The spectral decay of coda wave exhibits similar behavior irrespective of source location, path, and distance travelled. The coda power spectrum defined as function of frequency ω and time t can be expressed as $P(\omega|t) = S(\omega)C(\omega|t)$, where $S(\omega)$ is the source spectrum, and $C(\omega|t)$ accounts for the path and site effect that is independent of distance. For longer lapse times, coda energy is uniformly distributed in a volume surrounding the source and receivers, as empirically evidenced from the common shape of band-passed coda envelopes independent of distance; the amplitude of envelope varies with source strength and site amplification. Thus, by normalizing the source spectral amplitude by coda spectra at a fixed lapse time, one can separate the effects of seismic source, propagation path, and site from the observed data (coda normalization method; Aki 1980). Using this method, the relative site and source effects (spectral differences in source radiation and site amplification, respectively) can be evaluated using observations for different sources and receivers. The coda spectral ratio of event i recorded at two nearby sites j and k (k is reference site) can represent the relative site amplification (site effect) as $\frac{A_{ij}(\omega, t)}{A_{ik}(\omega, t)} = \frac{R_j(\omega)}{R_k(\omega)}$, where coda is defined at late lapse time ($t \geq 2t_s$). Similarly, the coda spectral ratio of two events i and l of the same area recorded at site j can yield relative source radiation (source effect) as $\frac{A_{ij}(\omega, t)}{A_{lj}(\omega, t)} = \frac{S_i(\omega)}{S_l(\omega)}$. The coda amplitude is estimated from band-passed envelopes.

There are several models of coda wave generation: scattered surface waves in a 2-D unbounded, homogeneous and isotropic media (Aki 1969), singly back-scattered (SBS) waves (both surface waves in 2-D and body waves in 3-D) from heterogeneities in an unbounded, homogeneous, and isotropic media for co-located source and receiver and for lapse time more than twice the S-wave arrival time (Aki and Chouet 1975), single isotropically scattered (SIS) body waves in 3-D for a finite distance by taking geometrical spreading correction into account (Sato 1977). The above scattering models assume weak single scattering of seismic waves (Born approximation), violating conservation of energy. For explaining coda in a highly heterogeneous media, we need to invoke multiple (higher orders) scattering for energy conservation to hold well, however. Based on the analyses of S-waves of local earthquakes in >1 Hz, global observations of surface waves in <1 Hz, and S-waves from the borehole seismograms of local earthquakes in >10 Hz, Aki (1980) conjectured a broad peak in Q_s^{-1} in 0.1–10 Hz corresponding to heterogeneities of length scale ranging from a few tens of km to a few hundreds of meters for seismically active regions.

The aforementioned coda models are purely phenomenological (base on empirical observations) lacking any theoretical basis. Thus, these models cannot tell about origin of coda and also cannot interpret coda attenuation in terms of medium

properties (intrinsic absorption versus scattering). Parameter Q_c obtained is a combination of scattering and intrinsic attenuation $\frac{1}{Q_c} = \frac{1}{Q_{sc}} + \frac{1}{Q_i}$, where Q_{sc} and Q_i denote the scattering and intrinsic Q , respectively. Scattering attenuation redistributes energy to other parts of wave field within the medium, while intrinsic attenuation loses energy from wave field. One of the models accounting for both the mechanisms is the energy flux model (EFM) that assumes uniform distribution of scattered coda energy in volume behind the advancing spherical wave front (Frankel and Wennerberg 1987). The coda energy envelope based on EFM can be modeled as $E^{EFM}(r, t|f) = Bt^{-3}e^{-\frac{\omega t}{Q_i}}\left[1 - e^{-\frac{\omega t}{Q_{sc}}}\right]$, where B is a constant. For negligible scattering (large Q_{sc}), EFM approximates to the SBS model of Aki and Chouet (1975) (i.e., $Q_{i,EFM} \sim Q_{c,SBS}$). Depending on frequencies, EFM for negligible scattering does not explain the observed coda envelope well (Fig. 2). Single scattering assumptions severely overestimate the coda decay in 3-D, especially at larger distances where multiple scattering is expected (Sato et al. 2012).

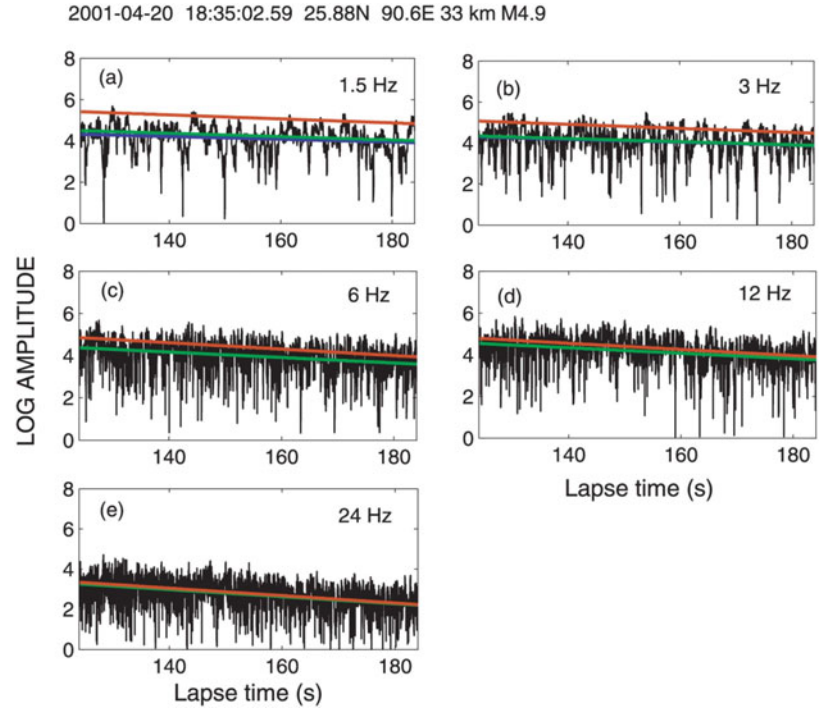
Wu (1985) modeled transport of energy during multiple scattering processes of coda wave formation by applying the radiative transfer theory (RTT) in seismology. RTT describes the spatial and temporal evolution of a wavefield's energy or intensity and ignores the wave interference effect. The general form of radiative transfer equation can be expressed as (Wu 1985):

$$\begin{aligned} \frac{\partial I(\vec{r}, \Omega, t)}{\partial t} + V\hat{n}(\Omega) \cdot \nabla I(\vec{r}, \Omega, t) &= S(\vec{r}, \Omega, t) \\ - \frac{1}{\tau_s} I(\vec{r}, \Omega, t) - \frac{1}{\tau_a} I(\vec{r}, \Omega, t) &+ \frac{1}{\tau_s} \int \frac{\partial \sigma_s}{\partial \Omega'} I(\vec{r}, \Omega', t) d\Omega', \end{aligned} \quad (15)$$

where $I(\vec{r}, \Omega, t)$ is the intensity at position $\vec{r}$ propagating in direction Ω , V is the group velocity of the average (coherent) wave field, $\hat{n}$ is the unit vector in the direction of propagation, and $S(\vec{r}, \Omega, t)$ is the source function. The differential scattering cross section $\frac{\partial \sigma_s}{\partial \Omega'}$ means the energy exchanged while travelling from one direction Ω into other Ω' , τ_s is the scattering mean free time defined as the time lapse between the energy exchanges. The left-hand side of Eq. (15) represents time variation of the intensity, while right side shows the growth and decay of energy with a source term, S . At longer lapse times (several mean free times), Eq. (15), for multiple scattering, approximates to diffusion regime, where coda waves are equi-partitioned among all modes and propagation directions with a stabilized S-to-P energy ratio, while with large fluctuations in ratio at background noise level (Fig. 3a and b) (Paul et al. 2005; Margerin et al. 2009). The diffuse coda waves are characterized by a prolonged duration and a very slow coda decay (strong scattering and very low intrinsic absorption) (Sato et al. 2012).

High-Frequency Seismology,

Fig. 2 Comparison of observed coda envelope with synthetic derived from EFM at different frequencies. The EFM synthetics are shown for three different values of Q_{sc} (100: blue, 1000: green, and 10,000: red) (after Padhy and Subhadra 2010; copyright of 2020 Oxford University Press)



Since Wu's introduction of RTT in seismology, the theory has extensively been developed based on the stochastic treatment of the wave equation and numerically solved to explain propagation of both acoustic and elastic waves in random media at different scales of investigation (Padhy et al. 2007; Przybilla et al. 2006; Sato et al. 2012 and references therein). The degree of heterogeneity of lithosphere expressed in terms of isotropic scattering coefficient (1/mean free path) of S-waves varies in the range 10^{-3} to $2 \times 10^{-2} \text{ km}^{-1}$ in 1–20 Hz and is $O(10^{-4} \text{ km}^{-1})$ in sublithosphere mantle at lower frequencies (0.1–1 Hz) (Fig. 4) (Sato 2019).

A more generalized energy density of a nonabsorbing 3-D multiple scattering processes satisfies the following integral equation for energy transport (Zeng et al. 1991):

$$E(r, t|f) = \frac{W(f)\Theta(\theta, \phi)}{4\pi V r^2} H(t) \delta\left(t - \frac{|r|}{V}\right) e^{-\eta_t(f)Vt} + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \eta_{sc}(f) E(r', t'|f) e^{-V(t-t')\eta_t(f)} \times \frac{4\pi |r - r'|^2}{4\pi |r - r'|^2} \times \delta\left(t - t' - \frac{|r - r'|}{V}\right) dt' dr', \quad (16)$$

where $W(f)$ is source energy in a given frequency band, $\eta_{sc}(f) = \frac{\omega}{Q_{sc}V}$ and $\eta_t(f) = \frac{\omega}{Q_t V}$, are scattering, and total

attenuation coefficients, respectively, $\eta_t = \eta_{sc} + \eta_i$ with $\eta_i(f) = \frac{\omega}{Q_i V}$ representing intrinsic absorption coefficient, $\Theta(\theta, \phi)$ is source radiation, and V is background velocity. The first term of Eq. (16) corresponds to direct wave propagation, while the second integral term scattered wave propagation. The second term can be approximated for the single isotropic scattering (SIS) as (Sato 1977):

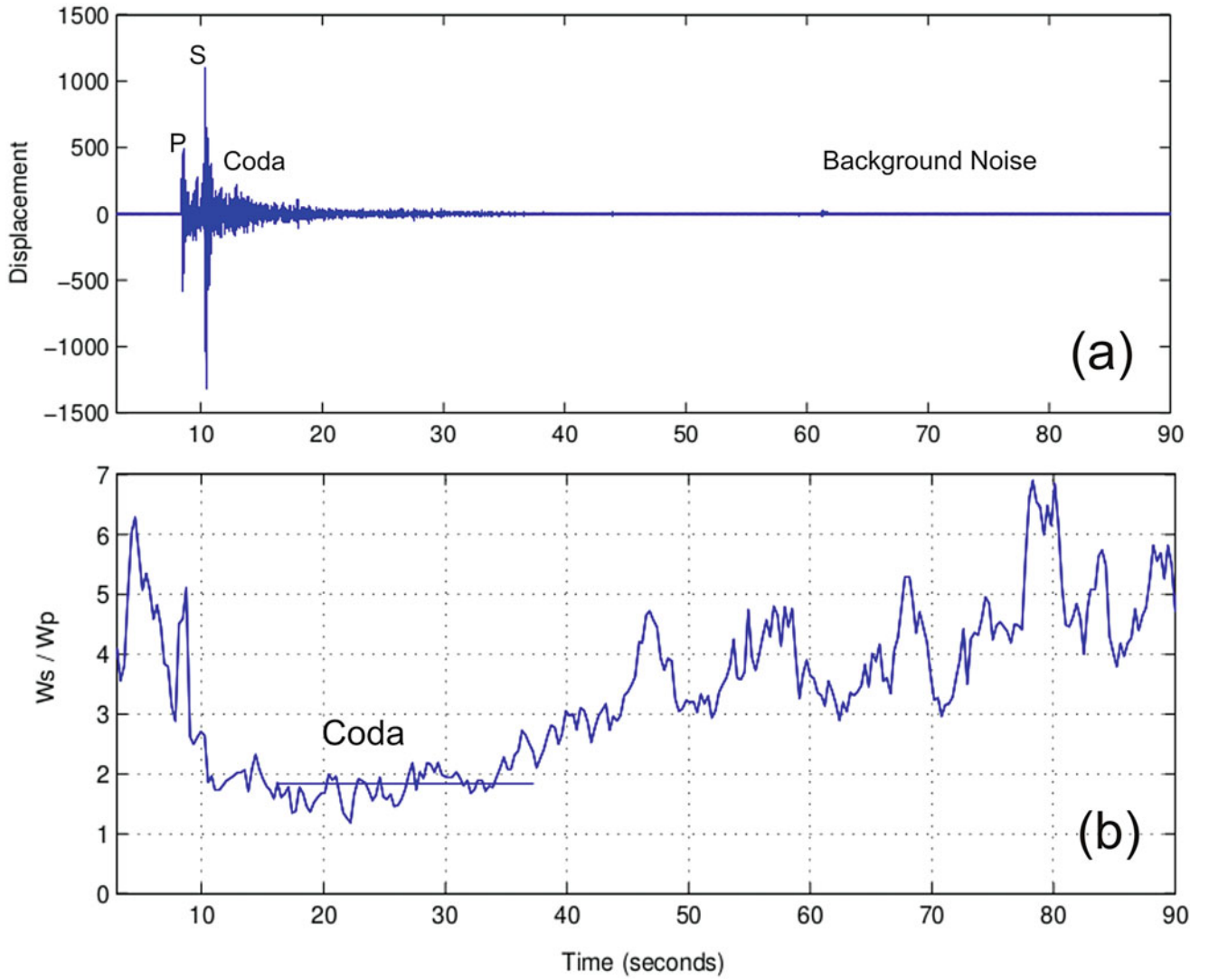
$$E^{SIS}(r, t|f) \approx \frac{W(f)\eta_{sc}e^{-\eta_{sc}Vt}}{4\pi r^2} \times \frac{r}{Vt} \ln \left[\frac{r + Vt}{r - Vt} \right] H\left(t - \frac{r}{V}\right), \quad (17)$$

For $t \gg \frac{2r}{V}$, energy density decays inversely with time squared as $\frac{W(f)\eta_{sc}}{2\pi V^2 t^2}$, in agreement with the SBS model of Aki and Chouet (1975) as:

$E^{SBS}(r, t|f) \approx \frac{W(f)\eta_{sc}}{2\pi V^2 t^2} e^{-Q_c^{-1}(f)2\pi f t}$. For longer lapse times, multiple scattering in Eq. (17) approaches to diffusion as:

$$E_{dif}(r, t|f) = \frac{e^{-\frac{r^2}{4Dt}}}{(4\pi Dt)^{3/2}} H(t), \quad (18)$$

where $D = \frac{V}{3\eta_{sc}}$ defines diffusivity. An approximate solution to Eq. (16) above was proposed by Paaschens (1997) as:



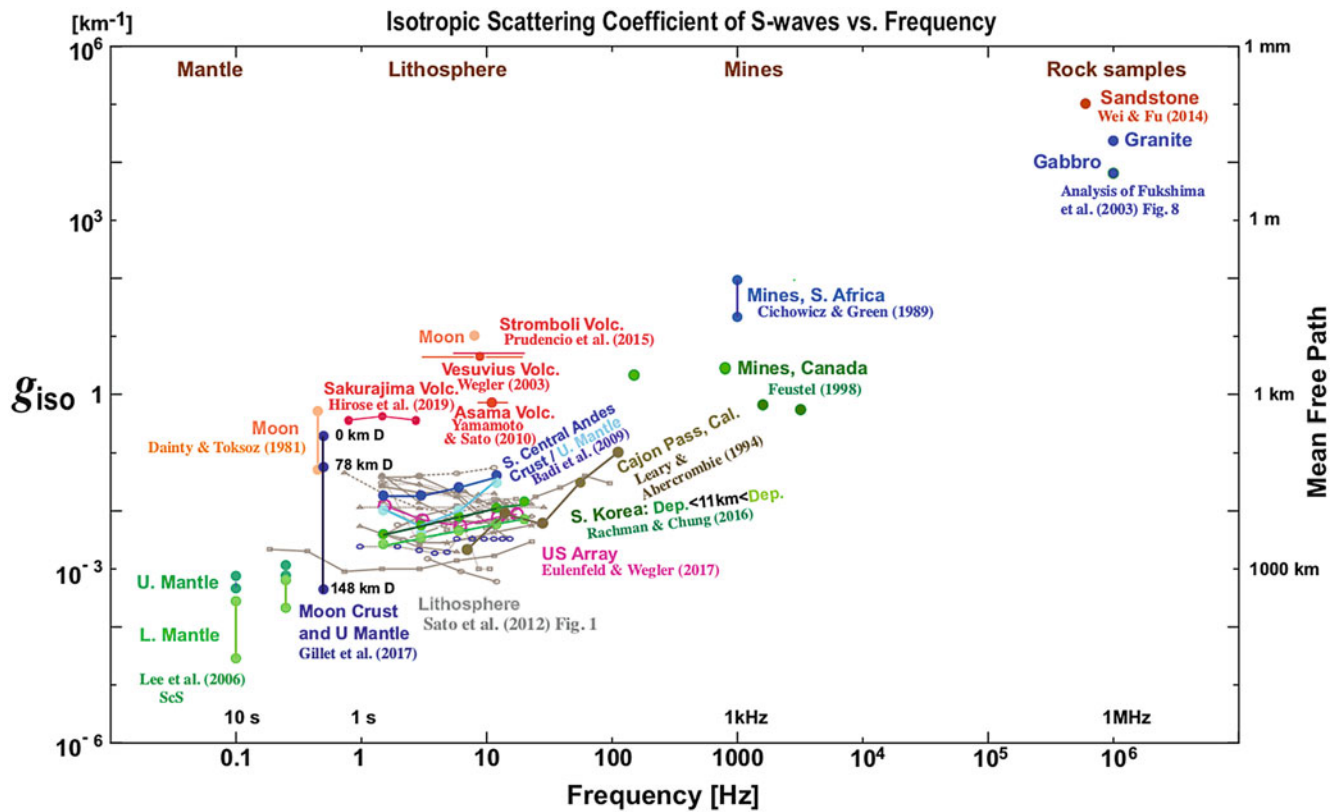
High-Frequency Seismology, Fig. 3 Equipartition of energy in coda. (a) Example of a vertical component displacement seismogram of a shallow local event (M_l 2.0). (b) Variation of S-to-P energy ratio with

time. Note the stabilization of energy ratio in coda with small fluctuations (after Margerin et al. 2009; copyright of 2020 Oxford University Press)

$$\begin{aligned}
 E(r, t|f) \approx & \frac{W(f)e^{-\eta_t V t}}{4\pi V r^2} \delta\left(t - \frac{r}{V}\right) \\
 & + \frac{W(f)\left(1 - \frac{r^2}{V^2 t^2}\right)^{1/8}}{\left(\frac{4\pi V t}{3\eta_t B_0}\right)^{3/2}} G\left(\eta_t B_0 V t \left(1 - \frac{r^2}{V^2 t^2}\right)^{3/4}\right) \\
 & \times e^{-\eta_t V t} H\left(t - \frac{r}{V}\right),
 \end{aligned}
 \tag{19}$$

where $G(x) \approx e^x \sqrt{1 + \frac{2.026}{x}}$, and $B_0 = \frac{\eta_{sc}}{\eta_t}$ is the seismic albedo.

The coda wave technique, such as the Multiple Lapse Time Window Analysis (MLTWA), uses any of the aforementioned multiple scattering models for separating Q_i and Q_{sc} . In this technique, integrated energy density is analyzed over consecutive time windows from the S-wave arrival time versus source-receiver distance. The observed energy density, after correction for geometrical spreading and the effects of source and site with coda-normalization, is compared with the synthesized one obtained with a multiple scattering model by solving the equation of RTT to separate Q_i and Q_{sc} (Hoshiba et al. 2001). The MLTWA has been used with uniform and layered velocity structure (Hoshiba et al. 2001; Padhy and Subhadra 2013; Sato et al. 2012 and references therein) to



High-Frequency Seismology, Fig. 4 Frequency-dependent isotropic scattering coefficient ($1/\text{mean free path}$) of S-waves for different tectonic regimes (after Sato 2019, Fig. 1; copyright of 2020 Oxford University Press)

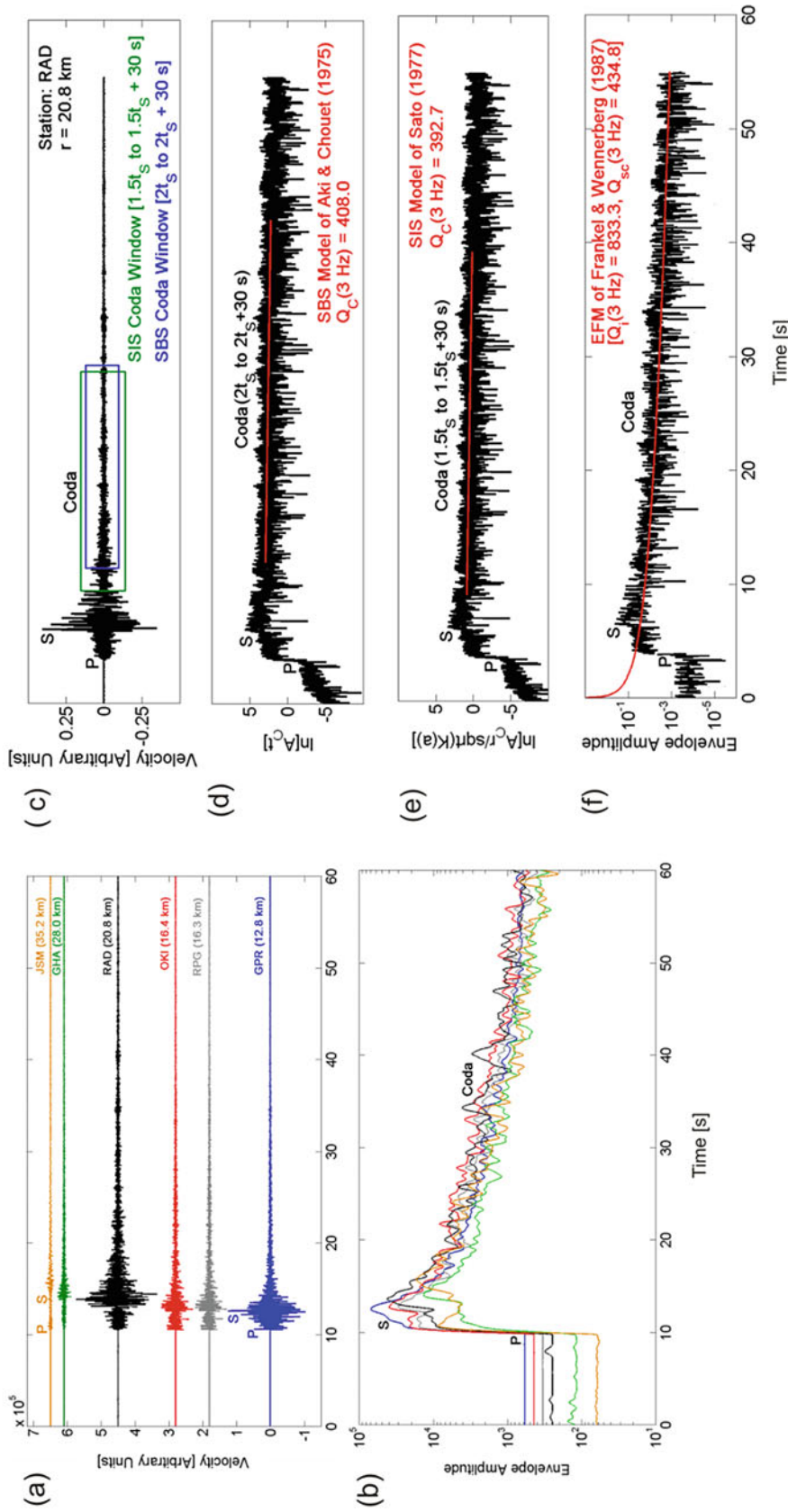
resolve mechanisms of attenuation; the results vary strongly with frequency depending on the model and type of data used. Because of the maximum length limitation of the chapter, I limit discussion of multiple scattering coda models here with a further reference to Sato et al. (2012) for details including more advanced realistic models of elastic waves in a layered earth structure.

Seismogram Analysis of the 1999 Chamoli Earthquake

As a case study, I analyzed few aftershocks of the 1999 Chamoli earthquake of India to examine the underlying assumptions of coda wave theory presented, and also to check the reliability of the Q_c , Q_i , and Q_{sc} factors obtained by comparing observations with synthetics for both the single and multiple scattering models of coda wave formation. Figure 5 summarizes the results. Figure 5a shows an original, vertical component of an aftershock (M_L 3.5) recorded by six stations in the source area of the 1999 Chamoli earthquake of north India. Figure 5b explains the fundamental property of coda waves: their smoothed envelope shapes are similar independent of distance, meaning that coda energy is uniformly distributed behind the

advancing wave front. The difference in envelope amplitude between different stations is related to the relative source/site effect. Figure 5c presents an example of band-passed (2–4 Hz) seismogram recorded by one of the stations RAD with different positions of coda window marked that are used by the two different single scattering models presented. Figure 5d and e presents the 3-Hz band-passed seismogram envelopes inverted for Q_c based on the SBS and SIS models, respectively, over the late (starting with twice the S-arrival time, $2t_s$) and early (starting with $1.5t_s$) coda windows marked with the best fit lines. It is worth mentioning that both the models yield almost similar Q_c values when similar coda window lengths are used. Figure 5f shows the application of the multiple scattering EFM to the 3-Hz observed envelope for resolving the intrinsic and scattering Q . The synthetic envelope (solid line), and thus, the corresponding Q_i and Q_{sc} values explain the observation reasonably good; the fit is not good near the direct S-arrival, however, which could be because of the nonisotropic source radiation that the current model does not consider. Thus, the data analyzed satisfy the assumptions of coda wave, and the Q estimates within limits can be considered reliable. We limited our application to local earthquake data, although some of the theories can be extended for analyses of regional earthquakes as well.

Methods of Coda Analysis - Aftershock records of the 1999 Chamoli earthquake
(1999-04-12 14:11:28.35 30.43N 79.2E 6.6km M 3.5)



High-Frequency Seismology, Fig. 5 Methods of analysis of coda waves for estimating Q_c . Q_c and Q_{sc} based on single and multiple scattering models. (a) Raw vertical component seismograms of an aftershock record of the 1999 event recorded by six seismic stations deployed in the source area. The parameters of the event are mentioned in the title. Phases P, S, and coda are marked. (b) Distances to each station are shown against each trace. Phases P, S, and coda are marked. (c) Smoothed envelopes of the seismograms shown in (a). Note the common shape of their envelopes; their amplitudes are adjusted for inferring the shape. (d) Band-passed (2–4 Hz) seismogram recorded at station RAD is shown with the varying positions of coda window used

for estimating Q_c based on the single scattering models. (d) Q_c estimation based on the SBS model. The best fit, based on the least-squares inversion, is shown with a solid line over a late coda starting from twice the S-arrival time ($2t_s$). (e) Same as (d), but with an early coda starting at $1.5t_s$ based on the SIS model. (f) The envelope is inverted for Q_i and Q_{sc} using a multiple scattering EFM of Frankel and Wennerberg (1987). The theoretical envelope is shown with a solid line. Note that the fit between the observation and theory is good over the coda (later) portion of the envelope, but is poor near the direct S-wave arrival. The corresponding Q values are shown against each frame

Summary and Outlook

I present basic concepts of seismic wave propagation and scattering in complex heterogeneous media. High-frequency seismic waves are essential for imaging of source or internal structure of earth at a sub-wavelength scale. They are very sensitive to small-scale heterogeneities of the medium, which scatter energy and broaden the waveforms producing coda waves. Starting with observational studies, I present here commonly used modeling of coda envelope based on energy transport in multiple scattering media with random heterogeneities. Several studies on theoretical and numerical modeling of coda energy are developed for both acoustic and elastic waves in random media (Sato et al. 2012). They can be used to invert envelopes for separating scattering and intrinsic absorption and for understanding the origin of coda wave types. Most of the phenomenological coda-based methods need to be validated with studies employing physics-based high-resolution numerical approaches, although they are computationally expensive for capturing high-frequency information of the waveforms to characterize small-scale heterogeneities. The assumptions underlying the coda wave theory and thus the reliability of the Q values obtained are examined by comparing observations with synthetics for few local earthquake recordings of the 1999 Chamoli event of India, and the results are encouraging within limits.

Coda-based methods have great implications in accurate ground-motion modeling by incorporating nonlinear multiple scattering effects of the medium. In imaging with scattered wave fields, the use of 1-D velocity structure to characterize surface wave propagation at short period (~ 1 s) does not fully exploit the fine details of shallow subsurface, which needs to be revisited by considering the lateral variation in physical properties with a more appropriate 3-D structure. The source rupture complexities, as inferred from inversion of high-frequency P-wave train from a dense array of stations, can be used for tsunami warning. Overall, all high-frequency features such as the appropriate 3-D velocity structure in complex media including undulated topography, sedimentary basins, volcanoes, fault zones, etc., source rupture details considering nonspherical source radiation, and characterization of sites accounting nonlinearity of soils are essential for improved shallow subsurface imaging and accurate hazard assessment. Multiple scattering-based imaging capabilities of coda waves have also found greater success in exploration such as in migration of multiples, amplitude corrections for transmission losses, etc., by treating locally scattered wave field, overcoming the limitations of linear single-scattering imaging that conventional migrations adopt.

Cross-References

- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Diffraction](#)
- [Seismic Waves, Scattering](#)
- [Seismogram Interpretation](#)

Acknowledgments Padhy sincerely acknowledges Professor Harsh K. Gupta (the Editor) for his kind invitation to contribute this chapter to Encyclopaedia of Solid Earth Geophysics. The editor and an anonymous reviewer are thanked for their insightful review comments. Ms. M. Kousalya is highly acknowledged for openly sharing the 1999 Chamoli earthquake data for research. Director, CSIR-NGRI is thanked for granting permission for the publication of this work (Ref. No. NGRI/Lib/2020/Pub-07).

Bibliography

- Aki K (1969) Analysis of the seismic coda of local earthquakes as scattered waves. *J Geophys Res* 74:615–631
- Aki K (1980) Scattering and attenuation of shear waves in the lithosphere. *J Geophys Res* 85:6496–6504
- Aki K, Chouet B (1975) Origin of coda waves: source, attenuation and scattering effects. *J Geophys Res* 80:3322–3342
- Aki K, Richards PG (1980) Quantitative Seismology – vol. 1: theory and methods. In: Freeman WH (ed), 2nd edn. University Science Books, 2009, New York 2002
- Ben-Menahem A, Singh SJ (1981) Seismic waves and sources. Springer, Berlin
- Emmerich H, Korn M (1987) Incorporation of attenuation into time domain computations of seismic wave fields. *Geophysics* 52:1252–1264
- Frankel A, Wennerberg L (1987) Energy-flux model of seismic coda: separation of scattering and intrinsic attenuation. *Bull Seismol Soc Am* 77(4):1223–1251
- Futterman WI (1962) Dispersive body waves. *J Geophys Res* 67:5279–5291
- Hoshiba M, Rietbrock A, Scherbaum F, Nakahara H, Haberland C (2001) Scattering attenuation and intrinsic absorption using uniform and depth dependent model – application to full seismogram envelope recorded in northern Chile. *J Seismol* 5(2):157–179
- Margerin L, Campillo M, Van Tiggelen BA, Hennino R (2009) Energy partition of seismic coda waves in layered media: theory and application to Pinyon flats observatory. *Geophys J Int* 177:571–585
- Paaschens JCJ (1997) Solution of the time dependent Boltzmann equation. *Phys Rev E* 56:1135–1145
- Padhy S, Subhadra N (2010) Attenuation of high-frequency seismic waves in Northeast India. *Geophys J Int* 181:453–467
- Padhy S, Subhadra N (2013) Separation of intrinsic and scattering seismic wave attenuation in Northeast India. *Geophys J Int* 195:1892–1903
- Padhy S, Wegler U, Korn M (2007) Seismogram envelope inversion using a multiple isotropic scattering model: application to after-shocks of the 2001 Bhuj earthquake. *Bull Seismol Soc Am* 97:222–233
- Paul A, Campillo M, Margerin L, Larose E, Derode A (2005) Empirical synthesis of time-asymmetrical green functions from the correlation of coda waves. *J Geophys Res* 110:08302
- Przybilla J, Korn M, Wegler U (2006) Radiative transfer of elastic waves versus finite difference simulations in two-dimensional random

- media. *J Geophys Res* 111:B04305. <https://doi.org/10.1029/2005JB003952>
- Sato H (1977) Energy propagation including scattering effects single isotropic scattering approximation. *J Phys Earth* 25:27–41
- Sato H (2019) Isotropic scattering coefficient of the solid earth. *Geophys J Int* 218:2079–2088
- Sato H, Fehler MC, Maeda T (2012) *Seismic wave propagation and scattering in the heterogeneous earth*. Springer Science and Business Media, Springer-Verlag Berlin Heidelberg
- Wu RS (1985) Multiple scattering and energy transfer of seismic waves – separation of scattering effect from intrinsic attenuation – I. theoretical modelling. *Geophys J R Astron Soc* 82:57–80
- Wu Z, Lay T, Ye L (2019) Shallow megathrust slip during large earthquakes that have high P coda levels. *J Geophys Res*. <https://doi.org/10.1029/2019JB018709>
- Zeng Y, Su F, Aki K (1991) Scattering wave energy propagation in a random isotropic scattering medium 1. Theory. *J Geophys Res* 96:607–619

Impact Craters on Earth

Richard A. F. Grieve^{1,2} and Gordon R. Osinski^{1,2}

¹Department of Earth Sciences, University of Western Ontario, London, ON, Canada

²Institute for Earth and Space Exploration, University of Western Ontario, London, ON, Canada

Definition

Impact crater	An approximately circular or polygonal depression formed by the hypervelocity impact of an interplanetary body (asteroid or comet) on a planetary surface.
Shock metamorphism	Irreversible changes to rocks and minerals, resulting from the passage of a shock wave produced by a hypervelocity impact.

The Terrestrial Record

Introduction

The results of planetary exploration have demonstrated that impact cratering is a ubiquitous process that has occurred throughout solar system history. On bodies that have retained portions of their early crust, such as the Moon, older surfaces are dominated by impact craters, attesting to the importance of impact cratering in early surface and crustal evolution. The Earth is the most geologically active of the terrestrial planets and is constantly renewing its surface. As impact craters are surface features, the net result of this high level of endogenous geological activity is that most of the original population of Earth's impact craters have been destroyed. What is left is a small preservation sample. The terrestrial record, however, is invaluable in the understanding of impact processes for it is the only source of ground-truth data on the three-dimensional structural and lithological character of natural impact craters.

Most recently, for example, the results of IODP-ICDP drilling at the Chicxulub crater has contributed to the understanding of the formation of so-called peak-rings in large impact basins (Morgan et al. 2016).

Since the Apollo missions in the 1960s, the ground-truth data supplied by terrestrial impact craters has served as analogues in understanding cratering processes on the other terrestrial planets; however, at the time, impact was generally not regarded as important to Earth's evolution. This changed in 1980, with evidence for a major impact as the cause of the mass extinction event at the Cretaceous-Paleogene (K-Pg) boundary ~65.5 million years ago. The actual crater, the ~180 km diameter Chicxulub crater, was identified in 1991, underneath ~1 km of sediments in the Yucatan peninsula, Mexico. The terrestrial impact record currently stands at approximately 200 odd or crater fields, with the largest being the ~300 km diameter Vredefort structure, South Africa. Evidence of distal ejecta from over two dozen impact events, some of which can be linked to their source craters, is also recorded in the stratigraphic record. They include four strewn fields of impact glasses known as tektites, which can occur over considerable areas. For example, the Australasian strewn field, which has currently no known source crater, is in excess of $50 \times 10^6 \text{ km}^2$.

Distribution

A listing of known terrestrial impact craters, along with their properties, can be found at <http://www.impactearth.com>. The spatial distribution of known terrestrial impact structures is not random, with concentrations on the geologically stable cratons of the Australian, European, and North American continents. They are areas with relative tectonic stability and low rates of erosion. That is, they are the best available surfaces for the preservation of terrestrial impact craters. These are also regions where there have been active programs to study impact craters. Geological processes, such as differential erosion, can result in terrestrial impact craters evolving from an originally negative to a positive landform, that is, by

definition, no longer a “crater.” Thus, it is more appropriate to use the term terrestrial impact “structure,” which is not limited to a particular landform. Approximately, one-third of the known terrestrial impact structures are buried by post-impact sediments. Detected initially as geophysical anomalies, their identity was confirmed by drilling. There are a number that are now underwater and several that were formed underwater, on relatively shallow continental shelves. At present, no impact structures are known on the true ocean floors, reflecting the relatively young age and poorer geologic knowledge of this realm.

At large (>20 km) diameters, the cumulative size-frequency distribution of terrestrial impact structures is similar to that on the other terrestrial planets. This likely represents some form of steady-state condition between the formation of impact structures and their removal by terrestrial geologic processes. At smaller diameters, however, there is a deficit in the number of known terrestrial impact structures. This is due to the crushing of small impacting bodies during

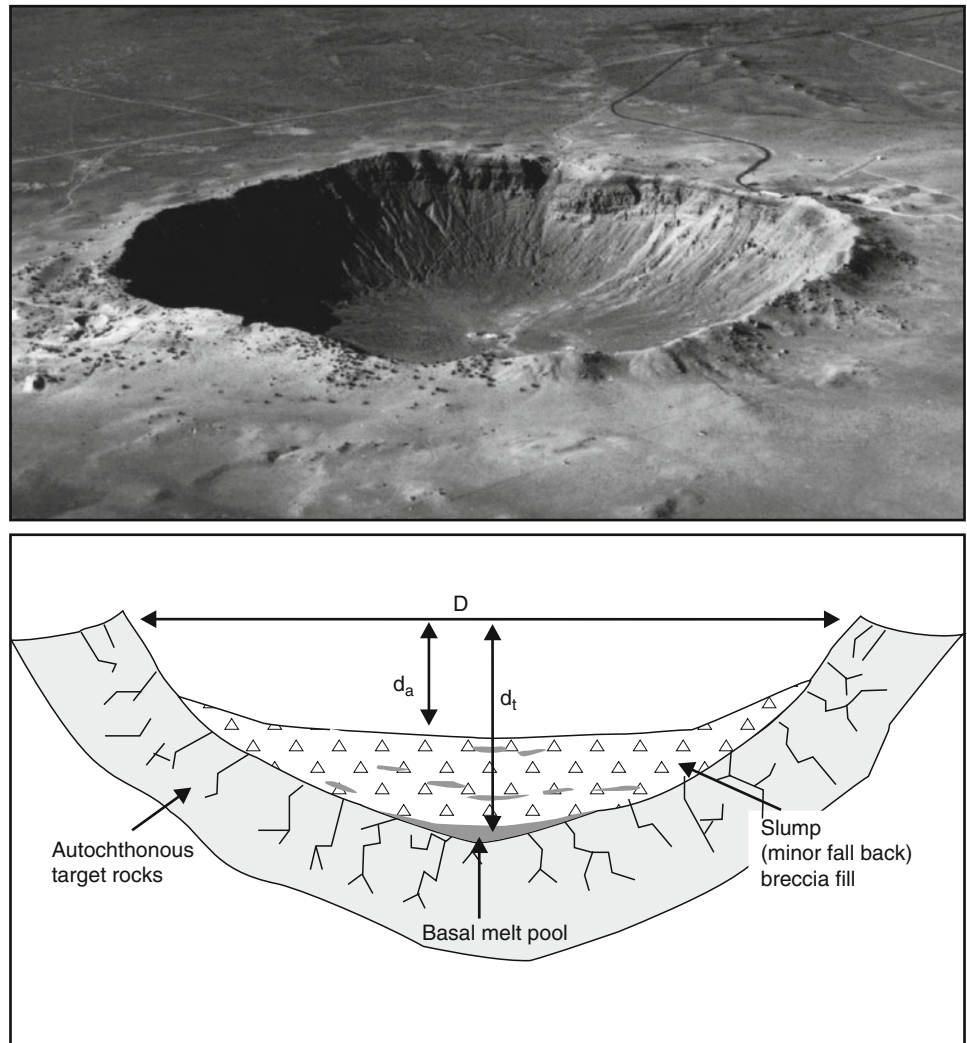
atmospheric passage, the relative ease with which smaller structures can be removed by erosion or buried, and the intrinsic difficulties in recognizing smaller structures. There is also a bias in the ages of known terrestrial impact structures, with most being <200 Ma old, which also reflects the problems of preservation in the terrestrial environment.

Morphology

Unless buried almost immediately, the forms of terrestrial impact structures are generally modified from their original. There is, however, the same overall progression in form, with increasing diameter, as on the other terrestrial planets. Smaller structures have the form of a bowl-shaped depression, with an upraised rim, and a depth of $\sim 1/6$ the diameter, and are termed simple structures (Fig. 1). The exposed rim, walls, and floor define what is called the apparent crater. Beneath the apparent floor is an allochthonous breccia lens, which is roughly parabolic in cross section and floored by parautochthonous fractured target rocks. This breccia lens—

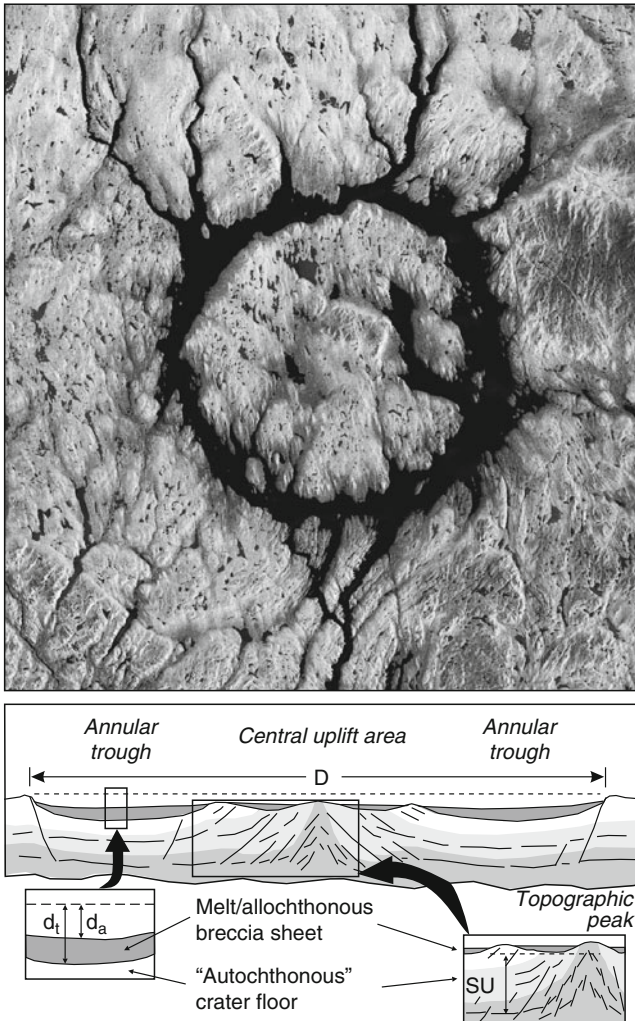
Impact Craters on Earth,

Fig. 1 Top: Oblique aerial photograph of the classic example of a terrestrial simple impact structure: The 1.2 km diameter Barringer or Meteor Crater, Arizona, USA. Bottom: Schematic cross section of a terrestrial simple impact structure, showing various lithological, structural, and morphologic attributes. Note allochthonous breccia lens, with impact-melted lithologies, partially filling the structure. D rim diameter, d_a apparent depth



parautochthonous target rock boundary defines what is called the true crater, with a depth of $\sim 1/3$ the rim diameter (Fig. 1). For simple structures, the rim diameter is ~ 10 times that of the impacting body.

As the diameter increases, there is a change in form, with increasing wall and rim collapse, as simple structures evolve into complex impact structures, with rims consisting of structurally faulted terraces (Fig. 2). Interior to the rim is a down-faulted annulus or trough, which is partly filled by allochthonous breccias and/or rocks melted by the impact. At



Impact Craters on Earth, Fig. 2 Top: RADARSAT image of the Manicouagan, Canada, complex impact structure. The annular lake is 65 km in diameter and is, in part, man-made, serving as reservoir for hydroelectric power generation. The original diameter of this partially eroded structure was ~100 km and the central peak (Mont de Babel) rises to 600 m above the level of the lake. Bottom: Schematic cross section of a terrestrial complex impact structure, indicating various lithological, structural, and lithological attributes. Example shown has a central peak and an annular ring, which may have been the original form of Manicouagan. D rim diameter, d_a apparent depth, d_t true depth, SU structural uplift, which is ~0.1 D

the center is a structurally complex central peak of parautochthonous target rocks, which is the topographic manifestation of a more extensive volume of uplifted target rocks that lie beneath the floor in the center (Fig. 2). The transition diameter from simple to complex structures is an inverse function of planetary gravity and target rock strength, with the transition occurring at a diameter of ~ 4 km in crystalline and ~ 2 km in sedimentary target rocks, in the terrestrial environment. A number of relatively young (i.e., only slightly eroded) terrestrial complex structures – Houghton, Canada; Ries, Germany; Zhamanshin, Kazakhstan – lack an emergent central peak. These structures are in mixed targets of sediments overlying crystalline basement, with the lack of a peak most likely due to target strength effects. For complex structures, the rim diameter is ~ 20 times that of the impacting body.

With still increasing diameter, a fragmentary inner ring appears, marking the transition from a central peak structure to peak-ring basin. The larger impact structures in the terrestrial environment, however, tend to be older, eroded structures that have been modified, with their current elements being structural features rather than topographic features. Chicxulub is the best-preserved large terrestrial impact structure, due to burial. As such, however, the definition of its morphological elements depends on the interpretation of geophysical data. It has an interior topographic peak-ring, a terraced rim area, and exterior ring faults and, thus, appears to correspond to the definition of a multi-ring basin, as on the Moon. Details of the physics of cratering mechanics and impact, in general, can be found in Melosh (1989) and Osinski and Pierazzo (2012).

Geology of Impact Structures

Only small, young, simple impact structures, where the atmosphere has partially slowed the impacting body, have physical evidence of the impacting body in the form of spalled fragments of iron or stony-iron meteorites, which are relatively quickly destroyed by weathering. (Equivalent-sized, weaker stony meteorites are generally crushed during atmospheric passage.) Larger impacting bodies (>100 m in diameter) impact the Earth with largely undiminished velocity. The average impact velocity of an asteroidal body is $20\text{--}25\text{ km s}^{-1}$ and cometary impacts have even higher velocities. Upon impact, most of the impacting body's considerable kinetic energy is partitioned into the target by means of a hemispherical propagating shock wave, which sets the target rocks in motion (kinetic energy) and compresses them (internal energy). A cratering flow-field is established in the target, leading to the excavation of target materials. During compression, a considerable amount of pressure-volume work is done. Not all this mechanical work is recovered on

decompression to ambient pressure and the excess is manifested as waste heat. This leads to melting and even vaporization of some of the target rocks and the destruction of the impacting body.

The transient compression, decompression, and heating of the target rocks lead to shock metamorphic effects, which record pressures, temperatures, and strain rates well beyond those produced in terrestrial regional or contact metamorphism. The exact physical conditions recorded depend of the peak pressure on impact, as determined by the impacting body's impact velocity and density, and distance from the point of impact. Given the highly transient nature of shock metamorphic processes, disequilibrium and metastable equilibrium are common. The nomenclature of shocked rocks or impactites (various melt rocks, breccias, and shocked target rocks) and the pressures required to generate particular shock metamorphic effects are summarized in Stöffler and Grieve (2007). Examples of impactites and shock metamorphic effects are well illustrated in French (1998).

The highest shock metamorphic effect preserved at impact structures is impact-melted materials (Osinski et al. 2018). They result from shock pressures >60 GPa and take the form of glass clasts in breccias, melt dikes in the floor of the structure, and lenses and coherent sheets of impact melt rock within the crater interior (Fig. 3). When crystallized, impact melt rocks have an igneous-looking matrix and are often charged with mineral and lithic debris. Compositionally, these melted lithologies can be modeled as a mixture of target lithologies, in geologically reasonable proportions. Isotopic systems, such as Sr, Nd, and Pb, mirror those of the initial ratios of the target rocks, while others (e.g., Ar) can be reset and record the time of impact. In some cases, siderophile-element enrichments above target rock levels occur, reflecting an admixture of material from the impacting body. In some cases, their relative abundances have identified the type of impacting body, for example, C-1 chondrite. In a few cases, Cr and Os isotopes have also been used to identify the impacting body. At lower shock pressures (~60–30 GPa), selective mineral melting occurs. In addition, the shock wave can destroy the internal crystallographic order of feldspars and quartz and convert them to solid-state glasses, which still have the original crystal shapes. These are “diaplectic” glasses (Fig. 3), with the required pressures being 30–45 GPa for feldspar and 35–50 GPa for quartz (diaplectic plagioclase glass is also known as maskelynite). The extremely rapid compression and then decompression also produces metastable polymorphs, for example, coesite and stishovite from quartz and diamond and lonsdaleite from graphite.

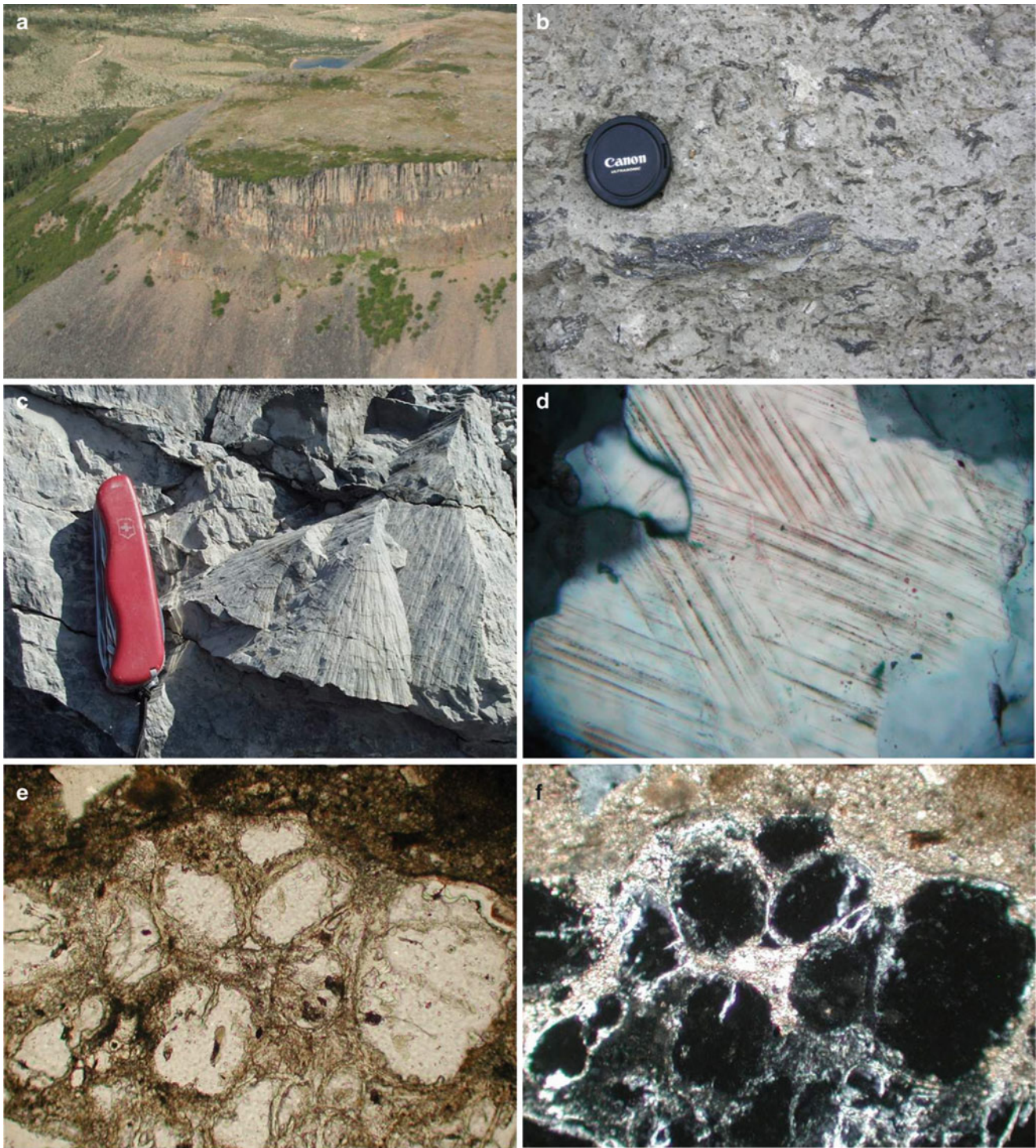
The most-documented shock metamorphic feature is the occurrence of so-called planar deformation features, particularly in quartz (Fig. 3), although they do occur in other minerals, for example, feldspar, zircon. Their utility is a function of the common occurrence and stability of quartz

in terrestrial rocks and the relative ease with which they can be observed and measured by optical microscopy. When fresh, planar deformation features are parallel planes of glass, with specific crystallographic orientations as a function of shock pressures of ~10–35 GPa. The only megascopic shock features are shatter cones, which are distinctive, striated, and horse-tailed conical fractures ranging in size from millimeters to meters (Fig. 3). They are best developed in fine-grained, structurally isotropic rocks, such as carbonates and quartzites, but more poorly developed variants can occur in coarser-grained rocks. They are most common in rocks shocked to relatively low levels, ~2–6 GPa, but can occur in rocks recording pressures as high as ~25 GPa.

Geophysics of Impact Structures

For buried impact structures, a geophysical anomaly is the first indicator of their existence, although confirmation of their genesis requires drilling and the documentation of shock metamorphic effects. The most common geophysical anomaly is a localized low in the regional gravity field (Fig. 4), due to lowering of rock density from brecciation and fracturing. In general, modeling of the gravity anomalies is consistent with the morphometric parameters of terrestrial impact structures, with the anomaly extending to the rim of the structure. The anomaly increases to a maximum of ~200–300 g.u. at rim diameters of 20–30 km (Pilkington and Grieve 1992), with the lack of further increases in magnitude at larger diameters due to lithostatic pressure closing any deep-seated impact-related fracturing. The gravity signature can be complicated by erosion and variations in target rock densities. Larger complex impact structures tend to have a central, relative gravity high, which can extend out to approximately half the diameter of the structure. This central, relative gravity high is from the uplift of originally deeper (more dense) lithologies in the structural uplift.

Magnetic anomalies are more varied. Most commonly, they are a magnetic low, with the disruption of any regional trends in the magnetic field. This is due to an overall lowering of magnetic susceptibility and the randomizing of pre-impact lithological trends in the target rocks. Larger structures commonly have high-amplitude, short wave-length anomalies in the center, due to changes to remnant magnetization in the target rocks from the effects of heat, shock, and/or post-impact hydrothermal alteration. Seismic velocities are reduced at impact structures, due to fracturing, and reflection seismic images are extremely useful in characterizing buried structures in sedimentary targets. Subsurface brecciation, fracturing, and incoherency of seismic reflectors increase toward the center of complex structures due to structural uplift, resulting in a central volume that is isotropic in terms

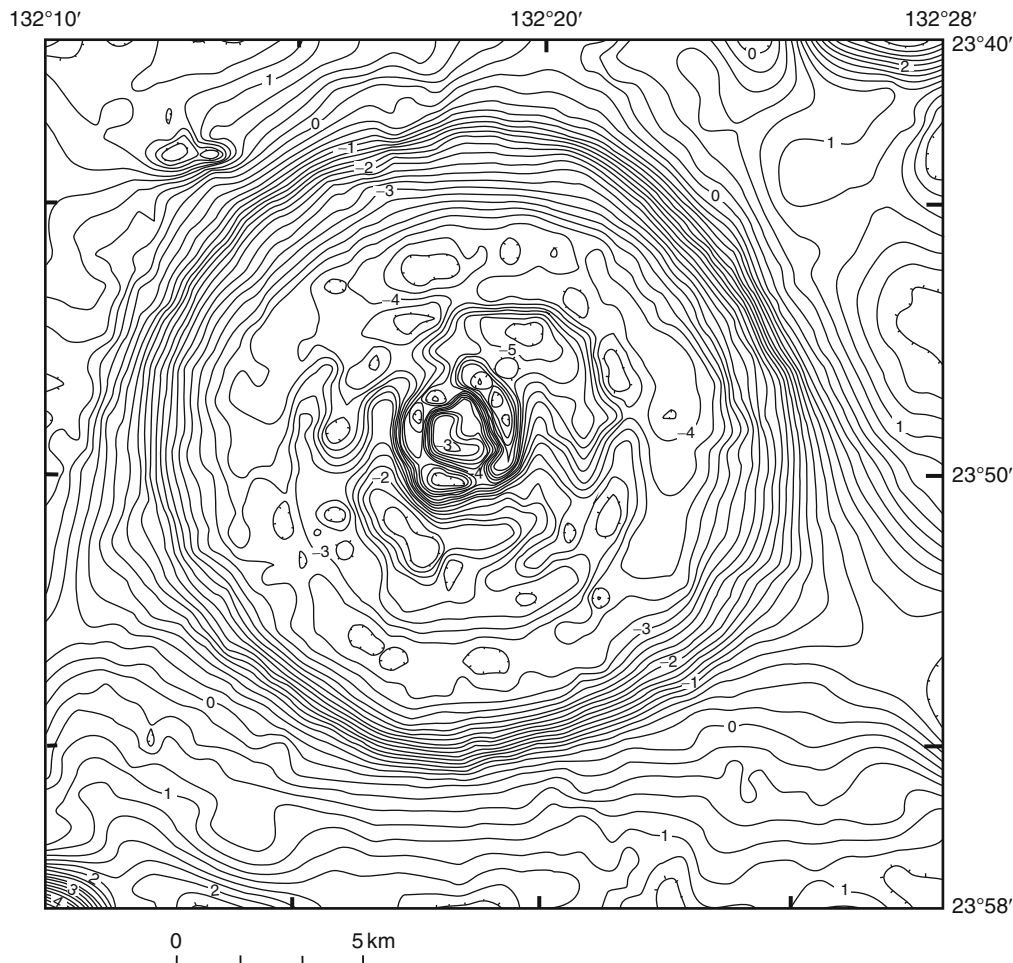


Impact Craters on Earth, Fig. 3 Montage of impactites. (a) The impact melt sheet of the 28-km diameter Mistastin impact structure, Canada. Note the columnar jointing in the melt sheet. (b) Impact melt-bearing breccias (glassy melt rock occurs as black clasts) at the Mistastin structure. Six-centimeter diameter camera lens for scale. (c) Shatter

cones from the Haughton impact structure, Canada. Eleven-centimeter long penknife for scale. (d) Photomicrograph (*plane light*) of Planar Deformation Features (PDFs) in quartz. Image is 0.35 mm across. (e) and (f) Photomicrograph of diaplectic quartz glass seen in plane (e) and cross-polarized light (f), respectively. Images are 5 mm across

of a reflection seismic signal. Electrical methods have been used less but, in general, there is a reduction in resistivity, which is related to fracturing. Resistivity contrasts can also

delineate lithological contacts, for example, between allochthonous breccia deposits and the fractured target rock and/or post-impact sediments.



Impact Craters on Earth, Fig. 4 Residual Bouguer gravity anomaly at the 22-km diameter Gosses Bluff impact structure, Australia, which is in the form of a symmetric gravity low, with a central, short-wave anomaly due to structural uplift. The target was a flat-lying sequence of

sedimentary rocks and the simplicity in the form of the anomaly results from the lack of major asymmetric lithological/density contrasts. Contour interval is 1 g.u.

Impacts and Earth History

The formation of the terrestrial planets by the accretion of small bodies through collisions and subsequent growth to proto-planets includes potential collisions between these proto-planets. One such collision between the proto-Earth and a Mars-sized body is currently the best working hypothesis for the origin of the Moon. Such an impact would result in an Earth-orbiting accretionary disk of material from the impacting body and proto-Earth, which was the source of volatile- and siderophile-depleted material for building the Moon. The consequences for the proto-Earth include massive remelting and the loss of any original atmosphere. Without a Moon and lunar tides, there would be little or no littoral zone, which is the most important area in the terrestrial ecosystem, to Earth's oceans. With no littoral zone, the development and

evolutionary path of Earth's biosphere would have been very different. Following planetary formation, the subsequent high rate of bombardment is attested to by the heavily cratered lunar highland crust and the formation of the very large (up to 2,500 km in diameter) multi-ring impact basins. The Earth, however, has a larger gravitational cross-section and will capture more incoming interplanetary bodies than the Moon. Thus, it has been subject to a greater number of impacts than the Moon throughout geological time, including the basin-forming impacts in the period of heavy bombardment. There have been reasoned speculation as to the potential effects of such impacts (e.g., Grieve et al. 2006), but no rocks are preserved from this period of the early Earth. Such impacts would retard any development of life on Earth but once they had subsided smaller impact events may have been beneficial for microbial life, with the generation of fertile hydrothermal

systems, impact crater lakes, and new habitats within shocked rocks and minerals (Cockell et al. 2012).

The temporal association of a large impact event (Chicxulub), world-wide ejecta, and a mass extinction at the K-Pg boundary is well established (Schulte et al. 2010). As other large impact structures have not been linked to mass extinctions in the terrestrial biosphere, it may be that there was a unique characteristic to the Chicxulub event. Sulfate-bearing rocks occur at Chicxulub, raising the possibility of massive amounts of sulfur aerosols in the atmosphere after the impact. They would reduce light levels below those required for photosynthesis for the best part of a year, with the collapse of the food chain. Chicxulub-scale impacts occur every 100 ± 50 Ma. Smaller impacts occur on shorter timescales and loading of the atmosphere with dust from craters in the 20-km size-range would be global in scale. This would not produce a mass extinction but could severely disrupt or even destroy the relatively fragile infrastructure that supports human civilization.

Past-impacts have also been beneficial for the human condition. Approximately 25% of known terrestrial impact structures have some form of natural resources, ranging from unexploited local deposits to world-class mining camps (Grieve 2012). For example, the ~ 1.5 billion tonnes of ores at the Sudbury impact structure, Canada, have so far produced over \$ 300 billion worth of metals. Approximately 50% of impact structures in hydrocarbon-bearing sedimentary basins are producers. The world-class Campeche Bank oilfield in the Gulf of Mexico produces most of its hydrocarbon products (~ 12 billion barrels of oil and > 3 trillion cubic feet of gas, since discovery) from Chicxulub-related breccias and has reserves exceeding the entire known conventional reserves of the USA.

Concluding Remarks

Although the first known recorded observations of impact craters were made by Galileo Galilei in 1609, it was not until the 1960s that the importance of impact cratering as a geological process began to be recognized. This came with the recognition of shock metamorphism and its effects on rocks and minerals and with impetus provided by the Apollo landings on the Moon. It is now widely acknowledged that impact cratering is one of the most important surface geological processes in the solar system. It has also become apparent that impact events have also played an important role throughout Earth's history, shaping the geological landscape, producing economic benefits, and affecting the evolution of life. The study of impact craters and related phenomena is relatively young when compared to other fields of scientific

study and, as such, there is still considerable potential for new and exciting discoveries.

Cross-References

- [Deep Scientific Drilling](#)
- [Gravity Anomalies, Interpretation](#)
- [Magnetic Anomalies: Interpretation](#)
- [Seismic Imaging, Overview](#)
- [Seismic Properties of Rocks](#)

Bibliography

- Cockell CS, Osinski GR (2012) Geomicrobiology of impact craters. In: Osinski GR, Pierazzo E (eds) *Impact cratering: Processes and products*. Wiley Blackwell, Oxford, pp 157–176
- French BM (1998) *Traces of Catastrophe: A Handbook of Shock-Metamorphic Effects in Terrestrial Meteorite Impact Structures*. LPI Contribution no. 954. Lunar and Planetary Institute, Houston
- Grieve RAF (2012) Economic deposits at terrestrial impact structures. In: Osinski GR, Pierazzo E (eds) *Impact Cratering: Processes and Products*. Wiley-Blackwell, New York, pp 177–193
- Grieve RAF, Cintala MJ, Theriault AM (2006) Large-scale impact and evolution of the Earth's crust: the early years. In: Reimold WU, Gibson RL (eds) *Processes on the early Earth*. Geological Society of America special paper 405, Boulder, pp 23–31
- Melosh HJ (1989) *Impact cratering: A geologic process*. Oxford University Press, Oxford
- Morgan JV, Gulick SPS, Bralower T, Chenot T, Christeson G, Claeys CC, Collins GS, Coolen MJL, Ferrière L, Gebhardt C, Goto K, Jones H, Kring DA, Le Ber E, Lofi J, Long X, Lowery C, Mellett C, Ocampo-Torres R, Osinski GR, Perez-Cruz L, Pickersgill A, Poelchau M, Rae A, Rasmussen C, Rebolledo-Vieyra M, Riller U, Sato H, Schmitt DR, Smit J, Tikoo S, Tomioka N, Urrutia-Fucugauchi J, Whalen M, Wittmann A, Yamaguchi KE, Zylberman W (2016) The formation of peak-rings in large impact craters. *Science* 354:878–882
- Osinski GR, Pierazzo E (2012) *Impact cratering: processes and products*. Wiley-Blackwell, New York
- Osinski GR, Grieve RAF, Bleacher JE, Neish CD, Pilles EA, Tornabbee LL (2018) Igneous rocks formed by hypervelocity impact. *J Volcanol Geotherm Res* 353:25–54
- Pilkington M, Grieve RAF (1992) The geophysical signature of terrestrial impact craters. *Rev Geophys* 30:161–181
- Schulte P, Alegret L, Arenillas I, Arz JA, Barton PJ, Bown PR, Bralower TJ, Christeson GL, Claeys P, Cockell CS, Collins GS, Deutsch A, Goldin TJ, Goto K, Grajales-Nishimura JM, Grieve RAF, Gulick SPS, Johnson KR, Kiessling W, Koeberl C, Kring DA, MacLeod KG, Matsui T, Melosh J, Montanari A, Morgan JV, Neal CR, Nichols DJ, Norris RD, Pierazzo E, Ravizza G, Rebolledo-Vieyra M, Reimold WU, Robin E, Salge T, Speijer RP, Sweet AR, Urrutia-Fucugauchi J, Vajda V, Whalen MT, Willumsen PS (2010) The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. *Science* 327:1214–1218
- Stöffler D, Grieve RAF (2007) A systematic nomenclature for metamorphic rocks: impactites. In: Fettes D, Desmons J (eds) *Metamorphic rocks: a classification and glossary of terms*. Cambridge University Press, Cambridge, pp 82–92

Instrumentation, Electrical Resistivity

M. H. Loke¹, O. Kuras², J. E. Chambers², D. F. Rucker³ and P. B. Wilkinson²

¹Geotomo Software Sdn. Bhd., Penang, Gelugor, Malaysia

²Geophysical Tomography Team, British Geological Survey, Keyworth, Nottingham, UK

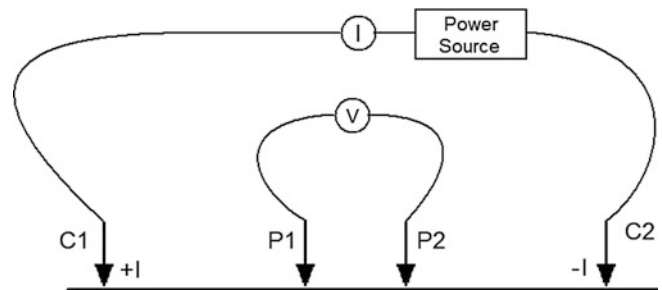
³HydroGEOPHYSICS, Inc., Tucson, AZ, USA

Definition

Electrical survey	Mapping subsurface resistivity by injecting an electric current into the ground.
Resistivity meter	Instrument used to carry out resistivity surveys that usually has a current transmitter and voltage measuring circuitry.
Electrode	A conductor planted into the ground through which current is passed, or which is used to measure the voltage caused by the current.
Apparent resistivity	The apparent resistivity is the resistivity of an equivalent homogeneous earth model that will give the same potential value as the true earth model for the same current and electrodes arrangement.
Multicore cable	A cable with a number of independent wires.

Introduction

The resistivity survey method is more than 100 years old and is one of the most commonly used geophysical exploration methods. It has been used to image targets from the millimeter scale to structures with dimensions of kilometers. It is widely used in environmental, engineering, and mineral exploration surveys (Loke et al. 2013). There have been many recent advances in instrumentation and data interpretation resulting in more efficient surveys and accurate earth models. In its most basic form, the resistivity meter injects a current into the ground through two metal stakes (electrodes) and measures the resulting voltage difference on the ground surface between two other points (Fig. 1). The current (I) and voltage (V) values are normally combined into a single quantity, the apparent resistivity, which is given by the following relationship.



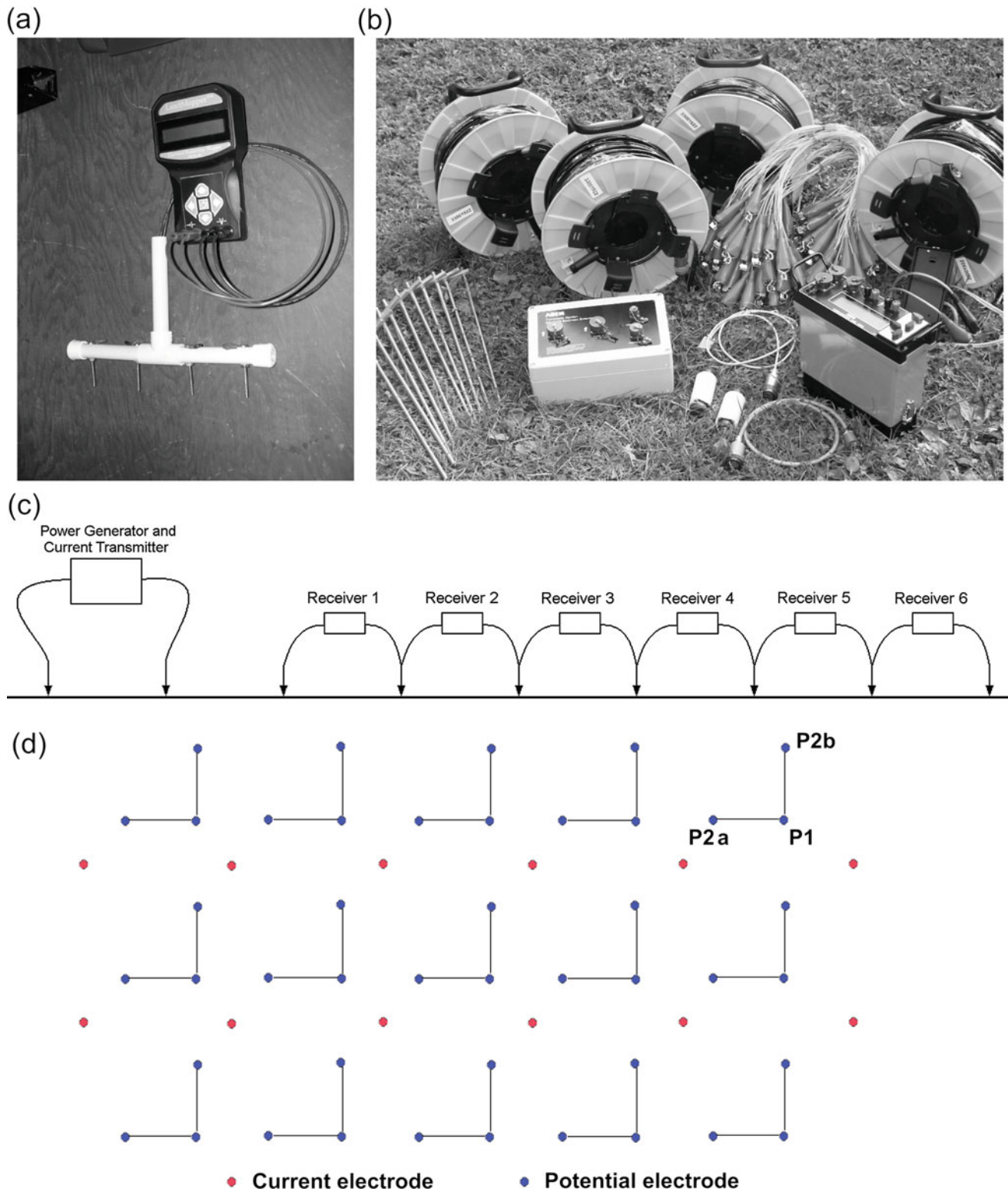
Instrumentation, Electrical Resistivity, Fig. 1 Basic setup for an electrical resistivity survey. The current is injected into the ground through the C1 and C2 electrodes, and the resulting voltage difference is measured by the P1 and P2 electrodes

$$\rho_a = k V/I \quad (1)$$

The apparent resistivity is the resistivity of an equivalent homogeneous earth model that will give the same potential value as the true earth model for the electrodes arrangement. The geometric factor k depends on the arrangement of the electrodes (*electrical resistivity surveys and data interpretation*). The flow of the current through the ground depends on the resistivity of the material (rocks) through which it passes and is reflected in the measured voltage values. Subsurface resistivity distribution can be mapped by making measurements with the current and voltage electrodes at different positions.

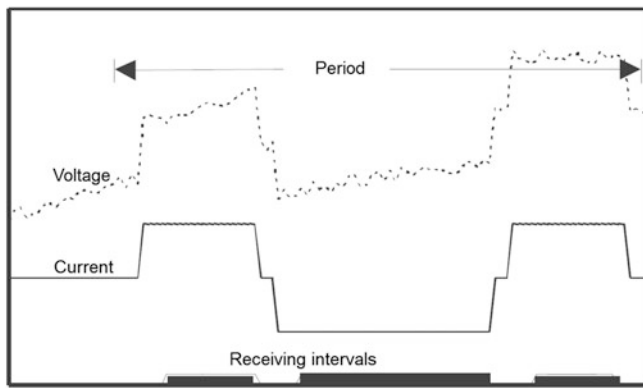
Basic Resistivity Instruments

The basic parts of a resistivity measurement system include a source of electrical current, a voltage measuring system, and the cables to connect these components to the electrodes. These components have undergone major modifications over the years to improve the efficiency of the survey procedure, the quality of the data, and to increase the depth of investigation. The power source for the current is typically a battery for shallow surveys. Some small portable resistivity meter systems (Fig. 2a) for very shallow surveys of up to a few meters depth weigh less than a kilogram (Pozdnyakov et al. 2009). A typical system (with associated cables and electrodes) for environmental and engineering surveys that uses an internal battery is shown in Fig. 2b. Such systems usually weigh between 10 and 50 kg. The current source and voltage measuring circuitry are integrated into a single unit (the resistivity meter). Such battery-based systems can provide currents of up to about 2 A and have been used for survey depths of up to about 200 m. For deeper surveys where



Instrumentation, Electrical Resistivity, Fig. 2 Different types of resistivity instruments. (a) A light-weight portable system (courtesy of Landviser LLC). The distance between the electrodes in the above picture is 5 cm. (b) A typical system used for shallow environmental

and engineering surveys (courtesy of Torleif Dahlin and Abem Instruments AB), (c) a schematic diagram of a system for deep surveys, and (d) schematic diagram of a vector array survey layout



Instrumentation, Electrical Resistivity, Fig. 3 Example of input current and resulting voltage signal. (Courtesy of Torleif Dahlin and Abem Instruments AB)

currents of up to 10 A are used, a petrol/diesel engine powered electric generator is usually used. Such systems can weigh several hundred kilograms and are commonly used for induced polarization (IP) surveys in mineral exploration (Truffert et al. 2019). The current transmitter is separate from the receivers that measure the potential signal (Fig. 2c).

In the interpretation of the survey data, it is assumed that a direct current (DC) is used. However, in practice, a low frequency square-wave alternating current (AC) is normally used (Fig. 3). The frequency is sufficiently low (typically less than 100 Hz) such that EM effects are negligible. Figure 3 shows an example of the shape of the input current used and the resulting voltage signal. The amplitude of the voltage signal is typically in the millivolt range, but smaller voltages can be resolved if the signal-to-noise ratio is sufficiently high or signal stacking techniques are employed. A 24-bit ADC (Analog-to-digital converter) is widely used in modern systems to digitize the voltage signal. In the absence of an input current, the ground has a natural self-potential (SP) voltage between different points on the surface (*Geoelectricity*). This can be caused by variations in the chemical properties of the soil or currents induced by changes of the geomagnetic field with time. The SP voltage can vary significantly with time. The voltage measuring system in the resistivity meter has a SP buck-out mechanism to remove the SP voltage signal. It usually also has notch filters (e.g., 50, 60, 16.7 Hz) to remove noise due to electric power lines and other technical infrastructure such as electrified railway systems.

IP surveys are widely used in mineral exploration, and recently there has been increasing interest in the environmental and engineering sector (Dahlin and Leroux 2012). While both time and frequency domain (Telford et al. 1990) measurement techniques have been used in mineral exploration, the time domain method appears to be the norm for

the less powerful equipment used in environmental and engineering surveys.

Electrode Types

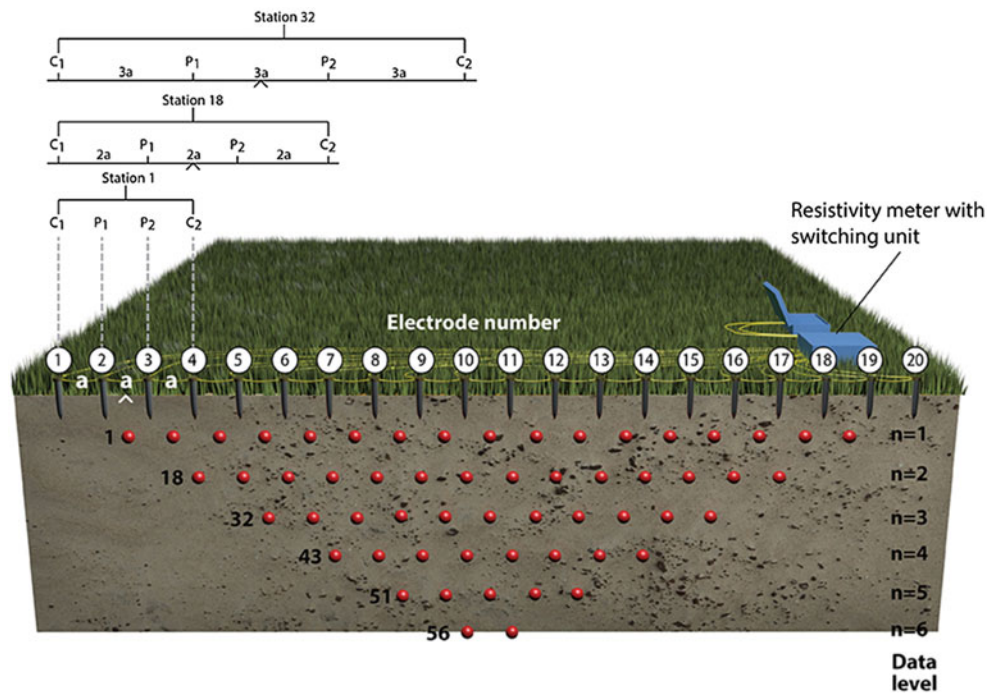
Metal stakes (Lu and Macnae 1998) are commonly used for the current electrodes as well as the potential electrodes in resistivity surveys. For some ground conditions where it is difficult to insert a stake electrode, flat-base (or plate) electrodes have been used (Athanasios et al. 2007); galvanic contact with the ground is achieved using an electrically conductive gel or mud at the base of the electrode. In IP surveys, nonpolarizable electrodes (Telford et al. 1990) are widely used as the potential electrodes to reduce SP noise. There has, however, been significant progress in using normal steel stakes as the potential electrodes (Dahlin and Leroux 2012) with multielectrode systems for IP surveys.

Multielectrode and Multichannel Systems

The development of multielectrode systems over the past 25 years has sparked a revolution in resistivity surveying. The advent of 2-D and 3-D (three-dimensional) resistivity tomography has opened up whole new application areas to electrical methods. Before the early 1990s, the electrical resistivity method was mainly used in resistivity sounding, profiling, and mapping surveys (Telford et al. 1990). Quantitative interpretation was mainly confined to 1-D (one-dimensional) structure of the subsurface consisting of horizontal layers (Keofoed 1979). The multielectrode systems made it practical to carry out two-dimensional imaging surveys that give a more accurate picture of the subsurface (Loke et al. 2013). A multielectrode system has about 25 or more electrodes connected to the resistivity meter via a multicore cable (Fig. 4). A switching circuitry controlled by an internal microcomputer within the resistivity-meter automatically selects the appropriate four electrodes for each measurement using a software program. Several variants of the multielectrode cable system exist, where the distinguishing factor tends to be the physical location of the switching circuitry. Some resistivity meters contain all the switching capability within the main unit, so that two or more long multicore cables are attached directly to the meter. The cables merely provide a galvanic contact between the meter and the ground. These systems have a practical limit on the number of electrode takeouts (commonly 32) in each cable. In second type of system, an external switch box is connected to the main meter via a

Instrumentation, Electrical

Resistivity, Fig. 4 Schematic diagram of a multielectrode system used for a two-dimensional electrical survey and an example sequence of measurements used to build up a pseudosection using the Wenner array



system bus. The switching capacity can then be extended by daisy-chaining switch boxes, each connecting to cable segments with multiple “dumb” electrodes. This design can vastly increase the possible number of electrodes available for simultaneous measurements, but it can also be less portable than the first arrangement, particularly in forested and rugged terrain. The third type of system uses “intelligent” electrodes, where the switching circuitry is de-centralized and placed at individual electrodes. Even the actual voltage measurement and analog-to-digital conversion may be carried out locally in such systems, potentially reducing interference problems (Stummer and Maurer 2001). Parsekian et al. (2017) have conducted experiments comparing many common multielectrode systems.

A recent development is multichannel capability for multielectrode systems that can greatly reduce the survey time. Only two electrodes can be used as the current electrodes at a single time, but the voltage can be measured between different pairs of potential electrodes. Commercial systems with four to ten channels are available and some research systems have more than 100 channels (Zhe et al. 2007; Rucker et al. 2014). New systems have been developed which transmit multiple current signals and can make several hundred measurements at a single time (Inazaki et al. 2015).

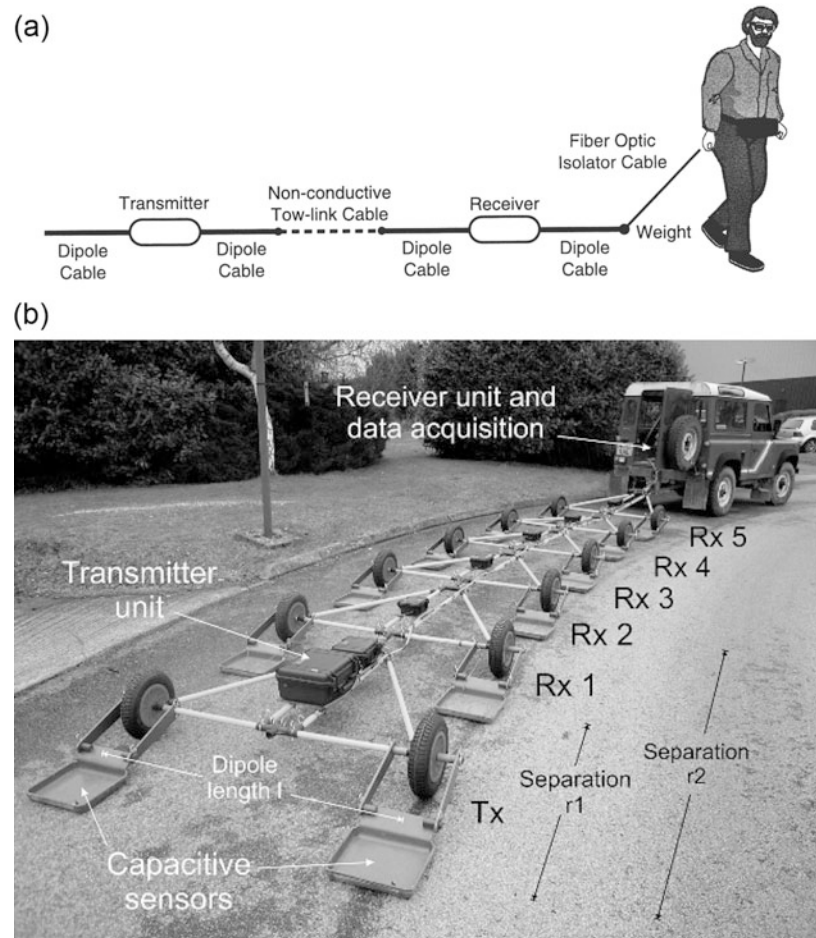
The multichannel multielectrode systems have been adapted for continuous profiling in water-covered areas

using electrodes attached to a cable pulled by a boat (Rucker and Noonan 2013; Simyrdanis et al. 2018). A PC coordinates the resistivity meter system data acquisition together with a GPS and water depth sounder. Continuous measuring systems for land surveys using cylindrical electrodes or spiked wheels to achieve continuous galvanic contact with the soil have also been developed (Loke et al. 2013).

Full-Waveform Systems

There has been considerable research on systems that record the full-waveform of the potential signal in an effort to extract more information from the data and to improve the signal-to-noise ratio using signal processing techniques (Zhe et al. 2007; Truffert et al. 2019). This is often the only practical approach for resistivity surveys in which large depths of investigation are required or where the nature of the target requires large dipole offsets. Efforts have been made to extract additional information about the subsurface IP characteristics from the full decay curve (Fiandaca et al. 2013). There have been renewed interest in vector array IP surveys (Truffert et al. 2019) using a distributed layout where each potential station has three electrodes that measure two components of electric field at right angles (Fig. 2d).

Instrumentation, Electrical Resistivity, Fig. 5 Types of capacitively coupled systems. (a) Line antenna (courtesy of Geometrics Inc.) and (b) plate-wire capacitive sensors (electrostatic quadrupole) arranged in a towed array. (Courtesy of British Geological Survey, NERC)



Capacitively Coupled Systems

Capacitively coupled systems use an oscillating, non-grounded electric dipole to generate current flow in the ground and a second similar dipole to measure the resulting potential distribution at the ground surface (Kuras et al. 2006) in areas where it is not possible to inject a galvanic current. Capacitively coupled sensor arrays allow the imaging of highly resistive targets in hardrock environments (Kuras et al. 2019; Murton et al. 2016). Mudler et al. (2019) describes making broadband spectral measurements with a new generation of capacitively coupled instrumentation, which can deliver complementary information for the characterization of the cryosphere compared to conventional low-frequency techniques.

Two major configurations have been used in mobile instruments with capacitively coupled sensors. The first configuration (line antenna type) uses cylindrical transmitters and receivers which are towed behind an operator (Fig. 5a). It gives measurements that are comparable to the galvanically coupled in-line dipole-dipole array (Kuras et al. 2006) and have maximum depths of investigation of 1–20 m. However, specially designed systems have been

used for surveys for depths of up to about 100 m in mineral exploration (Hildes et al. 2005). The second type (electrostatic quadrupole) uses flat metallic conductors (Fig. 5b) in an equatorial dipole-dipole configuration (Kuras et al. 2007) which has been used for survey depths of up to a few meters. Both configurations are dynamic measuring systems where the array configuration is fixed but the entire setup is moved during the survey for lateral coverage. Two-dimensional and three-dimensional surveys with dense lateral coverage can be rapidly conducted with these systems (Kuras et al. 2007).

Automated Resistivity Monitoring Systems

Automatic PC-controlled monitoring systems have been developed for detecting transient phenomena such as water seepage from dams, landslides, and solute transport (Uhlemann et al. 2017; Poje et al. 2018). These dedicated monitoring systems are becoming increasingly sophisticated and are now being deployed remotely with permanently installed electrode arrays, telemetric control, and data transfer, supported by automated data management systems to handle the large data

volumes generated through time-lapse data acquisition (Uhlemann et al. 2017). Other key developments include the production of low-power (e.g., 10 W) systems that are able to operate for long periods of time in the absence of mains electrical power and reactive monitoring capabilities where intensive resistivity measurements can be triggered by other environmental sensors such as rain gauges (Chambers et al. 2016).

Summary

There have been major advancements over the past few decades in the instrumentation for the electrical resistivity method. In addition to traditional one-dimensional surveys, the new multielectrode and multichannel systems have made it possible to efficiently carry out two-dimensional and three-dimensional imaging surveys that provide more accurate models of the subsurface geology. Computerized dynamic towed systems can map large areas on land and water rapidly. Capacitively coupled systems make it possible to survey areas with resistive surfaces where conventional galvanic systems cannot be used. Sophisticated PC-controlled monitoring systems are now available to automatically detect and record transient phenomena.

Cross-References

- [Electrical Properties of Rocks](#)
- [Electrical Resistivity Surveys and Data Interpretation](#)
- [Goelectromagnetism](#)

Acknowledgments We wish to thank Geometrics Inc., Landviser LLC, Abem Instrument AB and the British Geological Survey (NERC) for permission to use their illustrations in this paper. Kuras, Chambers and Wilkinson publish with the permission of the Executive Director, British Geological Survey (UKRI-NERC).

Bibliography

- Athanasiou EN, Tsourlos PI, Vargemezis GN, Papazachos CB, Tsokas GN (2007) Non-destructive DC resistivity surveying using flat-base electrodes. *Near Surf Geophysics* 5:263–272
- Chambers JE, Meldrum PI, Wilkinson PB, Gunn DA, Uhlemann SS, Kuras O, Swift RT, Inauen C, Butler S (2016) Remote condition assessment of geotechnical assets using a new low-power ERT monitoring system. EAGE near surface geoscience 2016 – 22nd European meeting of environmental and engineering geophysics, Barcelona
- Dahlin T, Leroux V (2012) Improvement in time-domain induced polarization data quality with multi-electrode systems by separating current and potential cables. *Near Surf Geophys* 10:545–565
- Fiandaca G, Ramm J, Binley A, Gazoty A, Christiansen AV, Auken E (2013) Resolving spectral information from time domain induced polarization data through 2-D inversion. *Geophys J Int* 192:631–646
- Hildes D, Dziuba F, Power M (2005) Kimberlite exploration using a capacitive-coupled resistivity system. In: Proceedings of 2005 CSEG national convention, May 16–19 2005. Calgary, Canada, pp 393–395
- Inazaki T, Imamura S, Sakanishi K (2015) A 4D ERT monitoring of water infiltration in the Vadose Zone by means of a high-speed resistivity measurement tool. In: Proceedings of the 28th annual symposium on the application of geophysics to engineering and environmental problems (SAGEEP 2015), 8 p
- Koefoed O (1979) *Geosounding principles 1: resistivity sounding measurements*. Elsevier Science Publishing, Amsterdam
- Kuras O, Beamish D, Meldrum PI, Ogilvy RD (2006) Fundamentals of the capacitive resistivity technique. *Geophysics* 71:G135–G152
- Kuras O, Meldrum PI, Beamish D, Ogilvy RD, Lala D (2007) Capacitive resistivity imaging with towed arrays. *J Environ Eng Geophys* 12:267–279
- Kuras O, Debouny T, Wilkinson PB, Field L, Milodowski AE, Metcalfe R, Norris S (2019) Investigating the saturation state of higher strength rock (HSR) by geoelectrical imaging at the core scale. 25th European meeting of environmental and engineering geophysics – near surface geoscience. EAGE, The Hague
- Loke MH, Chambers JE, Rucker DF, Kuras O, Wilkinson PB (2013) Recent developments in the direct-current geoelectrical imaging method. *J Appl Geophys* 95:135–156
- Lu K, Macnae JC (1998) The international campaign on intercomparison between electrodes for geoelectrical measurements. *Explor Geophys* 29:484–488
- Mudler J, Hördt A, Przyklenk A, Fiandaca G, Maurya PK, Hauck C (2019) Two-dimensional inversion of wideband spectral data from the capacitively coupled resistivity method – first applications in periglacial environments. *Cryosphere Discuss* 2019:1–24
- Murton JB, Kuras O, Krautblatter M, Cane T, Tschofen D, Uhlemann S, Schober S, Watson P (2016) Monitoring rock freezing and thawing by novel geoelectrical and acoustic techniques. *J Geophys Res-Earth Surf* 121:2309–2332
- Parsekian AD, Claes N, Singha K, Minsley BJ, Carr B, Voytek E, Harmon R, Kass A, Carey A, Thayer D, Flinchum B (2017) Comparing measurement response and inverted results of electrical resistivity tomography instruments. *J Environ Eng Geophys* 22:249–266
- Poje M, Salgado L, Rucker DF, Levitt M, Strohmeier J, McNeill M, Rucker K, Berumen R, Galbraith H (2018) Optimizing gold leaching with the aid of electrical resistivity geophysics. *Fast Times* 23(2):62–68
- Pozdnyakov AI, Rusakov AV, Shalaginova SM, Pozdnyakova AD (2009) Anisotropy of the properties of some anthropogenically transformed soils of Podzolic type. *Eurasian Soil Sci* 42:1218–1228
- Rucker DF, Noonan GE (2013) Using marine resistivity to map geotechnical properties: a case study in support of dredging the Panama Canal. *Near Surf Geophys* 11:625–637
- Rucker DF, Crook N, Winterton J, McNeill M, Baldyga CA, Noonan G, Fink JB (2014) Real-time electrical monitoring of reagent delivery during a subsurface amendment experiment. *Near Surf Geophys* 12:151–163
- Simyrdanis K, Moffat I, Papadopoulos N, Kowlessar J, Bailey M (2018) 3D mapping of the submerged Crowie barge using electrical resistivity tomography. *Int J Geophys* 2018(1):1–11
- Stummer P, Maurer HR (2001) Real-time experimental design applied to high-resolution direct-current resistivity surveys. In: Expanded abstracts, the international symposium on optical science and technology, SPIE 4491, pp 143–150
- Telford WM, Geldart LP, Sheriff RE (1990) *Applied geophysics*, 2nd edn. Cambridge University Press, Cambridge
- Truffert C, Gance J, Leite O (2019) State of the arts and new advanced ERT instrumentation and methods for urban planning challenges. EAGE-GSM 2nd Asia pacific meeting on near surface geoscience & engineering 22–26 April 2019, Kuala Lumpur, Malaysia, 1501–160

- Uhlemann S, Chambers J, Wilkinson P, Maurer H, Merritt A, Meldrum P, Kuras O, Gunn D, Smith A, Dijkstra T (2017) Four-dimensional imaging of moisture dynamics during landslide reactivation. *J Geophys Res Earth* 122(1):398–418
- Zhe J, Greenhalgh S, Marescot L (2007) Multichannel, full waveform and flexible electrode combination resistivity-imaging system. *Geophysics* 72:F57–F64

Instrumentation, EM

Steven Constable

Institute of Geophysics and Planetary Physics, Scripps
Institution of Oceanography, La Jolla, CA, USA

Definition

EM	Electromagnetic
CSEM	Controlled source EM
GDS	Geomagnetic depth sounding
MT	Magnetotelluric
SNR	Signal to noise ratio
ADC	Analog to digital conversion

Introduction

Here we look at instrumentation used for land and marine electromagnetic (EM) induction studies of Earth, which seek to characterize the electrical conductivity of the subsurface. Conductivity is driven by porosity and fluid content in the crust, temperature and mineralogy in the mantle, and minor proportions of highly conductive graphite, oxides, or sulfides in both. The principal tool for deep Earth conductivity sounding is the *magnetotelluric* (MT) method, in which time series measurements of the surface horizontal magnetic and electric fields are recorded and converted into a frequency-dependent transfer function. Loosely speaking, variations in the external magnetic field act as the input and induced electric fields in the earth act as the output. The frequency dependence of the amplitude and phase of the transfer function may be interpreted to recover electrical conductivity as a function of depth; longer periods propagate into deep structure while higher frequencies are absorbed in shallow structure. Station location constrains lateral variations in conductivity.

A related method is called *geomagnetic depth sounding* (GDS), in which a transfer function between the vertical and horizontal magnetic fields is used in a similar manner. In the past, arrays of portable GDS magnetometers were deployed to image lateral conductivity structure. Today, the main use of GDS is from data collected by magnetometers at

observatories and on magnetic satellites and used to image Earth structure at depths beyond the reach of the MT method. If assumptions are made about the morphology of the external magnetic source field, or estimates of the external field are made from the data themselves, then long period transfer functions can be interpreted to recover conductivity as a function of depth in a similar manner to MT data.

When near-surface crustal geology is of interest, man-made sources of EM energy (transmitters) can be used to probe conductivity, using either current passed through a grounded wire (an electric field source) or current passed through a loop (or loops) of wire (a magnetic field source). Electric or magnetic fields are measured as a function of transmitter frequency (frequency domain sounding) or, alternatively, transmitter turn-off time (time domain, or transient, sounding). Transmitter-receiver position, separation, and geometry (orientation, etc.) are also survey parameters. Measurements of this type fall into the category of controlled-source EM (CSEM) methods, although particular configurations often have distinct names or even trade names (e.g., Turam, LoTEM, SkyTEM).

Electric Field Measurement

Electrodes are used to make contact with the ground (or, in the marine case, seawater) for the purpose of making a measurement of potential difference, which is usually then divided by the electrode separation to form an estimate of electric field. Even with high input impedance amplifiers, some small current must be drawn to make the measurement. The simplest electrode is a metal rod in contact with damp earth or water, which will act as a source or sink of electrons e^- through the reaction.

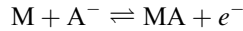


where M is the metal and M^+ is the metal cation. This is called an electrode of the first kind, and unless the metal cation is highly concentrated in the vicinity of the electrode, this reaction will not be reversible and the electrode will polarize, creating unwanted galvanic potentials and contributing to noise. This may be tolerable for high frequency measurements, and simple metal rods are used for high frequency MT measurements using man-made radio transmitters as the source field.

If the metal rod is immersed in a concentrated aqueous solution of a compatible salt, then the reaction is reversible and the electrode is non-polarizing. The most commonly used electrode of this type is the copper-copper sulfate “porous pot.” The “pot” containing the copper sulfate solution has a semipermeable barrier as its base, usually unglazed ceramic, which makes electrolytic contact with any moisture in the

soil. Such electrodes are commonly used for broadband MT measurements in the frequency range of 1 mHz to 100 Hz.

Non-polarizing electrodes of the second kind are made by coating a metal with a sparingly soluble salt and immersing in an aqueous electrolyte containing a compatible anion. The reaction here is



where A^- is the anion. On land the most commonly used electrode of this type is lead-lead chloride immersed in solution of potassium or sodium chloride (e.g., Petiau 2000). For marine work silver-silver chloride immersed in seawater, which contains about 30 g/l sodium chloride, is commonly used (e.g., Constable 2013), and recently silver chloride electrodes have been manufactured for use on land. Silver chloride electrodes have lower temperature coefficients and higher stability than those of the first kind, and are preferred for long-period MT work on land and almost all MT studies in the marine environment. Silver-silver chloride electrodes are also favored for marine CSEM soundings, although polarizing, capacitively coupled, electrodes such as the carbon fiber sensors of Crona et al. (2001) have been used.

Electrodes are separated by cables of order 100 m long on land to integrate the electric fields and provide a measurable potential difference. Contact impedance for electrodes can be as large as several kilohms, and so the input impedance of amplifiers should be large, but gain need not be very high since signals will be up to several millivolts across a 100-m antenna, and a reasonably well-designed amplifier will not contribute significantly to noise.

For marine measurements the situation is very different. The electrode contact impedance is much lower (around 10 Ω), the signal levels are much smaller (a few $\mu\text{V/m}$), and the antenna lengths are usually shorter as well (around 10 m). For long period MT simple amplifiers with gains of about 100 are adequate, but for broadband MT and CSEM measurements at frequencies of 0.1 Hz and above, capacitively coupled chopper amplifiers with very low noise and gains of 10,000 and higher are used (Constable 2013).

Magnetic Field Measurement

Magnetic field sensors fall into two broad categories: total field magnetometers and vector magnetometers. Total field proton and nuclear precession devices are used for magnetic mapping, but for EM induction vector measurements are needed, using either induction coils or fluxgate sensors. Superconducting (SQUID) magnetometers have been used for induction studies, but so rarely that they will not be discussed here.

A simple loop of wire can be used as a magnetic sensor based on Faraday's Law, in which a time-varying magnetic field B creating a magnetic flux through area A will induce a voltage V in a loop of wire with N turns:

$$V = -NA \frac{dB}{dt} = -NA\mu \frac{dH}{dt}$$

where H is the magnetizing field and μ is permeability, which for an air-cored loop is the free-space value of $47\pi \times 10^{-7}$ H/m.

For CSEM operations at high frequencies (kilohertz range) the derivative term is large enough that an adequate signal may be measured in this way, particularly for single frequency CSEM in which lock-in amplifiers can be used. Alternatively, for time domain studies in which a large loop of wire is used as the transmitter, the area may be as much as 10^4 m^2 and it is possible to make measurements using the transmitter loop after switch-off. Normally however, induction coils (also called search coils, Fig. 1a) are used, in which thousands of turns of wire are wound around a magnetically permeable core of soft ferrite or iron-nickel alloy (permalloy or mu-metal). The core captures the magnetic field and multiplies the flux by an effective permeability μ_e , determined by a combination of geometry and relative permeability, producing a voltage

$$V = -NA\mu\mu_e \frac{dH}{dt}.$$

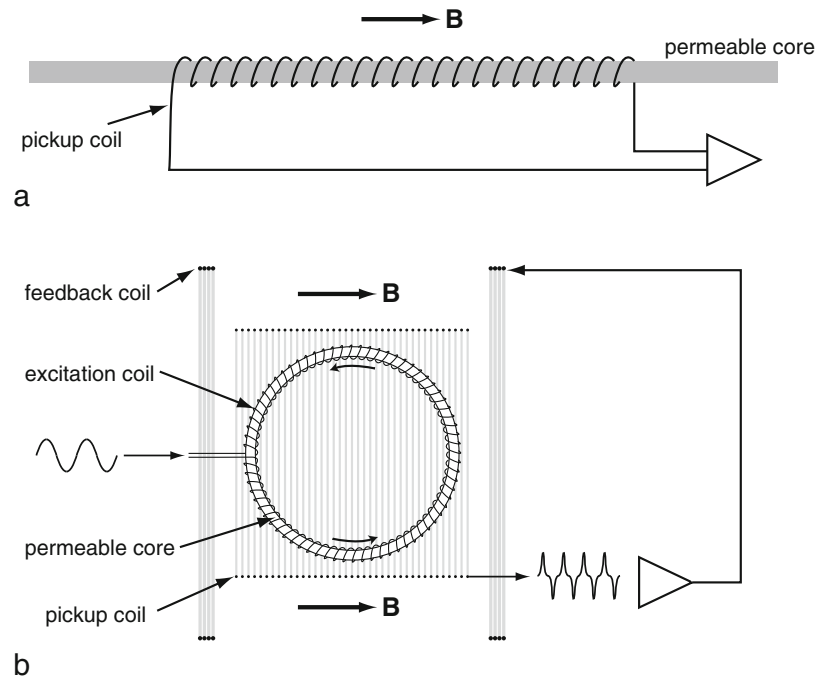
For long, cylindrical, highly permeable cores of length L and diameter d the effective permeability depends only on geometry and is given approximately by (Tumanski 2007):

$$\mu_e = \frac{(L/d)^2}{(\ln(2L/d) - 1)}.$$

We see that increasing length improves sensitivity, and typical broadband MT coils are of order 1 m long. The main source of noise for induction coils is the thermal (or Johnson) noise of the resistance in the wire windings, and so increasing N does not necessarily increase the signal to noise ratio (SNR) because it also increases resistance. Indeed, to a first approximation the SNR is proportional to the mass of copper used in the windings. However, if weight is an issue, as it may be for marine instruments, aluminum windings can be used for a significant improvement in weight at a modest cost in resistance noise. The noise of a good-quality induction coil used for EM induction studies is around $10^{-9} \text{ nT}^2/\text{Hz}$ at 1 Hz with a red $1/f$ type spectrum. Since motion in Earth's magnetic field can also produce a time-varying signal, on land induction coils are usually buried during operation to prevent

Instrumentation, EM,

Fig. 1 (a) Schematic depicting the operation of an induction coil magnetometer. A coil of wire is wound around a permeable core. (b) The ring-core fluxgate magnetometer. An excitation coil drives the core into saturation in two different directions. The direction aligned with the external magnetic field will saturate sooner than the direction apposed to the field, creating a pulse of unbalanced flux twice per excitation cycle, which is detected by a synchronous amplifier tuned to twice the excitation frequency. In practice, fluxgates are often used as null sensors with the output of the pickup coil used to null the field using a feedback coil; the current in the feedback coil becomes the measurement



movement. For marine MT and CSEM studies, the magnetometers must be stabilized by heavy anchors, usually made from concrete to avoid magnetic contamination.

Although integrating amplifiers or feedback can be used to flatten the frequency response of induction coils, the one-pole low-cut response of the dH/dt term is a fairly good match to the red spectrum of natural magnetic field variations. However, reasonable signal to noise performance for induction coils is still limited to periods shorter than several thousand seconds. For longer period MT and GDS studies, fluxgate magnetometers are preferred. Fluxgates use the principle that variations in magnetic flux may be generated for a fixed magnetic field by varying the effective permeability of the core material:

$$V = -NA \frac{dB}{dt} = -NA\mu \frac{d\mu_e}{dt} H.$$

This is achieved by winding the permeable core with an excitation coil which saturates the core in alternate directions using a sinusoidal or square wave current of around 1 kHz. A saturated core has an effective permeability of zero, since it is no longer influenced by the external field H , and so the excitation current generates the time-varying permeability required to get an inductive response.

Various core geometries can be used for fluxgates, but the most common type of geophysical sensor uses a ring-core geometry (Fig. 1b). The excitation coil saturates the two sides of the core symmetrically in opposite directions; the side of the coil excited in the same direction as the external magnetic

field saturates sooner, while the opposite side saturates later. This asymmetry produces a time varying net flux at twice the excitation frequency, which is detected by a pickup coil wound around the core; the orientation of the detection coil determines the direction of sensitivity. A lock-in amplifier tuned to twice the excitation frequency will thus have an output proportional to the external field. However, this output is only linear over a range of several hundred nanotesla, and also will be sensitive to variations in μ and μ_e caused by temperature, so most observatory quality instruments are operated as null sensors. The total magnetic field (usually the component of Earth's field aligned with the sensor) is cancelled by a solenoid or Helmholtz coils. The output of the fluxgate is fed back to control the current in the nulling coils, and this current then becomes the measurement of low frequency changes in the magnetic field.

The effective permeability of a ring-core of diameter L and thickness d is approximately

$$\mu_e = \frac{L}{d}$$

for large L/d ; ring-core sensors are usually about 2 cm in diameter and a millimeter or so thick, with quoted noise levels of order 10^{-4} nT²/Hz at 0.1 Hz (Primdahl 1979). By design, fluxgates sense the static magnetic field component in a given direction, which is typically 40 μ T on Earth (possibly as large as 60 μ T), and so the low-frequency response of fluxgates extends to DC. Excitation frequencies are high enough that variations up to a few tens of hertz can be recorded, but long

period MT instruments typically sample only at about 1 Hz, since the noise in fluxgates becomes higher than induction coils at a frequency of about 0.01 Hz. Also, making measurements of field variations in the presence of Earth's total field requires a large dynamic range, and with 24-bit analog to digital conversion (ADC), the least count will be of order 0.01 nT, and resolutions of 0.1 nT are more typical. Least count for an induction coil sensors are of order 0.01 pT.

EM Transmitters

Either a magnetic or electric field transmitter can be used for CSEM soundings. Electric transmitters consist of a long (hundreds of meters) wire grounded at both ends, through which a modulated current is driven. For large currents it is easier to switch square waves than generate sinusoidal waveforms, and if the switching is controlled rapidly then time domain measurements may be made. Alternatively, a swept or random sequence of binary pulses can be used and the receiver data deconvolved to generate a broadband response. For marine CSEM surveys, the high conductivity of seawater allows transmitter currents of up to a thousand amps to be used (Constable 2013). Magnetic transmitters consist of large loops of wire through which a current is passed, and are mainly used for time domain studies in land and airborne systems.

Data Acquisition Systems

The science and technology behind EM sensors have not changed significantly for many decades, but development of modern electronics has revolutionized the way MT and CSEM data are collected. The advent of microprocessors; 24 and 32 bit ADC modules; low power, high capacity, static digital memory; and the global positioning system (GPS), which allows time to be measured with microsecond accuracy anywhere on the surface of Earth, together allow continuous, time-synchronized, measurements to be made using small, portable, low power logging systems.

Time synchronization is important in MT data acquisition, since noise reduction in processing raw time series to frequency domain response functions depends on cross-spectra being estimated between two or more remotely separated sites. For CSEM soundings, synchronous stacking and estimation of phase lag between source and receiver both demand good timing for the transmitter and receiver. In the past this was done with wired or wireless links between sites for both CSEM and MT. Now, all sites can be individually clocked with GPS timing for later synchronous processing, at least on land. Marine MT and CSEM receivers must have installed clocks that are reset using GPS time before deployment, and

whose drift can be determined using GPS on recovery. In this way time accurate to better than a millisecond per day can be achieved. Since marine EM transmitters are attached to, and powered by, ships, GPS time can be used to control the output waveforms.

Similarly, in the past wired or wireless connections between the measurement equipment and bulky data acquisition and storage systems were needed, but now data can be stored site by site or even channel by channel, using integrated microprocessor controlled acquisition and solid state memory. Clearly, this has been of huge benefit to the collection of marine data, but land acquisition now also routinely uses this approach.

Navigation

The GPS system, as its name suggests, provides highly accurate locations and makes surveying land MT and CSEM stations straightforward. For marine MT and CSEM, the positioning of seafloor receivers is more difficult, since during deployment they will drift 50–100 m from the release point as they sink down through currents in the water column. For MT surveys the release point may provide sufficient accuracy, but this is not the case for CSEM surveys. In this case accurate positions need to be derived from acoustic ranging from the survey ship, whose position is known from GPS. Short (1–10 ms) and high frequency (10–100 kHz) pulses of sound are emitted from the ship and transponders on the receiver instruments reply to these with similar “pings.” The two-way travel time, along with estimates of sound speed in water derived from oceanographic tables or direct measurement, is used to calculate range. Two approaches are possible; long baseline surveys involve triangulation using ranges from three or more ship positions, and short baseline surveys use equipment which can estimate direction as well as range using multiple units or elements for the ship's transponders. For marine CSEM transmitter navigation, short baseline measurements are supplemented by depth, height, and possibly inertial sensors on the transmitter.

On land the orientation of sensors is measured or set using levels and magnetic compasses during installation (fluxgate sensors can act as their own compasses). For seafloor measurements, orientation needs to be estimated using recording tiltmeters and compasses.

Summary

Recordings of natural or man-made variations in the magnetic field may be made with fluxgate sensors or induction coils, depending on the frequency of interest. In terms of signal to noise, the crossover between the two sensors is about 0.01 Hz,

with induction coils used for measurements higher than 1 mHz and fluxgates for measurements lower than 1 Hz. Exact frequencies depend on the particular equipment and application. Electric field sensors are typically made using pairs of non-polarizing electrodes to make contact with ground or water, separated and connected to amplifiers by 10–100 m of wire. At high frequencies simple metal rods may suffice as electrodes. Modern electronics and timing derived from the global positioning system have made data collection using autonomous loggers standard for both land and marine MT and CSEM data acquisition.

Cross-References

- [Electrical Resistivity Surveys and Data Interpretation](#)
- [Electromagnetic Pulsations and Magnetic Storms](#)
- [Geomagnetic Field, Theory](#)
- [Paleomagnetism, Measurement Techniques and Instrumentation](#)

Bibliography

- Constable S (2013) Review paper: instrumentation for marine magnetotelluric and controlled source electromagnetic sounding. *Geophys Prospect* 61:505–532
- Crona L, Fristedt T, Lundberg P, Sigra P (2001) Field tests of a new type of graphite-fiber electrode for measuring motionally induced voltages. *J Atmos Ocean Technol* 18:92–99
- Petiau G (2000) Second generation of lead-lead chloride electrodes for geophysical applications. *Pure Appl Geophys* 157:357–382
- Primdahl F (1979) The fluxgate magnetometer. *J Phys E Sci Instrum* 12:241–253
- Ripka P (1992) Review of fluxgate sensors. *Sensors Actuators A* 33:129–141
- Tumanski S (2007) Induction coil sensors – a review. *Meas Sci Technol* 18:R31–R46

International Geophysical Year

Ralph W. Baird
Baird Petrophysical International, Houston, TX, USA

Definition

The International Geophysical Year (IGY 1957–1958) was the most successful global effort to coordinate the measurement and collection of geophysical data from around the world during a period of anticipated maximum solar activity. It was also timed as a continuation of a series of international efforts to collect data at the poles, called the International

Polar Years I and II (IPY 1882–1883 and IPY 1932–1933). The IGY was planned for 18 months, July 1, 1957, through December 31, 1958, and then a follow-on organization, the International Geophysical Cooperation (IGC 1959) was necessary to complete the work and assemble the results, from January 1, 1959, through December 31, 1959.

During 2007–2009, there was a fourth international effort referred to as IGY-2 or IGY + 50. But instead of a coordinated singular management initiative, it was organized as separate initiatives; the Electronic Geophysical Year (eGY 2007–2008), the International Polar Year (IPY 2007–2008), the International Heliophysical Year (IHY 2007–2009), and the International Year of Planet Earth (IYPE 2007–2009). To capture the work and materials from the IGY 1957–1958, organizers of the IHY created a legacy archive and recognition program entitled IGY Gold. During and after the IGY 1957–1958, the non-science public joined the scientists with pamphlets, magazine articles, films, children's toys, games, and books, all aimed at encouraging exploration and discovery.

Introduction and Concept

A lack of knowledge of natural Earth processes and the interest by thousands of Earth scientists to explore and discover relationships of the processes, combined with modern communications and global transportation improvements, inspired and motivated the organizers and participants of IGY in a postwar world (the early 1950s). The more measurements were studied, the more questions evolved. To get the answers, the world Earth science community needed to cooperate and collect data together and facilitate a means to share this new data to enable new Earth science discoveries.

The concept seems simple today in the twenty-first century: propose global efforts to collect physical and chemical measurements of the Earth on a semicontinuous basis over a specified period of time. In the early 1950s, this task was burdened with a lack of infrastructure and a lack of resources in all parts of the Earth, governed by separate regimes speaking hundreds of different languages and recovering from wartime conflicts. Just getting from point A to point B was a time-consuming task. All this was about to change.

History

While in Maryland, James Van Allen (1914–2006) and his wife Abigail hosted a dinner on April 5, 1950, for British geophysicist Sidney Chapman (1888–1970), a theoretical physicist interested in the Earth's magnetic phenomena and a participant in IPY II 1932–1933. Also present at dinner was Lloyd Berkner (1905–1967), a former radio engineer who had

been on Admiral Byrd's 1928–1930 Antarctic expedition. According to Van Allen, the dinner conversation ranged widely over geophysics and especially geomagnetism and ionospheric physics. Following dinner, as they were all sipping brandy in the living room, Berkner turned to Chapman and said, "Sydney, don't you think it is about time for another international polar year?" Chapman immediately embraced the suggestion, remarking that he had been thinking along the same lines himself (Van Allen 1998).

The time was ripe. Technological improvements in instrumentation and rocketry had enabled scientists to probe much farther into the atmosphere and deep into the Earth. In the process of enlisting support among the international scientific societies, Chapman and Berkner found a strong preference for a global program encompassing additional geographical regions and additional physical science disciplines.

Chapman first presented the idea for a third IPY to the member scientific unions of the International Council of Scientific Unions (ICSU). The unions, in turn, presented the proposal to the ICSU General Assembly, and ICSU, in turn, invited the World Meteorological Organization (WMO) to participate as well as the national organizations adhering to ICSU. By 1953, there were 26 countries signed up in what came to be known at the International Geophysical Year 1957–1958. The disciplines included most of the Earth, atmospheric, and oceanic sciences, covering many parts of the globe beyond the polar regions.

Scope

By 1954, the IGY organizing committee (set up by ICSU in 1952) was known as CSAGI after its French name, Comité Spécial de l'Année Géophysique Internationale. The CSAGI established goals for IGY projects to address several areas of specific interest:

- Problems requiring concurrent synoptic observations at many points involving cooperative observations by many nations.
- Problems in the geophysical sciences whose solutions would be aided by the availability of synoptic or other concentrated work during the IGY.
- Observations of all major geophysical phenomena in relatively inaccessible regions of the Earth that can be occupied during the IGY because of extraordinary effort during that interval (the Arctic and Antarctic).
- Epochal observations of slowly varying terrestrial phenomena.

These were not arbitrary or unreasonable criteria. Based on this defined planning and framework, the scope of the IGY program materialized, and the process to organize field

operations, data collection synchronization, data reporting, and assembly and archiving began.

Operations

When comparing the organization of IGY to today's professional and scientific working groups, the IGY was controlled by two separate bodies, the CSAGI and the ICSU. Areas of science emphasis covered:

Meteorology	Cosmic Rays	Gravity
Geomagnetism	Glaciology	Nuclear Radiation
Aurora and Air Glow	Oceanography	Latitude and Longitude
Ionosphere	Rockets and Satellites	World Days and Communication
Solar Activity	Seismology	

Participating countries represented in expeditions or contributing and sponsoring in-country data collection included:

Afghanistan	German Democratic Republic	Norway
Argentina	German Federal Republic	Pakistan
Australia	Ghana	Panama
Austria	Greece	Peru
Belgium	Guatemala	Philippines
Bolivia	Haiti	Poland
Brazil	Hawaii	Portugal
Bulgaria	Honduras	Rhodesia and Nyasaland
Burma	Hungary	Romania
Canada	Iceland	San Salvador
Ceylon	India	Saudi Arabia
Chile	Indonesia	Spain
China (Nationalist)	Iran	Sudan
China (People's Republic)	Ireland	Sweden
Colombia	Israel	Switzerland
Costa Rica	Italy	Thailand
Cuba	Japan	Tunisia
Czechoslovakia	Korea, Democratic Republic of	Turkey
Denmark	Libya	Union of South Africa
Dominican Republic	Malaya	Union of Soviet Socialist Republics
East Africa	Mexico	United Kingdom
Ecuador	Mongolia, People's Republic of	United States of America
Egypt	Morocco	Uruguay
Ethiopia	Netherlands	Venezuela
Finland	New Zealand	Vietnam, Democratic Republic of
France	Nicaragua	Vietnam, Republic of
		Yugoslavia

Achievements

General achievements of IGY were:

- Global cooperation for improved Earth physical and chemical measurements.
- Improved awareness and understanding of Earth processes.
- Coordinated collection and assembly of multidisciplinary data.
- Archive and sustainable safekeeping of collected and shared data.
- Inspiration and hope for future similar programs and initiatives.

Some of the most significant IGY achievements were:

- Defining the system of mid-ocean ridges that encircle the globe, furthering our understanding of the Earth's crust and the theory of plate tectonics.
- Discovery of the Van Allen radiation belts. These belts surround the Earth at altitudes of hundreds and at thousands of kilometers above the surface and are significant to present-day electronic communications.
- Collection of synoptic data, a comprehensive overview of global physical phenomena. These achievements were accomplished through organization of various scientific fields under the International Council of Scientific Unions (ICSU). This union created a series of technical panels with scientific goals and facilitated international cooperation.
- Under the collection of synoptic data, special attention was given to the Antarctic continent. Neither the race for the South Pole in the early 1900s nor the age of exploration in the 1930s brought the influx of humanity experienced during the IGY to the ice-covered continent.
- A new value for total abundance of water in the form of ice on the Antarctic continent. Ninety percent of the planet's ice is found on and around the continent, locking 68% of the world's fresh water in the Southern Hemisphere.
- Improved meteorological predictions by understanding the weather patterns of the Southern Hemisphere.
- Advancements in the theoretical analysis of glaciers.
- Seismology of the Southern Hemisphere.
- The scientific cooperation in Antarctica paved the way for the Antarctic Treaty. The treaty signed on December 1, 1959, created a continent free from nuclear weapons and open to scientific research; the first truly international territory.
- Improved science and math education. Through implied competition among countries, IGY generated a new sense of the importance of math and science to competitive problem-solving.
- Sputnik and satellite measurements. IGY provided the first peaceful use of military equipment and technology to enhance the measurement capability and later communications of people on Earth.
- World Days and Communication. During the IGY, a calendar was arranged to achieve simultaneous observations in most disciplines. During some periods, intensification in observations was considered, and the World Days program was established. There were three classes of special days: Regular World Days (RWD), World Meteorological Intervals (WMI), and Special World Intervals (SWI). RWD and WMI were picked in advance on the calendar. SWI were designed day-to-day and broadcast by the World Warning Agency (AGIWARN).
- World Data Centers. The IGY was predicated on full and open data exchange. The World Data Centers were created to provide equitable access for use by all qualified scientists for public good and for geophysical research as a tool for sustainability.
- World Gravity Map. Over 60,000 observations and additional 150,000 anomaly values results from the various IGY programs. The raw data were prepared for storage and further analysis on punched paper cards for machine analysis. The final values were published as a monograph and supplied to the World Data Centers.
- Postage stamps, films, songs, and pamphlets. Outside of the participating scientists, the new approach included outreach and generational attraction to science at all levels. These creative media and documentary projects contributed to the events during the IGY and continue through today to provide accounts and anecdotes about the progress made by participants in IGY.

Summary

The International Geophysical Year (IGY) was, *and is today*, inspirational and an excellent example of cooperation, collaboration, and communication among the different nations of the world. It's period of time for the event labeled a *Year* was planned as 18 months and was the start of further cooperation, collaboration, and communication among the founders and participants and added members over the years following the IGY. Seven decades of work that included volumes of data collection is now available for further research and study and planning for what may be next. Follow-on achievements are documented, but the success of the IGY must be etched into the progress made from the original works in 1957–1958. IGY was a giant step for mankind, and continued step-by-step knowledge from new measurements is expected.

Cross-References

[Electronic Geophysical Year](#)

[International Polar Year 2007–2008](#)

[International Year of Planet Earth](#)

Bibliography

- Baird RW, Hall JD (2007) Introduction to this special section: eGY. *Lead Edge* 26(10):1290–1325
- Berkner LV (1955) Letter to H.H. Hess. October 27. Box 4, Institute of Theoretical Physics Folder. In: E.H. Vestine Papers. Library of Congress, Manuscripts Division, Washington, DC
- Chapman S (1959) IGY: Year of discovery. University of Michigan Press, Ann Arbor
- Chapman S (1960) From polar years to geophysical year. *Stud Geophys Geod* 4:313–324
- Dickson P (2001) *Sputnik: The shock of the century*. Walker, New York
- Fagen D (1982) *I.G.Y. (What a Beautiful World)*. Warner Bros. Records Inc, Los Angeles
- IGY Special Section (1957) *Jr Scholast* 41(7):7–15. Junior Scholastic, New York
- International Council of Scientific Unions (1957–1970) *Annals of the international geophysical year*. Pergamon, London/New York
- International Council of Scientific Unions (1959) *Annals of the international geophysical year*, vol Vol. 1. Pergamon, New York
- Kelly W (1957) *G.O. Fizzickle*. Simon and Schuster, New York
- Korsmo FL (2004) Shaping up planet earth: the International Geophysical Year (1957–1958) and communicating science through print and film media. *Sci Commun* 26(2):162–187
- Korsmo FL (2009) The policy process and the International Geophysical Year, 1957–1958. In: *Smithsonian at the Poles: Contributions to the International Polar Year Science*. Smithsonian Institution Scholarly Press, Washington, DC, pp 23–34
- Korsmo FL (2010) The origins and principles of the World Data Center System. *Data Sci J* 8:IGY55–IGY65
- Marshack A (1958) *The world in space: The story of the International Geophysical Year*. Thomas, Nelson, New York
- National Academy of Sciences (1958) *Planet earth: The mystery with 100,000 clues*. NAS, Washington, DC
- National Academy of Sciences (1965) *Report on the U.S. Program for the International Geophysical Year*. NAS, Washington, DC
- Nicolet M (1984) The International Geophysical Year (1957–1958): great achievements and minor obstacles. *GeoJournal* 8(4):303–320
- Odishaw H, Ruttenberg S (eds) (1958) *Geophysics and the IGY: Proceedings of the Symposium at the Opening of the International Geophysical Year*. American Geophysical Union, Washington, DC
- Schanche D (1957) Space beyond Sputnik lies within our grasp. *Life Mag* 43(17):26–29
- Sputnik Mania (2007) DVD. Directed by David Hoffman. Varied Directions, Santa Cruz
- Sullivan W (1961) *Assault on the Unknown: The International Geophysical Year*. McGraw-Hill, New York
- Thompson B (2007) Preserving IGY's history and legacy. *Lead Edge* 26(10):1296–1297
- Van Allen JA (1997) Interview with B. Shoemaker, November 18, 1997. Transcript, Oral History Collection. Ohio State University, Columbus
- Van Allen JA (1998) Genesis of the International Geophysical Year. *The Polar Times* 2(11):5
- Wilson JT (1960) *I.G.Y.: The year of the new moons*. Alfred A. Knopf, New York

International Gravity Formula

Hans-Jürgen Götze

Institute of Geosciences, Working group Satellite- and Aerogeophysics, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

Definition

As a first approximation, the Earth is a rotating sphere. As a second approximation, it can be regarded as an equipotential ellipsoid of revolution. According to Moritz (1980), the theory of this equipotential ellipsoid was first given by P. Pizzetti in 1894. It was further elaborated by C. Somigliana in 1929 and served already as the basis for the International Gravity Formula adopted at the General Assembly of the International Union of Geodesy and Geophysics (IUGG) in Stockholm in 1930.

One particular ellipsoid of revolution, also called the “normal Earth” or “normal spheroid,” is the one having the same angular velocity and the same mass as the actual Earth, the potential U_0 on the ellipsoid surface equal to the potential W_0 on the geoid, and the center coincident with the center of mass of the Earth. The Geodetic Reference System 1967 (GRS 67), Geodetic Reference System 1980 (GRS 80), and World Geodetic System 1984 (WGS 84) all are “normal Earth” models.

The normal or theoretical gravity is the gravity effect due to an equipotential ellipsoid of revolution. Approximate formulas are used widely even though we can calculate the exact theoretical gravity analytically. The conventionally used International Gravity Formula is obtained by substituting the parameters of the relevant reference ellipsoid. Helmert's 1901 Gravity Formula and International Gravity Formulas 1930, 1967, and 1980 correspond, respectively, to the Helmert 1906, International 1924, GRS 67, and GRS 80 ellipsoids. For example, the recent 1980 International Gravity Formula is (Moritz 1980):

$$\gamma_{1980} = 9.780327 (1 + 0.0053024 \sin^2 \varphi - 0.0000058 \sin^2 2\varphi) \text{ ms}^{-2}$$

where φ is the geodetic latitude.

The resulting difference between the 1980 International Gravity Formula and the 1930 International Gravity Formula is:

$$\gamma_{1980} - \gamma_{1930} = -16.3 + 13.7 \sin^2 \varphi 10^{-5} \text{ m/s}^2.$$

Similarly, one can compare the 1967 formula to the 1980 formula in use today. The difference between the two is relatively small:

$$\gamma_{1980} - \gamma_{1967} = 0.8316 + 0.0782 \sin^2 \phi \cdot 10^{-5} \text{ m/s}^2.$$

The first term of the International Gravity Formula is the value of gravity at the equator on the ellipsoid surface. Unfortunately, in the 1930s no one really knew what it was. The most reliable estimate at that time was based on absolute gravity measurements made by pendulums at the Geodetic Institute Potsdam in 1906. The Potsdam gravity value served as an absolute datum for worldwide gravity networks from 1909 to 1971. In the 1960s, new measurements across continents were made by precise absolute and relative gravity meters and became the network of IGSN71 still in use today. A mean difference between the Potsdam datum and the IGSN71 reference has been found to be $14 \times 10^{-5} \text{ m/s}^2$ (Woollard 1979).

The theoretical gravity is the gravity effect due to an equipotential ellipsoid of revolution. Approximate formulas are used widely. In fact, the theoretical gravity at any position on, above, or below the ellipsoid surface can be calculated using closed-form expressions also. Such a closed-form expression for the theoretical gravity on the surface of the ellipsoid is given by the formula of Somigliana (Heiskanen and Moritz 1967, p. 76).

The Somigliana–Pizetti formula gives the magnitude of normal gravity on the surface of a geocentric reference ellipsoid:

$$\gamma = \gamma_e \frac{1 + k \sin^2 \phi}{\sqrt{1 - e^2 \sin^2 \phi}}$$

where k is the normal gravity constant (not to be confused with the universal gravitational constant, G), γ_e is the normal gravitational acceleration at the equator, and e^2 is the square of the first numerical eccentricity of the ellipsoid. Importantly, the geocentric latitude N , which is compatible with the GRS80, must be used in the Somigliana formula (Li and Götze 2001; Hackney and Featherstone 2003).

Parameters used to compute the normal gravity using the Somigliana–Pizetti formula are given values for the GRS80 ellipsoid from Moritz (1980) (Table 1).

The recent theoretical formula for gravity as a function of latitude φ using the Somigliana equation provides (Hinze et al. 2013):

$$\begin{aligned} \gamma(\varphi)_{1980} \\ = 9.7803267715 \left[\frac{1 + 0.001931851353 \sin^2 \varphi}{1 - 0.0066943800229 \sin^2 \varphi} \right] \text{ ms}^{-2} \end{aligned}$$

A later refinement of this normal gravity formula is based on WGS 84 (World Geodetic System of 1984) and gives:

$$\begin{aligned} \gamma(\varphi)_{1984} \\ = 9.7803253359 \left[\frac{1 + 0.00193185265241 \sin^2 \varphi}{1 - 0.00669437999013 \sin^2 \varphi} \right] \text{ ms}^{-2} \end{aligned}$$

International Gravity Formula, Table 1 Parameter definition and GRS80 values

A	Ellipsoid semimajor axis	6,378,137 m
B	Ellipsoid semiminor axis	6,356,752.3141 m
γ_e	Equatorial normal gravity	$9.7803267715 \text{ ms}^{-2}$
γ_p	Polar normal gravity	$9.8321863685 \text{ ms}^{-2}$
k	$(b\gamma_p/a\gamma_e) - 1$	0.001931851353
e^2	$(a^2 - b^2)/a^2$	0.00669438002290
f	$(a - b)/a$	0.00335281068118
m	$(\omega^2 a^2 b)/GM$	0.00344978600308
ω	Angular velocity	$7.292115 \times 10^{-5} \text{ rad s}^{-1}$
GM	Geocentric gravitational constant	$3.986005 \times 10^{14} \text{ m}^3 \text{ s}^{-2}$

with: $\gamma_p = 9.8321849378 \text{ ms}^{-2}$ (polar normal gravity). The small differences between the GRS 80 ellipsoid and the current WGS 84 ellipsoid have virtually no practical consequence.

Cross-References

[Gravity Anomalies, Interpretation](#)
[Gravity Field of the Earth](#)
[Gravity Method, Airborne](#)
[Gravity, Data to Anomalies](#)

Bibliography

- Hackney RI, Featherstone WE (2003) Geodetic versus geophysical perspectives of the “gravity anomaly”. *Geophys J Int* 154:35–43
- Heiskanen WA, Moritz H (1967) *Physical geodesy*. W. H. Freeman, San Francisco
- Hinze WJ, von Frese RRB, Saad AH (2013) *Gravity and Magnetic Exploration: Principles, Practices, and Applications*. Cambridge University Press, p 130. ISBN 978-1-107-32819-8.
- Li X, Götze H-J (2001) Ellipsoid, geoid, gravity, geodesy and geophysics. *Geophysics* 66:1660–1668
- Moritz H (1980) Geodetic reference system. *Bull Geod* 54:395–405
- Woollard GP (1979) The new gravity system – changes in international gravity base values and anomaly values. *Geophysics* 44:1352–1366

International Polar Year 2007–2008

David J. Carlson
 British Antarctic Survey, IPY International Programme
 Office, Cambridge, UK

Concept

The International Polar Year 2007–2008 (IPY) was an intensive, internationally coordinated scientific research campaign in the Arctic and the Antarctic sponsored by the

International Council for Science (ICSU) and the World Meteorological Organization (WMO). Working together through international projects, IPY researchers examined all aspects of polar regions. IPY highlighted the global importance of polar processes and the urgent need to understand and track extremely rapid changes occurring at high latitudes.

History

By most accounts, the 2007–2008 IPY represented the fourth IPY: 1882–1883, 1932–1933, 1957–1958 (that event grew from a Polar Year into a Geophysical Year – the IGY), and 2007–2008. Although topics and technologies have changed over 125 years, the challenges and difficulties of working in polar regions have not, and the most recent IPY still adheres strongly to the original concept that substantial advances in polar science require, periodically, extraordinary levels of enhanced and focused international cooperation and coordination. With news of sea ice retreat, ice sheet collapse and threats to polar species almost constantly in media headlines, the scientific need for this IPY became obvious.

Scope

The IPY Framework Document (Rapley et al. 2004) specified four goals:

- Make major advances in polar knowledge and understanding.
- Leave a legacy of new or enhanced observational systems, facilities, and infrastructure.
- Inspire a new generation of polar scientists and engineers.
- Elicit keen interest and participation from polar residents, schoolchildren, the general public, and decisionmakers worldwide.

IPY emerged as one of the most ambitious international research efforts in the past 50 years. It engaged the intellectual resources of thousands of scientists – many more than expected and often from nonpolar countries – representing an unprecedented breadth of specialties, from geophysical to biological to social sciences. IPY stimulated international and interdisciplinary collaboration with over 160 endorsed science projects assembled from the ideas of researchers in more than 60 countries. IPY stimulated substantial new funding for polar science – more than US\$ 400 million, which coordinated with and supplemented ongoing polar research and monitoring programs.

Achievements

Major Advances in Polar Knowledge and Understanding

Major IPY-stimulated advances in polar science will emerge in the time period 2009–2015 and beyond, as IPY researchers explore, analyze, evaluate, and share their data sets. Already, we can see increased and more pervasive evidence of change in ice sheets and sea ice; we can see and foresee alteration, migration, and depletion of polar ecosystems; we can see hints of large changes in crucial polar components (sinks and sources) of global carbon cycles; and we can see the early impacts of relentless environmental, economic, and social changes on polar communities. We can predict with confidence, as a major advance in understanding, the increasing recognition of the global importance and impact of polar change.

Infrastructure Legacy

This IPY has done a superior job of observing both polar oceans, through multiple ship voyages and extensive use of buoys, drifters, and autonomous gliders. It has produced, from spaceborne radars, lidars, and gravity sensors, unprecedented coverage and views of the polar ice masses. Four separate multinational teams made difficult traverses across Antarctica, covering some regions for the first time since IGY. Field stations around the Arctic upgraded their capabilities, established new networks, and supported twice and in some cases thrice as many researchers as in previous years.

A New Generation

A new Association of Polar Early Career Scientists (<http://www.apecs.is/>) has brought vibrant and critical energy and ideas to this IPY and to polar science. Now incorporated, internationally endorsed and engaged in all aspects of polar research, and with their own funded Directorate, APECS represents one of IPY's most significant accomplishments. With more than 1,400 members from more than 40 countries, with the ability to do online posters and career development workshops almost anywhere in the world, APECS has clearly succeeded in shaping the future of polar research.

Excite and Stimulate the Public

Through international collaboration and cooperation, by careful cultivation of a global community of enthusiastic volunteers, and by creative use of free technologies, this IPY stimulated active and engaged networks of hundreds of teachers, media officers, journalists, and students. Working with those partners and networks, quarterly Polar Days were organized that engaged more than 500 individual and institutional partners from 50 countries in polar activities. Teacher's workshops, summer schools, polar science weekends, and student expeditions in both hemispheres were conducted. A Polar Resource Book documenting these and many other IPY activities was published. The polar community was

expanded by leaving a legacy of polar information in many languages. This IPY has demonstrated that an effective, energetic, and enthusiastic team working full-time to promote science communication serves as the key partner in a science program's success.

Challenges Ahead

The variety, breadth, and volume of IPY data highlight the fact that international science data infrastructure, largely a legacy of IGY, has for the most part not kept pace with the modern expectations and needs for data access and data preservation. A few national and disciplinary data systems demonstrate admirable capabilities, and new approaches, focused on open access and an information commons approach, such as the Polar Information Commons (<http://www.polarcommons.org/>), may stimulate a data transformation. Identifying and sustaining the essential elements of long-term observation and monitoring systems for polar regions, as those regions undergo rapid change, will represent a severe challenge, one without obvious technical or resource solutions. A more urgent focus on integrated prediction skills and systems will provide improved structure and rationale for maintaining polar observational capabilities. The many science specialties in this IPY, and the wide variety of data, will make the task of synthesis extremely difficult, but they will also stimulate innovative integration processes and result in products of enormous impact and value.

Summary

Often the science community seeks to organize prominent “above and beyond” events, events that transcend business as usual and accelerate scientific processes. Presented as “Experiments” or “Observation Periods” or “Years,” these events share ambitious goals of enhanced research, increased resources, heightened attention and, not coincidentally, re-invigorated recruitment of new researchers. The IPY 2007–2008 succeeded in stimulating international and interdisciplinary cooperation. Through widespread international cooperation and extraordinary enthusiasm, it achieved (in large part) its research, resource, recruitment, and public attention goals, and will produce a positive legacy for participants, for science in general, and for the public.

Cross-References

- [Electronic Geophysical Year](#)
- [International Geophysical Year](#)
- [International Year of Planet Earth](#)

Bibliography

- Rapley C, Bell R, ICSU IPY 2007–2008 Planning Group (2004) A framework for the International Polar Year 2007–2008. ICSU, Paris

International Year of Planet Earth

Eduardo F. J. de Mulder¹ and Wolfgang Eder²

¹Earth Science Matters Foundation,
Haarlem, The Netherlands

²GeoCentre, Geobiology, University of Göttingen,
Göttingen, Germany

Definition

The “International Year of Planet Earth” (IYPE) was a joint initiative by the International Union of Geological Sciences (IUGS) and UNESCO and ran from 2007 to 2009.

Background and Activities

When, in the year 2000, the International Union of Geological Sciences (IUGS) tried to match current scientific knowledge about Planet Earth against the application of such knowledge in daily life, it found a major discrepancy. The Indian Ocean tsunami in 2004 and hurricane Katrina in 2005 only confirmed the existence of that gap; knowledge about natural hazards existed, but it was not applied – or not applied adequately – to protect vulnerable societies from devastation.

Another observation was that fewer and fewer future experts in geoscience had been recruited at universities since the early 1980s, despite the fact that society needs increasing numbers of them to help reduce the impact of such disasters and to assist governments in subsurface management issues. Today, industry needs more geo-experts to find the new earth resources for which the world is desperately searching. Those experts are the same ones that find innovative solutions to reduce the environmental impacts of such necessary resource exploitation.

These observations motivated us to explore how the public and the politicians could be urged to reverse these trends. We chose an international, rather than national, approach and began the long road toward the proclamation of an International Year of Planet Earth (IYPE) under the UN system. IYPE aimed to capture people's imagination with the exciting knowledge about our planet and to see that knowledge used more effectively to make the Earth a safer, healthier, and wealthier place for our children and

grandchildren. That, we believe, can be achieved by building on the knowledge accumulated by the world's 400,000 Earth scientists and is expressed by the IYPE subtitle: *Earth Science for Society*.

Two earlier events served as models. The *International Geophysical Year* (1957–1958) uncovered many facts about our planet, which was then seen for the first time from space. IGY attracted lots of public attention, and many youngsters were inspired to begin their professional careers in Earth sciences. More recently, in 2002, the Germans celebrated their “Jahr der Geowissenschaften” with many public events and with an attractive logo that was later donated to the IYPE. That public exposure resulted in a much higher enrollment of German students in the Earth sciences in subsequent years (Fig. 1). That information fuelled the outreach program of the International Year.

Soon after the inception of IYPE by the International Union of Geological Sciences (IUGS), UNESCO's (former) Earth Science Division joined in, followed by 25 Associate Partners and 11 Founding Partners. To achieve its aims and ambitions, and to collect public and political support for the IYPE, we then approached the United Nations system. That process went through various stages, beginning in UNESCO where the United Republic of Tanzania successfully launched the initiative in UNESCO's Executive Board in April 2005, followed by the General Conference in October. In December 2005, the General Assembly of the United Nations adopted Resolution 192 proclaiming 2008 to be the *International Year of Planet Earth* (United Nations General Assembly 2005).

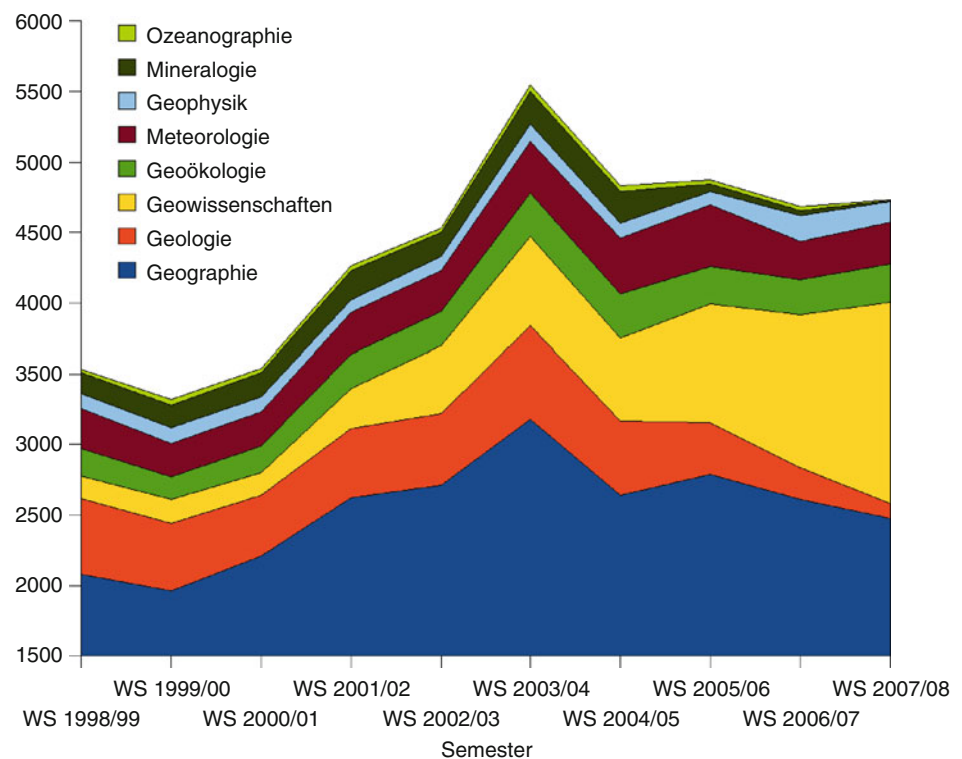
Shortly afterward, the IYPE Corporation was registered as a not-for-profit 501 (c) (3) Corporation in the USA, consisting of a Board of Directors and a Secretariat based at the Norwegian Geological Survey in Trondheim. Although the International Year was proclaimed by the UN for 2008, its initiators expanded that into a Triennium, starting in early 2007 and closing in June 2010. A Triennium was considered the minimum needed to implement at least part of our ambitious plan.

After the UN proclamation, additional support for the IYPE grew rapidly. By December 2009, 36 International Partners were providing financial and other support. Simultaneously, geoscientific communities in individual countries began to establish IYPE National Committees (Fig. 2) to a total number of 80 National and Regional Committees by the end of the IYPE Triennium. These countries are widely distributed across the globe.

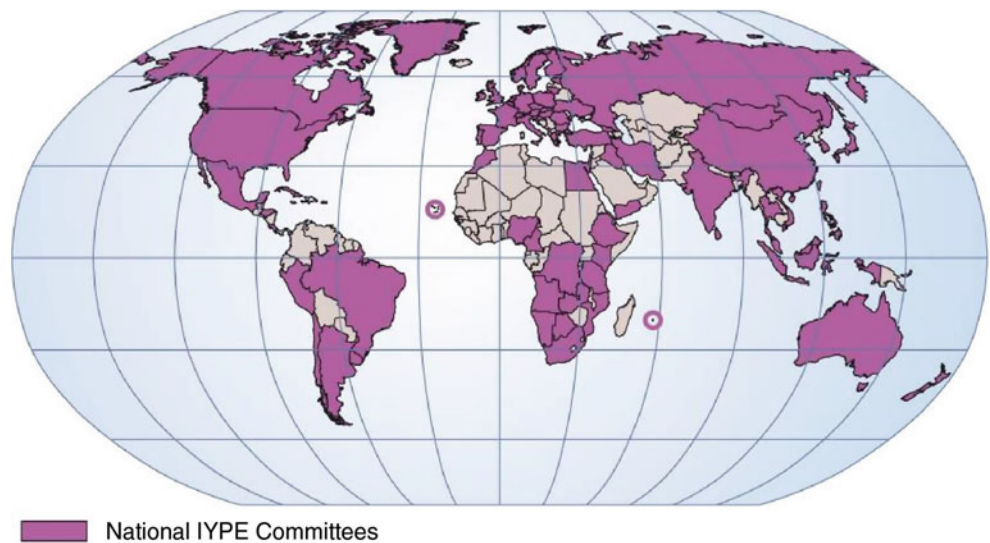
The main activities of the IYPE have been coordinated through Science and Outreach Programs. Both programs essentially operated in a “bottom-up” mode. The Science Program consisted of 10 broad, societal relevant, and multi-disciplinary themes: health, climate, groundwater, ocean, soils, deep Earth, megacities, resources, hazards, and life. Brochures on each of these themes are made available in hard copy in several languages or could be downloaded from IYPE’s Web site: www.yearofplanetearth.org.

The Outreach Program was mainly implemented at national levels. A global launch event was organized at UNESCO headquarters in Paris in February 2008. In addition,

International Year of Planet Earth, Fig. 1 Student enrolment in the Earth sciences in German universities between 1999 and 2007. Note the spike following the successful “Jahr der Geowissenschaften” (2002)



International Year of Planet Earth, Fig. 2 Countries with an established national committee for IYPE (in dark purple). One regional committee for IYPE had been established for East and Southeast Asia



continent-wide launch events took place for Africa (in Arusha, Tanzania) and for Latin America (in Brasilia, Brazil). Finally, a closing event was organized in Lisbon, Portugal, in November 2009. On an international level, brochures and flyers were produced and a global IYPE web-portal developed. Moreover, numerous activities in many countries with IYPE National Committees were monitored and registered.

Education has been an essential element in the International Year of Planet Earth. Many of the national and international activities involved students and focused on participation by secondary and primary school pupils. An example of student involvement was the IYPE global launch event in Paris. In preparation for that event, 18- to 22-year-old students from around the world submitted essays and poems each dealing with one of the International Year's themes. National Committees and the IYPE Secretariat evaluated these and selected the best. Through national and international sponsorship, some 200 winning students from almost every corner of the planet were awarded a free trip to Paris to participate actively in the launch event, including interacting with VIPs on pressing issues concerning Planet Earth. A comparable approach was taken for the Lisbon event, in 2009.

Legacy

The 80 National and Regional IYPE Committees belonged to the most important legacy items of the IYPE. At national levels, these Committees united key players of several, sometimes competing organizations into one joint ambition to raise awareness for the Earth sciences among decision makers and the public. An evaluation among the National Committees proved that this ambition was at least partly accomplished.

Since the onset of the IYPE Triennium in 2007, many thousands of activities all over the globe have taken place. A significant majority of these were related to education. Thousands of schools worldwide paid attention to the International Year of Planet Earth, each in their own languages and cultures. Probably such awareness raising activities under the UN-banner of IYPE have contributed to a strongly increased, more general concern about Planet Earth that we observe today.

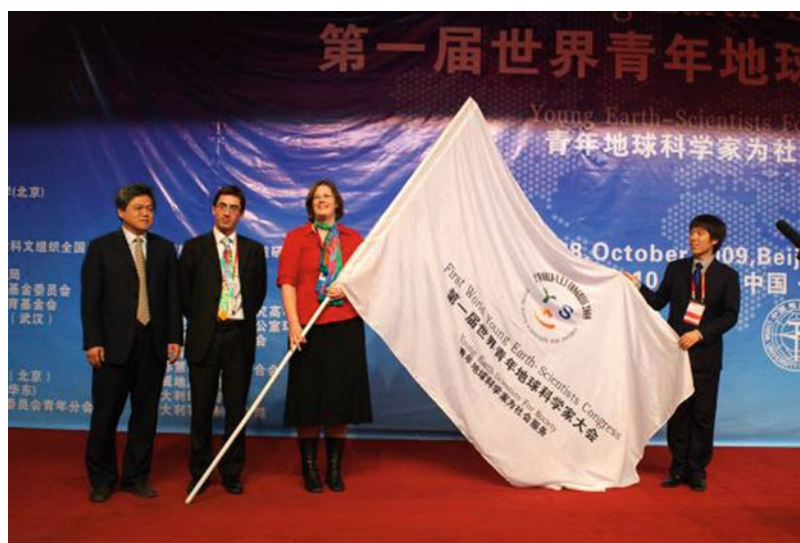
Another significant IYPE legacy item was the creation of a Young Earth Science Initiative (YES). This provides a platform for young professionals in the Earth sciences that found its origin at the gathering of young Earth scientists at the Global Launch Event of the IYPE in Paris (2008). From there the YES Initiative expanded, leading to a formal structure, a network of supporting organizations (including IYPE) and a series of international YES Conferences since. By today (2019) YES has some 5000, electronically connected, members (Fig. 3).

At a political level, a flood of public statements were given by leading politicians pleading for implementation of the aims and ambitions of the International Year of Planet Earth: to base decision making more on science and knowledge of the Earth for safer, healthier, and more prosperous societies. National launch events were the most common model for such involvements of decision-makers. There, politicians spoke in support of the IYPE. Such events were often coupled with cultural performances by school kids highlighting the importance of the Earth and sustainable development for future generations.

OneGeology is another major IYPE legacy item. That initiative was positioned under the IYPE banner in 2007. The ambition of that project was to bring together geological data of all nations into a digital database and translate these

International Year of Planet

Earth, Fig. 3 The first international conference for Young Earth Scientists (YES) was held in Beijing, China, in October 2009



into one single computer language. Eventually, that should result into a digital map of the subsurface of the entire world and may eventually provide a third dimension to popular geographical websites, as, e.g., Google Earth. Some 120 national Geological Surveys embarked on this initiative.

Many more legacy items developed during the IYPE Triennium. For example, cross-country geological hiking tours through the Alps: ViaGeoAlpina; an elegant method to assist teachers with simple materials to explain often complex geological processes: EarthLearningIdea; the production of a series of state-of-the art scientific books on each of the ten IYPE Themes, by Springer Verlag; and an UNESCO initiative to measure the conditions of Earth sciences education in Africa and to develop tools to improve these.

Student entries in the Earth sciences have been growing in a considerable number of countries since 2007. Although not all such positive developments may be attributed to the IYPE alone, the International Year of Planet Earth must have contributed at least to some extent to such growth, as that was often positively correlated with high levels of IYPE activities in such nations.

Finally, a Report on the entire IYPE operation had been produced for the United Nations and the IYPE Initiators (IUGS and UNESCO) as a legacy item in 2010.

Perspectives

Spurred by the National and Regional IYPE Committees who considered the IYPE as an excellent opportunity to raise public awareness to use knowledge of the Earth for better decision making and to encourage the youth to go for a professional career in the Earth sciences, and supported by the IYPE Board of Directors, some new initiatives were discussed since 2009. Two major new initiatives developed

with the specific intention to build on the IYPE legacy are described here.

The Earth Science Matters Foundation

The first initiative and offspring from the International Year of Planet Earth has a strong focus on outreach, the most successful and best exposed issue of the IYPE. Originally started in 2010 as the Planet Earth Institute (that was – for legal reasons – never materialized), the Initiative developed under a new name: The Earth Science Matters (ESM) Foundation in 2011. As in the IYPE, both authors of this article have been leading the ESM jointly from the onset. A Board of representatives of supporting organizations (International Partners) was created under the leadership of the second author while the first author operated as the ESM Director. In addition National Focal Points were identified in some 20 nations worldwide. The subtitle of the Earth Science Matters Foundation is “Bringing knowledge of the Earth to Everyone” and several projects dealing with Earth science outreach have been completed since, some of these together with the YES Network. The current ESM Director, Courtney Jermyn, was appointed in the ESM Board meeting in 2016.

GEOPARKS as a Tool

Earth science communication can be difficult, with explanations of complex geodynamic underground or surface processes and the use of scientific terminology often confusing those we want to take with us on a journey to understand and respect our restless planet. While science communication in general can be demanding, it is a prerequisite to convince people, decision-makers, the public at large, by providing up-to-date scientific knowledge on how to maintain and develop our environment in a sustainable way and sometimes to reduce mistrust towards science experts, as well as partly a bias towards industry and tourism (Eder 2019). A wonderful tool



EarthScienceMatters

Bringing knowledge of the Earth to everyone

to approach such challenging goals with respect to Earth Sciences are “Geoparks,” that – as a new international “brand” – are best defined through the criteria and guidelines of the “*UNESCO Global Geoparks*” designation that forms since November 2015 one part of UNESCO’s *International Geoscience and Geoparks Programme (IGGP)*. But even after more than 15 years having established Geoparks around the globe (the “Global Geoparks Network,” GGN, is existing since 2004; the “European Geoparks Network,” EGN, since 2000), it looks like that the ‘old’ prejudice “Geoparks are too popular and unscientific” is still alive. Still many geoscientists do not know sufficiently about Geoparks or do not perceive these as attractive platforms for research, conferences and outreach activities, but – vice versa – many geoscientists involved in the daily work within Geoparks are not familiar with big international scientific undertakings, programs, and conferences.

Global Geoscience Initiative

Another offspring of the IYPE aimed at identifying or designing “world-class” projects for a “Global Geoscience Initiative.” This was explored by an ad hoc group through a brainstorm meeting in July 2009, and the initiative was then presented to Earth scientific audiences at several major geoscientific gatherings around the world. Several topics for such a primarily science-based initiative were submitted and discussed by a broad platform of geoscientists and one topic that was considered both urgent and feasible for implementation emerged: availability and delivery of natural resources for future generations. This initiative was taken up by the International Union of Geological Sciences (IUGS) in 2013, who launched the Resources Future Generations Initiative with the aim of raising awareness of the challenges of sustainable resources supplies (Nickless 2017). Under the auspices of IUGS and organized by Canadian partners a major international Conference (RFG2018) was held on this issue, in June 2018 in Vancouver.

Summary

The International Year of Planet Earth has been particularly successful in its outreach and education programs. As fundraising started to bear fruit quite late in the Triennium, not sufficient funds were available to properly support a (costly) science program. Moreover, several of the 10 science

themes of the IYPE were addressed at national levels through current university curricula and special programs. UN proclamation has been crucial for the success of the IYPE, in particular for national implementation of the aims and ambitions through the 80 National and Regional Committees. These have also been the drivers for developing a new outreach-oriented initiative, the “Earth Science Matters” Foundation.

Cross-References

- [Electronic Geophysical Year](#)
- [International Geophysical Year](#)
- [International Polar Year 2007–2008](#)

Bibliography

- de Mulder EFJ et al (2008) The international year of planet earth: why and how? *Nature* 451(7176):305–305
- Eder W (2019) Foreword I, pp.v–vii; in: Lin, J.C & Su, S.J. *Geoparks of Taiwan*, Springer, ‘Geoheritage, Geoparks and Geotourism’ series (eds: Eder, W., Bobrowsky, P.T. & Martinez-Frias, J)
- Nickless E (2017) Resources future generations – a contribution by the Earth science community. *Nat Resour Res* 1–16. <https://doi.org/10.1007/s 11053-017-9331-x>.
- United Nations General Assembly (2005) The international year of planet earth, 2008 (A/RES/ 60/192), New York
- Woodfork L, de Mulder EFJ (2011) *International Year of Planet Earth, final report*. Earth Science Matters Foundation. Publication by IUGS and UNESCO

Inverse Theory, Artificial Neural Networks

William A. Sandham¹ and David J. Hamilton²

¹Scotsig, Glasgow, UK

²Consultraining, Edinburgh, UK

Definition

Geophysical inversion refers to mathematical and statistical techniques for recovering information on subsurface physical properties (magnetic susceptibility, density, electrical

conductivity, etc.) from observed geophysical data and involves the estimation of quantitative models of the Earth.

Artificial neural networks constitute a powerful paradigm which can emulate complex input/output mappings very efficiently.

Introduction

Neural networks have been recognized for many years as a powerful tool for solving complex problems associated with pattern recognition, classification, and generalization and have been used in geophysics in a large variety of different applications. The power of neural networks arises from their ability to emulate complex input/output mappings efficiently.

Comprehensive reviews may be found of all the major geophysical applications of neural networks in a number of excellent tutorial papers (McCormack 1991a, b; Cary and Upham 1992; Fogelman Soulié 1992; Sandham and Leggett 1998; Van der Baan and Jutten 2000; Poulton 2002; Peters et al. 2019). A number of textbooks on the subject have also been published in the last couple of decades (Poulton 2001; Nikravesh et al. 2003; Sandham and Leggett 2003).

In this present entry, the fundamentals of geophysical inversion using neural networks are reviewed. In section “[Geophysical Inversion](#)”, the physics of geophysical inversion are described, together with a short mathematical formulation. In section “[Artificial Neural Networks \(ANNs\)](#)”, the basics of artificial neural networks are discussed. This includes a review of the major types and the training strategies involved and a discussion of issues such as data preprocessing and performance. In particular, the important case of the multilayer perceptron trained with the backpropagation algorithm is considered in some detail. In section “[Geophysical Inversion Using Artificial Neural Networks](#)”, geophysical inversion applications of neural networks are reviewed, including computation of velocity models, split shear wave analysis, impedance inversion, pseudolog prediction, petrophysics, well logging, and potential field inversion.

Geophysical Inversion

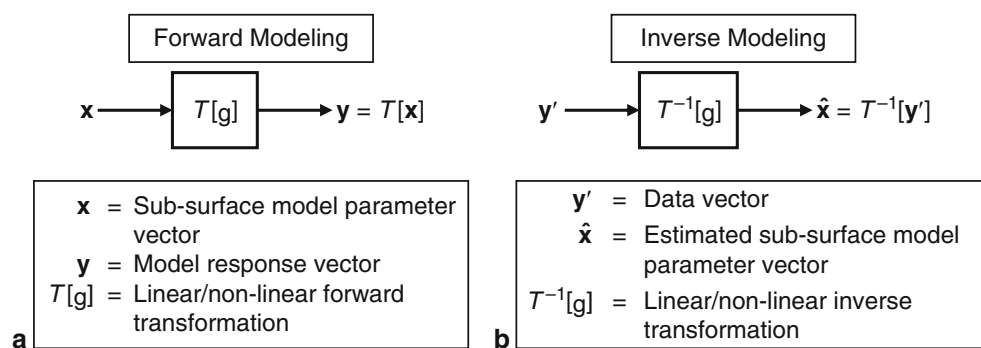
Geophysical inversion, or inverse modeling, may be defined as the estimation of quantitative subsurface Earth parameters using measured (observed) geophysical data (Backus and Gilbert 1967, 1968, 1970; Worthington 1987; Lines 1988; Russell 1988, 2019; Bednar et al. 1992; Parker 1994; Treitel and Lines 2001; Gubbins 2004; Ulrych and Sacchi 2005; Russell and Hampson 2006; Jones 2007; Kim and Nakata 2018; Zheng et al. 2019).

Generic definitions of forward and inverse modeling are given in Fig. 1. Mathematically, forward modeling may be described as the transformation $\mathbf{y} = T[\mathbf{x}]$, where $\mathbf{y}$ is the model response, $\mathbf{x}$ is a vector containing the set of subsurface model parameters, and $T[\cdot]$ is some transformation (linear or nonlinear) which describes an observed physical process. For example, in the seismic case, $T[\cdot]$ produces a model response in the form of a synthetic seismogram. Inverse modeling may also be written as $\hat{\mathbf{x}} = T^{-1}[\mathbf{y}']$, where $\hat{\mathbf{x}}$ is now a vector containing the set of estimated subsurface model parameters (the model space), derived from the data vector $\mathbf{y}'$ (the data space). The operator $T^{-1}[\cdot]$ then denotes the inverse transformation from data space to model space.

Unfortunately, analytical expressions for the inverse transformation seldom exist. Numerically, the inverse problem is solved using methods from mathematical optimization theory. A popular iterative solution is to assume an initial guess for a model, compute the model response $\mathbf{y}$, compare it with the observed response $\mathbf{y}'$, and then perturb the model with the hope of achieving a better fit. However, since the underlying physics in most cases is nonlinear, a straightforward inversion is, even under ideal conditions, inherently unstable. One possible solution is to try and linearize the inverse problem. Unfortunately, such methods are notoriously dependent on the choice of the initial guess for the model. Furthermore, the presence of noise in the data makes interpretation of the results additionally difficult and circumspect. A possible alternative is to use stochastic methods like Monte Carlo (MC), simulated annealing (SA), or genetic algorithms

Inverse Theory, Artificial Neural Networks, Fig. 1 (a)

Forward modeling illustrating the derivation of a model response vector $\mathbf{y}$, from the forward transformation $T[\cdot]$ of a subsurface model parameter vector $\mathbf{x}$. (b) Inverse modeling illustrating the derivation of an estimated subsurface model parameter vector $\hat{\mathbf{x}}$ from the inverse transformation $T^{-1}[\cdot]$ of a data vector $\mathbf{y}'$



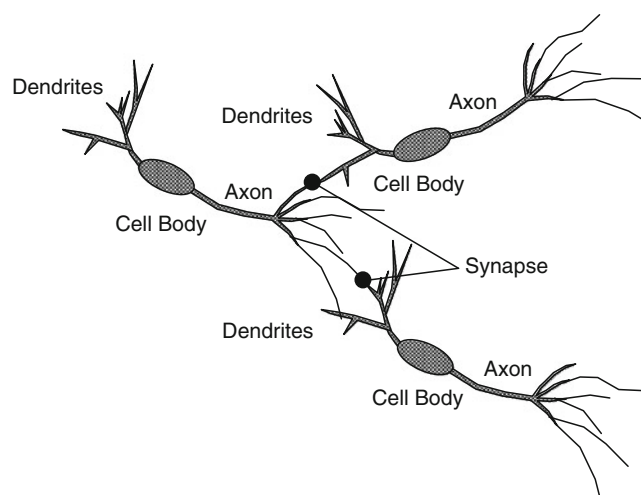
(GA), all of which provide some insight into the statistical nature of the inverse relationships. Unfortunately, MC is extremely inefficient and computing-intensive in multi-dimensional cases, and both SA and GA need subjective tuning parameters that determine the nature of the stochastic search. Most techniques also involve the assimilation of a priori knowledge during the inversion process, to constrain the result. All these techniques – linear and nonlinear – involve very expensive forward calculations and are necessarily specific to the particular data set used. Also, the solution is generally nonunique.

It is precisely these intractable features of geophysical inversion which has encouraged geophysicists to investigate solutions using artificial neural networks (ANNs).

Artificial Neural Networks (ANNs)

The expeditious pattern recognition and classification abilities of humans are a direct consequence of the human brain having a very large number of highly interconnected biological nerve cells or neurons. The basic anatomy of an individual neuron is shown in Fig. 2.

It consists of a cell body or soma, the output of which is a long, branching fiber called the axon which, on appropriate triggering by the cell, transmits an impulse. A human neuron “fires” when its electrical potential reaches a threshold. The output signal is transmitted by the axon and distributed to other neurons via connecting points termed synapses, located in the tree structure at the end of the axon, which modify the strength of the output signal. The input area of the cell is a set of branching fibers called dendrites. Input signals to the soma are attenuated with an increasing distance from the synapses



Inverse Theory, Artificial Neural Networks, Fig. 2 Biological neural network showing three neurons, each of which consists of a cell body or soma, an output or axon, and a series of inputs or dendrites, with the connecting point between an axon and a dendrite termed the synapse

to the soma. Since the human brain is a very dense neural network with a highly interconnected set of approximately 10^{11} neurons, the axon of each neuron is connected to around 10,000 other cells, resulting in approximately 10^{14} connections or synapses. It is this large number of interconnections, which is the basis for the human brain's immense computational power.

As the name implies, an artificial neural network (ANN), or simply neural network (NN), is a mathematical or computational model that tries to emulate the structure and/or functional aspects of a biological neural network, and as such, represents a technology that is rooted in many disciplines, including neurosciences, mathematics, statistics, physics, computer science, and engineering. ANNs consist of an interconnected group of artificial neurons (programming constructs that mimic the properties of biological neurons), which process information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during a learning phase, realized via suitable training iterations. ANNs may either be used to gain an understanding of biological neural networks or for solving artificial intelligence problems without necessarily creating a model of a real biological system. In the latter case, they may be defined as nonlinear statistical data modeling tools, and can be used to model complex relationships between inputs and outputs or to find patterns in data. Indeed, the power of neural networks arises from their ability to emulate complex input/output mappings efficiently.

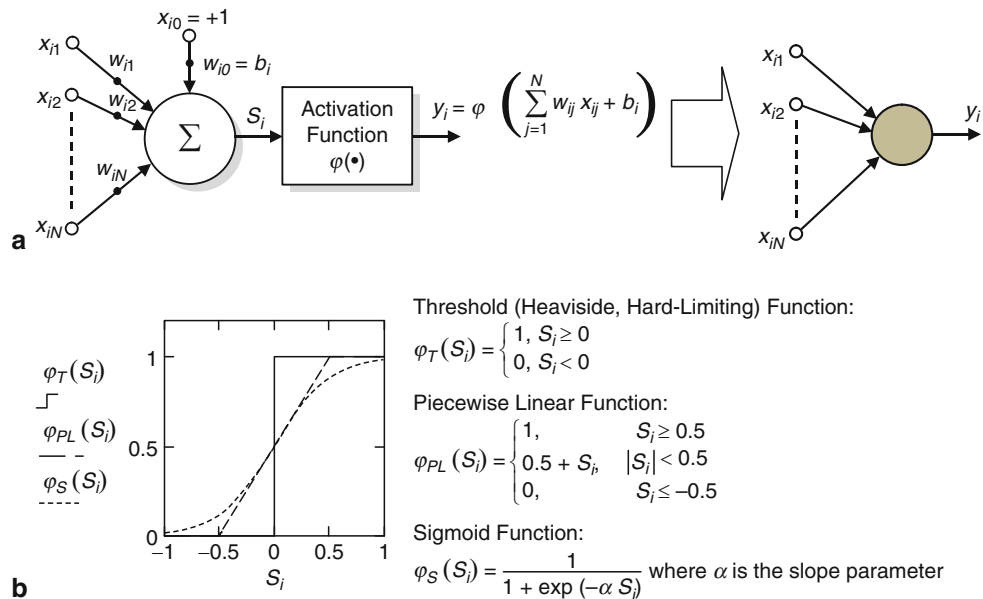
There are now some excellent general reviews of ANNs in the literature (Lippmann 1987, 1989; Eberhart and Dobbins 1990; Hush and Horne 1993; Zhang 2000, 2007; Meireles et al. 2003; Wilamowski 2009). A substantial number of books on the subject have also been published over the last number of decades (Aleksander and Morton 1990; Dayhoff 1990; Hertz et al. 1991; McCord Nelson and Illingworth 1991; Nelson and Illingworth 1991; Kosko 1992; Kung 1993; Freeman 1994; Haykin 1994, 1999, 2009; Bishop 1995; Masters 1995; Kasabov 1996; Priddy and Keller 2005). The texts by Haykin (including earlier editions) are probably the most comprehensive and definitive.

The basic building block of an ANN is the (mathematical) perceptron, which was conceived almost 70 years ago by McCulloch and Pitts to mimic the behavior of a biological neuron. The perceptron is also known as an artificial neuron, neurode, processing element/unit (PE/PU), or simply unit. A schematic of a perceptron is shown in Fig. 3a.

This depicts a number of inputs x_{ij} , analogous to the dendrites in a biological neuron, which are modified by corresponding weights w_{ij} , analogous to the synapses. The $w_{ij}x_{ij}$ products are then summed, along with bias b_i , and the result passed through a nonlinear activation function $\varphi(\bullet)$ to produce the final output y_i , analogous to the axon. The

Inverse Theory, Artificial Neural Networks,

Fig. 3 (a) Schematic of an artificial neuron, which modifies inputs x_{ij} by corresponding weights w_{ij} , sums the products along with bias b_i , and passes the result through a nonlinear activation function $\varphi(\cdot)$ to produce the final output y_i . The diagram on the right is a concise way of depicting the neuron. (b) Common activation functions used in artificial neural networks. The range of activation functions is usually $0 \leq \varphi(S_i) \leq 1$, as defined here; however, it is sometimes desirable for the range to be $-1 \leq \varphi(S_i) \leq 1$



activation function transforms the weighted/summed input S_i to a neuron into its activation. Some major types of activation function are shown in Fig. 3b. Historically, the threshold activation function was used. However, this particular function gives only a binary output (i.e., 1 or 0). Also, this particular function is not continuously differentiable, and so the optimum weights were very difficult to estimate. A more popular activation function is the sigmoid, which is continuously differentiable and monotonically increasing.

An ANN is formed by connecting (usually many) PEs together in some fashion. Some of the major ANN paradigms are depicted in Fig. 4.

Once an ANN has been chosen and the number of layers and neurons has been estimated, the various network weights w_{ij} have to be determined, through a network training procedure. Training can either be supervised or unsupervised (see Fig. 5a, b).

Supervised training involves presenting target answers for each input pattern, whereas with unsupervised training, the network adjusts its weights in response to input patterns without target answers, effectively classifying the input patterns into similarity categories. After training has been carried out, network performance is assessed using input patterns which have not been used previously, i.e., as training data (see Fig. 5c). This will indicate whether the network has been overtrained and cannot generalize, and also whether the network has been trapped during training in a local minimum. Performance is dependent on the quality of data, pre-processing involved, training algorithm, network complexity, generalization, and implementation.

One of the most popular ANNs is the so-called multilayer perceptron or MLP (Fig. 4a). This is a feedforward network

consisting of a layer of input distribution nodes, a number of “hidden” layers of perceptrons, and an output layer. In a fully connected version, all distribution nodes and perceptrons are connected to perceptrons in the following layers. The network configuration should allow for an adequate description of the underlying statistical distribution of the spread in the data. Since the number of input nodes and output neurons is usually dictated by the problem, good network design relies on estimating an optimum number of hidden layers and numbers of neurons in each layer. These are usually determined through trial-and-error procedures. A supervised training algorithm called backpropagation (or back-error propagation), popularized in the mid-1980s by Rumelhart and McClelland, is usually used to train a MLP. This iterative, steepest descent algorithm is outlined below.

Step 1: All w_{ij} set to small random values.

Step 2: Input presented together with target outputs t_i .

Step 3: Actual outputs y_i computed.

Step 4: Errors computed for all nodes in a layer,

$$\text{Output layer: } \delta_i = (t_i - y_i)\varphi'(S_i) \text{ hidden layer: } \delta_i = (t_i - y_i)\varphi'(S_i)$$

Step 5: Update weights:

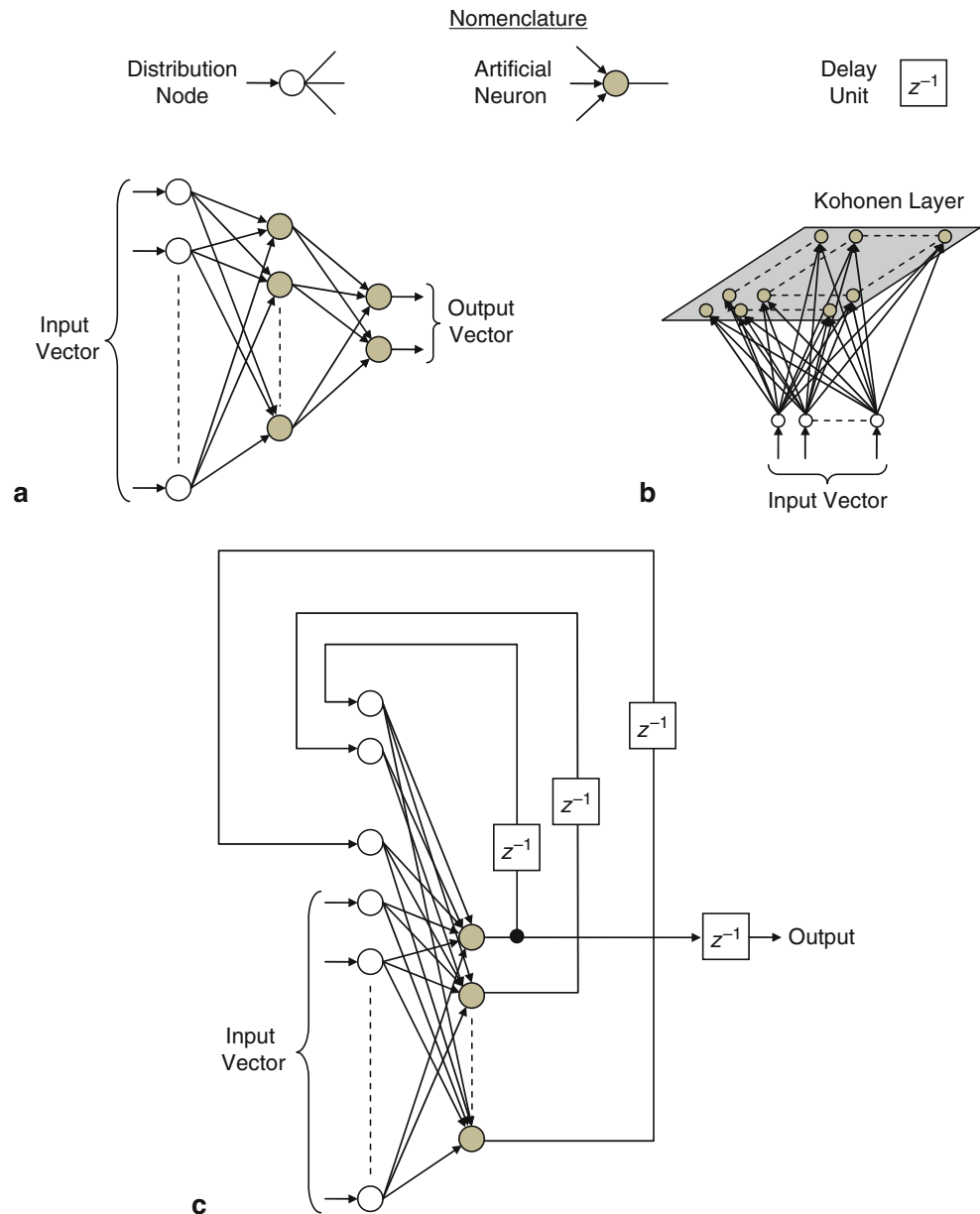
$$w_{ij}(t + 1) = w_{ij}(t) + \eta \delta_i y_j + \alpha [w_{ij}(t) - w_{ij}(t - 1)]$$

Step 6: Return to Step 4 or Step 2.

The weight update in Step 5 includes learning rate (η) and momentum (α) terms to improve the performance of the algorithm. The selection of a learning rate is of critical

Inverse Theory, Artificial Neural Networks, Fig. 4

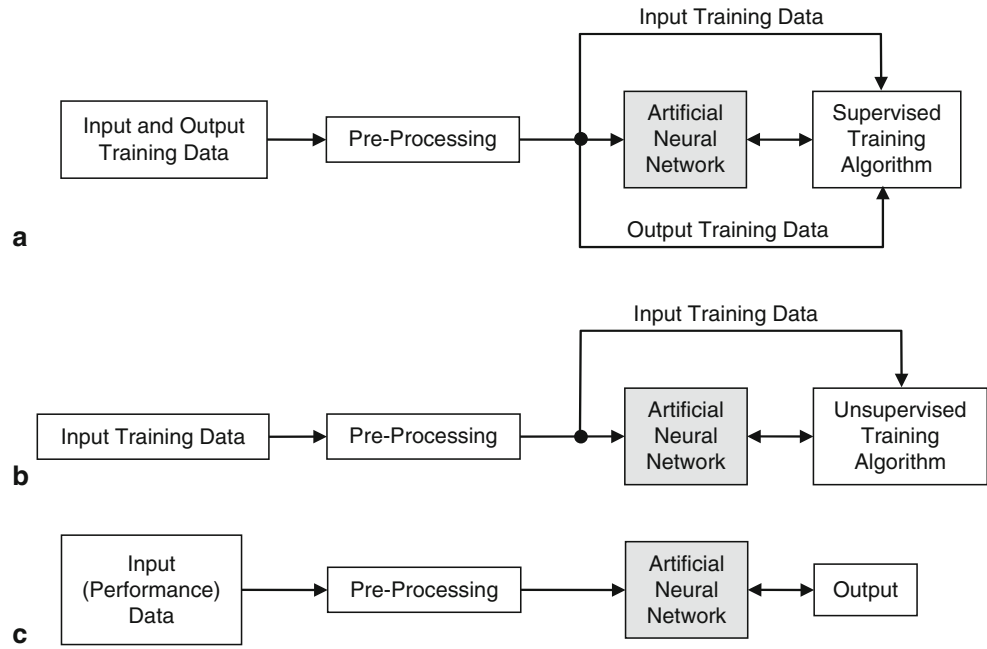
Some common neural network paradigms. (a) Fully connected multilayer perceptron (MLP) with one hidden layer and two output neurons. (b) Kohonen architecture trained using a self-organizing map (SOM) (unsupervised). (c) Fully connected recurrent network with one output neuron



importance in finding the true global minimum of the error distance. Backpropagation training with too small a learning rate will make agonizingly slow progress. Too large a learning rate will proceed much faster, but may simply produce oscillations between relatively poor solutions. Both of these conditions are generally detectable through experimentation and sampling of results after a fixed number of training epochs. Ideally, the largest learning rate should be used which still converges to the minimum solution. The use of a momentum term can be helpful in speeding algorithm convergence and avoiding local minima. This gives the system a certain amount of inertia since the weight vector will tend to continue moving in the same direction unless opposed by the gradient

term. The momentum term (i) smooths the weight changes and suppresses cross-stitching, i.e., cancels side-to-side oscillations across the error valley; (ii) amplifies the learning rate when all weight changes are all in the same direction, causing a faster convergence; and (iii) enables escape from small local minima on the error surface. Typical values for the learning rate and momentum parameter are $0 < \eta, \alpha < 1$. The momentum term will allow a larger learning rate and thus speed convergence and avoid local minima. On the other hand, a learning rate of 1 with no momentum will be much faster when no problem with local minima or non-convergence is encountered. Other popular training algorithms include quickprop and cascade correlation.

Inverse Theory, Artificial Neural Networks, Fig. 5 The two major network training strategies. (a) Supervised training. (b) Unsupervised training. (c) Post-training performance assessment



A very popular network which uses unsupervised training is Kohonen's self-organizing map or SOM (Fig. 4b). This is a two-layered feedforward network consisting of a layer of input distribution nodes and an output layer usually arranged as a 2-D map which shows the natural relationships among the patterns presented to the network. Essentially, neighboring neurons compete in their activation by means of mutual lateral interactions, and develop adaptively into specific detectors of different input patterns. The unsupervised training algorithm is outlined below.

Step 1: All w_{ij} set to small random values.

Step 2: Input presented.

Step 3: Euclidian distances δ_i between each input node (j) and output node (i), computed:

$$\delta_i = \sum [x_j(t) - w_{ij}(t)]^2$$

Step 4: Select output node i^* with minimum distance.

Step 5: Update weights to node i^* and neighbors:

$$w_{ij}(t + 1) = w_{ij}(t) + \eta(t) \times [x_j(t) - w_{ij}(t)], i \in NE_{i^*}(t)$$

Step 6: Return to Step 2.

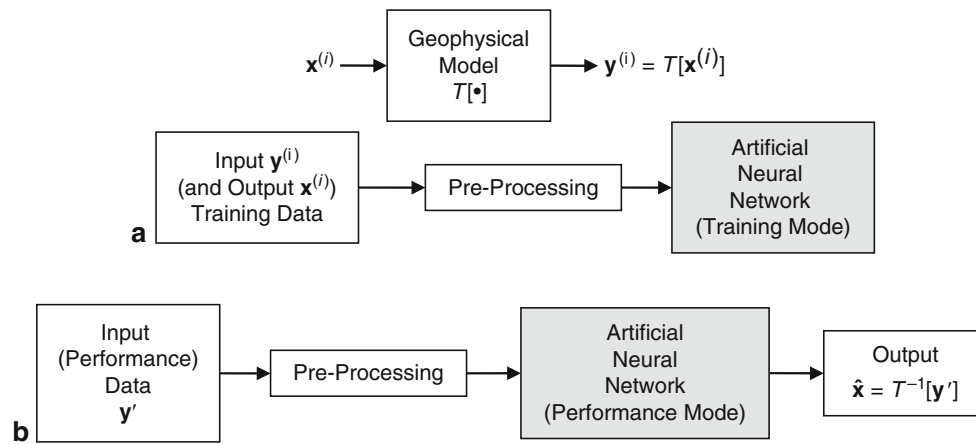
Recurrent networks, which involve the inclusion of feedback loops such as the one depicted in Fig. 4c, have a profound impact on the learning capability of the network and its performance.

Many other types of ANN exist, including Hopfield recurrent networks, those based on the adaptive resonance theory of Carpenter and Grossberg, probabilistic, modular and radial basis function ANNs, support vector machines, and more recently, networks including spiking ANNs and wavelet ANNs.

Geophysical Inversion Using Artificial Neural Networks

ANNs offer many advantages to geophysicists for inversion applications, and this has made them an intensely active research area. Indeed, many ANN inversion algorithms are now sufficiently mature and are routinely employed in their respective application area. A number of good reviews of geophysical inversion using ANNs are available in the literature (Raiche 1991; Röth and Tarantola 1991, 1992, 1994; Röth 1993; Roy-Chowdhury et al. 1997; Nunnari et al. 2001; Kappler et al. 2005; Maiti and Tiwari 2010; Russell 2019; Zheng et al. 2019).

The basis of ANN geophysical inversion methods is illustrated in Fig. 6. First, a geophysical model is used to derive many simulated model response vectors $\mathbf{y}^{(i)}$ from a large number N of subsurface model parameter vectors $\mathbf{x}^{(i)}$, where $0 \leq i \leq N$. Next, the response vectors act as input data training sets to an ANN. In supervised training the model parameter vectors are used as the target vectors. Finally, after appropriate training, a new survey data vector $\mathbf{y}'$ is input to the ANN and an estimated subsurface model parameter vector $\hat{\mathbf{x}}$ obtained,



Inverse Theory, Artificial Neural Networks, Fig. 6 The basis of geophysical inversion using ANNs. (a) A geophysical model is used to derive many simulated model response vectors $\mathbf{y}^{(i)}$ from subsurface model parameter vectors $\mathbf{x}^{(i)}$. The response vectors then act as input data training sets to an ANN. In supervised training the model parameter

vectors are also used. (b) After appropriate training, a new survey data vector $\mathbf{y}'$ is input to the ANN and an estimated subsurface model parameter vector $\hat{\mathbf{x}}$ obtained

i.e., geophysical inversion has been achieved. There are many variations of this basic technique.

It is important to emphasize that geophysical inversion using ANNs will only be successful if proper consideration and diligence are provided during the preprocessing and training phases. This is to ensure the prevention of over-training and local minima entrapment (Maiti and Tiwari 2010).

There are many examples in the literature of ANNs being employed across the complete spectrum of geophysical inversion applications, and these publications, many of which are cited below under their respective application domains, should be consulted for further information.

Seismic

General: (Michaels and Smith 1992, 1997; Michaels 1993; Zhang et al. 1995; Calderón-Macias 1997; Qu et al. 1998; Adriana and Gildardo 2004; Kelter 2005; Al-Khaled et al. 2006; Castro de Matos et al. 2007; Kononov et al. 2007; Dumitrescu 2009; Phan and Sen 2018, 2019; Biswas et al. 2019; Stuart et al. 2019).

Seismic Attributes: (Dumitrescu and Mayer 2006; Dumitrescu and Lines 2008, 2009; Dumitrescu 2009)

Porosity: (Wang et al. 1997; Ecoublet et al. 1998; Walls et al. 1999, 2000; Sun et al. 2001; Burnett et al. 2004; Pramanik et al. 2004; Gomez 2005; Iturrarán-Viveros and Spurlin 2005; Álvarez et al. 2009; Boateng et al. 2017).

Impedance: (Fu 1997, 2001, 2004; Fu et al. 1997; Cull 2006; Wang et al. 2019).

Facies Analysis: (West et al. 2001; Durrani and Baumel 2002; Aminzadeh and de Groot 2005; Gallop 2006; Castro de Matos et al. 2007; Álvarez et al. 2009; Maiti and Tiwari 2010; Dumitrescu et al. 2014; Ross 2017; Ross and Cole 2017)

Refraction Statics: (Taner et al. 1988).

Seismic Velocities: (Boadu 1998, 2000; Calderón-Macias et al. 1998).

Amplitude Variation with Offset (AVO): (Hampson and Todorov 1999; Sun et al. 2000; Mogensen and Link 2001; Chopra and Pruden 2003; Kuzma and Rector 2004; Royle et al. 2004)

Split Shear Wave: (Dai and MacBeth 1994)

Seismic Tomography: (Tselentis et al. 2007).

Well Logs: (Liu and Liu 1998; FitzGerald et al. 1999; Hampson et al. 2001; Lau et al. 2002; Dorrington and Link 2004; Aristodemou et al. 2005; Eftekharifar 2009; Huang et al. 2011).

Reservoir Geophysics: (Walls et al. 1999, 2000; Boadu 2001; Durrani and Baumel 2002; Garcia et al. 2002; Liu et al. 2002; Saggaf et al. 2003; Chopra et al. 2004; Devi and Cohen 2004; Aminzadeh and de Groot 2005; Bertrand et al. 2005; Schuelke et al. 2005; Aminzadeh and Brouwer 2006; Xiangjun et al. 2006; Modin and Breton 2007; Thakur et al. 2007; Veeken et al. 2009; Niu et al. 2015; Gogoi and Chatterjee 2018)

Electromagnetic: (Poulton et al. 1989; Winkler 1994; Zhang and Paulson 1997; Seiberl et al. 1998; El-Kaliouby et al. 1999; Spichak and Popova 2000; Ahl 2003; Daniels 2004; Anderson Kuzma and Rector 2005)

Resistivity: (Teles and Do Carmo 1999; Zhang et al. 1999, 2000a, b, 2002; Calderón-Macias et al. 2000; El-Qady and Ushijima 2001; Zhang and Zhou 2002; Singh et al. 2005; Sastry and Tesfakiros 2006; Neyamadpoura et al. 2009; Mayer et al. 2014).

Potential Field Inversion

General: (Fossati et al. 1992; Guo et al. 1992).

Gravity: (Taylor and Vasco 1990; Vasco and Taylor 1991; Cartabia et al. 1994; Kaftan and Şalk 2009; Xiang 2019)
Magnetic: (Al-Garni 2009)

Archaeology: (Bescoby et al. 2003, 2006)

Remote Sensing: (Tsang et al. 1992.

Summary

Geophysical inversion is an essential tool to help geophysicists solve practical environmental, engineering, or exploration problems. This is particularly relevant for situations which require detailed quantitative subsurface information and data. Traditional approaches to inversion are many and varied and include model perturbation/comparison and linearization usually of a nonlinear geophysical model, together with a variety of stochastic methods including Monte Carlo, simulated annealing, and genetic algorithms. These approaches do not perform well in the presence of noise, are very inefficient and computing-intensive, require subjective tuning parameters and a priori knowledge, and/or produce results which are nonunique.

Alternative geophysical inversion methods, based on artificial neural networks (ANNs), have been investigated over the past 15–20 years, to try and overcome the disadvantages of the traditional approaches. These ANN inversion methods have now found considerable success across many different geophysical domains, including seismic attributes, velocity fields, and tomography, to potential field inversion, archaeology, and remote sensing.

With the continuous introduction of improved ANN paradigms and training algorithms, coupled with the ever-increasing power of digital computing hardware [including field-programmable gate arrays (FPGAs)], there is no doubt that ANNs have a very positive future in the exciting and challenging world of geophysical inversion.

Cross-References

- Deep Seismic Reflection and Refraction Profiling
- Earth, Density Distribution
- Electrical Properties of Rocks
- Electrical Resistivity Surveys and Data Interpretation
- Geomagnetic Field, Measurement Techniques
- Geophysical Well Logging
- GPS, Data Acquisition, and Analysis
- Gravity Data, Advanced Processing
- Gravity Method, Airborne
- Gravity Method, Satellite
- Gravity Method, Surface
- Instrumentation, Electrical Resistivity

- Instrumentation, EM
- Inverse Theory, Linear
- Inverse Theory, Monte Carlo Method
- Inverse Theory, Singular Value Decomposition
- Magnetic Anisotropy
- Magnetic Methods, Airborne
- Magnetic Methods, Principles
- Magnetic Methods, Surface
- Magnetotelluric Interpretation
- Poroelectricity
- Propagation of Elastic Waves: Fundamentals
- Remote Sensing, Applications to Geophysics
- Sedimentary Basins
- Seismic Data Acquisition and Processing
- Seismic Velocity and Temperature Relationships
- Seismic Viscoelastic Attenuation
- Seismic, Velocity, and Density Relationships
- Seismic, Waveform Modeling and Tomography
- Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting
- Traveltime Tomography Using Controlled-Source Seismic Data
- Wavelet Analysis

Bibliography

- Adriana A-RR, Gildardo VV (2004) The impact of high technology to add value to existing data. In: SEG International Exposition and Annual Meeting, Denver, SEG
- Ahl A (2003) Automatic 1D inversion of multifrequency airborne electromagnetic data with artificial neural networks: discussion and a case study. *Geophys Prospect* 51(2):89–97
- Aleksander I, Morton H (1990) An introduction to neural computing. Chapman and Hall, London
- Al-Garni MA (2009) Interpretation of some magnetic bodies using neural networks inversion. *Arab J Geosci* 2(2):175–184
- Al-Khaled O, Al-Zuabi Y, Rahaman M, Al-Jenai J, Hussain T (2006) Seismic inversion of single sensor land 3D seismic survey: Kuwait, Magwa field, Wara reservoir. In: SEG International Exposition and Annual Meeting, New Orleans, SEG
- Álvarez P, Rangel J, Martínez M (2009) Mapping porosity distribution in a vuggy carbonate reservoir integrating seismic attributes with borehole image logs through a supervised facies analysis. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Aminzadeh F, Brouwer F (2006) Integrating neural networks and fuzzy logic for improved reservoir property prediction and prospect ranking. In: SEG International Exposition and Annual Meeting, New Orleans, SEG
- Aminzadeh F, de Groot P (2005) A neural networks based seismic object detection technique. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Anderson Kuzma H, Rector JW (2005) Support vector machines for EM data interpretation. Symposium on the Application of Geophysics to Engineering and Environmental Problems
- Aristodemou E, Pain C, de Oliveira C, Goddard T, Harris C (2005) Inversion of nuclear well-logging data using neural networks. *Geophys Prospect* 53(1):103–120

- Backus GE, Gilbert JF (1967) Numerical applications of a formalism for geophysical inverse problems. *Geophy J Roy Astron Soc* 13:247–276
- Backus GE, Gilbert JF (1968) The resolving power of gross earth data. *Geophy J Roy Astron Soc* 16:169–205
- Backus GE, Gilbert JF (1970) Uniqueness in the inversion of inaccurate gross earth data. *Phil Trans R Soc London* 266:123–192
- Bednar JB, Lines LR, Stolt RH, Weglein AB (eds) (1992) Geophysical inversion. Proceedings in Applied Mathematics. SIAM Workshop on Geophysical Inversion, Houston, Texas. SIAM, Philadelphia
- Bertrand A, McQuaid S, Bobolecki R, Leiknes S, Egil Ro H (2005) A high resolution workflow for 4D-friendly analysis: application to gas-oil contact monitoring at Troll West. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Bescoby DJ, Cawley GC, Chroston PN (2003) Interpretation of geophysical surveys of archaeological sites using artificial neural networks. International Joint Conference on Neural Networks
- Bescoby DJ, Cawley GC, Chroston PN (2006) Enhanced interpretation of magnetic survey data from archaeological sites using artificial neural networks. *Geophysics* 71(5):H45–H53
- Bishop C (1995) Neural networks for pattern recognition. Oxford University Press, Oxford
- Biswas R, Sen MK, Das V, Mukerji T (2019) Prestack and poststack inversion using a physics-guided convolutional neural network. *Interpretation* 7(3):SE161–SE174
- Boadu FK (1998) Inversion of fracture density from field seismic velocities using artificial neural networks. *Geophysics* 63(2):534–545
- Boadu FK (2000) Neural inversion for oil saturation from seismic velocities. In: SEG International Exposition and Annual Meeting, Calgary, SEG
- Boadu FK (2001) Petrophysical parameters affecting NMR relaxation times and cementation factor: artificial neural network analysis. *J Environ Eng Geophys* 6(3):107–114
- Boateng CD, Fu L-Y, Yu W, Xizhu G (2017) Porosity inversion by Caianiello neural networks with Levenberg-Marquardt optimization. *Interpretation* 5(3):SL33–SL42
- Burnett MD, Castagna JP, Camargo G, Chen H, Sanchez JJ, Santana A, Hernandez EM (2004) Synergistic porosity mapping in the upper cretaceous of the Chiapas region using spectral decomposition and neural network inversion. In: SEG International Exposition and Annual Meeting, Denver, SEG
- Calderón-Macias C (1997) Artificial neural systems for interpretation and inversion of seismic data. PhD, University of Texas at Austin
- Calderón-Macias C, Sen MK, Stoffa PL (1998) Automatic NMO correction and velocity estimation by a feedforward neural network. *Geophysics* 63(5):1696–1707
- Calderón-Macias C, Sen MK, Stoffa PL (2000) Artificial neural networks for parameter estimation in geophysics. *Geophys Prospect* 48:21–47
- Cartabia G, Zerilli A, Apolloni B (1994) Lineaments recognition for potential fields images using a learning algorithm for Boltzmann machines. In: SEG International Exposition and Annual Meeting, Los Angeles, Society of Exploration Geophysicists
- Cary P, Upham W (1992) An evaluation of neural networks. *Lead Edge* 11:45–47
- Castro de Matos M, Osorio PLM, Johann PRS (2007) Unsupervised seismic facies analysis using wavelet transform and self-organizing maps. *Geophysics* 72(1):P9–P21
- Chopra S, Pruden D (2003) Multiattribute seismic analysis on AVO-derived parameter – a case study. *Lead Edge* 22(10):998–1002
- Chopra S, Alexeev V, Pruden D (2004) Non-linearity in multi-attribute analysis – a comparative study. In: SEG International Exposition and Annual Meeting, Denver, SEG
- Cull P (2006) Caianiello and neural nets. In: *Imagination and rigor*. Springer, Milan, pp 47–61
- Dai H, MacBeth C (1994) Split shear-wave analysis using an artificial neural network. *First Break* 12:605–613
- Daniels DJ (ed) (2004) Ground penetrating radar. IEE radar, sonar, navigation and avionics series. The Institution of Electrical Engineers, London
- Dayhoff J (1990) Neural network architectures: an introduction. Van Nostrand Reinhold, New York
- Devi KRS, Cohen AJ (2004) Wavelet transforms and hybrid neural nets for improved pore fluid prediction and reservoir properties estimation. In: SEG International Exposition and Annual Meeting, Denver, SEG
- Dorrington KP, Link CA (2004) Genetic-algorithm/neural-network approach to seismic attribute selection for well-log prediction. *Geophysics* 69(1):212–221
- Dumitrescu CC (2009) Exploration of Lower Cretaceous sands in the Leland Area, Alberta, using seismically derived rock properties. *First Break* 27(10):53–60
- Dumitrescu CC, Lines L (2008) Seismic attributes used for reservoir simulation: application to a heavy oil reservoir in Canada. In: SEG International Exposition and Annual Meeting, Las Vegas, SEG
- Dumitrescu CC, Lines L (2009) Case study of a heavy oil reservoir interpretation using Vp/Vs ratio and other seismic attributes. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Dumitrescu CC, Mayer (2006) Case study of a gas reservoir (Leland) in the deep basin: from deterministic inversion to neural network analysis. In: CSEG Annual Conference and Exposition
- Dumitrescu CC, Vermeulen P, Gammie S, Li G (2014) Facies analysis using multicomponent seismic data in oil-sands reservoir: case study from Athabasca Oil Sands. In: SEG International Exposition and Annual Meeting, Denver, SEG
- Durrani JA, Baumel RT (2002) Reservoir characterization by Prestack inversion and neural net calibration. In: EAGE Annual Conference and Exposition, Florence, EAGE
- Eberhart RC, Dobbins RW (1990) Early neural network development history: the age of Camelot. *IEEE Eng Med Biol Mag* 9:15–18
- Ecoublet PO, Symes WW, Levin SA (1998) Porosity prediction from seismic inversion using artificial neural networks. In: EAGE Annual Conference and Exposition, EAGE
- Eftekharifar M (2009) Extrapolation of log properties by integrating fuzzy self-organizing maps and local linear modeling. In: SEG International Exposition and Annual Meeting, Houston, SEG
- El-Kaliouby HM, Poulton MM, El-Divany EA (1999) Inversion of coincident loop TEM data for layered polarizable ground using neural networks. In: SEG International Exposition and Annual Meeting, Houston, SEG
- El-Qady G, Ushijima K (2001) Inversion of DC resistivity data using neural networks. *Geophys Prospect* 49(4):417–430
- FitzGerald E, Bean C, Reilly R (1999) Fracture-frequency prediction from borehole wireline logs using artificial neural networks. *Geophys Prospect* 47:1031–1044
- Fogelman Soulié F (1992) Neural networks and their applications for the oil industry. *Review Institut Francais du Pétrole* 47:383–392
- Fossati M, Zerilli A, Ronchini G, Apolloni B (1992) Lineament analysis for potential-fields data using neural networks. In: SEG International Exposition and Annual Meeting, New Orleans, Society of Exploration Geophysicists
- Freeman JA (1994) Simulating neural networks with mathematica. Addison-Wesley, Reading
- Fu L-Y (1997) Application of the Caianiello neuron-based network to joint inversion. In: SEG International Exposition and Annual Meeting, Dallas, SEG
- Fu L-Y (2001) Caianiello neural network method for geophysical inverse problems. In: Poulton M (ed) *Computational neural networks for geophysical data processing*. Pergamon, Amsterdam, pp 187–215
- Fu L-Y (2004) Joint inversion of seismic data for acoustic impedance. *Geophysics* 69(4):994–1004

- Fu L-Y, Chen S, Duan Y (1997) ANNLOG technique for seismic wave impedance inversion and its application effect. *Oil Geophys Prospect* 32:34–44
- Gallop J (2006) Facies probability from mixture distributions with non-stationary impedance errors. In: SEG International Exposition and Annual Meeting, New Orleans, SEG
- Garcia HA, Martín N, Aldana M (2002) Elastic and acoustic inversions, discriminant function and neural networks for gas-sand detection. In: SEG International Exposition and Annual Meeting, Salt Lake City, SEG
- Gogoi T, Chatterjee R (2018) Neural network analysis and seismic-velocity model building in Upper Assam Basin, India: an aid to reservoir-characterization study. In: SEG International Exposition and Annual Meeting, Anaheim, SEG
- Gomez F (2005) Reservoir seismic characterization using rock physics, seismic attributes and spectral decomposition in Puerto Colon oil field, Colombia. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Gubbins D (2004) Time series analysis and inverse theory for geophysicists. Cambridge University Press, Cambridge
- Guo Y, Hansen R, Harthill N (1992) Feature recognition from potential fields using neural networks. In: SEG International Exposition and Annual Meeting, New Orleans, Society of Exploration Geophysicists
- Hampson DP, Todorov T (1999) AVO lithology prediction using multiple seismic attributes. In: SEG International Exposition and Annual Meeting, Houston
- Hampson DP, Schuelke JS, Quirein JA (2001) Use of multiattribute transforms to predict log properties from seismic data. *Geophysics* 66(1):220–236
- Haykin S (1994) Neural networks: a comprehensive foundation. Macmillan College Publishing Company, New York
- Haykin S (1999) Neural networks: a comprehensive foundation. Prentice-Hall, Englewood Cliffs
- Haykin S (2009) Neural networks and learning machines. Pearson Education, Upper Saddle River
- Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computation. Addison-Wesley, Redwood City
- Huang KY, Shen LC, Weng LS (2011) Well log data inversion using radial basis function network. In: SEG International Exposition and Annual Meeting, San Antonio, SEG
- Hush DR, Horne BG (1993) Progress in supervised neural networks – what's new since Lippmann? *IEEE Signal Process Mag* 10(1):8–39
- Iturrarán-Viveros U, Spurlin JH (2005) The gamma test applied to select seismic attributes to estimate effective porosity. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Jones F (2007) Inversion concepts. Retrieved 6 April 2010, from <http://www.eos.ubc.ca/ubcgif/iag/tutorials/invn-concepts/index.htm>
- Kaftan I, Şalk M (2009) Determination of structure parameters on gravity method by using radial basis functions networks case study: Seferihisar geothermal area (Western Turkey). In: SEG International Exposition and Annual Meeting, Houston, SEG
- Kappler K, Anderson Kuzma H, Rector JW (2005) A comparison of standard inversion, neural networks and support vector machines. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Kasabov NK (1996) Foundations of neural networks, fuzzy systems, and knowledge engineering. The MIT Press, Cambridge, MA
- Kelter A (2005) Estimation of Thomsen's anisotropy parameters from compressional and converted wave surface seismic traveltimes data using NMO equations, neural networks and regridding. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Kim Y, Nakata N (2018) Geophysical inversion versus machine learning in inverse problems. *Lead Edge* 37(12):894–901
- Kononov A, Gisolf D, Verschuur E (2007) Application of neural networks to traveltimes computation. In: SEG International Exposition and Annual Meeting, San Antonio, SEG
- Kosko B (1992) Neural networks for signal processing. Prentice-Hall, Englewood Cliffs
- Kung SY (1993) Digital neural networks. Prentice-Hall, Englewood Cliffs
- Kuzma HA, Rector JW (2004) Non-linear AVO inversion using support vector machines. In: SEG International Exposition and Annual Meeting, Denver, SEG
- Lau A, Gonzalez A, Mallick S, Gillespie D (2002) Prestack waveform inversion and neural network interpolation – a two-step approach to log prediction. In: EAGE Annual Conference and Exposition, Florence, EAGE
- Lines LR (ed) (1988) Inversion of geophysical data. Geophysics reprint series. SEG
- Lippmann RP (1987) An introduction to computing with neural networks. *IEEE ASSP Mag* 4(2):4–22
- Lippmann RP (1989) Pattern classification using neural networks. *IEEE Commun Mag* 27:47–64
- Liu Z, Liu J (1998) Seismic-controlled nonlinear extrapolation of well parameters using neural networks. *Geophysics* 63(6):2035–2041
- Liu Z, Castanga JP, Zhang X, Li Y (2002) Rock parameter modeling using well and prestack depth migration data by neural networks. In: SEG International Exposition and Annual Meeting, Salt Lake City, SEG
- Maiti S, Tiwari RK (2010) Automatic discriminations among geophysical signals via the Bayesian neural networks approach. *Geophysics* 75(1):E67–E78
- Masters T (1995) Advanced algorithms for neural networks – a C++ sourcebook. Wiley, New York
- Mayer F, Larson G, Dumitrescu CC, Talinga D (2014) Resistivity and density estimation from multicomponent seismic data: case study from the Lower Cretaceous McMurray Formation, Athabasca Oil Sands. In: SEG International Exposition and Annual Meeting, Denver, SEG
- McCord Nelson M, Illingworth WT (1991) A practical guide to neural nets. Addison-Wesley, Reading
- McCormack MD (1991a) Neural computing in geophysics. *Lead Edge* 10:11–15
- McCormack MD (1991b) Neural networks in the petroleum industry. In: SEG Annual Conference and Exposition, Houston, SEG
- Meireles MRG, Almeida PEM, Simoes MG (2003) A comprehensive review for industrial applicability of artificial neural networks. *IEEE Trans Ind Electron* 50(3):585–601
- Michaels P (1993) Surface wave inversion by neural networks. PhD, University of Utah
- Michaels P, Smith RB (1992) Recurrent neural network representation of the inelastic wave equation and full-waveform inversion without local minima. In: SEG International Exposition and Annual Meeting, New Orleans, SEG
- Michaels P, Smith RB (1997) Surface wave inversion by neural networks (radial basis functions) for engineering applications. *J Environ Eng Geophys* 2(1):65–76
- Modin D, Breton P (2007) Evaluation of various methods of pseudo-Vclay lithoseismic attribute construction. In: SEG International Exposition and Annual Meeting, San Antonio, SEG
- Mogensen S, Link CA (2001) Artificial neural network solutions to AVO inversion problems. In: SEG International Exposition and Annual Meeting, San Antonio, SEG
- Nelson MM, Illingworth WT (1991) A practical guide to neural nets. Addison-Wesley, Reading
- Neyamadpoura A, Taibb S, Abdullaha WATW (2009) Using artificial neural networks to invert 2D DC resistivity imaging data for high resistivity contrast regions: a MATLAB application. *Comput Geosci* 35(11):2268–2274
- Nikravesh M, Aminzadeh F, Zadeh LA (eds) (2003) Soft computing and intelligent data analysis in oil exploration. Developments in petroleum science. Elsevier Science B.V., Amsterdam

- Niu L, Ma J, Geng J, Zhou D, Yin X (2015) Mixture density network applied to reservoir parameter inversion in Bohai Oil field. In: SEG International Exposition and Annual Meeting, New Orleans, SEG, pp 2806–2810
- Nunnari G, Bertuccio L, Ferrucci F (2001) A neural approach to the integrated inversion of geophysical data of different types. *IEEE Trans Geosci Remote Sens* 39(4):736–748
- Parker RL (1994) *Geophysical inverse theory*. Princeton University Press, Princeton
- Peters B, Haber E, Granek J (2019) Neural networks for geophysicists and their application to seismic data interpretation. *Lead Edge* 38(7):534–540
- Phan S, Sen MK (2018) Hopfield networks for high-resolution prestack seismic inversion. In: SEG International Exposition and Annual Meeting, Anaheim, SEG
- Phan S, Sen MK (2019) Deep learning with cross-shape deep Boltzmann machine for pre-stack inversion problem. In: SEG International Exposition and Annual Meeting, San Antonio, SEG
- Poulton MM (ed) (2001) *Computational neural networks for geophysical data processing*. Handbook of geophysical exploration: seismic exploration. Elsevier Science, London
- Poulton MM (2002) Neural networks as an intelligence amplification tool: a review of applications. *Geophysics* 67(3):979–993
- Poulton M, Glass C, Sternberg B (1989) Recognizing EM ellipticity patterns with neural networks. In: SEG International Exposition and Annual Meeting, Dallas, SEG
- Pramanik AG, Singh V, Vig R, Srivastava AK, Tiwary DN (2004) Estimation of effective porosity using geostatistics and multiattribute transforms: a case study. *Geophysics* 69(2):352–372
- Priddy KL, Keller PE (2005) *Artificial neural networks: an introduction*. The International Society for Optical Engineering, Bellingham
- Qu S, Zhang F, Yin X (1998) A simulated annealing neural network and Lanczos inverse approach for impedance inversion. In: SEG International Exposition and Annual Meeting, New Orleans, SEG
- Raiche A (1991) A pattern recognition approach to geophysical inversion using neural nets. *Geophys J Int* 105(3):629–648
- Ross C (2017) Improving resolution and clarity with neural networks. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Ross C, Cole D (2017) A comparison of popular neural network facies classification schemes. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Röth G (1993) *Neural networks and seismic inverse problems*. PhD, University of Paris
- Röth G, Tarantola A (1991) Use of neural networks for inversion of seismic data. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Röth G, Tarantola A (1992) Inversion of seismic waveforms using neural networks. In: SEG International Exposition and Annual Meeting, New Orleans, SEG
- Röth G, Tarantola A (1994) Neural networks and inversion of seismic data. *J Geophys Res* 99(B4):6753–6768
- Roy-Chowdhury K, Devilee R, Curtis A (1997) Use of neural networks for non-linear non-unique inversion in geophysics. In: SEG International Exposition and Annual Meeting, Dallas, SEG
- Royle AJ, Logel JD, Lines LR (2004) AVO investigation of the Ben Nevis reservoir at the Hebron asset. In: SEG International Exposition and Annual Meeting, Denver, SEG
- Russell BH (1988) *Introduction to seismic inversion methods*. Society of Exploration Geophysicists. Course Notes vol 2. ISBN (Print) 9780931830655. ISBN (Electronic) 9781560802303
- Russell BH (2019) Machine learning and geophysical inversion – a numerical study. *The Leading Edge* 38(7):512–519
- Russell BH, Hampson D (2006) The old and the new in seismic inversion. *CSEG Recorder* 31(10):5–11
- Saggaf MM, Toksöz MN, Mustafa HM (2003) Estimation of reservoir properties from seismic data by smooth neural networks. *Geophysics* 68(6):1969–1983
- Sandham WA, Leggett M (1998) Geophysical applications of artificial neural networks and fuzzy logic. *First Break* 16:58–60
- Sandham W, Leggett M (eds) (2003) *Geophysical applications of artificial neural networks and fuzzy logic*. Modern approaches in geophysics. Kluwer, Dordrecht
- Sastry RG, Tesfakiros HG (2006) Neural network based interpretation algorithm for combined induced polarization and vertical electrical soundings of coastal zones. *J Environ Eng Geophys* 11(3):197–211
- Schuelke JS, Ruf A, Andersen J, Corwin L (2005) Volume-based rock property predictions and quantifying uncertainty. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Seiberl W, Ahl A, Winkler E (1998) Interpretation of airborne electromagnetic data with neural networks. *Explor Geophys* 29:152–156
- Singh UK, Tiwari RK, Singh SB (2005) One-dimensional inversion of geo-electrical resistivity sounding data using artificial neural networks – a case study. *Comput Geosci* 31(1):99–108
- Spichak V, Popova I (2000) Artificial neural network inversion of magnetotelluric data in terms of three-dimensional earth macro-parameters. *Geophys J Int* 142:15–26
- Stuart GK, Minkoff SE, Pereira F (2019) Enhanced neural network sampling for two-stage Markov chain Monte Carlo seismic inversion. In: SEG International Exposition and Annual Meeting, San Antonio, SEG, pp 1670–1674
- Sun Q, Castagna JP, Liu Z (2000) AVO inversion by Artificial Neural Networks (ANN). In: SEG International Exposition and Annual Meeting, Calgary, SEG
- Sun Q, Eissa M, Castagna JP, Cersosimo DS, Sun S, Decker C (2001) Porosity from artificial neural network inversion for Bermejo Field, Ecuador. In: SEG International Exposition and Annual Meeting, San Antonio, SEG
- Taner T, Lu L, Baysal E (1988) Unified method for 2-D and 3-D refraction statics with first break picking by supervised learning. In: SEG International Exposition and Annual Meeting, Anaheim, SEG
- Taylor CL, Vasco DW (1990) Inversion of gravity gradiometry data using a neural network. SEG international exposition and annual meeting. SEG, San Francisco
- Teles TN, Do Carmo EA (1999) DC resistivity data inversion using neural networks. In: EAGE Annual Conference and Exposition, Helsinki, EAGE
- Thakur RK, Duval F, Mustafa MR (2007) Integrated approach to identify complex reservoir sand channel/bar system. In: SEG International Exposition and Annual Meeting, San Antonio, SEG
- Treitel S, Lines L (2001) Past, present, and future of geophysical inversion – a Y2K analysis. *Geophysics* 66(1):21–24
- Tsang L, Chen Z, Oh S, Marks RJ, Chang ATCII (1992) Inversion of snow parameters from passive microwave remote sensing measurements by a neural network trained with a multiple scattering model. *IEEE Trans Geosci Remote Sens* GE-30:1015–1024
- Tselentis G-A, Serpetsidaki A, Martakis N, Sokos E, Paraskevopoulos P, Kapotas S (2007) Local high-resolution passive seismic tomography and Kohonen neural networks – application at the Rio-Antirio Strait, Central Greece. *Geophysics* 72(4):B93–B106
- Ulrych TJ, Sacchi MD (2005) *Information-based inversion and processing with applications*. Elsevier, Amsterdam
- Van der Baan M, Jutten C (2000) Neural networks in geophysical applications. *Geophysics* 65(4):1032–1047
- Vasco D, Taylor C (1991) Inversion of airborne gravity gradient data, southwestern Oklahoma. *Geophysics* 56:90–101
- Veeken PCH, Priezzhev II, Shmaryan LE, Shteyn YI, Barkov AY, Ampilov YP (2009) Nonlinear multitrace genetic inversion applied

- on seismic data across the Shtokman field, offshore northern Russia. *Geophysics* 74(6):WCD9–WCD19
- Walls J, Derzhi N, Dumas D, Guidish T, Taner T, Taylor G (1999) North Sea reservoir characterization using rock physics, seismic attributes, and neural networks: a case history. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Walls J, Taner T, Taylor G, Smith M, Derzhi N, Carr M, Drummonds J, McGuire D, Morris S, Bregar J (2000) Seismic reservoir characterization of a mid-continent fluvial system using rock physics, poststack seismic attributes and neural networks: a case history. In: SEG International Exposition and Annual Meeting, Calgary, SEG
- Wang B, Pann K, Schuelke JS, Shirley T, Ferguson B (1997) View of neural network training as constrained optimisation and applications to rock porosity prediction. In: SEG International Exposition and Annual Meeting, Dallas, SEG
- Wang Y, Ge Q, Lu W, Yan X (2019) Seismic impedance inversion based on cycle-consistent generative adversarial network. In: SEG International Exposition and Annual Meeting. San Antonio, SEG
- West B, May S, Eastwood J, Rossen C (2001) Interactive seismic facies classification of stack and AVO data using textural attributes and neural networks. In: SEG International Exposition and Annual Meeting, San Antonio, SEG
- Wilamowski BM (2009) Neural network architectures and learning algorithms. *IEEE Ind Electron Mag* 3(4):56–63
- Winkler E (1994) Inversion of electromagnetic data using neural networks. In: EAGE Annual Conference and Exposition, Vienna, EAGE
- Worthington MH (ed) (1987) *Deconvolution and inversion*. Blackwell Scientific Publications, Oxford, UK
- Xiang P (2019) Gravity inversion method based on quasi-radial basis function neural network. In: *International Workshop on Gravity, Electrical & Magnetic Methods and Their Applications*. Xi'an, China
- Xiangjun Z, Jianjun X, Qinxue L (2006) Rock parameter inversion using well and depth domain seismic data by SRM neural networks. In: SEG International Exposition and Annual Meeting, New Orleans, SEG
- Zhang GP (2000) Neural networks for classification: a survey. *IEEE Trans Syst Man Cybern Part C Appl Rev* 30(4):451–462
- Zhang GP (2007) Avoiding pitfalls in neural network research. *IEEE Trans Syst Man Cybern Part C Appl Rev* 37(1):3–16
- Zhang Y, Paulson KV (1997) Magnetotelluric inversion using regularized Hopfield neural networks. *Geophys Prospect* 45(5):725–743
- Zhang Z, Zhou Z (2002) Real time quasi-2-D inversion of array resistivity logging data using neural network. *Geophysics* 67(2):517–524
- Zhang X, Li Y, Liu F, Wang L (1995) Estimating reservoir's lithological parameters from seismic data using neural network. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Zhang L, Poulton M, Zhang Z, Chakravarthy S, Mezzatesta A (1999) Fast forward modeling simulation of resistivity well logs using neural networks. In: SEG International Exposition and Annual Meeting, Houston, SEG
- Zhang Z, Mezzatesta A, Painchaud S (2000a) Wellsite quasi-2D inversion of array resistivity logging data. *Explor Geophys* 31(3):503–507
- Zhang Z, Zhou Z, Frenkel MA, Chunduru R, Mezzatesta A (2000b) Real time inversion of array resistivity logging data using dimensional reduction and neural network simulation. In: SEG International Exposition and Annual Meeting, Calgary, SEG
- Zhang L, Poulton MM, Wang T (2002) Borehole electrical resistivity modeling using neural networks. *Geophysics* 67(6):1790–1797
- Zheng Y, Zhang Q, Yusifov A, Shi Y (2019) Applications of supervised deep learning for seismic interpretation and inversion. *Lead Edge* 38(7):526–533

Inverse Theory, Global Optimization

Mrinal K. Sen and Paul L. Stoffa

Department of Geological Sciences, UT Institute for Geophysics, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX, USA

Definition

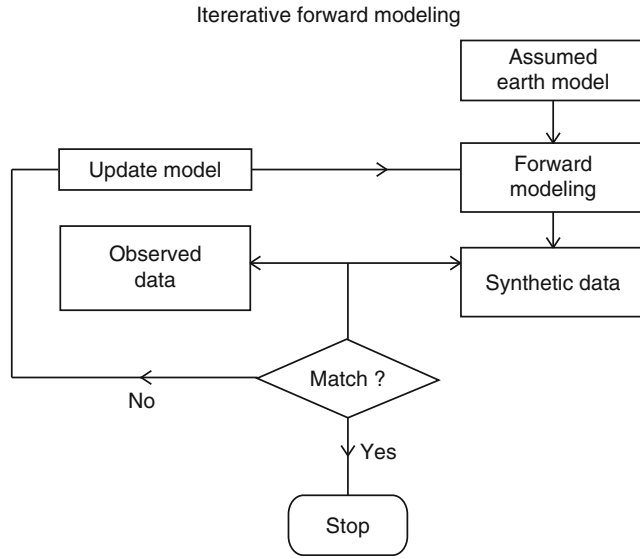
Inversion of geophysical data involves searching for optimal set(s) of earth model parameters that can be used to compute synthetic data that match with observations. The data misfit (or fitness) is often measured by a suitably defined objective function that is generally non-convex. Optimization methods search for a minimum (optimal) value of an objective function. In general, there can be solutions that are locally optimal. *Global optimization* refers to finding the best set of parameters corresponding to the global minimum of an objective function.

Global Optimization

Introduction

The primary goal of geophysics is to determine the properties of the earth's interior from a finite set of remotely sensed measurements from instruments placed on the surface of the earth with perhaps a few in boreholes. The forward problem involves computation of theoretical data for a given set of earth model parameters. Typically, this requires numerical solution of a partial differential equation. The inverse problem whose goal is to estimate earth model parameters from observations is nontrivial due to the fact the data are almost always inadequate, inconsistent, and insufficient (Jackson 1972). In a *model-based inversion* approach, synthetic data are generated for an assumed model and compared with the observed data. This iterative forward modeling procedure is repeated until an acceptable match is obtained between data and synthetics. Thus in this approach, inversion is viewed as *optimization* in which a model is sought that best explains the observations (Fig. 1). Therefore, searching for a model that best explains the field observation the best comprises one of the major tasks of inversion.

Inversion is not restricted to just finding the so-called best-fit model since there can be multiple solutions. In other words, often we have nonunique solutions. Thus, it is very important for a practitioner to have a clear understanding of the forward problem and to impose constraints and restrict the search for an optimal realistic solution within regions that are geologically



Inverse Theory, Global Optimization, Fig. 1 Principle of model-based inversion by optimization

meaningful. This can be addressed by a proper choice of an objective function. One approach to inversion may be termed “exploitation of model space” in which one simply seeks a best-fit model. Another approach may be termed “exploration of model space,” which casts the inverse problem in Bayesian framework (Tarantola 1987) and constructs a posteriori probability density function (PPD) in model space.

Details of geophysical inverse theory with some examples of application can be found in the texts by Menke (1984), Tarantola (1987), Parker (1994), Sen and Stoffa (1995), and Aster et al. (2005). In the following, I will provide a summary of global optimization methods with some examples of application to geophysical inversion.

Method

Background

The essential elements of a model-based inversion algorithm include data, model, forward problem, objective function, and optimization method. We will restrict ourselves to discrete data $\mathbf{d}$ and model $\mathbf{m}$ defined by the following vectors:

$$\begin{aligned} \mathbf{d} &= [d_1 \ d_2 \ d_3 \ \dots \ d_N]^T, \\ \mathbf{m} &= [m_1 \ m_2 \ m_3 \ \dots \ m_M]^T. \end{aligned} \quad (1)$$

In general, $N \neq M$. The forward problem can be represented by the following equation:

$$\mathbf{d}^{\text{syn}} = g(\mathbf{m}), \quad (2)$$

where g is a nonlinear forward modeling operator which operates on the model vector to generate synthetic data vector

$\mathbf{d}^{\text{syn}}$. The next step is to define an objective function that measures a misfit between the observed and synthetic data using a suitably defined norm. The data residual is given by:

$$\Delta \mathbf{d} = \mathbf{d}^{\text{obs}} - \mathbf{d}^{\text{syn}}, \quad (3)$$

and the data misfit norm can be written as:

$$\|\Delta \mathbf{d}\|_p = \left[\sum_{i=1}^N |\Delta d_i|^p \right]^{1/p}, \quad (4)$$

where p is an integer representing the order of the norm. The most commonly used norm is an L_2 norm for which $p = 2$ resulting in a minimization of the mean square error. This corresponds to the popular least squares minimization approach.

To address the issues of stability and nonuniqueness, constraints are incorporated by adding a model norm term, which can be used to impose smoothness, positivity, and bounds. A detailed description of this issue is beyond the scope of this chapter. A fairly general objective function with such constraints using an L_2 norm can be written as:

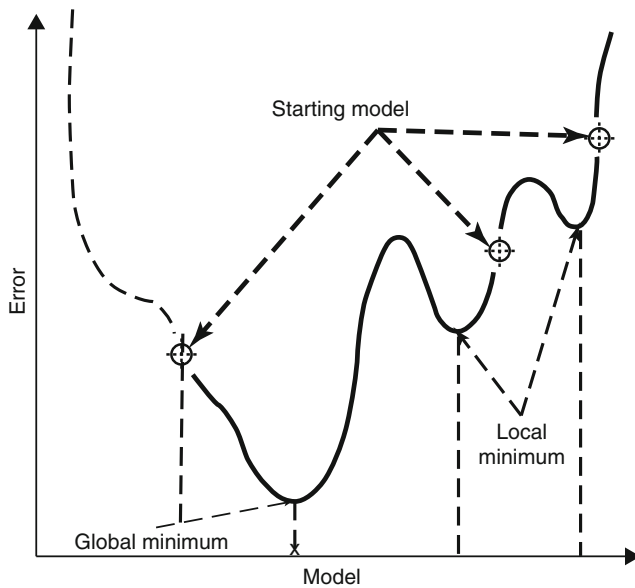
$$\begin{aligned} F(\mathbf{m}) &= (\mathbf{d}^{\text{obs}} - \mathbf{d}^{\text{syn}})^T \mathbf{W}_d (\mathbf{d}^{\text{obs}} - \mathbf{d}^{\text{syn}}) \\ &\quad + \alpha (\mathbf{m} - \mathbf{m}^{\text{prior}})^T \mathbf{W}_m (\mathbf{m} - \mathbf{m}^{\text{prior}}), \end{aligned} \quad (5)$$

where $\mathbf{W}_d$ and $\mathbf{W}_m$ are the data and model weighting matrices, $\mathbf{m}^{\text{prior}}$ is the prior model, and α is a regularization weight. The misfit function given by Eq. (5) is also called the objective function, cost function, error function, energy function, etc. in different applications.

Optimization

The goal of an optimization algorithm is to search for the minimum of an objective function such as the one defined by Eq. (5). Note that the operator g in Eq. (2) is, in general, a nonlinear operator. Therefore, the objective function (Eq. 5) is expected to be non-convex with multiple peaks and troughs (Fig. 2). In other words, the objective function can have multiple local minima; the minimum of all the local minima is called a “global” minimum (e.g., Gill et al. 1981). In most applications, once an objective function has been defined, the goal is to find the global minimum of the objective function.

Optimization methods vary depending on the method of search for the minimum. They can be broadly classified into two categories: local optimization and global optimization. The local optimization methods generally use calculus-based rules to find an update in their search. In other words, they make use of local gradient and/or curvature of the objective function to compute an update. Most of these methods are deterministic in nature and the success of these methods is



Inverse Theory, Global Optimization, Fig. 2 A hypothetical error function as a function of model (the model is assumed to have only one model parameter) showing several minima. The iterative gradient method will find the global minimum only when starting at position 1. Others will end up in secondary minima of the error function

largely dictated by the choice of the starting model. Only if the starting model is close to the globally optimal model, can we expect to reach the global minimum using local optimization methods. These techniques cannot generate or even use the global information needed to find the global minimum for a function with multiple local minima.

The global optimization methods, on the other hand, are mostly based on stochastic rules and use more global information about the misfit surface to update their current position. The interaction between computer science and optimization has yielded new practical solvers for global optimization problems, called *meta-heuristics*. The structures of meta-heuristics are mainly based on simulating nature and artificial intelligence tools. Meta-heuristics mainly invoke exploration and exploitation search procedures in order to diversify the search all over the search space and intensify the search in some promising areas. Therefore, meta-heuristics cannot easily be entrapped in local minima. However, meta-heuristics are computationally expensive due to their slow convergence. One of the main reasons for their slow convergence is that they may fail to detect promising search directions especially in the vicinity of local minima due to their random constructions.

The convergence of these methods to the globally optimal solution is not guaranteed for all the algorithms. Only for some of the algorithms under certain conditions is convergence to the globally optimal solution statistically guaranteed. Also, with real observational data, it is never possible to know whether the derived solution corresponds to the global

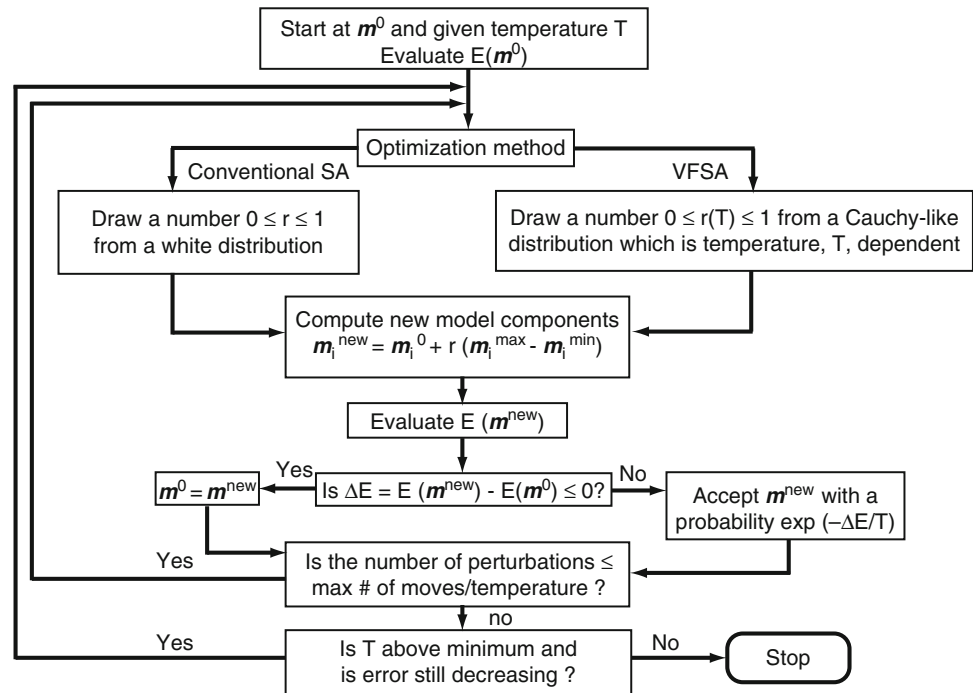
minimum or not. However, our experience indicates that we are able to find many good solutions starting with only poor initial models using global optimization methods (Sen and Stoffa 1995). They are less greedy than the well-known local optimization methods in that during iterative optimization worse solutions are occasionally accepted which allow these algorithms to avoid local minima. There are several variants of global optimization methods; detailed descriptions of some of these approaches as applied to geophysical inversion can be found in Sen and Stoffa (1995). Here I will provide brief descriptions of a few commonly used global optimization methods, namely the *simulated annealing* (SA), *genetic algorithms* (GA), *neighborhood algorithms* (NA), *particle swarm optimization* (PSO), and *trans-dimensional inversion*.

Simulated Annealing Simulated annealing was first proposed by Kirkpatrick et al. (1983). It is analogous to the natural process of crystal annealing when a liquid gradually cools to a solid state. The SA technique starts with an initial model $\mathbf{m}_0$, with an associated error or energy $E(\mathbf{m}_0)$. It draws a new model $\mathbf{m}_{\text{new}}$ from a flat distribution of models within the predefined limits. Note that each model parameter can be bounded by different limits. The associated energy (objective function value) $E(\mathbf{m}_{\text{new}})$ is then computed and compared against $E(\mathbf{m}_0)$. If the energy of the new state is less than the initial state, the new state is considered to be good. In this case, the new model is accepted and replaces the initial model unconditionally. However, if the energy of the new state is larger than the initial state, $\mathbf{m}_{\text{new}}$ is accepted with the probability of $\exp(-(E(\mathbf{m}_{\text{new}}) - E(\mathbf{m}_0))/T)$, where T is a control parameter called annealing temperature that controls if the “bad” model should be carried over to the new model. This completes one iteration. It is the rule of accepting with a probability that makes it possible for SA to be able to jump out of the local minima. The same process is repeated for a large number of times, with the annealing temperature gradually decreasing according to a predefined scheme. Hopefully with a carefully defined cooling schedule, a global minimum can be found. One may choose a linear or a logarithmically decreasing cooling scheme. The tradeoff here is between the computation cost and the accuracy of the result. Fast cooling will fail to produce a crystal (the algorithm gets stuck in a local minimum), and slow cooling takes a long time but may eventually find the global minimum.

To speed up the annealing process without much sacrifice in the solution, a variant of SA, called very fast simulated annealing (VFSA) was proposed by Ingber (1993). VFSA differs from SA in the following ways. The new model is drawn from a temperature-dependent Cauchy-like distribution centered on the current model (Fig. 3). This change has two fundamental effects. First it allows for larger sampling of the model space at the early stages of the inversion when the temperature is high, and much narrower sampling in the

Inverse Theory, Global Optimization, Fig. 3

A flowchart for global optimization by simulated annealing. Note that the VFSA draws models from a Cauchy-like distribution whose shape changes with iteration



model space as the inversion converges when the temperature decreases. Second, each model parameter can have its own cooling schedule and model space-sampling scheme. Therefore, it allows for the individual control for each parameter and the incorporation of a priori information. For many geophysical inversion applications, VFSA has been demonstrated to have excellent performance (Sen and Stoffa 1995).

Genetic Algorithm Genetic algorithms are so named because they emulate the biological processes of evolution and are based on the principle of survival of the fittest (e.g., Goldberg 1989). In a GA, the model parameters are coded in binary form (Fig. 4). The algorithm starts with a randomly chosen population of models called *chromosomes*. The second step is to evaluate the fitness values of these models. Note that unlike our preceding discussion on optimization in which we attempt to minimize an objective function, a GA searches for the maximum of a suitably defined fitness function that measures similarities between the observed and synthetic data. After the selection process, the three genetic processes of selection, crossover, and mutation are performed upon the models in sequence.

In *selection*, models are copied in proportion to their fitness values based on a probability of selection. Thus during this process, models with higher fitness values (lower error values) have higher probability of getting selected.

Crossover acts on the selected pairs of models. This operation picks a crossover site within the paired strings and exchanges, based on a crossover probability, the bits between

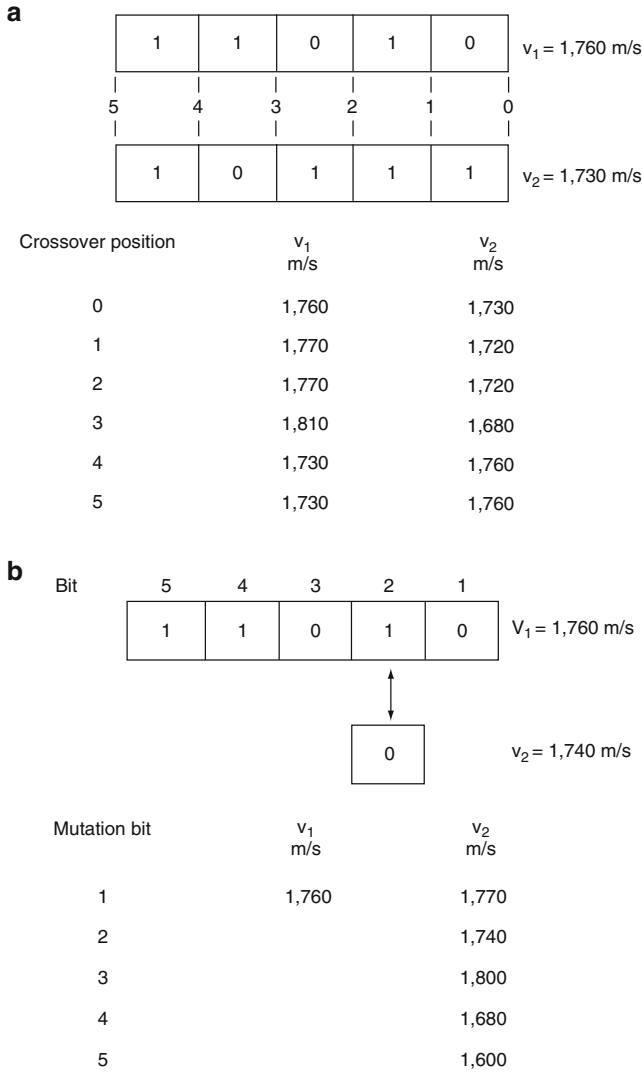
the two models to the right of the crossover site (Fig. 4). Thus the crossover process results in two new models (or children).

Mutation, which involves changing a bit at random based on a mutation probability, is applied to the models to maintain diversity.

After execution of these three processes, the new models are compared to the previous generation and accepted based on an update probability (Stoffa and Sen 1991). The procedure is repeated until convergence is reached, i.e., when the fitness of all the models becomes close to one another.

Note that the repeated applications of selection, crossover, mutation, and update do not necessary guarantee convergence to the global maximum of the fitness function. Nonetheless, the algorithm has been found to be successful in many optimization problems including those from Geophysics.

Neighborhood Algorithm (NA) NA was introduced by Sambridge (1999). The principal idea behind this algorithm is to generate a set of models, at each generation, whose sampling density is constructed from all previous models (Sambridge and Mosegaard 2002). In other words, the proposal distribution to draw the models from is updated at each iteration. One other unique feature of the algorithm is that the algorithm involves portioning of the model parameter space into nearest-neighbor or Voronoi cells about each of the models and thus the information in the previous model samples drives the search for new models. At regular intervals, the approximation is updated, and sampling can concentrate in multiple regions.



Inverse Theory, Global Optimization, Fig. 4 Genetic algorithms are based on binary coding of model parameters (a) and then processes of crossover and mutation (b)

Particle Swarm Optimization The particle swarm is an stochastic evolutionary computation technique (Kennedy and Eberhart 1995) used in optimization, which is inspired by observation of social behavior of individuals (called particles) in nature, such as bird flocking and fish schooling. This has been successfully used in many different science and engineering fields including Geophysics (Shaw and Srivastava 2007). PSO defines a swarm of particles (models) in an M-dimensional space. Each particle maintains memory of its previous best position, $\mathbf{p}_i$ and velocity, $\mathbf{v}_i$. At each iteration, a velocity adjustment of the particle is determined jointly by the previous best position occupied by the particle and the best position of the swarm. The new velocity is then used to compute a new position for the particle. The component of the velocity adjustment determined by the individual's previous best position has been termed as

the *cognition*, and the component influenced by the best in the population is the *social* part. The velocity update formula is given by:

$$\begin{aligned} \mathbf{v}_i^k &= \mathbf{v}_i^{k-1} + b \cdot \text{ran}(\cdot) (\mathbf{m}_i^l - \mathbf{m}_i^k) + c \cdot \text{ran}(\cdot) (\mathbf{m}_g - \mathbf{m}_i^k) \\ \mathbf{m}_i^{k+1} &= \mathbf{m}_i^k + a \mathbf{v}_i^k, \end{aligned} \quad (6)$$

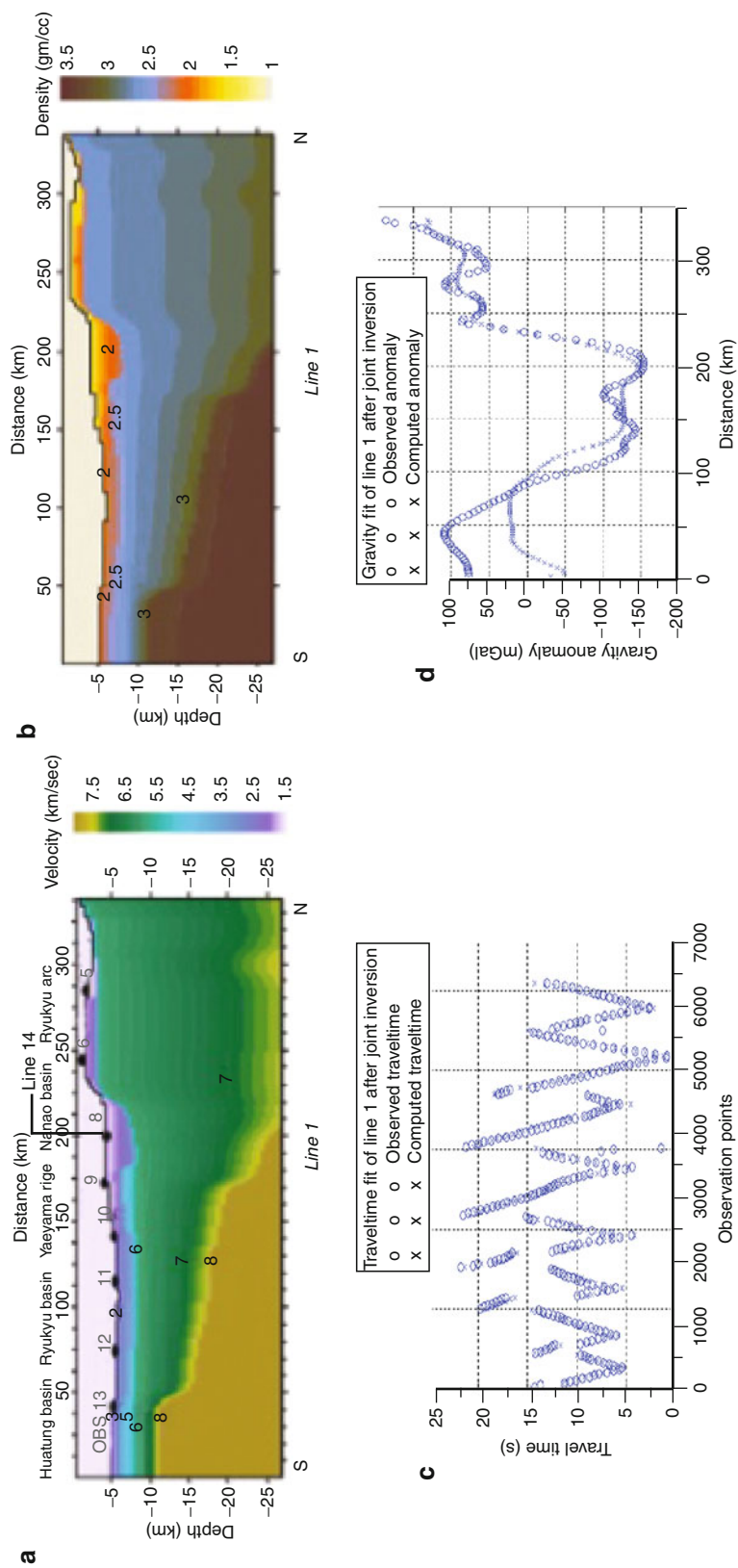
where, $\mathbf{m}_i^k$ is the current location, $\mathbf{v}_i^k$ is the current velocity, $\mathbf{m}_i^l$ is the best location so far, $\mathbf{m}_g$ is the best location achieved by the swarm prior to the kth iteration, $\text{ran}(\cdot)$ is a random number generator, and a, b, and c are constants. This algorithm is very intuitive and easy to program.

Trans-dimensional Inversion In most inverse problems, we assume that the number of model parameters is assumed known, which introduces bias in our estimates. In a trans-dimensional inversion, the model dimension is also treated as an unknown. This is generally formulated using a Bayesian framework and sampling is done using a reversible jump Monte Carlo method (RJMCMC, Malinverno 2002), which is a computationally expensive method. Most recently, Sen and Biswas (2017) developed an RJMCMC method that is augmented by large gradient steps resulting in a faster trans-dimensional algorithm. This new approach has also been applied successfully by Biswas and Sen (2017) to a computationally intensive full waveform inversion problem.

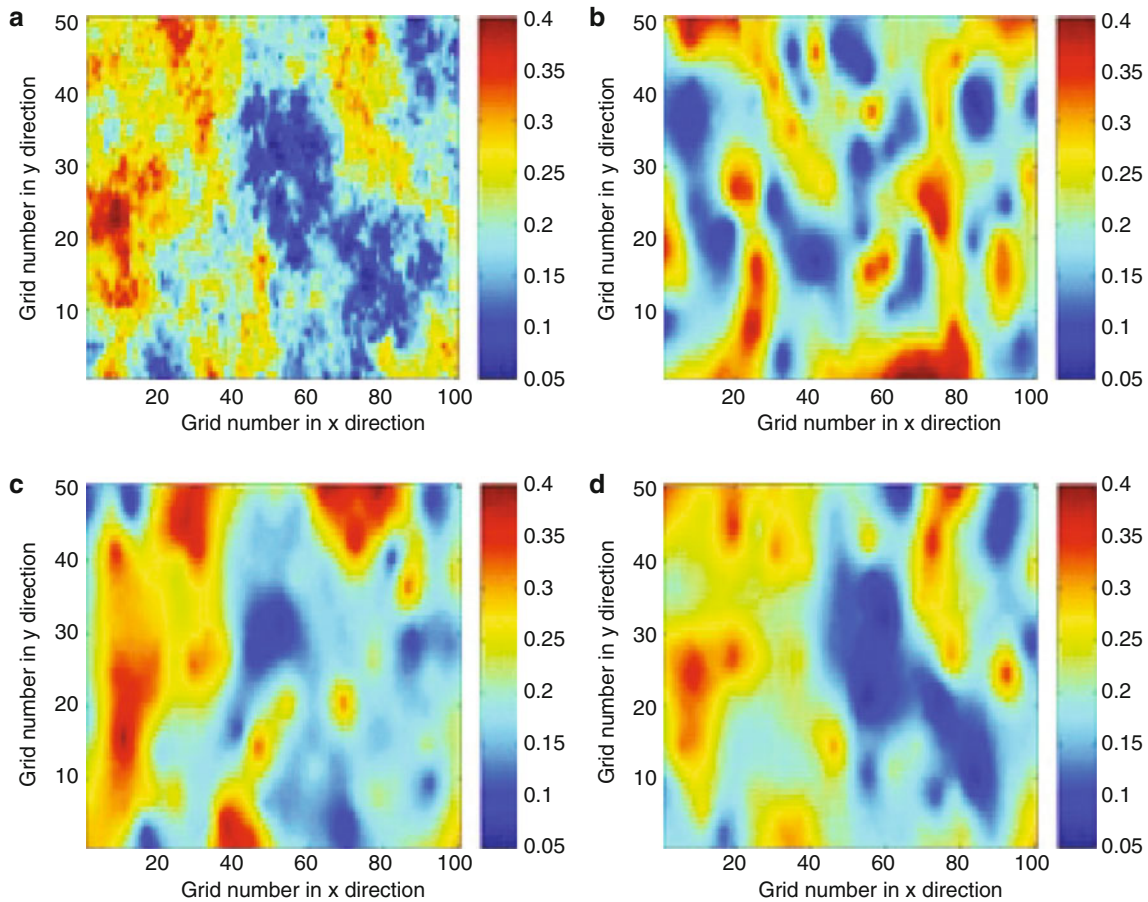
Applications

Several applications of global optimization to geophysical data can now be found in the literature. Due to space limitation, I include here only two examples, which are particularly intriguing since both of these applications address the issue of joint inversion of disparate datasets.

Roy et al. (2005) implemented a joint inversion of first arrival travel time and gravity data with application to field data from a geologically complex subduction zone. They employed a layer-based model description, in which interfaces (which may also be called iso-velocity lines) were defined by a summation of arc-tangent functions. Within each layer, the velocity is assumed to vary linearly with depth at each surface location. Because of the nonuniqueness of the gravity inversion, they used prior knowledge to relate velocity to density values. The nonlinear optimization problem was solved by a very fast simulated annealing (VFSA) technique. At each iteration, travel times were generated by the solution of the eikonal equation while the gravity anomalies are computed using a standard formula. The objective function consists of two parts: one measures the misfit in travel time and the other measures the misfit of gravity anomalies. The technique was applied to field data collected over the Ryukyu subduction zone offshore Taiwan during an ocean



Inverse Theory, Global Optimization, Fig. 5 Example of joint inversion of first arrival seismic travel time and gravity data (modified from Roy et al. 2005): the left panel shows the best fit P-wave velocity (top) and density (bottom) models; the right panel shows the travel time (top) and gravity anomaly (bottom) fit



Inverse Theory, Global Optimization, Fig. 6 Joint inversion of time-lapse seismic and well production data obtained by pilot point parameterization and very fast simulated annealing (modified from Jin

et al. 2009): true porosity (top left); inversion result using only well data (top right); inversion result using only seismic data (bottom left); and inversion result using both seismic data and well data (bottom right)

bottom seismometer (OBS) experiment (called TAICRUST) conducted in the year 1995. An example of results from a profile along NS direction is shown in Fig. 5. The left panel in Fig. 5 shows the velocity and density models derived by joint inversion of travel time and gravity data. The fit between observed and synthetic data is shown in the right panel of Fig. 5.

The second application (taken from Jin et al. 2009) is that of joint inversion of flow and seismic data for reservoir parameters. It is a challenging task in that these disparate datasets are sensitive to different physics and model resolutions for the forward problem. To overcome some of these challenges, Jin et al. (2009) developed a global optimization method based on very fast simulated annealing (VFSA) and a pilot-point-based model parameterization scheme. Reservoir simulation was used to create the saturation and pressure distribution with time. The simulation results were converted to seismic properties using an appropriate rock physics model. Seismic modeling was used to create the seismic response. The objective function was defined as a weighted sum of data misfit and model misfit and VFSA was used to

derive optimal model parameters. Results from application to a synthetic dataset are shown in Fig. 6. A comparison of results from using well data alone (top right panel), seismic data alone (bottom left panel), seismic and well data (bottom right) with the true model parameters (top left panel; a 2D porosity slice) shows marked improvement with the use of time-lapse seismic and well production data (Fig. 6).

Summary

Most geophysical inverse problems are nonlinear in nature and require searching for the minimum of an objective function that is often multimodal. Although linearization of the forward problems and use of local optimization methods have been fairly successful in many applications, many complex problems show suboptimal results using these methods. The global optimization methods are not sensitive to the choice of the starting models and make no assumption on the shape of the objective function. These methods, however, require numerous forward model evaluations and are therefore,

computationally expensive. With the advent of cheap fast computers with large memory, applications of these methods are becoming increasingly popular. In some applications, hybrid methods based on local and global optimization (e.g., Chunduru et al. 1997; Porsani et al. 2000) can be designed that can take advantage of both the approaches.

Cross-References

- [Inverse Theory, Artificial Neural Networks](#)
- [Inverse Theory, Linear](#)
- [Inverse Theory, Monte Carlo Method](#)
- [Inverse Theory, Singular Value Decomposition](#)

Bibliography

- Aki K, Richards PG (2000) Quantitative seismology. W. H. Freeman and Company, San Francisco
- Aster RC, Brochers B, Thurber CH (2005) Parameter estimation and inverse problems. Elsevier Academic Press, New York
- Biswas R, Sen MK (2017) 2D full-waveform inversion and uncertainty estimation using the reversible jump Hamiltonian Monte Carlo, SEG technical program expanded abstracts 2017, 2017
- Chunduru RK, Sen MK, Stoffa PL (1997) Hybrid optimization methods for geophysical inversion. *Geophysics* 62(4):1196–1207
- Gill PE, Murray W, Wright MH (1981) Practical optimization. Academic Press, London
- Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison Wesley Publishing Company, Reading
- Ingber L (1993) Simulated Annealing: Practice versus theory. *Mathl Comput Modeling* 12(8):967–993
- Jackson DD (1972) Interpretation of inaccurate, insufficient and inconsistent data. *Geophys J R Astron Soc* 28:97–100
- Jin L, Stoffa PL, Sen MK, Seif R (2009) Pilot point parameterization in stochastic inversion for reservoir properties using time-lapse seismic and production data. *J Seism Explor* 18:1–20
- Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of the IEEE international conference on neural networks, IV, 1942–1948
- Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680
- Malinverno A (2002) Parsimonious Bayesian Markov chain Monte Carlo inversion in a non-linear geophysical inversion problem. *Geophys J Int* 151(3):675–688
- Menke W (1984) Geophysical data analysis: discrete inverse theory. Academic Press, Amsterdam
- Parker R (1994) Geophysical Inverse Theory. Princeton University Press
- Porsani MJ, Stoffa PL, Sen MK, Chunduru RK (2000) Fitness functions, genetic algorithms and hybrid optimization in seismic waveform inversion. *J Seism Explor* 9(2):143–164
- Roy L, Sen MK, Stoffa PL, McIntosh K, Nakamura Y (2005) Joint inversion of first arrival travel time and gravity data. *J Geophys Eng* 2:277–289
- Sambridge M (1999) Geophysical inversion with a neighbourhood algorithm -I. searching a parameter space. *Geophys J Int* 138:479–494
- Sambridge M, Mosegaard K (2002) Monte Carlo methods in geophysical inverse problems. *Rev Geophys* 40(3), article no. 1009
- Sen MK (2006) Seismic inversion. Society of Petroleum Engineers Press, Richardson
- Sen MK, Biswas R (2017) Trans-dimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo. *Geophysics* 82(3): R119–R134
- Sen MK, Stoffa PL (1995) Global optimization methods in geophysical inversion. Elsevier Publishing Company, Amsterdam
- Shaw RK, Srivastava S (2007) Particle swarm optimization: a new tool to invert geophysical data. *Geophysics* 72(2):V75–F83
- Stoffa PL, Sen MK (1991) Nonlinear multiparameter optimization using genetic algorithms: inversion of plane wave seismograms. *Geophysics* 56(11):1794–1810
- Tarantola A (1987) Inverse problem theory, methods of data fitting and model parameter estimation. Elsevier Publishing Company, Amsterdam

Inverse Theory, Linear

Pravin K. Gupta

Department of Earth Sciences, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

Introduction

The activity of solving inverse problems is all pervading. We perform this activity throughout our life right since birth. As a new born baby the first inverse problem we solve, albeit through pattern recognition, is identification of mother in a group of persons. In this vein, the geophysical inverse problems are just a member of the class of inverse problems encountered and studied in science. Formally, the inverse problem can be defined as an estimation of system parameters together with their uncertainties from the observed system response to a given source excitation. In contrast, the forward problem is defined as computation of response of the system for a given model of its properties and for a given source excitation. Most of the geophysical inverse problems are nonlinear by virtue of the system response being a nonlinear function of system parameters. However, a widely employed class of methodologies solves the nonlinear inverse problems through quasi-linearization whereby the nonlinear problem is recast into a sequence of linear problems that are solved iteratively using linear methodologies. Further, since bulk of the nonlinearity signatures in a geophysical inverse problem arise from geometry of the distribution of physical property, the problem can be recast, in some cases, to a linear problem when the model geometry is considered known. For example, in gravity studies, if the domain of study is divided into small cubes of known volumes and the unknowns are only the density of cubes, the corresponding inverse problem is linear. Similarly, in case of a layered earth model, the problem can be cast into a linear one when the layer thickness is assumed to be known according to some criterion, for example, constant travel time in case of seismics, constant

thickness in DC resistivity, or constant skin depth in case of magnetotellurics (Kunetz 1972; Gupta et al. 1996, 1997). So, there exists an enormous interest in methodologies for solving the linear inverse problems. This article deals with general aspects of linear inverse theory, the description of well- and ill-posed problems, the least square, minimum norm, and generalized inverses of a matrix, the data and parameter resolution matrices, and the ways of handling noise in data. Lastly, the seminal method proposed by Backus and Gilbert (1967, 1968, 1970) for solving linear inverse problems is discussed in some detail. For rigorous readings on inversion theory one can refer the books by Aki and Richards (2002), Indira and Gupta (1998), Lines and Treitel (1984), Menke (1984), Scales et al. (2001), Sneider and Trampert (1999), Tarantola (2005), Tikhonov and Arsenin (1977), Twomey (1977), and Zhdanov (2002), while for numerical aspects of inverse problems one can refer the books by Bjorck (1996), Lawson and Hanson (1995), and Rao and Bhimasankaram (1992) and also the works of Rojas and Sorensen (2002) and Saltelli et al. (2006). For applications and case studies, one can refer Aki and Richards (2002) and Tarantola (2005) for seismological inverse problems, Borcea (2002) for electrical impedance tomography, Habashy and Abubakar (2004) for electromagnetic inversion, Mallat (2009) for signal processing, and Rawlinson et al. (2010) for seismic tomography. For general geophysical applications one can refer to Treitel and Lines (2001) and Oldenburg and Li (2004).

Linear Inverse Problems

The common linear inverse problems can be presented in the form of Fredholm's integral equations of first kind,

$$g(y) = \int_a^b K(x, y)f(x)dx, \quad c < y < d. \quad (1)$$

Here, $K(x, y)$ is the kernel function that transforms the system parameter function, $f(x)$, into the observable function $g(y)$. A special class of linear problems is the convolution problem where the kernel function has the special form $K(y-x)$.

The problem of obtaining a function, $f(t)$, from its Fourier transform function, $F(\omega)$, can be looked upon as an inverse problem as stated below

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt. \quad (2)$$

Oldenburg (1976) studied this linear inverse problem using Backus–Gilbert method.

Another example, the gravity anomaly problem can be stated as

$$g_z(\mathbf{r}) = \iiint_V \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} d^3\mathbf{r}'. \quad (3)$$

Here, $\rho(\mathbf{r}')$ is the unknown density distribution and the remaining factor defines the kernel function. The anomaly $g_z(\mathbf{r})$ is a linear function of density $\rho(\mathbf{r}')$. It may be noted that it is a convolution type problem.

Similarly, in DC resistivity method, the apparent resistivity function, $\rho_a(r)$, is linearly related to the resistivity transform function, $T(\lambda)$, through the kernel function as

$$\rho_a(r) = \int_0^{\infty} T(\lambda)J_n(\lambda r)d\lambda. \quad (4)$$

Ghosh (1971a, b) transformed this problem to a convolution problem by transformation of independent variables.

The discrete equivalent of linear problems can be cast in the form of a matrix equation as

$$\mathbf{Ax} = \mathbf{b}. \quad (5)$$

Here, $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{b}$ is the m -dimensional observation data vector and $\mathbf{x}$ is the n -dimensional parameter variable vector to be determined.

The traditional solution of this problem exists only for the case where $m = n$ and the coefficient matrix is nonsingular so that its exact inverse exists. However, in such special cases also there exist serious problems. Consider the simple example

$$\begin{aligned} x + y &= 2 \\ x + 1.0001y &= 2.0001 \end{aligned}$$

The solution of this problem is $x = y = 1$. However, if the right-hand side vector is an outcome of some observations where accuracy is only up to two decimal places, statement of the problem may become

$$\begin{aligned} x + y &= 2 \\ x + 1.0001y &= 2.00 \end{aligned}$$

The solution will now be $x = 2, y = 0$. Such drastic change in the solution vector for a very small change in data vector is a reflection of the problem being an *unstable* one. Similarly, if our coefficient is an outcome of some computations where the accuracy is only up to two decimal places, assuming the data vector to be error free, the problem becomes

$$\begin{aligned}x + y &= 2 \\x + 1.00y &= 2.0001\end{aligned}$$

In this case we find that the two equations are *inconsistent*.

Lastly, if in the previous case the data is also erroneous, the resulting equations are

$$\begin{aligned}x + y &= 2 \\x + 1.00y &= 2.00\end{aligned}$$

Now there exist infinite solutions and the problem is *nonunique*.

These problems of instability, inconsistency and nonuniqueness invariably crop up while solving inverse problems. If a problem has any one of these features it is termed as an ill-posed problem. Hadamard (1902) formalized the definition of a well-posed problem in terms of the three conditions:

1. The solution exists.
2. The solution is unique.
3. The solution is stable.

Since in applied mathematics one is interested in obtaining the solution, another condition was subsequently added to this list of conditions for a problem to be termed as well posed, that is,

4. The solution is constructible.

A linear or linearized inverse problem, when cast in a matrix form, can be solved as a least square problem. Here, we shall be discussing various ways of solving the matrix equation in its generality. The various classifications of matrix Eq. 5 can be

1. Case when $m > n = k$, is known as strictly overdetermined.
2. Case when $n > m = k$, is known as strictly underdetermined.
3. Case when $m = n = k$, is known as evenly determined.
4. Case when $k < \min(m, n)$ is known as underdetermined.

Here, k is the rank of matrix A .

In any real problem the observed data is, in general, inadequate and inaccurate. As a result the resulting inverse problem is inconsistent, underdetermined, and most of the time unstable. Under such adverse conditions when the exact solution does not exist, the basic task in solving the matrix Eq. 5 pertains to obtaining a solution that meets the desired objectives. Backus (1970, 1996), Backus and Gilbert (1967, 1968, 1970), Jackson (1972, 1979), and others studied these problems and developed methodologies to obtain a unique solution. It is inevitable that in such cases there would be a trade-off between the propagation of data error

to parameter estimation and the parameter resolution. The inevitable nonuniqueness is removed by incorporating the available a priori information about the unknown parameter vector $\mathbf{x}$. The various methods differ in handling the a priori knowledge about data error and model and in controlling the trade-off between error propagation and parameter resolution.

To be able to solve Eq. 5 one designs a matrix A^g , termed as generalized inverse, such that

$$\mathbf{x} = A^g \mathbf{b}. \quad (6)$$

It is obvious that matrix A^g will not possess all the properties of the exact matrix inverse, A^{-1} , which can be defined only for non-singular matrix. For various cases mentioned above, the generalized inverse will possess only a subset of these properties. One of the important properties of exact inverse is $AA^{-1} = A^{-1}A = I$. In case of a matrix A of order $m \times n$, the generalized inverse matrix, A^g , will lead to the product matrices, $S = AA^g$ of order $m \times m$ and $R = A^gA$ of order $n \times n$ and these, in general, will not be identity matrices. The former product matrix S is termed as *data resolution matrix* or *information density matrix*, while the latter product matrix R is known as *parameter resolution matrix*. These product matrices play an important role in the evaluation of quality of solution. Norms of the deviation of matrices S and R from respective Identity matrices provide quantitative measures of quality of information contained in data vector and of the degree of parameter resolution, respectively.

The matrix A^g can be synthesized in several ways for a given type of coefficient matrix A . The important approaches widely employed to solve the Eq. 5 lead to

1. Least square inverse
2. Minimum norm inverse
3. Damped least square (DLS) inverse
4. Rao–Mitra generalized inverse
5. Moore–Penrose generalized inverse based on singular value decomposition (SVD)

All these approaches are briefly presented below followed by Backus–Gilbert method which also provides a solution of the linear inverse problem stated as Eq. 1.

Generalized Inverses

Least Square Inverse

The least square (LS) solution of a strictly overdetermined problem was first proposed by Gauss in 1795. The LS solution exists even for inconsistent system of equations. It is obtained by minimizing the L2 norm of misfit error vector, that is,

$$\left\{ \begin{array}{l} \text{Minimize } \mathbf{e}^T \mathbf{e} = (\mathbf{b} - \mathbf{Ax})^T (\mathbf{b} - \mathbf{Ax}) \\ \text{Solution : } \mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} \\ \text{Inverse : } \mathbf{A}^g = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \\ \text{Information density matrix : } \mathbf{S} = \mathbf{AA}^g \\ \quad = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \neq \mathbf{I}_m \\ \text{Parameter resolution matrix : } \mathbf{R} = \mathbf{A}^g \mathbf{A} \\ \quad = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{A} = \mathbf{I}_n \end{array} \right\} \quad (7)$$

The expressions of matrices $\mathbf{S}$ and $\mathbf{R}$ reveal that the data quality is not optimal while all the parameters are resolved exactly.

Minimum Norm Inverse

The minimum norm (MN) solution of a strictly underdetermined problem is obtained through constrained minimization of the L2 norm of the unknown vector $\mathbf{x}$.

$$\left\{ \begin{array}{l} \text{Minimize } \mathbf{x}^T \mathbf{x} \\ \text{Subject to the constraint } \mathbf{b} - \mathbf{Ax} = \mathbf{0} \\ \text{Solution : } \mathbf{x} = \mathbf{A}^T (\mathbf{AA}^T)^{-1} \mathbf{b} \\ \text{Inverse : } \mathbf{A}^g = \mathbf{A}^T (\mathbf{AA}^T)^{-1} \\ \text{Information density matrix : } \mathbf{S} = \mathbf{AA}^g \\ \quad = \mathbf{AA}^T (\mathbf{AA}^T)^{-1} = \mathbf{I}_m \\ \text{Parameter resolution matrix : } \mathbf{R} = \mathbf{A}^g \mathbf{A} \\ \quad = \mathbf{A}^T (\mathbf{A}^T)^{-1} \mathbf{A} \neq \mathbf{I}_n \end{array} \right\} \quad (8)$$

The expressions of matrices $\mathbf{S}$ and $\mathbf{R}$ suggest that the data quality is optimal while all the parameters are not resolved exactly and there exists linear dependence amongst some components of the parameter vector.

Damped Least Square Inverse

The DLS solution of an overdetermined problem when rank $k < \min(m, n)$ was concurrently developed by Marquardt, Phillips, Tikhonov, and Twomey. It is obtained by minimizing the two objective functions defining norms of error $\mathbf{e}$ and of unknown vector $\mathbf{x}$, subject to the constraint of $\mathbf{x}$ satisfying the matrix Eq. 5.

$$\left\{ \begin{array}{l} \text{Minimize } \mathbf{e}^T \mathbf{e} = (\mathbf{b} - \mathbf{Ax})^T (\mathbf{b} - \mathbf{Ax}) \text{ and } \mathbf{x}^T \mathbf{x} \\ \text{Subject to the constraint } \mathbf{b} - \mathbf{Ax} = \mathbf{0} \\ \text{Solution : } \mathbf{x} = (\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{I})^{-1} \mathbf{A}^T \mathbf{b} \\ \quad = \mathbf{A}^T (\mathbf{AA}^T + \mu^2 \mathbf{I})^{-1} \mathbf{b} \\ \text{Inverse : } \mathbf{A}^g = (\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{I})^{-1} \mathbf{A}^T \\ \quad \text{or } \mathbf{A}^T (\mathbf{AA}^T + \mu^2 \mathbf{I})^{-1} \\ \text{Information density matrix : } \mathbf{S} = \mathbf{AA}^g \neq \mathbf{I}_m \\ \text{Parameter resolution matrix : } \mathbf{R} = \mathbf{A}^g \mathbf{A} \neq \mathbf{I}_n \end{array} \right\} \quad (9)$$

Here, the parameter λ is a control parameter that represents relative weighting given to the two objective functions. The expressions of matrices $\mathbf{S}$ and $\mathbf{R}$ suggest that neither the data quality nor resolution of parameters is optimal.

Weighted Damped Least Square Inverse

To incorporate any available information about the error characteristics of data and/or about smoothness characteristics of parameter vector, the two objective functions optimized in case of DLS inverse, can be redefined in terms of weighting matrices $\mathbf{W}_e$ and $\mathbf{W}_m$. $\mathbf{W}_e$ is the data error covariance matrix and $\mathbf{W}_m$ is the matrix defining smoothness characteristics of the desired model. $\mathbf{W}_m$ can be constructed from the low order differences of parameter vector components.

$$\left\{ \begin{array}{l} \text{Minimize } \mathbf{e}^T \mathbf{W}_e = (\mathbf{b} - \mathbf{Ax})^T \\ \quad \mathbf{W}_e (\mathbf{b} - \mathbf{Ax}) \text{ and } \mathbf{x}^T \mathbf{W}_m \mathbf{x} \\ \text{Subject to the constraint } \mathbf{b} - \mathbf{Ax} = \mathbf{0} \\ \text{Solution : } \mathbf{x} = (\mathbf{A}^T \mathbf{W}_e \mathbf{A} + \lambda^2 \mathbf{W}_m)^{-1} \mathbf{A}^T \mathbf{W}_e \mathbf{b} \text{ or} \\ \quad \mathbf{x} = \mathbf{W}_m^{-1} \mathbf{A}^T (\mathbf{A} \mathbf{W}_m^{-1} \mathbf{A}^T + \lambda^2 \mathbf{W}_e^{-1})^{-1} \mathbf{b} \\ \text{Inverse : } \mathbf{A}^g = (\mathbf{A}^T \mathbf{W}_e \mathbf{A} + \lambda^2 \mathbf{W}_m)^{-1} \mathbf{A}^T \mathbf{W}_e \text{ or} \\ \quad \mathbf{A}^g = \mathbf{W}_m^{-1} \mathbf{A}^T (\mathbf{A} \mathbf{W}_m^{-1} \mathbf{A}^T + \lambda^2 \mathbf{W}_e^{-1})^{-1} \\ \text{Information density matrix : } \mathbf{S} = \mathbf{AA}^g \neq \mathbf{I}_m \\ \text{Parameter resolution matrix : } \mathbf{R} = \mathbf{A}^g \mathbf{A} \neq \mathbf{I}_n \end{array} \right\} \quad (10)$$

All the above inverses can be considered as special cases of the generalized inverse defined by Rao and Mitra (1971) which is discussed next.

Rao–Mitra Inverse

Rao and Mitra (1971) put forth a definition of generalized inverse, $\mathbf{A}^g$, as any matrix that provides a solution of matrix Eq. 5 as

$$\mathbf{x} = \mathbf{A}^g \mathbf{b}. \quad (11)$$

It is evident that this inverse satisfies the relation

$$\mathbf{AA}^g \mathbf{A} = \mathbf{A}. \quad (12)$$

This condition leads to two more properties

$$\begin{aligned} \mathbf{AA}^g &\text{ is idempotent,} \\ \mathbf{A}^g \mathbf{A} &\text{ is idempotent.} \end{aligned}$$

It may be mentioned that idempotent matrices are those for which any power of the matrix is equal to itself.

Rao and Bhimasankaram (1992) presented a suit of algorithms for evaluating the generalized inverse using Gaussian elimination like steps based on elementary row and

column operations that reduce the coefficient matrix to a Hermite canonical form.

An earlier definition of generalized inverse, A^+ , given by Moore (1920) and rediscovered by Penrose (1955), was based on SVD of matrix A . It satisfies the conditions

$$\left\{ \begin{array}{l} AA^+A = bIA \\ A^+AA^+ = A^+ \\ AA^+ \text{ is idempotent} \\ A^+A \text{ is idempotent} \end{array} \right\} \quad (13)$$

Moore–Penrose inverse (MPI) is always unique while Rao–Mitra inverse (RMI) is, in general, not. In fact, MPI is one member of the set of all possible RMIs. To elaborate this point, it may be stated that for a given SVD of matrix A ,

$$A = U\Lambda V^T. \quad (14)$$

Here U and V are the modal matrices of orders $m \times m$ and $n \times n$, respectively and Λ is an $m \times n$ matrix whose leading k diagonal elements correspond to the nonzero singular values and all other elements are zero.

The MPI is given by

$$A^+ = V\Lambda^+U^T \quad (15)$$

Here Λ^+ is the $n \times m$ matrix whose first k diagonal elements correspond to the inverse of nonzero singular values and all other elements are zero.

In contrast, the RMI is given by

$$A^g = V\Lambda^gU^T. \quad (16)$$

Here matrix Λ^g is an $n \times m$ matrix whose leading $k \times k$ submatrix block is diagonal with nonzero diagonal elements corresponding to the inverse of nonzero singular values and all the other submatrices are arbitrary. This freedom of having arbitrary elements outside the main $k \times k$ block leads to the nonuniqueness of MRI. This, in turn, leads to nonunique solutions of matrix Eq. 5, which can be derived from a given solution, $x^{(1)}$, of Eq. 5 as

$$x = x^{(1)} + (A^gA - I)\xi \quad (17)$$

ξ being an arbitrary vector. One possible vector $x^{(1)}$ can be x + the solution obtained using MPI.

Backus–Gilbert Method

Backus and Gilbert (1967, 1968, 1970) studied in detail the problem of synthesizing solutions to a linear inverse

problem expressible as a Fredholm's equation of first kind like Eq. 1. The main feature of this methodology is the possible detailed study of the trade-off between parameter resolution and error propagation. The Backus–Gilbert method (BGM) is briefly described here both for continuous and discrete cases.

Continuous Inverse Problem

The solution of Equation 1 can be represented as

$$f(x) = \int_c^d G(x, y)g(y)dy, \quad a < x < b \quad (18)$$

For Eqs. 1 and 18 to coexist, it is essential that the functions $K(x, y)$ and $G(x, y)$ satisfy the relations

$$\int_a^b G(x, y)K(x, y')dx = \delta(y - y'), \quad (19)$$

and

$$\int_c^d G(x, y)K(x', y)dy = \delta(x - x'). \quad (20)$$

The function $\delta(r-r')$ is the transcendental Dirac-delta function.

Backus and Gilbert brought into focus the infinitely undetermined nature of the real continuous linear inverse problems represented by Eq. 1. They emphasized that from the finite observations, it would never be possible to exactly determine the infinite unknowns constituting the sought model property function $f(x)$. They circumvented the problem by seeking not the exact solution but its approximation as a linear superposition of the observed data values.

Let the finite observation set comprise m values $g_i = g(y_i)$ with $c \leq y_1 \leq y_2 \leq \dots \leq y_{m-1} \leq y_m \leq d$.

Further let the unknown function $f(x)$ need be estimated at the n argument values $a \leq x_1 \leq x_2 \leq \dots \leq x_{n-1} \leq x_n < b$. Now Eqs. 1 and 18 can be rewritten as

$$g_i = g(y_i) = \int_a^b K_i(x)f(x)dx, \quad \forall 1 < i < m \quad (21)$$

and

$$f_j = f(x_j) = \int_c^d G_j(y)g(y)dy, \quad \forall 1 < j < n \quad (22)$$

where

$$K_i(x) = K(x, y_i) \text{ and } G_j(y) = G(x_j, y). \quad (23)$$

Backus and Gilbert sought an approximation of Eq. 22 as

$$f_j \cong \hat{f}_j = f(x_j) = \sum_1^m c_i(x_j) g_i \quad (24) \quad \text{with}$$

Substituting for g_i from Eq. 21, we get

$$\hat{f}_j = \sum_1^m c_i(x_j) \int_a^b K_i(x) f(x) dx$$

For K_i , and f being square integrable, the order of integration and summation can be interchanged and the equation can be rewritten as

$$\hat{f}_j = \int_a^b \left[\sum_1^m c_i(x_j) K_i(x) \right] f(x) dx = \int_a^b A(x, x_j) f(x) dx \quad (25)$$

where

$$A(x, x_j) = \sum_1^m c_i(x_j) K_i(x), \quad (26)$$

is termed by Backus and Gilbert as averaging kernel and is also known in literature as resolving kernel or scanning function.

From Eq. 25, it is evident that for f_j to be exactly estimated, the averaging kernel should be the Dirac-delta function $\delta(x - x_j)$. This equality is, in general, not achievable. In order to make $A(x, x_j)$ as close to $\delta(x - x_j)$ as possible, the coefficients c'_i s are chosen to satisfy a judiciously chosen δ -ness criterion. One such criterion can be

$$q_j = q(x_j) = \int_a^b J(x, x_j) [A(x, x_j) - D(x, x_j)]^2 dx. \quad (27)$$

Several pairs of functions J, D have been used by different workers. Backus and Gilbert used $J(x, x_j) = 12(x - x_j)^2$ and $D(x, x_j) = \delta(x - x_j)$. They interpreted q_j as the half-width or spread of the resolving kernel A . Hence, c'_i s are so chosen that these minimize q_j subject to the condition that A is unimodular, that is,

$$\int_a^b A(x, x_j) dx = 1 \quad (28)$$

Expressing Eqs. 25, 27, and 28 in matrix form, we get

$$\begin{cases} \hat{f}_j = \mathbf{c}^T \mathbf{g} \\ q_j = \mathbf{c}^T \mathbf{M} \mathbf{c} + \mathbf{a}^T \mathbf{c} + b \\ \mathbf{u}^T \mathbf{c} = 1 \end{cases} \quad (29)$$

$$\begin{aligned} M_{ij} &= \int_a^b K_i(x) K_j(x) dx \\ a_i &= -2 \int_a^b K_i(x) J(x, x_j) D(x, x_j) dx \\ b &= \int_a^b D^2(x, x_j) dx \\ u_i &= \int_a^b K_i(x) dx \end{aligned}$$

Using Lagrange's method of undetermined multipliers, the solution of Eq. 29 is found to be

$$\mathbf{c} = \frac{\mathbf{M}^{-1} \mathbf{u}}{\mathbf{u}^T \mathbf{M}^{-1} \mathbf{u}} \quad (30)$$

Substituting these c -values in Eq. 24, one gets the estimate of unknown physical property function f at argument value x_j . All values $f_1, f_2, \dots, f_n$ are obtained by repeating the above steps for each $x_1, x_2, \dots, x_n$.

The quality of solution can be evaluated by analyzing the characteristics of the averaging kernel $A(x, x_j)$, which ideally should be the Dirac-delta function. Hence, the deviation of the center of $A(x, x_j)$ from x_j and the finite range of x over which $A(x, x_j)$ is nonzero may be taken as pointers to the quality of estimation.

The discussion till now was confined to error-free data. However, if the data is erroneous and its error covariance matrix is $\mathbf{E}$, the procedure need to be modified to get a solution. In such cases, the superposition coefficient vector $\mathbf{c}$ is obtained by minimizing besides q_j given by Eq. 27 and the error propagation measure e_j given by

$$e_j = \mathbf{c}^T \mathbf{E} \mathbf{c} \quad (31)$$

The two objective functions are combined into one as

$$r_j = \mathbf{c}^T \mathbf{N} \mathbf{c},$$

where

$$N = \lambda M + (1 - \lambda)E, 0 \leq \lambda \leq 1. \quad (32)$$

The procedure described above can be followed after replacing matrix M by N and the resulting estimate of coefficient vector will then be

$$c = \frac{N^{-1}u}{u^T N^{-1}u}. \quad (33)$$

Equations 21 and 22 make it obvious that for $\lambda = 1$, spread gets minimized but error propagation is uncontrolled while for $\lambda = 0$, the spread is unlimited and error propagation is minimized.

Discrete Inverse Problem

The implementation of BGM for the discrete case represented by Eq. 5 was given by Gupta (1998) where it is shown that, for error-free case, if the unimodularity constraint is not applied the BG estimate is same as the minimum norm estimate. However, when this constraint is applied the BG estimate has an additional correction term.

The solution of matrix Eq. 5 is synthesized as

$$x = A^g b. \quad (34)$$

To synthesize A^g some norm of $(A^g A - I)$ is minimized. In particular, for L2 norm, we can minimize

$$u^T (\{A^g A - I\} (A^T A^{gT} - I)) u.$$

Here $u^T = (1 \ 1 \ \dots \ 1)$. Minimization yields

$$A^g = A^T (A A^T)^{-1}. \quad (35)$$

This will lead to the minimum norm solution x_{min} .

When the unimodularity condition, asking for each row vector of the resolution matrix $R = A^g A = A^T (A A^T)^{-1} A$ to be of unit length, is imposed as constraint, the solution x_{BG} is found to be given as

$$x_{BG} = x_{min} + \frac{u - Ru}{u^T Ru} (u^T m_{min}) \quad (36)$$

Trade-Off Between Resolution and Error Propagation

In Backus–Gilbert (BG) method Eq. 32 demonstrates the trade-off between the two objective functions. For different values of λ one gets different solutions. Hence, it is expected that some value of λ will provide optimal trade-off solution. BG demonstrated in their paper of 1970 that when the values of objective functions q_j and r_j for each λ are plotted against

each other, one gets an L-shaped curve. This curve reveals that for large value of λ , error propagation is uncontrolled and the spread is optimized while for small value of λ , error propagation is optimized and the spread is uncontrolled. The tip of the L-curve provides the optimal value of λ .

In case of DLS method, the same exercise can be carried out to obtain the optimal value of λ . Lot of research has been put in to devise the strategies to find optimal λ values. Most of these works analyze the dependence of solution on λ through eigen analysis. For details, one can refer Calvetti et al. (2000), Hansen (1992), Reichel and Sadok (2008), and Rezghi and Hosseini (2009).

Summary

This article presents salient aspects of linear inverse problems. The LS formulations and the Backus–Gilbert formulation are discussed in brief. The information density and parameter resolution matrices and their importance are discussed. Finally, the methods for finding the optimal multi-objective optimization parameter λ are briefly mentioned. For more details on SVD one can refer to the article by A. Manglik, while for global optimization techniques for nonlinear inverse problems one can refer to the articles by M. K. Sen and by W. Sandham published in this series.

Cross-References

- [Inverse Theory, Artificial Neural Networks](#)
- [Inverse Theory, Global Optimization](#)
- [Inverse Theory, Singular Value Decomposition](#)

Bibliography

- Aki K, Richards PG (2002) Quantitative seismology. University Science Book, Sausalito
- Backus GE (1970) Inference from inadequate and inaccurate data. *Proc Natl Acad Sci* 65:1–7, 281–287; 67:282–289
- Backus GE (1996) Trimming and procrastination as inversion techniques. *Phys Earth Planet Inter* 98:101–142
- Backus GE, Gilbert JF (1967) Numerical applications of a formalism for geophysical inverse problems. *Geophys J R Astron Soc* 13:247–273
- Backus G, Gilbert JF (1968) Numerical applications of a formalism for geophysical inverse problems. *Geophys J R Astron Soc* 16:169–205
- Backus G, Gilbert JF (1970) Uniqueness in the inversion of inaccurate gross earth data. *Philos Trans R Soc Lond A* 266:123–192
- Bjorck A (1996) Numerical methods for least squares problems. SIAM, Philadelphia
- Borcea L (2002) Electrical impedance tomography – topical review. *Inverse Prob* 18(6):R99–R136
- Calvetti D, Morigi S, Reichel L, Sgallari F (2000) Tikhonov regularization and the L-curve for large discrete ill-posed problems. *J Comput Appl Math* 123:423–446

- Ghosh DP (1971a) The application of linear filter theory to the direct interpretation of geoelectrical resistivity sounding measurements. *Geophys Prospect* 19:192–217
- Ghosh DP (1971b) Inverse filter coefficients for the computation of resistivity standard curves for a horizontally stratified earth. *Geophys Prospect* 19:769–775
- Gupta PK (1998) Chapter 7: The Backus Gilbert method. In: Indira NK, Gupta PK (eds) *Inverse methods: general principles and applications in earth system sciences*. Narosa Publishing House, New Delhi, pp 60–68
- Gupta PK, Niwas S, Gaur VK (1996) Straightforward inversion scheme (SIS) for one-dimensional magnetotelluric data. *Proc Indian Acad Sci (EPS)* 105(4):413–429
- Gupta PK, Niwas S, Gaur VK (1997) Straightforward inversion of vertical electrical sounding data. *Geophysics* 62(3):775–785
- Habashy TM, Abubakar A (2004) A general framework for constraint minimization for the inversion of electromagnetic measurements. *Prog Electromagn Res* 46:265–312
- Hadamard J (1902) Sur les problemes aux derives partielles et leur signification physique. *Princeton Univ Bull* 13:49–52. Reprinted in his *Oeuvres*, Vol. III, Centre Nat. Recherche Sci., Paris, 1968, 1099–1105
- Hansen PC (1992) Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev* 34:561–580
- Indira NK, Gupta PK (eds) (1998) *Inverse methods: general principles and applications to earth system sciences*. Narosa, New Delhi
- Jackson DD (1972) Interpretation of inaccurate, insufficient, and inconsistent data. *Geophys J R Astron Soc* 28:97–109
- Jackson DD (1979) The use of a-priori data to resolve nonuniqueness in linear inversion. *Geophys J R Astron Soc* 57:137–158
- Kuntz G (1972) Processing and interpretation of magnetotelluric soundings. *Geophysics* 37:1005–1021
- Lawson CL, Hanson RJ (1995) *Solving least squares problems*. SIAM, Philadelphia
- Lines LR, Treitel S (1984) Tutorial: a review of least-squares inversion and its application to geophysical problems. *Geophys Prospect* 32:159–186
- Mallat S (2009) *A wavelet tour of signal processing: the sparse way*. Elsevier, Amsterdam
- Menke W (1984) *Geophysical data analysis: discrete inverse theory*. Academic, Orlando
- Moore EH (1920) On the reciprocal of the general algebraic matrices. *Bull Am Math Soc* 26:394–395
- Oldenburg DW (1976) Calculation of Fourier transform by the Backus-Gilbert method. *Geophys J R Astron Soc* 44(2):413–431
- Oldenburg DW, Li Y (2004) Inversion for applied geophysics – a tutorial, presented at EMI workshop held at Hyderabad
- Penrose R (1955) A generalized inverse for matrices. *Proc Camb Philos Soc* 51:406–413
- Rao AR, Bhimasankaram P (1992) *Linear algebra*. Tata McGraw-Hill, New Delhi
- Rao CR, Mitra SK (1971) *Generalized inverse of matrices and its applications*. Wiley, New York
- Rawlinson N, Pozgay S, Fishwick S (2010) Seismic tomography: a window into deep Earth – review. *Phys Earth Planet Inter* 178:101–135
- Reichel L, Sadok H (2008) A new L-curve for ill-posed problems. *J Comput Appl Math* 219:493–508
- Rezghi M, Hosseini SM (2009) A new variant of L-curve for Tikhonov regularization. *J Comput Appl Math* 231:914–924
- Rojas M, Sorensen DC (2002) A trust region approach to the regularization of large scale discrete forms of ill-posed problems. *SIAM J Sci Comput* 23(6):1843–1861
- Saltelli A, Ratto M, Tarantola S, Campolongo F (2006) Sensitivity analysis practices: strategies for model based inferences – review. *Reliab Eng Saf* 91:1109–1125
- Scales JA, Smith ML, Treitel S (2001) *Introductory geophysical inverse theory*. Smizdat, Golden, White River Junction
- Sneider R, Trampert J (1999) Inverse problems in geophysics. In: Wirgin A (ed) *Wavefield inversion*. Springer, New York, pp 119–190
- Tarantola A (2005) *Inverse problem theory and methods for model parameter estimation*. SIAM, Philadelphia
- Tikhonov AN, Arsenin VY (1977) *Solutions ill-posed problems*. Wiley, New York
- Treitel S, Lines L (2001) Past, present, and future of geophysical inversion – a new millennium analysis. *Geophysics* 66(1):21–24
- Twomey S (1977) *Introduction to the mathematics of inversion in remote sensing and indirect measurements*. Elsevier Scientific, Amsterdam
- Zhdanov MS (2002) *Geophysical inverse theory and regularization problems*. Elsevier, New York

Inverse Theory, Monte Carlo Method

Malcolm Sambridge¹ and Kerry Gallagher²

¹Seismology and Mathematical Geophysics, Research School of Earth Sciences, The Australian National University, Canberra, ACT, Australia

²UMR 6118 – Géosciences Rennes, Geosciences, Université de Rennes 1, Rennes Cedex, France

Definition

Monte Carlo method. A computational technique making use of random numbers to solve problems that are either probabilistic or deterministic in nature. Named after the famous Casino in Monaco.

Monte Carlo inversion method. A method for sampling a parameter space of variables representing unknowns, governed by probabilistic rules.

Markov chain Monte Carlo (MCMC). A probabilistic method for generating vectors or parameter variables whose values follow a prescribed density function.

Introduction

Because geophysical observations are made at (or very near) the Earth's surface, all knowledge of the Earth's interior is based on indirect inference. There always exists an inverse problem where models of physical properties are sought at depth that are only indirectly constrained by the available observations made at the surface. Geophysicists have been dealing with such problems for many years and in doing so have made substantial contributions to the understanding of inverse problems.

Pioneering work on linear inverse problems in the 1960s arose out of the need to understand how to use new surface observables from seismology to constrain radial variations in geophysical properties at depth within the Earth. Data were

few in number and attention was focused on the mathematical structure of the inverse problem and the ways in which reliable information could be recovered. This resulted in a series of important papers beginning with Backus and Gilbert (1967, 1968, 1970). Since that time the geosciences, like many other fields, have moved into a data-rich environment with increasing availability of computational power. Considerable progress has been made over 30 years utilizing the class of linear (typically least squares) parameter estimation algorithms, which are common to many areas of the physical sciences (Aster et al. 2005). In many of the inverse problems encountered the dependence of data on models is nonlinear and this must be taken into account for meaningful solutions. Often this is achieved by performing a local linearization and using *Inverse Theory, Linear*. As the mathematical relationship between data and unknowns becomes complex then linearized methods fail because they depend heavily on having a starting model for the iterative process which must be close enough to the solution for convergence.

Over the last 30 years there has been considerable progress in the solution of highly nonlinear inverse problems involving a limited number of parameters so that thorough exploration can be made of the character of models. Many algorithms have been devised, most of which make use of random numbers to make decisions, that is, in how to generate a set of values of the unknowns whose predictions can be compared to the available data. The original description of Monte Carlo methods by Hammersley and Handscomb (1964) is “the branch of experimental mathematics which is concerned with experiments on random numbers.” By this definition all inversion techniques that make use of random numbers are Monte Carlo methods. Many of the *Inverse Theory, Global Optimization* inversion methods fall within the class. The particular approach known as Markov chain Monte Carlo is the primary focus of the present entry.

Nonlinearity and Multimodal Fitness Functions

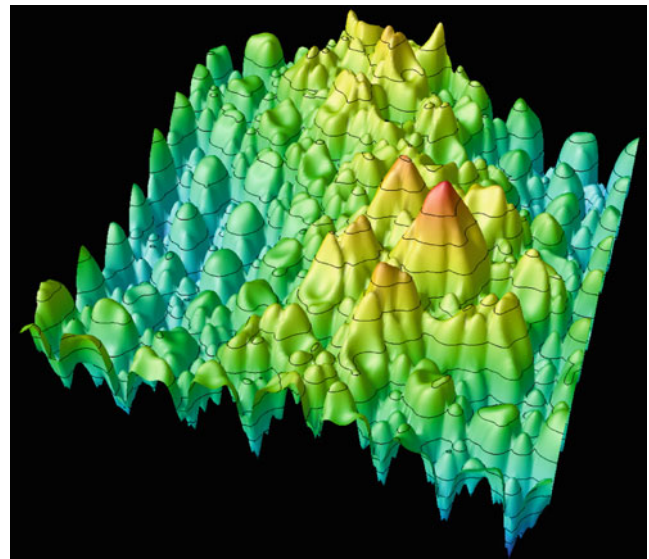
Figure 1 shows a fitness surface from an inverse problem that arises in the analysis of infrasound array data. (Infrasound arrays are used by the United Nations Comprehensive Test Ban Treaty Organization to monitor international adherence to the nuclear test ban treaty.) The height of the surface represents the degree of agreement between two oscillatory fields. There are just two unknowns in this case, which represent tuning parameters in the infrasound array. The object of the exercise is to best tune the array for sensitivity to incoming atmospheric signals, which means finding the point on the surface where the fitness is maximum. We see a curved rim of local maxima from background to foreground, with a broader valley of low fit to the left of the central maximum (red). The physics of the forward problem, that is, calculating the

oscillatory field corresponding to a pair of tuning parameters, as well as the nature of the data itself result in a complicated 2-D fitness function.

The set of unknowns that gives the best fit (i.e., smallest misfit) to data corresponds to the global maximum of the multimodal function and to find it one must employ *Inverse Theory, Global Optimization* techniques such as model space search. In this example the global maximum (at the red central peak) was efficiently found with the neighborhood algorithm of Sambridge (1999), which utilizes ideas from the field of computational geometry. Optimization techniques based on local *Inverse Theory, Linear* would only be suitable once a trial solution is found within the vicinity of the global peak (shaded red). Adaptive Monte Carlo-based direct search approaches like genetic algorithms, simulated annealing, and the neighborhood algorithm (see “► [Inverse Theory, Global Optimization](#)”) are able to solve this (two unknown parameters) problem relatively easily due to their ability to detect the variation of the fit and concentrate sampling where there is most benefit.

This example demonstrates the complexity of inverse problems in cases where the data are highly oscillatory waveforms, a common situation in fields such as acoustics and seismology, where the dimension of the problem is often much higher.

For the 2-D example in Fig. 1, the objective is to tune a particular instrument for maximum sensitivity, and it is appropriate to seek a global maximum. More generally in inverse problems the fitness landscape would represent the difference between observations and predictions made by a mathematical model. In this case simply finding the best-fit solution is inadequate. One needs to characterize the uncertainty in the



Inverse Theory, Monte Carlo Method, Fig. 1 A multimodal data fit surface arising from the mismatch between two oscillatory fields in the infrasound inversion problem (Kennett et al. 2003)

solution, for example, assess how noise in the data lead to errors and trade-offs in the estimated model. Linearized techniques (see “► [Inverse Theory, Linear](#)”) could be used, but all uncertainty estimates are then based on the assumption of local linearity and do not truly reflect the global nature of the data constraint.

Another issue that often arises in inverse problems is that of nonuniqueness (see “► [Inverse Theory, Singular Value Decomposition](#)”). In this case it is not possible to fully constrain the unknowns from the data. The model is unbounded and so best data fit solutions do not exist, and extra assumptions or independent information must be introduced to achieve a single optimal solution. In linearized inversion some form of regularization is used. An example is damping a solution back to some reference set of values (or model) (Aster et al. 2005). It is well known that in this case the details of the solution depend on the nature of regularization used. In addition, uncertainty estimates produced by linearized theory often reflect the choice of regularization. Typically, the least well-constrained components of the solution require the most regularization and resulting uncertainty estimates are severe underestimates of the real errors (Aster et al. 2005 for an example) potential leading to overconfidence in the results.

Bayesian Inference

An alternative approach to inversion is Bayesian inference. Many textbooks and review papers are available. Discussions within a geophysical context can be found in Tarantola and Valette (1982), Duijndam (1988a, b), Sambridge and Mosegaard (2002), and Mosegaard and Sambridge (2002). In Bayesian inference, all information on the unknowns is represented in terms of probability density functions (PDF). Within this framework it is accepted that all inference is relative. What one learns from the data gets added to what is known prior to collecting the data and represented in terms of an a posteriori PDF. The most commonly used form of Bayes’ rule is given below

$$p(m|d) = kp(d|m)p(m) \quad (1)$$

where $p(m|d)$ is the PDF of the model vector, m , containing the unknowns, given the data vector, d , containing the data; $p(d|m)$ is the likelihood function measuring the probability of the data, d , being observed given the model m ; $p(m)$ is the a priori PDF on the model (which is known or assumed about m before the data are collected), and k is a constant of proportionality. In a Bayesian framework, all information on the unknown variables in the model is represented by the posterior PDF, $p(m|d)$ and one usually sets about trying to generate an ensemble of candidate solutions to the inverse problem

whose density is distributed according to this function. This is termed sampling the posterior PDF.

The posterior PDF is the product of the likelihood and the prior PDF. Only the former contains the data vector, d . The likelihood increases as the model fits the data better relative to the noise in the data. The form of the likelihood depends on the statistical character of the data errors. A simple example is a multidimensional Gaussian function characterized by a mean and a covariance matrix, both of which are usually known or assumed. The prior PDF represents information known about the model before collecting the data represented in a probabilistic manner, and may take a variety of forms. Again a multidimensional Gaussian is the most simple, but rarely is real information in this convenient form. Prior PDFs can be the most controversial component of Bayesian inference as there is always a degree of subjectivity in any choice, and the only way to represent no prior information is to not have a prior PDF. Comparisons of Bayesian and alternate approaches can be found in Malinverno and Parker (2005).

We see then that instead of seeking a single optimal solution, in a Bayesian framework many samples are sought. Assessment of the constraints placed on the model is achieved by examining collective properties. Typically, this is done by plotting the distribution of samples as a function of one or more subsets of unknowns, calculating credible intervals to represent uncertainty and covariance matrices to examine the trade-offs between parameters.

The main task to be carried out is then to generate random samples that follow the multidimensional posterior PDF $p(m|d)$ arising from the inverse problem. MCMC methods are practical tools for dealing with complicated probability distributions. Used correctly they result in (quasi)-independent samples whose density follows any target PDF. They have been the subject of much research in fields from Theoretical Physics to Computational statistics. For summaries, see Smith (1991), Smith and Roberts (1993), and Bernardo and Smith (1994). Below we describe the MCMC method briefly and provide a simple illustrative example.

Markov Chain Monte Carlo

Fixed Dimension Approach

MCMC can be regarded as a combination of random Monte Carlo sampling and a Markov chain random walk strategy around the model space. The aim is to produce an ensemble of models from a probability distribution, that is, the posterior PDF, using only function evaluations. The basic approach was developed from the work of Metropolis et al. (1953), placed in a Bayesian framework by Hastings (1970), and a useful overview is given in Gilks et al. (1996). The practical applications lagged behind theoretical developments as a consequence of the need for many simulations. However,

the increase in computing power over the last 15 years or so has led to a rapid increase in use of this methodology in geophysics and other fields of Earth Sciences (e.g., Mosegaard and Tarantola 1995; Malinverno 2002; Sambridge et al. 2006; Gallagher et al. 2009).

The algorithm is as follows: first we choose an initial model from the prior distribution, and calculate its likelihood. Then we generate a new model by making a random perturbation (Monte Carlo) to the current model. This new model is known as the proposed model and depends only on the values of the current model (Markov chain). The final stage is to decide whether we replace the current model with the proposed model, or stay at the current model and repeat the whole process. This important step is determined from the acceptance criterion, which is defined below

$$\alpha = \min \left\{ 1, \frac{p(m')p(d|m')q(m|m')}{p(m)p(d|m)q(m'|m)} \right\} \quad (2)$$

where m' and m are the proposed and current models, respectively, $q(a|b)$ is the probability of proposing model a , given a current model b , and the other distributions are as defined earlier. The decision to accept or reject a proposed model is made by comparing the value of α (which is always between 0 and 1) to a uniform (between 0 and 1) random number, u . If $u < \alpha$ then we replace the current model with the proposed model, if not we discard the proposed model and stay at the current model. We then continue the sampling process (perturb the new current model and so on) for many iterations.

The choice of the proposal function is not critical to the correctness of the sampler, but does affect the efficiency, performance, and convergence. A typical choice might be a normal distribution, centered on the current model, and then we need to tune the performance through the scale parameter of this distribution (e.g., the variance). If we choose too small a scale parameter, the proposed model will be very similar to the current model, their likelihoods will be similar, and we will almost always accept the proposed model. If we choose too large a scale parameter, the proposed model will tend to be very different to the current model, and lead to large changes in the likelihood, which are more likely to be rejected. In practice, both situations mean that we tend to move slowly around the model space. The proposal functions need then to be tuned for particular problems to achieve a reasonable balance between accepting and rejecting the proposed models. A reasonable rate of acceptance is around 30–40%. Generally, we can choose proposal functions that are symmetrical, so that $q(a|b) = q(b|a)$, so these terms cancel out in the acceptance criterion. Also, if we choose uniform prior distributions, then the prior terms also cancel. The acceptance criterion then reduces to the original Metropolis et al. (1953) algorithm.

After an initial period sampling (known as the burn-in), the current model from each iteration is taken as representing a

sample from the posterior distribution (the sampling chain is then stationary). If the model space has N dimensions, and we are interested in the distribution on one of the parameters, m_i , for example, then formally we need to solve the following integral

$$p(m_i) = \int p(m_i, m_j) dm_j; \quad (3)$$

$$j = 1, \dots, i-1, i+1, \dots, N$$

that is we need to integrate out the variation in all parameters except m_i . This is known as marginalizing and $p(m_i)$ is the marginal probability distribution of m_i . Using the MCMC samples, we can just plot all value of m_i as a histogram as the sampling effectively deals with the integration.

Also, it is straightforward to calculate estimates of the expected (or average) value for any parameter. Formally, we have the expected value for parameter m_i defined as

$$E(m_i) = \int p(m_i) m_i dm_i \quad (4)$$

Using the MCMC samples, we simply average over all the samples accepted for that parameter, that is,

$$E(m_i) = \frac{1}{N_a} \sum_{j=1}^{N_a} m_i^j \quad (5)$$

where N_a is the number of samples accepted (post-burn-in) for model parameter m_i .

Transdimensional Approach

A major issue concerning most inverse problems, and the approaches used to solve them, is how best to balance the twin desires of fitting the observations and avoiding introduction of unjustified complexity in the resulting models. Green (1995) introduced a transdimensional form of MCMC (referred to as Reversible Jump), in which the inversion procedure involves the inference of the model complexity (see also birth–death MCMC, Geyer and Møller 1994). For finite dimension models (with a fixed number of unknowns) this then typically becomes a question of determining the dimension of the model. If we are dealing with two models with dimensions k and k' , then acceptance criterion can be written as

$$\alpha = \min \left\{ 1, \frac{p(k')p(m'|k')p(d|m', k')q(m|m')}{p(k)p(m|k)p(d|m, k)q(m'|m)} \right\} \quad (6)$$

Here we separate the prior on the number of dimension, $p(k)$, from the model parameter prior, $p(m|k)$. The proposal function $q()$ becomes more complex as we now want to

propose models with different dimensions. Moreover, we need to allow for the transformation from one model to another to ensure that theoretical probability requirements are maintained. In dealing with a situation where we are simply increasing or decreasing the number of parameters, we can write

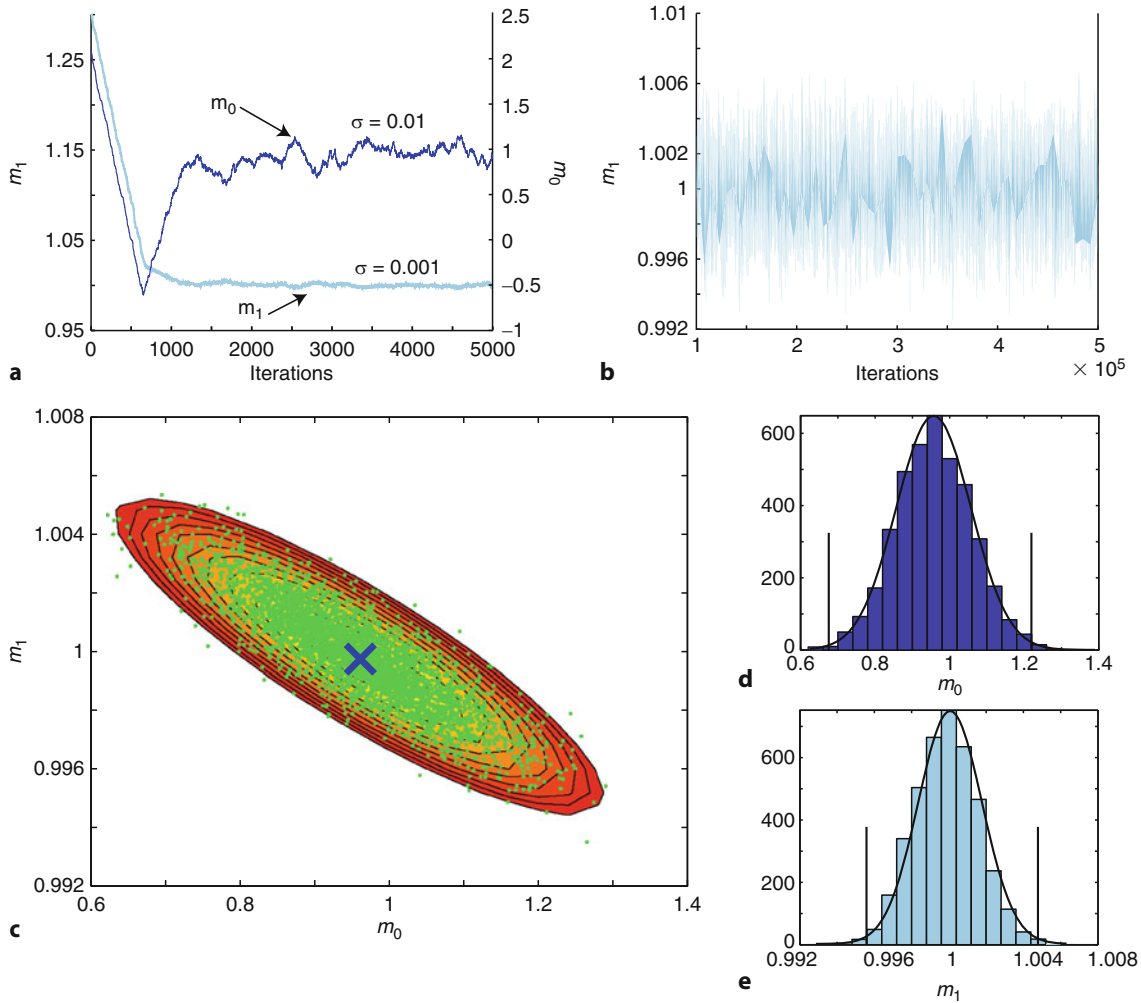
$$\alpha = \min \left\{ 1, \frac{p(k')p(m'|k')p(d|m',k')g(u^k)}{p(k)p(m|k)p(d|m,k)g(u^k)} |J| \right\} \quad (7)$$

Here u^k and u^k are vectors of random numbers of length r' and r , respectively, and used to transform from one model to another, such that $r + k = r' + k'$, and $g(\cdot)$ is the probability distribution used to generate these random numbers. The term

$|J|$ is the Jacobian, and allows for the transformation between the two models, that is,

$$|J| = \frac{\partial(m', u^k)}{\partial(m, u^k)} \quad (8)$$

The last equation for α is actually a general form for the acceptance criterion, although for fixed-dimensional problems the Jacobian is generally 1, and the proposal functions are of the form as described earlier. The details of the reversible jump acceptance criterion are discussed in more detail by Green (2003), Malinverno (2002), and Sambridge et al. (2006) and examples of the implementation algorithms for variable dimension problems are given in Jasra et al. (2006), Bodin and Sambridge (2009), Charvin et al. (2009), and Hopcroft et al. (2009). One important characteristic of the



Inverse Theory, Monte Carlo Method, Fig. 2 (a) Initial 5,000 iterations for sampling of two parameters for the linear regression problem. (b) Post-burn-in iterations for parameter m_1 . (c) The green points show the post-burn-in sampling, and the contours are the log likelihood function. The blue cross is the best solution (equivalent to the analytical

maximum likelihood or least squares solution). (d) Marginal distribution for parameter m_0 . The histogram is constructed from the post-burn-in accepted samples, and the curve is the analytical solution for the marginal distribution. The two vertical bars mark the upper and lower bounds on the 95% credible interval. (e) As (d) but for parameter m_1

Bayesian transdimensional formulation is that it is naturally parsimonious. For two models that fit the data equally well, it will tend to favor simpler models over complex ones as a consequence of the posterior probability effectively being penalized through the addition of more terms to the prior distribution.

A Simple Example

To demonstrate MCMC in action, we choose a simple two parameter linear regression problem, that is,

$$y_i = m_0 + m_1 x_i + \varepsilon_i, \quad i = 1, \dots, N \quad (9)$$

where m_0 and m_1 are the model parameters, y is an observed/measured value, and ε is the data error. In a Bayesian formulation, this problem has an analytical solution for a uniform prior on the model parameters, assuming the data error is known (Lee 1989, p. 180). We chose $m_0 = m_1 = 1$ and generated 100 synthetic data (y) for random values of x between 0 and 100, and added noise ($\varepsilon = 0.5$). We used Gaussian proposal distributions, with different scale parameters ($\sigma_{m_0} = 0.01$, and $\sigma_{m_1} = 0.001$). In Fig. 2 we show the sampling for the two parameters, starting from a randomly selected model. Figure 2a shows that the sampler has not reached the stationary state until at least 1,500–2,000 iterations. Figure 2b shows the sampling for parameter m_1 for later iterations which is clearly stationary. The sampling resembles a white noise spectrum about the mean (or expected) value, lacking any internal structure as a function of iteration. It is also clear that the sampler manages to move toward the upper and lower extreme values of the parameter range (determined by the posterior PDF and the proposal function scale parameter). These are diagnostic (but qualitative) characteristics of stationarity.

Figure 2c shows the 2-D distribution of samples of the post-burn-in accepted samples (here we thinned the chain taking every 100th sample), compared to the log likelihood function (which is proportional to the log of the posterior distribution as we use uniform priors). The density of the sampling increases around the high likelihood values, but there are still some samples from the lower likelihood regions. Figure 2d, e shows the marginal distributions for the two parameters as a frequency histogram, and also the analytical solutions (scaled to the same peak height). These are constructed simply by taking all the accepted values for a given model parameter, as the sampler effectively integrates out the other parameters. Also shown are the 95% credible interval ranges for each parameter. These are constructed by sorting all the samples for a given parameter in ascending order, and finding the indices for the credible values such that 2.5% of the samples are less than the lower credible value, and 2.5% of the values are greater than the upper credible value.

Again, we can see that the sampler has managed to distribute itself across the distribution according to the posterior probability and these histograms are good representations of the marginal distributions.

Summary

Monte Carlo sampling, relying on random numbers, has been used in geophysics for over 40 years, although the increase in computing power has seen a commensurate increase in applications in the last 15 years or so. This approach avoids the use of gradients, is robust to local minima, and so is suitable for nonlinear inverse problems which often have complex misfit (or fitness) surfaces in high dimensions. MCMC, particularly when used in a Bayesian formulation, provides a means of sampling a model space according to the (unknown) posterior distribution for the model parameters. Transdimensional (or reversible jump) Markov chain Monte Carlo generalizes this approach to allow models of different dimensions to be considered, and provides a means of choosing between models of differing complexity. Quantifying the posterior distribution with MCMC is then a solution to the inverse problem and various types of inference can be made from this distribution (e.g., expected values, marginal distributions, credible intervals) to characterize the model space.

Cross-References

- [Inverse Theory, Artificial Neural Networks](#)
- [Inverse Theory, Global Optimization](#)
- [Inverse Theory, Linear](#)
- [Inverse Theory, Singular Value Decomposition](#)

Acknowledgments We would like to thank FAST (French-Australia Science and Technology exchange program) for their support during the preparation of this entry. This project is supported by the Commonwealth of Australia under the *International Science Linkages program*.

Bibliography

- Aster R, Borchers R, Thurber CH (2005) Parameter estimation and inverse problems. International Geophysics Series, vol 90. Elsevier, Amsterdam
- Backus GE, Gilbert JF (1967) Numerical applications of a formalism for geophysical inverse problems. *Geophys J R Astron Soc* 13:247–276
- Backus GE, Gilbert JF (1968) The resolving power of gross Earth data. *Geophys J R Astron Soc* 16:169–205
- Backus GE, Gilbert JF (1970) Uniqueness in the inversion of inaccurate gross Earth data. *Philos Trans R Soc Lond A* 266:123–192
- Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, Chichester
- Bodin T, Sambridge M (2009) Seismic tomography with the reversible jump algorithm. *Geophys J Int* 178:1411–1436

- Charvin K, Gallagher K, Hampson G, Labourdette R (2009) A Bayesian approach to infer environmental parameters from stratigraphic data 1: methodology. *Basin Res* 21:5–25
- Duijndam AJW (1988a) Bayesian estimation in seismic inversion part I: principles. *Geophys Prospect* 36:878–898
- Duijndam AJW (1988b) Bayesian estimation in seismic inversion part II: uncertainty analysis. *Geophys Prospect* 36:899–918
- Gallagher K, Charvin K, Nielsen S, Sambridge M, Stephenson J (2009) Markov chain Monte Carlo (McMC) sampling methods to determine optimal models, model resolution and model choice for Earth science problems. *Mar Pet Geol* 26:525–535
- Geyer CJ, Møller J (1994) Simulation procedures and likelihood inference for spatial point processes. *Scand J Stat* 21:369–373
- Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo in practice. Chapman & Hall, London
- Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–732
- Green PJ (2003) Chapter 6: Trans-dimensional McMC. In: Green PJ, Hjort N, Richardson S (eds) *Highly structured stochastic systems*. Oxford statistical sciences series. Oxford University Press, Oxford, pp 179–196
- Hammersley JM, Handscomb DC (1964) Monte Carlo methods. Chapman & Hall, London
- Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109
- Hopcroft P, Gallagher K, Pain CC (2009) A Bayesian partition modelling approach to resolve spatial variability in climate records from borehole temperature inversion. *Geophys J Int* 178:651–666
- Jasra A, Stephens DA, Gallagher K, Holmes CC (2006) Analysis of geochronological data with measurement error using Bayesian mixtures. *Math Geol* 38:269–300
- Kennett BLN, Brown DJ, Sambridge M, Tarlowski C (2003) Signal parameter estimation for sparse arrays. *Bull Seismol Soc Am* 93:1765–1772
- Lee PM (1989) Bayesian statistics: an introduction. Edward Arnold, New York/Toronto
- Malinverno A (2002) Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophys J Int* 151:675–688
- Malinverno A, Parker RL (2005) Two ways to quantify uncertainty in geophysical inverse problems. *Geophysics* 71:15–27
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. *J Chem Phys* 21:1087–1092
- Mosegaard K, Sambridge M (2002) Monte Carlo analysis of inverse problems. *Inverse Probl* 18:R29–R54
- Mosegaard K, Tarantola A (1995) Monte Carlo sampling of solutions to inverse problems. *J Geophys Res* 100:12431–12447
- Sambridge M (1999) Geophysical inversion with a neighbourhood algorithm – I. Searching a parameter space. *Geophys J Int* 138:479–494
- Sambridge M, Mosegaard K (2002) Monte Carlo methods in geophysical inverse problems. *Rev Geophys* 40:3.1–3.29
- Sambridge M, Gallagher K, Jackson A, Rickwood P (2006) Trans-dimensional inverse problems, model comparison and the evidence. *Geophys J Int* 167:528–542
- Smith AFM (1991) Bayesian computational methods. *Philos Trans R Soc Lond A* 337:369–386
- Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J R Stat Soc Ser B* 55:3–23
- Tarantola A, Valette B (1982) Inverse problems = quest for information. *J Geophys* 50:159–170

Inverse Theory, Singular Value Decomposition

Ajay Manglik

National Geophysical Research Institute (Council of Scientific and Industrial Research),
Hyderabad, India

Definition

Condition number Ratio of the largest singular value to the smallest singular value of a matrix. For a singular matrix, it is infinite, and it is very large for an ill-conditioned matrix.

Description

Linear geophysical inverse theory (Inverse Theory, Linear) deals with solving a system of linear equations, represented in matrix form as

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \quad (1)$$

where $\mathbf{A}$ is a $m \times n$ Jacobean matrix, $\mathbf{x}$ a $n \times 1$ column vector of perturbations in unknown model parameters with respect to a reference model, and $\mathbf{b}$ the corresponding $m \times 1$ column vector of perturbations in observed responses. For a full rank square matrix ($m = n = p$: the rank of matrix $\mathbf{A}$), it is easy to find a unique inverse $\mathbf{A}^{-1}$ and estimate the model parameters as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. However, geophysical inverse problems (Tarantola 1987) mostly deal with inversion of rectangular matrix $\mathbf{A}$ wherein m is either greater than n (overdetermined system representing more number of observations than the unknown parameters, also known as least squares solution) (Lawson and Hanson 1974; Lines and Treitel 1984) or less than n (underdetermined system representing less number of observations compared to the unknown parameters, also called minimum norm solution). For many geophysical problems, the rank p of the matrix is less than m and n . In such situations, linear geophysical inverse theory aims at finding out a pseudo-inverse $\mathbf{A}^{-}$ of $\mathbf{A}$ such that

$$\hat{\mathbf{x}} = \mathbf{A}^{-}\mathbf{b}, \quad (2)$$

where $\hat{\mathbf{x}}$ represents column vector of the estimated model parameters. Moore (1920) and Penrose (1955), using a spectral decomposition form of matrix $\mathbf{A}$, showed that there exists a unique pseudo-inverse $\mathbf{A}^{-}$ of the matrix $\mathbf{A}$ such that (1) $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$, (2) $\mathbf{A}^{-}\mathbf{A}\mathbf{A}^{-} = \mathbf{A}^{-}$,

(3) $(\mathbf{A}\mathbf{A}^-)^T = \mathbf{A}\mathbf{A}^-$, and (4) $(\mathbf{A}^-\mathbf{A})^T = \mathbf{A}^-\mathbf{A}$, where superscript T denotes the transpose of matrix. Further details can be found in Inverse Theory, Linear.

Singular value decomposition (SVD) is a powerful tool to invert matrices that are either singular or numerically very close to singular (ill-conditioned system). Standard matrix equation solvers can fail in such situations. The advantage of SVD is that one need not a priori know the nature of the problem, that is, whether it is overdetermined, under-determined, or partially determined. SVD is very general in the sense that it can be applied to any arbitrary size matrix. The technique was discovered independently by Beltrami in 1873 and Jordan in 1874 as a tool to solve square matrices and was extended to rectangular matrices by Eckart and Young in the 1930s (Stewart 1993). Its use as a computational tool dates back to the 1960s (Golub and Kahan 1965; Golub and Reinsch 1970). Golub and van Loan (1996) demonstrated its usefulness and feasibility in a wide variety of applications.

Following SVD, a matrix $\mathbf{A}$ of size $m \times n$ is factorized in the form

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T, \quad (3)$$

where $\mathbf{U}$ is a $m \times m$ unitary matrix, $\mathbf{\Lambda}$ is a $m \times n$ diagonal matrix, and $\mathbf{V}$ is a $n \times n$ unitary matrix. The columns of $\mathbf{V}$ form a set of orthonormal vectors called parameter eigenvectors, and the columns of $\mathbf{U}$ form a set of orthonormal basis vectors called data eigenvectors. For a special case of $m = n$, the matrix $\mathbf{\Lambda}$ has the form

$$\mathbf{\Lambda} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix}, \quad (4)$$

where $\sigma_1, \dots, \sigma_n$ are called the singular values of matrix $\mathbf{A}$. The number of non-zero singular values is equal to the rank of $\mathbf{A}$. In the SVD representation, the singular values are arranged in the order of decreasing amplitude, that is,

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n. \quad (5)$$

For a real matrix, the columns of matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, that is, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and T is transpose of matrix. For complex matrices transpose is replaced by complex conjugate.

Inverse of $\mathbf{A}$

The expression for the inverse of $\mathbf{A}$ can be obtained by first substituting Eq. 3 into Eq. 1 to get

$$\mathbf{U}\mathbf{A}\mathbf{V}^T\mathbf{x} = \mathbf{b}, \quad (6)$$

and then applying the following orthogonal transformation

$$\mathbf{b} = \mathbf{U}\mathbf{b}'; \quad \mathbf{x} = \mathbf{V}\mathbf{x}', \quad (7)$$

and multiplying Eq. 6 by $\mathbf{U}^T$, which gives

$$\mathbf{\Lambda}\mathbf{x}' = \mathbf{b}'. \quad (8)$$

Since $\mathbf{\Lambda}$ is a diagonal matrix, we can easily get

$$\mathbf{x}' = \mathbf{\Lambda}^{-1}\mathbf{b}'. \quad (9)$$

From Eq. 7 and Eq. 9, and orthogonality property of data and parameter eigenvectors, we get

$$\mathbf{x} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^T\mathbf{b}, \quad (10)$$

or the inverse of $\mathbf{A}$ can be written as

$$\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^T. \quad (11)$$

Rank-Deficient Matrix and Its Inverse

The advantages of above spectral decomposition can be seen in the case of $\mathbf{A}$ being a singular matrix. Let us assume that only p of n singular values of $\mathbf{A}$ are non-zero. In this case, the data and parameter eigenvectors can be split into two subspaces representing non-zero eigen-space and null-space, respectively, that is,

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_p & 0 \\ 0 & 0 \end{bmatrix}; \mathbf{U} = [\mathbf{U}_p, \mathbf{U}_0]; \mathbf{V} = [\mathbf{V}_p, \mathbf{V}_0], \quad (12)$$

then

$$\mathbf{A} = [\mathbf{U}_p, \mathbf{U}_0] \begin{bmatrix} \mathbf{\Lambda}_p & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_p \\ \mathbf{V}_0 \end{bmatrix}^T, \quad (13)$$

which gives

$$\mathbf{A} = \mathbf{U}_p\mathbf{\Lambda}_p\mathbf{V}_p^T. \quad (14)$$

This indicates that matrix $\mathbf{A}$ can be reconstructed only from non-zero singular values and corresponding vector spaces. An inverse of a rank-deficient matrix can be written as:

$$\mathbf{A}^- = \mathbf{V}_p\mathbf{\Lambda}_p^{-1}\mathbf{U}_p^T, \quad (15)$$

by substituting Eq. 12 into Eq. 11.

Ill-Condition Matrix

Equation 15 assumes that the singular values of the system are either zero or positive scalars. If some of the singular values

are very small, then it is easy to see that the inverse of such values can be very large and, in the case of error in data, can lead to unstable solutions. Stability of the matrix inverse can be defined in terms of ratio of largest to smallest singular value, also called the condition number

$$\eta = \frac{\sigma_1}{\sigma_{\min}}. \quad (16)$$

The matrix is ill-conditioned if the condition number is very large. In such situations, the system can be stabilized by specifying a cutoff threshold η_0 for the condition number and ignoring all singular values that lead to the condition number larger than this threshold. Let there be p non-zero singular values arranged in decreasing order of their amplitudes of which only the first q singular values have the condition number smaller than η_0 . Then $(q + 1, \dots, p)$ singular values can be reset to zero thus reducing the rank of the matrix to q . This can be understood as ignoring finer details of the parameter space, which are associated with small singular values. Geophysical inverse problems normally fall under this category.

For least squares inverse problems, where an inverse of $\mathbf{A}$ is given by the expression

$$\mathbf{A}^- = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T, \quad (17)$$

Marquardt (1963) suggested that an ill-conditioned matrix $\mathbf{A}^T \mathbf{A}$ may be stabilized by adding an arbitrary damping factor λ to its diagonal elements, that is,

$$\mathbf{A}^- = (\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{I})^{-1} \mathbf{A}^T. \quad (18)$$

In terms of singular value decomposition, it amounts to boosting the small singular values of the matrix $\mathbf{A}^T \mathbf{A}$. This can be shown by expressing Eq. 18 in terms of singular values decomposition of the matrix $\mathbf{A}$, which gives

$$\mathbf{A}^- = \mathbf{V} \text{diag}\left(\frac{\sigma}{\sigma^2 + \lambda^2}\right) \mathbf{U}^T, \quad (19)$$

where $\text{diag}(\dots)$ represents the elements of a diagonal matrix. Here, it can be seen that adding λ to small singular values stabilizes the inverse.

Sensitivity to Errors in Data

Field observations normally contain noise superimposed on the actual response. If we assume a noise vector $\boldsymbol{\varepsilon}$, then we observe $\mathbf{b} + \boldsymbol{\varepsilon}$ instead of $\mathbf{b}$. This leads to an error $\delta \mathbf{x}$ in the estimated model parameters. Therefore, for noisy data Eq. 1 may be written as:

$$\mathbf{A}(\mathbf{x} + \delta \mathbf{x}) = \mathbf{b} + \boldsymbol{\varepsilon}. \quad (20)$$

Substituting from Eq. 1, we get

$$\mathbf{A} \delta \mathbf{x} = \boldsymbol{\varepsilon}. \quad (21)$$

Expressing matrix $\mathbf{A}$ in spectral form and applying orthogonal transformation $\delta \mathbf{x}' = \mathbf{V} \delta \mathbf{x}$ and $\boldsymbol{\varepsilon}' = \mathbf{U} \boldsymbol{\varepsilon}$, Eq. 21 gives

$$\delta \mathbf{x}' = \boldsymbol{\Lambda}^{-1} \boldsymbol{\varepsilon}'. \quad (22)$$

This shows that in the case of small singular values, small error in data yields large error in estimated model parameters. This problem of instability of estimated model parameters can be overcome by either truncating the small singular values or boosting these by adding a Marquardt damping factor, as discussed in the previous section. Such treatment nevertheless results in loss of finer details of the model parameters.

Resolution Matrices

In linear inverse theory, the concepts of parameter resolution and data information density analyses are used to analyze the resolution of estimated model parameters and the regions of data contributing maximum to these model estimates. These resolution matrices can be expressed in terms of data and parameter eigenvectors. Parameter resolution matrix for a rank-deficient system can be obtained by substituting Eqs. 1, 14, and 15 into Eq. 2, which gives

$$\mathbf{R} = \mathbf{V}_p \mathbf{V}_p^T. \quad (23)$$

For a well-resolved system, $\mathbf{R}$ is an identity matrix $\mathbf{I}$. For a rank-deficient system, this matrix provides information about the equivalence of various model parameters.

Similarly, data information density matrix can be written as

$$\mathcal{J} = \mathbf{U}_p \mathbf{U}_p^T, \quad (24)$$

by substituting Eqs. 2, 14, and 15 into Eq. 1.

Jupp and Vozoff (1975) provided an elegant description of the concepts of stability of solution and resolution of model parameters for geophysical data contaminated by noise, and Vozoff and Jupp (1975) showed its application to direct current (DC) electrical resistivity and magnetotelluric (MT) data interpretation.

Example

The following example demonstrates the use of SVD for an overdetermined system and ill-conditioned matrix

$$\mathbf{A} = \begin{bmatrix} 3.0 & 1.0 & 2.9999 \\ 1.01 & 2.9 & 0.9999 \\ 0.5 & 2.0 & 0.5 \\ 0.3 & 0.31 & 0.29 \end{bmatrix}, \quad (25)$$

which has rank $p = 3$. It is decomposed into U , Λ , and V with three non-zero singular values as 5.2573, 2.5613, 0.0078, and

$$U = \begin{bmatrix} -0.7700 & 0.6311 & -0.0932 & 0.0115 \\ -0.5380 & -0.6086 & 0.2595 & -0.5224 \\ -0.3285 & -0.4809 & -0.4604 & 0.6699 \\ -0.0988 & -0.0056 & 0.8438 & 0.5275 \end{bmatrix},$$

$$V = \begin{bmatrix} -0.5796 & 0.4047 & 0.7073 \\ -0.5740 & -0.8188 & -0.0019 \\ -0.5784 & 0.4071 & -0.7069 \end{bmatrix}. \quad (26)$$

First three columns of U form matrix U_p whereas V_p is equal to V . Following Eq. 15, the pseudo-inverse of A is obtained as

$$A^+ = \begin{bmatrix} -8.2480 & 23.4520 & -41.7228 & 76.3959 \\ -0.0952 & 0.1905 & 0.3010 & -0.1915 \\ 8.6133 & -23.5143 & 41.6213 & -76.3367 \end{bmatrix}, \quad (27)$$

with the resolution matrix $\mathfrak{R}$ as the identity matrix I .

In this example, third singular value is small having the condition number of 672.89. In order to stabilize the solution, we can reset this singular value to zero (new rank = 2). Thus, the data and parameter eigenvectors become

$$U_p = \begin{bmatrix} -0.7700 & 0.6311 \\ -0.5380 & -0.6086 \\ -0.3285 & -0.4809 \\ -0.0988 & -0.0056 \end{bmatrix} \text{ and} \quad (28)$$

$$V_p = \begin{bmatrix} -0.5796 & 0.4047 \\ -0.5740 & -0.8188 \\ -0.5784 & 0.4071 \end{bmatrix},$$

yielding the pseudo-inverse as

$$A^+ = \begin{bmatrix} 0.1846 & -0.0368 & -0.0398 & 0.0100 \\ -0.1177 & 0.2533 & 0.1896 & 0.0126 \\ 0.1850 & -0.0375 & -0.0403 & 0.0100 \end{bmatrix}, \quad (29)$$

and the resolution matrix $\mathfrak{R}$ as

$$\mathfrak{R} = \begin{bmatrix} 0.4997 & 0.0013 & 0.5 \\ 0.0013 & 1.0 & -0.0013 \\ 0.5 & -0.0013 & 0.5003 \end{bmatrix}. \quad (30)$$

It can be seen that stabilizing of the inverse of the matrix comes at the cost of compromising on the resolution of model parameters. Here, first and third parameters are coupled when the rank of the matrix is reduced.

Summary

SVD is a powerful tool to solve geophysical linear inverse problems which often deal with rank-deficient singular matrices. The advantage of SVD is that one need not a priori know the nature of the problem, i.e., whether it is overdetermined, under-determined, or partially determined. It is very general in the sense that it can be applied to any arbitrary size matrix.

Cross-References

► [Inverse Theory, Linear](#)

Acknowledgments The example is computed using Matlab. CSIR-NGRI contribution number NGRI/Lib/2019/Pub-81 under the project MLP6404-28(AM).

Bibliography

- Golub GH, Kahan W (1965) Calculating the singular values and pseudo-inverse of a matrix. *J Soc Indus Appl Math Series B* 2:205–224
- Golub GH, Reinsch C (1970) Singular value decomposition and least squares solutions. *Numer Math* 14:403–420
- Golub GH, Van Loan CF (1996) *Matrix computations*, 3rd edn. Johns Hopkins, Baltimore
- Jupp DLB, Vozoff K (1975) Stable iterative methods for the inversion of geophysical data. *Geophys J R Astron Soc* 42:957–976
- Lawson CL, Hanson RJ (1974) *Solving least squares problems*. Prentice-Hall, Englewood Cliffs
- Lines LR, Treitel S (1984) A review of least-squares inversion and its application to geophysical problems. *Geophys Prospect* 32:159–186
- Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. *SIAM J Appl Math* 11:431–441
- Moore EH (1920) On the reciprocal of the general algebraic matrices. *Bull Am Math Soc* 26:394–395
- Penrose R (1955) A generalized inverse for matrices. *Math Proc Camb Philos Soc* 51:406–413
- Stewart GW (1993) On the early history of the singular value decomposition. *SIAM Rev* 35:551–566
- Tarantola A (1987) *Inverse problem theory: methods for data fitting and model parameter estimation*. Elsevier Science, Amsterdam
- Vozoff K, Jupp DLB (1975) Joint inversion of geophysical data. *Geophys J R Astron Soc* 42:977–991

Isostasy

A. B. Watts

Department of Earth Sciences, University of Oxford, Oxford, UK

Synonyms

Isostacy – (Dutton 1882).

Definition

Isostasy – a principle or general law (Heiskanen 1931). *Isostasy considers there is a certain surface within the Earth, known as the depth of compensation, on which the vertical stresses due to an overlying column of rock are equal (isos = Greek ἴσος “equal,” stasis = Greek στάσις “a standing still”). Isostasy implies a state of hydrostatic equilibrium such that the Earth’s crust and mantle float on their substrate and light regions have a greater elevation than dense regions.*

Isostatic equilibrium – an idealized state that the outermost layers of the Earth tend toward following their disturbance by the addition or removal of loads associated, for example, with the waxing and waning of ice sheets, the growth and decay of volcanoes and deposition, sliding and slumping of sediments.

Local isostasy – a hypothesis that the outermost layers of the Earth are weak such that the deformation caused by loading and unloading is localized, intense, and may involve faulting.

Regional (or flexural) isostasy – a hypothesis that the outermost layers of the Earth have an intrinsic strength such that they resist the deformation associated with loading and unloading and bend (or flex) over a broad area rather than break.

Thermal isostasy – a hypothesis that the outermost layers of the Earth form a cooling thermal boundary layer and that regional differences in topography are controlled by differences in the temperature structure such that hot regions have a greater elevation than cold regions.

Historical Background

Isostasy is a principle that has played a major role in our understanding of the structure, morphology, and tectonic evolution of Earth’s outermost layers. The term was first coined by Dutton (1889), but there is evidence (Delaney 1940; Vai

2006) that the equilibrium state of the Earth’s crust was being considered much earlier by, for example, the Italian scientists Leonardo da Vinci (1452–1519) who pondered the occurrence of marine fossils at high elevations in mountain ranges and by Luigi Fernando Marseli (1658–1730) who was puzzled about the juxtaposition of ocean deeps and mountains. J. F. W. Herschel (1792–1871) was one of the first to link subsidence in one area to uplift in another by some sort of subcrustal mantle flow and both C. Lyell (1797–1875) and C. Babbage (1790–1871) through observations at the Temple of Serapis appreciated that the same area of the crust could experience both subsidence and uplift.

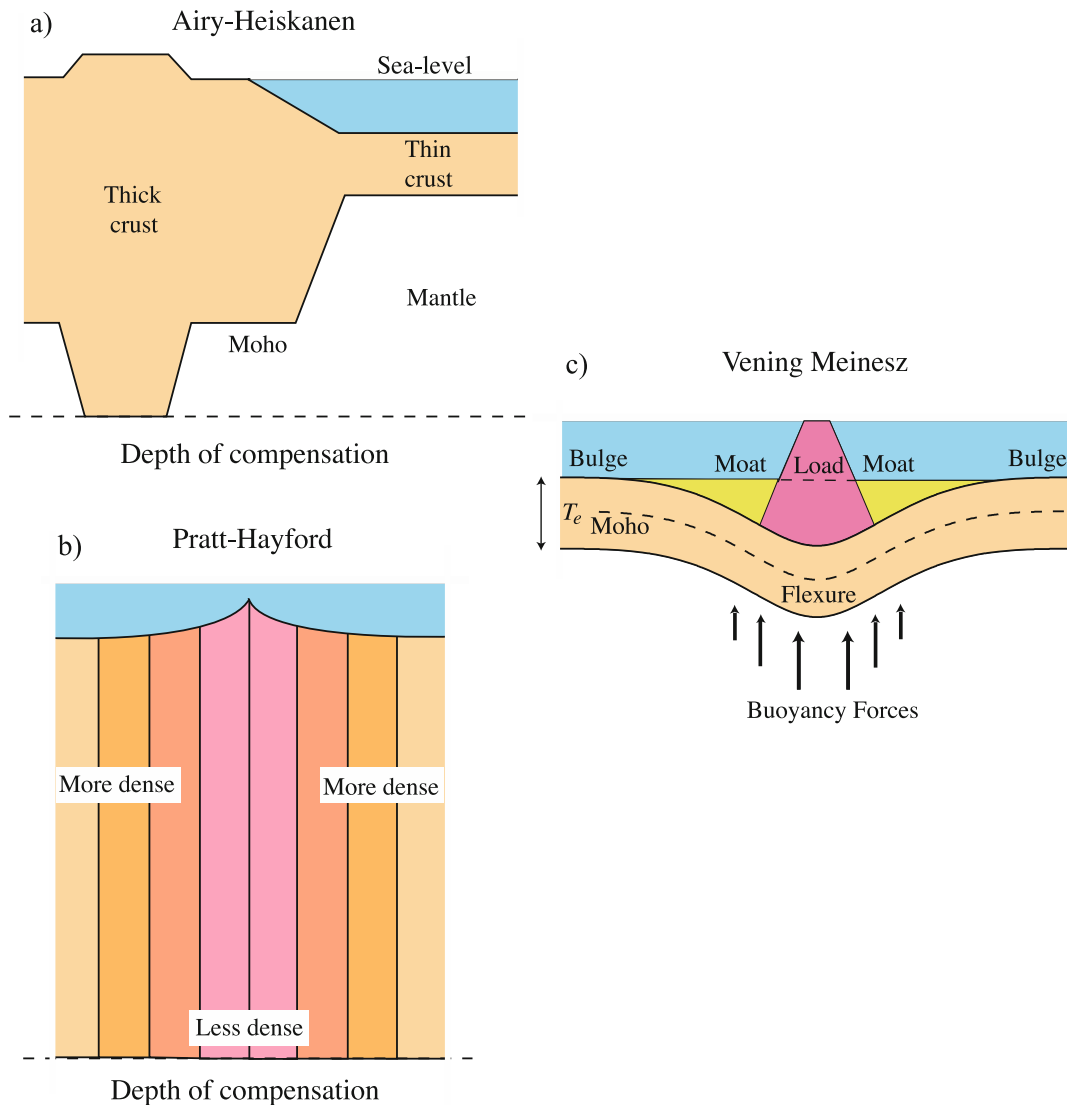
The break through that led to the formulation of the principle of isostasy came following George Everest’s pioneering geodetic work in South Africa and India. Airy (1855) and then Pratt (1855) used Everest’s deflection of the vertical data in northern India to address the question of how the Himalaya mountains were supported at depth. Fisher (1881), who favored Airy’s hypothesis that the Himalaya were supported by a deep crustal “root,” concluded that the Earth’s crust was in a state of hydrostatic equilibrium and floated on its fluid substrate, much like an iceberg on seawater.

Today, we tend to think of isostasy as a restraining force rather than a driving force like ridge push or trench pull. Isostasy is a process that acts on the topography created, for example, at plate boundaries by continental rifting, break-up, and mountain building so as to reshape it and redistribute the stresses that have built up. The existence of regions of large topographic relief, unusually thick and thin crust, and large-amplitude gravity and geoid anomalies are all pointers that isostatic equilibrium does not prevail everywhere. The main challenge in isostatic studies has been to quantify the degree to which hydrostatic equilibrium of Earth’s crust is achieved and the time and spatial scales over which it operates.

A number of isostatic models have been proposed, three of which are illustrated in Fig. 1. Airy and Pratt are local models in which changes in the Earth’s topography are supported by either change in the thickness of a uniform density crust or by lateral change in density of the crust and/or mantle. Both models imply that the crust and mantle respond to loads (e.g., volcanoes) or unloads (e.g., erosion) locally such that neighboring regions are not involved in their support or removal. While the Airy and Pratt models are a potential source of stress in the crust (Bott and Dean 1972; Bott and Kusznir 1979) and may drive extension and compression they imply an intrinsically weak crust that responds to loads and unloads locally and in some cases by faulting. Vening Meinesz, on the other hand, is a regional model in which loads and unloads are supported by flexure of a thin elastic plate and stresses accumulate by bending.

It is important to remind ourselves that the models in Fig. 1 are highly idealized and only represent the state that the Earth's outermost layers would tend toward in the absence of disturbing forces. These disturbing forces include many of the great cycles of geology such as the waxing and waning of ice-sheets, the growth and decay of volcanoes, and sedimentation and erosion. Each process is potentially a disturbing force that could act to delay or even prevent isostatic equilibrium from being achieved.

There is now a considerable body of work that indicates Earth is a dynamic planet that is continuously being deformed. Isostasy, on the other hand, is a static concept that refers to a crust and mantle that has already moved and is now in equilibrium. It therefore appears at odds with current geodynamic views. But isostatic studies of the deformation that follows, for example, megathrust earthquakes, late-glacial rebound, volcano emplacement, and sediment accumulation provide “snapshots” of the way that the crust and



Isostasy, Fig. 1 The main types of isostatic models. Each model implies a state of hydrostatic equilibrium such that the Earth's outermost layers are in a state of flotation on their more fluid substrate. Names (e.g., Airy-Heiskanen) refer to the workers who first conceived and quantified the models. Vertical (lithostatic) stresses are constant on the depth of compensation in the Airy-Heiskanen (Airy 1855; Heiskanen 1931) and Pratt-Hayford (Pratt 1855; Hayford 1909) models and vary laterally in the Vening-Meinesz model (Vening Meinesz 1931). (a) Airy model: isostatic equilibrium is achieved by variations in the thickness of a

uniform low-density crust that overlies a high-density mantle. (b) Pratt model: isostatic equilibrium is achieved by lateral changes in the density. (c) Vening-Meinesz model: topography is considered as a surface load, and isostatic equilibrium is achieved by flexural downwarping and upwarping of the crust and mantle over a broad area. The load is supported by the strength of the lithosphere (which is determined by its effective elastic thickness, T_e) and the buoyancy of the underlying asthenosphere

mantle respond to loads and unloads over certain temporal and spatial scales (Watts 2007). Together, these “snapshots” provide information on whole lithosphere behavior and the clues as to how the crust and mantle would actually deform in response to continually applied loads and unloads.

A goal of isostatic studies has been to determine the vertical motion history of the Earth’s crust and mantle on short, through intermediate to long time-scales. Therefore, isostasy is central to studies of global change, especially those that are aimed at determining the relative role of tectonics and sea-level change in controlling the position of the shoreline today and in the geological past. Most significantly, isostatic studies provide information on how the lithosphere responds to the load and unload shifts that occur during climate-driven glacial and inter-glacial cycles. We are presently in an inter-glacial, for example, yet the isostatic effects of the Last Glacial Maxima (LGM) are still clearly visible today in tide gauge and satellite Global Positioning System (GPS) data. Isostasy is a key to understanding the lithosphere on which we live and how it interacts with the asthenosphere below and the cryosphere, hydrosphere, and atmosphere above.

Concepts and Applications

The isostatic models illustrated in Fig. 1 imply different gravity anomalies and crust and mantle structures. Airy and Pratt, for example, predict that free-air gravity anomalies will generally be small over volcanoes because they are underlain by either a thick, uniform-density, crustal “root” or a constant thickness low-density crust, the negative gravitational effect of which will compete with the positive effect of the volcano. Vening Meinesz, on the other hand, predicts a shallower depth to Moho beneath a volcano than Airy or Pratt and, hence, the free-air gravity anomaly will be generally large over a volcano because its gravity effect will now dominate over that of the compensation.

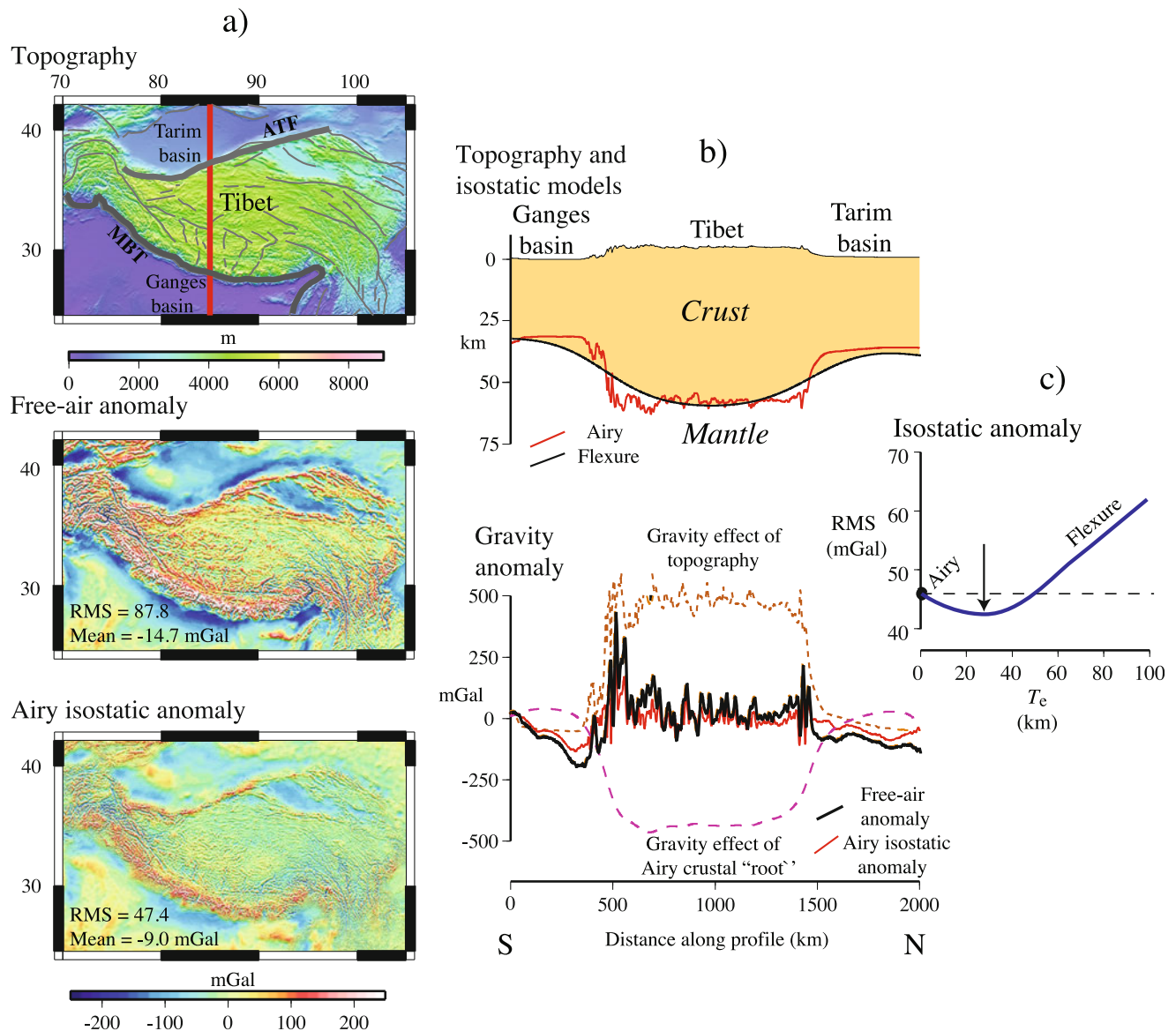
Figure 2 shows the topography, free-air, and isostatic gravity anomalies in the region of northern India where the early geodesists carried out their classical work. The isostatic anomaly has been calculated by subtracting (and thereby removing) the positive gravity effect of the topography and the negative gravity effect of the Airy compensation from the observed free-air gravity anomaly. The Root Mean Square (RMS) of the free-air gravity anomaly variation about its mean is 87.8 mGal while the RMS of the Airy isostatic gravity anomaly is 45.9 mGal. These RMS values show that the free-air gravity anomaly is substantially reduced after a correction for isostasy and so we may conclude that equilibrium prevails to a significant degree in the northern India region.

Despite this success, Fig. 2a shows that significant departures remain in the Airy isostatic anomaly map. The most

striking are the positive-negative anomaly “couples” that correlate with the edge of the Tibetan Plateau, particularly, its boundary with the Ganges and Tarim foreland basins. The negative anomalies are usually on the basin side while the positive anomalies are on the plateau side, which implies that the Moho is deeper than expected for Airy on the basin side and shallower on the plateau side. Such a Moho geometry can be explained if we assume that the load of the Himalaya and Kunlun Mountains is supported by the strength of the underlying crust and mantle which flexes over a broad region that extends beyond the mountains themselves. The Vening Meinesz model takes into account these ideas of loads and bending and Fig. 2c shows that when an elastic plate flexure model is used to compute the compensation rather than an Airy model then the RMS of the isostatic anomalies are reduced even further. A useful parameter in the Vening Meinesz model is the effective elastic thickness, T_e , since it is a proxy for the long-term strength of the lithosphere. Figure 2c shows that there is a well-developed minimum in the RMS of the isostatic anomaly for $T_e \sim 30$ km, suggesting this as the best fit T_e estimate for the region (see also Crosby 2007 who found a similar best fit T_e for the northern India region). This is, of course, only an average and so there will be areas that are stronger and areas that are weaker than this. Indeed, Braitenberg et al. (2003) and Jordan and Watts (2005) have already shown, using inverse methods, that T_e is <30 km over the central and northern Tibetan Plateau and >30 km over the Ganges and Tarim foreland basins.

Another way to test isostasy on a regional scale is to compare compilations of the seismically constrained crustal structure directly to the predictions of isostatic models. Figure 3 shows, for example, a 18,000 km long transect of the Indian, Atlantic, and Pacific Oceans along which observed and calculated depths to Moho are compared. The observed Moho is based on CRUST 2.0 (Bassin et al. 2000), a $2^\circ \times 2^\circ$ grid of P-wave velocity and crustal layer thickness data derived from active source seismic data. The calculated Moho is based on an Airy model with the same density and crustal thickness as used in Fig. 2. The figure shows that the calculated Airy Moho tracks the observed seismic Moho well, especially so at rifted margins and beneath some continents.

The main departures are in the central Atlantic Ocean where the observed Moho depth is shallower than predicted by Airy and beneath the Eastern US Appalachians the Gulf of Mexico and the Persian Gulf where the observed Moho depth is deeper. The shallower depths in the can be explained by our assumption of a relatively high subcrustal oceanic mantle density (3330 kg m^{-3}) which if reduced (e.g., to 3180 kg m^{-3}) beneath a mid-oceanic ridge would result in a shallower Moho depth and better agreement with the seismic Moho depths. The idea of a relatively low-density subcrustal mantle beneath the ridge suggests a Pratt rather than an Airy

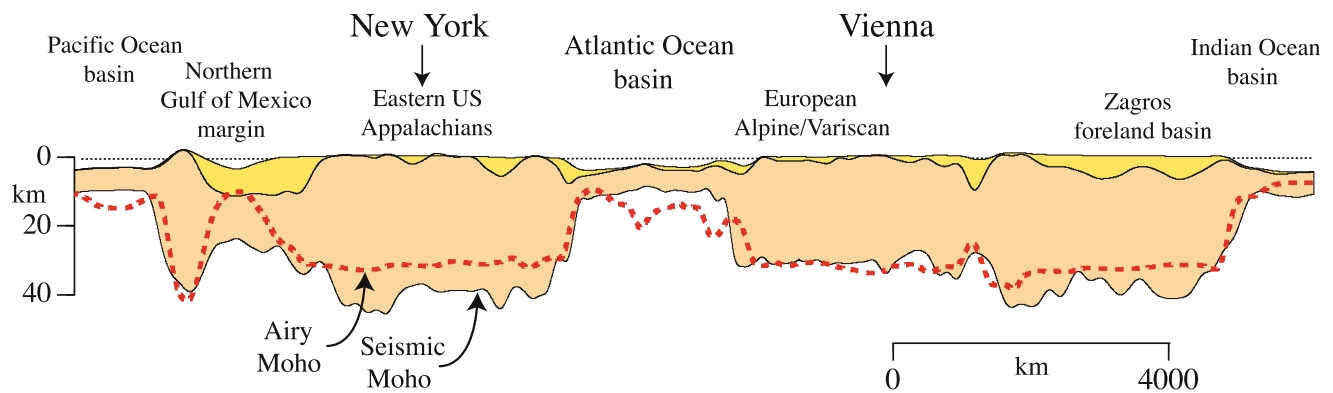


Isostasy, Fig. 2 Topography, free-air, and isostatic gravity anomalies in the region of the Tibetan plateau. (a) Upper panel: Topography based on GEBCO (2003). The map shows the plateau and the flanking Tarim and Ganges foreland basins, the major faults (ATF = Altyn Tagh Fault, MBT = Main Boundary Thrust), and the north-south profile (thick red line) plotted in (b). Middle panel: Free-air gravity anomaly based on the EGM2008 2.5×2.5 min grid (Pavlis et al. 2008). Lower Panel: Airy isostatic anomaly calculated by subtracting the gravity effect of the topography and its compensating "root" from the observed free-air gravity anomaly. Calculations assume a density of water, crust, and mantle of 1030, 2800, and 3330 kg m^{-3} , respectively, and a zero

elevation crustal thickness of 31.2 km. (b) Predicted crustal structure based on the Airy and flexure models and free-air and isostatic gravity anomaly profiles of the Tibetan plateau at longitude 85°E . (c) Root Mean Square (RMS) of the isostatic anomaly as a function of elastic thickness, T_e , which is a proxy for the long-term strength of the lithosphere. The Moho depth for $T_e = 0$ km corresponds to the predictions of an Airy model which implies a plate with no strength. The minimum RMS and hence the isostatic model that best explains free-air gravity anomaly data in the vicinity of the Tibetan plateau is for a flexure model with $T_e = 30$ km

model of isostatic compensation. Indeed, the Pratt model has been assumed in thermal cooling models (Sclater and Francheteau 1970) for the variation of seafloor depth and heat flow with age and the model may well apply to other cases of thermal isostasy, for example, at mid-plate oceanic swells (Nakiboglu and Lambeck 1985) and fracture zones (Sibuet and Veyrat-Peinet 1980).

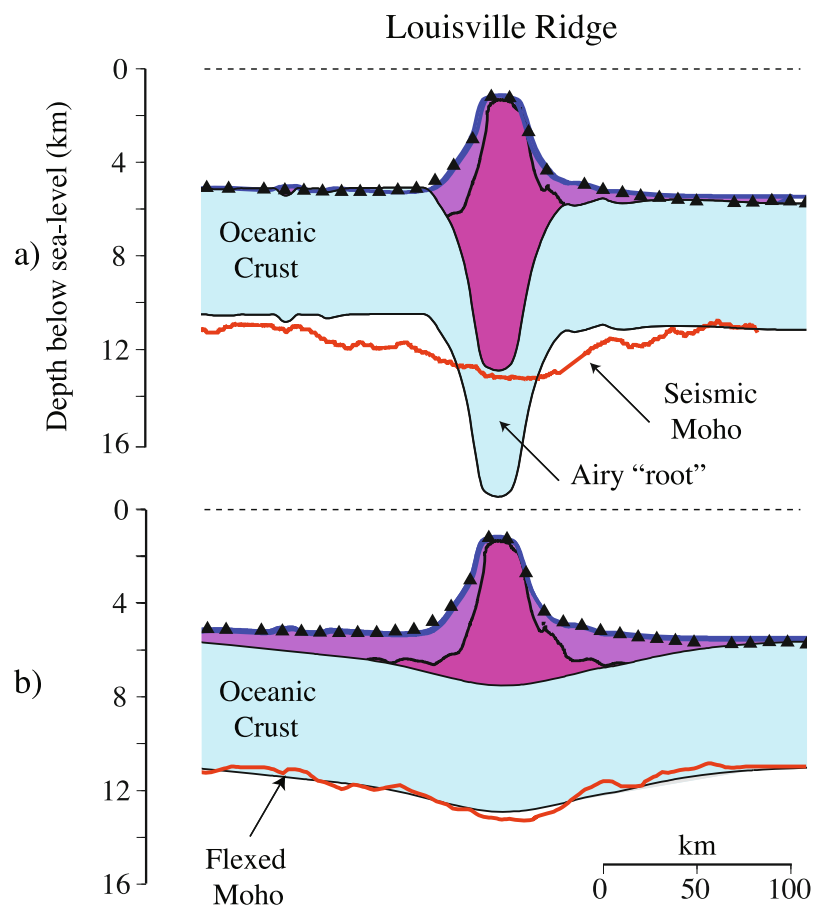
The deeper depths than predicted by Airy in Fig. 3 correlate with thick sediment accumulations in the Mississippi River Delta and the Zagros foreland basin. Here, sediment loads due to rifting and thermal contraction and thrusting and folding providing the accommodation space have flexed the Moho downwards over a broad area such that Moho is deeper than would be expected for an Airy model. The deeper depths



Isostasy, Fig. 3 Comparison of the crustal structure based on seismic refraction data to the predicted crustal structure assuming an Airy model along a 18,000 km long great circle profile that extends from the Pacific Ocean in the west across the Atlantic Ocean to the Indian ocean in the

east. The seismic refraction data is based on CRUST2.0 <http://igppweb.ucsd.edu/~gabi/rem.html> and the Airy model is based on the same parameters as assumed in Fig. 2

Isostasy, Fig. 4 Comparison of the observed crustal structure derived from seismic refraction data (Contreras-Reyes et al. 2010) to the predictions of an Airy and flexure model of isostasy along a NNE-SSW trending profile of the Louisville Ridge seamount chain at latitude 27.5° S, southwest Pacific Ocean. Black filled triangles indicate the location of the ocean bottom seismometers that were used to constrain the seismic structure of the crust. Thick solid red lines show the depth to the seismically constrained Moho. Thin solid black line separating the purple and pink filled regions show the 6.0 km/s P wave iso-velocity contour. Purple fill = volcanoclastic material. Pink fill = volcano edifice. Blue fill = deformed oceanic crust. (a) Airy model based on the same parameters as assumed in Fig. 2. (b) Flexure model with $T_e = 10$ km



beneath the Eastern US Appalachians are puzzling and may reflect a downward flexure related to thinner lithosphere, as defined by teleseismic studies (Rychert et al. 2005).

The role of flexure in contributing to crustal structure is well illustrated in Fig. 4 which shows the results of a recent seismic reflection and refraction survey of the Louisville

Ridge seamount chain near its intersection with the Tonga-Kermadec trench (Contreras-Reyes et al. 2010). The survey utilized a large-volume air gun source array and 39 ocean bottom seismometer receivers spaced ~15 km apart. The geometry was such that full ray coverage of the seamount and the underlying oceanic crust and uppermost ~10 km of

the mantle was achieved. Figure 4a shows that the seismic Moho (which is resolved here to ± 100 m) cannot be explained by an Airy model which predicts too deep a Moho beneath the seamount chain and too shallow a Moho in flanking regions. The best fit between observed and calculated Moho depths (Fig. 4b) is for a flexure model of isostasy with $T_e = 10$ km. Such a model explains well the observed depth to Moho beneath *both* the seamount edifice and flanking moat regions.

Current Investigations

One of the most important results to have emerged from isostatic studies over the past 40 years has been to show that a link exists between the T_e derived from flexural modelling and the long-term strength of the lithosphere. Data from experimental rock mechanics, for example, suggest that the strength of oceanic lithosphere is limited by brittle deformation in its uppermost part and by ductile flow in its lowermost part (Goetze and Evans 1979). This strength “profile” implies that loads applied to the surface (or base) of the lithosphere will be supported partly by the brittle and ductile strength of the lithosphere and partly by a central “core” which deforms elastically. It has been shown that oceanic T_e reflects the thickness of this “core” which will be small in regions of young crust, large loads, and high curvature of bending and high in regions of old crust, small loads, and low curvature (Watts and Burov 2003). Oceanic T_e is in the range 0–40 km and since the oceanic crust is only ~ 8 km thick, we may conclude that not only the crust but also the mantle is involved in supporting long-term geological loads.

Figure 5 shows that T_e at seamounts and ocean islands and deep-sea trench and outer rise systems increases with the age of the lithosphere at the time of loading, being small at volcano and trench loads that form on young seafloor (e.g., on-ridge) and large at loads on old seafloor (e.g., off-ridge). Thus, as oceanic lithosphere cools, subsides, and increases in age following its creation at a spreading center, it becomes more rigid in the way that it responds to loads. Interestingly, most seamount and oceanic island values fall within the 300–600 °C isotherm envelope based on the cooling plate model of Parsons and Sclater (1977), while most trench and outer rise values require a higher controlling isotherm in the region 450–800 °C. This difference in the range of controlling isotherm can be explained by difference in the load age, which is small (essentially present day) at deep-sea trench-outer rises and large (~ 1.5 to 100 Myr) at seamounts and oceanic islands. Indeed, a load age dependency explains much of the scatter in plots of T_e vs. age of the lithosphere at the time of loading. Other possible contributors are uncertainties in load, infill and mantle densities, thermal

perturbations due to hot and cold spots, and yielding in regions of large loads and high curvature.

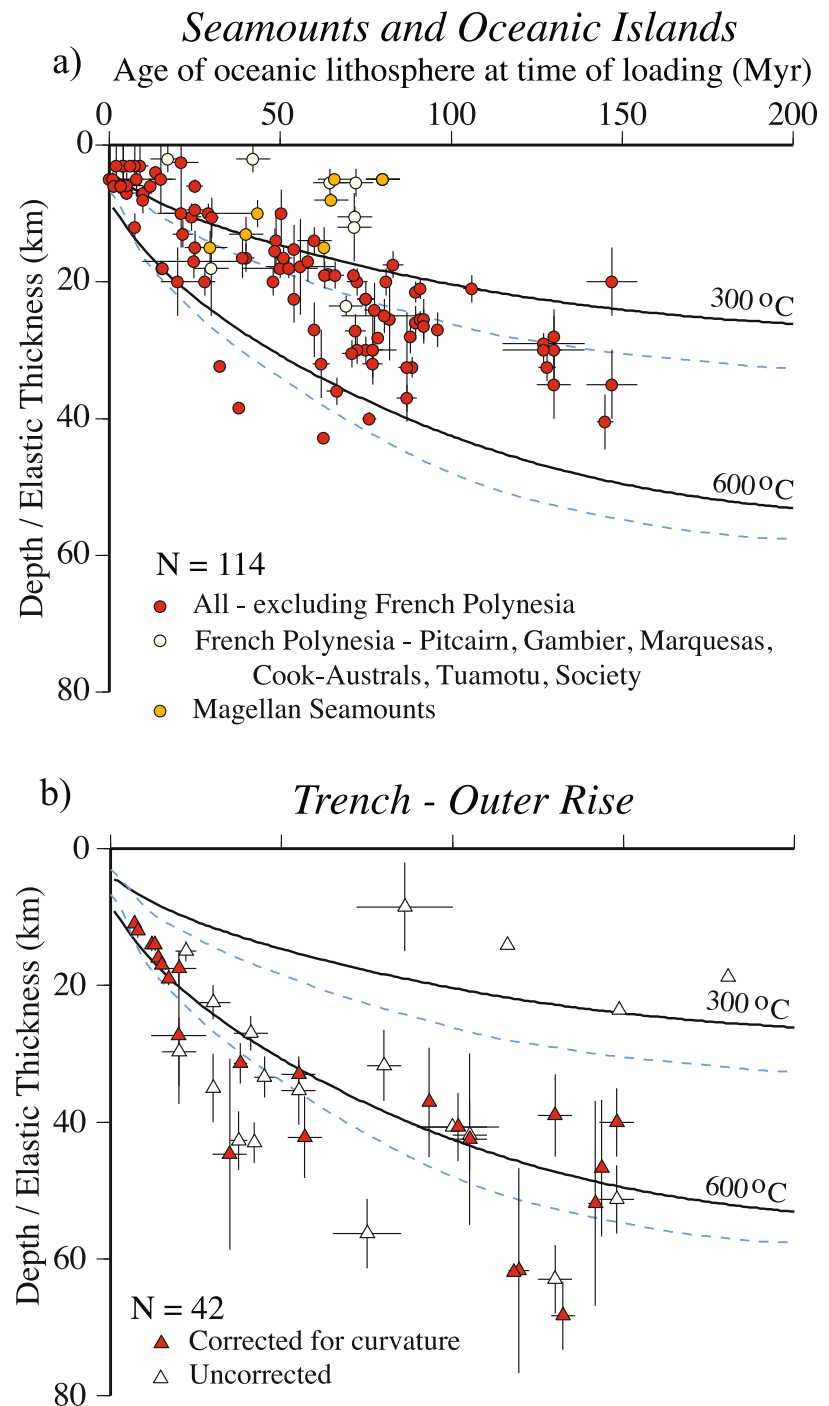
Most of the T_e estimates plotted in Fig. 5 are based on a forward modelling approach in which the gravity anomaly, surfaces of flexure, or vertical motion history are calculated and compared to observations. The best fit T_e is then determined by a “trial and error” comparison of the observed and calculated. While forward modelling is a satisfactory way to estimate T_e , the number of sites where information on both load and plate age are available is limited.

Technological improvements in satellite radar altimetry and shipboard multibeam (swath) bathymetry data acquisition systems have led to the development of new analysis techniques and an increase in T_e estimates. Watts et al. (2006), for example, used satellite-derived gravity data together with a reciprocal admittance technique to predict bathymetry for different values of T_e . By comparing observed and calculated bathymetry, these workers obtained 9758 individual estimates of T_e , 291 of which were at sites where both load and plate age were known. They found a greater than expected scatter in global plots of T_e vs. age of the lithosphere at the time of loading with no clear indication of a controlling isotherm. Kalnins and Watts (2009) used satellite-derived gravity and shipboard bathymetry data and a sliding window admittance technique and found the controlling isotherm varied regionally. This may help explain some of the scatter in global plots of T_e against load and plate age. Bry and White (2007) used satellite-derived gravity and shipboard bathymetry data and an inversion technique to determine T_e along a number of fixed length profiles of deep-sea trench outer rise systems and found no evidence of a controlling isotherm. Hunter and Watts (2016), however, used a similar circum-Pacific data set and inverse modelling approach and found a strong dependence of T_e and plate age. They attributed the difference with Bry and White (2007) to a combination of factors which include: the manner of the long wavelength gravity field removal; the method used to calculate gravity anomalies; the constraint placed on bending moment; and the location of the trench axis.

One problem with spectral methods is the influence of window size on the recovered T_e (Crosby 2007). Windows need to be large enough to resolve the largest T_e and small enough to avoid “contamination” from more than one type of tectonic setting (e.g., on-ridge, off-ridge).

The physical meaning of T_e and its relationship to load and plate age is less clear in the continents than it is in the oceans. Unlike oceanic lithosphere, continental lithosphere has a multilayered rheology (Burov and Diament 1995; Meissner and Strehlau 1982) and there may be more than one brittle and ductile layer that supports an individual load or unload. In this situation, T_e is not a particular depth in the lithosphere but a measure of its integrated strength. Continental T_e will depend on both the curvature of flexure (which depends in turn on the

Isostasy, Fig. 5 Plot of T_e vs. age of the oceanic lithosphere at the time of loading. (a) Seamounts and oceanic islands and (b) Deep-sea trench and outer rise systems. The data is based on Table 6.1 in Watts (2001) with additional values from Freedman and Parsons (1986), McQueen and Lambeck (1989), and 15 other studies that have been carried out since 2000. Solid and dashed lines show the calculated depth of the 300 °C and 600 °C isotherms based on the cooling plate models of Parsons and Sclater (1977) and McKenzie et al. (2005), respectively



size of the load or unload) and how well individual competent layers are coupled, being low for high curvature and poorly coupled systems and high for low curvature and well-coupled systems (Watts and Burov 2003).

As in the oceans, continental T_e has been estimated using both forward and spectral modelling techniques. Most forward modelling estimates are based on fitting either the gravity anomaly or the surfaces of flexure (as inferred, for example, from basement depth data) at rift-type, strike-slip,

intra-cratonic, and foreland basins to model predictions. Sometimes, both gravity anomaly and depth to basement data are available and this yields most reliable estimates (Haddad and Watts 1999). Continental T_e estimates range from 5–100 km and no simple relationship has yet been found between T_e and load and plate age (Watts 2001). The data is suggestive, however, of some age dependence. Archaean and Proterozoic cratons, for example, are generally associated with high T_e values (>70 km) while Phanerozoic

and younger mountain belts generally have lower values (25–60 km) (Pérez-Gussinyé and Watts 2005). Young rifted margins (e.g., South China Sea) appear to have lower T_e values than old margins (e.g., Amazon) (Watts et al. 2009) and this has been interpreted as the result of heating during continental break-up, which weakens the lithosphere, and cooling following rifting that strengthens it (Burov and Poliakov 2001). Plume-influenced rifts (e.g., Afar, North Atlantic) are associated with the lowest T_e values (Ebinger et al. 1989; Watts and Fairhead 1997) but it is not clear how much strength the weakened rifted lithosphere in such tectonic settings is capable of regaining following a rifting event.

According to the Wilson Cycle, ocean basins open and close and compressional structures in mountain belts develop on or close to the site of extensional rifted margins. Isostatic studies of the foreland basins suggest that they may inherit the T_e structure of the underlying rifted margin (Lin and Watts 2002; Ali and Watts 2009). The foreland basin may be either narrow or wide, depending on whether it formed on a young or old margin. The very high (>70 km) T_e values that are associated with some foreland basins (e.g., central Andes) may be explained by the fact that the fold and thrust belts, the main driving load for foreland basin subsidence, will sometimes over-ride the rifted margin altogether and “telescope” onto the more rigid cratonic interior, thereby inheriting its value.

As in the oceans, spectral approaches have provided new insights into continental T_e , especially its spatial and temporal variation. The first studies (e.g., Dorman and Lewis 1970; Banks et al. 1977; McNutt and Parker 1978) were based on a Bouguer admittance, Fourier transform periodogram method, which generally yielded low values. Forsyth (1985), however, pointed out that the low values might be a result of subsurface (i.e., buried) loads that were uncorrelated with surface topographic loads. Subsurface loads are difficult to define and so he suggested using the Bouguer coherence, rather than the Bouguer admittance, because it was less sensitive to the ratio of surface to subsurface loading. Forsyth’s method has been widely applied (e.g., Zuber et al. 1989) and as Bechtel et al. (1990) showed for North America, it yields low as well as high values, especially in cratonic regions. Subsequent studies have incorporated new techniques, including maximum entropy estimators (Lowry and Smith 1995) and multitapers (McKenzie and Fairhead 1997). McKenzie and Fairhead (1997) recommend using the free-air admittance rather than the Bouguer coherence because the surface topography, unlike buried loads, is a known load, the gravity effect of which should not really be removed from the free-air gravity anomaly. Forsyth’s Bouguer coherence technique they argued maybe biased by the effects of erosion and his method will

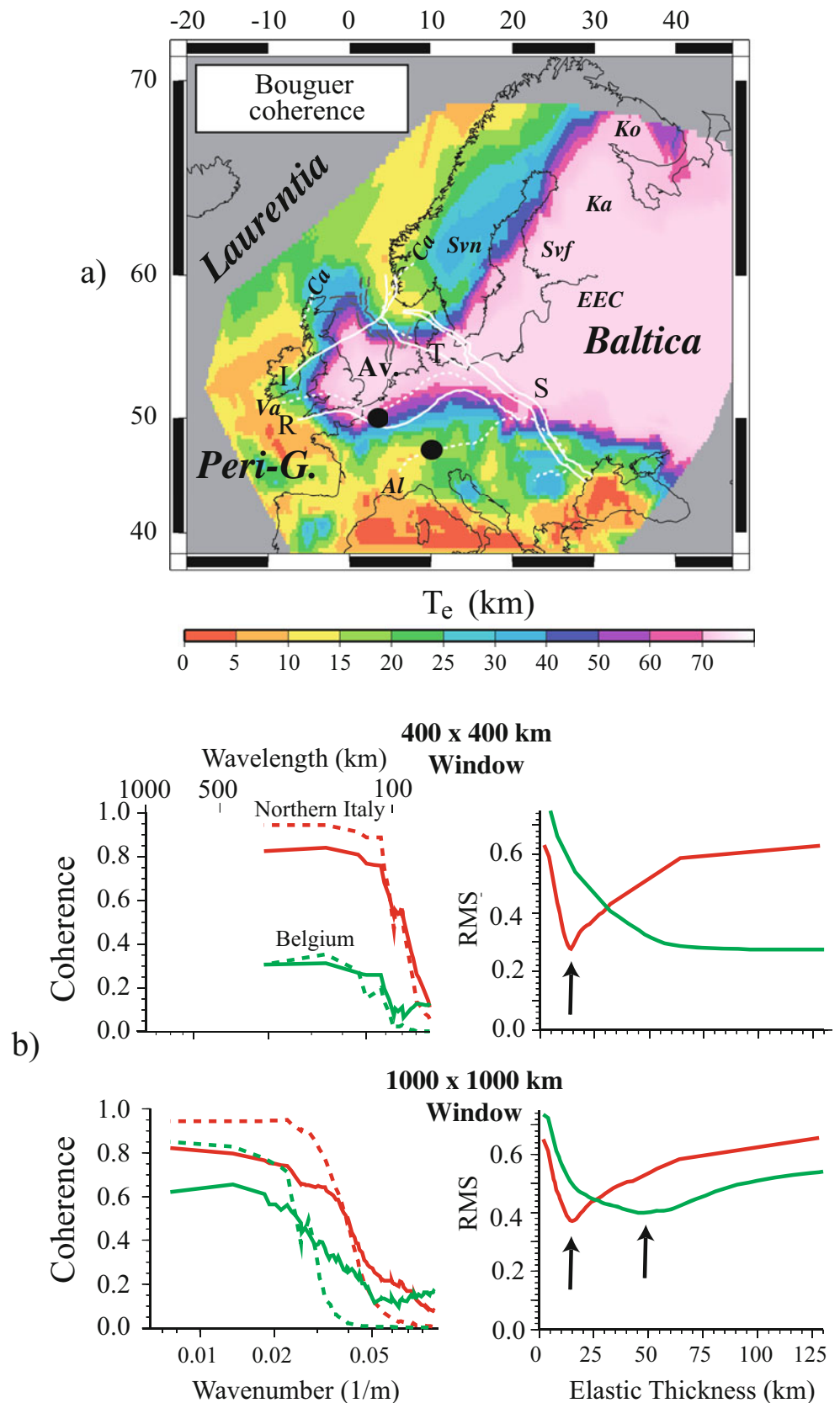
yield over-estimates of T_e rather than true values. Pérez-Gussinyé and Watts (2005) used both Bouguer coherence and free-air admittance methods and found the highest values in Europe (>70 km) over the Archaean and Early Proterozoic (<1.5 Ga) East European craton (which is part of Baltica) and the lowest values (10–45 km) over the flanking Caledonian, Variscan, and Alpine orogenic belts (Fig. 6). Similar results were found for North and South America (Kirby and Swain 2009; Pérez-Gussinyé et al. 2007), Australia (Simons et al. 2000), and Africa (Pérez-Gussinyé et al. 2009) and a consensus seems to be emerging that the continental lithosphere is capable of showing both great weakness and great strength. Interestingly though, cratons are not always associated with a uniformly high T_e , and there exist regions within them that have low values.

In another development, wavelets have been used instead of spectral methods. Wavelets have an advantage over spectral methods in that they are local in space and scale so that the relationship between gravity anomaly and topography can, in principal, be examined over a range of wavelengths without the need for annular averaging over discrete windows. They generally yield similar results as spectral methods, although the recovered values of T_e tend to be smoother (Audet and Mareschal 2007; Sacek and Ussami 2009; Stark et al. 2003; Tassara et al. 2006).

Irrespective of the methods used, caution needs to be applied when interpreting continental T_e data. This is because it is not clear whether the T_e deduced from spectral studies refers to the same loading and unloading schemes as are assumed in forward modelling. In forward modelling, it is possible to discretize the loads (e.g., fold and thrusts) and flexures being considered, and the T_e deduced usually refers to a particular thermo-tectonic event in the geological past (e.g., a rift or orogenic event). Spectral methods, by way of contrast, “see” *all* the loads that are acting on the lithosphere, and these include not only loads in the past but sedimentary loads and erosional unloads, some of which continue to shape Earth’s surface topography up to the present day.

What is clear from all these studies is that Earth’s lithosphere has a rich T_e structure. In the oceans, T_e varies by up to ± 15 km over horizontal distances of 500–1000 km (Watts et al. 2006). Spatial variations are even larger in the continents. The compilations of Pérez-Gussinyé and co-workers, for example, suggest ± 25 km over similar distances. In many places (e.g., western Canada and USA), weak regions with T_e as low as 10 km abut strong regions with $T_e > 70$ km (Fluck et al. 2003). The consequences of such strength variations are profound, especially as they impact the computation of the isostatic gravity anomaly (Kirby 2019), the patterns of glacial isostatic rebound (Whitehouse et al. 2006), the solid earth tide

Isostasy, Fig. 6 The T_e structure of Europe as derived from the coherence between the observed Bouguer gravity anomaly and topography as a function of wavelength. Modified from (Pérez-Gussinyé and Watts 2005). (a) T_e and its relationship to the terrane structure. Thin white lines show major sutures. I = Iapetus, R = Rheic, A = Alpine. T = Thor. S = Sorgenfrei-Tornquist. The tectonic provinces are Ko, Kola; Ka, Karelia; and Svf, Svecofennian. Thin dashed lines show the Caledonian (Ca), Variscan (Va), and Alpine (Al) deformation fronts. Peri-G. = Peri-Gondwanaland. Av = Avalonia. EEC = East European Craton. Black filled circles show the location of the coherence and RMS plots in (b). (b) Examples of the observed and predicted Bouguer coherence and RMS between the observed (solid line) and calculated (dashed line) coherence for a 400×400 km and 1000×1000 km analysis windows centered on Belgium and northern Italy. Note that northern Italy appears as a region of low T_e in both windows, but the high T_e region of Belgium only has a clear RMS minima in the larger window



(Mantovani et al. 2005), and the terrane structure of continental lithosphere (Pérez-Gussinyé et al. 2007).

Current Controversies and Gaps in Knowledge

Isostatic studies have never been far from controversy as is evidenced by the vigorous debates on local vs. regional models that took place at the turn of the twentieth century. Central to these debates was the applicability of isostatic models to particular geological features and certain “schools of thought” emerged among geologists, geophysicists, and geodesists. There were even debates either side of the Atlantic when an Airy model was preferred by Europeans and a Pratt model by North Americans. Today, there is now general agreement on the relative role of local and regional models of isostasy. Debate is focused instead on the flexure model, particularly T_e , what it means and what it tells us about the long-term strength of the lithosphere, the time-scales of isostatic adjustment, and the geological processes that shape the top and base of the lithosphere.

The Long-Term Strength of the Lithosphere

In the oceans, there is widespread agreement on oceanic T_e and what it indicates about the long-term strength of the lithosphere. While there is scatter in plots of oceanic T_e vs. plate age, most of it can be attributed to viscoelastic stress relaxation (Watts and Zhong 2000), curvature and yielding (McNutt and Menard 1982), and spatial variations in the controlling isotherms that determine T_e (Kalnins and Watts 2009).

There has not, however, been the same agreement concerning continental T_e estimates and their relationship to age. The controversy came to a head in the so-called *crème brûlée* vs. *jelly sandwich* debate (Burov and Watts 2006). Proponents of the *crème brûlée* model follow McKenzie and Fairhead (1997) and Jackson (2002) who believe that because T_e is small and similar in thickness to the seismogenic layer, the strength of the lithosphere resides mainly in the uppermost part of the crust. In their view, the lower crust and uppermost mantle are weak which explains why deep earthquakes are rare. Proponents of the *jelly sandwich* model, however, follow Meissner (1986) who argue that the strength resides mainly in the upper crust and the uppermost mantle. In his view, only the lower crust is weak. This model is compatible with the results of seismic reflection profiling which show a layered lower continental crust and with the results of flexure studies which show a wide range of T_e values. Low T_e is indicative of thin competent layers while high T_e values reflect thick competent layers. Very high T_e values (>70 km) are also permissible, depending on composition, and occur when the upper and lower crust and/or the lower crust and mantle are coupled. Deep earthquakes are rare, not because the mantle is

too weak, but because it is too strong and the regional tectonic stresses required to initiate faulting and earthquakes are not large enough.

The debate has led to a reexamination of the methodologies used to estimate continental T_e , the physical meaning of T_e and the geological implications of a weak or strong mantle. Pérez-Gussinyé et al. (2004), for example, showed that it is necessary when comparing observed and calculated free-air admittances to first window them using the same multitaper estimator while Crosby (2007) showed that in order to recover T_e in cratonic shield regions, large window sizes (up to 1500 km) are needed. Otherwise, T_e maybe underestimated. Watts and Burov (2003) pointed out that while T_e may well be low and similar to the seismogenic layer thickness in some tectonic settings (e.g., Basin and Range, western USA), the two parameters are not the same. T_e reflects the integrated strength of the lithosphere while the seismogenic layer is the thickness of the uppermost part of the crust that responds to stresses by faulting and earthquakes. Finally, Handy and Brun (2004) argued that a strong uppermost mantle is required in order to explain the integrity of sinking slabs in subduction zones and the deep structure of rifted margins and orogens.

The continental T_e debate has proved a useful discussion point in isostatic studies. It highlights the importance of T_e while at the same time illustrates gaps in our knowledge about the Earth's lithosphere, especially regarding its long-term rheology. Other gaps are:

Time-Scales of Isostatic Adjustment

Isostasy varies temporally, as well as spatially. Exactly how slow or fast isostasy operates is of fundamental importance to our understanding of the Earth's vertical motion history as well as to topics such as sea-level change. For example, Bloom (1967) pointed out that the height of a wave-cut notch above present day sea-level depends not only on the amplitude of sea-level change but how quickly isostatic equilibrium is achieved. If isostasy is fast, then a sea-level rise will have time to load the crust and upper mantle and so the height of the notch will be low due to regional subsidence. If, on the other hand, isostasy is slow then there will be time and a notch will be cut at a higher elevation.

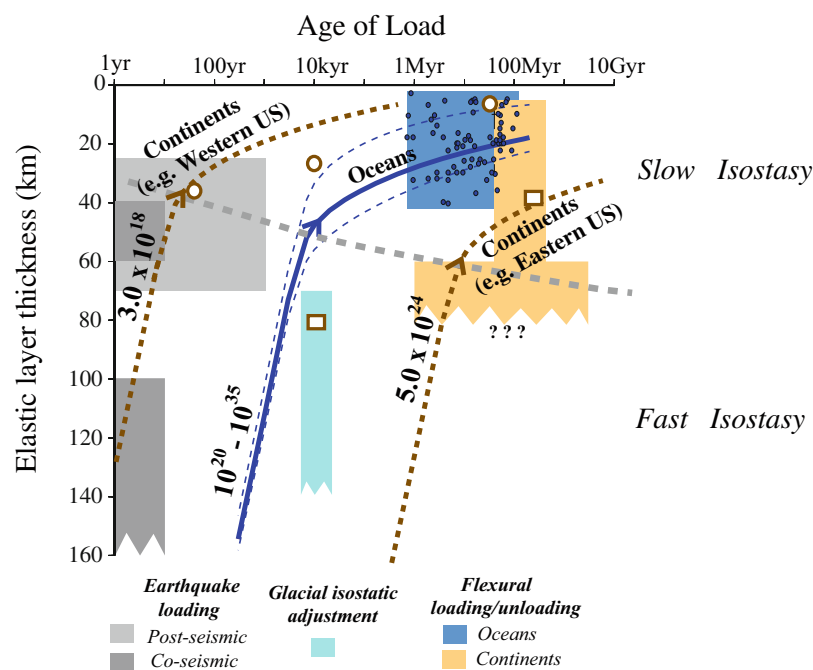
Several attempts have been made to estimate the time scales of isostatic adjustment associated with load shifts on the top or base of the Earth's crust. Watts (1978), for example, compared the T_e derived from oceanic flexure studies to the thermal and seismic thickness of the lithosphere and concluded that on loading there must be a rapid reduction in the thickness of the mechanical layer that supports a load as the lithosphere relaxes from its short-term seismic to its long-term elastic thickness. Beaumont (1979) used an elastic-plastic model based on the creep laws of olivine to argue that while oceanic lithosphere was essentially elastic on long-time scales, continental lithosphere would continue to relax on

time-scales of up to several tens of Myr. Willett et al. (1985) used a viscoelastic plate that overlies an inviscid substrate to model continental T_e data and showed that on loading there would be a weakening of the lithosphere that reduced the thickness of the mechanical layer by as much as a factor of 3–8, depending on the initial thermal structure of the lithosphere that was assumed. And finally, Watts and Zhong (2000) and Watts et al. (2013) have used the actual temperature structure of the oceanic lithosphere constrained from seafloor depth and heat flow data to calculate a viscosity profile and then examined how the equivalent T_e would vary following volcano loading. They showed that because viscosity increases from low values in the lower lithosphere to high values in the upper lithosphere, the lithosphere relaxes from the bottom-up such that stresses migrate upwards and the reduction in the thickness of the mechanical layer that supports a surface or subsurface load is initially fast and then slows with increase in the load age.

Figure 7 compares the T_e for a number of seamounts and ocean islands in the world's ocean basins to predicted curves based on a multilayered viscoelastic model. The model

assumes an asthenosphere viscosity of 10^{21} Pa s and a lithosphere viscosity that increases from 10^{21} Pa s to 10^{35} Pa s, depending on thermal age. The figure shows a rapid decrease in the thickness of the lithosphere that supports a load and hence a decrease in its strength. This weakening results in an increase in subsidence and a narrowing of the zone of subsidence with increase in load age. Eventually, the strength reduction slows and the thickness that supports a load approaches a steady state value.

The continents display a wider range of load and unload ages than the oceans and elastic thickness estimates have now been published for short-term (a few s to a few hundred s) earthquake loading (2–100 s) through intermediate (~10 ka) glacial isostatic rebound to long-term (>1 Ma) flexure. There are regions (e.g., western and eastern USA) where elastic layer thickness has been estimated for two or more timescales. These data suggest that the relaxation observed in oceanic lithosphere may also characterize continental lithosphere, albeit over a much wider range of relaxation time-scales. For example, the Hebgen Lake (Nishimura and Thatcher 2003), Lake Bonneville (Nakiboglu and Lambeck 1983),



Isostasy, Fig. 7 Plot of elastic layer thickness vs. age of load for oceanic and continental lithosphere. Modified from Watts (2007). The thick solid blue line and thin dashed blue lines show the predicted change in elastic layer thickness for young, intermediate, and old oceanic lithosphere according to Watts and Zhong (2000). Filled blue circles show the data from individual seamount and ocean island loads which range in age from ~1–100 Myr. The data indicate that by ~10 kyr oceanic loads have undergone most of their relaxation. The thick dashed brown lines show the predicted change in elastic thickness for continental lithosphere. Filled grey rectangles are based on earthquake loading (post-seismic and co-seismic), filled blue rectangles are based on glacial isostatic

adjustment and filled beige and yellow rectangles are based on continental flexure studies (see Watts 2007 for references). Open symbols show elastic layer thickness estimates from the same tectonic region (circles = western USA, squares = eastern USA) based on the data in Nakiboglu and Lambeck (1983), Di Donato et al. (2000), Lowry et al. (2000), Nishimura and Thatcher (2003), and Pazzaglia and Gardner (1994). The data suggest two modes of relaxation that correspond to “fast” and “slow” isostasy. Interestingly, the relaxation time separating these modes varies between oceans and continents with western USA relaxing much quicker (~10 yr) than the oceans and eastern USA much slower (~70 Myr)

and western Cordillera (Lowry et al. 2000) loads which reflect ages of 0–120 s, 10 ka, and ~60 Ma are associated with elastic layer thicknesses of 38, 30, and 5 km, respectively. T_e therefore decreases with increase in load age. The decrease is similar in form to the ocean loading cases modelled by Watts and Zhong (2000) and Watts et al. (2013), but the viscosity of the asthenosphere required is less and $\sim 3 \times 10^{18}$ Pa s. This viscosity results in a rapid initial and a slow final weakening of the lithosphere and a transition between the two weakening regimes at load ages of ~30 a. Interestingly, the transition age appears to vary regionally such that loads on tectonic western North America relax quicker than those emplaced on the more stable parts of eastern North America (Watts et al. 2013).

Isostasy and Landscape Evolution

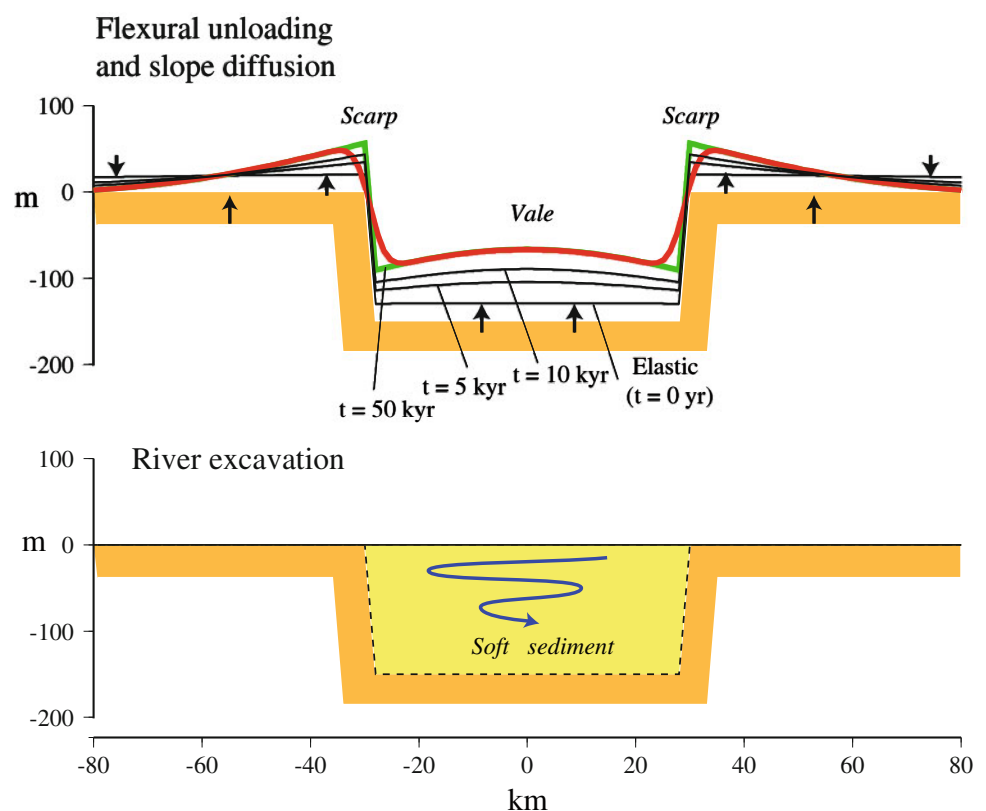
The behavior of the lithosphere and how it interacts with the cryosphere, hydrosphere, and atmosphere above and the asthenosphere below is a fundamental question that needs to be addressed if we are to fully understand geological processes and construct realistic models. We know, for example, from flexure studies that the lithosphere behaves as a low-pass filter in the way that it responds to loads and unloads, suppressing the short-wavelength deformation and passing the long-wavelengths. The physical properties of the lithosphere therefore smooth out the effects of surface and subsurface processes such as those involved in landscape evolution

and mantle convection, making it difficult to see the full effect of these processes.

It is being increasingly recognized that isostasy plays a major role in the development of landscapes and seascapes, both in the constructive forces such as tectonics that build them and the destructive forces such as erosion and large-scale slope failures that destroy them. Examples include the flexures that follow mechanical unloading on normal faults (ten Brink and Stern 1992), fault scarp retreat (Gilchrist and Summerfield 1990), and glacial erosion (Pelletier 2004; Stern et al. 2005). Because flexure involves rock uplift and subsidence it is capable of modifying the landscape, especially river knickpoints, base levels and pre-existing drainage basins. This is particularly well seen in the rim uplifts that flank rifts (Tucker and Slingerland 1994) where flexure is a major factor in controlling the drainage divide that separates scarp and dip streams. Other examples have been described from the Cotswold Hills, southern England (Lane et al. 2007) where rim uplifts flank regions of fluvial incision and excavation.

Figure 8 illustrates the isostatic adjustment that would occur following fluvial incision and excavation of a 150 m deep, 60 km wide, region of relatively low-density sediments. The calculations assume that the lithosphere behaves as a viscoelastic plate, similar to that implied by the T_e data in Fig. 7 and used by Watts and Zhong (2000) to calculate the time-dependant flexure that follows seamount loading. The

Isostasy, Fig. 8 Simple model for flexure of the lithosphere due to river incision and excavation. The model assumes that the lithosphere responds to loads (and unloads) initially as a thick elastic plate ($T_e = 80$ km), then as a viscous material and finally as a thin, essentially elastic, plate ($T_e = 10$ km). The response shows an initial long-wavelength, low-amplitude uplift a transient short-wavelength high-amplitude uplift and subsidence and a final short-wavelength high-amplitude uplift and subsidence that together reflects the thinning of the lithosphere from its initial short-term thickness to its long-term elastic thickness. Thick solid red line shows the final topography after it has been modified by slope diffusion (assumed erosion time = 50 ka, diffusion constant = $50 \text{ m}^2 \text{ a}^{-1}$)



main model features are an initial thick-plate elastic response, a viscous time-dependent response, and a final thin-plate, essentially elastic, response. The figure shows that lithospheric weakening due, for example, to load-induced stress relaxation has the capacity to create topography that locally either enhances or subdues the more regional deformation associated with the initial short-term response. Karner (1984) showed that topography is also created during lithospheric strengthening when, for example, the oceanic lithosphere is thermally rejuvenated (and weakened) at midplate hotspot swells and subsequently cools. The topography in this case subdued the original swell.

Incision and excavation, together with refilling, are common features of the fluvial response to glacial and inter-glacial cycles. Isostasy will therefore have a major role to play in the deformations that result. For example, incision due to melt-water charged rivers is a feature of glacial periods when sea-level was low while valley filling and delta construction are features of inter-glacials when sea-level was high. Repeated episodes of incision and valley fill will therefore re-surface a landscape, creating in the process isostatic rim uplift and subsidence, as seems to have occurred in the region of the Mississippi River where it debauches in to the Gulf of Mexico (Blum et al. 2008). One consequence of this resurfacing is that apparent sea-level may vary along-strike the northern Gulf of Mexico rifted margin such that the Mississippi River valley in Texas is presently subsiding relative to the adjacent Alabama coast.

The Earth's lithosphere is deformed not only by surface geological loads and unloads but also by subsurface loads, some of which originate within the crust while others have their source in the subcrustal mantle. Like surface loads, subsurface loads are associated with mass excesses and deficiencies and create flexures of the Moho and other crustal layers that are manifest in topography and gravity anomaly data. For example, subsurface loads due to sinking down-going slabs, intra-crustal thrusting, obduction, and ophiolite emplacement cause flexures that contribute to both topography and gravity anomalies (e.g., Royden and Karner 1984; Lin and Watts 2002; Ali and Watts 2009).

Long-Wavelength Gravity and Topography Anomalies, Isostasy, and Mantle Dynamics

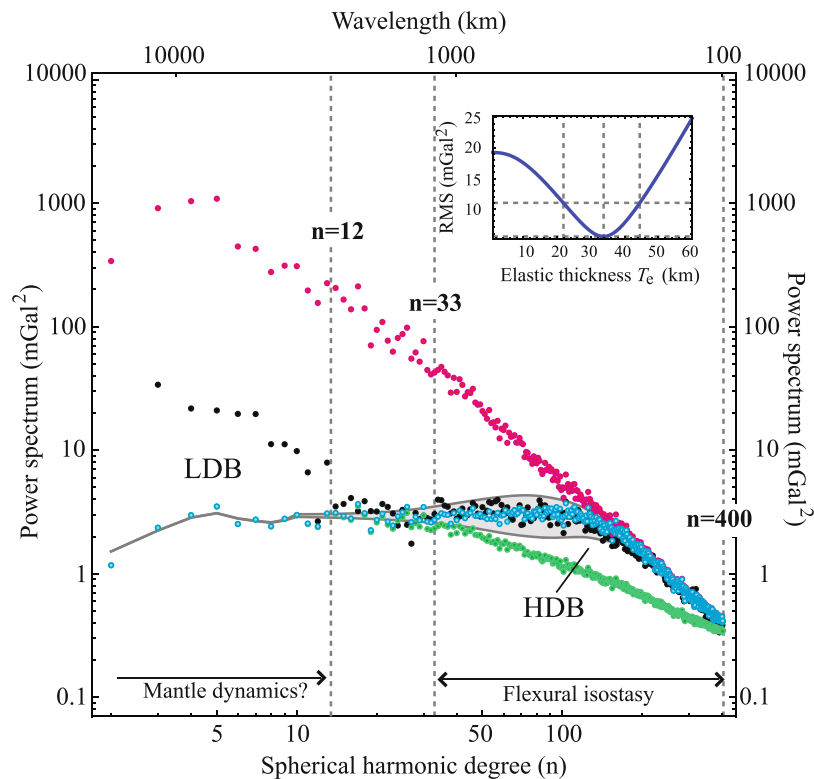
It has been known since the pioneering studies of Kaula (1967) that Earth's gravity field spectrum is rich and has significant power at short as well as long wavelengths. Isostatic loads and unloads, however, are compensated so that their associated gravity anomaly will be reduced and may approach zero at the longest wavelengths. This is clearly seen in the theoretical free-air admittance which shows that Airy, Pratt, and flexure models have a distinct "roll-over" that approaches zero at the longest wavelengths because of isostasy (Watts 2001). The existence in oceanic regions of

observed admittances in the range 5–20 mGal km⁻¹ at wavelengths of 1000–4000 km (Crosby 2007; Watts 2007; Wicczorek 2007) is therefore of much geodynamic interest. Watts (2007) has shown that the long-wavelength admittance remains approximately constant at ~18 mGal km⁻¹ in the central Pacific Ocean as the analysis window size is increased from 1000 to 4000 km. He argued that the long-wavelength admittance cannot be attributed to isostasy and so must reflect dynamic processes in the subcrustal mantle.

As McKenzie (2010) has shown, continents, like the oceans, are characterized by a high admittance at long wavelengths. The higher values he obtained (35–50 mGal km⁻¹) are expected since topography displaces air in the continents rather than water thereby increasing the contribution of the gravity effect of the topography. The observation of high, positive, admittances at long wavelengths in both oceans and continents is interesting, especially as it suggests a degree of correlation between gravity and topography at these wavelengths.

A correlation between long wavelength gravity and topography was noted by Kaula (1967) and as McKenzie (1977) points out, makes a strong argument for some form of mantle dynamics. Density-driven convection, for example, has the potential to contribute to the free-air gravity anomaly and the surface topography in both upwelling and downwelling regions. Long-wavelength topography that cannot be attributed to isostasy has been dubbed a "depth anomaly" by Menard (1973) and "dynamic topography" by (Hager et al. 1985) and convection maybe a significant contributor to the depth of mid-plate swells (Watts 1976), old (Huang and Zhong 2005) and young (Marks et al. 1990; Crosby and McKenzie 2009) ocean floor, the subsidence and uplift history of foreland-type (Liu and Nummedal 2004), rift-type (Muller et al. 2000), and some intra-cratonic sedimentary basins (Coakley and Gurnis 1995). An outstanding question is how best to separate the effects of isostasy from gravity anomaly and topography data and isolate the planform of motions in Earth's mantle.

Some progress on the problem has now been made by better defining the wavelength band of flexural isostasy. Figure 9 shows, for example, a plot of the power spectral density (in mGal²) of Earth's free-air gravity anomaly as a function of spherical harmonic degree and equivalent wavelength. The figure shows the power spectra increases at short wavelengths (<400 km), flattens at intermediate wavelengths (400–1100 km), and then rises again at long wavelengths (>1100 km). The rise at short wavelengths follows the predictions of a model in which topography is uncompensated. The flattening at intermediate wavelengths suggests some degree of compensation of topography, although not in the form as predicted by the Airy model of isostasy. The observed spectra are clearly higher than the calculated spectra in this wavelength band and suggest instead a flexure model of



Isostasy, Fig. 9 Comparison of the power spectrum of Earth's free-air gravity anomaly field (filled black circles) to calculated spectra based on the gravity effect of uncompensated topography (Red filled circles) and the gravity effect of the topography and its Airy compensation (Green filled circles) and flexural compensation (Light blue filled circles). From Watts and Moore (2017). The observed spectra have been calculated from the spherical harmonic coefficients that describe the EGM2008 gravity field model (Pavlis et al. 2008) and the uncompensated and isostatic spectra have been calculated using equations in Kaula (1967)

and Watts and Moore (2017) and the Earth2014 topography model of Hirt and Rexer (2015). Grey band = twice the minimum Root Mean Square (RMS) difference between the observed spectra and the calculated spectra based on a flexure model of isostasy, used to define a region of permissibility for the best fit elastic thickness, T_e . HDB = High Degree Band. LDB = Low Degree Band. Inset shows the RMS difference between observed spectra and calculated spectra of the gravity effect of the topography and its flexural compensation as a function of T_e . The best fit global T_e is 34.0 ± 4.0 km

isostasy with a T_e of 34.0 ± 4.0 km. This range of T_e is interesting, especially as it is a global mean and indicates that there are regions of Earth with lower and higher values. The rise at long wavelengths (>1100 km) cannot be explained by the gravity effect of the flexure model of isostasy, or any other isostatic model, and therefore defines for us the limits of isostasy and the role played by non-isostatic processes such as those associated with mantle convection in contributing to Earth's gravity field spectra. Improved gravity anomaly, topography and seismic field data, localized spectral analysis techniques, and better geodynamical models that include a lithosphere offer the most promise of better quantifying the contribution of isostasy to Earth's gravity and topography fields and crustal structure in the future.

Conclusions

Isostasy is a principle that is central to the Earth Sciences. It was first formulated in the 1880s to describe the tendency of

the crust to float, in hydrostatic equilibrium, on its denser substrate such that light areas stand at a greater elevation than dense areas. Today, we consider isostasy as a highly idealized state that the crust and upper mantle approaches, but rarely achieves. Nevertheless, the comparison of observed gravity anomalies to the predictions of local and regional isostatic models have led to a new understanding about the thermal and mechanical properties of the lithosphere. We now know, for example, that the lithosphere responds to long-term (i.e., $>10^5$ years) loading and unloading in a similar manner as would a strong elastic plate that overlies a weak fluid substrate. In the oceans, the thickness of the elastic layer that supports a volcano load increases with the age of the lithosphere at the time of loading. In the continents, the relation between elastic thickness and age is not as clear. However, there is evidence that older Archaen and Early Proterozoic cratons have high values of the elastic thickness (i.e., >70 km) while younger Middle/late Proterozoic and Phanerozoic orogenic belts and rifts have lower values. Data from experimental rock mechanics suggest that the elastic

thickness derived from isostatic studies is a proxy for the long-term strength of the lithosphere. The elastic thickness differs from the seismogenic layer thickness, which is the thickness of the uppermost part of the lithosphere that responds to stresses by faulting and earthquakes and the seismic “lid” thickness, which is the mechanical thickness that supports short-term loads and unloads. Isostatic studies suggest that the time-scales of isostatic adjustment varies regionally such that tectonically active areas respond to loads and unloads much quicker than cratonic areas. The behavior of the lithosphere on seismic to geologic time-scales is important to take into account when modelling geological processes, especially those linking dynamic topography and mantle convection.

Bibliography

- Airy GB (1855) On the computation of the effect of the attraction of mountain-masses, as disturbing the apparent astronomical latitude of stations of geodetic surveys. *Phil Trans R Soc A* 145:101–104
- Ali MY, Watts AB (2009) Subsidence history, gravity anomalies and flexure of the United Arab Emirates (UAE) foreland basin. *Geo Arabia* 14:17–44
- Audet P, Mareschal J-C (2007) Wavelet analysis of the coherence between Bouguer gravity and topography: application to the leaistic thickness anisotropy in the Canadian shield. *Geophys J Int* 168:287–298
- Banks RJ, Parker RL, Huestis SP (1977) Isostatic compensation on a continental scale: local versus regional mechanisms. *Geophys J Roy Astro Soc* 51:431–452
- Bassin C, Laske G, Masters G (2000) The current limits of resolution for surface wave tomography in North America. *Eos Trans AGU* 81: F897
- Beaumont C (1979) On rheological zonation of the lithosphere during flexure. *Tectonophysics* 59:347–365
- Bechtel TD, Forsyth DW, Sharpton VL, Grieve RAF (1990) Variations in effective elastic thickness of the north American lithosphere. *Nature* 343:636–638
- Bloom A (1967) Pleistocene shorelines: a new test of isostasy. *Geol Soc Am Bull* 78:1477–1494
- Blum MD, Tomkin JH, Purcell A, Lancaster RR (2008) Ups and downs of the Mississippi Delta. *Geology* 36:675–678
- Bott MHP, Dean DS (1972) Stress systems at young continental margins. *Nature* 235:23–25
- Bott MHP, Kusznir NJ (1979) Stress distribution associated with compensated plateau uplift structures with application to the continental splitting mechanism. *Geophys J R Astron Soc* 56:451–459
- Braitenberg C, Wang Y, Fang J, Hsu HT (2003) Spatial variations of flexure parameters over the Tibet-Quinghai plateau. *Earth Planet Sci Lett* 205:211–224
- Bry M, White N (2007) Reappraising elastic thickness variation at oceanic trenches. *J Geophys Res* 112. <https://doi.org/10.1029/2005JB004190>
- Burov EB, Diamant M (1995) The effective elastic thickness (T_e) of continental lithosphere: what does it really mean? *J Geophys Res* 100:3895–3904
- Burov EB, Poliakov A (2001) Erosion and rheology controls on synrift and postrift evolution: verifying old and new ideas using a fully coupled numerical model. *J Geophys Res*, 106 461–16(481):16
- Burov EB, Watts AB (2006) The long-term strength of continental lithosphere: jelly sandwich or crème brûlée. *GSA Today* 16:4–10
- Coakley B, Gurnis M (1995) Far-field tilting of Laurentia during the Ordovician and constraints on the evolution of a slab under an ancient continent. *J Geophys Res* 100:6313–6327
- Contreras-Reyes E, Grevemeyer I, Watts AB, Planert L, Flueh ER, Peirce C (2010) Crustal intrusion beneath the Louisville hotspot track. *Earth Planet Sci Lett* 289:323–333
- Crosby AG (2007) An assessment of the accuracy of admittance and coherence estimates using synthetic data. *Geophys J Int* 171:25–54
- Crosby AG, McKenzie D (2009) An analysis of young ocean depth, gravity and global residual topography. *Geophys J Int* 178:1198–1219. <https://doi.org/10.1111/j.1365-246X.2009.04224.x>
- Delaney JP (1940) Leonardo da Vinci on Isostasy. *Science* 91:546–547
- Di Donato G, Mitrovica JX, Sabadini R, Vermeersen LLA (2000) The influence of a ductile crustal zone on glacial isostatic adjustment: geodetic observables along the U.S. East Coast. *Geophys Res Lett* 27:3017–3020
- Dorman LM, Lewis BTR (1970) Experimental isostasy 1: theory of determination of the Earth's response to a concentrated load. *J Geophys Res* 75:3357–3365
- Dutton CE (1882) Physics of the earth's crust; by the Rev. Osmond Fisher. *Amer J Sci* 23:283–290
- Dutton CE (1889) On some of the greater problems of physical geology. *Bull Phil Soc Washington* 2:51–64
- Ebinger CJ, Bechtel TD, Forsyth DW, Bowin CO (1989) Effective elastic plate thickness beneath the east African and Afar plateaus and dynamic compensation of the uplifts. *J Geophys Res* 94:2883–2901
- Fisher O (1881) Physics of the Earth's crust. Macmillan and Co., London. 299 pp
- Fluck P, Hyndman RD, Lowe C (2003) Effective elastic thickness T_e of the lithosphere in western Canada. *J Geophys Res* 108:2430. <https://doi.org/10.1029/2002JB002201>
- Forsyth DW (1985) Subsurface loading and estimates of the flexural rigidity of continental lithosphere. *J Geophys Res* 90:12,623–612,632
- Freedman AP, Parsons B (1986) Seasat-derived gravity over the musician seamounts. *J Geophys Res* 91:8325–8340
- GEBCO (2003) The GEBCO digital atlas, centenary edition. CD-ROM
- Gilchrist AR, Summerfield MA (1990) Differential denudation and flexural isostasy in formation of rifted-margin upwarps. *Nature* 346:739–742
- Goetze C, Evans B (1979) Stress and temperature in the bending lithosphere as constrained by experimental rock mechanics. *Geophys J R Astron Soc* 59:463–478
- Haddad D, Watts AB (1999) Subsidence history, gravity anomalies, and flexure of the northeast Australian margin in Papua New Guinea. *Tectonics* 18:827–842
- Hager BH, Clayton RW, Richards MA, Cromer RP, Dziewonski AM (1985) Lower mantle heterogeneity, dynamic topography and the geoid. *Nature* 313:541–545
- Handy MR, Brun J-P (2004) Seismicity, structure and strength of the continental lithosphere. *Earth Planet Sci Letts* 223:427–441
- Hayford JF (1909) The figure of the earth and isostasy from measurements in the United States, 178 pp. Government Printing Office, Washington
- Heiskanen W (1931) Isostatic tables for the reduction of gravimetric observations calculated on the basis of Airy's hypothesis. *Bull Géodésique* 30:110–129
- Hirt C, Rexer M (2015) Earth2014: 1 arc-min shape, topography, bed-rock and ice-sheet models – available as gridded data and degree-10,800 spherical harmonics. *Int J Appl Earth Obs Geoinf* 39:103–112. <https://doi.org/10.1016/j.jag.2015.03.001>
- Huang J, Zhong S (2005) Sublithospheric small-scale convection and its implications for the residual topography at old ocean basins and the

- plate model. *J Geophys Res* 110. <https://doi.org/10.1029/2004JB003153>
- Hunter J, Watts AB (2016) Gravity anomalies, flexure and mantle rheology seaward of circum-Pacific trenches. *Geophys J Int* 207:288–316. <https://doi.org/10.1093/gji/ggw275>
- Jackson, J. (2002) Strength of the continental lithosphere: time to abandon the jelly sandwich? *GSA Today* September, 4–10
- Jordan TA, Watts AB (2005) Gravity anomalies, flexure and the elastic thickness structure of the India-Eurasia collisional system. *Earth Planet Sci Letts* 236:732–750
- Kalnins LM, Watts AB (2009) Spatial variations in effective elastic thickness in the western Pacific Ocean and their implications for Mesozoic volcanism. *Earth Planet Sci Letts* 286:89–100
- Karner GD (1984) Thermally induced residual topography of oceanic lithosphere. *Nature* 318:327–331
- Kaula WM (1967) Geophysical implications of satellite determinations of the Earth's gravitational field. *Space Sci Rev* 7:769–794
- Kirby J (2019) On the pitfalls of Airy isostasy and the isostatic gravity anomaly in general. *Geophys J Int* 216:103–122. <https://doi.org/10.1093/gji/ggy411>
- Kirby JF, Swain CJ (2009) A reassessment of spectral Te estimation in continental interiors: the case of North America. *J Geophys Res* 114. <https://doi.org/10.1029/2009JB006356>
- Lane, N., A. B. Watts, and A. Farrant (2007), An analysis of the Cotswolds topography: Insights into the landscape response to denudational isostasy, *J. Geological Soc. London*, 165, 85–103, <https://doi.org/10.1144/0016-76492006-76492179>
- Lin AT, Watts AB (2002) Origin of the West Taiwan basin by orogenic loading and flexure of a rifted continental margin. *J Geophys Res* 107. <https://doi.org/10.1029/2001JB000669>
- Liu S-F, Nummedal D (2004) Late cretaceous subsidence in Wyoming: quantifying the dynamic component. *Geology* 32:397–400
- Lowry AR, Smith RB (1995) Strength and rheology of the western U.S. Cordillera. *J Geophys Res* 100:17,947–17,963
- Lowry AR, Ribe NM, Smith RB (2000) Dynamic elevation of the cordillera, western United States. *J Geophys Res* 105(23):371–323,390
- Mantovani MSM, Shukowsky W, de Freitas SRC, Neves BBB (2005) Lithosphere mechanical behaviour inferred from tidal gravity anomalies: a comparison of Africa and South America. *Earth Planet Sci Letts* 230:397–412
- Marks KM, Vogt PR, Hall SA (1990) Residual depth anomalies and the origin of the Australian-Antarctic discordance zone. *J Geophys Res* 95:17325–17337
- McKenzie D (1977) Surface deformation, gravity anomalies and convection. *Geophys J Int* 48:211–238
- McKenzie D (2010) The influence of dynamically supported topography on estimates of Te. *Earth Planet Sci Lett* 295:127–138. <https://doi.org/10.1016/j.epsl.2010.03.033>
- McKenzie DP, Fairhead D (1997) Estimates of the effective elastic thickness of the continental lithosphere from Bouguer and free-air gravity anomalies. *J Geophys Res* 102:27,523–527,552
- McKenzie D, Jackson J, Priestley K (2005) Thermal structure of oceanic and continental lithosphere. *Earth Planet Sci Letters* 233:337–349
- McNutt MK, Menard HW (1982) Constraints on yield strength in the oceanic lithosphere derived from observations of flexure. *Geophys J R Astron Soc* 71:363–394
- McNutt MK, Parker RL (1978) Isostasy in Australia and the evolution of the compensation mechanism. *Science* 199:773–775
- McQueen HWS, Lambeck K (1989) The accuracy of some lithospheric bending parameters. *Geophys J* 96:401–413
- Meissner, R. (1986), Twenty years of deep seismic reflection profiling in Germany - a contribution to our knowledge of the nature of the lower Variscan crust, *J. Geological Soc. London Special Publication No.* 24, 1–10
- Meissner R, Strehlau J (1982) Limits of stresses in continental crusts and their relation to the depth-frequency distribution of shallow earthquakes. *Tectonics* 1(1):73–89
- Menard HW (1973) Depth anomalies and the bobbing motion of drifting islands. *J Geophys Res* 78:5128–5137
- Muller RD, Lim VSL, Isern AR (2000) Late tertiary tectonic subsidence on the northeast Australian passive margin: response to dynamic topography? *Mar Geol* 162:337–352
- Nakiboglu SM, Lambeck K (1983) A reevaluation of the isostatic rebound of Lake Bonneville. *J Geophys Res* 88:439–447
- Nakiboglu SM, Lambeck K (1985) Comments on thermal isostasy. *J Geodynamics* 2:51–65
- Nishimura T, Thatcher W (2003) Rheology of the lithosphere inferred from postseismic uplift following the 1959 Hebgen Lake earthquake. *J Geophys Res* 108. <https://doi.org/10.1029/2002JB002191>
- Parsons BE, Sclater JG (1977) An analysis of the variation of ocean floor bathymetry and heat flow with age. *J Geophys Res* 82:803–827
- Pavlis NK, Holmes SA, Kenyon SC, Factor JK (2008) An earth gravitational model to degree 2160: EGM 2008. EGU General Assembly 2008
- Pazzaglia FJ, Gardner TW (1994) Late Cenozoic flexural deformation of the middle U.S. Atlantic passive margin. *J Geophys Res* 99:12,143–12,157
- Pelletier JD (2004) Estimate of three-dimensional flexural-isostatic response to unloading: rock uplift due to late Cenozoic glacial erosion in the western United States. *Geology* 32:161–164
- Pérez-Gussinyé M, Watts AB (2005) The long-term strength of Europe and its implications for plate-forming processes. *Nature* 436. <https://doi.org/10.1038/nature03854>
- Pérez-Gussinyé M, Lowry AR, Watts AB, Velicogna I (2004) On the recovery of effective elastic thickness using spectral methods: examples from synthetic data and from the Fennoscandia shield. *J Geophys Res* 109. <https://doi.org/10.1029/2003JB002788>
- Pérez-Gussinyé M, Lowry AR, Watts AB (2007) Effective elastic thickness of South America and its implications for intracontinental deformation. *Geochem Geophys Geosyst* 8. <https://doi.org/10.1029/2006GC001511>
- Pérez-Gussinyé M, Metois M, Fernandez M, Verges J, Fulla J, Lowry AR (2009) Effective elastic thickness of Africa and its relationship to other proxies for lithospheric structure and surface tectonics. *Earth Planet Sci Lett* 287:152–167
- Pratt JH (1855) On the attraction of the Himalaya mountains, and of the elevated regions beyond them, upon the plumb line in India. *Phil Trans R Soc A* 145:53–100
- Royden L, Karner GD (1984) Flexure of the lithosphere beneath Apennine and Carpathian foredeep basins: evidence for an insufficient topographic load. *Am Assoc Pet Geol* 68:704–712
- Rychert CA, Fischer KM, Rondenay S (2005) A sharp lithosphere-aesthenosphere boundary imaged beneath eastern North America. *Nature* 436. <https://doi.org/10.1038/nature03904>
- Sacek V, Ussami N (2009) Reappraisal of the effective elastic thickness for the sub-Andes using 3-D finite element flexural modelling, gravity and geological constraints. *Geophys J Int* 179:778–786
- Sclater JG, Francheteau J (1970) The implications of terrestrial heat flow observations on current tectonic and geochemical models of the crust and upper mantle of the earth. *Geophys J R Astron Soc* 20:509–542
- Sibuet J-C, Veyrat-Peinet B (1980) Gravimetric model of the Atlantic equatorial fracture zones. *J Geophys Res* 85:943–954
- Simons FJ, Zuber MT, Korenaga J (2000) Isostatic response of the Australian lithosphere: estimation of effective elastic thickness and anisotropy using multitaper spectral analysis. *J Geophys Res* 105:19,163–119,184

- Stark CP, Stewart J, Ebinger CJ (2003) Wavelet transform mapping of the effective elastic thickness and plate loading: validation using synthetic data and application to the study of the south African tectonics. *J Geophys Res* 108. <https://doi.org/10.1029/2001JB000609>
- Stern TA, Baxter AK, Baxter PJ (2005) Isostatic rebound due to glacial erosion within the Transantarctic Mountains. *Geology* 33:221–224
- Tassara A, Swain C, Hackney R, Kirby J (2006) Elastic thickness structure of South America estimated using wavelets and satellite-derived gravity data. *Earth Planet Sci Lett* 253:17–36
- ten Brink U, Stern T (1992) Rift flank uplifts and hinterland basins: comparison of the Transantarctic Mountains with the great escarpment of southern Africa. *J Geophys Res* 97:569–585
- Tucker GE, Slingerland RL (1994) Erosional dynamics, flexural isostasy, and long-lived escarpments: a numerical modeling study. *J Geophys Res* 99:12,229–12,243
- Vai GB (2006) Isostasy in Luigi Ferdinando Marsili's manuscripts, in *The origins of geology in Italy*, edited. Geol Soc, Amer
- Vening Meinesz FA (1931) Une nouvelle methode pour la réduction isostatique régionale de l'intensité de la pesanteur. *Bull Géod* 29:33–51
- Watts AB (1976) Gravity and bathymetry in the Central Pacific Ocean. *J Geophys Res* 81:1533–1553
- Watts AB (1978) An analysis of isostasy in the world's oceans: 1. Hawaiian-emperor seamount chain. *J Geophys Res* 83:5,989–6,004
- Watts AB (2001) *Isostasy and flexure of the lithosphere*, 458 pp. Cambridge University Press, Cambridge
- Watts AB (2007) An overview, in *Treatise of Geophysics*. In: Watts AB (ed) *Crust and lithosphere dynamics*, vol 6. Elsevier, pp 1–48
- Watts AB, Burov EB (2003) Lithospheric strength and its relationship to the elastic and seismogenic thickness. *Earth Planet Sci Lett* 213:113–131
- Watts AB, Fairhead JD (1997) Gravity anomalies and magmatism at the British Isles continental margin. *J Geol Soc Lond* 154:523–529
- Watts AB, Moore JDP (2017) Flexural isostasy: constraints from gravity and topography power spectra. *J Geophys Res* 122. <https://doi.org/10.1002/2017JB014571>
- Watts AB, Zhong S (2000) Observations of flexure and the rheology of oceanic lithosphere. *Geophys J Int* 142:855–875. <https://doi.org/10.1046/j.1365-246x.2000.00189.x>
- Watts AB, Sandwell DT, Smith WHF, Wessel P (2006) Global gravity, bathymetry, and the distribution of submarine volcanism through space and time. *J Geophys Res* 111. <https://doi.org/10.1029/2005JB004083>
- Watts AB, Rodger M, Peirce C, Greenroyd CJ, Hobbs RW (2009) Seismic structure, gravity anomalies, and flexure of the Amazon continental margin, NE Brazil. *J Geophys Res* 114. <https://doi.org/10.1029/2008JB006259>
- Watts AB, Zhong SJ, Hunter J (2013) The behavior of the lithosphere on seismic to geologic timescales. *Annu Rev Earth Planet Sci* 41. <https://doi.org/10.1146/annurev-earth-042711-105457>
- Whitehouse P, Latychev K, Milne GA, Mitrova JX, Kendall R (2006) Impact of 3-D earth structure on Fennoscandian glacial isostatic adjustment: implications for space-geodetic estimates of present-day crustal deformations. *Geophys Res Lett* 33. <https://doi.org/10.1029/2006GL026568>
- Wieczorek MA (2007) Gravity and topography of the terrestrial planets, in *Treatise of Geophysics*. Volume 10: Planets and Moons, edited by T. Spohn, pp. 165–206, Elsevier
- Willett SD, Chapman DS, Neugebauer HJ (1985) A thermo-mechanical model of continental lithosphere. *Nature* 314:520–523
- Zuber MT, Bechtel TD, Forsyth DW (1989) Effective elastic thickness of the lithosphere and mechanisms of isostatic compensation in Australia. *J Geophys Res* 94:13,919–13,930

Isostasy, Thermal

Derrick Hasterok¹ and David S. Chapman²

¹Earth Sciences Department, University of Adelaide, Adelaide, SA, Australia

²Department of Geology and Geophysics, The University of Utah, Salt Lake City, UT, USA

Definition

Isostasy

(Greek: isos “equal,” stasis “standstill”) term used in geology to refer to the condition of gravitational equilibrium such that the rigid outer part of Earth's crust and uppermost mantle “float” at an elevation which depends on its density. This concept explains how different topographic heights can exist at Earth's surface.

Compositional isostasy

The part of isostasy traceable to density variations arising from compositional (rock type and mineralogy) differences.

Thermal isostasy

The part of isostasy traceable to density variations arising from temperature differences and thermal expansion within the rock column.

Isostatic equilibrium

The condition whereby a certain area of crust and upper mantle reaches the state of isostasy such that a hydrostatic equilibrium exists at depth below a level of compensation. The region is said to be in *isostatic equilibrium*.

Introduction

The condition in Earth whereby variations in crustal thickness and density determine the elevation of Earth's solid surface is known as isostasy and has been appreciated for a century and a half (Pratt 1855; Airy 1855; Watts 2001, for general review). Local isostasy requires a hydrostatic equilibrium condition at a depth below a compensation level, implying that the integral of density over a rock column is constant. When density differences arise from differences in rock composition (rock type, mineralogy), the isostatic condition is referred to as compositional isostasy. It is compositional isostasy, and in particular the density contrast across the crust–mantle boundary, that dominates isostasy discussions in textbooks. If, on the other hand, density differences arise from different thermal states of the lithosphere and thermal expansion of rock, the condition is properly called thermal isostasy. This entry is

devoted to thermal isostasy, computation of thermal isostasy effects, and examples of thermal isostasy in oceanic regions and on the continents.

Thermal Isostasy

Rock density is influenced by temperature through thermal expansion. If rock composition is identical in two vertical columns through the lithosphere, the warmer column has a lower density than the cooler lithosphere. Applying the isostatic condition of hydrostatic equilibrium at some depth of compensation leads to a prediction of elevation changes for the two regions.

Consider two regions of different lithospheric thermal states. Temperature differences between the two regions represented by a regional geotherm $T(z)$ and a reference geotherm $T_{\text{ref}}(z)$, respectively, multiplied by a thermal expansion coefficient α_V and integrated over depth predict an elevation change, $\Delta\epsilon_T$, given by

$$\Delta\epsilon_T = \alpha_V \int_0^{z_{\text{max}}} [T(z) - T_{\text{ref}}(z)] dz. \quad (1)$$

The maximum depth of integration, z_{max} , is the depth at which the colder geotherm converges to a mantle adiabat assumed to be identical for both lithospheric regions.

Most studies use a coefficient of thermal expansion of $3.0 \times 10^{-5} \text{ K}^{-1}$ within the crust, consistent with the values for major crustal forming rocks including granite and gabbro. Within the mantle, a value of thermal expansion of $3.2 \times 10^{-5} \text{ K}^{-1}$ is used.

Oceanic Thermal Isostasy

The clearest example of thermal isostasy occurs in oceanic regions as hot, near-molten lithosphere at a spreading ridge

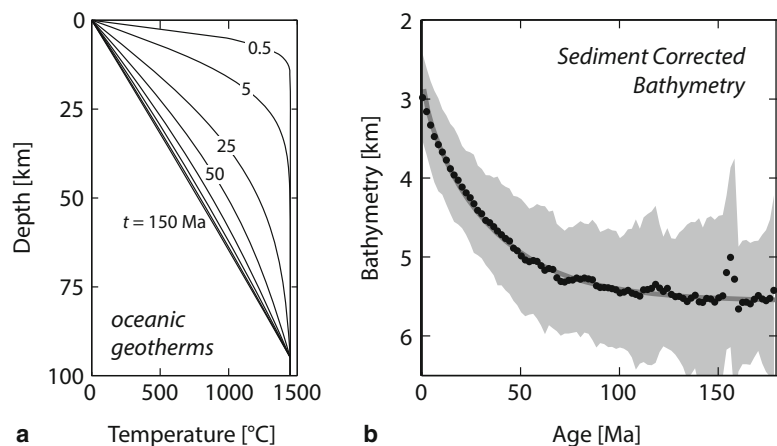
cools over tens of millions of years leading to thermal contraction, increasing rock density, and subsidence of the seafloor by about 3,000 m. It is important to note that lateral compositional variations are small in oceanic regions. Oceanic crust about 7 km thick is formed at a spreading center and its thickness and composition are modified only in minor ways over its maximum 165 My existence. The thermal state of oceanic lithosphere, on the other hand, varies from a near molten state at the ridge to extremely cold lithosphere after up to 165 My of cooling. These contrasting thermal states have consequences for the rigidity of the lithosphere but in particular for its density and therefore elevation.

Almost since the discovery of plate tectonics and seafloor spreading, thermal isostasy has been invoked to explain the bathymetry of the oceans as a function of age. Figure 1a illustrates the time evolution of an oceanic geotherm for a cooling plate model. The initial condition is 1,450 °C (Stein and Stein 1993); other cooler values lead to similar subsidence curves if thermal expansion values are adjusted accordingly. The ocean lithosphere initially cools rapidly as shown by the large difference between the 0.5 My geotherm and the 5 and 25 My geotherms (Fig. 1a). At longer times, the cooling slows and the geotherms reach near steady state shown by the 150 My geotherm. The average temperature change in the lithosphere between zero age at the ridge and about 150 My seafloor is 725 K.

Bathymetry is proportional to the integrated cooling of the oceanic lithosphere (Eq. 1), additionally incorporating the changing load of water as the seafloor subsides. This subsidence (Fig. 1b) is explained very well by a simple one-dimensional (1-D) model of lithospheric cooling. A midocean ridge, or zero age lithosphere, has on average a depth of 2.7 km, a consequence of both its thermal state, crustal thickness, and composition. But as the lithosphere ages and cools, temperature changes but crustal thickness and composition do not. Cooling over the first 9 My causes a subsidence of 1 km to a depth of 3.7 km. At a lithospheric age of 36 My, the solid surface has subsided a further kilometer to a water depth of

Isostasy, Thermal,

Fig. 1 Thermal evolution of oceanic lithosphere after formation at a spreading oceanic ridge system. (a) Temperature-depth profiles as a function of lithosphere age in millions of years. (b) Subsidence of the seafloor as a function of seafloor age. Solid dots are data in 2 My age bins. Gray shading represents ± 1 standard deviation about the mean bathymetry.



4.7 km. As the lithosphere reaches its thermal equilibrium, the bathymetry also approaches its equilibrium value of 5.5 km.

Continental Thermal Isostasy

On continents, thermal isostasy has been used successfully to examine the evolution of regions that mimic oceanic spreading, such as continental rifts and elements of provinces with extensive volcanism or back-arc regions (McKenzie 1978; Jarvis and McKenzie 1980; Brott et al. 1981; Lachenbruch and Morgan 1990; Hyndman et al. 2005). However, direct thermal effects on continental elevation are difficult to discern because of the potential masking effect of compositional variations in lithospheric density and crustal thickness.

The elevation effect due to compositional variation can be estimated by a simple isostatic calculation using continental extremes. A mountainous region with a crustal thickness of 50 km and a density of $2,800 \text{ kg m}^{-3}$ (granodiorite) would have an elevation 5 km higher than a rift province with a crustal thickness of 25 km and a density of $2,900 \text{ kg m}^{-3}$ (gabbro). Both columns assume a similar mantle density of $3,340 \text{ kg m}^{-3}$. The potentially large compositional effect can easily mask the effect of thermal isostasy. Therefore, in order to isolate the effect of thermal isostasy in the continents, it is necessary first to remove compositional isostatic effects.

Normalizing Compositional Elevation

Compositional variations involving both crustal thickness and density are removed with a simple isostatic correction to

the observed elevation (Han and Chapman 1995). The adjustment normalizes any crustal column to an arbitrary crustal standard. In refining this method, Hasterok and Chapman (2007a) used a standard crustal thickness h'_c of 39 km and a standard crustal density ρ'_c of $2,850 \text{ kg m}^{-3}$. These values represent the average crustal thickness and density of North American provinces. A lithospheric mantle density of $3,340 \text{ kg m}^{-3}$ is used based on the xenolith derived estimate by Griffin et al. (1999) for Proterozoic lithosphere.

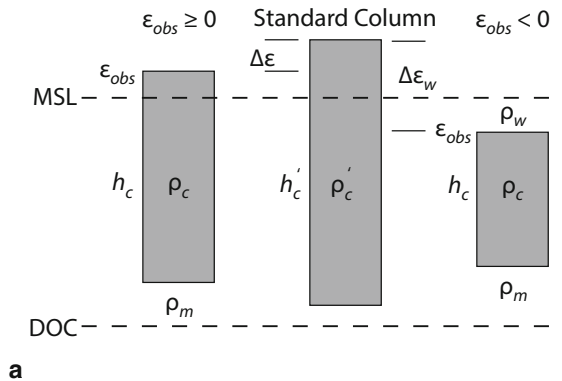
Consider a crustal column with a crustal thickness, h_c , and crustal density, ρ_c , at an elevation of ϵ_{obs} (Fig. 2a). If the density and thickness of the observed crustal column are adjusted to match the standard density and thickness, the elevation would change by an amount $\Delta\epsilon_c$ (Han and Chapman 1995) given by

$$\Delta\epsilon_c = h'_c \left(1 - \frac{\rho'_c}{\rho_m} \right) - h_c \left(1 - \frac{\rho_c}{\rho_m} \right) \quad (2)$$

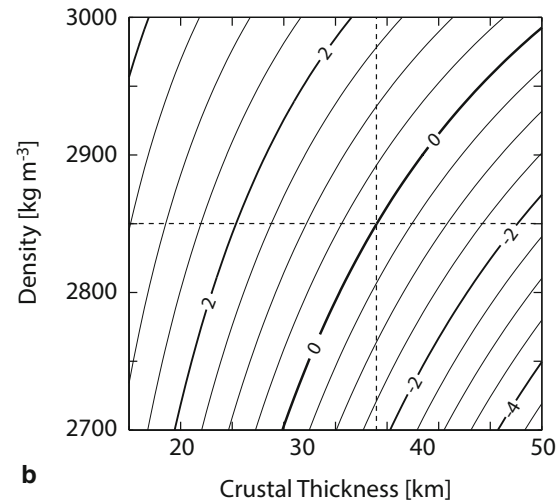
where ρ_m is the density of the mantle. If the province surface is below sea level (e.g., continental shelf), an additional term involving the water depth (where ϵ_{obs} is the bathymetry) and seawater density, ρ_w , is required and the elevation adjustment becomes

$$\Delta\epsilon_c = h'_c \left(1 - \frac{\rho'_c}{\rho_m} \right) - h_c \left(1 - \frac{\rho_c}{\rho_m} \right) - \frac{\epsilon_{\text{obs}} \rho_w}{\rho_m}. \quad (3)$$

By adding the elevation adjustment to the observed elevation, one arrives at the final adjusted elevation, ϵ_{adj} , given by



Isostasy, Thermal, Fig. 2 Compositional isostasy adjustments to continental elevation. (a) Cartoon illustrating parameters used in compositional correction for elevation. MSL is mean sea level and DOC is the depth of compensation. The observed crustal columns (*left* for regions above sea level and *right* for continental margins below sea level) are adjusted to an arbitrary standard crust (*center*) of thickness 39 km and



density $2,850 \text{ kg m}^{-3}$. (b) Nomogram contouring the compositional elevation adjustment for observed crustal thickness and average crustal density. Contour interval is 0.5 km. Dashed lines for crustal thickness of 39 km and density of $2,850 \text{ kg m}^{-3}$ represent the standard crustal column. Zero line is locus of values for which the elevation adjustment is zero.

$$\varepsilon_{\text{adj}} = \varepsilon_{\text{obs}} + \Delta\varepsilon_{\text{C}}. \quad (4)$$

Three important physical parameters must be estimated in order to make the compositional adjustment: ε_{obs} , h_{C} , and ρ_{C} . The observed elevation is obtained from GTOPO30 for elevations above sea level and from bathymetric maps for elevations below sea level. Crustal thickness is obtained from 1-D whole crustal V_{P} models. Crustal density is estimated by using empirical velocity–density (V_{P} – ρ) relationships. The lateral dimension of each province analyzed varies from ~ 70 km in some provinces in the Western USA to >500 km in shield provinces. See Hasterok and Chapman (2007a, b) for details.

The magnitudes of compositional elevation adjustments made using Eqs. 2 and 3 are shown in Fig. 2b. The zero contour is the locus of crustal thickness and density combinations for which no crustal adjustment is made and therefore passes through the standard crust of thickness 39 km and density $2,850 \text{ kg m}^{-3}$. Over the extreme range of crustal thicknesses (20–60 km) and densities ($2,700$ – $3,000 \text{ kg m}^{-3}$), compositional elevation adjustments can be as much as 4 km. Regions that are thinner and/or more dense than the standard crust receive positive elevation adjustments, whereas regions that are thicker and/or less dense than the standard crust receive negative elevation adjustments. Most adjustments in Fig. 2b fall in the range ± 1 km. From this nomogram we can also determine the uncertainty in elevation adjustment introduced by variations in crustal thickness and density. For example, to achieve an uncertainty of ± 250 m requires an uncertainty in crustal thickness less than 3 km and in density less than 43 kg m^{-3} .

Continental Thermal State

The thermal state of continental lithosphere is described by a set of geotherms. Whereas oceanic geotherms may be identified in terms of age, continental geotherms do not correlate well with rock age. Instead, the most important controlling parameter for continental geotherms is surface heat flow (Pollack and Chapman 1977b; Chapman and Pollack 1977; Blackwell 1971; Lachenbruch and Sass 1977; Rao et al. 1982).

There exist a number of geotherm models of the continental crust (Blackwell 1971; Lachenbruch and Sass 1977; Chapman 1986; Chapman and Furlong 1992). Many other models are based on locally derived estimates of thermal conductivity and/or heat production. Although locally and xenolith-derived models may be best in the region for which they are developed, they may be poor estimates when extended to other regions. Therefore, we feel it is best to use a consistent geothermal model based on the fewest parameters necessary to demonstrate the usefulness of this method. We chose the method of Chapman (1986) and Chapman and Furlong (1992) to compute our geotherms because they yield a set of geotherms that are well documented, commonly

used, and yield temperatures which fall between other warm and cool models.

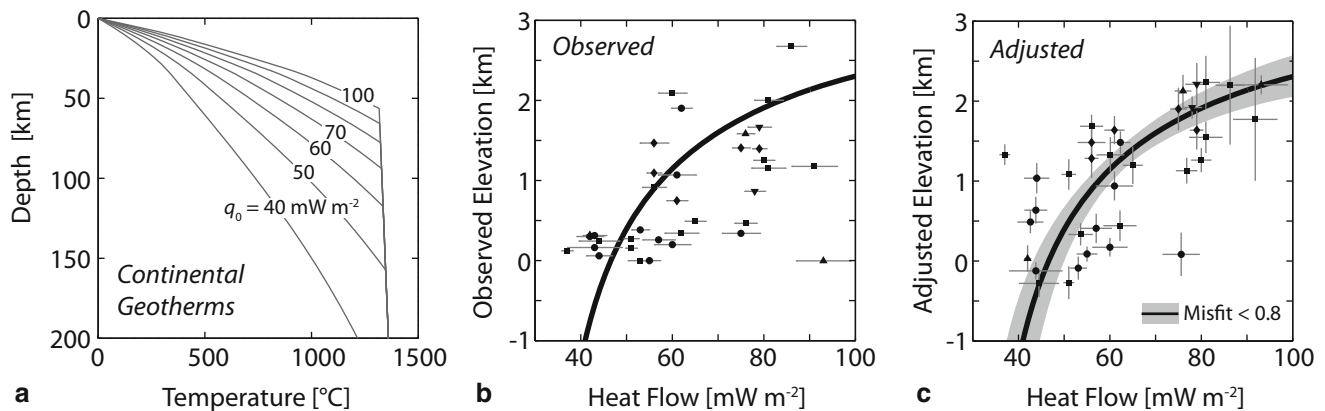
The properties of the geotherm model include a lithospheric structure divided into an upper and lower crust and mantle lithosphere. Thermal conductivity within the crust is determined using a pressure-/temperature-dependent relationship of Chapman (1986). Thermal conductivities within the upper and lower crust are initially set as 3.0 and $2.65 \text{ W m}^{-1} \text{ K}^{-1}$, respectively, corresponding generally to felsic and mafic crystalline rocks at STP conditions. Mantle lithospheric thermal conductivities include a radiation term at temperatures above 500 K . Surface heat production is determined assuming a 40% radiogenic contribution to surface heat flow with a characteristic depth of 8 km (Pollack and Chapman 1977a). This characteristic depth is also used as the decay length of an exponentially decreasing function used to compute heat generation within the upper crust. Heat production of the lower crust is set as $0.4 \mu\text{W m}^{-3}$ consistent with many studies of exposed granulite terranes. Mantle heat production is assumed to be $0.02 \mu\text{W m}^{-3}$, a lower value justified by Chapman (1986) and Chapman and Furlong (1992). Temperatures within the asthenosphere are computed using an adiabat with a potential temperature of $1,300^\circ\text{C}$ and increase with depth of $0.3^\circ\text{C km}^{-1}$. Steady state, 1-D, conductive continental geotherms computed for surface heat flow values between 40 and 100 mW m^{-2} are shown in Fig. 3a. Details of the geotherm calculation and boundary conditions used are given by Chapman (1986) and further discussed in Hasterok and Chapman (2007a).

The thermal structure in some regions differs from the geotherm models used in this study as a result of nonsteady state processes and differences in thermophysical properties. However, we have chosen to keep our thermal models as simple as possible, interpreting the more difficult and uncertain parameters, that is, heat production and thermal conductivity, as residuals from a reference thermal isostatic model.

Revealing Thermal Isostasy on Continents

Because the thermal state of continents is not clearly related to the age of the lithosphere as it is in the oceans, one cannot confirm continental thermal isostasy by simply plotting elevation of continental regions against age of the province. Other approaches must be explored. One such approach examined elevation as a function of heat flow for several tectonic provinces around the globe (Han and Chapman 1995; Nagihara et al. 1996). Hasterok and Chapman (2007a, b) followed the same approach in much greater detail and made an assessment of the uncertainties related to removing compositional isostatic effects.

A thermal isostatic curve for the continents, derived from the geotherm family in Fig. 3a and Eq. 1, is shown in Fig. 3b. We use a reference geotherm corresponding to a surface heat flow of 47 mW m^{-2} and assign a lithosphere having this



Isostasy, Thermal, Fig. 3 Thermal isostasy on continents. (a) Generalized family of continental geotherms parametric in surface heat flow values from 40 to 100 mW m⁻². Geotherms are truncated by the 1,300 °C mantle adiabat. (b) Observed elevation versus heat flow for individual tectonic provinces of North America. Symbols denote dominant style of tectonic province: circle (shield), star (basin), square (collision), diamond (volcanic), triangle (extension). Error bars are one standard

deviation of the mean for heat flow. The geotherms in (a) together with Eq. 1 are used to calculate a theoretical thermal isostasy relationship shown as a bold line. (c) Elevation adjusted for compositional effects versus heat flow for the same tectonic provinces in (b). Elevation error bars are predicted from a Monte Carlo analysis of uncertainties in the seismic velocity to density conversion. Data trend confirms expected 3,000 m of thermal isostasy effect.

thermal state an elevation of 0 km. Although the zero elevation reference heat flow at 47 mW m⁻² is assumed here, the actual zero elevation could easily be set at another value. This initial assumption for the zero-elevation reference heat flow is refined further by Hasterok and Chapman (2007b). Within the range of surface heat flow for continental tectonic provinces (40–90 mW m⁻²), the predicted elevation range resulting from thermal isostasy is approximately 3 km. This prediction is similar to the oceanic bathymetry difference between hot ridges and cold abyssal plains.

Figure 3b also demonstrates how much continental thermal isostasy is obscured by compositional effects. There is little correlation between observed elevation of 36 North American tectonic provinces and the continental thermal isostatic curve. Much of this scatter is the result of variations in crustal density and thickness between the different tectonic provinces.

Compositional elevation adjustments for the 36 tectonic provinces of North America range from 763 ± 221 m in the Middle Rocky Mountains to $2,207 \pm 119$ m in the Gulf of California (Hasterok and Chapman 2007b). Nearly all of the provinces have compositional elevation adjustments between -1 and +1 km. The results suggest that the compositional contribution to the elevation of North America accounts for ~3 km of observed elevation variation. Patterns are evident in variations of elevation adjustment as a function of crustal thickness and density (Fig. 3c). The magnitude and range of elevation adjustments correlate with the type of tectonic regime that characterizes each North American province. Precambrian cratons have compositional adjustments between -0.5 and +0.5 km. Cenozoic rifts have elevation adjustments >0.5 km, and adjustments for volcanic provinces range from just under 0–1.0 km. Collisional orogens have a

very large range of elevation adjustments relative to the other tectonic environments, ranging from 0.75 to 1.0 km.

The elevation adjustment is added to the observed elevation to determine the compositionally adjusted elevation (Fig. 3c). Although the data shown in Fig. 3c exhibit much scatter, two trends are clear. First, the adjusted elevation difference between hot and cold provinces is ~3 km, a magnitude roughly equal to the elevation range observed in the oceans between hot spreading centers and cold abyssal plains. Second, whereas observed elevations do not define a clear trend with heat flow, the adjusted province elevations are less scattered and define a pattern that more closely resembles our predicted thermal isostatic curve (black line). The considerable scatter in Fig. 3c may be caused by imperfect compositional adjustment (seismic velocity characterization, velocity to density transform), by regions not being in isostatic equilibrium, or by surface heat flow not being perfectly representative of the thermal state of the lithosphere as described by the geotherm model used.

The North American Cordillera

The elevation of the North American Cordillera provides an excellent case study for thermal isostasy. Conventional wisdom suggests that many continental mountain belts, the result of continent–continent collision, are elevated because the collision process has thickened the continental crust. Airy isostasy considerations predict terrain elevations of 1,000–2,000 m as a result of thickened crust. But Hyndman (2010) point out that modern seismic data reveal that many of these collision provinces, in particular the North American Cordillera, do not have crustal roots. The Canadian Cordillera, for example, has a crustal thickness of only 35 km in comparison to the Canadian Shield crustal thickness of 40 km. But the elevation of the

thinner crust Cordillera is more than 1,000 m higher than the Canadian Shield, exactly opposite to the Airy compositional isostasy prediction.

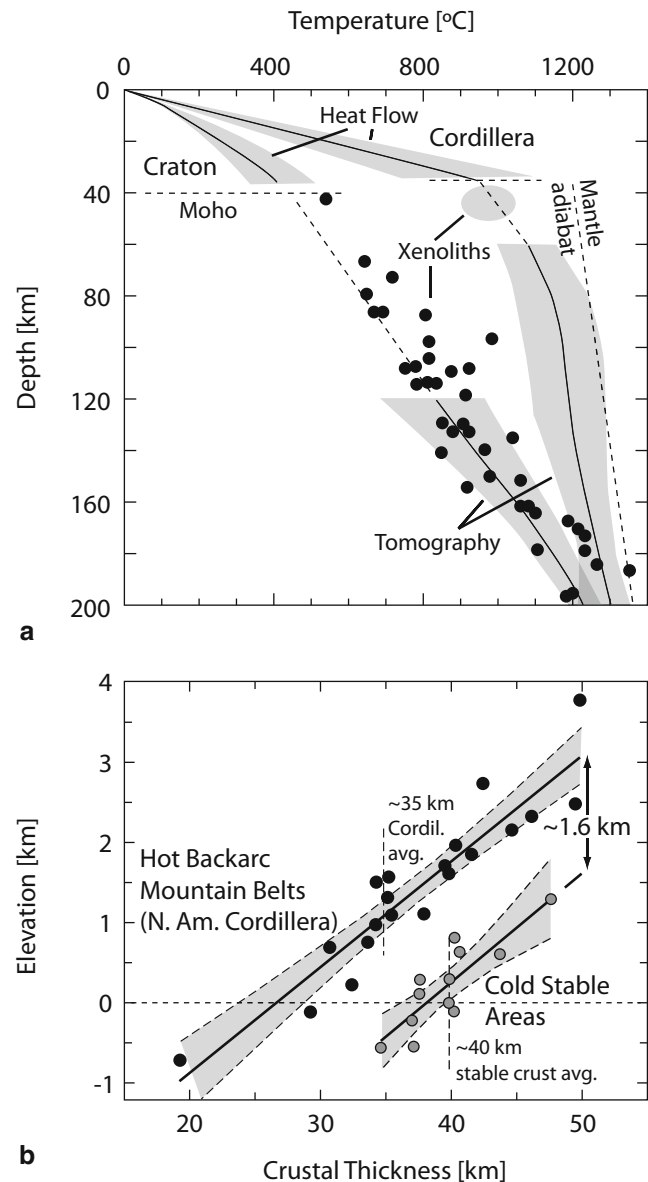
The solution to this apparent conundrum is found in thermal isostasy (Hyndman 2010). Hyndman et al. (2005) and Currie and Hyndman (2006) have shown that high upper mantle temperatures characterize most subduction back-arc regions (see also Lewis et al. 1992, 2003). And that many mountain belts also occur in current or recent back-arc locations. The temperature-depth field for the Cordillera, therefore, is much hotter than for a cratonic region (Fig. 4a). The Cordillera geotherm approaches a mantle adiabat at a depth of about 60 km, defining the lithospheric thickness for the Cordillera, while the Craton geotherm does not approach the mantle adiabat until a depth of more than 200 km. If Eq. 1 is applied to these two regions, the estimated difference in elevation, once compositional effects have been removed should be about 1,600 m. Figure 4b (from Hyndman 2010) shows exactly that effect. Elevations for both back-arc mountain belts and cold stable areas corrected for compositional density effects are plotted against local crustal thickness. Each group of data exhibit the expected trend of increasing elevation with increasing crustal thickness, but the Cordillera elevations are consistently 1,600 m higher than stable area elevations for the same crustal thickness.

Australia

Not all variations in heat flow between provinces can be easily attributed to increases in temperature. Proterozoic Australia, for instance, has typical heat flow values $>80 \text{ mW m}^{-2}$, which is more consistent with active tectonic zones than shields or cratons. Numerous studies have attributed these high heat flow anomalies to very high crustal radioactivity (e.g., McLaren et al. 2003). Surface heat flow, q_s , can be decomposed into a radiogenic heat flow, q_r , and a sub-lithospheric heat flow, q_m ,

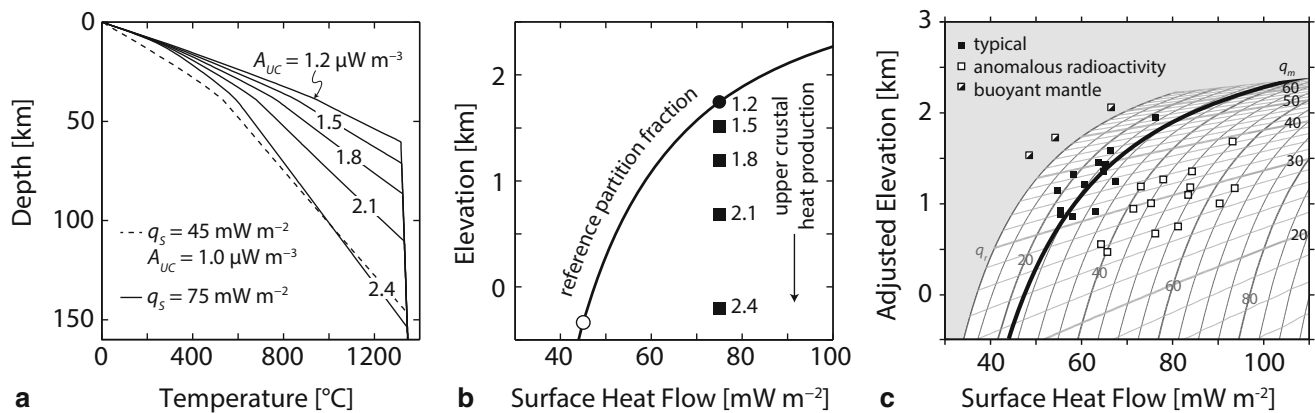
$$q_s = q_r + q_m \quad (5)$$

Figure 5a illustrates several geotherms expected for a set of regions with identical surface heat flow but differing upper crustal heat production. The integrated lithospheric temperatures is much lower than expected by the reference geotherm family, instead isostatically equivalent to a typical region with low surface heat flow (Fig. 5b). In Fig. 3, a typical partitioning of surface heat flow between a radiogenic upper crustal component (40%) and a basal component (60%) is assumed. If this model is reconsidered to allow for variations in partitioning fraction, the surface heat flow and adjusted elevation plot (Fig. 5c) can be used to identify regions of anomalous radioactivity or sub-lithospheric heat flow. Hasterok and Gard (2016) used thermal isostasy in this manner to estimate the heat production variations within the Australian continent (Fig. 5c).



Isostasy, Thermal, Fig. 4 Thermal isostasy case study for North America Cordillera (after Hyndman 2010). (a) Temperature-depth plot for the northern Cordillera in comparison to the adjacent craton. Shaded regions in the mantle are general results from tomography; solid dots are temperature-depth loci from xenolith studies (see Hyndman 2010 for details). (b) Elevation adjusted for composition (density) versus crustal thickness (Table 3 of Hasterok and Chapman 2007b) for the hot Cordillera and cold stable areas of North America. The ~1.6 km elevation difference can be explained by thermal isostasy using Eq. 1

The decomposition method can also aid interpretation of Australian non-thermal isostatic variations. Some regions sit outside the bounds of allowable heat flow–adjusted elevation pairs which indicate a violation of the basic model (Fig. 5c). The three provinces that sit outside this region in Australia include the Pilbara Craton, Yilgarn Craton, and the Albany-Fraser belt (a reworked portion of the Yilgarn Craton). The



Isostasy, Thermal, Fig. 5 Decomposition of surface heat flow using thermal isostasy. (a) A demonstration of radiogenic effect on geotherms with the same surface heat flow and differing upper crustal heat production. (b) The differences in geotherms lead to elevation anomalies with respect to the reference thermal isostatic curve. (c) Case study of Australian heat flow and elevation (after Hasterok and Gard 2016) used to estimate radiogenic and mantle heat flow contributions to the surface heat flow. The sub-horizontal contours are lines of constant mantle heat flow whereas the sub-vertical contours are lines of constant radiogenic heat flow. The individual geological provinces separate into three

distinct sets, one with typical ranges of radioactivity and mantle heat flow when compared with North America, a second set of provinces with anomalously low elevation/high surface heat flow consistent with high radiogenic heat flow, and a third set outside the thermal bounds allowable by the model—a consequence of low-density Archean lithospheric roots. The error bars on the individual provinces have been removed to emphasize the background contours. The geotherms used to compute these models have been updated, but the thermal elevation, in Eq. 1, and the process of elevation adjustment are the same.

lithospheric mantle beneath these regions are often assumed to be depleted – compositionally buoyant – Archean lithosphere, violating the thermal isostatic model that only includes compositional adjustment for crustal buoyancy.

Thermal Isostasy on Terrestrial Planets

Venus and Mars also have convecting mantles beneath rigid lithosphere and presumably have significant variations in heat flow across their surfaces. Indeed, volcanism on Mars and Venus may be plume-related, but the cooling of the lithosphere following volcanic periods may result in significant temporal changes in topography. Cooling may be responsible for anomalies in ancient coastline elevations observed on Mars today (Ruiz et al. 2004).

Summary

Thermal isostasy, that part of isostasy traceable to density differences being caused by temperature differences and thermal expansion of rock, is an important but underappreciated condition in crustal geophysics. The condition/process of thermal isostasy received much attention in oceanic regions with the discovery of plate tectonics and seafloor spreading. It was quickly recognized that systematic cooling of the lithosphere formed at a spreading ridge could explain the 3,000 m of seafloor subsidence away from a ridge. On continents, thermal isostasy effects are obscured by compositional isostasy, principally crustal thickness, which produces over 3,000 m of elevation effects. Recent analyses, however, have developed

methods to adjust elevations for compositional effects. When those adjustments are made, there is a clear trend of increasing elevation of a continental province with surface heat flow – a confirmation of thermal isostasy. The magnitude of elevation differences between hot and cold continental regions is also about 3,000 m, comparable to thermal isostasy effects in marine areas. An excellent example of thermal isostasy effects is drawn from the North American Cordillera where Airy isostasy is unable to explain the elevation of the Cordillera terrains relative to stable shield areas. Instead, the hot upper mantle associated with current or recent back-arc activity, conveniently explains the 1,600 m of elevation difference that is observed.

Cross-References

- [Heat Flow, Continental](#)
- [Lithosphere, Continental: Thermal Structure](#)
- [Lithosphere, Oceanic: Thermal Structure](#)
- [Radiogenic Heat Production in the Continental Crust](#)

Bibliography

- Airy G (1855) On the computation of the effect of the attraction of mountain-masses, as disturbing the apparent astronomical latitude of stations in geodetic surveys. *Philos Trans R Soc Lond* 145:101–103
- Blackwell D (1971) The thermal structure of the continental crust. In: Heacock J (ed) The structure and physical properties of the earth's crust. Geophysical monograph series, vol 14. AGU, Washington, DC, pp 169–184

- Brott C, Blackwell D, Ziagos J (1981) Thermal and tectonic implications of heat flow of the western Snake River Plain, Idaho. *J Geophys Res* 86:11709–11734
- Chapman D (1986) Thermal gradients in the continental crust. In: Dawson J, Carswell D, Hall J, Wedepohl K (eds) *The nature of the lower continental crust*, vol 24. Geological Society Special Publication, Chapman, London, pp 63–70
- Chapman D, Furlong K (1992) Thermal state of the continental lower crust. In: Fountain D, Arculus R, Kay R (eds) *Continental lower crust geotectonics*, vol 23. Elsevier, New York, pp 179–199
- Chapman D, Pollack H (1977) Regional geotherms and lithospheric thickness. *Geology* 5:265–268
- Currie R, Hyndman R (2006) The thermal structure of subduction backarcs. *J Geophys Res* 111:B08404. <https://doi.org/10.1029/2005JB004024>
- Griffin W, O'Reilly S, Ryan C (1999) The composition and origin of sub-continental lithospheric mantle. In: Fei Y, Bertka C, Mysen B (eds) *Mantle petrology: field observations and high pressure experimentation: a tribute to Francis R. (Joe) Boyd*, vol 6. Special Publication of the Geochemical Society, Griffin, Lancaster, PA, pp 13–45
- Han U, Chapman D (1995) Thermal isostasy: elevation changes of geologic provinces. *J Geol Soc Korea* 31:106–115
- Hasterok D, Chapman D (2007a) Continental thermal isostasy: 1. Methods and sensitivity. *J Geophys Res* 112:B06414. <https://doi.org/10.1029/2006JB004663>
- Hasterok D, Chapman D (2007b) Continental thermal isostasy: 2. Application to North America. *J Geophys Res* 112:B06415. <https://doi.org/10.1029/2006JB004664>
- Hasterok D, Gard M (2016) Utilizing thermal isostasy to estimate sub-lithospheric heat flow and anomalous crustal radioactivity. *Earth Planet Sci Lett* 450:197–207. <https://doi.org/10.1016/j.epsl.2016.06.037>
- Hyndman R (2010) The consequences of Canadian Cordillera thermal regime in recent tectonics and elevation: a review. *Can J Earth Sci* 47:621–632. <https://doi.org/10.1139/E10-016>
- Hyndman R, Currie C, Mazzotti S (2005) Subduction zone backarcs, mobile belts, and orogenic heat. *Geol Soc Am Today* 15:4–10. [https://doi.org/10.1130/1052-5173\(2005\)015<4:SZBMBA>2.0.CO;2](https://doi.org/10.1130/1052-5173(2005)015<4:SZBMBA>2.0.CO;2)
- Jarvis G, McKenzie D (1980) Sedimentary basin formation with finite extension rates. *Earth Planet Sci Lett* 48:42–52
- Lachenbruch A, Morgan P (1990) Continental extension, magmatism, and elevation; formal relations and rules of thumb. *Tectonophysics* 174:39–62
- Lachenbruch A, Sass J (1977) Heat flow in the United States and the thermal regime of the crust. In: Heacock J (ed) *The Earth's crust. Its nature and physical properties*, geophysical monograph series, vol 20. AUG, Washington, DC, pp 625–675
- Lewis T, Bentkowski W, Hyndman R (1992) Crustal temperatures near the Lithoprobe Southern Cordillera Canada transect. *Can J Earth Sci* 29:1197–1214
- Lewis T, Hyndman R, Flueck P (2003) Heat flow, heat generation, and crustal temperatures in the northern Canadian Cordillera: thermal control of tectonics. *J Geophys Res* 108:2316–2334. <https://doi.org/10.1029/2002JB002090>
- McKenzie D (1978) Some remarks on the development of sedimentary basins. *Earth Planet Sci Lett* 40:25–32
- McLaren S, Sandiford M, Hand M, Neumann N, Wyborn L, Bastrakova I (2003) The hot south continent: heat flow and heat production in Australian Proterozoic terranes. In: Hillis RR, Müller RD (eds) *Evolution and dynamics of the Australian plate*. Geological Society of America Special Paper, vol 22. Geological Society of America, Boulder, pp 151–161. <https://doi.org/10.1130/0-8137-2372-8.157>
- Nagihara S, Lister C, Sclater J (1996) Reheating of old oceanic lithosphere: deductions from observations. *Earth Planet Sci Lett* 139:91–104
- Pollack H, Chapman D (1977a) Mantle heat flow. *Earth Planet Sci Lett* 34:174–184
- Pollack H, Chapman D (1977b) On the regional variation of heat flow, geotherms, and lithospheric thickness. *Tectonophysics* 38:279–296
- Pratt J (1855) On the attraction of the Himalaya mountains, and the elevated regions beyond them, upon the plumb-line in India. *Philos Trans R Soc Lond* 145:53–100
- Rao R, Rao G, Reddy G (1982) Age dependence of continental heat flow – fantasy and facts. *Earth Planet Sci Lett* 59:288–302
- Ruiz J, Fairén AG, Dohm JM, Tejero R (2004) Thermal isostasy and deformation of possible paleoshorelines on Mars. *Planet Space Sci* 52:1297–1301. <https://doi.org/10.1016/j.pss.2004.06.003>
- Stein C, Stein S (1993) Constraints on Pacific midplate swells from global depth–age and heat flow – age models. In: Pringle M (ed) *The Mesozoic Pacific. Geology, tectonics, and volcanism*, geophysical monograph series, vol 77. AGU, Washington, DC, pp 53–76
- Watts A (2001) *Isostasy and the flexure of the lithosphere*. Cambridge University Press, Cambridge

KTB Depth Laboratory: A Window into the Upper Crust

Ulrich Harms and Jochem Kück
Scientific Drilling, GFZ German Research Centre for
Geosciences, Potsdam, Germany

Definition

Depth observatory or *depth laboratory* is an up to several kilometer-deep borehole or underground facility enabling observations of geophysical, geological, mineralogical, geochemical, and biological phenomena in the deep Earth. It is equipped with various sensors allowing for long-term data acquisition at crustal depth level.

KTB Kontinentales Tiefbohrprogramm der Bundesrepublik Deutschland was the Continental Deep Drilling Program of the Federal Republic of Germany with two, 4,000 m and 9,101-m-deep, boreholes drilled during the years 1987–1994.

Borehole monitoring is either the systematic direct collection of physical, chemical, rock mechanical and biological data from formations penetrated by boreholes thus covering various geological timescales or the indirect observation of information from greater depth, e.g., through elastic (seismic) waves. Installations of downhole instruments and subsequent data transmission to data centers for further processing are either permanent, e.g., cemented in place, or retrievable at the end of the observation period.

Scientific drilling encompasses an Earth science research program that utilizes drilling machines to produce a deep well, retrieve samples and acquire data from depth to gain novel scientific insights.

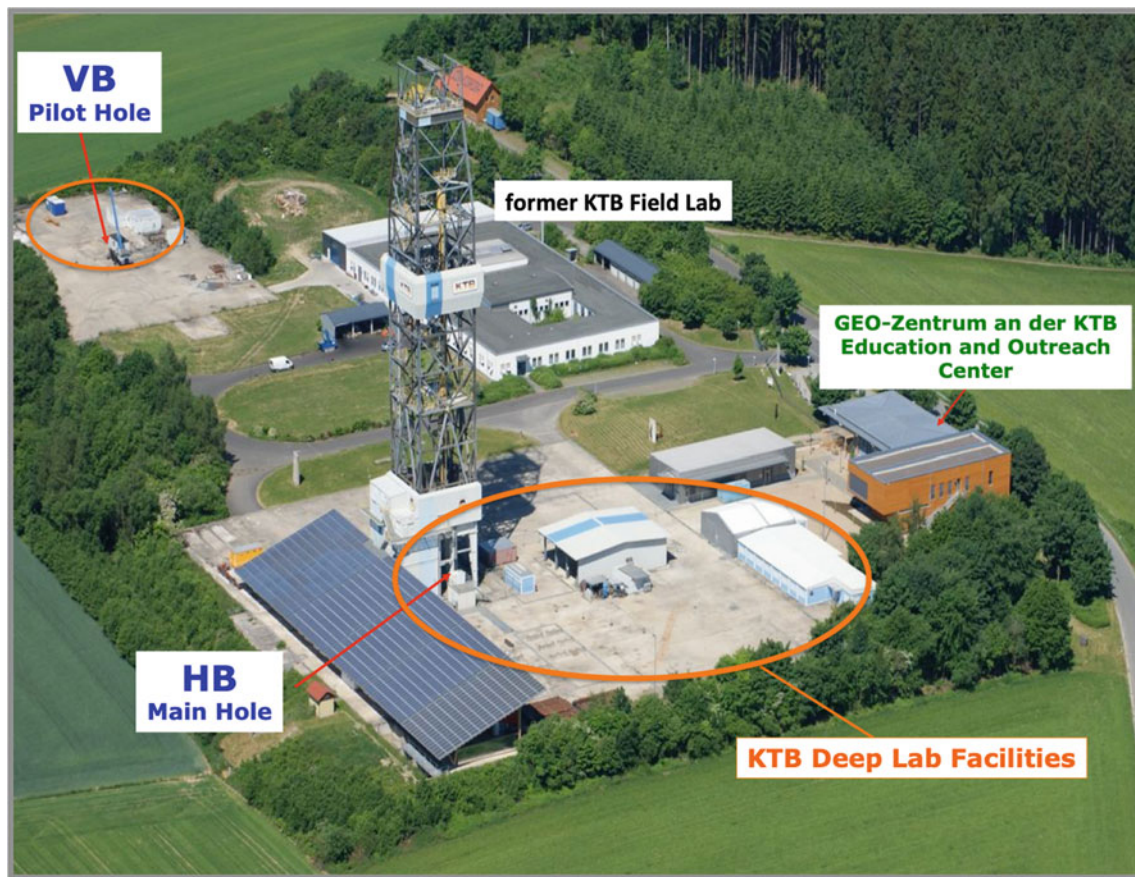
Introduction

Long-lasting and successful international scientific drilling projects are not only striving to provide direct access to the

deep underground for geoscientific research purposes but also to monitor in situ physical, chemical, and rock mechanical properties over prolonged time periods to understand the dynamics and processes inside the Earth. One of the paramount examples of the utilization of very deep research boreholes is the two 4- and 9.1-km-deep wells of the Continental Deep Drilling Program of Germany, in short KTB (Fig. 1). These wells have been drilled not only to understand the complex geological terrane structure of Central Europe and its related physical and thermoelastic properties but also to utilize them as windows into the deep underground for monitoring purposes. After 25 years of operating the KTB Depth Laboratory (KTB-DL) with still ongoing monitoring and testing activities, an assessment of possibilities and challenges for utilizing boreholes for Earth science is timely.

Scientific Background

The KTB project is based on a German Earth science community effort that started in 1978. Embedded in the framework of the International Lithosphere Program, research was focused on the continental crust through deep geophysical sounding and direct access through deep drilling. The German Federal Ministry for Research and Technology granted funding for KTB in 1982 with the objective to create a window into the upper crust of Central Europe. Research started with a 2-year preparatory phase including comparative geoscientific studies at five locations in West Germany and project studies for a deep well drilling program. Data gained and geological hypothesis advanced in this phase then resulted in a decision for a more detailed site survey at two locations, namely, the Upper Palatine (Oberpfalz) and the Black Forest (Schwarzwald). Studies at both locations, including drilling of several shallow boreholes, were executed and compiled from 1984 to 1986 and finally discussed in 1986, leading to a decision for the Upper Palatine site near the town of Windischeschenbach, Eastern Bavaria, SE Germany.



KTB Depth Laboratory: A Window into the Upper Crust, Fig. 1 Aerial view of the KTB-DL site with the KTB-VB pilot hole (upper left corner), KTB-HB main hole below the drill rig (foreground),

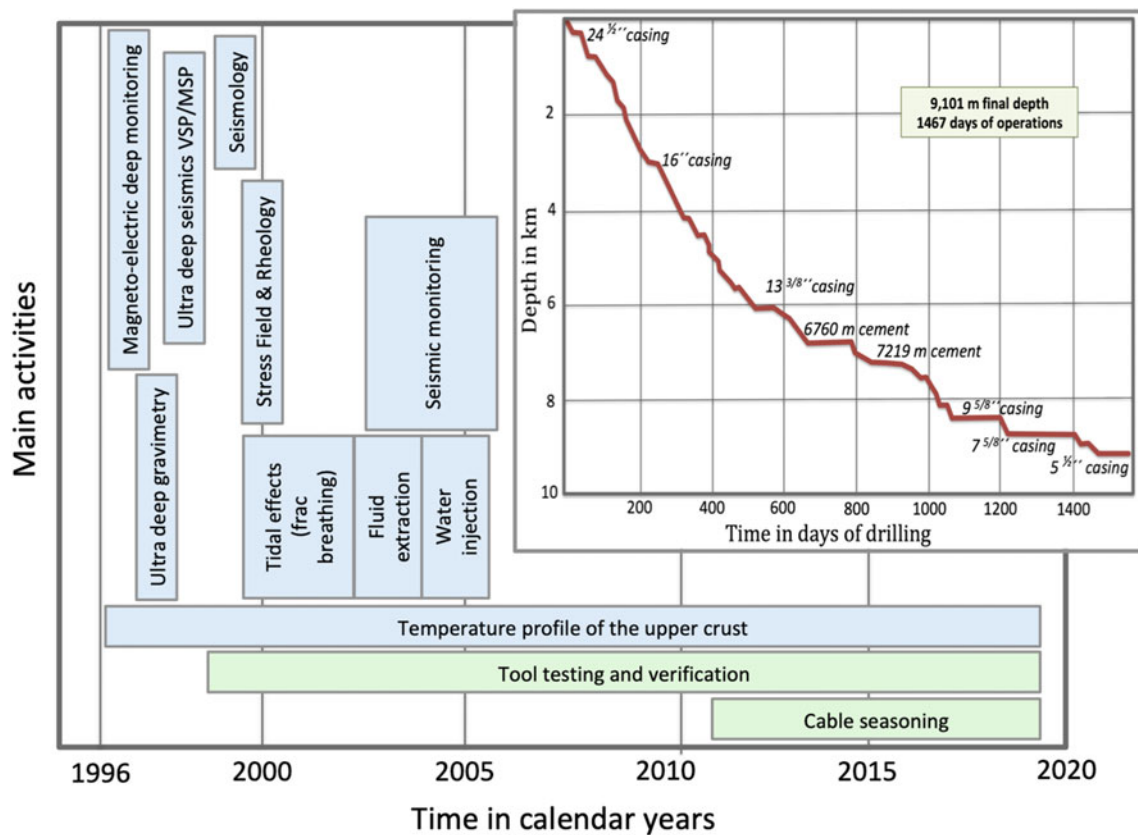
the former field lab in the main building behind the rig. The logging installations are within the orange circles

The first operational phase lasted from 1987 to 1990 with drilling of the 4,000.1-m-deep KTB-VB pilot hole, followed by a 1-year long phase to conduct measurements, testing, and evaluation, resulting in research concepts and a detailed engineering and management plan for the KTB-HB main borehole. Following this multiphase approach, the ultradeep KTB-HB (9,101 m) was then drilled from 1990 to 1994. Drilling, coring, and downhole logging of the deep KTB-HB (Figs. 2 and 3) required extensive and detailed technical planning and key technology developments. The two most outstanding advances were a fully automated drill pipe handling system reducing trip times about half and the vertical drilling systems providing an almost vertical borehole path to drastically minimize drill string friction. Both technologies, pipe handler and steering systems, have since become standards in deep drilling (Engeser 1996).

Ultradeep borehole measuring necessitated a wire line downhole logging system capable of safely handling depths beyond 8 km. This was achieved by installing a cable-friendly custom-made friction winch between the drum winch and the borehole and by running deep logs in tapered cable mode: a 1–2 km long, lightweight, HT temperature-resistant cable at

bottom and the long, heavy, and strong cable above (Kück et al. 2020).

The finalization phase in 1995 comprised the dismantling of most components of the drilling facilities, except for the logging equipment and the 80-m-high drill rig itself. Data evaluation, documentation, publication, and administrative phaseout completed this phase. In the following years, the two fully accessible, cased, and cemented deep wells were prepared as a deep crustal laboratory (KTB-DL) with concrete pads around the wells and a complete logging infrastructure including office and workshop space (Fig. 1). From 1996 on, the two KTB wells with logging and testing infrastructure were operated by the GFZ German Research Centre for Geosciences as depth laboratory. During the first 10 years, more than 250 wire line operations were performed at the KTB-DL, resulting in more than 1,000 km of logging runs. The establishment of a KTB-DL was an indispensable part of the overall KTB concept from the very beginning, because only long-term observations (months to years) allow for investigations of temporally varying processes under natural in situ conditions years after all drilling related disturbances have decayed. Such a borehole lab enables carrying out even



KTB Depth Laboratory: A Window into the Upper Crust, Fig. 2 Overview of the main activities in the KTB-DL since its start in 1996; blue = scientific experiments, green = technical activities,

insert: drilling progress chart of the KTB main hole showing delays during technical challenges below 6760 m depth

very long experiments at far lower costs compared to the very high daily costs of an active drilling.

The key scientific objectives of the KTB as defined already in 1983 (Emmermann and Lauterjung 1997) were:

1. To improve the understanding of the geological structure and evolution of the Central European Earth's crust
2. To study, sample, and monitor the geophysical structures and heterogeneities at depth and to compare them with results from surface investigations
3. To shed new light on the fluid system at depth, its pathways, and its influence on geological processes such as metamorphism
4. To determine the upper crustal temperatures, the geothermal gradient, and the crustal heat flow
5. To obtain data on stress and strain in the upper crust and better understand earthquake nucleation and propagation
6. To establish a "deep crustal laboratory" and allow for long-term observations at depth

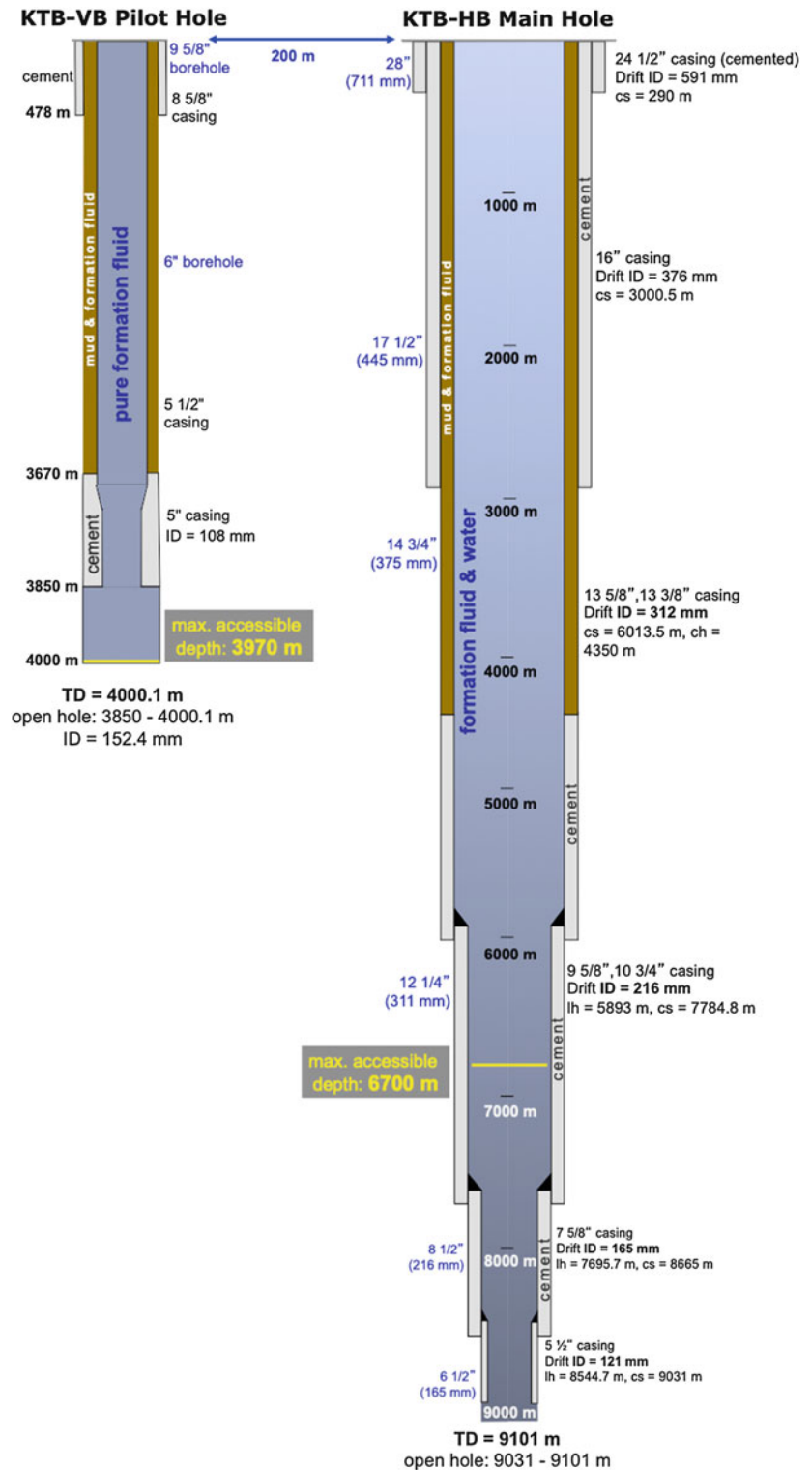
These targets were complemented by a large number of specific science goals as defined during the course of this long-lasting project by the various Earth science disciplines

involved. The overarching goal was to conduct basic geoscientific research on the physical and chemical conditions in the deeper continental crust and the acting processes aiming to provide new insight into the structure, the dynamics, and the evolution of an intracontinental crustal area. The different research fields directed their efforts to the verification of fundamental concepts in Earth sciences. One key example was the transition from brittle to ductile deformation in a temperature range from 250 °C to 300 °C where a fundamental change in rock parameters was expected to occur (Brudy et al. 1997). Others such as the proof of several kilometer-wide networks of permeable fractures in the deeper crust and the related temperature profile and heat flow are shown in Emmermann and Lauterjung (1997) and detailed references therein.

KTB Data Infrastructure

The utilization of the KTB boreholes as deep laboratory is based on the complex digital data record available for the site. Cores, cuttings, and fluids were regularly sampled in short intervals and are mostly still available to interested science parties.

**KTB Depth Laboratory:
A Window into the Upper
Crust, Fig. 3** Drilling and casing
scheme of the KTB boreholes
showing the current KTB Depth
Laboratory layout



Sample materials were investigated in great detail for a wealth of critical parameters in mineralogy, geochemistry, petrophysics, and geomechanics. These data and complementing downhole logging data, borehole tests, and surface geophysics including deep geophysical sounding and their joint

interpretation have been published and are digitally available for download. The data set was mainly acquired during the drilling phase in the KTB field laboratory (Emmertmann and Lauterjung 1997). KTB samples and data served and still serve to publish a large number of scientific papers. This wealth of

information available makes the cuboid around KTB with its size of about 10 km depth times several kilometers lateral extension one of the best explored crystalline basement blocks on Earth. The complete digital data sets including metadata and the related links to all published KTB reports are available from Kück et al. (2020).

KTB Depth Laboratory Utilization

The KTB-DL layout with two deep wells at just 200 m distance is shown in detail in Fig. 3, while the supporting surface infrastructure surrounding of the boreholes (downhole logging tools, winches, cables, workshops, offices, etc.) is indicated in Fig. 1. The main well is still accessible down to 6,700 m with c. 200 mm diameter instruments ($T_{\max} = 190^\circ\text{C}$ and $p_{\max} = 66\text{ MPa}$), while the pilot well is still fully accessible down to 3,970 m with 90 mm diameter devices (at $T_{\max}/p_{\max}$ of $115^\circ\text{C}/20\text{ MPa}$). This setup allowed for experiments and tests over more than two decades. Over the first 5 years, more than 200 weeks of experiments were performed in the wells, while over the last 5 years (2013–2018), about 60 weeks of operations happened on site. In summary all post-drilling activities at the KTB-DL can be ordered into four categories:

Geoscientific Experiments

During the first KTB-DL phase from 1996 to 2006, the KTB holes were intensely used with an uninterrupted sequence of scientific experiments that were planned ahead already during the KTB drilling time. Experiments carried out investigated eight main research areas: deep temperature gradient, gravity field (deepest gravity meter survey), tidal effects (frac breathing at 4 km) (Schulze et al. 2000), deep magnetic survey through steel casing, seismic imaging in crystalline rocks/deepest VSP/MSP (Rabbel et al. 2004), seismology (seismometer at 4 km for almost 2 years), rheology of the Earth's crust/fracture propagation from brittle to ductile below 9 km (Bohnhoff et al. 2004), and fluid transport. The duration of these experiments varied between a few weeks and 4 years. More than 200 scientists and technicians from 22 institutes and companies worked at the KTB-DL and the surrounding area during this time. The longest experiment lasting from 2002 to 2005 was a pumping and injection test, each part one and a half year long with the aim to investigate the fracture network intersected by the pilot hole at c. 3,995 m and the natural rock fluids therein. An amount of $23,000\text{ m}^3$ was produced from the well, and later almost $85,000\text{ m}^3$ of freshwater was injected into the crystalline rock (Fig. 2). Key results such as the far-reaching fluid pathways at depth with high fracture permeabilities have been published in a special volume of *Geofluids* (Erzinger and Stober 2005) and by Kümpel (2006).

Test and Calibration of New and Existing Instruments, Methods, and Installations

Since 2006 the boreholes are used mainly by GFZ and other research institutes for verification and testing of new and experimental instruments and methods as well as for technical works such as tool testing and equipment preparation before and after campaigns elsewhere. In the framework of GFZ's and other research institutes' downhole tool development programs and the ongoing utilizations of geophysical borehole sondes, the wells are regularly used for performance tests and to calibrate instruments, e.g., XRF-density/elements sonde, memory logging-while-tripping sondes, borehole heat exchange reactor, and active seismic prediction sonde. Future experiments will target at fracture-fluid systems in deeper regions ($>6,000\text{ m}$) and the deep biosphere in the formation fluid filled and since many years undisturbed 4-km-deep pilot hole with an open hole temperature at 110°C .

Commercial Utilization of the KTB Wells

Aside from the academic utilization of the KTB-DL, the holes are also used by a number of companies to test or verify new downhole methods, tools, cables, and winches under in situ conditions (temperature, pressure, in-hole cable length). Logging service companies frequently use the deep well to season new logging cables (reduce cable torque) before using them in actual logging jobs.

Outreach and Education

Based on the public outreach facility installed during drilling, a vibrant environmental and geoscientific education center with an on-site museum, the "GEO-Zentrum an der KTB," has been established at the KTB site since 1998. It is running continuously with communal and local state support as well as European Union funding. The number of visitors as of 2018 is in the range of 25,000 per year. The center serves as KTB museum with the KTB drill rig as landmark and focuses in addition on Earth and environmental education for teachers and school classes. It is also used for academic education such as for courses in the framework of the International Continental Scientific Drilling Program (ICDP).

The access to the KTB wells with conditions of utilization has been detailed by Harms and Kück (2016) at <http://www.gfz-potsdam.de/en/section/geomechanics-and-rheology/infrastructure/ktb-deep-crustal-lab/>.

Long-Term Utilization of a Deep Borehole Site

The majority of scientific drilling projects worldwide as well as commercial drillings have strict legal obligations concerning well completion and abandonment. This commonly requires closing the borehole safely by cutting casings, cementing the well, and other procedures to ensure hydraulic and gas sealing of

intersected reservoirs. For long-term use of boreholes, it is therefore necessary that a company or institute has to take responsibility for safety and final borehole closure and to commit to its enduring mining law responsibility. In the case of the KTB-DL the GFZ German Research Centre for Geosciences is legally responsible.

A crucial lifetime determining factor of borehole observatories is the long-term accessibility of the borehole, both logistically at surface and inside the well due to borehole wall instability, together with the availability of funds to work over collapsed hole sections. Only few deep drill holes stay stable over many years, and even a steel casing is in danger to be damaged and deformed by failing rock in a strong triaxial stress field. In general, the deeper the hole is, the higher is the risk of collapse which has to be expected.

Unsurpassed until today is the Kola Superdeep SG-3 in NW Russia with more than 12 km depth which had an observatory phase after the end of the drilling operations since 1994. It offered 8,500 m of the mainly cased well (Popov et al. 1999) also for international utilization. Despite highly qualified and experienced experts, the difficult and uncertain conditions after end of the Soviet Union led to an insignificant utilization and withdrawal of governmental support. The facilities were abandoned in the year 2008, and the infrastructure deteriorated soon after.

Fortunately, other borehole laboratories have derived from scientific drilling projects in various geological environments because the drilled rocks proved to be stable and bearing no special risks. Organizations like geological surveys, research institutes, or universities committed to remain responsible, to maintain sites, and to offer access to wells for protracted periods. Examples are the 2,500-m-deep Outokumpu borehole laboratory OKU-1 in Finland since 2006, the 2,500-m-deep COSC-1 underground lab near Åre in Sweden since 2014, and the 3,000-m-deep KFD1 hole of the Koyna drilling project in India since 2017, all supported by the ICDP. More depth laboratories can be expected in the coming years because long-term monitoring has become an inherent component of deep scientific drilling projects.

Summary

The KTB-DL after 25 years in operation continues to offer deep access to the upper crystalline crust. The infrastructure available and the stable conditions of the vertical 9191-m-deep KTB main hole (accessible to 6,700 m) and the 4000-m-deep KTB pilot hole (open hole section from 3,850 to 3,970 m) at only 200 m lateral distance invite for both ultra-deep geoscientific experiments and for technical use as test bed. The KTB-DL was used in the first 10 years after drilling primarily for further scientific investigations of fluids and seismicity while tool development and testing purposes prevail in the experiments in the recent past. Multi-purpose

utilization including science, commercial use and education and outreach as well as stable funding conditions have ensured 25 years of lifetime of the KTB wells. New depth laboratories and observatories have emerged in the past 10 years and more can be expected from upcoming deep scientific drilling projects.

Cross-References

- [Borehole Seismic Networks and Arrays](#)
- [Deep Scientific Drilling](#)
- [Seismic Monitoring of Nuclear Explosions](#)
- [TOPO-EUROPE: From the Deep Earth to the Surface of Continental Europe and Its Margins](#)

Bibliography

- Bohnhoff M, Baisch S, Harjes H-P (2004) Fault mechanisms of induced seismicity at the superdeep German Continental Deep Drilling Program (KTB) borehole and their relation to fault structure and stress field. *J Geophys Res* 109:B02309. <https://doi.org/10.1029/2003JB002528>
- Brudy M, Zoback MD, Fuchs K, Rummel F, Baumgärtner J (1997) Estimation of the complete stress tensor to 8 km depth in the KTB scientific drill holes: implication for crustal strength granites. *J Geophys Res* 102:18453–18475
- Emmerrmann R, Lauterjung J (1997) The German Continental Deep Drilling Program KTB: overview and major results. *J Geophys Res* 102:B818,179–B818,201. <https://doi.org/10.1029/96JB03945>
- Engeser B (ed) (1996) KTB report 95-3. Das Kontinentale Tiefbohrprogramm der Bundesrepublik Deutschland, KTB. Bohrtechnische Dokumentation, the continental deep drilling project of the Federal Republic of Germany. Documentation of the used drilling technology, Niedersächsisches Landesamt für Bodenforschung. ISBN 978-3-928-55916-4
- Erzinger J, Stober I (2005) Introduction to special issue: long-term fluid production in the KTB pilot hole, Germany. *Geofluids* 5:1–66. <https://doi.org/10.1111/j.1468-8123.2004.00107.x>
- Harms U, Kück J (2016) Superdeep tests and experiments at 9.1 km and 4 km. *J Large Scale Res Facil* 2:A75. <https://doi.org/10.17815/jlsrf-2-132>
- Kück J, Conze R, Harms U (2020) Data from the German Continental Deep Drilling Project (KTB, Kontinentale Tiefbohrung). GFZ Data Services. <https://doi.org/10.5880/GFZ.KTB.top>
- Kümpel H-J, Erzinger J, Shapiro S (2006) Two massive hydraulic tests completed in deep KTB pilot hole. *Sci Drill* 3:40–44. <https://doi.org/10.2204/iodp.sd.3.05.2006>
- Popov Y, Pevzner S, Pimenova V, Romushkevicha R (1999) New geothermal data from the Kola superdeep well SG-3. *Tectonophysics* 306:345–366. [https://doi.org/10.1016/S0040-1951\(99\)00065-7](https://doi.org/10.1016/S0040-1951(99)00065-7)
- Rabbel W, Beilecke T, Bohlen T, Fischer D, Frank A, Hasenclever J, Borm G (2004) Superdeep vertical seismic profiling at the KTB deep drill hole (Germany): seismic close-up view of a major thrust zone down to 8.5 km depth. *J Geophys Res* 109(Issue B9). <https://doi.org/10.1029/2004JB002975>
- Schulze K, Kümpel H-J, Huenges E (2000) In-situ petrohydraulic parameters from tidal and barometric analysis of fluid level variations in deep wells: some results from KTB. In: Stober I, Bucher K (eds) *Hydrogeology of crystalline rocks. Water science and technology library*, vol 34. Kluwer, Rotterdam, pp 79–104. https://doi.org/10.1007/978-94-017-1816-5_4

Legal Continental Shelf: Geology, Geophysics, and Tectonics

Elana Geddis¹, Ray Wood² and Vaughan Stagpoole³

¹Harbour Chambers, Wellington, New Zealand

²CRP-OCS Ltd, Haumoana, New Zealand

³GNS Science, Lower Hutt, New Zealand

Synonyms

Continental margin; Continental shelf; Extended continental margin; Extended continental shelf

Definition

The legal continental shelf comprises the submerged prolongation of the land mass of a State beyond 200 nautical miles (M) from the territorial sea baselines, consisting of the seabed and subsoil of the shelf, the slope, and the rise. Formulae to determine the outer limits of the legal continental shelf are in Article 76 of the United Nations Convention on the Law of the Sea (UNCLOS). Application of the Article 76 definition frequently turns on an understanding of the geology, geophysical structure, and tectonic evolution of the sea floor.

Introduction

A State has exclusive rights over the seabed resources of its legal continental shelf as the natural prolongation of its land territory. The assertion of those rights is an inherent entitlement, not a claim.

In 1982 UNCLOS imposed a legal framework on the scientific concept of the continental shelf, creating a clear definition of the outer limits of the legal continental shelf

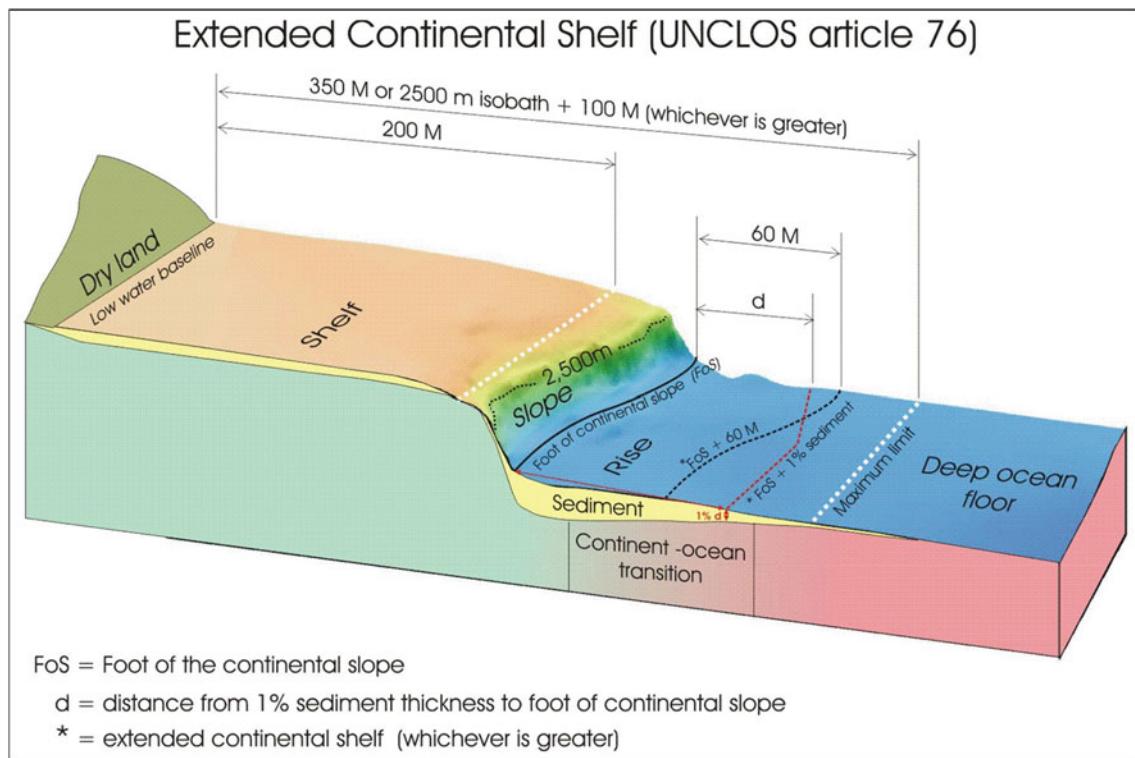
and a process for international verification of those limits. These are contained in Article 76 of UNCLOS and Annex II of the Final Act of the Third United Nations Conference on the Law of the Sea (the Statement of Understanding). Under Article 76 the continental shelf is defined as extending to the outer edge of the continental margin, subject to two constraints.

Article 76 was drafted to deal with the vast majority of continental margins, but it does not adequately deal with margins characterized by a narrow and steep morphological shelf and slope and an extraordinarily extensive rise. The Statement of Understanding extends the principles of Article 76 to those margins.

Where the outer limits of the continental shelf extend beyond 200 M from the coast, a State must make a submission to the UN Commission on the Limits of the Continental Shelf (CLCS) to verify that the outer limits have been calculated correctly. The recommendations issued by the CLCS in response to States' submissions address the real world complexities of continental margins and continue to refine the understanding of the legal continental shelf and the scientific concepts that underpin it.

Defining the Outer Edge of the Continental Margin and the Limits of the Continental Shelf

Article 76 defines the continental margin to include the shelf, the slope, and the rise as the “submerged prolongation of the land mass” distinct from the “deep ocean floor.” The outer edge of the continental margin is determined through the application of two formulae, both calculated from the foot of the continental slope (Fig. 1). The outer edge of the margin lies at points 60 M from the foot of the slope or where the thickness of sediment is 1% of the distance from the foot of the slope. In the absence of evidence to the contrary, the foot of the slope is identified at the point of maximum change in seafloor gradient.



Legal Continental Shelf: Geology, Geophysics, and Tectonics, Fig. 1 Cartoon showing a hypothetical continental margin, summarizing the formulae and constraints described in Article 76 that determine the outer limits of the continental shelf. (Modified from Kapoor and Kerr 1986)

Article 76 also constrains the legal continental shelf to no more than 350 M from the coast or 100 M from the 2500 m isobath, whichever is greatest. On margins with expansive, relatively shallow slopes the 2,500 m + 100 M constraint can lie seaward of the 350 M line. Significantly, the 2,500 m + 100 M constraint applies to “submarine elevations that are natural components of the margin” but does not apply to “submarine ridges” (limiting these to 350 M). The distinction between “submarine elevations” and “submarine ridges” remains both conceptually and analytically difficult. Although the CLCS Technical Guidelines (1999) reveal general principles that are considered, each feature must be assessed individually in terms of its morphological, geological, and tectonic continuity (Brekke and Symonds 2011).

Applying the Article 76 definition requires four steps (Fig. 2): (1) the outer edge of the continental margin is identified using the two formulae; (2) the two constraint lines are determined; (3) the most seaward constraint line is applied; and (4) the resulting line composed of formulae and/or constraint lines is connected by lines up to 60 M long (Wood et al. 2011). These steps are described in the CLCS Technical Guidelines (1999).

Margins with a narrow morphological shelf, steep slope, and a very extensive rise may meet the criteria under the

Statement of Understanding. The Statement of Understanding has four elements that identify continental margins to which it applies. For States that meet its criteria, the Statement of Understanding, in conjunction with the CLCS Technical Guidelines (1999), defines the outer limits of the continental shelf at points where the sediment thickness is at least 1 km.

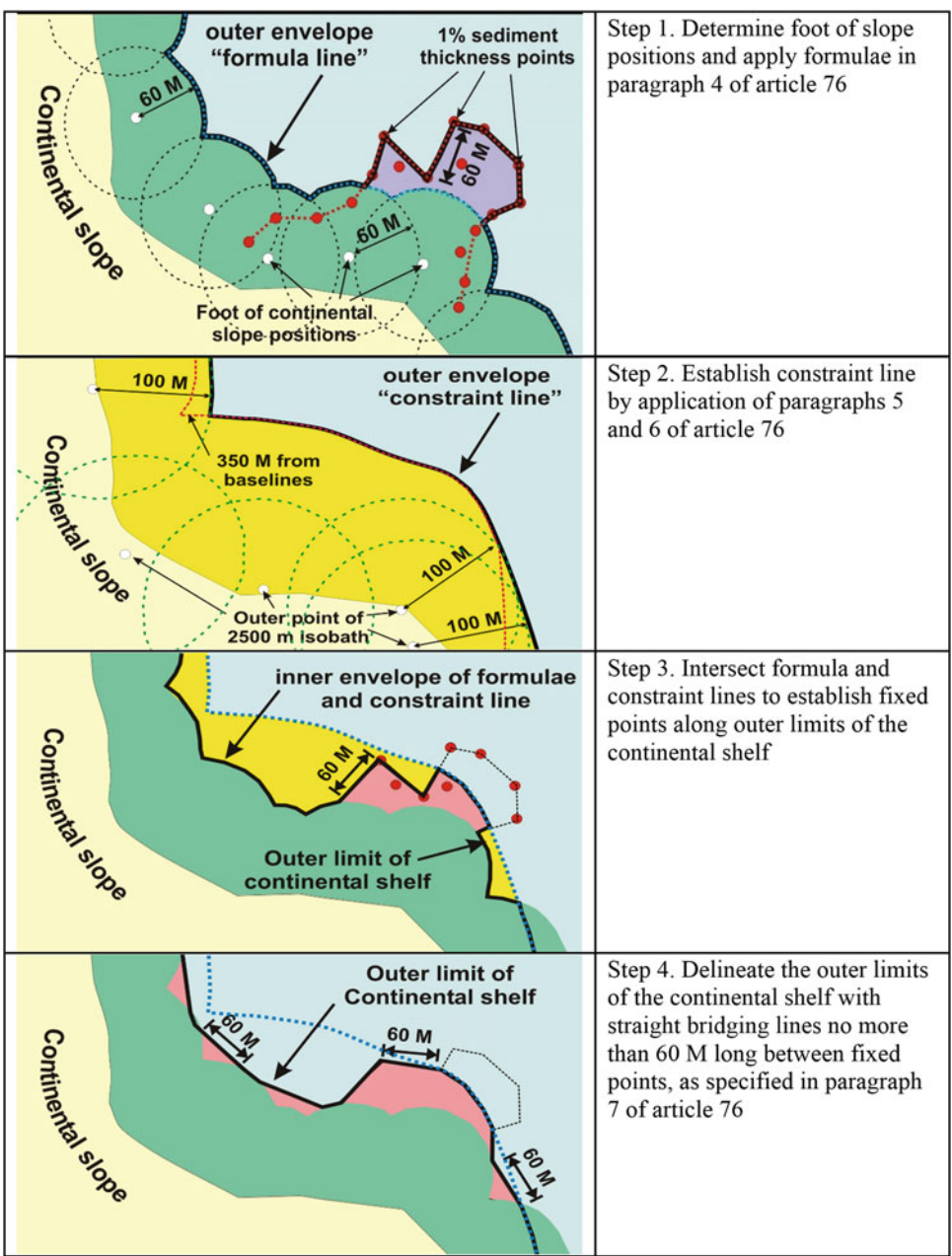
Prolongation of the Land Mass

The concept of the submerged (or natural) prolongation of the land mass is central to the definition of the legal continental shelf. Establishing where prolongation occurs and the nature of that prolongation are of primary importance because they determine (a) whether a legal continental shelf exists and (b) how far it extends beyond 200 M (Wood et al. 2011).

Natural prolongation has three aspects (Wood et al. 2011):

- Morphologic – the seafloor shape is a continuation of the land mass morphology.
- Geologic – the rocks beneath the seafloor are the same as, or related to, those of the land mass.
- Tectonic – the rocks beneath the seafloor share their history with those of the land mass.

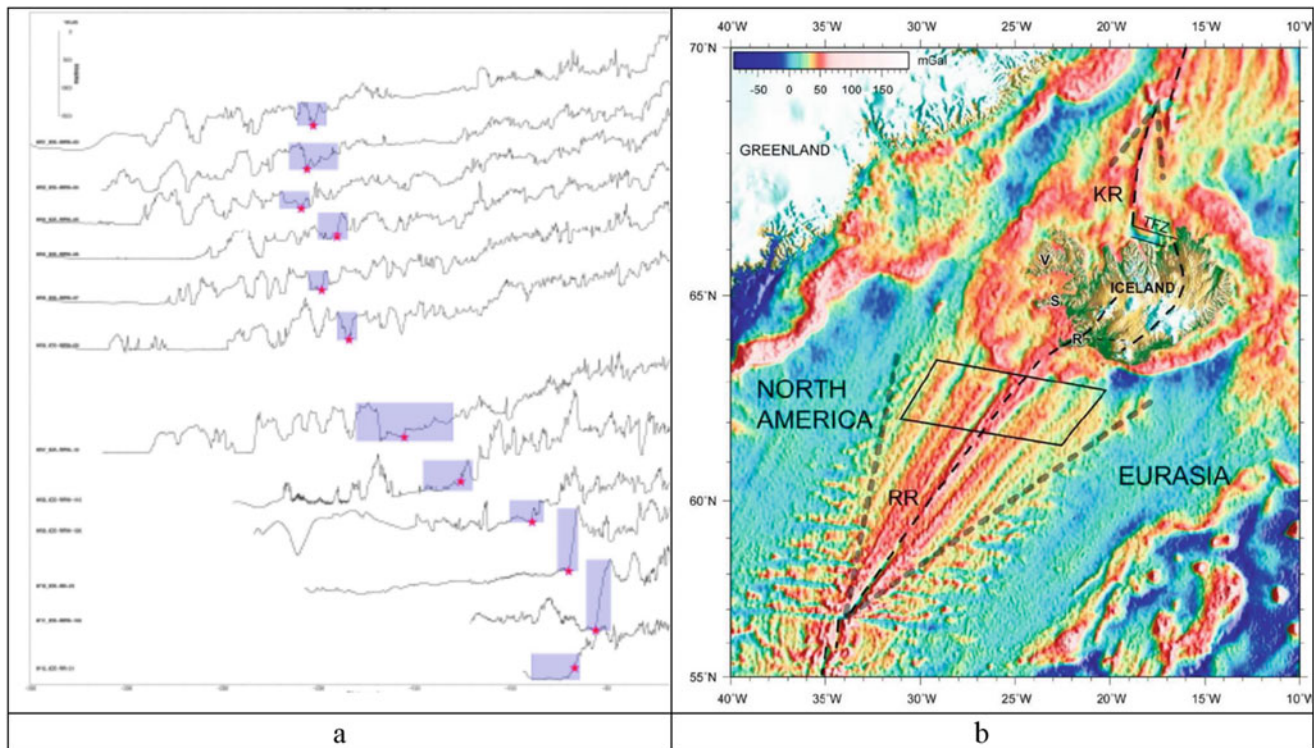
Legal Continental Shelf: Geology, Geophysics, and Tectonics, Fig. 2 The four steps for establishing the outer limits of the continental shelf. (From Wood et al. 2011)



The transition from land mass to deep ocean floor is the result of the composition and density of the rocks beneath the seafloor, the geological processes that form and shape them, and the tectonic forces that act on them (Wood et al. 2011). The morphological transition is relatively simple on many margins, with a shelf dipping gently to a depth of several 100 m, a steeper slope deepening to a depth of several 1,000 m, and a gently dipping rise lying between the slope and the flat deep ocean floor (Fig. 1). Other margins have a complex morphology with ridges, seamounts, and canyons, and the transition with the deep ocean floor can be harder to recognize (Fig. 3).

Role of Geology and Tectonics

The Article 76 definition divides the Earth into land territory, the continental margin, and the deep ocean floor with its oceanic ridges. Applying that definition to a particular margin often requires complex scientific analysis. On simple margins the shelf, the slope, and the rise may be easily identified from the seafloor morphology. The primacy of seafloor morphology is reinforced by both the CLCS Technical Guidelines (1999) and subsequent CLCS recommendations. But frequently geology, geophysics, and tectonics will be important in interpreting that morphology to:



Legal Continental Shelf: Geology, Geophysics, and Tectonics, Fig. 3 (a) Foot of slope positions along Reykjanes Ridge determined from analysis of (b) multibeam data and regional tectonics. (From CLCS 2016)

- Distinguish between areas that are part of the continental margin and those that are part of the deep ocean floor
- Distinguish between “submarine ridges” and “submarine elevations that are natural components of the margin”
- Clarify the location of the foot of the continental slope

The role of geology and tectonics in the application of Article 76 is demonstrated by examples from the recommendations on the submissions by Iceland, Ascension Island (the United Kingdom), Bouvetøya and Shaka Ridge (Norway), Ogasawara Composite High (Japan), and the Kerguelan Plateau (Australia).

Iceland is the product of voluminous volcanism related to the interaction of a mantle hotspot with a seafloor spreading ridge. Because its origin is similar to that of the volcanism that forms the deep ocean floor, geologically it is difficult to identify the boundary of the continental margin. In its recommendations the CLCS concluded that hotspot-ridge interaction had significantly changed the seafloor spreading process and the morphology of parts of the Reykjanes Ridge (CLCS 2016). It therefore considered that the region defined by the Iceland hotspot interaction with the seafloor spreading process is part of the continental margin of Iceland for the purposes of Article 76.

Even given this criterion for identifying the outer limits of the continental margin, finding those limits is not

straightforward. Figure 3a shows the foot of slope points along the Reykjanes Ridge that were accepted by the CLCS. It is apparent that their choice is not based entirely on morphology but also on identification of the V-shaped seafloor region formed by hotspot activity (Fig. 3b). Significantly, the CLCS could not agree on whether the continental margin is a “submarine ridge” or a “submarine elevation” for the purposes of Article 76.

A similar argument was used by the United Kingdom for prolongation from Ascension Island. The United Kingdom submitted that Ascension Island is an integral component of the Mid-Atlantic Ridge and its natural prolongation therefore extends along the ridge. The CLCS agreed that islands surmounting ridges rising from the deep ocean floor are entitled to a legal continental shelf. But it determined that the rugged seafloor surrounding the Ascension Island volcanic edifice is part of the deep ocean floor and there is no morphological feature in the vicinity of Ascension Island that could be characterized as continental margin extending beyond 200 M (CLCS 2010).

In summary, Iceland was able to demonstrate that the volcanic processes that formed the land mass affected areas of the surrounding seafloor so that they are morphologically and geologically distinct from the adjacent deep ocean floor. Iceland could therefore show the continental margin extended beyond 200 M. Ascension Island, on the other hand, could not

identify a geological distinction between the volcanic edifice on which the island sits and the Mid-Atlantic Ridge.

Similar to Ascension Island, Bouvetøya is a volcanic island on the Southwest Indian Ridge in the South Atlantic, a seafloor spreading ridge. Bouvetøya and the elevated area of seafloor extending northeast from it to the eastern end of Shaka Ridge are interpreted to have formed by hotspot related volcanism.

On the basis of seafloor morphology and the geochemistry of the basement rocks, the CLCS agreed that there was natural prolongation of Bouvetøya to the eastern end of Shaka Ridge (CLCS 2019). In addition, the CLCS relied on geochemical analysis to conclude that the Shaka Ridge is a “submarine elevation” that forms a natural part of the Bouvetøya margin, distinct from the oceanic basalts of the deep ocean floor of the Southwest Indian Ridge.

Geology and tectonics also played an interesting role in determining the outer limits of Japan’s continental shelf around the Ogasawara Composite High, on the eastern flank of the Izu-Ogasawara Arc southeast of the islands of Japan. The Ogasawara Composite High is composed of several east-west trending seamounts, and straddles an active subduction trench and volcanic arc.

The presence of an active subduction trench could be considered a fundamental obstacle to establishing prolongation. But the CLCS identified an elevated morphological saddle across the subduction trench, which it considered sufficient to demonstrate prolongation to the high (CLCS 2012). This was reinforced by its acceptance that, based on the presence of thrust faulting, the western part of the high had been accreted to the Izu-Ogasawara Arc, and thus is a “submarine elevation” forming a natural part of the Japanese margin. The eastern part of the high, by contrast, had not been accreted and was concluded by the CLCS to be a “submarine ridge.”

In Australia’s continental shelf submission on the Kerguelan Plateau region, geology was used to distinguish between elements of the plateau region which are “submarine elevations” and those which are “submarine ridges.”

Heard and MacDonald Islands are embedded within the late Cretaceous magmatic crust of the Central Kerguelan Plateau. These basement rocks and those of the morphologically connected Southern Kerguelan Plateau show affinity with continental crust based on geochemical evidence. The CLCS therefore agreed that the Central and Southern Kerguelan Plateau regions are “submarine elevations” and natural components of the continental margin of Heard and MacDonald Islands (CLCS 2008).

However, the Commission decided that there is insufficient data to establish the geological origin of Williams Ridge, a ridge extending southeast from the Central Kerguelan Plateau. Therefore, even though it is morphologically part of the margin, the CLCS considered it is a “submarine ridge,” not a natural component of the continental margin.

Summary

The definition of the legal continental shelf in Article 76 enables States to define the outer limits of the continental shelf with geographic precision. The elements of the definition, and the discipline imposed by the requirement to submit outer limits to the CLCS for review, have required States to study the geology, geophysical structure, and tectonic evolution of the continental margin in new detail. This is particularly important for complex margins where the shelf, slope, and the rise cannot be easily identified from the seafloor morphology, and there is uncertainty in distinguishing the transition from land mass to deep ocean floor.

The examples from Iceland, Ascension Island, Bouvetøya and Shaka Ridge, the Ogasawara Composite High, and the Kerguelan Plateau show the value of tectonic and geological evidence when applying the Article 76 definition. Although seafloor morphology remains the primary evidence, there is in many cases a proven need to consider a range of geological, geochemical, and geophysical evidence to determine the outer limits of the legal continental shelf.

Bibliography

- Brekke H, Symonds P (2011) Submarine ridges and elevations of Article 76 in light of published summaries of recommendations of the Commission on the Limits of the Continental Shelf. *J Ocean Dev Int Law* 42:289–306
- Commission on the Limits of the Continental Shelf (1999) Scientific and technical guidelines of the Commission on the Limits of the Continental Shelf (CLCS/11). https://www.un.org/Depts/los/clcs_new/commission_documents.htm. Accessed 31 July 2019
- Commission on the Limits of the Continental Shelf (2008) Recommendations of the Commission on the Limits of the Continental Shelf (CLCS) in regard to the submission made by Australia on 15 November 2004. https://www.un.org/Depts/los/clcs_new/submissions_files/submission_austr.htm. Accessed 31 July 2019
- Commission on the Limits of the Continental Shelf (2010) Summary of recommendations of the Commission on the Limits of the Continental Shelf in regard to the submission made by the United Kingdom of Great Britain and Northern Ireland in respect of Ascension Island on 9 May 2008. https://www.un.org/Depts/los/clcs_new/submissions_files/submission_gbr.htm. Accessed 31 July 2019
- Commission on the Limits of the Continental Shelf (2012) Summary of recommendations of the Commission on the Limits of the Continental Shelf in regard to the submission made by Japan on 12 November 2008. https://www.un.org/Depts/los/clcs_new/submissions_files/submission_jpn.htm. Accessed 31 July 2019
- Commission on the Limits of the Continental Shelf (2016) Summary of recommendations of the Commission on the Limits of the Continental Shelf in regard to the submission made by Iceland in the Ægir Basin area and in the western and southern parts of Reykjanes Ridge on 29 April 2009. https://www.un.org/Depts/los/clcs_new/submissions_files/submission_isl_27_2009.htm. Accessed 31 July 2019
- Commission on the Limits of the Continental Shelf (2019) Summary of recommendations of the Commission on the Limits of the Continental Shelf in regard to the submission made by Norway in respect of Bouvetøya and Dronning Maud Land on 4 May 2009. https://www.un.org/Depts/los/clcs_new/submissions_files/submission_nor_2009.htm. Accessed 31 July 2019

un.org/Depts/los/clcs_new/submissions_files/submission_nor_30_2009.htm. Accessed 31 July 2019

Kapoor DC, Kerr AJ (1986) A guide to maritime boundary delimitation. Carswell, Toronto, p 123

Wood R, Henrys S, Stagpoole V, Davy B, Wright I (2011) Legal continental shelf. In: Encyclopedia of solid Earth geophysics, 1st edn. Springer, Heidelberg

Lithosphere, Continental

David E. James

Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington, DC, USA

Definition and Introduction

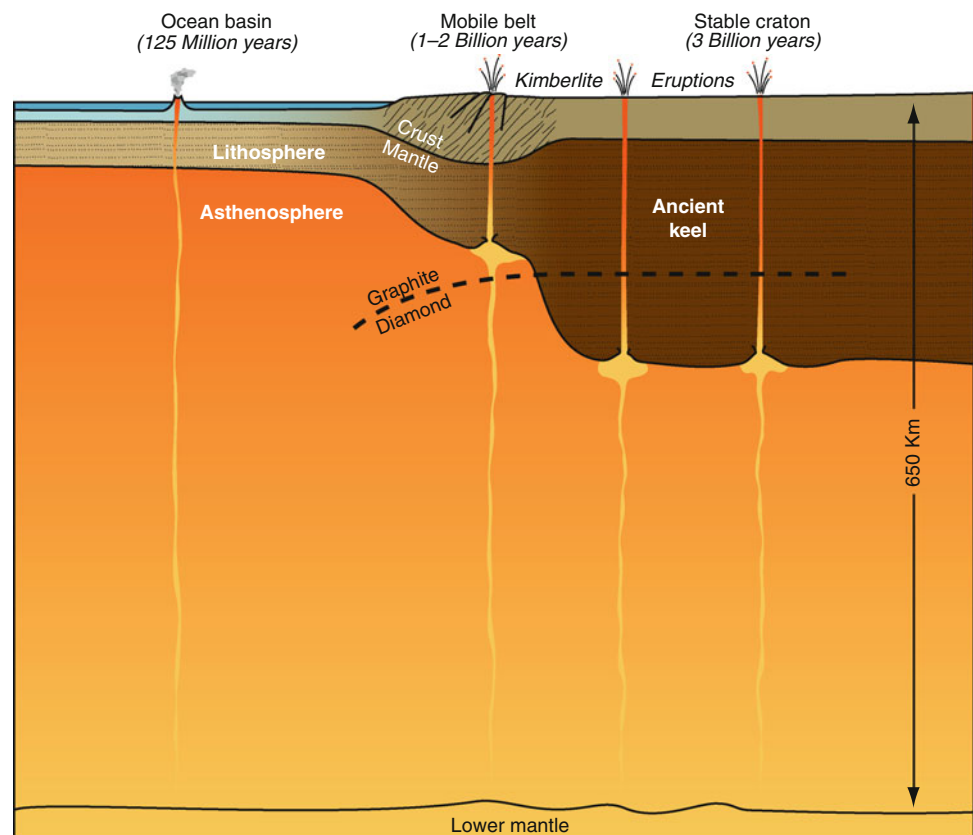
The continental lithosphere consists of the continental crust and, typically, some nonconvecting part of the underlying upper mantle (Fig. 1). In plate tectonics terms, the continental lithosphere is part of the rigid outer rind of the Earth, which is segmented into several major plates. The cold lithosphere lies atop a hotter, more mobile (low strength) asthenosphere. In this context, oceanic lithosphere for the most part obeys comparatively simple thermal models, with the boundary

between lithosphere and asthenosphere represented by an isotherm marking the transition in mantle peridotite between elastic and ductile behavior (but see ► [“Lithosphere, Oceanic”](#) for the growing list of complications). By contrast, the continental lithosphere is heterogeneous and its structure highly variable. That contrast reflects the fact that oceanic lithosphere is formed in much the same way worldwide and that it is continually recycled into the interior of the Earth via the plate tectonic “conveyor belt.” Continental lithosphere, on the other hand, is generally too buoyant to be subducted, although an increasing number of instances have been reported in which parts of continental lithospheric mantle show evidence of gravitational instability, forming “drips” where lithosphere delaminates and sinks into the deeper mantle (Zandt et al. 2004). The detachment of continental mantle from the crust in such circumstances may be an important factor in intraplate deformation and volcanism (Carlson et al. 2005). Thus, while no known oceanic lithosphere is older than about 200 my, the continental lithosphere has evolved over billions of years, one consequence of which is that it is seen today as the resulting collage of many superimposed tectonic events, including rifting, subduction, continental collision, accretion, and hot spot magmatism, that recur repeatedly over geologic time.

The study of continental lithosphere has been revolutionized over the past two decades with the advent of portable

Lithosphere, Continental,

Fig. 1 Schematic cross section showing transition from oceanic lithosphere to younger continental lithosphere and into the thick, cratonic lithosphere (tectosphere) beneath the ancient continental nuclei. Melts rising from depth ascent through relatively thin oceanic and young continental lithosphere to produce volcanoes at the surface, whereas thick tectosphere acts as a barrier to melts percolating from depth, requiring evolved, volatile-rich kimberlitic magmas to bore a conduit through the mantle to the surface



broadband seismology. Passive array experiments (see ► [“Seismological Networks”](#)) have made it possible to obtain high-resolution images of deep lithospheric structures beneath every continent (James 2007). The greatest concentration of high-density broadband deployments has been in western North America, Europe, and Asia, with large-scale experiments also in Africa, Central and South America, and Australia. Within the USA, the Transportable Array of the national EarthScope project has already resulted (at the time of this writing) in uniformly high-resolution seismic imaging of the mantle beneath the western half of the nation. The lithosphere so revealed consists of a collage of overlapping structures produced by ongoing subduction, trench migration, ridge override, regional extension, mantle upwellings and volcanism, and lithospheric delamination. In some tectonically active regions, a thin crust is the only lithosphere remaining intact (Carlson et al. 2005). This complex and disrupted lithosphere is in marked contrast to that of stable cratonic regions, including the continental nuclei of north-central North America, western Australia, and southern Africa. The cratons of southern Africa are perhaps the best studied of the Archean cratons of the world, where results from large-scale passive array experiments show that lithospheric roots as deep as 250 km or more underlie the undisturbed parts of the cratons (James and Fouch 2002; Ritsema and van Heijst 2000).

Despite remarkable progress in imaging the continental lithosphere over the past 2 decades, a precise definition of continental lithosphere remains elusive. No single definition of continental lithosphere can be made to fit all circumstances. Particularly controversial is the definition of what constitutes the base of the lithosphere. Many definitions of continental lithosphere appear in the literature – seismological, mechanical, rheological, thermal, and compositional. Depending on one’s view, the continental lithosphere may be a *mechanical boundary layer*, a *thermal boundary layer*, or a *chemical boundary layer*. In the remainder of this entry, we shall explore these various concepts in some detail. Our goal here is to describe the continental lithosphere in terms of global Earth structure, plate tectonics, and the long-term evolution of the continents. By this view, the continental lithosphere is envisaged to be a long-lived plate tectonic unit, consisting both of continental crust and, typically, a very significant mantle “keel” that is attached to and translates with the continent. In this sense, it is seen as a stable component of the continent and does not participate in the convective processes of the deeper dynamic mantle.

Mechanical Models

A simple expression of the lithosphere is *elastic thickness*. While rarely used to describe continental lithospheric in a

geologic or structural sense, elastic thickness is a useful concept in that it models the mechanical response of the lithosphere as an elastic plate overlying a weak (asthenospheric) substrate. As discussed in detail elsewhere in this volume (see ► [“Lithosphere, Mechanical Properties”](#)), the elastic thickness of the lithosphere may be determined through topographic response to surface loads (such as volcanoes), the relationship between topography and Bouguer gravity anomalies, or the shape of depositional surfaces in basins. The thickness of elastic lithosphere so determined is rarely as much as 100 km and in areas of extension or hotspot activity may be less even than crustal thickness. The elastic thickness of the continents is chiefly controlled by the thermal state of the lithosphere (see ► [“Lithosphere, Mechanical Properties”](#)). As Forsyth (1989) points out, however, the apparent elastic thickness is also affected by the state of stress and the rheological stratification in the crust and mantle, so mechanical thickness may not bear a simple relationship to a particular isotherm.

One important aspect of the mechanical structure of the continental lithosphere as it relates to lithospheric evolution is the large decrease in strength of crustal rocks in the deep crust. The steep geothermal gradient in the crust produces increasing ductility of crustal rocks with depth. The brittle zone in tectonic areas typically extends only to about 15 km, below which the rocks deform plastically. The topmost part of the underlying mantle, however, is relatively much stronger as temperatures there are still far from the melting point of mantle rocks. Thus, it is widely posited that there is a zone of weak lower crust sandwiched between strong upper crust and strong upper mantle, creating a possible zone of crust–mantle decoupling (see ► [“Lithosphere, Mechanical Properties”](#)).

While the flexural lithosphere is a useful *mechanical* concept, it fails as a plate tectonic description of lithosphere. In many tectonic regions, the flexural lithosphere is not even as thick as the continental crust. As low-density sialic crust is not easily subducted into the mantle, it must persist as the primary long-lived component of continental lithosphere. From the perspective of plate tectonics, it may be argued that the full definition of the lithosphere should account both for the long-term structural coherence of continental plates and the impact of asthenospheric processes that affect either the formation or the removal of lithospheric mantle beneath the crust.

Thermal Models

A convenient definition for *thermal lithosphere*, or the thermal boundary layer, is the outer layer of the Earth in which heat transfer is dominated by conduction (see ► [“Heat Flow, Continental”](#) and ► [“Lithosphere, Continental: Thermal Structure”](#)). Thus, the stronger lithosphere acts as a barrier

to thermally induced buoyancy forces that drive convective heat transfer in the underlying asthenosphere (Morgan 1984). In the case of the oceanic lithosphere, this conductively cooled layer of the Earth is thought to have finite thickness (i.e., the lithosphere ceases to undergo thermal contraction as it ages beyond about 80 Ma, implying that heat is supplied to the lithosphere from the underlying asthenosphere) (see ► “Lithosphere, Oceanic” and ► “Lithosphere, Oceanic: Thermal Structure”). For the continents, the situation is far more complex. The continental lithosphere does not undergo simple monotonic cooling, but may be subjected to repeated episodes of thermal or tectonic disturbances. Thus, superimposed upon the geological age of the lithosphere is the thermal or tectonic age of the lithosphere, factors that have led to significant compositional heterogeneity in both crust and lithospheric mantle.

We consider first the thickness of the thermal boundary layer beneath the stable cratonic cores of the continents. Here, the lithosphere is both at its thickest and its strongest. If the effective viscosity of the asthenosphere beneath the cratonic lithosphere is taken to be about 10^{21} Pa s, based upon glacial unloading and gravity/topography correlations (see ► “Mantle Viscosity”), it implies a temperature of transition from conductive to convective heat transfer of about 1300–1400 °C for typical mantle compositions (Morgan 1984). If crustal heat production is taken into account, the thickness of the thermal lithosphere of the stable continental cratons as calculated from surface heat flow may range from about 90 to 220 km (Rudnick et al. 1998).

The lithospheric thickness obtained by thermal modeling depends upon other factors, the most important of which is the composition, including volatile content, of the subcrustal mantle. If the continental lithospheric mantle consists of low-density peridotite depleted of its low-melting basaltic fraction and devolatilized over time due to metamorphic and magmatic events (see discussion below), the continental mantle will be less dense and substantially more refractory than the rest of the mantle. The chemical boundary layer represented by this depleted subcontinental mantle peridotite may stabilize the continental mantle root against thermal disruption to depths below the thermal boundary layer (Jordan 1978, 1981).

Tectono-thermal events, such as those that are widespread across the western USA, will have the effect of thinning the existing lithosphere. Regions where prior thermal disturbances have occurred, therefore, may be more prone to reactivation, whereas stable cratons, particularly those buffered by bordering Proterozoic mobile belts as in southern Africa, should be comparatively less vulnerable to thermal perturbations from the underlying asthenosphere (Pollack 1986). Tectonic reactivation seems to be particularly true of continental rifting, where repeated episodes apparently recur along the same long-lived zones of weakness. The greatly

expanded scope of tomographic studies in recent years has revealed a number of localities in which various forms of lithospheric delaminations, or “drips” appear to be occurring, including a number in the western USA (e.g., West et al. 2009; Zandt et al. 2004). Such processes suggest that lithospheric mantle beneath tectonically active areas of the continent may be thinned or removed and later regenerated. Indeed, a wide range of epeirogenic (vertical) movements within continental interiors have been interpreted to be due to thermal disturbances that produce lithosphere heating and thinning, followed by conductive relaxation and asymptotic thickening of the lithospheric thermal boundary layer.

Seismological/Compositional Models: Tectosphere

It is well known that average seismic velocities, especially shear velocity, are much higher under cratons than under oceanic or tectonically active areas. This was first shown by surface wave dispersion measurements and later by measurements of vertically travelling ScS phases. Sipkin and Jordan (1980) showed that ScS one-way travel-time anomalies associated with the upper mantle beneath stable continental interiors could be as large as 3.5 s (fast) relative to the upper mantle beneath the western Pacific. Moreover, the high-velocity continental paths are also associated with high Q_s , suggesting that the anomalies are due to thermal and/or compositional variations in the upper mantle (Jordan 1981). Since the early studies of Sipkin and Jordan, numerous other studies based on three-dimensional inversion of travel-time anomalies for velocity structure have been carried out regionally across the globe (see summaries in Carlson et al. 2005; James 2007). The results nearly all show that regional velocity contrasts, even between provinces entirely within the continents, extend to at least 200–250 km (James 2007).

The notion of a long-lived thermochemical boundary layer in the upper mantle beneath vast regions of Precambrian crust raised many difficult questions as to how they formed and how they survived over geologic time and to a significant extent these questions remain only partially answered. There is general agreement, however, that the longevity and stability of ancient cratonic roots must be due, at least in part, to a combination of factors, primarily compositional buoyancy and strength and secondarily buffering of cratons by marginal Proterozoic mobile belts. The concept of a stabilizing chemical boundary layer is at the heart of the *temperature-depletion compensation* hypothesis laid out by Jordan in an important series of papers on the continental *tectosphere* (see Jordan (1981) for references) and discussed in some detail below. The term tectosphere was introduced by Jordan in an attempt to circumvent some of the difficulties associated with the conflicting definitions of lithosphere alluded to above. It is

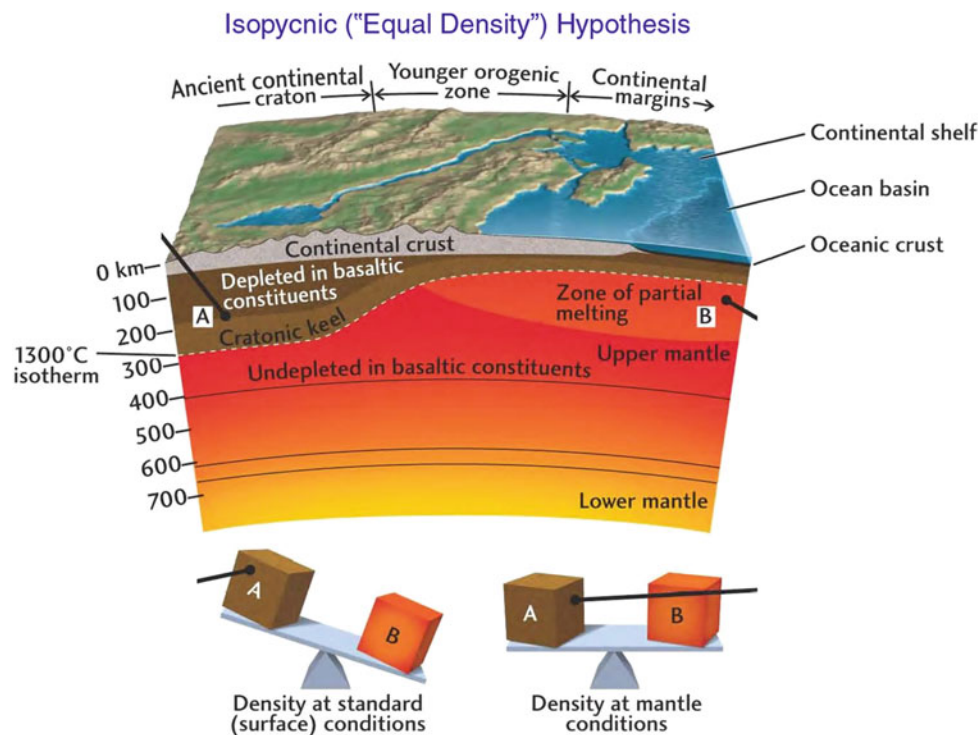
a particularly convenient concept for understanding those seismological and petrological/geochemical aspects of deep continental structure that appear to indicate continent-ocean heterogeneity to depths of at least 200 km, and perhaps considerably more.

There is a large body of knowledge about the composition of deep cratonic mantle from the study of xenoliths that have been carried to the surface in explosive volcanic pipes. The development of accurate mineralogical geothermometers and geobarometers, which allow the equilibrium depth and temperature of xenolith source regions to be determined, made it possible to “map” xenolith compositions as functions of position in the mantle. A comprehensive summary of results from the geochemical study of mantle xenoliths can be found in Carlson et al. (2005). Most of the ultramafic nodules associated with older cratonic areas are highly depleted in their basaltic melt fraction to a depth of at least 200 km and are approximately the same age as the overlying crust (Carlson et al. 2005; James et al. 2004). Depleted mantle xenoliths, which are by far the most common, are characterized by significantly lower density and higher seismic velocities than the more typical fertile suboceanic mantle (Boyd and McCallister 1976). This fact led Jordan (1981) to

postulate a thick *chemical boundary layer* for the continental tectosphere.

The effect of removing a basaltic partial melt from a fertile mantle peridotite (such as the hypothetical pyrolite of sub-oceanic mantle) is to reduce the density of the residue and to increase seismic velocities. Jordan compared densities of depleted peridotite nodules with densities of a hypothetical pyrolite composition and shows that average depleted peridotite of subcontinental mantle has a normative density about 1.3% less than that of pyrolite. This reduction in density will approximately compensate for the density contrast in a homogeneous mantle that would result from the temperature contrast of about -400° between a typical shield geotherm and the average oceanic adiabat. The model then, is based on the supposition that the lower density of the subcontinental mantle due to basalt depletion is gravitationally balanced by the density decrease in the oceanic mantle caused by a much higher geotherm, hence the expression *temperature-depletion compensation* (Fig. 2). The thickness of the chemical boundary layer is generally taken to be about 150–200 km.

The tectosphere model remains controversial. The mechanisms by which a chemical boundary layer can be built up to form a deep mantle keel beneath ancient continental crust are



Lithosphere, Continental, Fig. 2 Schematic diagram (adapted from Jordan, e.g., (Jordan 1988) with reference to previous work) showing the principal components of the isopycnic (“equal density”) hypothesis. The cold, strong lithosphere (A) extends to at least 250 km beneath the ancient cratons and migrates with the craton during plate motions. At comparable depths beneath the oceans (B), the mantle is hot, weak, and mobile. The depleted mantle peridotite that makes in the cratonic keel is

depleted in heavier constituents (e.g., Fe) and is therefore intrinsically less dense than more fertile oceanic mantle at the same temperature and pressure. At the same time, the sub-cratonic mantle is colder than the suboceanic mantle so the densities are approximately equal (the “temperature-depletion compensation” model). The chemical buoyancy of the cratonic root, along with its anhydrous strength, helps stabilize it against convective disruption over geologic time

poorly understood, although there are many modes by which magma can be extracted from continental mantle over time (Carlson et al. 2005). The most difficult aspect of deep mantle roots to model, however, has been their remarkable dynamical stability over billions of years of a chemical boundary layer whose thickness varies laterally by 150 km or more (Shapiro et al. 1999). Convective instability should act to disperse and thin the zone of depleted mantle, and the fact that this does not occur appears to be due largely to the anhydrous, highly viscous, nature of the depleted mantle keel.

One aspect of lithospheric formation that may be important in some cases involves underplating of continental lithosphere by oceanic lithosphere. There is evidence, for example, that among the eclogites brought up in kimberlite pipes in Africa some may be relicts of basaltic oceanic crust emplaced beneath the continental lithosphere in Precambrian time (MacGregor and Manton 1986). The data on which this conclusion is based come primarily from measurements of oxygen isotopic ratios combined with trace element and radiogenic isotopic ratios. The oxygen isotopic ratios (low δO^{18}), in particular, are difficult to explain as originating anywhere other than in oceanic crust. Similar evidence for underplating of cratonic mantle by oceanic lithosphere has also been found in sulfide inclusions in eclogitic diamonds from southern Africa, where episodic diamond formation has been linked to the accretion of oceanic lithosphere onto early cratonic nuclei, with subsequent stabilization of the tectospheric mantle (Shirey et al. 2001).

Lithosphere–Asthenosphere Boundary

A fundamental concept of plate tectonic theory is that the cold, rigid lithospheric plates of the Earth's outer rind are decoupled from the hot, ductile underlying mantle of the convecting Earth at the lithosphere–asthenosphere boundary (LAB). The detailed nature of this global boundary, despite many years of intense study, is still very poorly known. Yet understanding the seismic and rheological character of the decoupling zone between lithosphere and asthenosphere is a critical element in understanding the dynamical interaction between the lithospheric plates moving across the surface of the globe and the convecting asthenospheric mantle beneath. While seismic studies of the LAB beneath the ocean basins show that in many cases – although by no means all – the LAB can plausibly be related to an isotherm or rheological interface, the LAB beneath continents is far more problematic, particularly beneath the stable continental interior, where the LAB has been remarkably difficult even to detect, and its specific nature remains largely unknown.

Much of the information gleaned over the past 2 decades on the velocity structure and thickness of the continental lithosphere comes from seismic tomography, which images

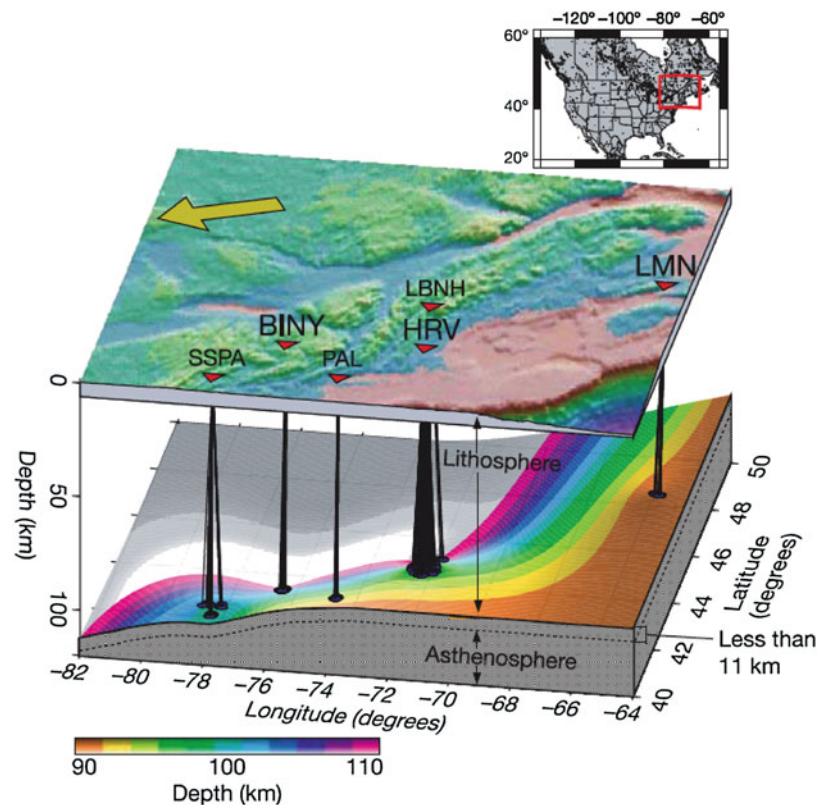
3-D variations in velocity distribution in the upper mantle. Where continental lithospheric mantle is present, it is seen as a higher velocity layer overlying a lower velocity asthenosphere. On the other hand, tomographic imaging, while an invaluable tool for 3-D mapping of velocity structure, provides little or no information as to the sharpness of the lithosphere–asthenosphere boundary or the velocity contrast across it. This is a significant limitation, as knowledge of the fine structure of the LAB is crucial to sorting out the competing effects of varying geothermal gradients and the presence or absence of hydrous/carbonate fluids or partial melts for decoupling the continental lithosphere from the mobile mantle beneath.

Direct approaches to determining velocity structure at the base of the lithosphere generally involve investigating the small seismic signals produced either by the *conversion* or the *reflection* of teleseismic body waves incident on the underside of the lithosphere–asthenosphere discontinuity (e.g., Rychert et al. 2005; Sacks and Snoke 1977). A converted phase is generated when some fraction of the energy of a compressional/shear wave that is obliquely incident on a velocity discontinuity in the Earth is converted to a shear/compressional wave. One widely used method, *receiver function analysis* (see ► “[Seismic, Receiver Function Technique](#)”), can effectively isolate the small signals on seismic records that represent conversions from the LAB. In principle, the waveform and amplitude of converted phases can provide detailed information about the structure (sharpness) and the velocity contrast – including the sign of the velocity contrast – across a discontinuity. Similarly, underside reflections from the LAB, which are seen on seismic records as small precursors to the large amplitude SS or PP surface reflections from distant earthquakes can be used to determine the sharpness and velocity contrast across the LAB. Unlike conversions, however, underside reflections sample substantial areas of the LAB (typically ~ 1,000 km in the case of SS phases) and, therefore, measure average regional, rather than local, structure.

Unlike most discontinuities in the Earth, the LAB represents a reversal in velocity with depth, the effect of which is to produce a converted or reflected signal with opposite polarity to that produced by a velocity increase. Results of receiver function studies of converted phases from the LAB have been decidedly mixed. An important study published in 2005 (Rychert et al. 2005) revealed a sharp LAB at a depth of about 100 km beneath northeastern USA (Fig. 3). The authors suggested on the basis of the magnitude of the negative shear-wave velocity contrast across the LAB (~5.10%) and the sharpness of the velocity transition (~ 10 km or less) that the LAB cannot be produced by thermal or compositional contrast alone, but must represent a transition from solid (melt-free) lithosphere to a porous asthenosphere containing at least a few percent of partial melt or fluid phase. Thus, the LAB

Lithosphere, Continental,

Fig. 3 A schematic 3-D rendering of LAB conversion points beneath six permanent seismic stations (*red triangles*) in northeastern USA (From Rychert et al. 2005. With permission). Inset map at top shows location of 3-D rendering. Topography is shown as shaded relief on the upper 3-D panel. The lower surface represents the base of the lithosphere, which ranges in depth from 90 km (*orange*) to 110 km (*pink*). Blue circles on the LAB discontinuity surface indicate the conversion points of the Ps phases. Black lines connect piercing points to the station at which the conversion is observed (Rychert et al. 2005)



could correspond to the solidus, or point of incipient melting, in the mantle beneath the continent, or it could mark the depth at which an impermeable barrier (the lithosphere) blocks the upward migration of fluids or partial melts from deeper in the mantle. In either case, the LAB in the Rychert et al. study represents a remarkably sharp rheological boundary, implying an abrupt transition from rigid lithosphere to ductile asthenosphere. Subsequent studies employing similar methodologies to those of Rychert et al., however, have proven difficult to interpret (e.g., Rychert and Shearer 2009), in that LAB depths so measured, particularly those beneath the stable continental interior, are incompatible either with tomographic images or with expected lithospheric thickness based on xenolith data or heat flow analysis (Pollack 1986; Romanowicz 2009). Thus, while the lithosphere–asthenosphere boundary beneath continents may be sharp and readily identified in some areas, notably tectonic regions or those of younger age, it would appear that beneath large parts of the stable continents the LAB may simply be too diffuse and/or heterogeneous to be identified as a single coherent discontinuity.

Summary and Conclusions

Lithosphere, as it is understood in plate tectonic phraseology, is a term that was developed almost entirely in the context of

oceanic plates. There, the discontinuity separating oceanic lithosphere from asthenosphere is widely recognized to be a rheological boundary between a strong melt-free plate and a weak, probably partially molten asthenosphere. In many important respects, we now know that this easily visualized oceanic model – universally displayed in textbooks both elementary and advanced – is a cartoonish oversimplification when applied to the continental lithosphere. Although there do appear to be regions where the continental lithosphere is almost as well defined as it is beneath ocean basins, the distinction between lithosphere and asthenosphere at depth beneath the continents, specifically the boundary between the two, is still shrouded in uncertainty. The failure of our ability seismologically to determine where lithosphere ends and asthenosphere begins has major implications for plate tectonic models in a dynamical earth. If we do not understand the coupling (or lack thereof) between lithosphere and asthenosphere, and the depth at which that occurs, then the task of sorting out the geodynamical drivers for plate tectonics becomes an exceedingly difficult exercise. While seismology ultimately has the ability to ferret out the truth of how continental plates are coupled or decoupled from the mobile mantle beneath, we are not quite at that point in the science. Nonetheless, the study of the continental lithosphere, particularly the lithospheric mantle, is entering a period of unprecedented expansion. In the USA, the EarthScope program's Transportable Array of 400 broadband seismic systems has

migrated to the country's midsection, with nearly 1,000 stations occupied at the time of this writing. Similar large-scale programs are well underway in Japan, China and Europe, giving hope that within the next several years the continental lithosphere and the lithosphere–asthenosphere boundary will have been mapped at high resolution over much of the globe's surface.

Cross-References

- Heat Flow, Continental
- Lithosphere, Continental: Thermal Structure
- Lithosphere, Mechanical Properties
- Lithosphere, Oceanic
- Lithosphere, Oceanic: Thermal Structure
- Mantle Viscosity
- Seismic, Receiver Function Technique
- Seismological Networks

Bibliography

- Boyd FR, McCallister RH (1976) Densities of fertile and sterile garnet peridotites. *Geophys Res Lett* 3(9):509–512
- Carlson RW, Pearson DG, James DE (2005) Physical, chemical and chronological characteristics of continental mantle. *Rev Geophys* 43, RG1001, 1–24
- Forsyth DW (1989) Lithosphere: mechanical properties. In: James D (ed) *The encyclopedia of solid earth geophysics*. Van Nostrand Reinhold, New York, pp 655–660
- James DE (2007) Crust and lithospheric structure – natural source portable array studies of the continental lithosphere. In: Dziewonski AM, Romanowicz B (eds) *Treatise on geophysics*. Elsevier, New York, pp 479–531
- James DE, Fouch MJ (2002) Formation and evolution of Archaean cratons: insights from Southern Africa. In: Ebinger C, Fowler CMR, Hawkesworth CJ (eds) *The early earth: physical, chemical and biological development*. Geological Society, London, pp 1–26
- James DE, Boyd FR, Schutt D, Bell DR, Carlson RW (2004) Xenolith constraints on seismic velocities in the upper mantle beneath southern Africa. *G-cubed* 5:1–32. [https://doi.org/10.1029/2003GC000551\(Q01002\)](https://doi.org/10.1029/2003GC000551(Q01002))
- Jordan TH (1978) Composition and structure of the continental tectosphere. *Nature* 274:544–548
- Jordan TH (1981) Continents as a chemical boundary layer. In: *The origin and evolution of the Earth's continental crust*. Trans Roy Soc London Ser A 301(1461):359–373
- MacGregor ID, Manton WI (1986) Roberts Victor eclogites: ancient oceanic crust. *J Geophys Res* 91:14063–14079
- Morgan P (1984) The thermal structure and thermal evolution of the continental lithosphere. *Phys Chem Earth* 16:107–193
- Pollack HN (1986) Cratonization and thermal evolution of the mantle. *Earth Planet Sci Lett* 80:175–182
- Ritsema J, van Heijst H (2000) New seismic model of the upper mantle beneath Africa. *Geology (Boulder)* 28(1):63–66
- Romanowicz B (2009) The thickness of tectonic plates. *Science* 324:474–476
- Rudnick RL, McDonough WF, O'Connell RJ (1998) Thermal structure, thickness and composition of continental lithosphere. *Chem Geol* 145:395–411
- Rychert CA, Shearer P (2009) A global view of the lithosphere–asthenosphere boundary. *Science* 324:495–498
- Rychert CA, Fischer KM, Rondenay S (2005) Scattered wave imaging of a sharp lithosphere–asthenosphere boundary beneath eastern North America. *Nature* 436:542–545
- Sacks IS, Snoke JA (1977) The use of converted phases to infer the depth of the lithosphere–asthenosphere boundary beneath South America. *J Geophys Res* 82:2011–2017
- Shapiro SS, Hager BH, Jordan TH (1999) Stability and dynamics of the continental tectosphere. In: van der Hilst RD, McDonough WF (eds) *Composition, deep structure and evolution of continents*. Elsevier, Amsterdam, pp 115–133
- Shirey SB, Carlson RW, Richardson SH, Menzies AH, Gurney JJ, Pearson DG, Harris JW, Wiechert U (2001) Emplacement of eclogite components into the lithospheric mantle during craton formation. *Geophys Res Lett* 28(13):2509–2512
- Sipkin SA, Jordan TH (1980) Multiple ScS travel times in the Western Pacific: implications for mantle heterogeneity. *J Geophys Res* 85:853–861
- West JD, Fouch MJ, Roth JB, Elkins-Tanton LT (2009) Vertical mantle flow associated with a lithospheric drip beneath the Great Basin. *Nat Geosci* 2:439–444
- Zandt G, Gilbert HJ, Owens TJ, Ducea M, Saleeby J, Jones CH (2004) Active foundering of a continental arc root beneath the southern Sierra Nevada in California. *Nature* 431:41–46

Lithosphere, Continental: Thermal Structure

Claude Jaupart¹ and Jean-Claude Mareschal²

¹Université de Paris Institut de Physique du Globe, Paris, France

²Centre GEOTOP-UQAM, University of Québec, Montréal, QC, Canada

Definition

Thermal conduction is the dominant form of heat transport across the lithosphere. The thermal structure of the continental lithosphere and its thickness are determined by the distribution of heat-producing elements in the crust and lithospheric mantle and the boundary condition at its base. Vertical seismic velocity profiles, geothermobarometry on mantle xenoliths, and sedimentary records of subsidence provide additional constraints on lithospheric thickness and temperature. The base of the lithosphere can be defined as the intersection of the conductive temperature profile with the mantle isentrope.

Introduction

The lithosphere is the superficial shell or *boundary layer* that lies at the top of the convecting mantle and experiences little internal deformation. The concept of lithosphere has evolved

over the years, leading to various definitions and estimates of its thickness. The base of the lithosphere was initially defined by a seismic low-velocity zone associated with partial melting. In the oceans, it was also determined by the variations of seafloor depth and surface heat flux as a function of age. It was clear from the beginning that all these definitions are related to the thermal structure. For instance, partial melting requires temperatures above the solidus of mantle rocks. These definitions were introduced, and tested, in the oceans with varying degrees of success but could not be readily extended to the continents. Over large continental areas, no seismic low-velocity zone can be detected, and, for a long time, the meager scattershot heat flux data set allowed conflicting interpretations.

In the oceans, the lithosphere is made out of homogeneous starting material that is well characterized. It is affected by few thermal perturbations and tectonic events during its short residence time at Earth's surface. Its thermal structure and thickness are essentially functions of only one variable, age, and evolve in simple ways that can be studied with several independent geophysical techniques. In comparison, the continental lithosphere is heterogeneous and thick. Not only was it formed a long time ago by processes that are still debated today, but it has been affected by a succession of perturbations and tectonic events which have left their imprints. Its thermal structure is sensitive to the large amounts of heat-producing elements in the crust and cannot be characterized by age only.

Because the continental lithosphere is thick, its thermal relaxation time is very large, which offers opportunities to study mechanisms that are no longer active today. Because of its old age, it preserves structural and chemical records of ancient geological processes. From a purely thermal standpoint, it plays an important role in regulating the Earth's heat loss and in storing large quantities of radioactive elements that are no longer available to power mantle convection.

The Thermal Boundary Layer of Mantle Convection

The lithosphere can be defined from a purely thermal standpoint. There are two basic mechanisms of heat transport within the Earth, conduction and convection. Convection refers to the transport of energy by matter that is set in motion by buoyancy forces. Without mass transport, energy can be transferred by conduction in association with a temperature gradient. These mechanisms are important in different parts of the Earth. At shallow depths beneath the surface, the cold upper layer does not deform easily on geological timescales, so that conduction is the dominant heat transport mechanism. The lithosphere may be defined from this perspective, but this definition does not account for heat that is supplied by convection from below. One must add a relatively thin convective

boundary layer at the base of the lithosphere, which connects the rigid and purely conductive upper region to the well-mixed mantle below. In this basal boundary layer, heat transport by conduction is not negligible, as discussed below in more detail.

Thermal Structure

In the continental lithosphere, heat is in part supplied from below by convection and in part generated by the decay of radioactive elements. In steady state and assuming that heat transport occurs in the vertical direction only, the heat balance equation is:

$$0 = -\frac{dq}{dz} + H \quad (1)$$

where z is the depth, q is the vertical heat flux, and H is the rate of heat production. In the upper part of the thermal boundary layer, there is no convection, and the heat flux is given by the Fourier law of heat conduction:

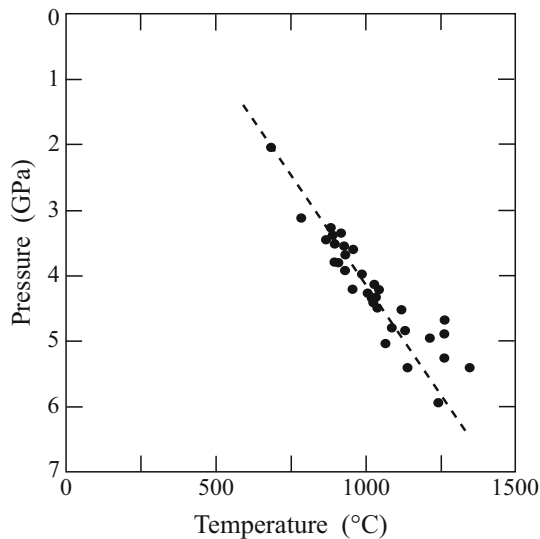
$$q = -\lambda \frac{dT}{dz} \quad (2)$$

where T is temperature and λ thermal conductivity. The temperature gradient can be measured at Earth's surface in deep boreholes and determined in the lithospheric root below the crust using the thermodynamic equilibrium conditions between mineral assemblages found in the sample from the mantle, mantle xenoliths found in kimberlite pipes. Typical values in stable continents are 15 K km^{-1} and 5 K km^{-1} , respectively (Fig. 1). The difference between these two values is due to radiogenic heat production in the continental crust. Below the lithosphere, the temperature gradient is dictated by the convective regime and is expected to be close to an isentrope. By definition, entropy is conserved when going up or down an isentrope, such that:

$$\frac{dT_S}{dz} = \frac{\alpha g T_S}{C_p} \quad (3)$$

where α is coefficient of thermal expansion, g is acceleration of gravity, T_S is temperature, and C_p is specific heat at constant pressure. For the Earth's mantle, the isentropic gradient is typically 0.5 K km^{-1} , which is much less than the conductive gradient in the lithospheric mantle. One identifies the isentrope by its *potential temperature*, i.e., the value of its temperature at atmospheric pressure.

Three different depths may be defined in the thermal boundary layer (Fig. 2). The shallowest boundary, h_1 , corresponds to the base of the rigid upper part and of what we shall call the thermal lithosphere. The deepest boundary, at depth h_3 , corresponds to the top of the well-mixed mantle and lies



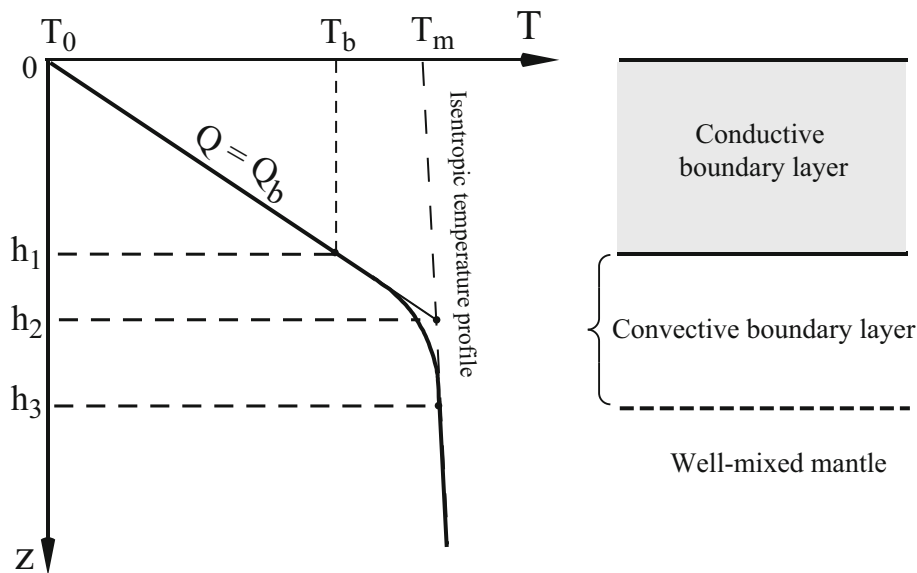
Lithosphere, Continental: Thermal Structure, Fig. 1 Pressure and temperature estimates from studies of mineral assemblages in xenoliths from the lithospheric mantle beneath the Kirkland Lake kimberlite pipe, Abitibi subprovince, Canadian Shield. Data are taken from Vicker (1997). The best fit (dashed line) indicates a temperature gradient of 5 K km^{-1} . Note that it intersects Earth's surface at a temperature of about 400 °C , showing that the temperature gradient is larger than 5 K km^{-1} in the crust

along the mantle isentrope. With no knowledge of boundary layer characteristics and heat transport mechanisms, one cannot determine h_1 . An intermediate depth, h_2 , is obtained by downward extrapolation of the conductive geotherm to the mantle isentrope. h_3 is obtained from seismic velocity anomalies, as it is such that temperatures do not deviate significantly from those beneath oceans or from the average mantle velocity profile. These thickness determinations say nothing about h_1 . Yet, it is h_1 which defines the mechanically coherent unit (the “plate”) which moves at Earth's surface and sets the thermal relaxation time which follows tectonic and magmatic perturbations. Uncertainty on this thickness has severe consequences because the diffusive relaxation time is $\propto h^2/\kappa$, where κ is thermal diffusivity.

In steady state, Q_0 , the heat flux at Earth's surface, is obtained by integrating the heat balance Eq. (1) and can be broken down in three components:

$$Q_0 = Q_{\text{crust}} + Q_{\text{lith}} + Q_b \quad (4)$$

where Q_{crust} and Q_{lith} stand for the contributions of heat sources in the crust and in the lithospheric mantle and Q_b is the heat flux at the base of the lithosphere. To describe heat transport mechanisms and the lithosphere structure, one must



Lithosphere, Continental: Thermal Structure, Fig. 2 Left: Schematic vertical temperature profile through the continental lithosphere, illustrating three different thicknesses and temperatures. Thickness h_1 corresponds to the rigid upper part of the thermal boundary layer. h_2 is determined by downward continuation of the temperature profile in the rigid upper part to the isentrope that approximates the temperature profile

in the well-mixed convective mantle. h_3 denotes the base of the thermal boundary layer, such that temperature is along the well-mixed isentrope. Right: three different regions with different heat transport characteristics. Heat is transported by conduction only above depth h_1 and by convection only below depth h_3 . Between these two depths, a convective boundary layer is such that both heat transport mechanisms are important

introduce three temperatures, T_0 at the upper boundary, which, for all practical purposes, may be taken as fixed and equal to 0 °C, T_b at the base of the lithosphere, and T_m along an isentrope in the well-mixed convective interior (Fig. 2). One can reasonably assume that the mantle potential temperature is the same beneath continents and oceans and hence that it can be deduced from the composition of mid-ocean ridge basalts (Kinzler and Grove 1992). Today, this temperature is about 1325 °C. T_b , the temperature at the base of the thermal lithosphere (the rigid lid), cannot be determined from geophysical measurements.

In the oceans, one may ignore Q_{crust} and Q_{lith} , and, for seafloor older than 100My, the surface heat flux is equal to the basal heat flux. In this case:

$$Q_0 = Q_b = \lambda \frac{T_b - T_0}{h_1} = \lambda \frac{T_m - T_0}{h_2} \quad (5)$$

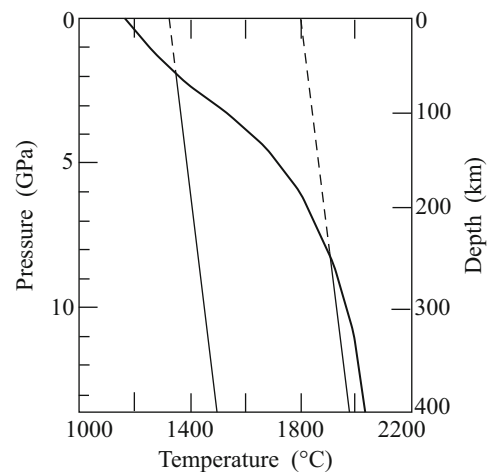
Well-established convection theory leads to a closure equation relating Q_b to the temperature difference across the convective boundary layer, $(T_m - T_b)$ (Davaille and Jaupart 1993; Solomatov 1995). In this case, therefore, measurement of the surface heat flux Q_0 allows determination of the basal heat flux and of the temperature difference $(T_m - T_b)$. In turn, this leads to the values of thicknesses h_1 and h_2 as well as of basal temperature T_b . In the oceans, therefore, the values of all the variables can be deduced from surface measurements. In the continents, the problem is more complicated because the surface heat flux includes the crustal contribution Q_{crust} which may vary by large amounts. This crustal component is typically about 30 mW m⁻² out of a total of 42 mW m⁻² in Archean (i.e., older than 2.5Gy) cratons. In such conditions, surface heat flux measurements do not provide a direct measure of the heat flux at the base of the lithosphere, which must be inferred from other data with methods that will be discussed below. Q_{crust} depends on the composition of the continental crust, which itself reflects the past history of magma emplacement and tectonic deformation/accretion. Observations show that it varies by large amounts even in a single geological province. It cannot be calculated from first physical principles and must be determined on a case-by-case basis using a host of different measurements.

The lithosphere owes its physical properties to partial melting and melt extraction processes which leave a solid residue that is both less dense and mechanically stronger than the starting mantle material. Mechanical strength is due to both the lower temperature in the lithosphere than in the convective mantle and dehydration that occurred during partial melting (Pollack 1986). Melting may be achieved by two different mechanisms operating in different settings. One mechanism is decompression in an upwelling and has been

documented in considerable detail in mid-ocean ridges (McKenzie and Bickle 1988). Decompression melting begins at a depth which depends on mantle temperature and water content (e.g., 60–80 km for the oceans today). Continents are not linear features and were probably not generated by linear mantle upwellings such as present-day mid-ocean ridges. Their formation has been linked to large mantle plumes or subduction zones (Carlson et al. 2005). For mantle plumes, the mechanism is basically the same as that beneath mid-ocean ridges, and the depth where melting starts dictates the thickness of the buoyant solid residue. In a dry mantle, high potential temperatures are required for the generation of thick lithospheric roots (Herzberg 1983; Fig. 3). For a thickness of 250 km, for example, the mantle potential temperature must be about 1800 °C. On the other hand, if the Archean mantle was wet, melting occurred at shallower depths and at lower temperatures (Grove and Parman 2004). In this case, a second event is required to account for the large thickness of continental lithosphere, involving large-scale compression and/or thrusting. For both types of models, the lithosphere thickness provides a strong constraint on the mechanism of continent formation.

Heat Flux at the Base of the Continental Crust and at the Base of the Lithosphere

Constraints on the thermal structure of the continental lithosphere may be obtained with several geophysical methods. To obtain continuous temperature profiles, however, one can only use surface heat flux data and downward continue



Lithosphere, Continental: Thermal Structure, Fig. 3 Solidus temperature of dry mantle peridotite as a function of pressure and depth, from Herzberg (1983). Also shown are two isentropes that intersect the solidus at depths of 60 and 250 km

shallow temperature measurements. To this aim, one needs to specify the heat flux at the base of the lithosphere. A first step is to determine the amount of heat produced in the continental crust or, alternatively, the heat flux at the Moho discontinuity.

Uranium and thorium are the main heat-producing elements in rocks. They are located mostly in accessory minerals and grain boundaries, which depend weakly on the bulk chemical composition. Thus, their concentrations are not related to physical properties such as density and seismic velocity and cannot be retrieved from large-scale geophysical studies. In addition, they vary by large amounts at all scales, from that of a petrological thin section to that of a whole massif. Within an apparently homogeneous pluton, they can change in both vertical and horizontal directions due to fluid migration and late-stage alteration. In the Bohus granite, Sweden, for example, thorium concentrations vary by a factor of 5 over horizontal distances as small as a few tens of meters (Landstrom et al. 1980).

In a geological province, radio-element concentrations cannot be estimated using data from other provinces and must be measured in all the major rock types present. They depend on lithology and are high in granites and meta-sediments, so that a geological map gives a rough idea of the spatial distribution of heat production. With the very wide ranges that exist for each rock type, however, the standard deviation of the heat production distribution is large and often larger than the mean. As a consequence, a single thermal model for a province would gloss over the important lateral temperature variations that occur due to the heterogeneous crustal structure. For a reliable thermal calculation in a specific area, one cannot use the local heat production values of a few rock types found on the outcrop because temperatures are sensitive to heat that is generated over rather large volumes. Average heat production values must be determined on a scale that is intermediate between the dimensions of individual plutons and the size of the province.

Vertical variations of heat production can be determined in deep boreholes or from exposed crustal sections. On a vertical scale of ~ 10 km, data from the deep boreholes at Kola, in the Russian part of the Baltic Shield (Kremenensky et al. 1989), and the KTB, Germany (Clauser et al. 1987), show no systematic change with depth. At KTB, heat production between 8 and 9 km depth is the same as between 1 and 2 km and higher than above 1 km. At Kola, heat generation in the Archean rocks between 8 and 12 km is higher ($1.47 \mu\text{W m}^{-3}$) than in the shallower Proterozoic section ($0.4 \mu\text{W m}^{-3}$).

Over a crustal thickness scale, heat production is lower in mid-crustal assemblages than in the upper crust. This vertical variation is not monotonous and cannot be described by a simple function valid everywhere, as shown by exposed crustal sections such as the Vredefort in South Africa (Nicolaysen et al. 1981), the Cordilleran core complexes of

Lithosphere, Continental: Thermal Structure, Table 1 Element concentration and heat production in the continental crust. (From Rudnick and Gao (2014)). Element concentrations are given in ppm

	U	Th	K	$\mu\text{W m}^{-3}$
Upper crust	2.7	10.5	23,300	1.65
Middle crust	1.3	6.5	19,200	1.00
Lower crust	0.2	1.2	5,080	0.19
Total crust	1.3	5.6	15,080	0.89

Lithosphere, Continental: Thermal Structure, Table 2 Estimates of bulk continental crust heat production from heat flux data. (From Jaupart and Mareschal (2014))

Age group	A^a	Q_{crust}^b
Archean	0.56–0.73	23–30
Proterozoic	0.73–0.90	30–37
Phanerozoic	0.95–1.21	37–47
Total continents	0.79–0.99	32–40

^aRange of heat production in $\mu\text{W m}^{-3}$

^bRange of the crustal heat flow component in mW m^{-2}

Arizona (Ketcham 1996), or the Pikwitonei-Sachigo and Kapuskasing-Wawa areas of Canada (Fountain et al. 1987; Ashwal et al. 1987). One robust feature is the depletion of lower crustal assemblages in comparison to upper and middle crustal ones (Table 1).

Table 1 lists estimates derived from global geochemical models of the crust and a very large number of measurements. The crust is split into three different crustal horizons corresponding to average regional seismic models. The model crust is 40 km thick with middle crust in the 12–23 km depth range and is such that $Q_{\text{crust}} = 36 \text{ mW m}^{-2}$. Another method relies on heat flux data and will be explained below (Table 2). With this method, one may also evaluate heat production variations between provinces. The two independent methods lead to comparable results and show that crustal heat generation typically accounts for more than half of the surface heat flux.

Heat Flux and Heat Production

With the large variations in the concentrations of heat-producing elements that have been documented in crustal rocks, it is not surprising that, in continents, the surface heat flux changes by large amounts over small horizontal distances of a few tens of kilometers. The largest anomalies occur in association with enriched granitic plutons. These plutons are usually of small size, and the local value of the heat flux is not representative of the average crust that lies below the measurement site. In such conditions, studies of the thermal structure of continents require large data sets. The challenge is to obtain average values of heat flux and heat production that are representative of the whole crust in

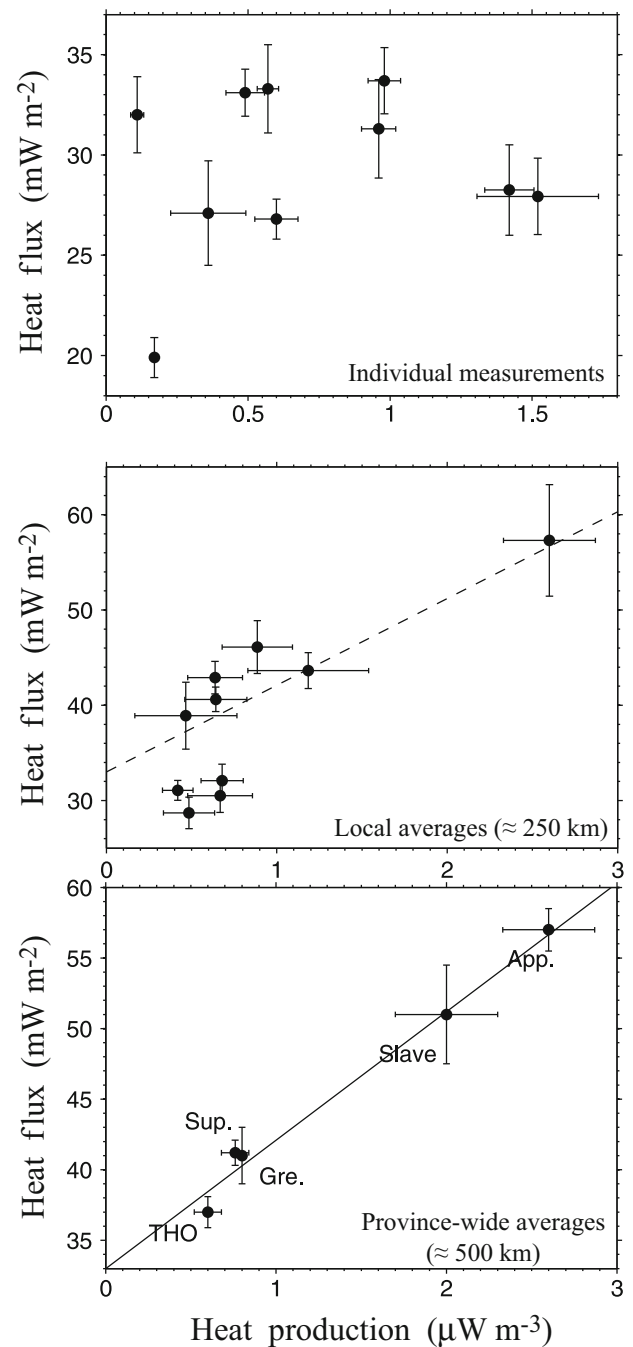
an area. We discuss the scales over which this can be done in a reliable manner.

There is only a weak relationship between local values of heat flux and heat production in a geological province (Fig. 4). For a representative thermal model, one must determine an average heat flux value over an area that is sufficiently large for the smoothing of small-scale variations due to isolated anomalous massifs. Figure 4 also shows data for ten $\approx 250 \text{ km} \times 250 \text{ km}$ windows distributed in North America. These are located in the same area as the individual heat flux values discussed previously and show that a relationship between heat flux and heat production begins to emerge at this scale. Even at this scale, the relationship is not defined tightly due to insufficient sampling of heat production. A single heat flux value is sensitive to crustal heat production over large horizontal distances and records the contribution of many different geological units. In contrast, the heat production values correspond to a small number of rocks found at shallow depth in boreholes used for heat flux determinations. The limited sampling achieved at a scale of 250 km does not account for the full range of rock types and heat production rates in the crust. For representative heat production averages, one needs to work at a larger scale. Figure 4 shows data for five major geological provinces in North America with typical dimensions of $\approx 500 \text{ km} \times 500 \text{ km}$. These provinces cover the geological history of the stable continent and represent different types of continental crust. Furthermore, they have been extensively deformed and eroded, so that rocks from a large range of depths can be found at the surface. The metasedimentary-plutonic belts of the western Superior Province, Canadian Shield, for example, include rocks from all metamorphic grades up to granulite facies. The province-wide averages exhibit a remarkable linear relationship, with a heat flux intercept of $\sim 33 \text{ mW m}^{-2}$ corresponding to crust with zero surface heat production. The data for the ten $\approx 250 \text{ km} \times 250 \text{ km}$ windows lie close to this global linear relationship (Fig. 4).

The Moho heat flux is equal to:

$$Q_M = Q_0 - Q_{\text{crust}} = Q_{\text{lith}} + Q_b \quad (6)$$

which depends on heat production in the lithospheric mantle and on the heat flux at the base of the lithosphere. Both components come from depths larger than the crustal thickness, so that lateral variations, if they exist, get smoothed out by horizontal diffusion. For an average lithospheric thickness of 250 km, variations of the basal heat flux Q_b over wavelengths smaller than 500 km are not detectable in the surface heat flux (Mareschal and Jaupart 2004). A similar statement can be made for the contribution of heat sources in the lithospheric mantle, Q_{lith} . Thus, on scales that are smaller than 500 km, one must consider that the Moho heat flux is



Lithosphere, Continental: Thermal Structure, Fig. 4 Relationship between heat flux and radiogenic heat production in North America at different scales, from Levy et al. (2010). Top panel: for individual measurements, there is clearly no relationship between the two. Middle panel: at an intermediate scale of about 200 km, a relationship between the average values of heat flux and heat production begins to emerge. The dashed line corresponds to the best-fit linear relationship defined by province-wide averages. Bottom panel: it shows a strong correlation between the average values of heat flux and heat production for the five major geological provinces of North America. The best-fit linear relationship has a slope of 9 km and an intercept of 33 mW m^{-2} . Symbols are as follows: App, Appalachian Province (400My); Gre, Grenville Province (1100My); THO, Trans-Hudson Orogen (1800My); Sup, Superior Province (>2500My); Slave, Slave Province (>2500My)

uniform and that variations of the surface heat flux come solely from changes of crustal heat production.

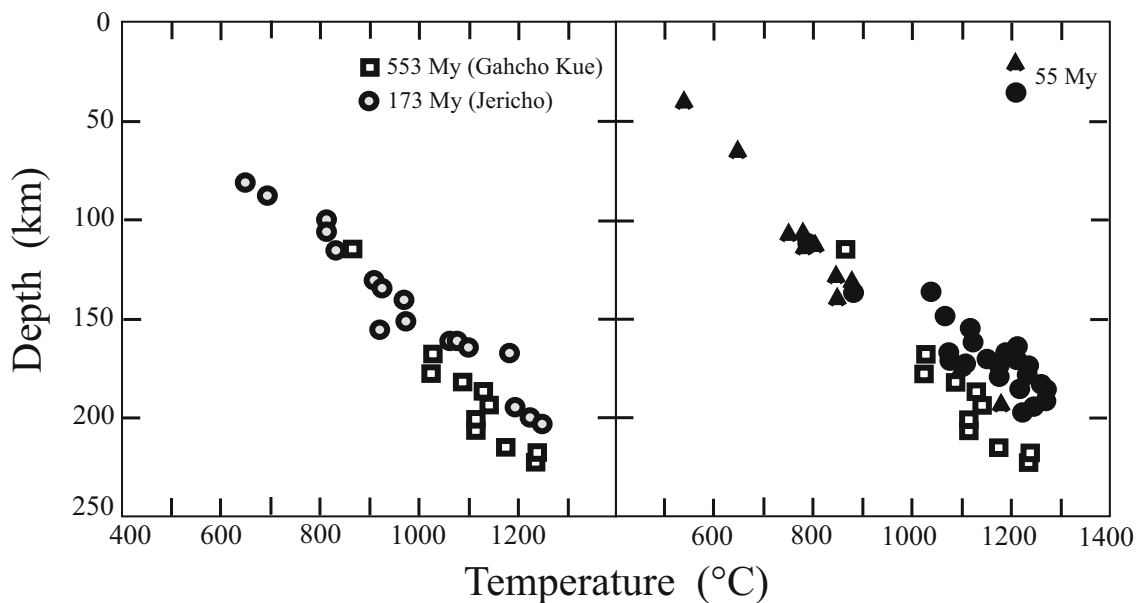
The strong correlation that is observed between the province-wide average surface heat flux and heat production indicates that variations of the Moho heat flux are small throughout the whole North American continent (Fig. 4). If there were large variations of the Moho heat flux, they would need to be compensated by opposite variations of the average lower crustal heat production. No physical mechanism can explain how such independent variables can be linked to one another. Variations of the Moho heat flux are therefore less than the intrinsic uncertainty on the heat flux measurements and the magnitude of departures from the heat flux-heat production relationship, which is about $2\text{--}3 \text{ mW m}^{-2}$ (Mareschal and Jaupart 2004). The absolute value of the Moho heat flux is not specified by this analysis, however, and this issue is discussed in the next section.

The Moho Heat Flux

Direct determination of the heat flux through the lithospheric mantle can be made using xenolith samples brought from lithospheric depths by powerful *kimberlitic* eruptions. Coexisting mineral assemblages and mineral compositions allow estimates of pressure and temperature and of the geothermal gradient. Together with a thermal conductivity estimate, they lead to values of the heat flux beneath the crust. For the Kirkland Lake pipe, within the Superior Province, Canada (Fig. 1), Rudnick and Nyblade (1999) obtained a best-fit

Moho heat flux estimate of $\approx 18 \text{ mW m}^{-2}$ within a total range of $17\text{--}25 \text{ mW m}^{-2}$. Russell and Kopylova (1999) derived a best fitting value of 15 mW m^{-2} , within a range between 12 and 24 mW m^{-2} , for the Jericho kimberlite area of the Archean Slave Province, also in Canada. Figure 5 shows data from several kimberlite pipes in the Slave Craton (2008). There are significant temperature differences between the different data sets. Surprisingly, the youngest eruptions, which provide recent samples of the lithospheric mantle, yield the highest temperatures. A similar result is obtained in the Kaapvaal Craton, South Africa (Michaut et al. 2007). If these temperature variations are indeed due to age and reflect the time evolution of lithospheric temperatures, they are not consistent with secular cooling of the Earth's mantle. The pipes do not overlap so that the data may in fact record lateral variations of temperature in the lithospheric mantle, which can only be explained by changes of heat production. Nevertheless, these data lead to similar values of the temperature gradient and heat flux.

Bounds on the Moho heat flux can be obtained by taking advantage of the variations of heat flux and crustal structure in a geological province. Combined with heat production data for the various rock types involved, one can isolate the variable crustal component and the uniform Moho heat flux. In Canada, Moho heat flux values of $10\text{--}15 \text{ mW m}^{-2}$ have thus been derived for the Grenville Province, the Trans-Hudson Orogen (THO), for the Abitibi belt of the Superior Province (Jaupart and Mareschal 2015). Crustal thickness and



Lithosphere, Continental: Thermal Structure, Fig. 5 Relationship between temperature and depth in the lithospheric mantle of the Slave Province, Canada, from Canil (2008). The Slave Province is small and has been pierced repeatedly by kimberlite pipes in the same area. Left panel: data from the two oldest pipes. Right panel: data for two young

pipes compared to those for the oldest pipe in the area. On average, temperature estimates from the two young pipes are larger than those of the older pipes by about 150 K . Such differences may indicate lateral temperature variations in the lithospheric mantle

composition can also be constrained by combining seismic refraction, gravity, and heat flux. Such methods lead to values of Q_M between 7 and 15 mW m⁻².

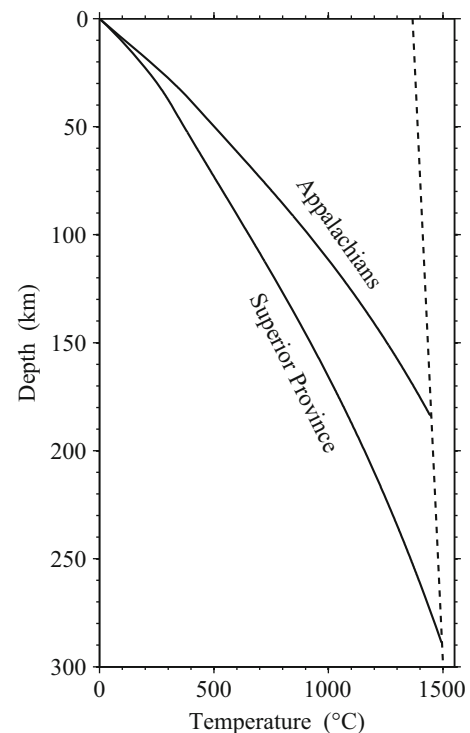
Another method relies on P_n velocities from seismic refraction surveys (Perry et al. 2006). For given values of the surface heat flux, the Moho temperature depends on the amount of crustal heat production and hence on the Moho heat flux. One can therefore obtain a relationship between Moho temperature and seismic velocity with the Moho heat flux as control parameter. One can also derive independently a relationship between temperature and seismic velocity using laboratory measurements of elastic properties and seismic velocities, with the bulk mantle composition as control parameter. The two relationships are only consistent with one another over a restricted range of composition and Moho heat flux. For depleted mantle compositions appropriate to the Superior Province, a good fit between predicted and observed P_n velocity values is obtained if the Moho heat flux is within a range of 12–25 mW m⁻² (Perry et al. 2006).

The three different and independent methods that we have described have all been applied to the Abitibi belt in the Canadian Shield and can be combined to tighten the final range to 12–15 mW m⁻² for that subprovince. Lower and upper bounds on the Moho heat flux have been derived using other arguments. Upper bounds on the mantle heat flux are obviously provided by the lowest heat flux measured. Values of 20–23 mW m⁻² have been reported within the Canadian Shield (Mareschal et al. 2000). Regional values as low as 18 mW m⁻² have also been reported for the Baltic or Siberian Shield. One can refine this estimate to 18 mW m⁻² using lower bounds on crustal heat production.

The methods that we have described here rely on different data and hence are associated with different sources of uncertainty. That such completely independent methods converge to similar results allows some confidence in the final range obtained. Values lower than 12 mW m⁻² are not consistent with the xenolith data, whereas values higher than 18 mW m⁻² can be excluded because of the heat flux data. This upper bound appears to be valid for all the shields (Jaupart and Mareschal 2015).

Lithospheric Geotherms

Using information on heat flux values at the surface and at the Moho, one may determine geotherms through the lithosphere. The calculation requires values for thermal conductivity that are discussed in the Appendix. Fig. 6 shows two such geotherms for two provinces in North America, which are consistent with both seismic and heat flux data (Levy et al. 2010). These two geotherms intersect the well-mixed mantle isentrope at different depths, indicating large lateral thickness variations for the continental lithosphere. The maximum thickness is reached beneath the Archean Superior Province and is estimated to be 280 ± 30 km.

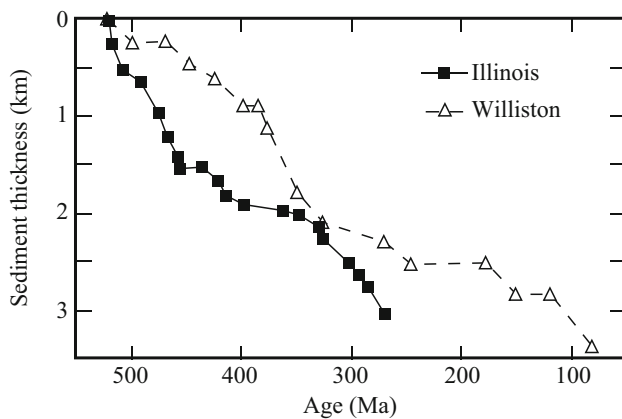


Lithosphere, Continental: Thermal Structure, Fig. 6 Two steady-state lithospheric geotherms beneath the Appalachian and Superior provinces of North America, from Levy et al. (2010). These two geotherms were deduced from measurements of the surface heat flux and heat production and were constrained by requiring consistency with seismic data. The dashed line is the isentrope for a potential temperature of 1350 °C

Thermal Transients

Because the continental lithosphere is thick, it is characterized by a long thermal relaxation time. For a thickness of 250 km and a thermal diffusivity κ equal to 10^{-6} m² s⁻¹, the characteristic diffusion time $\tau = h^2/\kappa$ is ≈ 2 Gy. This is much longer than that of the oceanic lithosphere, which is only about 200 My (for a thickness of 80 km). We shall now discuss two types of thermal transients in thick lithosphere.

Thermal relaxation of the lithosphere following a deep thermal perturbation can lead to subsidence, which has been recorded in intracratonic sedimentary basins located away from active plate boundaries. Subsidence histories for two such basins of North America, the Michigan and Williston basins, are shown in Fig. 7 and emphasize their long durations exceeding 200 My. On the basis of models for heat transport and cooling in the vertical direction only, the thickness of the North American lithosphere is found to about 115 km beneath the Michigan basin and 270 km beneath the Williston (Haxby et al. 1976; Ahern and Mrkvicka 1984). There is no obvious explanation for such a large thickness difference in the middle of the North American continent, which suggests that the subsidence models may be oversimplified. These models



Lithosphere, Continental: Thermal Structure, Fig. 7 Subsidence histories for the Michigan and Williston basins, North America

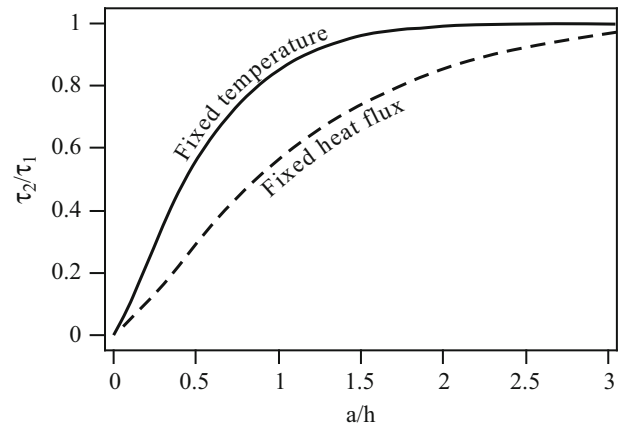
rely on 1-D calculations and hence are only valid if the thermal anomaly stretches over a horizontal distance that is much larger than the lithosphere thickness. The Michigan and Williston basins have been attributed in large part to plate flexure driven by a deep load (Nunn and Sleep 1984; Ahern and Mrkvicka 1984). This load is related to the initial thermal anomaly and has a radius of about 120 km beneath the Michigan basin (Haxby et al. 1976), which is smaller than the thickness of the lithosphere beneath the North American craton. In these conditions, the assumption of purely vertical heat transfer is not tenable.

Kaminski and Jaupart (2000) have reevaluated thermal models for the continents by taking into account lateral heat transport within thick lithosphere. They considered a lithosphere of thickness h with a cylindrical thermal anomaly of radius a between depths h and b . At the surface ($z = 0$), temperature is fixed at T_o . At steady state, the base of the lithosphere is at $T = T_o + \Delta T$. Assuming, for example, that the thermally perturbed region is initially at a uniform temperature $T_o + \Delta T$, temperature can be written as the sum of the equilibrium temperature and a dimensionless perturbation θ :

$$T(r, z, t) = T_o + \Delta T \left[\frac{z}{h} + \theta(r, z, t) \right], \quad (7)$$

where r is the radial distance. The temperature perturbation may be expressed in terms of a vertical component θ_z , which corresponds to the 1-D solution, and a radial component which can be calculated using Bessel functions. For thermal anomalies with planforms that are not circular, lateral heat transfer is dominated by the smallest horizontal dimension.

The acceleration of cooling due to lateral heat loss can be measured simply by the time to achieve 90% of the final subsidence, noted τ_2 , scaled to that for an infinitely wide initial perturbation, τ_1 (corresponding to the 1-D calculation) (Fig. 8). Results are sensitive to the boundary condition at the base of the lithosphere and show that the impact of lateral heat



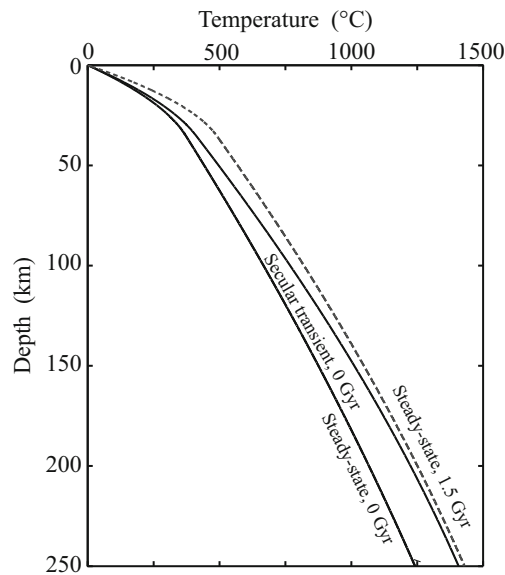
Lithosphere, Continental: Thermal Structure, Fig. 8 Time to achieve 90% subsidence in thick lithosphere as a function of the radius of the deep lithospheric thermal anomaly, a (scaled to the lithosphere thickness h). The subsidence time has been scaled to the value obtained with a 1-D thermal model for purely vertical heat transport. Results have been obtained for two different boundary conditions at the base of the lithosphere. (Adapted from Kaminski and Jaupart (2000))

transport is significant for $a/h < 1$, corresponding to $a < \approx 250$ km in the North American craton. This is the case of all the major intracratonic basins. In comparison to the oceans, thermal relaxation in the continents is lengthened by the larger lithosphere thickness and shortened by lateral heat transport. Different subsidence histories and durations are generated by the same basic process operating over different lateral distances.

Secular Transients

We have shown above that continental temperatures depend strongly on radiogenic heat production. Thus, continents cool down continuously due to the rundown of radioactivity with time. One should also add that conditions at the base of the lithosphere also evolve with time due to secular cooling of the whole Earth. This is ill-constrained and we focus our attention to the effects of decaying heat sources. For the Appalachian geotherm of Fig. 6, for example, crustal heat production contributes about 300 °C. For average Th/U and K/U ratios, radiogenic heat production decreases by a factor of 2 in about 2.7 Gy. We thus deduce that continental temperatures decrease at a rate of about 50 K Gy⁻¹, which is comparable to the bulk mantle cooling rate (Abbott et al. 1994). This simple argument relies on a quasi-steady-state approximation, such that ancient geotherms are calculated using the steady-state heat equation together with past values of heat production. For thick lithosphere, this approximation is not valid and may lead to significant errors.

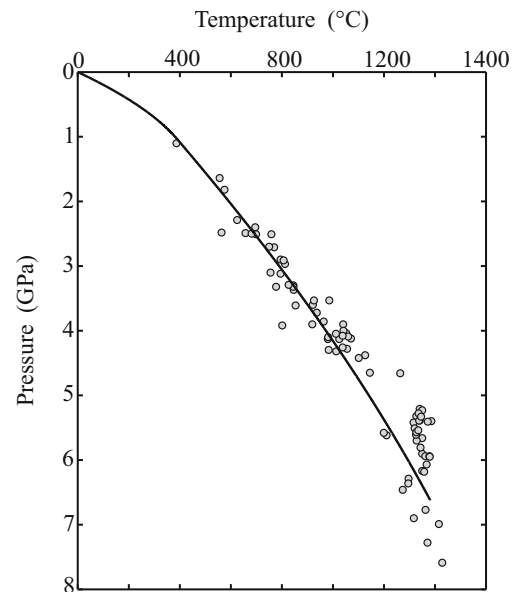
In thick continental lithosphere, the timescale for diffusive heat transport is comparable to the half-lives of uranium, thorium, and potassium. As noted above, the characteristic time for heat diffusion equal to h^2/κ is 2 Gy for 250-km-thick



Lithosphere, Continental: Thermal Structure, Fig. 9 Present-day vertical temperature profile for transient thermal model with decaying heat sources (thick line) compared with two steady-state geotherms (solid line for present-day heat production; dashed line, for heat production rate values 1.5 Gy ago) for the same set of parameters ($h = 250$ km, crustal heat production $A_c = 1.0 \mu\text{W m}^{-3}$, lithospheric mantle heat production $A_m = 0.02 \mu\text{W m}^{-3}$, basal heat flux $Q_b = 10 \text{ mW m}^{-2}$). Surface heat flow is of 49 mW m^{-2} for the transient solution, 46 mW m^{-2} for the present-day steady-state profile, 62 mW m^{-2} for the 1.5 Ga steady-state profile

lithosphere, which is close to the decay time of radiogenic heat production (≈ 2.7 Gy). Thus, heat diffusion is not efficient enough to keep up with the decaying heat sources, and lithospheric temperatures are not in equilibrium with the instantaneous rate of heat production. In these conditions, the vertical temperature profile exhibits significant curvature and may be hotter than a steady-state profile by as much as 150 K depending on the absolute value of the heat generation rate (Fig. 9). Another consequence is that forcing a steady-state temperature profile through xenolith (P-T) data leads to an overestimate of the mantle heat flux.

For illustration purposes, Fig. 9 shows the instantaneous vertical temperature profile with decaying sources and a steady-state geotherm derived for the same present-day values of crust and mantle heat production. The difference between the two geotherms increases with depth because of increasingly inefficient transport of heat to the Earth's surface. For comparison, we have also calculated a steady-state geotherm corresponding to heat production values 1.5 Gy ago. The transient geotherm falls between the steady-state geotherms calculated for heat production rates at 0 and 1.5 Ga, showing that it is sensitive to heat generation in the past. At the base of the lithosphere, all geotherms are parallel because of the constant heat flux bottom boundary condition. Secular transient behavior is not important at shallow depths,



Lithosphere, Continental: Thermal Structure, Fig. 10 Xenolith P-T data from the Kimberley, Jagersfontein, Premier and Frank Smith kimberlite pipes, Kaapvaal Craton, South Africa (Rudnick, personal communication, 2003). Thick line: vertical temperature profile for a transient thermal model with decaying heat sources and for a lithosphere thickness of 225 km. Parameter values are crustal heat production $A_c = 0.8 \mu\text{W m}^{-3}$, mantle heat production $A_m = 0.025 \mu\text{W m}^{-3}$, and basal heat flux $Q_b = 12 \text{ mW m}^{-2}$. Note that the geotherm terminates above several deep data points, due to the starting assumption on lithosphere thickness

due to the efficiency of diffusive heat transport over small distances. In the crust, differences between the transient and present-day steady-state profiles are very small and can be neglected for all practical purposes. Between 50 and 150 km depth, heat flux is about 13 mW m^{-2} for the present-day steady-state model and 16 mW m^{-2} for the transient one. Maximum departure from thermal equilibrium is achieved at the base of the lithosphere. The magnitude of transient effects essentially depends on mantle heat production as well as on lithosphere thickness. Large values of heat production in the lithosphere do not introduce large transients at shallow depths. In contrast, even small values of heat production lead to significant effects in a thick lithosphere.

The secular transient geotherm is compared to xenolith (P-T) data from Kaapvaal Craton, South Africa, for $h = 225$ km in Fig. 10. For this solution, the surface heat flux is 47 mW m^{-2} , which is close to measurements in this region, and heat production in the lithospheric mantle is $0.025 \mu\text{W m}^{-3}$, which lies in the lower range of values proposed by Rudnick and Nyblade (1999). The basal heat flux is 12 mW m^{-2} . The Kaapvaal xenolith (P-T) data suggest a change of slope in the geotherm at a pressure of about 5 GPa (Bell et al. 2003). According to some authors, this marks the transition between lithosphere and asthenosphere, with the deepest data points belonging to the convecting mantle below

the lithosphere. Rudnick and Nyblade (1999) and Bell et al. (2003) have argued that all data points belong to unperturbed lithosphere. Significant curvature of the secular transient geotherm allows a good fit to all the data and does not suggest a break in the geotherm.

Due to the intrinsic decay of radioactive heat sources, the lithospheric mantle therefore undergoes secular cooling even when thermal conditions at the base of the lithosphere remain steady. Predicted cooling rates are in the range of 50–150 K Gy⁻¹. Such transient behavior also has implications for heat flow studies. Assuming that the decay of radiogenic heat production can be represented by a single exponential, such that $A_m(t) = A_o h \exp(-\alpha t)$, the contribution of lithospheric heat production to the surface heat flux is:

$$\Delta Q_{\text{lith}}(t) = A_o h \frac{\tan \sqrt{\alpha \tau}}{\sqrt{\alpha \tau}} \exp(-\alpha t) \quad (8)$$

This is compared to the instantaneous amount of heat produced in the lithosphere:

$$\Delta Q_{\text{lith},i}(t) = A_o h \exp(-\alpha t) \quad (9)$$

Ratio $Q_{\text{lith}}/Q_{\text{lith},i}$ increases as a function of lithosphere thickness and is 1.5 for $h = 300$ km, for example. This shows that surface heat flux measurements record some time-average of the deep lithospheric heat production.

Summary

The continental lithosphere is much thicker than its oceanic counterpart, and its thermal structure depends strongly on radiogenic heat production in both the crust and in the mantle. Because of these two characteristics, its behavior depends on its geological history and is not a simple function of age. Beneath Archean shields, it can be as thick as 300 km.

Cross-References

- Deep Seismic Reflection and Refraction Profiling
- Energy Budget of the Earth
- Heat Flow Determinations, Continental
- Heat Flow, Continental
- Heat Flow, Seafloor: Methods and Observations
- Lithosphere, Mechanical Properties
- Lithosphere, Oceanic: Thermal Structure
- Mantle Convection
- Radiogenic Heat Production of Rocks
- Sedimentary Basins
- Seismic Velocity and Temperature Relationships
- Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

Appendix: Thermal Conductivity

For an isotropic medium, Fourier's law is:

$$\mathbf{q} = -\lambda \nabla T \quad (10)$$

where λ is the thermal conductivity measured in W m⁻¹ K⁻¹. More generally, if the medium is not isotropic, the thermal conductivity must be defined as a second-order tensor. Thermal conductivity can be broken into lattice and radiative components, which have different behaviors as temperature and pressure change.

At low temperatures, heat transport is in large part affected by lattice vibrations. Lattice conductivity decreases with increasing temperature.

At larger temperatures, electromagnetic waves transport energy in continuous media. Radiation is rapidly attenuated but the medium re-emits energy if the temperature is sufficiently high. In a temperature gradient, the net energy flux will be $\propto T^3 \nabla T$, which has the same form as the flux for lattice conduction. One can therefore define a radiative component of conductivity, λ_R , such that the radiative heat flux is:

$$\mathbf{q}_R = -\lambda_R \nabla T \quad (11)$$

The gray body law states that:

$$\lambda_R = \frac{16}{3} \frac{n^2}{\epsilon} \sigma T^3 \quad (12)$$

where n is the refractive index, ϵ is the opacity, and σ is the Stefan constant of blackbody radiation ($\sigma = 5.67 \times 10^{-8}$ W m⁻² K⁻⁴).

Lattice Conductivity

For individual crystals, theoretical arguments suggest that, below the Debye temperature, the temperature dependence of lattice conductivity takes the following form:

$$\lambda(T) = \lambda_{298} \sqrt{\frac{298}{T}} \quad (13)$$

For polycrystalline assemblages, this equation is no longer valid and one relies on empirical fits to the data. Reasonable agreement can be obtained with a relationship of the type:

$$\lambda(T) = A + \frac{B}{350 + T} \quad (14)$$

where T is in °C and where constants A and B depend on the rock type (Table 3).

The effect of pressure on conductivity can be treated independently of temperature, so that one can write:

Lithosphere, Continental: Thermal Structure, Table 3 Constants for calculating the thermal conductivity of different rock types. (From Clauser and Huenges (1995))

Rock type	T range (°C)	A	B
Metamorphic rocks	0–1200	0.75	705
Felsic rocks	0–1400	0.64	807
Mafic rocks	50–1100	1.18	474
Ultramafic rocks	20–1400	0.73	1293

Lithosphere, Continental: Thermal Structure, Table 4 Pressure dependence of lattice conductivity

Material	$\beta^a \times 10^{-2} \text{ GPa}^{-1}$	Reference
Mafic granulites (Finland)	2.4 ± 0.8	Kukkonen et al. (1999)
Olivine	3.2	Xu et al. (2004)

^aCoefficient in Eq. 15 for thermal conductivity

$$\lambda(T, P) = \lambda_o(T)(1 + \beta P) \quad (15)$$

where β is a constant coefficient (Table 4). The effect of pressure increases conductivity by about 20% between the surface and 200 km.

For olivine, several independent laboratory measurements suggest that:

$$\lambda_R = 0.368 \times 10^{-9} T^3 \quad (16)$$

This equation leads to $0.13 < \lambda_R < 1.8 \text{ W m}^{-1} \text{ K}^{-1}$ for $700 < T < 1700 \text{ K}$. It is only valid in a single crystal if the mean free path of photons is independent of temperature. For mantle rocks, one must account for scattering and for the effect of interfaces in a mineral assemblage. Such complications led Marton et al. (2005) to use a constant radiative conductivity component $\lambda_R = 1 \text{ W m}^{-1} \text{ K}^{-1}$ for temperatures higher than 700 K.

Bibliography

- Abbott D, Burgess L, Longhi J (1994) An empirical thermal history of the Earth's upper mantle. *J Geophys Res* 99:13835–13850
- Ahern JL, Mrkvicka SR (1984) A mechanical and thermal model for the evolution of the Williston basin. *Tectonics* 3:79–102
- Ashwal LD, Morgan P, Kelley SA, Percival J (1987) Heat production in an Archean crustal profile and implications for heat flow and mobilization of heat producing elements. *Earth Planet Sci Lett* 85:439–450
- Bell DR, Schmitz MD, Janney PE (2003) Mesozoic thermal evolution of the southern African mantle lithosphere. *Lithos* 71:273–287
- Canil D (2008) Canada's craton: a bottoms-up view. *GSA Today* 18:4–10
- Carlson RW, Pearson DG, James DE (2005) Physical, chemical, and chronological characteristics of continental mantle. *Rev Geophys* 43:2995–3007
- Clauser C, Huenges E (1995) Thermal conductivity of rocks and minerals. In: Ahrens TJ (ed) *A handbook of physical constants: rock physics and phase relations*. AGU, Washington, DC, pp 105–126
- Clauser C, Gieses P, Huenges E, Kohl T, Lehmann H, Rybach L, Safanda J, Wilhelm H, Windlow K, Zoth G (1987) The thermal regime of the crystalline continental crust: implications from the KTB. *J Geophys Res* 102:18,417–18,441
- Davaille A, Jaupart C (1993) Transient high-Rayleigh-number thermal convection with large viscosity variations. *J Fluid Mech* 253: 141–166
- Fountain DM, Salisbury MH, Furlong KP (1987) Heat production and thermal conductivity of rocks from the Pikwitonei-Sachigo continental cross section, Central Manitoba: implications for the thermal structure of Archean crust. *Can J Earth Sci* 24:1583–1594
- Grove TL, Parman SW (2004) Thermal evolution of the earth as recorded by komatiites. *Earth Planet Sci Lett* 219:173–187
- Haxby WF, Turcotte DL, Bird JM (1976) Thermal and mechanical evolution of the Michigan basin. *Tectonophysics* 36:57–75
- Herzberg CT (1983) Solidus and liquidus temperatures and mineralogies for anhydrous garnet lherzolite to 15 GPa. *Phys Earth Planet Inter* 32:193–202
- Jaupart C, Mareschal JC (2014) Constraints on crustal heat production from heat flow data. In: Rudnick RL (ed) *Treatise on geochemistry*, 2nd ed, The crust, vol 4. Elsevier, New York, pp 65–84
- Jaupart C, Mareschal JC (2015) Heat flow and thermal structure of the lithosphere. In: Watts AB (ed) *Treatise on geophysics*, 2nd ed The lithosphere, vol 6. Elsevier, New York, pp 218–253
- Kaminski E, Jaupart C (2000) Lithosphere structure beneath the Phanerozoic intracratonic basins of North America. *Earth Planet Sci Lett* 178:139–149
- Ketcham RA (1996) Distribution of heat-producing elements in the upper and middle crust of southern and west Central Arizona: evidence from the core complexes. *J Geophys Res* 101: 13,611–13,632
- Kinzler RJ, Grove TL (1992) Primary magmas of Mid-Ocean ridge basalts 2. Applications. *J Geophys Res Solid Earth* 97(5):6907–6926
- Kremenetsky AA, Milanovsky SY, Ovchinnikov LN (1989) A heat generation model for the continental crust based on deep drilling in the Baltic shield. *Tectonophysics* 159:231–246
- Kukkonen IT, Jokinen J, Seipold U (1999) Temperature and pressure dependencies of thermal transport properties of rocks: implications for uncertainties in thermal lithosphere models and new laboratory measurements of high-grade rocks in the central Fennoscandian shield. In: *Surveys in geophysics*, vol 20. Kluwer, Dordrecht, pp 33–59
- Landstrom O, Larson SA, Lind G, Malmqvist D (1980) Geothermal investigations in the Bohus granite area in southwestern Sweden. *Tectonophysics*. 64:131–162
- Levy F, Jaupart C, Mareschal JC, Bienfait G, Limare A (2010) Low heat flux and large variations of lithospheric thickness in the Canadian shield. *J Geophys Res Solid Earth* 115:408
- Mareschal JC, Jaupart C (2004) Variations of surface heat flow and lithospheric thermal structure beneath the North American craton. *Earth Planet Sci Lett* 223:65–77
- Mareschal JC, Poirier A, Rolandone F, Bienfait G, Gariépy C, Lapointe R, Jaupart C (2000) Low mantle heat flow at the edge of the North American continent, Voisey Bay, Labrador. *Geophys Res Lett* 27:823–826
- Marton FC, Shankland TJ, Rubie DC, Xu Y (2005) Effects of variable thermal conductivity on the mineralogy of subducting slabs and implications for mechanisms of deep earthquakes. *Phys Earth Planet Inter* 149:53–64
- McKenzie D, Bickle MJ (1988) The volume and composition of melt generated by extension of the lithosphere. *J Petrol* 29:625–679
- Michaut C, Jaupart C, Bell DR (2007) Transient geotherms in Archean continental lithosphere: new constraints on thickness and heat

- production of the subcontinental lithospheric mantle. *J Geophys Res Solid Earth* 112(11):04408
- Nicolaysen LO, Hart RJ, Gale NH (1981) The Vredefort radioelement profile extended to supracrustal strata at Carletonville, with implications for continental heat flow. *J Geophys Res* 86:10,653–10,661
- Nunn JA, Sleep NH (1984) Thermal contraction and flexure of intracratonal basins: a three-dimensional study of the Michigan basin. *Geophys J R Astron Soc* 79:587–635
- Perry HKC, Jaupart C, Mareschal JC, Shapiro NM (2006) Upper mantle velocity-temperature conversion and composition determined from seismic refraction and heat flow. *J Geophys Res Solid Earth* 111(10):07301
- Pollack HN (1986) Cratonization and thermal evolution of the mantle. *Earth Planet Sci Lett* 80:175–182
- Rudnick RL, Gao S (2014) Composition of the continental crust. In: Rudnick RL (ed) *Treatise on geochemistry* (2nd ed), The crust, vol 4. Elsevier, New York, pp 1–51
- Rudnick RL, Nyblade AA (1999) The thickness and heat production of Archean lithosphere: constraints from xenolith thermobarometry and surface heat flow. In: Fei Y, Bertka CM, Mysen BO (eds) *Mantle petrology: field observations and high pressure experimentation: a tribute to Francis R. (Joe) Boyd*. The Geochemical Society, Houston, pp 3–12
- Russell JK, Kopylova MG (1999) A steady-state conductive geotherm for the north central slave, Canada: inversion of petrological data from the Jericho kimberlite pipe. *J Geophys Res* 104:7089–7101
- Solomatov VS (1995) Scaling of temperature- and stress-dependent viscosity convection. *Phys Fluids* 7:266–274
- Vicker PA (1997) Garnet peridotite xenoliths from kimberlite near Kirkland Lake, Canada. In: Unpublished MS thesis. University of Toronto, p 127
- Xu Y, Shankland TJ, Linhardt S, Rubie DC, Langenhorst F, Klasinski K (2004) Thermal diffusivity and conductivity of olivine, wadsleyite and ringwoodite to 20 GPa and 1373 K. *Phys Earth Planet Inter* 143:321–336

Lithosphere, Mechanical Properties

Evgueni Burov
(Deceased)

Synonyms

Geosphere

Definition

Lithosphere (mechanical). This is the rigid (“litho” = stone) outer layer of the Earth that remains mechanically strong over geological time spans. *Mechanical lithosphere* includes rigid layers of crust and outermost mantle capable to maintain high differential tectonic stresses, from 10 MPa to 1 GPa. *Mechanical lithosphere* is 1.5–2 times thinner than “seismic,” “thermal,” or “chemical” lithosphere.

Mechanical properties of the lithosphere. This term refers to the *integrated strength* of lithospheric plates, their *rheological structure* and parameters, and mechanical behavior in response to various tectonic loads.

Introduction

Mechanical properties of the lithosphere are of primary importance for local and global geodynamics. In particular, compared to the convective mantle, high long-term mechanical strength makes the lithosphere a unique stress/strain guiding and accumulating envelope with lasting mechanical memory. High strength prohibits internal *heat advection*, so *thermal conduction* is main heat transfer mechanism in the lithosphere, in contrast to the convective mantle. High strength also stabilizes vertical lithological structure of lithosphere making it a *stagnant layer*. In contrast to viscous mantle, long-term *rheology* of the lithosphere is strongly influenced not only by its *ductile* but equally *elastic* and *brittle* properties. It is probably the nonviscous properties of the lithosphere that shape it in the characteristic *plate tectonics* patterns.

The term *lithosphere* has been introduced in the second half of the nineteenth century, while the notion of *mechanical lithosphere* appeared in early twentieth century, in conjunction with that of *seismic lithosphere* (see ► “[Earth’s Structure, Global](#)”), after formulation of the *continental drift theory* by Wegener and first interpretations of *regional isostasy* by J. Barrell and Vening-Meinesz (Watts 2001; see entry ► “[Isostasy](#)”). The fact that the lithosphere has finite measurable strength has been demonstrated from observations and models of *regional isostatic compensation* of large topographic loads. Before that, the lithosphere was considered either as a very strong solid layer (*Pratt’s model*) or, in turn, a weak fractured layer (*Airy’s model*). *Postglacial rebound studies* of early twentieth century have contributed to the definition of the *mechanical lithosphere* as the uppermost layer of the solid earth characterized by slow *viscoelastic relaxation*, in contrast to the underlying, relatively low-viscosity asthenosphere. The long-term mechanical base of the lithosphere, h_m , is limited by the depth to isotherm 500–600 °C in oceans and 700–800 °C in continents, compared to almost twice as deep 1,330 °C isotherm delimiting the thermal lithosphere (see entries ► “[Lithosphere, Oceanic: Thermal Structure](#)” and ► “[Lithosphere, Continental: Thermal Structure](#)”). As suggested on the basis of recent mantle–lithosphere interaction models (e.g., Schmeling et al. 2008), it is the elastic and plastic properties of the lithosphere that essentially determine the geometry of lithospheric plates and the mechanisms of formation of constructive, destructive, and transform plate boundaries at global scale. At smaller scale, *mechanical properties* of the lithosphere control formation and evolution of major geological structures such as rifts, passive margins, foreland basins, mountain ranges,

plateau or strike-slip faults. They also control short-term processes such as seismicity (Watts and Burov 2003).

Mechanical Properties at Different Timescales

Mechanical properties of the lithosphere are timescale dependent (e.g., Watts 2001). At seismic timescales (<0.5 h), the lithosphere behaves as an elastic or brittle-elastic layer of a thickness largely exceeding h_m . At *post-seismic timescales* (from days to years), the uppermost crust relaxes a part of seismically induced stress evoking some rather poorly understood short-timescale viscoelastic properties. At *postglacial rebound* timescales ($\sim$ several 1,000 year), lithosphere remains largely elastic, with thickness of the elastic domain almost as important as its *seismic thickness*. At geodynamic timescales (>1 Myr), the thickness of the *elastic core* is reduced to values smaller than h_m , and the stresses are slowly relaxed at timescales on the order of several Myr. The long-term properties of the lithosphere were first assessed from geodynamic-scale observations, first of all, *regional isostasy*. According to these data, the lithosphere exhibits a large spectrum of long-term behavior from quasi-elastic to brittle and viscous. Observation of crustal and lithosphere scale faults and *distributed seismicity* shows that some domains within the lithosphere deform in brittle-plastic regime over long time spans, while relaxation of deformation below, for example, oceanic volcanic islands, points to long-term viscoelastic strength. However, the long-term properties of the lithosphere are more generally studied indirectly using extrapolations from *rock mechanics data*.

Sources of Information on the Mechanical Properties of the Lithosphere

Flexural studies and *experimental rock mechanics data* are major sources of quantitative information on long-term behavior and strength of the lithosphere. In structured *viscous-elastic-plastic* media, all rheological properties are interrelated, and various time- and scale-dependent factors may cause variations in the effective elastic, brittle, and ductile deformation. Consequently, observations of long-term deformation are of primary importance for assessment of the effective mechanical properties of the lithosphere. These and other sources of information are summarized below:

1. Observations of long-term ($t > 1,000$ year) lithosphere response to tectonic loading, in the order of importance:
 - Observations of *regional isostatic compensation: gravity-flexural studies* providing estimates for the *equivalent elastic thickness* of the lithosphere, T_e
 - Vertical motions due to *postglacial rebound, lake, and volcanic island* loading
 - Observations of *lithosphere folding* (*folding wavelength* is a function of plate strength)
 - Field observations of ductile and brittle deformation in the outcrops
 - Interpretations of *deformational microstructures* and *paleo-stresses*
 - *Seismic tomography* (e.g., evidence for more or less strong slabs at depth)
 - *Borehole* stress measurements
2. Intermediate timescale observations ($t > 3\text{--}10$ year):
 - *Slow-rate rock mechanics experiments* (strain rates 10^{-9} to 10^{-4} s $^{-1}$, for ductile properties)
 - Geodetic (GPS-INSAR) data over >5 year time spans (strains, viscoelastic properties)
 - *Inter-seismic deformation; slow earthquake data*
3. Observations of deformation in response to short-term loading ($t < 1$ year):
 - Short-term *rock mechanics experiments* (elastic and brittle properties)
 - Distribution of intraplate seismicity (brittle properties)
 - Tidal deformation (viscoelastic properties)
 - Post-seismic relaxation (viscoelastic properties)
 - Geodetic (GPS-INSAR) data (strains, viscoelastic properties)
 - Attenuation of S waves (proxy to low-viscosity zones)
 - Magnetotelluric sounding (reduced electrical resistivity is proxy to low-viscosity zones)
4. Physical considerations and self-consistent thermo-mechanical models:
 - Estimates of the minimal *integrated strength* of the lithosphere required for lifetime stability of geological structures, *subduction* or transmission of tectonic stresses, and forces over large spatial scales, including horizontal pressure gradients caused by lateral variations in lithospheric density structure and topography (*gravity potential energy* theory). For example, lithosphere must be strong enough to transmit *ridge push* and *slab pull* forces on the order of $10^{11}\text{--}10^{13}$ N per unit length.
 - Lithosphere scale *numerical thermo-mechanical models* of tectonic processes integrating multidisciplinary data, which allows for testing the validity of data and hypotheses on lithosphere rheology.

Observations of long-term deformation provide key parameters such as the *integrated strength* of the lithosphere. These parameters are needed to constrain *rock mechanics data* obtained at laboratory conditions, because they are too far from geological conditions: short timescales ($t < 5$ years), small spatial scales ($l \sim 0.1$ m), high strain rates ($\dot{\epsilon} > 10^{-9}$ s $^{-1}$), small strains ($\epsilon < 10$), high temperatures, simple deformation, largely mono-phase samples. *Rock*

mechanics data allow only for assessment of general form of rheology laws, their sensitivities, and relative strengths of different kinds of rocks. Their extrapolation to geodynamic scales ($t > 10^6$ years, $l \sim 1,000$ km, $\dot{\epsilon} < 10^{-14} \text{ s}^{-1}$, $\epsilon > 100$, cold temperatures, aggregate rocks) needs re-parameterization and validation using real-scale observations and models.

Interpretation of short-term data (seismic, satellite geodesy, INSAR) is not straightforward. In particular, interpretation of *intraplate seismicity* (see entry ► “[Seismicity, Intraplate](#)”) and *post-seismic relaxation* data is questioned due to the lack of evidence that mechanisms of this short-term deformation can be physically linked to those of long-term deformation.

Observations of Flexural Behavior and Effective Long-Term Strength of the Lithosphere

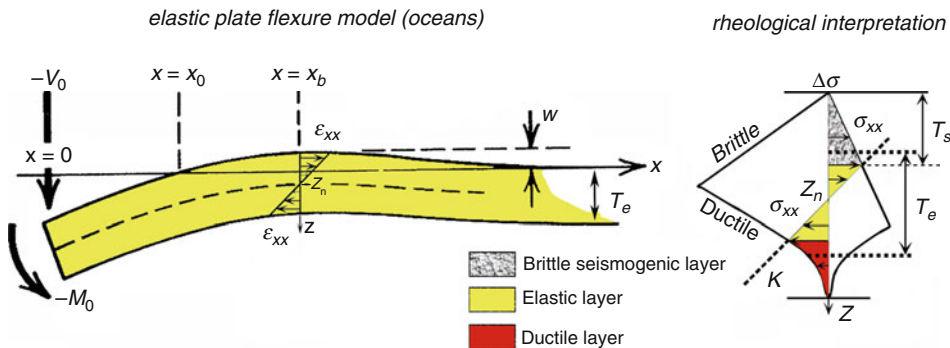
Observations of *regional isostatic compensation* (e.g., Watts 2001; see entry ► “[Isostasy](#)”) have shown that the lithosphere has substantial long-term *elastic rigidity* that allows for transferring and maintaining intraplate stress levels (10 MPa–1 GPa) over geodynamic time spans (>several Myrs). Studies of gravity anomalies observed over mountain ranges and subduction zones have demonstrated that lithospheric plates bend like *thin elastic plates* of finite stiffness in response to tectonic, topography, and sedimentary loads. With improvement of geophysical measurement techniques during the second half of twentieth century, multiple studies (specifically, forward *flexural models*) have produced robust estimates of *flexural rigidity*, D , and *equivalent elastic thickness*, T_e , of lithospheric plates. These data arguably present a major source of information on the long-term *mechanical properties* of lithosphere. *Flexural studies* reveal strong variations of lithosphere strength, from near zero at ocean ridges to 100–150 km thick quasi-elastic cores detected within old and cold cratons (Burov and Diamant 1995; Kirby and Swain

2009). In *flexural models* of *regional isostatic compensation* (Watts 2001), D is varied until the model-predicted basement or *Moho topography* provides optimal fit to observations. *Gravity data* (see entry ► “[Gravity Anomalies, Interpretation](#)”) are used when basement topography is hidden (e.g., by sediments) or not representative of flexure (e.g., modified by erosion). The most robust *gravity models* are *forward models*. *Inverse models* are widespread due to the ability to cover large zones, but their results are more sensible to errors and should be cross-checked with *forward models*. For example, if not properly formulated, *gravity admittance* techniques may generate spurious results in continents (e.g., Kirby and Swain 2009).

D provides a direct measure for the *integrated long-term strength* of the lithosphere and is linked to the *equivalent elastic thickness* of the lithosphere, $T_e: D = ET_e^3(12(1 - \nu^2))^{-1}$, where E and ν are Young’s modulus and Poisson’s ratio, respectively. *Plate bending or flexure* is characterized by its vertical deflection, $w(x)$ and local radius of curvature, $R_x(x)$ or curvature, $K(x) = -R_x^{-1} = \partial^2 w / \partial x^2$ (Fig. 1). The flexural equation, when expressed using *bending moment* $M_x(x)$ is rheology independent and is valid for all, elastic and inelastic plates:

$$\frac{\partial^2}{\partial x^2} \left(\overbrace{\frac{ET_e^3}{12(1 - \nu^2)} \frac{\partial^2 w(x)}{\partial x^2}}^{M_x(x)} \right) + \frac{\partial}{\partial x} \left(F_x \frac{\partial w(x)}{\partial x} \right) + \Delta \rho g w(x) = \rho_c g h(x) + p(x) \quad (1)$$

where F_x is horizontal *fiber force*, $\Delta \rho$ is the density contrast between surface and subsurface material (i.e., between mantle density ρ_m and topography/sediment/water density ρ_c), $h(x)$ is topography elevation, and $p(x)$ is additional surface or subsurface load. The elasticity is used just as simplest



Lithosphere, Mechanical Properties, Fig. 1 Classical flexural model of oceanic lithosphere (left). Right: brittle-elastic-ductile yield stress envelope (YSE) and interpretation of the equivalent elastic thickness T_e of the lithosphere. $\epsilon_{xx}(z)$ is flexural strain, $\sigma_{xx}(z)$ is flexural stress,

$K(x)$ is local plate curvature, $\Delta\sigma$ is differential stress, $Z_n(x)$ is neutral plane, and T_s is brittle seismogenic layer. $w(x)$ is vertical plate deflection. V_0 and M_0 are boundary cutting force and flexural moment, respectively

rheological interpretation of bending strength: elastic bending stress is linear function of curvature and depth: $\sigma_{xx}(x, z) \approx (0.5T_e - z)EK(1 - \nu^2)$. T_e is thus effective parameter and should not be automatically related to any real layer within the lithosphere. For inelastic plates, T_e and D have a sense of “condensed” plate strength and are direct proxies for the long-term *integrated strength*, B , of the lithosphere (Watts 2001). For example, for a single-layer oceanic plate ($T_e = T_{e_ocean}$, Fig. 1):

$$B = \int_0^{h_m} \sigma^f(x, z, t, \dot{\epsilon}) dz$$

$$\text{while } T_{e_ocean} = \left(12 \left(\frac{\partial \sigma_{xx}^f}{\partial y} \right)^{-1} \left(\int_0^{h_m} \overbrace{\sigma_{xx}^f(z - Z_n) dz}^{M_x(x)} \right) \right)^{\frac{1}{3}};$$

$$T_{e_ocean} < h_m$$

(2)

where σ_{xx}^f is brittle-elastic-ductile bending stress (Burov and Diamant 1995); T_e is usually much smaller than h_m . M_x and, hence, D can be obtained from depth integration of σ_{xx}^f . D and T_e may spatially vary due to their dependence on local bending that leads to localized plate weakening (called *plastic or ductile hinging*) in the areas of utmost flexure, for example, near *subduction zones* or below mountains and islands.

Rheological Properties of Lithosphere According to Rock Mechanics Data

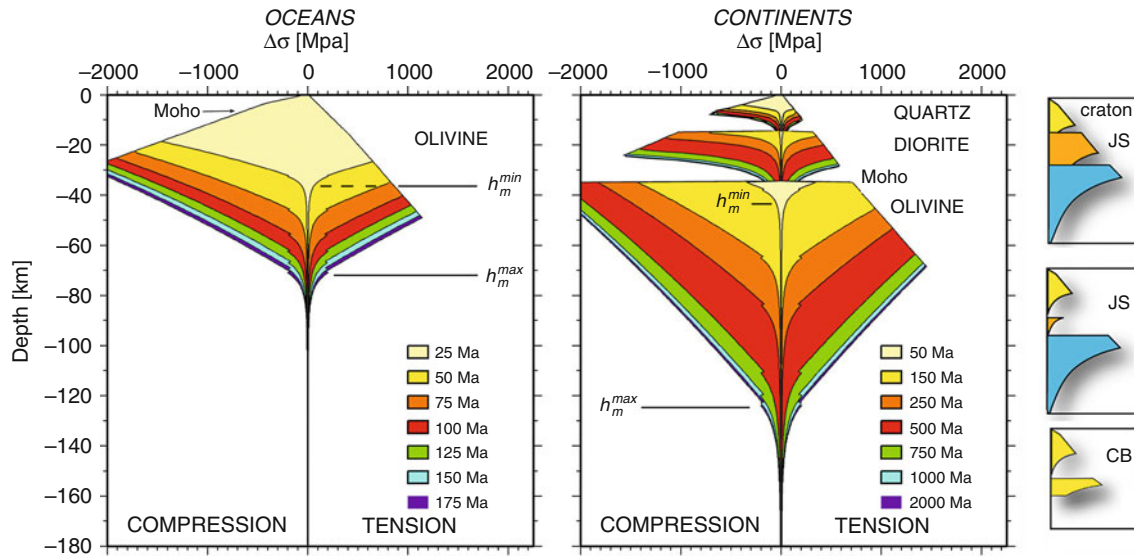
The long-term mechanical behavior of rocks is represented by *extended Maxwell solid*, in which total strain increment equals a sum of elastic, viscous (ductile), and plastic (brittle) increments while the elastic, viscous, and plastic stresses are mutually equal. The weakest rheological term thus defines the effective behavior of this solid. Goetze and Evans (1979) used *Maxwell solid* and rock mechanics data to introduce the *yield stress envelope* (YSE) of the lithosphere (Figs 1 and 2). This approach consists in predicting, for a representative *background strain rate* and depth–pressure–temperature profile, the maximal *yield strength* $\Delta\sigma_{\max}(z)$ as function of depth, z . If elastic *differential stress* estimate $\Delta\sigma^e(z) < \Delta\sigma_{\max}(z)$, the deformation is elastic and *differential stress* $\Delta\sigma(z) = \Delta\sigma^e(z)$. If $\Delta\sigma^e(z) \geq \Delta\sigma_{\max}(z)$, then $\Delta\sigma(z) = \Delta\sigma_{\max}(z)$ and the deformation is brittle or ductile depending on z .

Elastic Properties

The *elastic behavior* is described by linear Hooke’s law:

$$\sigma_{ij} = \lambda \epsilon_{ii} \delta_{ij} + 2G \epsilon_{ij} \quad (3)$$

where λ and G are Lamé’s constants. Repeating indexes mean summation, δ is Kronecker’s operator. For most rocks $\lambda \approx G = 30$ GPa (Turcotte and Schubert 2002).



Lithosphere, Mechanical Properties, Fig. 2 Commonly inferred (brittle) rheological *yield stress envelopes* (YSEs) as function of *thermotectonic age* for oceans and continents. In mantle, maximal strength can be limited by *Peierls* or *GBS* law instead of

Byerlee’s law. For continents, variations in crustal composition and fluid content result in various *Jelly Sandwich*, *Jelly Sandwich*, and more rare *Crème Brûlée*. (After Burov (2007))

Brittle-Plastic Properties

Brittle resistance, τ , is a linear function of normal stress, σ_n and, that is, of pressure (Byerlee 1978):

$$\begin{aligned} \tau &= 0.85\sigma_n, \quad \sigma_n \leq 200 \text{ MPa} \\ \tau &= 50 \text{ MPa} + 0.6\sigma_n, \quad 1,700 \text{ MPa} > \sigma_n > 200 \text{ MPa} \end{aligned} \quad (4)$$

Byerlee's law is equivalent of Mohr-Coulomb (Coulomb-Navier) plastic failure criterion (e.g., Burov 2007):

$$\tau = C_0 + \tan(\phi)\sigma_n \quad (5)$$

where C_0 is cohesive strength (<20 MPa) and ϕ is the internal friction angle (30 – 33°).

The Griffith criterion extends (5) to pure tensile domain:

$$\tau^2 = 4T_0^2 + 4T_0\sigma_n \quad (6)$$

where $T_0 = C_0/2$ is tension cut-off.

Viscous-Ductile Properties

Effective viscosity of the lithosphere varies from 10^{19} Pa s at the lithosphere-asthenosphere boundary to $>10^{26}$ Pa s in cold mantle near Moho depth. Several ductile mechanisms such as diffusion creep, grain boundary sliding (GBS), pressure solution, and cataclastic flow may play important role at appropriate conditions, but the leading part belongs to the dislocation creep (Kohlstedt et al. 1995).

$$\begin{aligned} \dot{\epsilon}^d &= Af_w \Delta\sigma^n \exp(-H(RT)^{-1}) \\ &\text{for } \Delta\sigma < 200 \text{ MPa (Power law)} \\ \dot{\epsilon}^d &= Af_w \exp(-H(1 - \Delta\sigma/\sigma_p)^2/RT) \\ &\text{for } \Delta\sigma > 200 \text{ MPa (Harper – Dorn law)} \end{aligned} \quad (7)$$

where $\dot{\epsilon}^d$ is shear strain rate, A is material constant, n is power law constant, f_w is water fugacity factor, $\Delta\sigma$ is differential stress, R is universal gas constant, $H = Q + PV$ is creep activation enthalpy, Q is activation energy (100 – 600 kJ/mol), P is pressure, V is activation volume, T is temperature in K. σ_p is Peierls-like stress ($\sigma_p \sim$ several GPa). For tectonically relevant $\Delta\sigma/\sigma_p$ ratios (<0.1), Harper-Dorn's flow refers to Peierls plasticity that limits rock strength in high stress regime. Dislocation creep is strongly nonlinear non-Newtonian viscous flow ($n \sim 2$ – 4).

The second most important deformation mechanism, low-stress diffusion creep (Nabarro-Herring, Coble), results from directional diffusivity of rocks under applied stress (Kohlstedt et al. 1995):

$$\dot{\epsilon}^d = Aa^{-m}f_w\Delta\sigma^n \exp(-H(RT)^{-1}) \quad (8)$$

where a is grain size and m is diffusion constant. For olivine, $m \sim 3$ and $n \sim 1$ and the constitutive law is linear Newtonian.

The Peierls super-plasticity is likely to replace Byerlee's law in the mantle below 30 – 40 km depth (Kameyama et al. 1999), that is, at high confining pressures (>1 GPa). Since Peierls plasticity includes strong water-weakening ($H = Q + p(V - \beta\Delta V_w)$), where ΔV_w is molar volume change due to incorporation of hydroxyl ions in the main rock and β is experimental parameter, it may play an important role in localization of deformation in subduction zones. This role may be shared with GBS creep, which might be responsible for aseismic localization of deformation in the lithosphere mantle.

Mechanical Properties of Oceanic Lithosphere Versus Continental Lithosphere

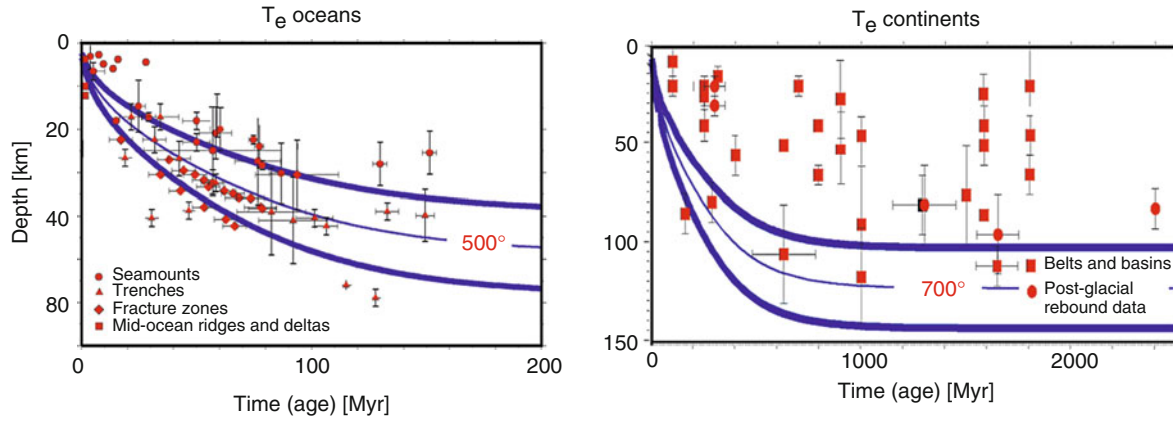
Due to its temperature dependence, ductile strength limits the depth to the base of mechanical lithosphere, h_m , to that of the isotherm 500 – 600 °C in oceans and 700 – 800 °C in continents (in continents, higher pressure for given temperature increases yield strength via the H term in Equation 7).

Oceanic YSEs ($0 < T_e < 50$ – 70 km) derived from (Eqs. 3, 4, 5, 6, 7, and 8) predict brittle seismogenic behavior in the upper parts of lithosphere, where depth of brittle-ductile transition, BDT, varies from few km near ridges to 40 km near subduction zones (Figs. 2 and 3). The competent domains below the elastic core are dominated by aseismic ductile creep, where important ductile strength is preserved down to the depths of 80 – 100 km near subduction zones.

Continental YSEs (0 – $5 < T_e < 150$ km) reflect strong rheological stratification between the upper, intermediate, lower crust and mantle lithosphere (Figs. 2 and 3). There might be several BDT depths, typically at 15 – 25 and 30 – 45 km (Watts and Burov 2003; Burov 2007, 2010). The depth to the mechanical bottom of strong ductile lithosphere may vary from 30 to 200 km.

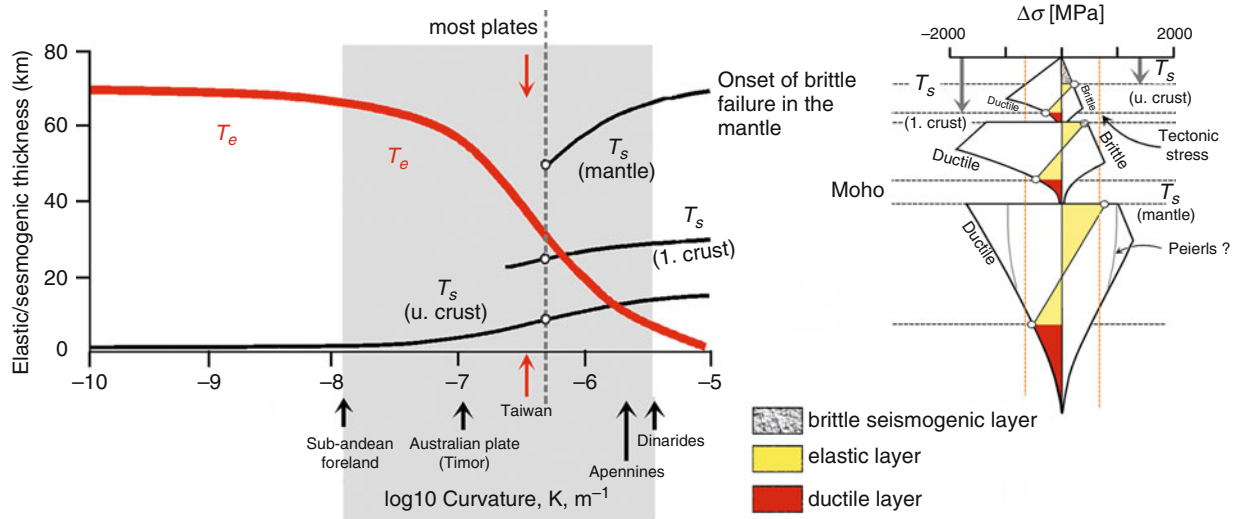
T_e estimates (Fig. 3) reveal important difference between the mechanical properties of oceanic and continental lithosphere. In oceans, T_e (<70 km) follows the geotherm of 500 ± 100 °C and correlates with age. It fits into the mechanically competent core (Fig. 1).

In continents, maximal T_e values in cratons are 110 – 150 km, while T_e distribution reveals complex “bimodal” behavior (clustering around 25 – 30 and 70 km), which does not allow for geometrical interpretation of T_e . This complexity is explained by rheological stratification of continental plates, specifically of their 35 – 70 km thick crust (see entry ► “Lithosphere, Continental”). Continental plates consist of several layers of contrasting strength (upper, lower,



Lithosphere, Mechanical Properties, Fig. 3 Observed T_e distribution in oceans and continents as function of *thermotectonic age* of the lithosphere, compared to depth of geotherms of 500 °C and 700 °C defining, respectively, the mechanical base of oceanic and continental lithosphere (Burov and Diamant 1995; Watts 2001; see also “► *Isostasy*”). Assumed thermal thickness of continental lithosphere is 250 km.

Depending on assumed thermal thickness and boundary conditions, predicted depth to 500–700 °C may shift up or down by up to 25%. In continents, crust-mantle decoupling leads to ~50% T_e reduction compared to the depth to its mechanical base at 700 °C. This explains occasional correlation of T_e with 400 °C depth



Lithosphere, Mechanical Properties, Fig. 4 Mechanical and rheological interpretation of continental T_e and seismogenic layer thickness T_s (right) computed for a representative YSE as function of local plate

curvature K (left). After Burov (2010). Homogeneous horizontal stresses will shift T_e and T_s up or down in opposite directions, but will not change the character of their dependence on plate curvature

intermediate crust, mantle) that can be mechanically coupled or uncoupled. For plates younger than 200–400 Ma, T_e grows with age yet exhibiting large scatter of values between 0 and 70 km. Older plates (>500–700 Ma) are in stationary thermal regime; the age dependence of T_e is expectedly small, but T_e scatter is still strong revealing influence of factors other than thermal age. Using rock mechanics data (Fig. 4), it can be shown that continental T_e should be strongly controlled by the crustal thickness and mechanical state of crust-mantle interface (Burov and Diamant 1995). When the lower crust is mechanically strong, crust and mantle are mechanically coupled yielding a single strong layer with high T_e :

$$T_e \approx h_1 + h_2 \dots = \sum_n h_i \quad (9)$$

where h_i are thicknesses of crustal and mantle competent layers. Mechanical decoupling between crust and mantle leads to *structural weakening*, that is, dramatically smaller T_e (Burov and Diamant 1995):

$$T_e \approx (h_1^3 + h_2^3 \dots)^{1/3} = \sqrt[3]{\sum_n h_i^3} \approx \max(h_i) < \sum_n h_i \quad (10)$$

“Decoupled” T_e roughly equals to the *mechanical thickness* of strongest layer in the lithosphere. In particular, $T_e \geq h_c$ (h_c is crustal thickness) identifies mantle as the strongest layer. Based on T_e data it has been shown that continental crust is systematically mechanically decoupled from mantle, except for cratons, and that mantle lithosphere is generally stronger or at least as strong as the crust (Burov and Diamant 1995). *Numerical thermo-mechanical models* have later confirmed these assertions (e.g., Burov 2007, 2010).

Mechanical Properties of the Lithosphere and Four Modes of Horizontal Deformation

In addition to flexure under normal loads, four modes of lithosphere deformation and their combinations can develop under extensional or compressional conditions: (1) *pure shear* (extension: *McKenzie rifting model*; compression: *plate collision*), (2) *simple shear* (extension: *Wernicke rifting model*; compression: *subduction*), (3) *tensional/compressional instability* (extension: *boudinage*; compression: *folding-buckling*), and (4) *gravitationally unstable pure shear* (slow passive margins, *Rayleigh-Taylor instabilities* due to thickening of a colder and thus denser lithosphere mantle overlying hotter and thus lighter asthenosphere). Ductile properties largely control these deformation modes, including brittle fault spacing observed at surface (e.g., slow oceanic spreading centers vs. fast spreading centers). Brittle properties play an important role as well, specifically in localization of small and large-scale deformation, dyking, stabilization of cratons, transmission of tectonic stresses, and seismicity. Some of brittle mechanisms such as *low-angle faulting*, localization of large-scale *strike-slip* or *transform faults* and large-scale *dyke propagation* remain enigmatic.

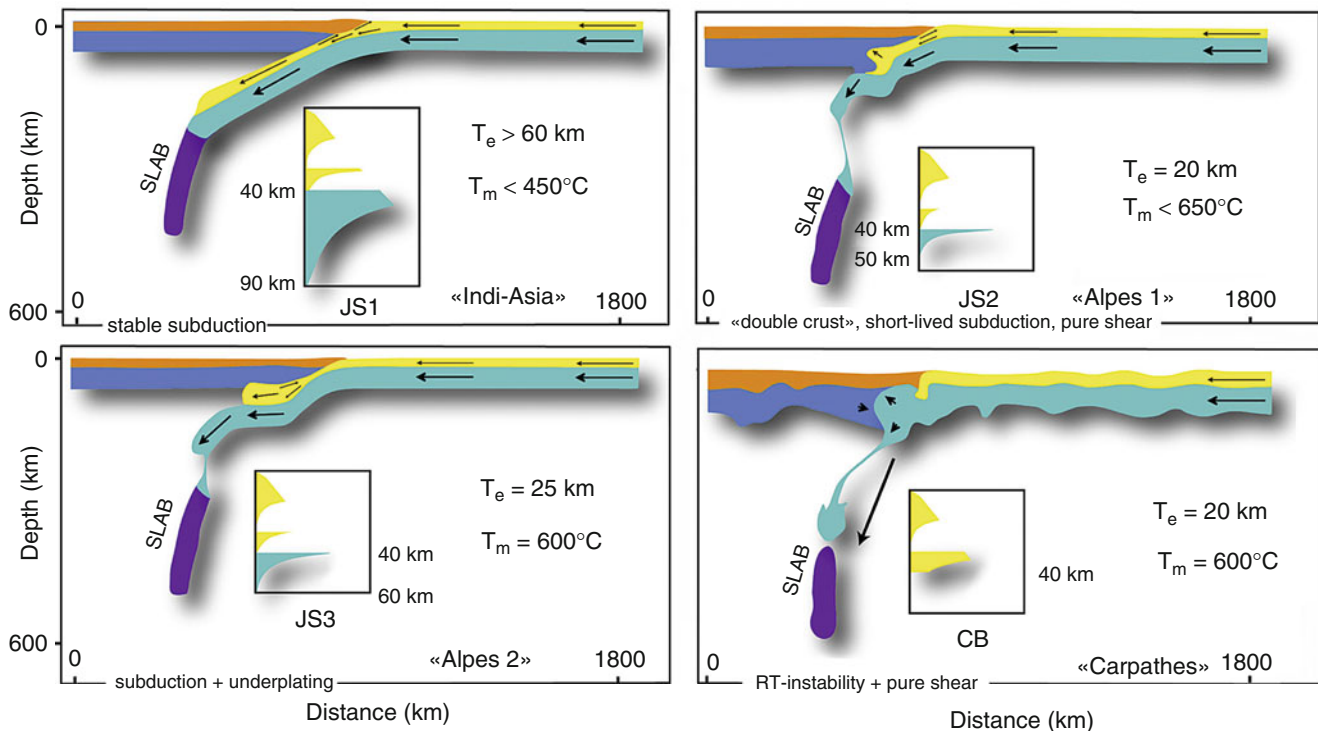
Rheological stratification of continents is likely to influence the response of the lithosphere in all deformation modes. *Crust-mantle coupling* results in formation of narrow deep rifts while decoupling at different crustal levels may produce a variety of basins, from shallow large basins to narrow rifts. Under compression, *crust-mantle decoupling* plays a major role for continental subduction, UHP rock exhumation, slab detachment, formation of plateaus, or *lithosphere folding*. Lateral lower crustal flow is believed to be one of the key mechanisms affecting orogenic building and rift/passive margin formation or assuring feedback between surface and subsurface processes. For *mantle-continental lithosphere interactions*, such as *plume impingement*, the presence of ductile lower crust results in damping of the long-wavelength *dynamic topography* at surface, which is replaced by short-wavelength tectonic-scale deformation.

Uncertainties of Data on the Mechanical Properties of the Lithosphere

T_e studies play a key role in interpretation of *rock mechanics data*, allowing for parameterization of laboratory-based *flow laws* to geological time and spatial scales. Indeed, due to the large differences between time (10 orders) and spatial scales (6 orders) of the laboratory experiments and nature, *rock mechanics data* cannot be solely used to quantify the long-term *mechanical properties* of the lithosphere. Yet, since the *integrated strength B* of the lithosphere is directly related to T_e , one can use inelastic *flexural models* in combination with *rock mechanics data* (Eqs. 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10) to constrain, parameterize, and validate rheology parameters for geological timescales. In particular, one can derive YSEs that are compatible with the host of different observations such as T_e , heat flow, seismic imagery, gravity, and so on. There is generally a good correlation between the observed T_e and T_e predicted from experimental YSE (Watts 2001, Fig. 4). In oceans, T_e values scale well with *rock mechanics data* and *thermal models of the lithosphere* (e.g., McNutt and Menard 1982, Fig. 3). In continents, T_e rheology parameterization is delicate but can be done if one respects the physics of underlying processes (Burov and Diamant 1995; Burov 2010). Three types of *rheological envelopes* can be proposed for continental lithosphere: *Cratonic Jelly Sandwich*, *Jelly Sandwich*, *Crème brûlée* (with its variant hydrated *Banana Split* rheology) (Watts and Burov 2003; Burov 2010, Fig. 2). The first two types evoke strong mantle lithosphere that can be either mechanically coupled with the crust (*Cratonic type*) or decoupled from it by weak lower or intermediate crustal layer (pure *Jelly Sandwich*). These first two rheologies dominate and correspond to 80–90% of all cases, as confirmed by matching T_e predicted from YSEs with the observed T_e and mechanical models. The *Crème brûlée* rheology infers weak lithosphere mantle, and applies to few cases of hot lithospheres (e.g., *Basins and Range Province*).

Mechanical Properties of the Lithosphere and Styles of Tectonic Deformation

Important constraints on lithosphere rheology and properties come from physical considerations including *thermo-mechanical numerical models* of geodynamic processes. One can verify the validity of rheological assumptions by comparing predictions of thermo-mechanical models based on these assumptions with the observed tectonic evolution in various contexts such as collision and mountain belts (Fig. 5), rifts and passive margins, and subduction zones and lithosphere folding. The analytical and *numerical thermo-*



Lithosphere, Mechanical Properties, Fig. 5 Numerical thermo-mechanical models of long-term geological processes (here, continental collision) allow for validation of inferred rheology profiles (*Jelly Sandwich*, JS1–JS3, and *Crème Brûlée*, CB) by testing the compatibility of model-predicted and observed tectonic evolution. Models JS2, JS3, and CB have similar T_e (20–25 km) but different mantle strength, leading to

strong differences in resulting collision styles. The models show that strong mantle lithosphere is needed to drive continental subduction while strong crust cannot play same role (its low density prohibits subduction if there is no mantle drag). T_m – temperature at Moho. Colors: orange, yellow – crust; blue, azure – mantle; purple – oceanic slab. (Modified from Burov (2010))

mechanical models of geological processes show that predicted evolution is highly sensitive to rheology choices (Fig. 5). It has been demonstrated that in most cases mantle lithosphere has to be rheologically strong to insure stability of geological structures and the observed styles of deformation. In particular, stability of mountain ranges, continental *subduction* (e.g., Fig. 5, Burov 2010) and most observed *rifting modes* (Buck 1991; Bassi 1995) need initially strong mantle lithosphere with *mechanical thickness of the mantle part* more than 30–50 km. Collision models (Fig. 5) quantify relationships between the collision styles and deep lithosphere structure and rheology. They show that tectonic evolution depends not only on the integrated strength (T_e) but as much on strength distribution with depth, that is, on which particular lithological layer (upper crust, lower crust, or mantle lithosphere) provides main contribution to T_e . Rift models show that *narrow rifts* are associated with initially strong mantle lithosphere, while metamorphic core complexes “need” warm lithospheres with negligible mantle strength. Wavelengths of *boudinage* and deca-kilometric upper crustal fault spacing observed in rift zones also appear to be a direct function of the ductile properties of the underlying lower

crust. In compressional settings, observed *folding wavelengths* provide constraints on plate rheology since these wavelengths (50–1,000 km) are proportional to 5–10 times thickness of competent layers in the lithosphere and correlate with T_e values (e.g., Burov 2007).

Links Between Different Timescales: Burger's Rheology Model

The abundance of short-timescale observations such as *earthquake distributions* or *post-seismic relaxation* explains attempts to interpret these data in terms of long-term rheology. Yet, these attempts are typically based on misinterpretation of Goetze's YSEs that are valid only for geodynamic strain rates. Earthquakes and relaxation occur at locally and temporarily high strain rates that are not representative of long-term rates. Hence, the fact that continental lithosphere below Moho depth is mainly aseismic cannot be interpreted as a sign of weak ductile behavior (Fig. 4). It rather indicates that either mantle stress levels are insufficient to induce brittle sliding, or that seismicogenic *Byerlee's law* does not operate at high

pressure (>36–70 km depth), or that the frictional behavior is strain-rate dependent. At seismic strain rates (10^{16} times higher than geodynamic strain rates), the entire lithosphere acts as a brittle-elastic body and no ductile flow can occur. Strictly speaking, T_e and *maximal seismogenic depth* T_s should rather anti-correlate (Watts and Burov 2003, Fig. 4). However, T_s cannot be used to constrain T_e without knowing intraplate stress level. The fact that T_s distributions in oceans and continents are similar, while their rheological and mineralogical structures are different, adds to the argument that seismicity is primarily related to stress levels.

Post-seismic relaxation data provide controversial results yielding effective viscosities of deforming domains about 1–2 orders of magnitude smaller than *postglacial rebound* data that provide minimal viscosity (10^{19} to 5×10^{19} Pa s) of the Earth's weakest layer – asthenosphere. Since these estimates are based on inversion of *surface deformation*, they strongly depend on initial assumptions on lithosphere structure and properties, while it is impossible to determine which layer in the lithosphere relaxes post-seismic deformation. Nevertheless, whatever is this layer, the estimated viscosities appear too low for long-term properties. To explain this controversy, one can consider *Burger's model* of solids. According to this model, lithospheric behavior is described by two independent serially connected terms, one of which is *Kelvin solid* responsible for the primary creep (seismic, post-seismic) and the second one is *Maxwell solid* responsible for secondary longterm creep (geodynamic timescale). The viscosity of the first term is independent of that of the second term. In other words, the physics of deformation mechanisms activated at seismic-scale strain rates is different from that of the mechanisms acting at geodynamic strain rates.

Summary

Lithosphere has important long-term ductile, elastic, and brittle-plastic strength and is capable of maintaining (not relaxing) differential stresses at geological timescales. These properties can be accessed mainly from estimates of the *equivalent elastic thickness*, T_e , combined with *rock mechanics data*, validated by analytical and *numerical thermo-mechanical models* of geological and geodynamic processes. In oceans, T_e fits in the mechanical core of the lithosphere, but in continents, it generally does not represent any particular layer due to strong *rheological stratification*. The brittle-plastic properties of the lithosphere are governed by *Byerlee's law* to 30–40 km depth and likely by *Peierls* plasticity and/or *GBS* creep below. The ductile properties are dominated by *dislocation power-law creep*, which is strongly nonlinear and rock-type dependent. Highest *integrated strength* ($T_e \sim 100$ –150 km) is detected in cratons (*Cratonic Jelly Sandwich rheology*) where strong crust is mechanically coupled with strong mantle. In warmer

continental lithospheres, mechanical decoupling between crust and mantle leads to *structural weakening* resulting in ~50% reduction of T_e (*Jelly Sandwich rheology*). In most cases, mantle lithosphere provides main contribution to the integrated plate strength. The cases where mantle is weaker than crust (*Crème-Brûlée rheology*) refer to hot lithospheres such as metamorphic core complexes. There is probably no exploitable link between short-term deformation such as earthquake data or post-seismic relaxation and long-term properties of the lithosphere.

Cross-References

- Earth's Structure, Global
- Gravity Anomalies, Interpretation
- Isostasy
- Lithosphere, Continental
- Lithosphere, Continental: Thermal Structure
- Lithosphere, Oceanic: Thermal Structure
- Seismicity, Intraplate

Bibliography

- Bassi G (1995) Relative importance of strain rate and rheology for the mode of continental extension. *Geophys J Int* 122:195–210
- Buck WR (1991) Modes of continental extension. *J Geophys Res* 96:20161–20178
- Burov E (2010) The equivalent elastic thickness (T_e), seismicity and the long-term rheology of continental lithosphere: time to burn-out “crème brûlée”? Insights from large-scale geodynamic modeling. *Tectonophysics* 484:4–26
- Burov EB (2007) Plate rheology and mechanics. In: Schubert G (ed) *Treatise on geophysics*, vol. 6: crust and lithosphere dynamics (vol ed: Watts AB), vol 99, no 152. Elsevier, Amsterdam, 611 pp. TOGP00102. ISBN:978-0-444-51928-3
- Burov EB, Diament M (1995) The effective elastic thickness (T_e) of continental lithosphere: what does it really mean? *J Geophys Res* 100:3895–3904
- Byerlee JD (1978) Friction of rocks. *Pure Appl Geophys* 116:615–626
- Goetze C, Evans B (1979) Stress and temperature in the bending lithosphere as constrained by experimental rock mechanics. *Geophys J R Astron Soc* 59:463–478
- Kameyama M, Yuen D, Karato S-I (1999) Thermal-mechanical effects of low-temperature plasticity (the Peierls mechanism) on the deformation of a viscoelastic shear zone. *Earth Planet Sci Lett* 168:159–172
- Kirby JF, Swain CJ (2009) A reassessment of spectral T_e estimation in continental interiors: the case of North America. *J Geophys Res* 114: B08401. <https://doi.org/10.1029/2009JB006356>
- Kohlstedt DL, Evans B, Mackwell SJ (1995) Strength of the lithosphere: constraints imposed by laboratory experiments. *J Geophys Res* 100:17587–17602
- McNutt MK, Menard HW (1982) Constraints on yield strength in the oceanic lithosphere derived from observations of flexure. *Geophys J R Astron Soc* 71:363–395
- Schmeling H, Babeyko AY, Enns A, Faccenna C, Funicello F, Gerya T, Golabek CJ, Grigull S, Kaus BJP, Morra G, Schmalholz SM, van Hunen J (2008) A benchmark comparison of spontaneous subduction models – toward a free surface. *Phys Earth Planet Inter* 171:198–223

- Turcotte DL, Schubert G (2002) *Geodynamics*. Cambridge University Press, Cambridge, p 456
- Watts AB (2001) *Isostasy and flexure of the lithosphere*. Cambridge University Press, Cambridge, p 458
- Watts AB, Burov E (2003) Lithospheric strength and its relationship to the elastic and seismogenic layer thickness. *Earth Planet Sci Lett* 213:113–131

Lithosphere, Oceanic

James S. McClain

Department of Earth and Planetary Sciences, University of California, Davis, Davis, CA, USA

Definition

The term “lithosphere” is applied to a variety of concepts and observations, and the actual definitions for the term often depend on the type of observations made. Most generally, lithosphere refers to a strong and cool outer carapace for the Earth. It lies above a weaker, hotter, and more mobile asthenosphere. The oceanic lithosphere comprises a crustal component and a mantle component, where the latter is significantly thicker, but models for the former are better constrained. The boundary between the lithosphere and asthenosphere (LAB) is the subject of a great deal of study and controversy and remains unresolved. Models for the oceanic lithosphere are intimately connected to the ideas about the nature of LAB, and to the theory of seafloor spreading. Because the crust and upper mantle must undergo cooling as they spread away from the mid-ocean and back-arc spreading centers, it is anticipated that temperature-dependent properties of the LAB deepen as the oceanic plates age. Thus, virtually all models for the oceanic lithosphere include the constraint that it must thicken with age. Modern models allow for a maximum thickness for older lithosphere, because cooling diminishes at ages approaching 100 Ma. Constraints for thickening of the oceanic lithosphere, and the nature of the lithosphere/asthenosphere boundary, are supplied by observations of age-dependent changes in seafloor depth and heat flow, by observations of a seafloor deformation, and by a variety of seismic and magnetotelluric observations.

Introduction

The term “lithosphere” is applied to a variety of concepts and observations, and actual definitions for the term often depend directly or indirectly on the type of observations made. Most generally, lithosphere refers to a relatively strong and cool outer carapace for the Earth. The concept was initially put forth by Barrell (1914), who considered the need for a strong

outer layer of the Earth to sustain tectonic loads over geologic timescales. More broadly, the lithosphere is a cold outer layer that is physically, compositionally, and/or rheologically distinct from the warmer, next deepest part of the Earth’ upper mantle, known as the “asthenosphere.” Often, lithosphere is considered a primarily conductive boundary layer, and the asthenosphere is considered an adiabatically convective layer. Furthermore, most models lead to the conclusion that the older lithosphere is denser than the underlying asthenosphere, thus leading to a density instability that contributes to the formation of subduction zones. However, in detail, these ideas remain controversial, and full acceptance of any one model or definition remains elusive.

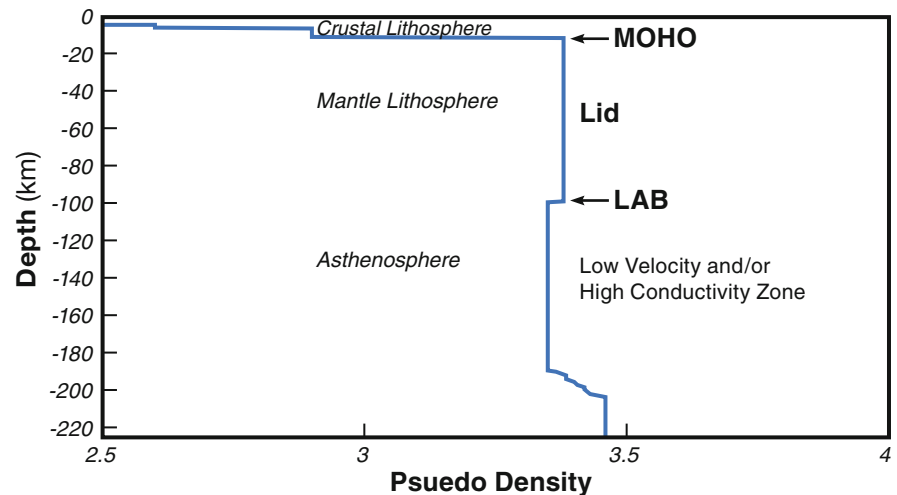
Ideas about the nature of the lithosphere are closely linked to understanding the nature of the lithosphere-asthenosphere boundary, or LAB (see Fig. 1). For example, as we shall see below, the LAB is often observed because the lithosphere displays higher upper mantle seismic velocities (known as the seismic “lid”), while the asthenosphere is associated with the seismic “low-velocity zone” (LVZ) that is often observed, particularly for shear waves.

The oceanic lithosphere shares many of the above concepts with its continental counterpart, but there are significant differences between the two. The term “lithospheric plate” is sometimes used to describe that part of the Earth that moves as a contiguous unit coherent plate. However, the vast majority of models for the oceanic lithosphere are intimately connected to the theory of seafloor spreading. Because the crust and upper mantle must undergo cooling as they spread away from the mid-ocean ridges and back-arc spreading centers, it is anticipated that temperature-dependent properties of the mantle and crust, including the rheology, physical properties (such as seismic velocity), and compositional changes, should deepen as the oceanic plates age. Thus, virtually, all models for the oceanic lithosphere include the constraint that it must thicken with age. Also, as we shall see, all modern models *allow* for a maximum thickness for older lithosphere, as cooling seems to slow or cease at ages approaching 100 Ma. This means that the uppermost young asthenosphere also must move along with the plate.

Constraints for thickening of the oceanic lithosphere, the nature of the LAB, and the difference between the lithosphere and asthenosphere are supplied by changes in seafloor depth and heat flow, by observations of a seafloor deformation, and by a variety of seismic and magnetotelluric observations. We look in detail at each of these below. The nature of the LAB remains controversial. Most models argue for a boundary between the completely solid lithosphere and partially melted asthenosphere. Others argue for a hydration boundary, with a lithosphere depleted in water, while the asthenosphere retains its water (e.g., Karato and Haemyeong 1998). Others argue for a boundary that includes variation in the crystal fabric of the mantle, or solid-solid transitions.

Lithosphere, Oceanic,

Fig. 1 Cartoon defining different levels of the oceanic upper mantle and crust, as used in this entry. Density is only approximate, based on the PREM model of Dziewonski and Anderson (1981). The lower boundary of the asthenosphere or low velocity zone is poorly defined for the oceans



Different authors have used the term lithosphere to be restricted to the uppermost mantle, while others include the crust. For the purposes of this entry, we shall include the crust and uppermost mantle as components of the lithosphere. The inclusion of the crust is critical, because the crust plays an important compositional role in the creation of the lithosphere at the ridges, and because the crust plays a major role in the hydration of the lithosphere as it spreads away from the axes, and dewatering of the plate as it is subducted. For this entry, we will exclude oceanic sediments, and anomalous regions such as islands, continental margins, and arcs. We include that part of the lithosphere that is generated on seafloor spreading centers and at back-arc spreading centers.

The Mantle Component of the Lithosphere

Despite the importance of the crust to lithospheric processes, it is the uppermost mantle that makes up the bulk of the observations constraining models for the lithosphere and LAB (see Fig. 1). In fact, a vast number of studies have been made of the nature of the mantle lithosphere, the mantle asthenosphere, and the boundary between them. An exhaustive review is beyond the scope of this entry, but we will try to discuss some of the key elements.

Directly or indirectly, the LAB for many models depends on temperature. While a slight change in the temperature gradient in the Earth can result in a low velocity zone in the mantle, many studies suggest that the LAB is a fairly sharp transition, and thus it is a result of a more fundamental change in the mantle. Many workers argue for a boundary between a totally solid lithosphere and an asthenosphere consisting of a small but significant percentage of melt (e.g., Anderson and Spetzler 1970; Kennett and Furumura 2015; Eilon and Abers 2017). The effects of such melt may be enhanced by its

concentration at the LAB, or by the geometry of melt layering.

However, partial melt models are not universally accepted (e.g., Karato and Park 2019). Others argue that the LAB may represent a hydration boundary (e.g., Hirth and Kohlstedt 1996; Karato and Haemyeong 1998). Other transitions have been proposed, including a change in the orientation of the anisotropy at the LAB, deformation-related grain size effects, and the presence of elastically accommodated grain boundary sliding (see Karato and Park 2019). Solid-state phase boundaries also may explain the transition.

Most of these mechanisms for the LAB are temperature sensitive, and this suggests that the depth of the LAB must change as the lithosphere ages and cools. As we shall see below, most models include such age-related changes. Many models also imply an upper limit on the thickness of the lithosphere. To the extent that these models converge, the thickness of the lithosphere (or depth to the LAB) appears to lie between 90 and 135 km.

Models Constrained by the Cooling of the Lithosphere as It Ages

Measurements of heat flow and seafloor depth are directly related to the temperatures in the ocean lithosphere, and to the changes in temperature as it ages. The ocean forms an isothermal boundary layer (near 0 °C) at the seafloor, and temperatures in the lithosphere are controlled by thermal conduction and convection of heat to the seafloor from the mantle. With the recognition of the theory of seafloor spreading it was predicted that the seafloor should contract (deepen) and heat flow should decrease as it ages. Early models focused on conductively cooled models. Davis and Lister (1974) considered a simple half-space cooling model and calculated the depth as a function of age. They show that conductive cooling of the lithosphere and resulting

contraction lead to an ocean depth anomaly (Δh) that is proportional to the square root of age and can be represented as,

$$\Delta h(t) = \frac{2}{\sqrt{\pi}} \alpha_e T_m \sqrt{\kappa t}, \quad (1)$$

where age is given by t , the thermal diffusivity of the “lithosphere” is given by κ , and the emplacement temperature of new crust and mantle ($t = 0$) is given by T_m . The effective thermal expansion coefficient, α_e , is adjusted for the replacement of rocks with lithosphere density (ρ_l), by water (ρ_w), or

$$\alpha_e = \left(\frac{\rho_l}{\rho_l - \rho_w} \right) \alpha_l. \quad (2)$$

In addition to increasing in depth, it was found that the conductive model predicts a decrease in measured heat flow (q) with

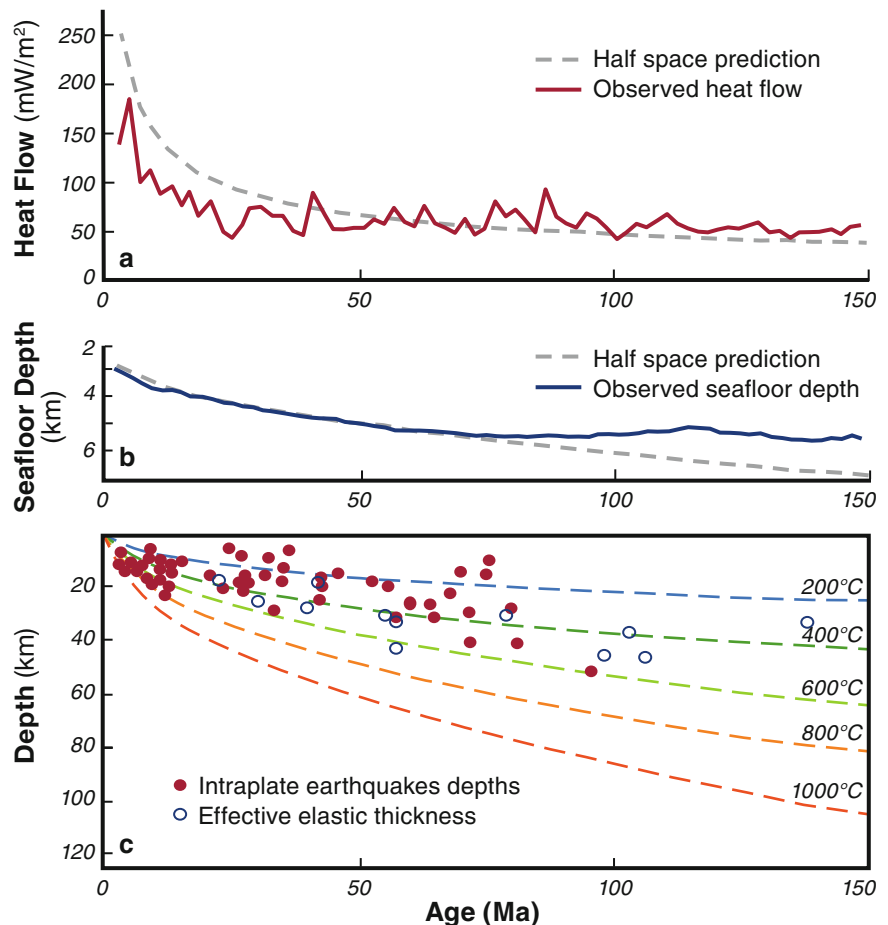
$$q(t) = k T_m \frac{1}{\sqrt{\kappa t}}, \quad (3)$$

where k is the thermal conductivity of the lithosphere. Davis and Lister (1974) and Parsons and Sclater (1977), and others, pointed out significant discrepancies between the above predictions and actual measurements made on the seafloor. The indication that seafloor depths were too deep and heat flow too low on the young seafloor was most prominent. It was suggested that hydrothermal circulation at the ridges essentially cooled the crust and upper mantle far more efficiently than conductive models, thus accounting for the observations. This was confirmed by the discovery of very high temperature venting (e.g., Corliss et al. 1979) that spectacularly confirmed the role of hydrothermal circulation in cooling the lithosphere, as well as accommodating chemical alteration of the oceanic crust (see Fig. 2a and b).

In addition to the anomalous depths and heat flow in younger plate ages, studies revealed that *older seafloor* exhibits just the opposite effects, with seafloor depths being shallower and heat flow higher than the half-space predictions (see Fig. 2a and b). These observations suggest that cooling of older (>70 Ma) lithosphere is greatly reduced or eliminated. That is, the average seafloor depth approaches a maximum, while the heat flow approaches a minimum above older

Lithosphere, Oceanic,

Fig. 2 Different observables used to constrain models for the oceanic lithosphere, primarily its mantle component. **(a)** Observed heat flow data (in red) averaged in two million year bins (modified from Stein and Stein (1992) compared to predictions for a conductively cooled half space (in gray-dashed line). Heat flow is too low and scattered for ages less about 40 Ma, indicating nonconductive (hydrothermal circulation) cooling must be taking place. **(b)** Observed seafloor depth (in blue) compared to predictions for a conductively cooled half space (in gray-dashed). Depths are too shallow beyond 80 Ma, indicating temperatures have stabilized and heat is being supplied from below. **(c)** Plot of earthquake depths and effective elastic thickness on top of expected isotherms for a half space model



lithosphere. A common explanation is that the lithosphere is a zone of conductive cooling, while the underlying asthenosphere is undergoing adiabatic convective heat flow, maintaining a limit to cooling, and thus yielding a maximum thickness to the lithosphere. Parsons and Sclater (1977) found that the model that best fit their large data set was a lithosphere-asthenosphere boundary that reached a depth of 125 km under older lithosphere, with a temperature of 1350 °C.

These early models have been greatly refined by better data sets and the inclusion of more constraints. In their widely cited paper, Stein and Stein (1992) found an improved fit for a larger data set, with an older lithosphere thickness of 95 km (± 15 km), and a weakly constrained basal temperature of 1475 °C (± 250 °C). Their model, labeled GDH1, has become a standard of comparison for more recent efforts to study the evolution of the lithosphere. Stein and Stein (1994) also considered in detail the role of hydrothermal circulation, both high temperature circulation at the spreading centers, and more diffuse and lower temperature circulation on seafloor that may persist to >60 Ma.

Afonso et al. (2007) further refined models for the lithosphere by including gravity and geoid height measurements, as well as consideration of seismic velocities as constraints on the LAB and lithosphere. In their model, they include the buoyancy effects of the crust, and the effects of solid phase transitions in the mantle, and using a basal temperature of 1300 °C to find a “thermal lithosphere” thickness of 105 km. Very important is their finding that the density of the lithosphere *may* not be greater than the underlying asthenosphere, contrary to accepted models.

Grose and Afonso (2013) consider the insulating effect of the crust, and radiative heat transport in their analysis of the temperature effects on lithosphere evolution. They find their models that include consideration of pressure-dependent thermal properties, hydrothermal circulation, and insulating oceanic crust achieve good fits to various observations. They find an asymptotic lithosphere thickness of 109 (± 6) km if a radiative contribution is small, and 123 km if it is significant. They also find significantly lower heat flow values for young ocean lithosphere, when compared to GDH1.

Korenaga and Korenaga (2016) argue that the presence of a relatively low density crust may add buoyancy to the lithosphere, not accounted for by many models. More recently, Richards et al. (2018) reexamine data after corrections for sediment and crustal thickness. They also include constraints from laboratory measurements on the temperature and pressure dependence of thermal parameters such as conductivity and thermal expansion coefficients. In their model, they predict the maximum lithosphere thickness of 135 km.

Models Constrained by Rheological Observations

One of the major concepts of the mantle lithosphere is that it behaves with a rheology that is distinct from that of the

underlying asthenosphere. Specifically, it is widely considered that the lithosphere can be defined by elastic or brittle behavior while the asthenosphere has a viscoelastic or ductile rheology. These differences have been exploited in a number of studies and give distinct values for the long-term thickness of the “lithosphere.”

It has long been known that vertical loads added to the Earth’s crust, such as volcanoes, mountain ranges, and glacial ice sheets, will cause the surface to subside. Similarly, removal of the load by erosion or melting ice will cause uplift. This leads to the concept of an isostatic compensation depth, at which the pressures exerted by the overlying loads are equal. However, if the lithosphere is strong and elastic, it can support and distribute that load over a region that is broader than that of the load applied. That is, there is less subsidence than would be normally expected. The resulting bending of the lithosphere under the load can be thought of as a response of an elastic plate overlying a viscoelastic asthenosphere. The wavelength and amplitude of bending are reflected in deviations in seafloor depth, and measurement of the seafloor can be used to determine an “effective plate thickness” (Watts 2001). This thickness is not necessarily the same as that of the lithosphere. The effective plate thickness as a function of age is shown in Fig. 2c (modified from Stein and Stein 1992). We see the equilibrium lithospheric thickness is far less than the estimates from overall seafloor depths and heat flow values discussed above, with maximum values reaching about 50 km, roughly corresponding to the 600 °C isotherm in the half-space cooling model.

Similarly, workers have considered the depth of “intra-plate” earthquakes (earthquakes occurring away from plate boundaries). The maximum depth of seismicity can be considered the boundary for brittle and nonbrittle behavior (e.g., Wiens and Stein 1983; Stein and Stein 1992), and that in turn could be considered a proxy for lithospheric thickness. As can be seen in Fig. 2C, the result yields a “lithospheric” thickness that also corresponds to the 600 °C isotherm for the half space cooling model, and a maximum thickness of about 60 km for a 100 Ma crust. However, Afonso and others (2007) argue that temperatures at the base of the seismogenic zone would be higher (700–800 °C) than those predicted by simpler models, but still much below shallower than any melting boundary.

Models Constrained by Seismic and Magnetotelluric Observations

The best information about the Earth’s interior structure comes from observations exploiting seismic wave propagation. This is true for the oceanic lithosphere and asthenosphere as well, although often in the context of high seismic velocity “lid” overlying a low velocity zone (LVZ). Thus, the characterization of the lithosphere is closely linked to the observations of the underlying LVZ, and the boundary

between them, the LAB (see Fig. 1). Over the years, the major constraints on the upper mantle structure have resulted from observations of surface waves. With the development of better and more ocean bottom seismometers (OBS), body wave phases converted at the LAB have provided new, more detailed constraints. In addition, certain diffuse seismic phases, typically labeled Po and So, are believed to be scattered P and S waves that are trapped in the lid, and they can be used to further characterize the lithosphere.

Surface Waves Studies

Seismic surface waves have long provided the firm evidence that a low velocity zone (LVZ) exists in the upper mantle between the Moho and the 400 km discontinuity (e.g., Forsyth 1975). Surface waves are primarily sensitive to shear wave velocities. Low velocity zones are frequently reported for shear waves and appear to be particularly well-defined under the oceans at a depth of 100–150 km. While not observed for all studies, the LVZ was present in models often enough to be included in for horizontally polarized shear waves (S_h) in the widely cited and used Preliminary Earth Reference Model (PREM, Dziewonski and Anderson 1981). Dziewonski and Anderson allow a negative velocity gradient with respect to depth below the Moho (Fig. 3). Vertically polarized shear waves (S_v) do not share this feature in PREM, requiring significant seismic anisotropy in the uppermost mantle. PREM includes this negative velocity gradient down to the Lehman discontinuity at 220 km, which is a

relatively sharp increase in velocity. Thus, the PREM does not include a LAB or separate lithosphere and asthenosphere.

In general, by virtue of their long wavelengths, surface wave models do not resolve details of the lithosphere-asthenosphere boundary (LAB) or of the lithosphere itself. The presence of the LVZ obviously implies a high velocity lid, however, the nature of the LAB remains controversial.

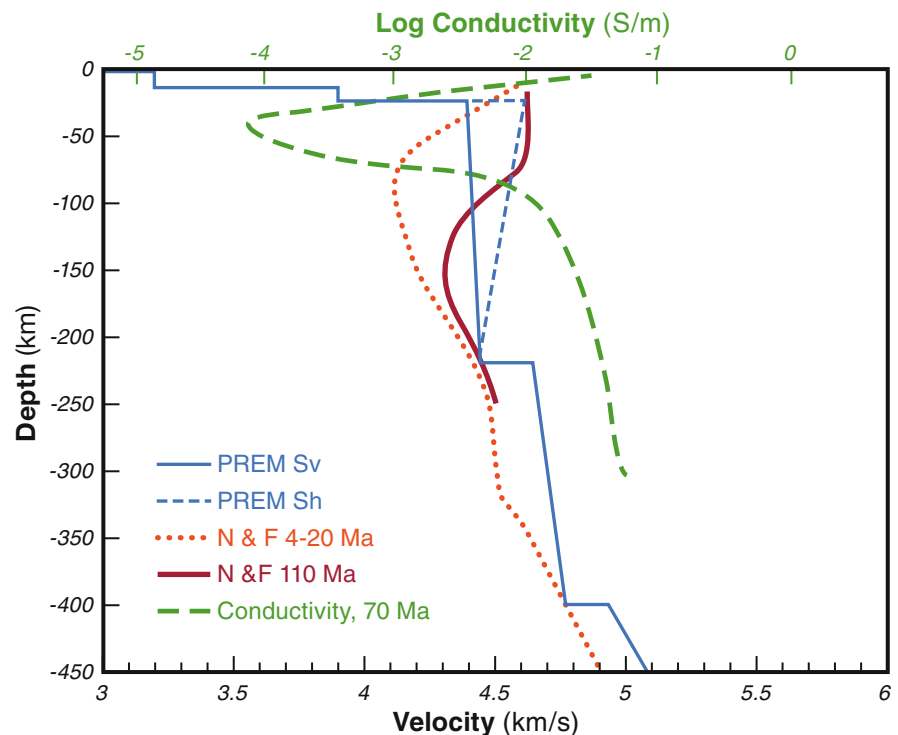
Several important features have been observed using surface waves. Perhaps most important has been the observation that the depth of the seismic LVZ increases with seafloor age (e.g., Forsyth 1975, 1977; Nishimura and Forsyth 1989) (see Figs. 3 and 4). This implies that, whatever the nature of the LAB, it must be at least partly controlled by temperature. The fact that it deepens with age, and that the lithosphere thickens with age, requires that the deeper parts of the older lithosphere initially were part of the young asthenosphere.

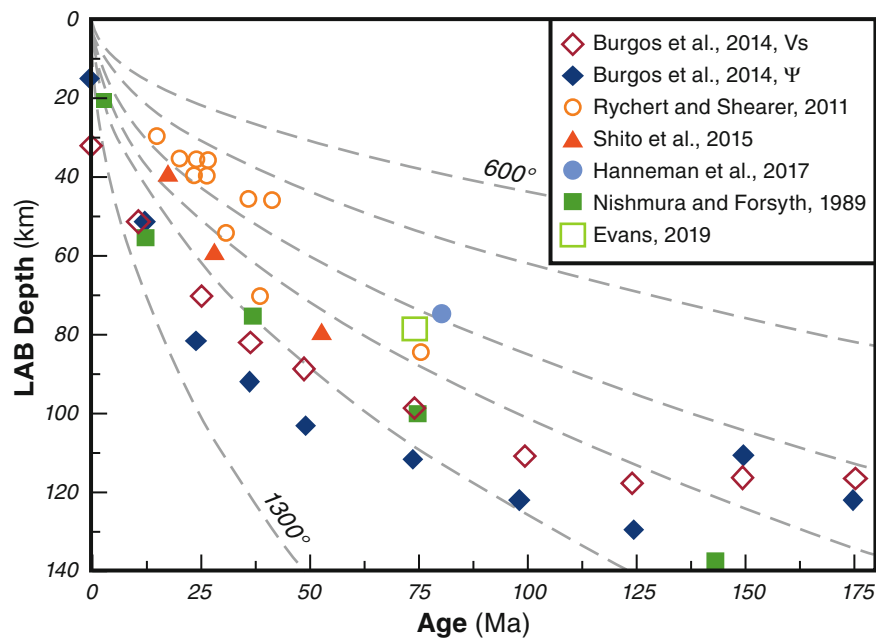
Burgos et al. (2014) used an inversion data set comprising some 300,000 different dispersion measurements for group and phase velocities for surface Rayleigh and Love waves. They found regional variations in structure of the upper mantle, including age-related thickening of the lithosphere. In order to quantify those changes, they chose four different “proxies” for the depth of the LAB. They chose the maximum negative velocity gradient for S_v waves (the vertically polarized shear wave velocities) as one of their proxies.

For another proxy, Burgos et al. (2014) used the azimuthal seismic anisotropy revealed by the shear wave inversions, and its relationship to present-day plate motion. Seismic

Lithosphere, Oceanic,

Fig. 3 Examples of shear wave velocities (km/s) and the log of electrical conductivity (S/m) for profiles extending through the crust and upper mantle. The Preliminary Reference Earth Model (PREM) of Dziewonski and Anderson (1981) is shown, for both vertically and horizontally polarized shear waves (S_v and S_h) are shown in blue. Results of Nishimura and Forsyth (1989) are shown in red for two different ages in the Pacific, and a conductivity profile for the NoMELT experiment is shown in Green (Evans et al. 2019)





Lithosphere, Oceanic, Fig. 4 Lithosphere-Asthenosphere Boundary (LAB) depth for a variety of seismic studies using different techniques. We also include one magnetotelluric experiment result. Regardless of the technique, all of the studies are consistent with the concept of a lithosphere that thickens with age because of progressive cooling of the upper oceanic mantle. Most allow models where cooling is reduced or ceases in lithosphere older than 100 Ma. Burgos et al. (2014) used surface waves, and a variety of proxies to represent the boundary. Included here are results using the maximum shear wave negative velocity gradient (for vertically polarized shear waves) (Vs). Another of their proxies was the

transition from azimuthal anisotropy (Ψ) (see text), which gives a slightly deeper LAB at older ages. Nishimura and Forsyth (1989) also use surface waves to define a LAB. Hannemann et al. (2017) used instrument receiver functions to determine LAB depth beneath the Atlantic. The Rychert and Shearer (2011) results were from shear body wave multiples (SS) reflecting from the LAB. We have only replotted their results from “normal” plate locations. Shito et al. (2015) modeled propagation of phases (Po and So) propagating in the lithosphere to constrain LAB depths. Evans et al. (2019) present an estimate of LAB depth from a magnetotelluric survey southeast of Hawaii

anisotropy both for compressional (P) and shear (S) waves has been widely observed in the oceans, particularly for the uppermost mantle (e.g., Raitt et al. 1971; Blackman and Kendall 2002; Lin et al. 2016). In the lithosphere, it is widely accepted that the anisotropy coincides with the preferred fabric of olivine in the mantle rocks. This preferred fabric is the result of strains related to plate motion at the time for formation; it is “frozen” in and preserves a record of plate motion during formation at the ridge axis. For the more fluid asthenosphere, they suggested that the anisotropy could change and reflect the present-day deformation, perhaps related to subsequent changes in present-day plate motion or to other smaller scale convection. Thus, they choose the depth where the anisotropy diverts away from present-day plate motion as the depth of the LAB.

Burgos et al. (2014) find that both of the above proxies, as well as others, reveal a deepening of the LAB with age from 0 to 150 Ma. However, it is important to note that different proxies give different depths, again illustrating the problem of defining what is meant by the term “lithosphere” (see Fig. 4). For example, Burgos found a mature lithospheric thickness of about 110–120 km for the proxies shown in Fig. 4, but he also found using the maximum radial anisotropy as a proxy led to a lithospheric thickness of 70–80 km.

In addition to seismic velocities, a number of studies have shown that the seismic LVZ roughly correlates with a zone of high seismic attenuation (or low Q) (e.g., Anderson and Hart 1978; Eilon and Abers 2017). The lithosphere or lid forms a zone of low attenuation or high Q. Hammond and Humphreys (2000) argue that attenuation measurements argue for a partial melt in the asthenosphere, and thus in their model the LAB is the melt boundary.

Studies Using Seismic Body Wave Phases

All of the above-discussed observations share in the fact that the depth resolution of resulting models is poor because of the long wavelengths required for deeply sampling surface waves. This characteristic tends to “smooth out” boundaries if they are present (see the excellent review of Karato and Park (2019) for more detail). Such smooth boundaries are consistent with changing varying variables such as pressure and temperature. For example, in the absence of any other type of transition seismic, velocity gradient will depend on the ratio of the pressure gradient to the temperature gradient. Higher pressure gradients will drive the seismic velocity gradient positive, while higher temperature gradients will lead to a negative velocity gradient. The critical temperature gradient is about 6.5 °C/km (Birch 1952), and with a correction for the

Earth's curvature, the critical gradient is about 8.5 °C/km. These are quite reasonable geotherms for the mantle.

In recent years, new advances in ocean instrumentation and data analysis techniques have improved our understanding of the oceanic upper mantle. New techniques allow studies of faint seismic body wave arrivals that may provide better vertical resolution than those provided by surface waves studies.

Many recent studies have made use of the so-called “instrument response functions” to improve the resolution for models for the lithosphere, asthenosphere, and LAB. Such studies exploit the body waves traveling nearly vertically to a receiving seismometer or array of seismometers. As the body waves intersect a boundary, conversions between P waves and vertically polarized shear waves (Sv) may occur. In general, only a fraction of the incident wave energy will be converted. The size of these converted phases depends on the sharpness and velocity changes across the boundary. The resulting phases are denoted P_s (where a small fraction of the incident P wave converts to a shear wave at a target interface) and S_p (S to P conversions). For the former, the arriving P wave will be followed by a low amplitude shear wave for stations that overlie the targeted boundary. In the case of S_p , the arriving shear wave will be preceded by a low-amplitude P wave that formed at the boundary.

Such converted phases can be difficult to discern from scattered energy, but with proper processing of the resulting seismograms, they can be detected and analyzed. For example, Kawakatsu et al. (2009) argue that such conversions observed in the western Pacific require a sharp LAB that cannot be explained by gradual temperature/pressure variations. They argue the data require a LAB to be less than 15 km thick and include a velocity reduction of 7–8%. The large change leads them to suggest that boundary's effect is enhanced by partial melt that occupies horizontal layers. Such a sharp boundary probably precludes simple changes in temperature gradient as the cause of the lithosphere-asthenosphere boundary.

Hannemann et al. (2017) made a careful study of potential P to S conversions for a broad-band seismic array deployed in the Atlantic over 79 Ma lithosphere. Using a variety of techniques, they attempted to identify P_s conversions for the mantle, using data for both single stations, and stacked data from multiple stations. They made a tentative identification of the LAB for single-station data. If correct, it suggests the LAB is 70–80 km deep (see Fig. 4), and also argues for a relatively sharp transition.

In addition to converted phases, a number of workers have examined precursors to shear wave arrivals such as those reflected from the core (ScS) or from the surface (SS) (e.g., Goes et al. 2013). Rychert and Shearer (2011) examine differences between SS phases reflecting from the Earth's surface and those that may reflect from the LAB. Those

differences will be subtle. However, they can be enhanced by stacking multiple seismograms from ray paths that reflect from the LAB at similar geographic locations and ages. They find that the LAB thickens with age as expected, but at depths corresponding to temperatures of 820–1020 °C (see Fig. 4). Rychert and Shearer (2011) allow for, but do not necessarily prove, the presence of a sharp boundary. They suggest the IAB may correspond to a hydration boundary, a melt boundary, or a permeability boundary. However, these temperatures are low for normal melting.

Other Geophysical Studies

Additional constraints about mantle lithology are derived from the response of the Earth to electromagnetic induction caused by long wavelength electromagnetic waves that are vertically incident from sources such as fluctuations in the Earth's magnetosphere or lightning strikes trapped by the ionosphere. In these studies, the asthenosphere correlates with a high conductivity zone (e.g., Evans et al. 2019) (see Fig. 3), while lithosphere would be represented by relatively high conductivities.

In addition to the importance of seismic surface waves and body waves, observations have been made of unusual compressional (Po) and shear (So) phases that seem to be trapped in the lithosphere. These are typically characterized by a long, diffuse coda that follows the initial P and S waves propagating along the Moho. They tend to be of very high frequency compared to regular body wave phases. (Note there is some confusion of notation. Some workers specify these waves a Po and So, while some use notation as Pn and Sn. We use the former notation, and restrict the use of Pn and Sn for the first arriving mantle phases, that tend to be low amplitude.)

Po and So are most prominent for over long travel paths, indicating that reverberation in a high Q medium, such as the lithosphere, is important. Sereno and Orcutt (1987) point out that reverberation in the water and sediment columns may also contribute to the complexity of the coda. However, Kennett and Furumura (2013) and Shito et al. (2013) argue for a propagation path best represented by scattering from three-dimensional heterogeneities. Kennett and Furumura (2015) model both Po and So phases using a stochastic model for the scale of heterogeneities. They find that scattering from “quasi-laminate” features with a horizontal characteristic scale of 10 km, and vertical characteristic scale of 0.5 km, provides the best fit.

More work is needed in the study of these phases, but they do provide constraints of a lithospheric thickness. Using synthetic seismogram modeling, Shito et al. (2015) found that these trapped phases generated by earthquakes in the western Pacific demonstrate a systematic thickening (to 80 km.) of the lithosphere with age to at least 50 million years (see Fig. 4).

The Crustal Component of the Oceanic Lithosphere

The top few kilometers of the oceanic lithosphere comprise the ocean crust. To a first approximation, the crust is viewed as the mafic member of the lithosphere and is generated by processes at the mid-ocean ridges (or at back-arc spreading centers). Three lines of evidence dominate our knowledge of the crust. These include direct sampling of seafloor rocks where they are exposed, seismic experiments that reveal the seismic velocity structure of the crust, and the study of crustal analogues exposed above sea level. These studies have converged on the so-called Penrose or “ophiolite” model (Fig. 5). While this model is highly successful in its predictions, the research in the last 20 years has led to the recognition, at least for some crust generated at ridges with slow to ultra-slow spreading rates, that the ophiolite models must be revised to include newly discovered crustal features.

Samples of Crustal Rocks

Oceanic crustal rocks are generally recovered from near the mid-ocean ridges before sediments have a chance to accumulate on the aging seafloor. The vast majority of these rocks are mafic basalts, however, other rocks such as gabbros and serpentinites are sometimes exposed at transform faults or on rifted mid-ocean ridges.

From the Challenger expedition in the 1870s until the 1980s, most seafloor samples were collected by dredging, where a chain bag, with a steel jaw, is dragged behind a drifting or slowly moving ship. The problem with dredging, in addition to the uncertainty of success, is that because the

ship was towing the dredge bag, it was never certain where the dredge was when it collected the sample.

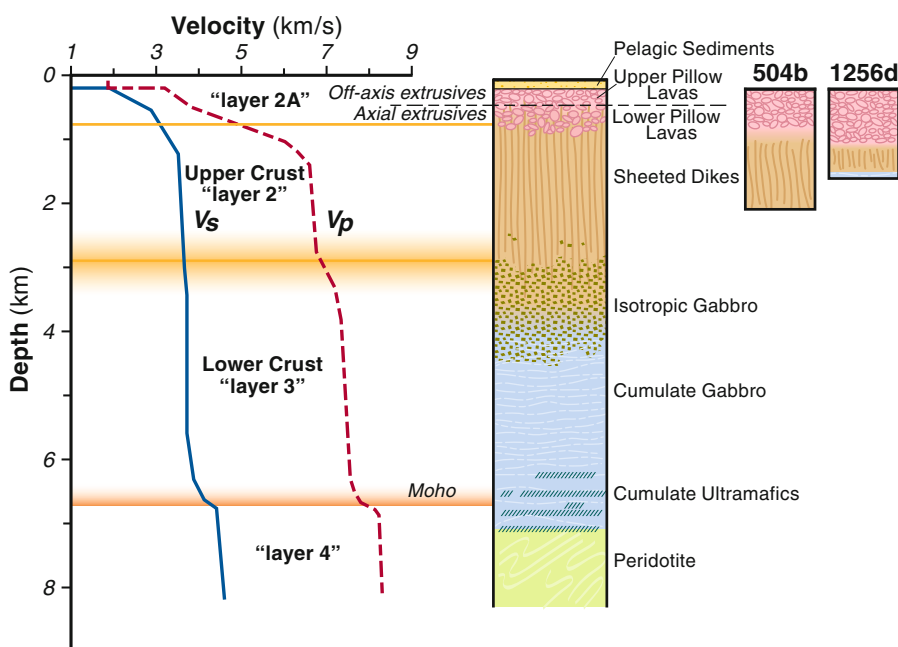
Recently, more precisely located samples have become possible through the use of well-navigated submersibles and remotely operated vehicles (ROVs). The location of these vehicles, and any samples they recover, can be determined very precisely through acoustic techniques and the use of GPS in the ship above. In addition, a relatively new technique, rock coring or “wax coring,” has gained usage. With this tool, a weight is lowered rapidly from the ship, nearly straight down. The turnaround time is rapid compared to dredging or submersibles, so numerous samples can be collected. Because the corer is dropped straight down, rather than dragged, the location of the samples can be determined from the ship’s GPS position. The samples are small and usually volcanic glasses from the surface of the seafloor basalts, and with newer analytical tools and techniques, they can be analyzed for the detailed chemistry of the glass.

Regardless of the sampling method, all of the samples are examined and described on the ship. When they are returned shore, they undergo extensive and higher precision analysis using a variety of techniques. Even from the earliest studies, basalts are the most commonly observed rock type, and they result from partial melting of ultramafic peridotites in the mantle.

For rocks recovered that originated on the mid-ocean ridges, the vast majority are so-called MORBs or mid-ocean ridge basalts. MORBs are characterized by a composition that includes depletion in the incompatible elements (e.g., K, Ti, P, Ba, Pb, Th, U, Zr, Rb, and Sr) and light rare earth elements (atomic numbers between 57 and 71). These observations,

Lithosphere, Oceanic,

Fig. 5 Comparison of seismic velocities with an idealized ophiolite, and with drilling results for two Pacific ODP (504b) and IODP (1256d) drill holes that successfully penetrated to midcrustal levels (modified from McClain 2003). The seismic structure is for a specific profile on 16 Ma crust generated on the fast spreading southern East Pacific Rise. The midcrustal inflection is used here as the boundary between layer 2 and layer 3 for this particular crustal model. The inflection could be caused by the lithological boundary between sheeted dikes and gabbros as shown here, and/or by changes in porosity or metamorphic grade. From velocity alone, we cannot tell. However, all three causes could occur at the same depth



coupled with a relatively low $^{87}\text{Sr}/^{86}\text{Sr}$ ratio, indicate that the mid-ocean ridge basalts originate from partial melting of a depleted upper mantle beneath the ridges. An additional constraining factor from seafloor basalts was that the detailed composition required that magma undergoes partial fractionation at shallow (i.e., lower crustal) pressures. This led to models that included a magma chamber in the lower crust.

With the improved analytical instrumentation and techniques, the compositions of MORBs have been further parsed into N-MORB (normal-MORB) and E-MORB (enriched-MORB), and intermediate compositions (transitional or T-MORB). These observations generally are interpreted to indicate they originated in a heterogeneous, but still depleted, mantle source.

Ophiolites, Ocean Drilling, and the Ocean Crust

Ophiolite sequences have been studied for over 100 years and have been the source of substantial controversy. Complete ophiolites (see Fig. 5) include a coherent stratigraphy extending from submarine lava flows into sheeted dikes, representing the feeders for the lava flows. Beneath the sheeted dikes are gabbros, often divided between isotropic gabbros (no preferred fabric) at the top, and layered (cumulate) gabbros at the base. Beneath the gabbros are ultramafic peridotites, varying between harzburgites, lherzolites, and dunites. In general, the peridotites are serpentinized to a greater or lesser extent, but it is unclear whether the serpentinization occurs during the original formation of the crust at the mid-ocean ridges or during emplacement and uplift onto land.

Early on, ophiolites were thought to originate as oceanic crust and uppermost mantle, and to have been thrust up onto the continents by convergent processes (see, Moores 1982). Substantial debate centered on whether particular ophiolites originated in arc or back-arc settings, or whether some may have originated on midocean ridges. The former argument is primarily based on the detailed geochemistry of the ophiolite lavas, which seem to mimic those dredged from arcs. In contrast, some ophiolites are stratigraphically overlain by pelagic sediments, with no evidence for the volcanoclastic sediments observed in arc settings.

Regardless of the origin of any particular ophiolite, it is clear that ophiolites provide an important analog for ocean crust generated on midocean ridges. The presence of sheeted dikes to deliver magma from a crustal magma chamber, and the presence of gabbros originating from the freezing of the magma chamber, leads to a model that is consistent with the petrologic models derived from seafloor sampling as described above.

The Ocean Drilling Project, its predecessor Deep Sea Drilling Project, and successor the Integrated Ocean Drilling Project provide support for the “ophiolite” model. In particular, drilling results in the Costa Rican rift (Hole 504b) and off

southern Mexico (Hole 1257) have penetrated through the seafloor lavas and into the sheeted dikes. For the latter site, gabbros intercalated with sheeted dikes have been reached at the base of the sheeted dike column.

Seismic Studies of the Ocean Crust

Except for studies of subareal ophiolites, rare exposures on the seafloor, and a few deep drilling efforts, seismic studies provide the only information about the oceanic crust below the seafloor. They provide information on structure and seismic velocities that lead to inferences about the petrology of the upper mantle and crust.

After World War II, marine institutions began to exploit the seismic refraction technique to explore the Earth beneath the sea. These experiments usually yielded observations of compressional or P-wave seismic velocities, although converted shear waves were sometimes observed, and shear velocity structures could be inferred. In these early studies, theoretical and computational tools necessarily restricted models to laterally homogeneous structures, and a consistent four-layer model for the crust and overlying sediments was developed (e.g., Shor et al. 1969). The first layer was that of sediments, with velocities averaging about 2 km/s. Beneath that was “Layer 2,” and because the seafloor was known to reveal lavas in unsedimented areas, it was expected that layer 2 was the “volcanic layer.” The thickest part of the oceanic crust is “Layer 3,” with velocities of about 6.8–7.2 km/s and thicknesses on the order of 2–4 km. The lithology of layer 3, also known as the “oceanic layer,” has been controversial, with some models preferring a gabbroic composition, and others a serpentinite composition. The former model has gained general acceptance, partly because of the presence of gabbros in ophiolites (e.g., Christensen and Salisbury 1975; McClain 2003). Underlying layer 3 is layer 4, with compressional wave velocities of about 8 km/s, and corresponds with the upper mantle. The lithology of the mantle is generally taken to be peridotite, consistent with the above-mentioned models for partial melting of the mantle leading to the rise and formation of a mafic crust. Thus, the crust comprises layers 2 and 3, and the boundary between the crust and mantle, layers 3 and 4, is known as the Moho or Mohorovicic discontinuity.

With better experiments, the original four-layered model was parsed into more finely divided structures. The earliest refinement was the observation of “Layer 2A.” This was the shallowest and lowest velocity portion of the upper crust, and the low seismic velocities were attributed to the high porosities of basalt flows. This interpretation is further supported by the observation that the velocities of layer 2A tend to increase to normal layer 2 values. That leads models where layer 2A thins over a few million years. This was interpreted as infilling of pores by metasomatic precipitation of minerals as the crust ages (e.g., Houtz and Ewing 1976). Such an explanation

requires fluid circulation and chemical alteration in the crust that must continue over millions of years after the crust forms.

In the 1970s, the development of new theoretical and computation tools allowed workers to move beyond the simple layered models to those that continuously vary, but usually vertically stratified models. As models grew more sophisticated, it became possible to correlate them with ophiolite structures, using both velocities and thicknesses (e.g., Christensen and Salisbury 1975; McClain 2003), and the so-called Penrose model or “ophiolite model” for the oceanic lithosphere has become the accepted (Fig. 5). For the crust, this led to models where Layer 2 was divided into an upper zone with relatively low velocities (layer 2A) discussed above, underlain by a high velocity gradient extending to layer 3. The high-gradient zone, sometimes labeled layer 2B, was attributed to a decrease in crustal porosity and is correlated with the transition to sheeted dikes of the ophiolite model.

At mid-crustal levels, some experiments revealed a slight inflection in seismic velocities between the lower and upper crust (layer 2 and layer 3), while others did not (see Fig. 2). The origin of this slight transition was (and is) attributed either to a metamorphic boundary between lower and higher grade metasomatic changes in lithology, or to a transition from diabase dikes to the gabbros in the lower crust. Finally, the Moho is more gradational in these refined models, and the mantle velocities are well correlated with those expected for peridotites.

The Midocean Ridges

The oceanic crust is generated by a complex interaction of magmatic, tectonic, and hydrothermal processes that take place at the midocean ridges where seafloor spreading occurs. The end result, at least on intermediate- to fast-spreading ridges, should be an ophiolite-like structure. The majority of seismic, petrological, and observational experiments over the last 2–3 decades have focused on the tectonic plate boundaries, with particular attention on the ridges. Interpretation of the data from these studies required models that allowed for two- or three-dimensional structure, rather than the laterally homogenous models from earlier. This required a breakthrough in seismic studies. Early ray-tracing techniques have given way to more robust tomographic inversion. In addition, the development of longer seismic hydrophone streamers and new processing and imaging algorithms now permits detection of subbasement (crustal and upper mantle) reflections using multichannel reflection experiments.

One of the critical controversies about the structures beneath midocean ridges was the presence of a seismically resolvable axial magma chamber (AMC). As stated above, petrologic data seemed to require some sort of magma chamber in the lower crust, and the ophiolite model includes a thick gabbro layer. However, models resulting from seismic

refraction surveys were equivocal, with some revealing crustal AMC (e.g., Orcutt et al. 1975). McClain et al. (1985) were able to provide convincing evidence for a narrow magma chamber on the East Pacific Rise at 13° north. The seismic velocity required for the chamber was 4.5 km/s, suggesting a partial melt rather than a pure magma. Further evidence was that the resolved AMC was found in an extensive tomographic experiment on the East Pacific Rise at 9°30'N (Toomey et al. 1990).

For fast-spreading ridges, the question of magma chamber existence was largely resolved when a large multichannel experiment was conducted along the northern East Pacific Rise (e.g., Detrick et al. 1987). Using the long hydrophone streamers, researchers were able to resolve reflections from the base of Layer 2A, as well as the Moho. However, most striking was the presence of a reflection at midcrustal levels along most of the length of the ridge. Analysis of the reflection amplitude and the change in the sign of the reflection require that the reflections result from a sudden decrease in seismic velocity, and the explanation for this velocity inversion was the presence of an axial magma chamber. Furthermore, the large amplitudes of the reflections require a rather low velocity compared to models obtained from refraction experiments. This suggests that reflections result from pure or nearly pure melt sills. The discrepancy results because refraction experiments do not resolve the sills, but instead image an average structure made up of the sills and partial melt zones. Early models have suggested these sills form at the top of the magma chamber, and may be responsible for the isotropic gabbros observed in some ophiolites. The larger partial melt zones would form the cumulate gabbros that make up the bulk of the lower crust.

Subsequent to the initial experiment of Detrick et al. (1987), similar magma reflections have been imaged on other segments of the fast or ultrafast spreading East Pacific Rise (e.g., Singh et al. 2006), as well as ridges with intermediate spreading rates (e.g., Canales et al. 2009). Most studies suggest that the crustal AMC is generally narrow.

Hydrothermal Activity

One of the most spectacular discoveries in the annals of earth science was the observation of vigorous hydrothermal circulation at spreading centers (e.g., Corliss et al. 1979). While the presence of hydrothermal circulation was not surprising, the exceptionally high temperatures, the deposition of hydrothermal minerals, and a vigorous chemosynthetic biological community were a major breakthrough in the study of midocean ridges. It is clear that the chemical, biological, and geological contribution of these vents is felt throughout the world's oceans. Thus far, over 200 hydrothermal vent fields have been discovered, and as many as 1000 may lie on the world's midocean ridges (Baker and German 2004).

Early studies suggested that a substantial portion of the Earth's heat was lost along the high temperature hydrothermal circulation on midocean ridges. More recent analyses indicate that more heat is lost through lower temperature circulation in older crust. However, it is clear that near-axis circulation is a major component of the crust forming process, and their discovery has led to a new appreciation for the interdisciplinary nature of midocean ridge science.

High temperature venting is observed on all ridges, and at all spreading rates. Such systems require a heat source and a permeability structure that permits a focused flow of fluids to the surface. For fast-spreading ridges, heat sources are plentiful, with the presence of axial magma chambers and frequent eruptive events. Indeed, the narrowness of the axial magma chamber requires that circulation penetrates at least to the top of the chamber and/or into the gabbros at its flanks (e.g., Lister 1983). The geometry of the down-going limb of the seawater into the crust remains uncertain. It may be that it penetrates along-axis, perhaps along faults or fissures, or it may penetrate from the sides, also along small normal faults.

For slower spreading ridges, the heat source is unclear. Certainly volcanic eruptions do occur, so a heat source is present at least some of the time. However, it may be that water also can penetrate to the deeper hot, but not molten, mantle rocks. The large normal faults often present on slow spreading ridges may accommodate this deeper penetration.

Along-Axis Variations in Ridge Processes and Slow-Spreading Ridges

It has long been known that midocean ridges display a systematic variation in cross-sectional topography, varying from rifted ridges, with rift valleys of 1–2 km deep, and nonrifted ridges that display topographic highs. In general, rifting is associated with slower-spreading ridges, while nonrifted ridges exhibit rapid seafloor spreading. There are a number of exceptions to the rule. For the Reykjanes Ridge adjacent to the Iceland Hot Spot, the slow-spreading ridge is unrifted, and on the Galapagos Ridge the ridge segments closest to the hot spot also are unrifted. For ridges spreading at intermediate rates, both rifted and unrifted ridge segments are observed (e.g., the Gorda and southern Juan de Fuca Ridges) and are often separated by only a transform fault.

These observations suggest that the presence or absence of a rift valley is more likely controlled by magma supply. On slow spreading ridges, where the isotherms are narrow, less melt may reach the surface unless excess temperatures and magma are supplied by a nearby mantle plume. For intermediate spreading ridges, local variations in magma supply, perhaps along-axis flow, may influence the formation of a rift valley.

With the advent of multibeam bathymetric systems, it was found that ridges are segmented on a variety of scales, with offsets between segments also varying in character and

length. Macdonald and Fox (1988) described a correlation between segment depths and cross-sectional geometry, and they suggested that these variations were linked to focused magma supply at segment centers and subsequent magma flow along axis to the segment ends. It is logical to assume that the magma supply would be tied to the resulting crustal structure and thickness, with thicker crust at the centers. The seismic reflection results of Detrick et al. (1987) seemed to support that linkage. However, it appears very little correlation or variation in thickness for spreading rates greater than 30 mm/year, and it appears that segmentation does not have a major influence on crustal thickness. For ridges that spread slowly, or very slowly, crustal thicknesses are far more variable.

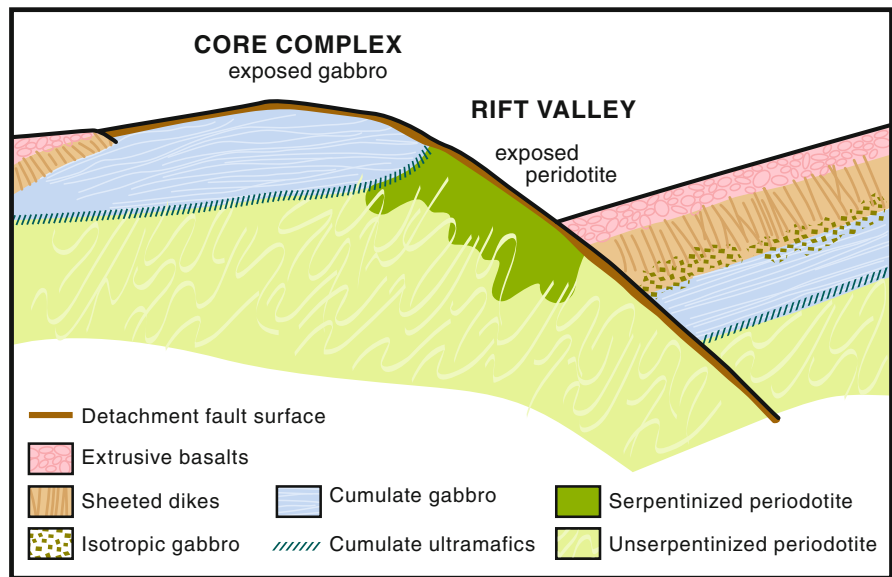
With the improvement of techniques, midocean ridge researchers have been able to overcome the complexities caused by topography on rifted ridges. For such ridges, typically spreading at slow rates, crustal thickness is closely tied to location along a segment, and it is likely that magma and mafic crust production are enhanced at segment centers. Tolstoy et al. (1993) correlated a thick crust in the shallow center of a rifted segment on the midAtlantic with gravity “bulls eyes,” or negative values centered at its center. The crust is thin at the segment ends, where the ridge intersects a transform.

The observation is that thin crust exists at the intersections of rifted ridges and transforms, where we expect the lowest mantle temperatures, and a suppressed magma supply (e.g., Cannat 1996). The discovery of oceanic core complexes, with exposures of gabbro and/or peridotite on the seafloor, provided observational support models for “new” crustal processes at ridges with slow to ultra-slow spreading centers (Cann et al. 1997) (see Fig. 6). It was recognized that the ophiolite model may not suffice for such locations. Instead, one has to incorporate substantial rifting that dissects the crust, and exposes the lower crust and upper mantle. This rifting leads to a complex laterally heterogeneous structure in the oceanic crust and upper mantle (e.g., Canales et al. 2008).

It appears that magma intrusions are not continuous beneath oceanic core complexes, and gabbros may not form a continuous layer as it does for the ophiolite model. During nonmagmatic rifting phases, we would argue that the mafic crustal thickness could be zero; i.e., the mantle peridotites exist at the seafloor. However, depending on the depth of serpentinization in the peridotites, with the consequent lowering of velocities, the seismic crust (velocities less than or equal to 7 km/s) may have a nonzero thickness (Fig. 6). The discovery of spectacular, serpentine-hosted hydrothermal circulation, at least one core complex, confirms the importance of rifting and serpentinization on the slow and ultra-slow spreading ridges (Kelley et al. 2001). It is clear that models for the oceanic crust must be refined to allow for slow to ultraslow spreading rates.

Lithosphere, Oceanic,

Fig. 6 Cartoon of rift-dominated processes on slow or ultraslow spreading ridges. Substantial normal faulting along a detachment fault exposes the deeper gabbros and peridotites, as volcanic/magmatic processes cannot “keep up” with the rifting. Rock units are the same as those in Fig. 5. Serpentinization of exposed or shallow peridotites will slow the seismic velocity, perhaps to crustal values. Thus, even where there are no crustal rocks, a thin “seismic crust” may be observed



Summary

The oceanic lithosphere forms the relatively cool and strong surface layer under the oceans, and it has been the subject of numerous studies and speculation over many decades. It includes two components, the crust and uppermost mantle. However, a strict description and characterization of the lithosphere remains elusive and depends on the types of observations that are used to constrain models.

Higher seismic velocities, low seismic attenuation, and low electrical conductivity characterize the mantle component of the lithosphere. It is underlain by an asthenosphere that is characterized by lower seismic velocities (the seismic low velocity zone), higher seismic attenuation (low Q), and higher electrical conductivity. The boundary between lithosphere and the asthenosphere deepens with age, i.e., the oceanic lithosphere thickens with age, a product of cooling of the plate as it spreads away from the ocean spreading center. That cooling continues for some distance but appears to decrease or cease when the lithosphere has aged on the order of 100 million years, after it has reached a thickness of 80–140 km. These maxima suggest that temperatures at the base of the older plate are maintained by convection in the asthenosphere.

The oceanic crust represents the mafic component of the lithosphere and is generated by a complex interaction of magmatic/volcanic processes, tectonic processes (seafloor spreading), and hydrothermal processes. The “competition” between magmatic processes that build the crust and rifting processes that spread the ridges is age dependent. At intermediate to fast spreading ridges, the crust is best described by the so-called “Penrose” model with a consistent igneous stratigraphy (see Fig. 5). For slow to ultraslow spreading ridges, the

crust is laterally heterogeneous as rifting brings deeper layers of the crust to or near to the surface.

Cross-References

- [Crustal Reflectivity \(Oceanic\) and Magma Chamber](#)
- [Isostasy](#)
- [Lithosphere, Continental](#)
- [Seismic Waves, Scattering](#)

Bibliography

- Afonso JC, Ranalli G, Fernandez M (2007) Density structure and buoyancy of the oceanic lithosphere revisited. *Geophys Res Lett* 34: L10302. <https://doi.org/10.1029/2007GL029515L1>
- Anderson DL, Hart RS (1978) Q of the Earth. *J Geophys Res* 83:5869–5882
- Anderson DL, Spetzler H (1970) Partial melting and the low-velocity zone. *Phys Earth Planet Inter* 4:62–64
- Baker ET, German CR (2004) On the global distribution of hydrothermal vent fields. In: German CR, Lin J, Parson LM (eds) *Mid-ocean ridges: hydrothermal interactions between the lithosphere and oceans*. Geophysical monograph series, vol 148. American Geophysical Union, Washington, DC, pp 245–266
- Barrell J (1914) The strength of the Earth’s crust. *J Geol* 22:425–433
- Birch F (1952) Elasticity and constitution of the Earth’s interior. *J Geophys Res* 57:227–286
- Blackman DK, Kendall JM (2002) Seismic anisotropy in the upper mantle 2. Predictions for current plate boundary flow models. *Geochem Geophys Geosyst* 3:8602
- Burgos G, Montagner JP, Beucler E, Capdeville Y, Mocquet A, Drilleau M (2014) Oceanic lithosphere/asthenosphere boundary from surface wave dispersion data. *J Geophys Res Solid Earth* 119:1079–1093. <https://doi.org/10.1002/2013JB010528>

- Canales JP, Tucholke BE, Xu M, Collins JA, DuBois DL (2008) Seismic evidence for large-scale compositional heterogeneity of oceanic core complexes. *Geochem Geophys Geosyst* 9:1–22. <https://doi.org/10.1029/2008GC002009>
- Canales JP, Nedimovic MR, Kent GM, Carbotte SM, Detrick RS (2009) Seismic reflection images of a near-axis melt sill within the lower crust at the Juan de Fuca Ridge. *Nature* 460:89–94. <https://doi.org/10.1038/nature08095>
- Cann JR, Blackman DK, Smith DK, McAllister E, Janssen B, Mello S, Avgerinos E, Pascoe AR, Escartin J (1997) Corrugated slip surfaces formed at North Atlantic ridge-transform intersections. *Nature* 385:329–332
- Cannat M (1996) How thick is the magmatic crust at slow spreading oceanic ridges? *J Geophys Res* 101:2847–2857
- Christensen NI, Salisbury SH (1975) Structure and constitution of the lower oceanic crust. *Rev Geophys Space Phys* 13:57–86
- Corliss JB, Dymond J, Gordon LI, Edmond JM, von Herzen RP, P R, Ballard RD, Green KL, Williams D, Brainbridge AL, Crane K, van Andel TH (1979) Submarine thermal springs on the Galapagos Rift. *Science* 203:1073–1083
- Davis EE, Lister CRB (1974) Fundamentals of ridge crest topography. *Earth Planet Sci Lett* 21:405–413
- Detrick RS, Buhl P, Vera E, Mutter J, Orcutt J, Madsen J, Brocher T (1987) Multi-channel imaging of a crustal magma chamber. *Nature* 326:35–41
- Dziewonski AM, Anderson DL (1981) Preliminary reference Earth model. *Phys Earth Planet Inter* 25:297–356
- Eilon ZC, Abers GA (2017) High seismic attenuation at a mid-ocean ridge reveals the distribution of deep melt. *Sci Adv* 3:e1602829. <https://doi.org/10.1126/sciadv.1602829>
- Evans RL, Sarafian E, Sarafian AR (2019) The evolution of the oceanic lithosphere: an electromagnetic perspective. In: Yuan H, Romanowicz B (eds) *Lithospheric discontinuities*. Geophysical monograph, vol 239, 1st edn. American Geophysical Union, Washington, DC
- Forsyth DW (1975) The early structural evolution and anisotropy of the oceanic upper mantle. *Geophys J R Astron Soc* 43:103–162
- Forsyth DW (1977) The evolution of the upper mantle beneath the mid-ocean ridges. *Tectonophysics* 38:89–118
- Goes S, Eakin CM, Ritsema J (2013) Lithospheric cooling trends and deviations in oceanic PP-P and SS-S differential traveltimes. *J Geophys Res Solid Earth* 118:996–1007. <https://doi.org/10.1002/jgrb.50092>
- Große CJ, Afonso JC (2013) Comprehensive plate models for the thermal evolution of oceanic lithosphere. *Geochem Geophys Geosyst* 14:3751–3778. <https://doi.org/10.1002/ggge.20232>
- Hammond WC, Humphreys ED (2000) Upper mantle seismic wave attenuation: effects of realistic partial melt distribution. *J Geophys Res-Solid Earth* 105:10987–10999. <https://doi.org/10.1002/2000JB900042>
- Hannemann K, Krüger F, Dahm T, Lange D (2017) Structure of the oceanic lithosphere and upper mantle north of the Gloria Fault in the eastern mid-Atlantic by receiver function analysis. *J Geophys Res Solid Earth* 122:7927–7950. <https://doi.org/10.1002/2016JB013582>
- Hirth G, Kohlstedt DL (1996) Water in the oceanic upper mantle: implications for rheology, melt extraction, and the evolution of the lithosphere. *Earth Planet Sci Lett* 144:93–108
- Houtz RE, Ewing J (1976) Upper crustal structure as a function of plate age. *J Geophys Res* 81:2490–2498
- Kanamori H, Press F (1970) How thick is the lithosphere? *Nature* 226:330–331
- Karato S, Haemyeong J (1998) Water, partial melting and the origin of the seismic low velocity and high attenuation zone in the upper mantle. *Earth Planet Sci Lett* 157:193–207. [https://doi.org/10.1016/S0012-821X\(98\)00034](https://doi.org/10.1016/S0012-821X(98)00034)
- Karato S, Park J (2019) Chapter 1, On the origin of upper mantle seismic discontinuities. In: Yuan H, Romanowicz B (eds) *Lithospheric discontinuities*. Geophysical monograph, vol 239, 1st edn. American Geophysical Union, Washington, DC
- Kawakatsu H, Kumar P, Takei Y, Shinohara M, Kanazawa T, Araki E, Suyehiro K (2009) Seismic evidence for sharp lithosphere-asthenosphere boundaries of oceanic plates. *Science* 324(5926):499–502. <https://doi.org/10.1126/science.1169499>
- Kelley DS, Karson JA, Blackman DK, Fruh-Green GL, Butterfield DA, Liley MD, Olson EJ, Schrenk MO, Roe KK, Lebon GT, Rivizzigno P, AT3-60 Shipboard Party (2001) An off-axis hydrothermal vent field discovered near the Mid-Atlantic Ridge at 30°N. *Nature* 412:145–149
- Kennett BLN, Furumura T (2013) High-frequency Po/So guided waves in the oceanic lithosphere: I – long distance propagation. *Geophys J Int* 195:1862–1877
- Kennett BLN, Furumura T (2015) Toward the reconciliation of seismological and petrological perspectives on oceanic lithosphere heterogeneity. *Geochem Geophys Geosyst* 16:3129–3141. <https://doi.org/10.1002/2015GC006017>
- Korenaga T, Korenaga J (2016) Evolution of young oceanic lithosphere and the meaning of seafloor subsidence. *J Geophys Res Solid Earth* 121:6315–6332. <https://doi.org/10.1002/2016B013395>
- Lin P-YP, Gaherty JB, Jin G, Collins JA, Lizarralde D, Evans RL, Hirth G (2016) High-resolution seismic constraints on flow dynamics in the oceanic asthenosphere. *Nature* 535:538–541
- Lister CRB (1983) On the intermittency and crystallization mechanisms of sub-seafloor magma chambers. *Geophys J R Astron Soc* 73:351–365
- Macdonald KC, Fox PJ (1988) The axial summit graben and cross-sectional shape of the East Pacific Rise as indicators of axial magma chambers and recent volcanic eruptions. *Earth Planet Sci Lett* 88:119–131
- McClain JS (2003) Ophiolites and the interpretation of marine geophysical data: how well does the ophiolite model work for the Pacific Ocean crust? *Geol Soc Am, Special Paper* 373:173–185
- McClain JS, Orcutt JA, Burnett M (1985) The East Pacific Rise in cross section: a seismic model. *J Geophys Res* 90:8627–8629
- Moore EM (1982) Origin and emplacement of ophiolites. *Rev Geophys Space Phys* 20:735–760
- Nishimura CE, Forsyth DW (1989) The anisotropic structure of the upper mantle in the Pacific. *Geophys J Int* 96:203–229
- Orcutt JA, Kennett BLN, Dorman LM, Prothero WA (1975) Evidence for a low velocity zone underlying a fast-spreading rise crest. *Nature* 256:475–476
- Parsons B, Sclater J (1977) An analysis of the variation of ocean floor bathymetry and heat flow with age. *J Geophys Res* 82:803–827
- Raitt RW, Shor GG Jr, Morris GB, Kirk HK (1971) Mantle anisotropy in the Pacific Ocean. *Tectonophysics* 12:173–186
- Richards FD, Hoggard MJ, Cowton LR, White NJ (2018) Reassessing the thermal structure of oceanic lithosphere with revised global inventories of basement depths and heat flow measurements. *J Geophys Res Solid Earth* 123:9136–9161. <https://doi.org/10.1029/2018JB015998>
- Rychert CA, Shearer PM (2011) Imaging the lithosphere-asthenosphere boundary beneath the Pacific using SS waveform modeling. *J Geophys Res* 116:B07307. <https://doi.org/10.1029/2010JB008070>
- Sereno TJ, Orcutt JA (1987) Synthetic Pn and Sn phases and the frequency dependence of Q of oceanic lithosphere. *J Geophys Res* 92:3541–3566
- Shito A, Suetsugu D, Furumura T, Sugioka H, Ito A (2013) Small-scale heterogeneities in the oceanic lithosphere inferred from guided waves. *Geophys Res Lett* 40:1708–1712. <https://doi.org/10.1002/grl.50330>
- Shito A, Suetsugu D, Furumura T (2015) Evolution of the oceanic lithosphere inferred from Po/So waves traveling in the Philippine

- Sea Plate. *J Geophys Res Solid Earth* 120:5238–5248. <https://doi.org/10.1002/2014JB011814>
- Shor GG Jr, Menard HW, Raitt RW (1969) Structure of the Pacific Basin. In: Maxell AE (ed) *The sea*, vol 4. Wiley, New York, pp 3–27
- Singh SC, Harding AJ, Kent GM, Sinha MC, Combier V, Bazin S, Tong CH, Pye JW, Barton PJ, Hobbs RW, White RS, Orcutt JA (2006) Seismic reflection images of the Moho underlying melt sills at the East Pacific Rise. *Nature* 442:287–290. <https://doi.org/10.1038/nature04939>
- Stein CA, Stein S (1992) A model for the global variations in oceanic depth and heat flow with lithospheric age. *Nature* 359:123–129
- Stein CA, Stein S (1994) Constraints on hydrothermal heat flux through the oceanic lithosphere from global heat flow. *J Geophys Res* 99(B2):3081–3095
- Tolstoy M, Harding AJ, Orcutt JA (1993) Crustal thickness on the Mid Atlantic Ridge: bull's gravity anomalies and focused accretion. *Science* 262:726–729
- Toomey DR, Purdy GM, Wilcock WSD (1990) The three dimensional structure of the East Pacific Rise, near latitude 9°30'N. *Nature* 347:639–645
- Watts AB (2001) *Isostasy and flexure of the lithosphere*. Cambridge University Press, Cambridge, UK, p 458. <https://doi.org/10.2277/0521006007>
- Wiens DA, Stein S (1983) Age dependence of oceanic intraplate seismicity and implications for lithospheric evolution. *J Geophys Res* 88:6455–6468

Lithosphere, Oceanic: Thermal Structure

Earl E. Davis¹ and David S. Chapman²

¹Pacific Geoscience Centre, Geological Survey of Canada, Sidney, BC, Canada

²Department of Geology and Geophysics, The University of Utah, Salt Lake City, UT, USA

Definition

Lithosphere

The outermost shell of the Earth that is sufficiently strong to support long-term geologic loads and transmit stress. These properties are attributed to low temperature and possibly to composition.

Asthenosphere

The mobile layer underlying the lithosphere where heat transfer is dominated by convection. Its properties are attributed to high temperature and trace quantities of water.

Plate tectonics

Lateral motion of large intact pieces of lithosphere (plates) that include both continental and oceanic regions. Mountain building, volcanoes, and earthquakes are the consequences of the relative motion between plates at their convergent, divergent, and transform boundaries.

Boundary-layer cooling

Conductive cooling of an initially hot semi-infinite medium, applied to the vertical cooling of aging oceanic lithosphere created at a seafloor spreading center.

Plate cooling

The description of the cooling of an initially hot layer of uniform thickness having a lower boundary maintained at a uniform temperature.

Rayleigh number

A dimensionless number reflecting the ratio of the forces driving and resisting convection.

Nusselt number

A dimensionless number reflecting the vigor of convection, equal to the ratio of the total heat transport relative to the transport that would take place by conduction alone across the same temperature differential.

Introduction

The thermal structure of the oceanic lithosphere is of profound importance. One of the dominant mechanisms of heat loss from the Earth is the creation at seafloor spreading centers and subsequent cooling of oceanic lithosphere. Ocean lithosphere constitutes roughly 60% of the Earth's surface, and the total heat flow through its surface comprises more than 80% of the global mantle heat flow (i.e., excluding the contribution from radiogenic heat production in continental crust; Jaupart et al. 2007). The thermal structure of the lithosphere plays key roles in constraining the rigidity and ultimate strength (seismic rupture potential) of the outermost layer of the Earth and influencing many geodynamic processes including the Earth's most fundamental and defining one, plate tectonics. Driving forces for the motion of tectonic plates, and numerous consequent geological processes, are derived from the integrity and density of the oceanic lithosphere as controlled to a large degree by its thermal structure. And the volume of ocean basins, and hence the continental freeboard relative to sea level, is influenced directly by the thermal structure of oceanic lithosphere.

“Lithosphere” has multiple meanings. It was first defined as the outer layer of the Earth, including the crust and upper mantle, having sufficient strength to support loads such as those imposed by mountains, volcanoes, and major river deltas (Barrell 1914). It was later identified as the high seismic velocity lid that overlies a deeper low-velocity zone in the mantle; its base was inferred to correspond to the closest approach of the geotherm to the mantle melting temperature (Gutenberg 1959). More recently and most fundamentally, it has become a geodynamic term that refers to the part of the mantle that forms relatively rigid plates, and within which

heat is transported by conduction. Flexural strength, long-term immobility, seismic properties, and possibly composition all are properties that are used to define the lithosphere and its thickness, and all are tied to temperature. Beneath the lithosphere is the relatively mobile asthenosphere, where heat transport is dominated by vigorous (high Rayleigh number, high Nusselt number) convection. The oceanic lithosphere forms a thermomechanical upper boundary layer to a “conveyor-belt” convection system, in which the Earth’s interior heat is liberated through the seafloor as the layer moves laterally away from the locus of formation at a seafloor spreading center and is absorbed by the cooled layer when it sinks back into the asthenosphere at a subduction zone. Convection in the asthenosphere is to a large extent decoupled from the lithosphere. It is characterized by Rayleigh and Nusselt numbers of the order of 10^7 and 200, respectively, and is driven both by cooling at the top by heat lost through the lithosphere and by heating from the bottom and internally by deep and distributed sources derived from secular cooling and radiogenic heat production.

Oceanic lithosphere includes a c. 7-km-thick basaltic oceanic crust, a product of partial melting of the upper mantle at a seafloor spreading center, and underlying upper mantle that has been cooled by an amount that gives the material its short- and long-term mechanical and elastic properties that contrast with those of the underlying mobile asthenosphere. Contrasting composition, resulting from the refractory process of producing the oceanic crust and from water extraction during asthenospheric ascent, may also play a part in defining the properties of the oceanic lithosphere. These various aspects are summarized schematically in Fig. 1. Previous reviews can be found in Davis (1989) and Jaupart and Mareschal (2007).

Key Observations

Because the creation and cooling of ocean lithosphere is a transient process, observations used to constrain its thermal structure are most commonly placed in the context of lithospheric age and considered in the context of a model for transient cooling. Various observations are discussed in this section, and theoretical models are discussed in the next.

Heat Flux

Heat flux through the seafloor provides the most direct constraint on the thermal structure of the underlying lithosphere, although observational uncertainties are often large (Fig. 2a). Detailed studies of the thermal regime just below the seafloor commonly reveal a hydrologic regime that in young areas is dominated by high Rayleigh- and Nusselt-number pore-water convection (see discussions in Harris and Chapman 2004 and

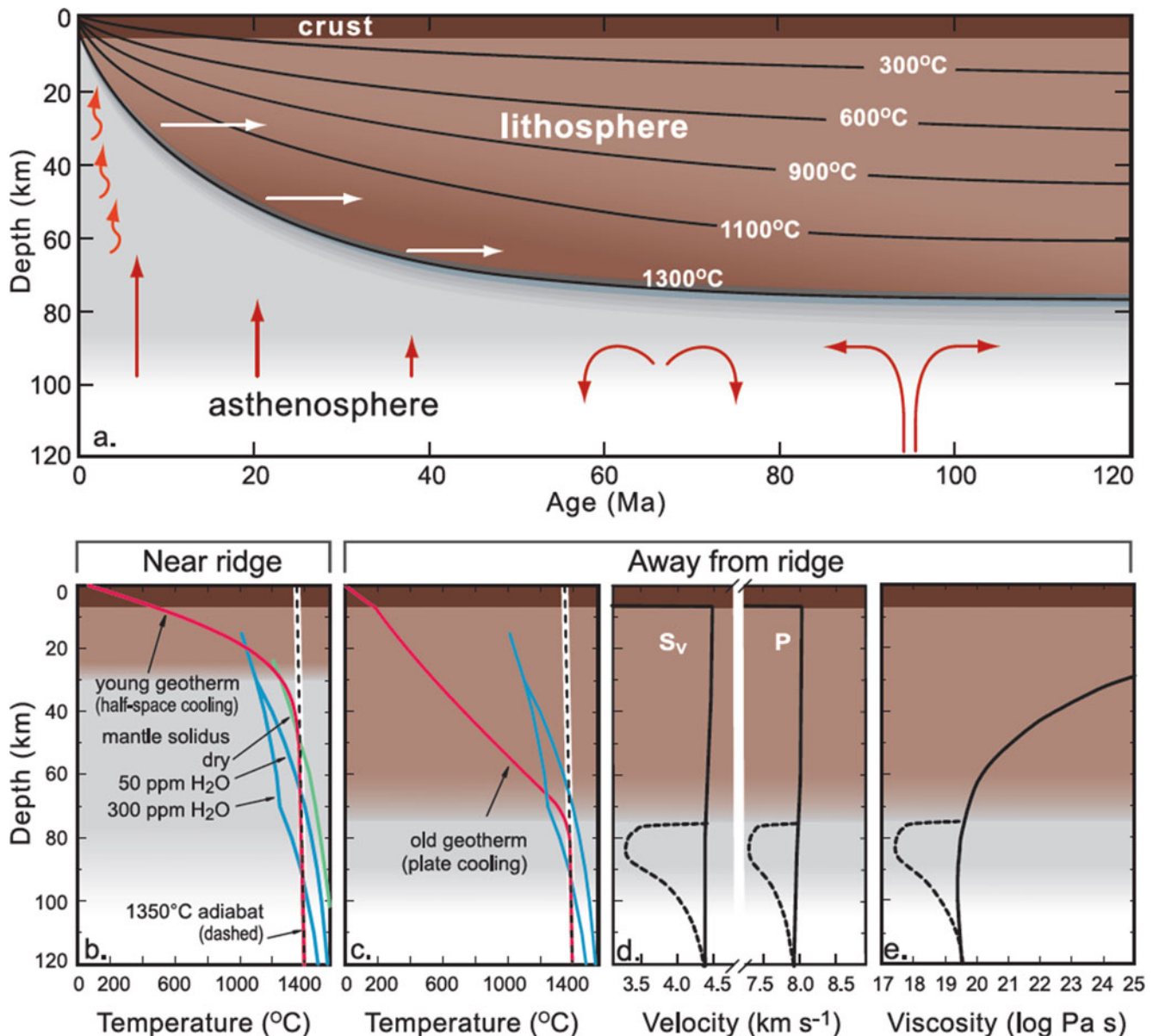
► [Heat Flow, Seafloor: Methods and Observations](#) entry, this volume), with high permeability in the uppermost few hundred meters of young igneous crust resulting from the volcanic and tectonic processes involved in crustal creation. Hydrothermal alteration of the crust slowly reduces the permeability, but the effects of convective heat transport are observed to persist for tens of millions of years (Von Herzen 2004). Heat passing advectively through the seafloor is virtually impossible to assess quantitatively and often goes unrecognized. Hence, wholesale compilations of seafloor heat flux observations are strongly biased (e.g., Stein and Stein 1992), particularly in young, sparsely sedimented areas where igneous crustal outcrops, serving as permeable ventilation points, are common. Where thick and extensively continuous accumulations of low-permeability sediment are present, advective ventilation ceases, and the heat flow becomes fully conductive and thus measurable. If circulation within the igneous crust persists, however, local heat flux variations can still be present, and large numbers of measurements must be made over a large area to obtain a meaningful average of the conductive seafloor heat flux at any given age.

To avoid the biasing effects of advective heat loss, two approaches have been used. Values from global data sets have been selected to be far from areas of basement outcrop (relying on statistics for a reliable mean to emerge; e.g., Sclater et al. 1976; Hasterok et al. 2011). A more direct approach involves surveys carried out with large numbers of observation points positioned carefully in context of local and regional sediment/igneous hydrologic structure to rule out effects of ventilation and properly account for the effects of sub-sedimentary fluid flow that produces scatter (e.g., Lister et al. 1990; Davis et al. 1999; Fisher et al. 2003). From such studies, a reliable relationship between lithospheric heat flux and age has emerged, with heat flux decreasing proportionately with age^{-1/2} out to an age of roughly 100 Ma (Fig. 2b) and becoming fairly constant at greater ages at roughly 48 mW m⁻².

Seafloor Depth

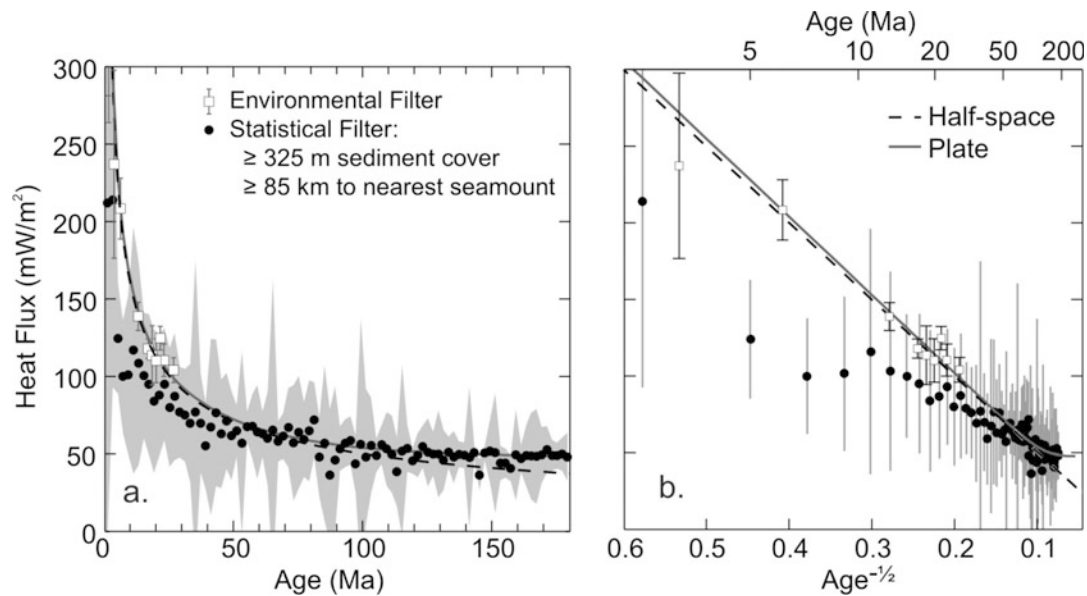
The most reliably measured consequence of lithospheric cooling is seafloor subsidence (Fig. 3a). Whereas heat flow determinations are affected by hydrothermal circulation in the upper oceanic crust, seafloor subsidence with increasing age away from a ridge results from lithospheric thermal contraction and hence reflects the integrated temperature state of the entire lithosphere. Bathymetry thus provides a robust constraint for understanding lithospheric thermal structure.

As in the case of seafloor heat flux, seafloor depths must be considered with some care, with attention given to several potential perturbations. This vigilance requires checks on



Lithosphere, Oceanic: Thermal Structure, Fig. 1 Schematic illustration of the creation and subsequent thermal evolution of oceanic lithosphere (a). Mechanical properties of the lithosphere are likely influenced primarily by temperature (brown shading) and possibly composition (light gray shading). Red arrows depict the migration of partial melt supplying crustal formation. Dark red arrows depict three modes of convection in the asthenosphere, including (1) upwelling beneath migrating lithosphere wherever it is thickening, (2) small-scale convection driven by cooling and associated instability below the base of the lithosphere, and (3) plumes driven by deep-seated thermal instabilities. Wherever temperatures are higher than the solidus (b), water will be absorbed by partial melt. Any migration of this melt will leave a dried residual asthenosphere with anomalous melting temperature and mechanical properties. Conductive cooling at great age may be limited to a compositionally established depth or arrested by regulated convection in the asthenosphere (as shown by the plate-cooling geotherm in (c)). Properties such as seismic velocity (d) and viscosity (e) are likely

controlled by a combination of temperature, water content, and locally ponded partial melt. The zone of partial melt in the asthenosphere rising to supply material for crustal creation and lithospheric thickening has been discussed by Plank and Langmuir (1992) and others; the solidi (for dry and damp peridotite) are adopted from the summary of Asimow et al. (2004); the old ocean geotherm is computed using the temperature-dependent thermal conductivity structure summarized by Hasterok and Chapman (2011) including radiative heat transport (e.g., Hofmeister 1999) and an assumed radiogenic heat production rate of $0.5 \mu\text{W m}^{-3}$ in the oceanic crust; effects of sediment accumulation or crustal hydrothermal circulation are not included; velocity structure is adopted from surface wave tomographic results of Schaeffer and Lebedev (2013) with the localized low-velocity zone following results of Mehrouachi and Singh (2018); viscosity, based on Mitrovica and Forte (2004), is highly schematic and includes an anomalous zone corresponding to the seismic low-velocity layer



Lithosphere, Oceanic: Thermal Structure, Fig. 2 Oceanic heat flux. Data (from Hasterok et al. 2011) are plotted versus lithospheric age in (a) and versus age^{-1/2} in (b). Because raw heat flux data in young, sparsely sedimented areas are systematically biased by hydrothermal circulation, the data in these plots have been filtered to exclude sites where sediment thickness is less than 325 m and which are within 85 km of a known seamount. Data are drawn from all oceans. Solid dots show median heat

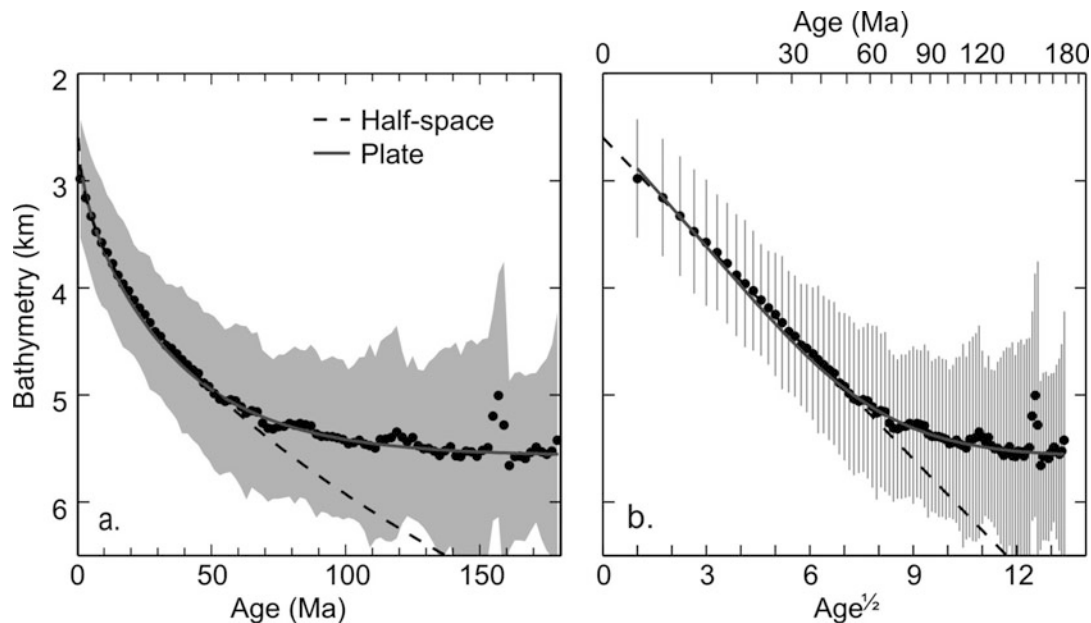
flux for 2 My age bins; gray shading is ± 1 standard deviation for the measurements in a bin. Open symbols represent high-quality data sets drawn from experiments where the environment of the heat flux sites is known from seismic imaging of the seafloor and other geophysical observations. The dashed and solid lines represent heat flux for half-space and plate cooling models, respectively

isostatic compensation, sediment loading, crustal thickness variations, and sub-lithospheric density anomalies. Early data compilations were made with only modest regard to these factors (e.g., Davis and Lister 1974; Parsons and Sclater 1977); later studies, such as those of Schroeder (1984), Marty and Cazenave (1989), Johnson and Carlson (1992), Hillier and Watts (2005), and Crosby et al. (2006), have put more effort into “filtering” data to define depth/age transects that are relatively free of the effects of known or suspected major mantle convective upwellings and downwellings, crustal thickness anomalies, and thick sediment accumulations. Despite the various potential sources of “noise” (which, as in the case of heat flux data, add bias, in this case typically causing observed depths to be anomalously shallow), all compilations have shown a similar result, with seafloor depths increasing proportionately with age^{1/2} until 70–80 Ma (Fig. 3b). Depths at greater ages fall off of this trend and generally stabilize at about 5.6 km. Other systematic behavior has been gleaned from the compilations, most notably that the rates of subsidence vary both locally and regionally and that the rate of subsidence is related to the depth of the local ridge axis (Davis and Lister 1974; Marty and Cazenave 1989; Hillier and Watts 2005; Crosby et al. 2006). This behavior is believed to reflect variations in the temperature of the asthenosphere supplying material at and near spreading centers (Klein and Langmuir 1987).

Depth of the Lithosphere-Asthenosphere Boundary

A third measureable indication of the thermal state of oceanic lithosphere is lithospheric thickness itself, identified principally by observations from seismology. Laboratory observations and field calibrations show that shear-wave velocity in mantle rocks is a strong function of temperature, with values decreasing rapidly (up to 0.1% K⁻¹) with approach to the melting temperature (Gibb and Cooper 1998; Priestley and McKenzie 2006). Hence, shear-wave velocities determined from dispersive surface waves provide a useful constraint on temperature at depth. An even greater influence derives from the presence of partial melt, which is controlled by temperature and the presence of water. Partial melt also reduces compressional wave velocity and electrical resistivity.

Surface wave data show a clear age-dependent structure for the uppermost mantle. A low-velocity zone is distributed widely in the oceans; it is thick and rises to the surface at mid-ocean ridges and extends beneath a high-velocity lid (the “seismic lithosphere”) that thickens with age out to roughly 70 Ma (Ritzwoller et al. 2004; Schaeffer and Lebedev 2013). Overall it centered at roughly 100 km. The long wavelength of surface waves poses a severe limitation on the depth resolution of the structure defined. This limit is being overcome by use of receiver function analyses permitted by a small but growing number of long-term broadband



Lithosphere, Oceanic: Thermal Structure, Fig. 3 Oceanic bathymetry. Data (from Hasterok et al. 2011) are plotted versus lithospheric age in (a) and versus age^{1/2} in (b). Bathymetry plotted is the position of the top of the igneous crust, 1-D isostatically adjusted for the effect of local

sediment loading. Solid dots are median values for 2 My bins of seafloor age; gray shading is ± 1 standard deviation for the measurements in a bin. Dashed and solid lines represent bathymetry for half-space and plate cooling models, respectively

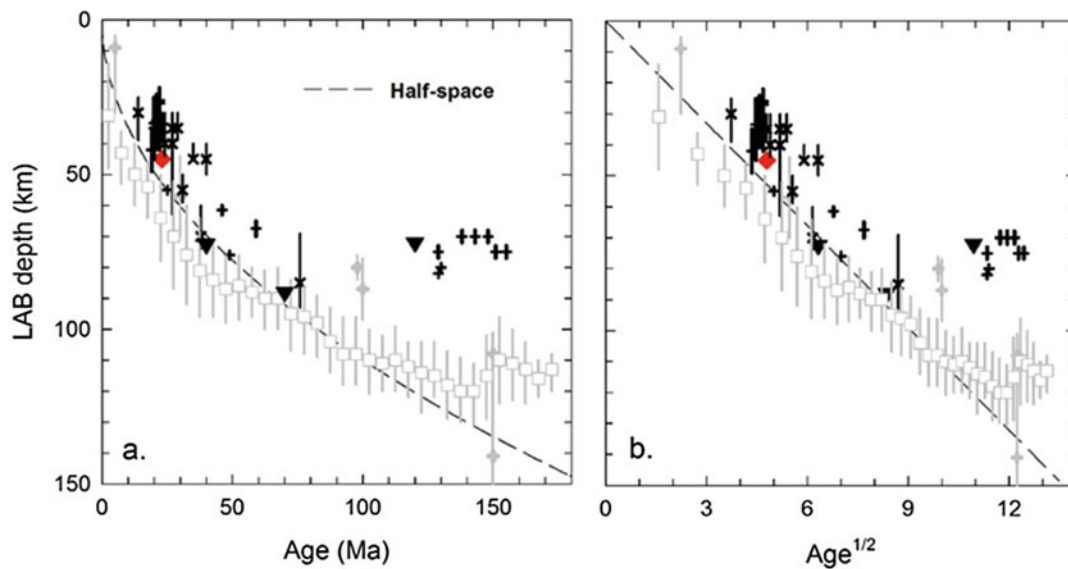
seismometer deployments (e.g., Kawakatsu et al. 2009), by analysis of SS waveforms that contain a reflection from the base of the lithosphere (Rychert and Shearer 2011), and by active-source seismic reflection data where recording has been extended to several tens of seconds (e.g., Stern et al. 2015; Mehouchi and Singh 2018). Observations made using these methods, particularly the last, require a relatively sharp lower boundary to the lithosphere and a thin zone hosting partial melt beneath (although this inference has been debated; e.g., Olugboji et al. 2016; Niu and Green 2018). Magnetotelluric data also strongly suggest the presence of partial melt at the base of the lithosphere, distributed in a way that is sufficiently interconnected to cause anomalously high electrical conductivity with a bias in the direction of plate motion (e.g., Naif et al. 2013).

Depths to the base of the lithosphere defined using these various techniques vary significantly, with a systematic difference between those estimated from shear-wave dispersion observations (gray symbols in Fig. 4) and those determined from seismic reflections and other techniques (black symbols). Both groups, however, suggest an age dependence of lithospheric thickness in younger settings and stabilization in older settings. All methods show great promise; additional studies will be needed to define just how consistent or variable the relationship between lithospheric thickness and age might be and to determine whether seismic reflections define the thickness of the lithosphere more accurately than surface wave dispersion.

Other Observational Constraints

Additional constraints on lithospheric thermal structure are provided by observations of flexural rigidity, gravity, and geoid. Gravity data provide little information about the very long-wavelength signature of lithospheric cooling except to confirm that the topography is isostatic. At the shorter wavelengths of abyssal hills, seamounts, and swells, gravity is sensitive to the vertical mass distribution and can be used to discriminate whether topography is supported or unsupported by lithospheric strength and to constrain the depth of compensation of isostatic topography (e.g., Watts 2007; Crosby and McKenzie 2009). The geoid is sensitive to long-wavelength density structure and provides a useful constraint on the thermally influenced density of oceanic lithosphere. A linear relationship between geoid height and age is seen over young seafloor, with a proportionality ranging between 0.1 and 0.14 m Myr⁻¹, whereas the response is flat at ages greater than 80 Ma (e.g., Haxby and Turcotte 1978; Sandwell and Schubert 1980; De Laughter et al. 1999).

Flexural rigidity (usually expressed as effective elastic thickness) and ultimate strength are strongly controlled by temperature. The first of these related properties is reflected by the long-term load-bearing capacity and is determined using the wavelength of flexure and the relationship between gravity and topography in the vicinity of large seamounts, trenches, fracture zones, and other crustal or tectonic loads. As in the case of the thickness of the high shear-wave-velocity lid, a clear age dependence has been revealed that shows that



Lithosphere, Oceanic: Thermal Structure, Fig. 4 Oceanic lithosphere thickness measured as the depth to the seismic low-velocity zone. Data, primarily from the Pacific Basin, are plotted versus lithospheric age in (a) and versus age^{1/2} in (b). Symbols indicate types of analysis and uncertainties: gray symbols indicate estimates derived from surface wave dispersion (Leeds et al. 1974, Leeds 1975, crosses; Burgos et al. 2014, squares); black crosses are from receiver function analyses

(Kawakatsu et al. 2009 and Olugboji et al. 2016); black x's are from SS precursor analysis (Rychert and Shearer 2011); inverted triangles are from active-source seismic reflection profiling (Mehouachi and Singh 2018); and the red diamond is from magnetotelluric data analysis (Naif et al. 2013). The dashed line shows the rate of thickening expected for half-space cooling

the effective elastic thickness is limited by temperature, although the scatter in thicknesses and inferred maximum temperatures is large, with the latter falling between 200 and 600 °C. This reflects in part the dependence of this property on the duration of loading.

A limiting constraint on ultimate strength is provided by the occurrence of earthquakes, and when placed in a context of age, a relationship is also seen. The maximum depth of hypocenters increases with increasing age, and the limiting temperatures for seismogenic failure falls at roughly 600 °C (McKenzie et al. 2005). Again, the rate of loading that leads to seismogenic failure is an important factor (in addition to temperature) controlling this property. Settings in which deep earthquakes occur are normally complex (such as in subduction zones), however, and this restricts the quantitative utility of the depth distribution of earthquakes for defining the thermal structure of the lithosphere.

Ways of Understanding Lithospheric Thermal Structure

Because the lithosphere is not in an equilibrium state in most oceanic regions, the local thermal structure cannot be defined directly from seafloor heat flux, subsidence, or other observations. It must be estimated using transient cooling models, constrained by the observations described above along with estimates of material properties at depth.

Boundary-Layer Cooling

The simplest physical description of the thermal structure of oceanic lithosphere is provided by boundary-layer cooling theory (Parker and Oldenberg 1973; Davis and Lister 1974). Except very near the ridge axis, cooling is vertical and equivalent to cooling of an initially uniform-temperature half space. The lithosphere is defined simply as cooled asthenosphere with the boundary defined by an isotherm that deepens with age and thus distance from the ridge. Heat flux through the seafloor, Q , decreases with age, t , as $Q = \lambda T_a (\pi \kappa t)^{-1/2}$, and seafloor depth, h , increases as $h = 2 \alpha_{eff} T_a (\kappa t / \pi)^{1/2}$, where T_a = the initial temperature of the asthenospheric material from which the lithosphere forms (relative to seawater at $T \approx 0$ °C), κ = thermal diffusivity, $\alpha_{eff} = \alpha \rho_0 / (\rho_a - \rho_w)$ is the effective volumetric thermal expansivity adjusted for seawater loading, and ρ_a and ρ_w are the densities of the asthenosphere and seawater, respectively. The thermal properties are known to be sensitive to temperature (e.g., Hofmeister 1999); this dependence must be considered when estimating temperatures at depth (e.g., McKenzie et al. 2005), but it does not change the simple predicted dependence of subsidence and heat flux on age. The simplest application of boundary-layer cooling theory produces a heat flux singularity at the ridge axis; this is dealt with by applying a heat-balance boundary condition at the ridge axis (Davis and Lister 1974). Two important predictions of the theory are that the heat flux integrated from $t = 0$ to $t = a$ is $2 Q(a)$, and that the heat flux is related to seafloor depth as $Q = (\rho c_p / \alpha_{eff}) dh/dt$,

where c_p is the specific heat at constant pressure and α_{eff} is appropriate for 0.7 of T_a (Lister 1977). Wherever cooling follows boundary-layer behavior, this relationship can be used to obtain a robust estimate of the heat loss from the oceanic lithosphere without a need for relying on sparse, scattered, and potentially biased seafloor heat flux data.

When plotted as functions of $t^{-1/2}$ and $t^{1/2}$ as they are in Figs. 2b and 3b and considered in light of the heat flux/subsidence relationship given above, the heat flux and depth observations clearly demonstrate the efficacy of boundary-layer cooling theory for characterizing lithosphere younger than 70–80 Ma. The same is true for the geoid, which is predicted by boundary-layer cooling theory to decline at a rate of 0.16 m Myr^{-1} (Haxby and Turcotte 1978), and for the thickening of the lithosphere itself, which (as defined by seismic surface wave dispersion analysis) occurs at a rate of roughly $11 \text{ km Ma}^{-1/2}$ (Fig. 4b). The cessation of simple subsidence, the flattening of the geoid, and the stabilization of the lithosphere/asthenosphere boundary depth beyond about 70 Ma, as well as the relatively constant heat flux beyond an age of roughly 100 Ma, all require something more complex than heat loss via transient cooling, however.

Plate Cooling Models

That ocean depths at great ages are relatively uniform and consistently shallower than those predicted by boundary-layer cooling provided the fundamental rationale for plate cooling models, which describe the lithosphere as a layer of constant thickness having a fixed lower boundary temperature (McKenzie 1967). This description was initially used out of mathematical convenience. There was little physical justification for the lithosphere to have an intrinsic thickness, or for constant heat flux to be supplied by asthenospheric convection beneath old lithosphere, until physical arguments were made for the latter by Parsons and McKenzie (1978) by way of an age-dependent onset of convection in the growing thermal boundary layer below the cooled, mechanically immobile lithosphere. Physical modeling of convection using material having a strongly temperature-dependent viscosity has shown how this phenomenon can serve to stop and possibly even reverse subsidence and later stop the decline in heat flux with age (Davaille and Jaupart 1994; Crosby et al. 2006). The key factor in this model is the growth of the thermal boundary layer beneath the conductively cooled plate. Once this layer reaches a sufficient thickness to become convectively unstable, convection begins to bring heat advectively from the large reservoir of the asthenosphere beneath. A balance is ultimately reached between the heat conducted through the lithosphere and that carried by convection in the asthenosphere.

Alternate means suggested for maintaining an effectively steady-state lithosphere at great age include convection driven from below (plumes, distributed radiogenic heat, and

secular cooling) and creation near ridge axes of a compositionally distinct layer (Oxburgh and Parmentier 1977; Morgan 1997). The first by itself is precluded as a general case by both topographic and heat flux observations which show no statistically significant influence of augmented heat flux at intermediate ages where the evolution of the lithosphere is well matched by boundary-layer cooling theory. Regulated heat flux by asthenospheric convection is also argued against by the contrast between heat flux at the base of the lithosphere beneath old oceanic and old continental regions, the latter being less than half of the former. The second seems unlikely in light of the observations of seismic reflections at the base of the lithosphere (e.g., Fig. 4), which suggest that there is a weak layer at the lithosphere-asthenosphere boundary (depicted schematically in Fig. 1e). Any compositional differences resulting from the extraction of partial melt and water from the asthenosphere in the source zone for the oceanic crust (e.g., Plank and Langmuir 1992; Asimow and Langmuir 2003; see Fig. 1a) must be concentrated in a volume near the ridge axis and not carried laterally for lithospheric underplating at greater age. The effects of composition and water content must not to be discounted in terms of their influence on viscosity and melting temperature, however (e.g., Hirth and Kohlstedt 1996; Asimow and Langmuir 2003; Asimow et al. 2004). Preconditioning via partial melt of the asthenosphere from which the lithosphere is created may happen wherever asthenospheric ascent occurs to replace material lost to the thickening and laterally migrating lithosphere.

Influence of Plumes

Convective plumes in the asthenosphere, with defining characteristics of localized upwelling and anomalously high temperatures (possibly more than 200 K higher than typical asthenospheric temperatures; Putirka 2005), have been argued to thin or “reset” the thermal age of oceanic lithosphere (see review in Detrick et al. 1989), with the eventual collective contribution from plumes resulting in the departure from boundary-layer cooling behavior (Heestand and Crough 1981) and possibly in the supply of much of the heat flux to the base of plates (Malamud and Turcotte 1999), although these possibilities remain controversial. Locally the influence of plumes can certainly be large; surface volcanism from plumes is typically highly focused, but the associated topographic swells are large and extensive, often exceeding 1 km in amplitude and having a wavelength approaching 1000 km. A similarly broad thermal anomaly is implied for a source of the swells, but given the large thermal time constant of the lithosphere, seafloor heat flux provides an inherently poor constraint on the depth of the thermal source; attempts have been made to infer anomalies at depth with understandably mixed results (e.g., Von Herzen et al. 1989). A study of the geoid signature over the Hawaiian Swell by Moore and

Schubert (1997) suggests that the lower part of the lithosphere is involved, but a study of the topography and gravity over swells by Crosby and McKenzie (2009) reveals an admittance (30 mGal km^{-1}) that is consistent with the thermal density anomaly being below the base of the plate.

Driving Forces for Lithospheric Motion

An element that is essential to a complete description of oceanic plates is the source of the force that drives their motion. Great attention has been given to the subject of plate driving forces ever since the formulation of the concept of plate tectonics (McKenzie 1969; Forsyth and Uyeda 1975; Runcorn 1980; Richardson 1992). Possible forces considered have been traction from large-scale convection at the base of the lithosphere (both oceanic and continental), negative buoyancy at subduction zones, and gravitational sliding away from mid-ocean ridges. Drive from basal traction forces is probably insignificant. Convection in the asthenosphere is generally too small in scale and directionally random, outside of the general flow required to maintain a mass balance between lithospheric creation at ridges and destruction at subduction zones which would generate a bias in the wrong direction. In addition, the low seismic shear velocity, high electrical conductivity, and possible presence of melt (Fig. 1) suggest that the lithosphere is mechanically decoupled from the asthenosphere. Such decoupling reduces any force transfer at the base of the asthenosphere, including near mid-ocean ridges where there could be a potential for “ridge push.”

Most of the driving force for plate tectonics probably arises from the lithosphere itself, specifically the thermally derived density contrast between the lithosphere and asthenosphere. Wherever boundary-layer cooling governs the thickness of the lithosphere (i.e., where plate age is less than 70–80 Ma), gravitational forces associated with the thermally elevated lithosphere drive it down along its sloping lower boundary. This results in a body force that is constant with age (e.g., Lister 1975). The greatest role the lithosphere plays in moving itself is believed to occur at subduction zones, where cooled oceanic lithosphere sinks into the asthenosphere. The dominance of this negative buoyancy force is apparent in the global variations of plate velocities: plates bounded by subduction zones that are lengthy and consume old oceanic lithosphere, specifically the Pacific and Australian/Indian plates, move at the fastest rates. The negative buoyancy of subducting plates also provides the force for the genesis of mountain belts.

Summary

A broad suite of observations gathered in context of powerful physical models have brought plate tectonics from a largely

geometrical description of the history of the surface of the Earth to a robust theory that explains many geological processes. Among other things, this theory addresses the mechanisms by which plates form and are driven, the thermal structure of oceanic lithosphere that influences its mechanical and geodynamic behavior, and the dominant modes of heat loss from the Earth. In particular, the mathematics of half-space cooling – applied to the physical model of seafloor spreading and constrained by carefully considered observations – leads to simple but profound approximations for the heat loss, bathymetry, and lithospheric thickness in young ($< 70\text{--}80 \text{ Ma}$) oceanic regions of the Earth and suggests the following relationships for predicting heat flow through the seafloor (Q , in mW m^{-2}), depth below sea level of the seafloor (d , in km), and depth below seafloor of the lithosphere-asthenosphere boundary (l , in km) as functions of age (t , in My):

$$\begin{aligned} Q &= 500t^{-1/2} \\ d &= 2.6 + 0.34t^{1/2} \\ l &= 11t^{1/2} \end{aligned}$$

Departures from these relationships occur in regions of anomalous upper mantle temperatures and in regions of greater age, but with modifications for seafloor greater than 80 My to account for the limits of conductive cooling, these simple approximations explain heat flow, bathymetry (thermal subsidence), and lithosphere thickness for more than two thirds of the Earth's surface.

Uncertainties remain, however, and further work is required to resolve several lingering questions about lithospheric creation, cooling, and eventual stabilization. Among these are (1) to what degree is the mechanical behavior of oceanic lithosphere (particularly its resistance to thinning by convection of any form in the asthenosphere) established compositionally by the extraction of partial melt and water at ridge axes or in young regions; (2) to what degree does asthenospheric convection stimulated by top-down cooling regulate heat loss through and limit the growth of the lithosphere at great age; and (3) to what degree do plumes, the consequence of convective instabilities generated deep in the asthenosphere, contribute to heat flux at the base of plates, and do they convectively thin or merely conductively warm the plates beneath which they ascend? Addressing the first will improve constraints on plate drive generated by gravitational sliding forces, which are maximum for the case of a mechanical lithosphere being defined purely as a cooling boundary layer. Addressing the second and third will help resolve just what the primary regulating factor is that constrains the rate of heat loss from the Earth, such as the form and vigor of asthenospheric convection, the quotient of asthenospheric temperature and lithospheric thickness, or the rate of

production and destruction of lithosphere at ridges and subduction zones.

It is noteworthy that in addressing the primary topic of this chapter, the thermal structure of the oceanic lithosphere, we have had to lean heavily on indirect constraints. Even with the help of boundary-layer and plate cooling models, uncertainties of the parameters required for estimating deep lithospheric thermal structure from seafloor heat flux – specifically thermal conductivity, heat capacity, and heat production rate – lead to equivalent uncertainties in estimates of temperatures at depth. For this reason we have chosen to integrate other observations that constrain temperatures, as reflected in Figs. 1 and 4, and we urge readers to accept any estimates of deep temperatures with suitable caution.

Cross-References

- [Heat Flow, Seafloor: Methods and Observations](#)
- [Lithosphere, Continental: Thermal Structure](#)

Bibliography

- Asimow PD, Langmuir CH (2003) The importance of water to oceanic mantle melting regimes. *Nature* 421:815–820
- Asimow PD, Dixon JE, Langmuir CH (2004) A hydrous melting and fractionation model for mid-ocean ridge basalts: application to the mid-Atlantic ridge near the Azores. *Geochem Geophys Geosyst* 5(1):24pp
- Barrell J (1914) The strength of the Earth's crust. I. Geologic tests of the limits of strength. *J Geol* 22:28–48
- Burgos G, Montagner J-P, Beucler E, Capdeville Y, Mocquet A, Drilleau M (2014) Oceanic lithosphere-asthenosphere boundary from surface wave dispersion data. *J Geophys Res* 119:1079–1093
- Crosby AG, McKenzie D (2009) An analysis of young ocean depth, gravity and global residual topography. *Geophys J Int* 178:1,198–1,219
- Crosby AG, McKenzie D, Sclater JG (2006) The relationship between depth, age, and gravity in the oceans. *Geophys J Int* 166:555–573
- Davaille A, Jaupart C (1994) Onset of thermal convection in fluids with temperature-dependent viscosity: application to the oceanic mantle. *J Geophys Res* 99:19853–19866
- Davis EE (1989) Thermal aging of oceanic lithosphere. In: Wright JA, Loudon KE (eds) *Handbook of seafloor heat flow*. CRC Press, Boca Raton, pp 145–167
- Davis EE, Lister CRB (1974) Fundamentals of ridge crest topography. *Earth Planet Sci Lett* 21:405–413
- Davis EE, Chapman DS, Wang K, Villinger H, Fisher AT, Robinson SW, Grigel J, Pribnow D, Stein JS, Becker K (1999) Regional heat flow variations on the sedimented Juan de Fuca Ridge eastern flank: constraints on lithospheric cooling and lateral hydrothermal heat transport. *J Geophys Res* 104:17675–17688
- De Laughter J, Stein S, Stein CA (1999) Extraction of a lithospheric cooling signal from oceanwide geoid data. *Earth Planet Sci Lett* 174:173–181
- Detrick RS, White RS, Courtney RC, Von Herzen RP (1989) Heat flow on midplate swells. In: Wright JA, Loudon KE (eds) *Handbook of seafloor heat flow*. CRC Press, Boca Raton, pp 169–190
- Fisher AT, Stein CA, Harris RN, Wang K, Silver EA, Pfender M, Hutnak M, Cherkaoui A, Bodzin R, Villinger H (2003) Abrupt thermal transition reveals hydrothermal boundary and role of sea-mounts within the Cocos plate. *Geophys Res Lett* 30:1550. <https://doi.org/10.1029/2002GL016766>
- Forsyth D, Uyeda S (1975) On the relative importance of the driving forces of plate motion. *Geophys J R Astron Soc* 40:163–200
- Gibb TT, Cooper RF (1998) Low-frequency shear wave attenuation in polycrystalline olivine: grain boundary diffusion and the physical significance of the Andrade model for viscoelastic rheology. *J Geophys Res* 103:27267–27279
- Gutenberg B (1959) *Physics of the Earth's interior*. Academic Press, New York
- Harris RN, Chapman DS (2004) Deep-seated oceanic heat flux, heat deficits, and hydrothermal circulation. In: Davis EE, Elderfield H (eds) *Hydrogeology of the oceanic lithosphere*. Cambridge University Press, Cambridge, pp 311–336
- Hasterok D, Chapman DS (2011) Heat production and geotherms for the continental lithosphere. *Earth Planet Sci Lett* 307:59–70
- Hasterok D, Chapman DS, Davis EE (2011) Oceanic heat flow: implications for global heat loss. *Earth Planet Sci Lett* 311:386–395
- Haxby WF, Turcotte DL (1978) On isostatic geoid anomalies. *J Geophys Res* 83:5473–5478
- Heestand RL, Crough TS (1981) The effect of hot spots on the oceanic age-depth relation. *J Geophys Res* 86:6107–6114
- Hillier JK, Watts AB (2005) Relationship between depth and age in the North Pacific Ocean. *J Geophys Res* 110. <https://doi.org/10.1029/2004JB003406>
- Hirth G, Kohlstedt DL (1996) Water in the oceanic upper mantle: implications for rheology, melt extraction and the evolution of the lithosphere. *Earth Planet Sci Lett* 144:93–108
- Hofmeister A (1999) Mantle values of thermal conductivity geotherm from phonon lifetimes. *Science* 283:1699–1709
- Jaupart C, Mareschal J-C (2007) Heat flow and thermal structure of the lithosphere. *Treat Geophys* 6:217–251
- Jaupart C, Labrosse S, Mareschal J-C (2007) Temperatures, heat, and energy in the mantle of the Earth. *Treat Geophys* 7:253–303
- Johnson HP, Carlson RL (1992) Variations of sea floor depth with age: a test of models based on drilling results. *Geophys Res Lett* 19:1971–1974
- Kawakatsu H, Kumar P, Yasuko T, Shinohara M, Kanazawa T, Araki E, Suyehiro K (2009) Seismic evidence for sharp lithosphere-asthenosphere boundaries of oceanic plates. *Science* 324:499–502
- Klein EM, Langmuir CH (1987) Global correlations of ocean ridge basalt chemistry with axial depth and crustal thickness. *J Geophys Res* 92:8089–8115
- Leeds AR (1975) Lithosphere thickness in the Western Pacific. *Phys Earth Planet Inter* 11:61–64
- Leeds AR, Knopoff L, Kausel EG (1974) Variations of upper mantle structure under the Pacific Ocean. *Science* 186:141–143
- Lister CRB (1975) Gravitational drive on oceanic plates caused by thermal contraction. *Nature* 257:663–665
- Lister CRB (1977) Estimators for heat flow and deep rock properties based on boundary layer theory. *Tectonophysics* 41:157–171
- Lister CRB, Sclater JG, Davis EE, Villinger H, Nagihara S (1990) Heat flow maintained in ocean basins of great age: investigations in the north-equatorial West Pacific. *Geophys J Int* 102:603–630
- Malamud BD, Turcotte DL (1999) How many plumes are there? *Earth Planet Sci Lett* 174:113–124
- Marty JC, Cazenave A (1989) Regional variations in subsidence rate of oceanic plates: a global analysis. *Earth Planet Sci Lett* 94:301–315
- McKenzie D (1967) Some remarks on heat flow and gravity anomalies. *J Geophys Res* 72:6261–6273
- McKenzie D (1969) Speculations on the consequences and causes of plate motions. *Geophys J R Astron Soc* 18:1–32

- McKenzie D, Jackson J, Priestley K (2005) Thermal structure of oceanic and continental lithosphere. *Earth Planet Sci Lett* 233:337–349
- Mehouachi F, Singh SC (2018) Water-rich sublithospheric melt channel in the equatorial Atlantic Ocean. *Nat Geosci* 11:65–69
- Mitrovica JX, Forte AM (2004) A new inference of mantle viscosity based upon joint inversion of convection and glacial isostatic adjustment data. *Earth Planet Sci Lett* 225:177–189
- Moore WB, Schubert G (1997) Lithospheric thinning and chemical buoyancy beneath the Hawaiian Swell. *Geophys Res Lett* 24:1287–1290
- Morgan JP (1997) The generation of a compositional lithosphere by mid-ocean ridge melting and its effect on subsequent off-axis hotspot upwelling and melting. *Earth Planet Sci Lett* 146:213–232
- Naif S, Key K, Constable S, Evans RL (2013) Melt-rich channel observed at the lithosphere-asthenosphere boundary. *Nature* 495:356–359
- Niu Y, Green DH (2018) The petrological control on the lithosphere-asthenosphere boundary (LAB) beneath ocean basins. *Earth Sci Rev* 185:301–307
- Olugboji TM, Park J, Karato S, Shinohara M (2016) Nature of the seismic lithosphere-asthenosphere boundary within normal oceanic mantle from high-resolution receiver functions. *Geochem Geophys Geosyst* 17:1265–1282
- Oxburgh ER, Parmentier EM (1977) Compositional and density stratification in oceanic lithosphere – causes and consequences. *J Geol Soc* 133:343–355
- Parker RL, Oldenberg DW (1973) Thermal model of ocean ridges. *Nature* 242:137–139
- Parsons B, McKenzie D (1978) Mantle convection and the thermal structure of the plates. *J Geophys Res* 83:4485–4496
- Parsons B, Sclater JG (1977) An analysis of the variation of ocean floor bathymetry and heat flow with age. *J Geophys Res* 82:803–827
- Plank T, Langmuir CH (1992) Effects of the melting regime on the composition of the oceanic crust. *J Geophys Res* 97:19749–19770
- Priestley K, McKenzie D (2006) The thermal structure of the lithosphere from shear wave velocities. *Earth Planet Sci Lett* 244:285–301
- Putirka KD (2005) Mantle potential temperatures at Hawaii, Iceland, and the mid-ocean ridge system, as inferred from olivine phenocrysts: evidence for thermally driven mantle plumes. *Geochem Geophys Geosyst* 6:Q05L08. <https://doi.org/10.1029/2005GC000915>
- Richardson RM (1992) Ridge forces, absolute plate motions, and the intraplate stress field. *J Geophys Res* 97:11739–11748
- Ritzwoller MH, Shapiro NM, Zong S-J (2004) Cooling history of the Pacific lithosphere. *Earth Planet Sci Lett* 226:69–84
- Runcorn K (1980) Some comments on the mechanism of continental drift. In: Davies PA, Runcorn SK (eds) *Mechanisms of continental drift and plate tectonics*. Academic Press, London
- Rychert CA, Shearer PM (2011) Imaging the lithosphere-asthenosphere boundary beneath the Pacific using SS waveform modeling. *J Geophys Res* 116:B07307. <https://doi.org/10.1029/2010JB008070>
- Sandwell D, Schubert G (1980) Geoid height versus age for symmetric spreading ridges. *J Geophys Res* 85:7235–7241
- Schaeffer AJ, Lebedev S (2013) Global shear speed structure of the upper mantle and transition zone. *Geophys J Int* 194:417–449
- Schroeder W (1984) The empirical age-depth relation and depth anomalies in the Pacific Ocean basin. *J Geophys Res* 89:9873–9883
- Sclater JG, Crowe J, Anderson RN (1976) On the reliability of ocean heat flow averages. *J Geophys Res* 81:2997–3006
- Stein CA, Stein S (1992) A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature* 359:123–129
- Stern TA, Henrys SA, Okaya D, Louie JN, Savage MK, Lamb S, Sato H, Sutherland R, Iwasaki T (2015) A seismic reflection image for the base of a tectonic plate. *Nature* 518:85–88
- Von Herzen RP (2004) Geothermal evidence for continuing hydrothermal circulation in older (>60 M.y.) oceanic crust. In: Davis EE,

- Elderfield H (eds) *Hydrogeology of the oceanic lithosphere*. Cambridge University Press, Cambridge, pp 414–447
- Von Herzen RP, Cordery MJ, Detrick RS, Fang C (1989) Heat flow and the thermal origin of hot spot swells: the Hawaiian swell revisited. *J Geophys Res* 94:13783–13799
- Watts AB (2007) Crust and lithosphere dynamics: an overview. *Treat Geophys* 7:1–53

Lithospheric Magnetic Anomalies from Satellite Data

Stavros Kotsiaros

Planetary Magnetospheres Lab, NASA Goddard Space Flight Center, Greenbelt, MD, USA

University of Maryland College Park, College Park, MD, USA

Technical University of Denmark (DTU), Kongens Lyngby, Denmark

Synonyms

Crustal (magnetic) field; Lithospheric (magnetic) field

Definition

Geomagnetic field is Earth's magnetic field. Units are nT (nanoTesla). The geomagnetic field at the Earth's surface has a magnitude of 25,000 nT to 65,000 nT. *LEO satellites* are spacecraft orbiting above Earth's surface with an altitude of less than 2000 km. Typical altitudes of satellites measuring Earth's magnetic field are between 300 km and 700 km. *Magnetic observatories* are land-based stations collecting measurements of Earth's magnetic field. Magnetic observatories collect absolute measurements in contrast to variometer stations which measure magnetic field variations. *Geomagnetic storms* are perturbations in the Earth's magnetic field caused by temporary disturbances of Earth's magnetosphere usually caused by fluctuations in the stream of charged particles originating from the sun known as the solar wind. *Space weather* refers to the electromagnetic conditions in Earth's space environment such as in the solar wind and within the magnetosphere and ionosphere.

Introduction

The Earth possesses a rich magnetic field generated mainly by the motion of conducting metals, mostly molten iron and nickel, in our planet's liquid (outer) core. The liquid core is found in a layer between ~2900 km and ~5150 km (Busse

2011) deep inside Earth. This self-sustaining process which maintains Earth's magnetic field is known as the geodynamo and produces a field with a strength of 25,000 nT near the equator and 65,000 nT near the poles making up about 95% of Earth's total magnetic field. The remaining few percent is produced by magnetized material in the Earth's lithosphere as well as by electric currents in the ionosphere at about 90 km to 1000 km altitude and magnetosphere, which extends to distances larger than several thousands of km, creating fields ranging from a few nT to thousands of nT at polar latitudes (Olsen and Stolle 2012). And lastly, a very small fraction is produced by induced electric currents in the Earth's mantle, lithosphere, and ocean creating fields of a few nT.

The Earth's crust and the underlying uppermost part of the mantle make up the lithosphere which is about 100 km thick (Langel and Hinze 1998). The lithospheric magnetic field exhibits high regional variability because it depends on the local geological history, petrochemical composition, and physical properties of the magnetized rocks of the region. It can locally reach magnitudes of several hundreds or even thousands of nT, and since it is a small fraction of Earth's total field, it is often referred to as the anomaly field and the associated magnetic signatures of the magnetized rocks are the lithospheric magnetic anomalies. Therefore, to determine the lithospheric field globally, detailed magnetic field observations with global coverage are needed. Low Earth Orbiting (LEO) satellites provide the most effective means of acquiring such observations and provide efficient coverage for the determination of Earth's lithospheric field on a global scale. As an example, Fig. 1 shows the global distribution of magnetic ground observatories (red dots) collecting measurements of Earth's magnetic field versus the measurements taken along a satellite orbit during 24 h (yellow ground track). It is clear that the present ground observatory network provides a sparse data distribution compared to the spatial coverage (even from only 24 h) of satellite observations. Continuous satellite data for several months or even years provide a dense net of magnetic field observations far superior to what is possible with observations close to Earth's surface. However, due to orbital mechanics, the orbits of LEO satellites used in geomagnetism are usually inclined, that is, the angle between the orbit plane and the equatorial plane is not exactly 90° (it is usually $\sim 90^\circ \pm 3^\circ$ to 6°), which results in leaving a small cap around the two geographic poles unsampled (see polar views on Fig. 1). These unsampled regions are known as "polar gaps."

The Lithospheric Field from Space

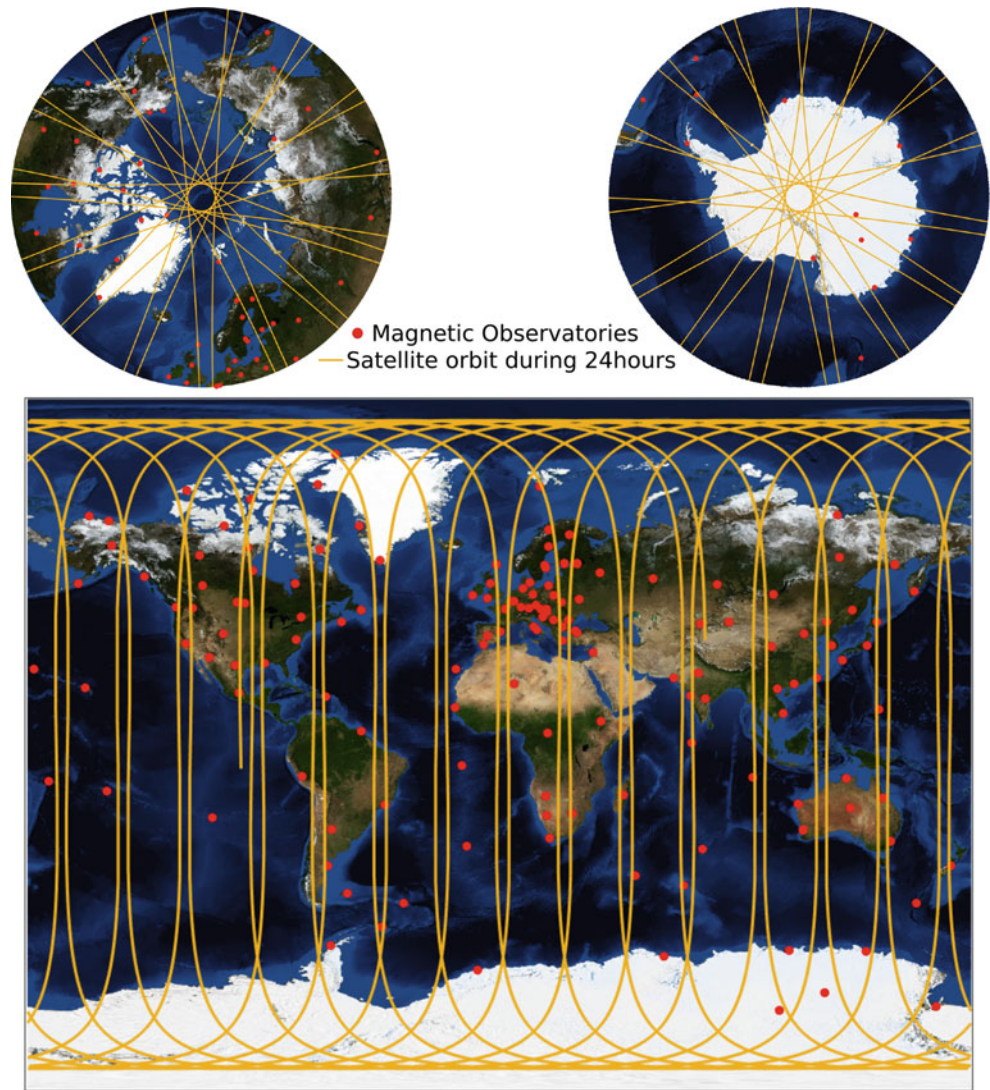
An extensive review of measuring the Earth's lithospheric field with satellites is given by Langel and Hinze (1998) from the Polar Orbiting Geophysical Observatories (POGO) launched between 1965 and 1971, until the Ørsted satellite

which was launched in 1999. POGO were the first satellites to provide global measurements of the Earth's magnetic field whereas Ørsted marked the start of the International Decade of Geopotential research, an international effort to promote and coordinate a continuous monitoring of the geopotential (magnetic and gravity) field variability in the near-Earth environment. Two missions followed after the launch of Ørsted which provided the most detailed data for lithospheric field modeling and enhanced our knowledge of the global contour of Earth's magnetized lithosphere. The first mission, CHAMP, was launched in 2003 and the second one, called Swarm, was launched in 2013. Ørsted, CHAMP, and Swarm have provided continuous measurements with an extended time-space coverage that helped improve our ability to characterize and understand the various sources that contribute to Earth's magnetic field. Currently, Swarm is the only operational mission dedicated to measuring Earth's magnetic field. It is a constellation of three identical satellites two of which are separated by ~ 150 km (at the equator and < 150 km in higher latitudes) in the east-west direction and orbit as a pair at lower altitudes (≤ 460 km) and a third flying at higher altitude (≤ 530 km) sampling different local times than the lower pair. The Swarm satellite mission provides a new dimension of detecting lithospheric anomalies from space. By using simultaneous (or near-simultaneous) measurements from the lower satellite pair, cross-track gradients (which approximate east-west gradients at nonpolar latitudes) of the magnetic field can be estimated. Gradients suppress the signals from external sources and specifically amplify the small-scale lithospheric anomalies. A full basis of magnetic space gradiometry, that is, the measurement of magnetic field gradients from satellites, and their advantages in the lithospheric field exploration is given by Kotsiaros and Olsen (2012). Specifically, east-west gradients contain valuable information on North-South oriented features of the lithospheric anomalies. Similarly, information on east-west features requires knowledge of the North-South gradient. North-South gradients can be approximated by along-track gradients (at nonpolar latitudes) which in turn can be estimated by finite differences along the satellite's track. However, on the contrary to cross-track gradients, along-track gradients are not instantaneous since they are estimated from measurements from a single satellite taken at different time instances. Therefore, the spacing of the along-track finite differences is important for maximizing the signal to noise ratio. Kotsiaros et al. (2015) has shown the optimal estimation of along-track gradients and their added benefit for lithospheric anomaly detection using data from the CHAMP mission.

Figure 2 shows how lithospheric anomalies are seen by (a) vector data (the radial component B_r) (b) across-track $\left(\frac{dB}{d\phi}\right)$, and (c) along-track $\left(\frac{dB}{d\theta}\right)$ gradients. While across-track gradients amplify north-south structures, along-track gradients amplify

Lithospheric Magnetic Anomalies from Satellite Data,

Fig. 1 Ground track of 24 h of a satellite orbiting at ~450 km altitude (yellow curves). The red dots indicate the locations of ground-based magnetic observatories. (Source of Earth's background image: NASA)



east-west structures. Evidently, both gradients contain additional information and show anomalies with a higher level of detail compared to vector data.

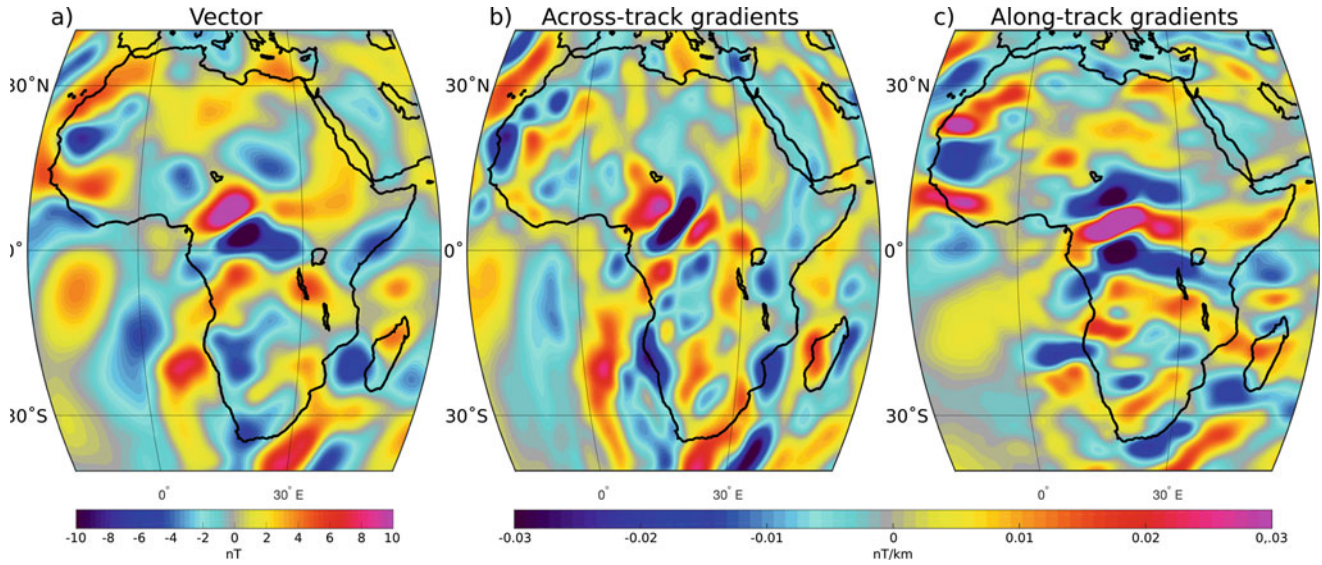
In general, both the magnitude and spatial scale of the lithospheric anomaly signals are attenuated at satellite altitude. The process of propagating the measured anomaly signals from satellite altitude down to the Earth's surface is called downward continuation. Although downward continuation is a very useful tool for estimating the magnetic field at the Earth's surface, systematic errors are also propagated and amplified together with the signal. This creates artifacts in the mapping of lithospheric anomalies.

Mathematical Background

The magnetic field measured by a satellite is a superposition of contributions from various sources and their separation based on magnetic field measurements is challenging. The

lithospheric magnetic field signal is therefore entangled with the signals originating from the Earth's core, ionosphere, and magnetosphere. The first attempt to separate the various magnetic field contributions was performed more than 100 years ago by Gauss (1877) (a revised translation from the German text is provided by Glassmeier and Tsurutani (2014)). Gauss introduced the concept of spherical harmonic analysis and the various sources can be separated based on the fact that each of their contributions has a specific spatial and temporal characteristic.

For example, with Earth's surface as reference, the magnetic field sources can be separated into "internal" originating in the Earth's core and lithosphere and "external" originating in the ionosphere and magnetosphere. The magnetic field due to internal sources varies as a function of $(\frac{a}{r})$, with $a = 6371.2$ km (the mean Earth's radius) and $r > a$ the radial distance of the measurement location, whereas the field due to external sources varies as a function of $\sim(\frac{r}{a})$.



Lithospheric Magnetic Anomalies from Satellite Data, Fig. 2 Lithospheric anomalies at satellite altitude (450 km) detected by (a) vector data (radial component, B_r), (b) across-track $\left(\frac{dB}{d\phi}\right)$, and (c) along-track $\left(\frac{dB}{d\theta}\right)$ gradients

In regions without electric currents, such as along most part of the satellite orbits, the magnetic field, $\mathbf{B}$, can be written as the (negative) gradient of a scalar potential V ,

$$\mathbf{B} = -\nabla V \quad (1)$$

and because of $\nabla \cdot \mathbf{B} = 0$, the potential V has to obey Laplace's equation: $\nabla^2 V = 0$. The scalar potential V can be split into an internal part V^{int} and an external part V^{ext} , which correspond to the internal and external magnetic field sources, that is,

$$V = V^{\text{int}} + V^{\text{ext}} \quad (2)$$

and from Eq. (1) we have:

$$\mathbf{B} = -\nabla V^{\text{int}} - \nabla V^{\text{ext}} = \mathbf{B}^{\text{int}} + \mathbf{B}^{\text{ext}}. \quad (3)$$

The internal and external potentials V^{int} and V^{ext} can be expanded in series of spherical harmonics (which are solutions of Laplace's equation):

$$V^{\text{int}} = a \sum_{n=1}^{N^{\text{int}}} \sum_{m=0}^n \times (g_n^m \cos m\phi + h_n^m \sin m\phi) \left(\frac{a}{r}\right)^{n+1} P_n^m(\cos \theta) \quad (4)$$

$$V^{\text{ext}} = a \sum_{n=1}^{N^{\text{ext}}} \sum_{m=0}^n \times (q_n^m \cos m\phi + s_n^m \sin m\phi) \left(\frac{r}{a}\right)^n P_n^m(\cos \theta), \quad (5)$$

where $a = 6371.2$ km is the Earth's mean radius, (r, θ, ϕ) are geocentric spherical coordinates, $P_n^m(\cos \theta)$ are the associated Schmidt semi-normalized Legendre functions of degree n and order m , and g_n^m , h_n^m and q_n^m , s_n^m are the spherical harmonic expansion coefficients describing the internal and external sources respectively. For more details see Chapman and Bartels (1940) and Langel (1987). Theoretically, the series extend up to infinite degree, but since in practice there is only a limited number of measurements at discrete points, the series are truncated at a maximum degree N^{int} (N^{ext}) for the expansion of the internal (external) potential. The estimation of the internal (or external) spherical harmonic coefficients g_n^m , h_n^m (or q_n^m , s_n^m) from discrete satellite measurements of $\mathbf{B}$ is known as spherical harmonic analysis and the set of the estimated coefficients is referred to as a (global) magnetic field model. From the estimated spherical harmonic coefficients (the model), we can calculate the magnetic field values $\mathbf{B}$ at any given point. This process is called spherical harmonic synthesis. LEO satellites that measure Earth's magnetic field orbit above the highly conducting layer of the ionosphere where the ionospheric currents flow (at about 110 km altitude) therefore they see the ionospheric field as an "internal" field. The separation of the lithospheric and ionospheric fields is challenging since both sources are internal to satellite observations. Therefore, lithospheric magnetic anomalies are often contaminated by ionospheric field signals in the satellite data. A standard technique to reduce contamination is to perform data selection and choose night time data when the conductivity of the ionosphere is low and the associated ionospheric magnetic field is significantly weak. An alternative, but also more complicated approach is to model the ionospheric magnetic field

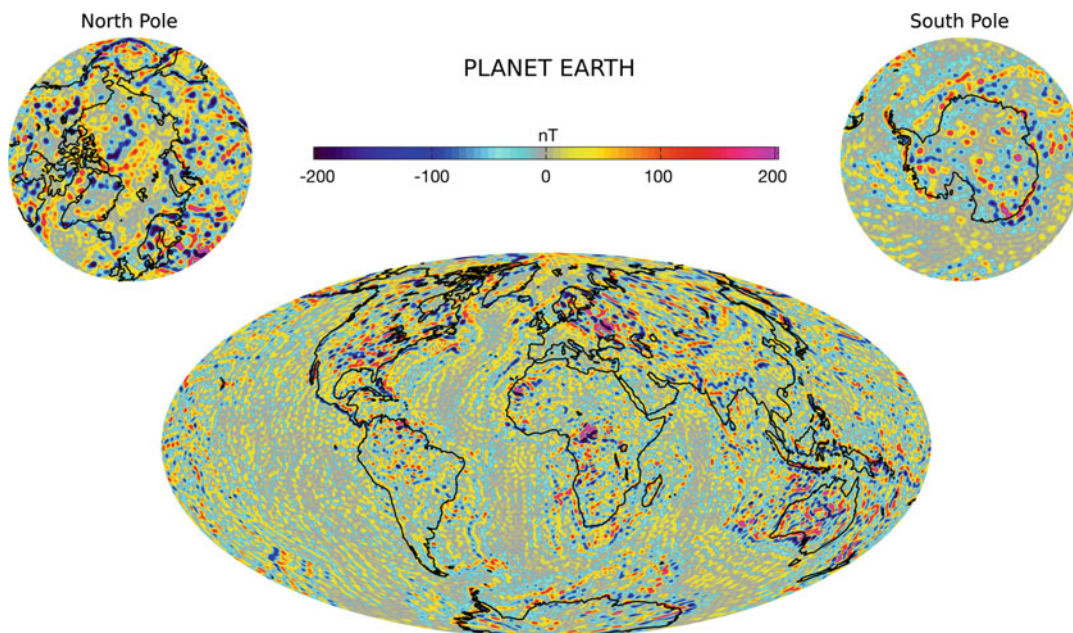
simultaneously to the lithospheric field, an approach commonly known as comprehensive modeling (e.g., Sabaka et al. 2004). Another difficulty is the amalgamation of the lithospheric field with that originating in the Earth's core. Specifically, at spatial scales of ~ 3000 km the core field overlaps the lithospheric field making the derivation of separate high-quality magnetic models for the core and lithosphere at those spatial scales very challenging. Various techniques for modeling the lithospheric field are reviewed by Thébault et al. (2010).

Global Models of Lithospheric Magnetic Anomalies

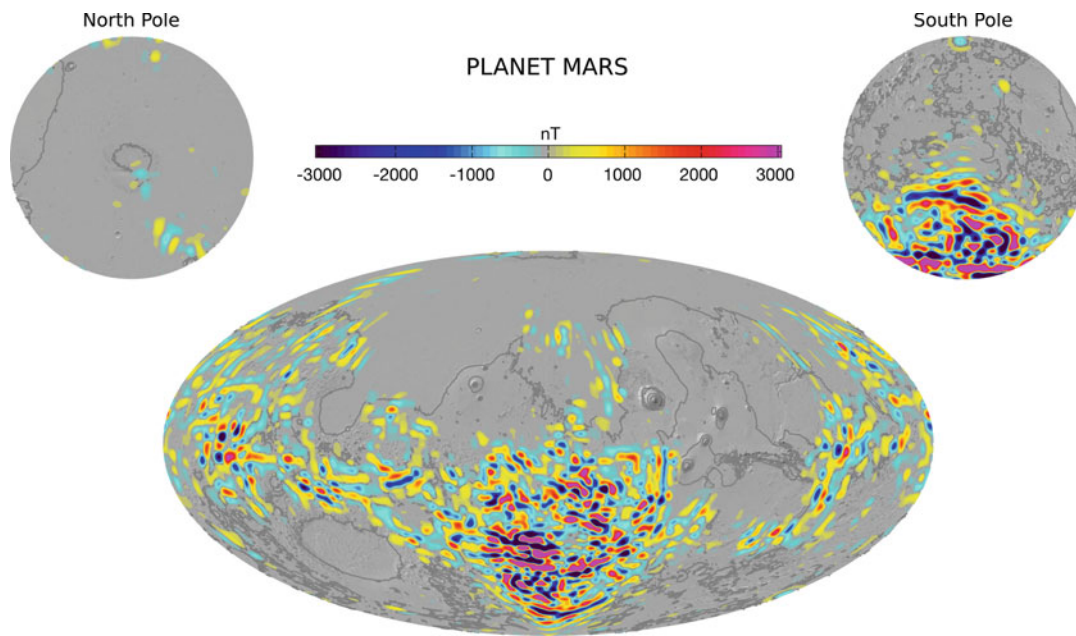
Global models of the Earth's magnetic field were derived from satellite observations more than 50 years ago using measurement of the field intensity. However, intensity measurements alone do not provide enough information to derive unique global models and vector observations are required. The CHAMP mission has provided high-quality vector data with increased sensitivity to Earth's lithospheric field due to CHAMP's low altitude (< 350 km) during the last 4 years of its mission. Detailed models able to capture lithospheric magnetic anomalies at spatial scales of ~ 300 km, corresponding to spherical harmonic degree 133, were derived using CHAMP data (e.g., Maus 2010). However, Swarm significantly improved the lithospheric field mapping from space. Models derived from Swarm data are able to capture anomalies at spatial scales ~ 216 km, corresponding to spherical harmonic degree 185 (e.g., Olsen et al. 2017). Despite the present higher

altitude (~ 450 km) of Swarm compared to CHAMP and therefore its lower sensitivity to the lithospheric magnetic anomalies, this resolution enhancement is possible mainly due to the east-west gradient information provided by the unique Swarm constellation. Upcoming Swarm data taken at lower altitudes will further increase the resolution of global lithospheric field models. Figure 3 shows the global lithospheric magnetic field (radial component) based on a model derived exclusively from Swarm data (Olsen et al. 2017). Earth's lithospheric anomalies are stronger on the continental regions where the lithosphere is thick and weaker in the oceanic regions where it is thin. A strong anomaly in central Africa, known as the Bangui anomaly, stands out. The exact origin of the Bangui anomaly remains unclear, although two theories have been suggested. One proposes a large igneous intrusion and the other a meteorite impact. The largest magnetic anomaly on Earth in south-west of Russia, known as the Kursk anomaly, is associated with massive iron-ore formations in this region.

Earth is not the only planet in our solar system that bears a lithospheric magnetic field. In fact, Mars has a very interesting lithospheric field. The Martian lithospheric anomalies are primarily concentrated in the southern hemisphere and are ~ 10 times stronger than Earth's lithospheric anomalies. The Martian lithospheric field is the best surveyed planetary lithospheric field from space after Earth's. Mars Global Surveyor (MGS), which was in orbit around the red planet during 1997–2006, and Mars Atmosphere and Volatile Evolution (MAVEN), which was launched in 2013 and is still in orbit, have acquired accurate magnetic field observations which are sufficient to derive detailed global models of the Martian



Lithospheric Magnetic Anomalies from Satellite Data, Fig. 3 Earth's lithospheric magnetic anomalies determined by satellite data



Lithospheric Magnetic Anomalies from Satellite Data, Fig. 4 Mars' lithospheric magnetic anomalies determined by satellite data

lithospheric field. Figure 4 shows the magnetic anomalies of the Martian lithosphere based on a model derived from MGS data. The peculiarity of the Martian lithospheric anomalies gathered in the southern hemisphere is evident. The source of the anomalies is either large volumes of magnetized material, very strong remanence, or both.

Summary

Continuous high-precision magnetic field observations taken by satellites have enabled us to make considerable progress in the study of the magnetization of Earth's lithosphere. In particular, the capability of Swarm satellites to estimate magnetic field gradients has provided the means for mapping Earth's lithospheric magnetic anomalies from space with remarkable detail. The Swarm mission is foreseen to carry on surveying Earth's magnetic field at least until the end of 2021 and possibly for several years after that. However, presently, there is no mission planned to continue the geomagnetic field exploration from space after Swarm.

Uninterrupted observations of the magnetic field are important in order to capture the temporal evolution of the field generated in the Earth's liquid core as well as to separate the different magnetic field sources. The separation of the sources is done physically with the help of data selection and mathematically with the help of spherical harmonic series based on the different temporal and spatial characteristics of

the internal and external field contributions. A significant challenge still remains at those spatial scales where the lithospheric field overlaps with the core field. A proper full separation of the core and lithospheric field could perhaps be achieved if knowledge of the full geological history and in-depth chemical composition of Earth's lithosphere was known which is of course unrealistic on a global scale and perhaps even on a local scale.

Satellite observations provide excellent spatial data coverage making it possible to derive global models of Earth's magnetic field. However, the worldwide network of magnetic observatories is providing magnetic measurements that are equally important despite their inferior spatial coverage compared to satellites. For example, data from magnetic observatories are used to derive indices which characterize Earth's external field activity and describe space weather related events such as geomagnetic storms. Data from magnetic observatories can be used to complement satellite data for modeling Earth's magnetic field.

Cross-References

- ▶ [Geomagnetic Field, IGRF](#)
- ▶ [Geomagnetic Field, Measurement Techniques](#)
- ▶ [Geomagnetic Field, Theory](#)
- ▶ [Magnetic Anomalies: Interpretation](#)
- ▶ [Magnetic Gradiometry](#)
- ▶ [Magnetic Methods, Principles](#)

Bibliography

- Busse FH (2011) Geomagnetic field, theory. Springer Netherlands, Dordrecht, pp 394–401
- Chapman S, Bartels J (1940) Geomagnetism, vol I+II. Clarendon, Oxford/London
- Gauss CF (1877) Allgemeine Theorie des Erdmagnetismus. Springer Berlin Heidelberg, Berlin/Heidelberg, pp 119–193
- Glassmeier K-H, Tsurutani BT (2014) Carl Friedrich Gauss – general theory of terrestrial magnetism – a revised translation of the German text. *Hist Geo Space Sci* 5(1):11–62
- Kotsiaros S, Olsen N (2012) The geomagnetic field gradient tensor. *GEM Int J Geomath* 3:297–314
- Kotsiaros S, Finlay C, Olsen N (2015) Use of along-track magnetic field differences in lithospheric field modelling. *Geophys J Int* 200(2):878–887
- Langel RA (1987) The main field. In: Jacobs JA (ed) *Geomagnetism*, vol 1. Academic, London, pp 249–512
- Langel RA, Hinze WJ (1998) The magnetic field of the Earth's lithosphere: the satellite perspective. Cambridge University Press, Cambridge
- Maus S (2010) Magnetic field model MF7. Available at www.geomag.us/models/mf7/mf7.html
- Olsen N, Stolle C (2012) Satellite geomagnetism. *Annu Rev Earth Planet Sci* 40:441–465
- Olsen N, Ravat D, Finlay CC, Kother LK (2017) LCS-1: a high-resolution global model of the lithospheric magnetic field derived from CHAMP and Swarm satellite observations. *Geophys J Int* 211(3):1461–1477
- Sabaka TJ, Olsen N, Purucker ME (2004) Extending comprehensive models of the Earth's magnetic field with Ørsted and CHAMP data. *Geophys J Int* 159:521–547
- Thébault E, Purucker M, Whaler KA, Langlais B, Sabaka TJ (2010) The magnetic field of the Earth's lithosphere. *Space Sci Rev* 155(4-4):95–127

Magnetic Anisotropy

Leonardo Sagnotti

Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Definition

The magnetic properties of rocks and minerals are generally anisotropic, that is, they are directionally dependent. Magnetic anisotropy of minerals arises either from fundamental anisotropy in the crystal structure or from the shape of non-spherical ferromagnetic grains. The most common cause of magnetic anisotropy in rocks is the preferential distribution-orientation of the constituting minerals, in other words the rock fabric.

Introduction

Pioneering works on the magnetic anisotropy of rocks were carried out during the 1940s and 1950s (Ising 1942; Graham 1954). These authors first realized that magnetic methods may be used to characterize the preferred orientation of minerals within the rock samples. Ising studied varved clays in Sweden and noticed that the magnetic susceptibility was higher on the bedding plane than orthogonally to it. Graham recognized that the anisotropy of magnetic susceptibility may be regarded as a petrofabric element; he later extended the analysis to various sedimentary rocks of the Appalachian Mountains and pointed out the existence of distinct and systematic relationships of the magnetic fabric with structural setting (Graham 1966). The studies progressively developed in the following decades, and a first comprehensive review on magnetic anisotropy and its application in geology and geophysics was published by Hrouda (1982). Researches on magnetic anisotropy then gained widespread use and were extended to examine the fabric in a variety of sedimentary, igneous, and

metamorphic rocks (e.g., see reviews by Jackson 1991; Rochette et al. 1992; Tarling and Hrouda 1993; Borradaile and Jackson 2010; Parés 2015; Bilardello 2016). Presently, the study of the magnetic anisotropy of rocks is still one of the most promising research issues in the field of rock magnetism.

Theoretical Principles

Two principal mechanisms control the magnetic anisotropy of rocks: (1) lattice alignment of crystals with magnetocrystalline anisotropy and (2) shape alignment of ferromagnetic grains.

Magnetocrystalline Anisotropy

This is an intrinsic property of all minerals since it arises from the ordered structure of elementary particles in the crystal lattice; it is defined as the energy per atom needed to move the crystal magnetization from one crystallographic direction to another. Magnetocrystalline anisotropy thus reflects the dependence of the free energy of a magnetic system on the direction of magnetization. In the atomic structure of a crystal, the spin-orbit coupling between the magnetic spins and the lattice establish “easy” and “hard” directions of magnetization and determine the orientation of the spin moments with respect to the crystal structure. Within ferromagnetic crystals (the term ferromagnetic will be used in the following in its broader sense, that is, to indicate all minerals with ordered magnetic spin structures, including ferrimagnetic and antiferromagnetic ordering), the exchange energy in the $3d$ electronic orbitals of transition elements (e.g., Fe, Co, Ni) is, unlike the s orbitals, anisotropic. Hence, within these crystals, electron spins will be easier to coordinate in some directions than in others. Along these directions, the elemental electron spins align more readily, and the magnetization acquired in an applied field is greatest. In magnetite the direction of the lowest magnetocrystalline anisotropy energy (i.e., the easy axis of magnetization) is along the body diagonal, that

is, the [111] crystallographic direction. In hematite, magnetocrystalline anisotropy is very weak in the basal plane, but it is very high along the crystallographic c -axis, perpendicular to this plane. Consequently, hematite crystals may be easily magnetized by magnetic fields applied along the basal plane and are almost unaffected by fields applied perpendicular to it.

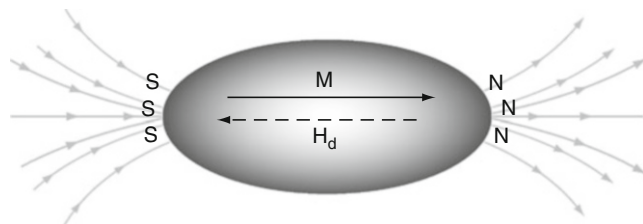
Shape Anisotropy

Shape anisotropy mostly occurs in ferromagnetic minerals with high intrinsic susceptibility and low degree of magnetocrystalline anisotropy (e.g., magnetite, maghemite), where just a slight deviation from the isometric shape produces notable magnetic anisotropies, in order to minimize the magnetostatic energy due to the formation of surface magnetic poles at the extremities of the mineral grains. In ferromagnetic minerals, the alignment of electron spins produced by exchange energy or by the application of an external magnetic field generates an external magnetic field that is proportional to the spontaneous or induced magnetic moment and may be represented as a field produced by an apparent distribution of north and south magnetic poles on the surface of the crystal. In turn, the surface magnetic poles generate an internal field whose field lines start on the north magnetic poles on the surface and end on the south poles (Fig. 1). This internal field is, therefore, antiparallel to the alignment of the magnetic moments within the crystal, and it is known as the demagnetizing field $\mathbf{H}_d$. Such a demagnetizing field is proportional to the magnetization $\mathbf{M}$ of the crystal and is sensitive to its shape:

$$\mathbf{H}_d = -N\mathbf{M}$$

where N is a demagnetizing factor which is determined by the shape of each crystal.

Shape anisotropy arises from the dependence of the demagnetization factor N on the direction of magnetization with respect to the particle shape. Different directions of magnetization results in different arrays of surface poles. N is a tensor depending on the grain shape: for a spherical grain, $N = 1/3$; for an ellipsoid grain with principal axes $a \geq b \geq c$, $N_a + N_b + N_c = 1$.



Magnetic Anisotropy, Fig. 1 Surface magnetic poles and internal demagnetizing field $\mathbf{H}_d$ for an ellipsoidal grain with uniform axial magnetization $\mathbf{M}$

By analogy with electrostatics, magnetostatic forces decrease with distance, and the internal demagnetizing field is reduced when the surface poles are furthest apart. In the case of an ellipsoid magnetized parallel to the elongation axis a , the demagnetizing field, which depends on $1/r^2$, is weaker than in the case of a sphere (i.e., $N_a < 1/3$). Similarly, if the ellipsoid is magnetized along the shortest axis c , the demagnetizing field is stronger (i.e., $N_c > 1/3$). In general, the demagnetizing factor N is minimum for very elongated (“pencil shape”) grains and maximum for platy grains (“disc shape”). In an ellipsoid of revolution, N_a approaches 0 for an infinite needle ($N_b = N_c = 1/2$), and N_c approaches 1 for a disc of infinite thinness ($N_a = N_b = 0$).

The demagnetizing factor N determines the intensity of the magnetic field that effectively acts ($\mathbf{H}_{\text{eff}}$) within a grain subjected to an external applied field $\mathbf{H}_{\text{ex}}$:

$$\mathbf{H}_{\text{eff}} = \mathbf{H}_{\text{ex}} - N\mathbf{M}$$

The magnetization $\mathbf{M}_j$ of a grain along an arbitrary direction j is given by

$$\mathbf{M}_j = k_o \mathbf{H}_{\text{ex}} = k_i \mathbf{H}_{\text{eff}} = k_i (\mathbf{H}_{\text{ex}} - N_j \mathbf{M}_j)$$

where k_o is the observed magnetic susceptibility along the direction j ; k_i is the intrinsic magnetic susceptibility of the grain, which is isotropic and dependent only on the chemical composition of the mineral; and $\mathbf{H}_{\text{ex}}$ is the intensity of the external magnetic field.

The observed magnetic susceptibility along a given direction of a grain is related to the intrinsic magnetic susceptibility of the mineral by

$$k_o = k_i / (1 + N_{ki})$$

For grains with weak intrinsic magnetic susceptibility ($k_i \ll 1$) – that is, for diamagnetic, paramagnetic, and anti-ferromagnetic particles – the Nk_i factor is negligible, so that the observed and the intrinsic magnetic susceptibility are practically equal ($k_o = k_i$). The magnetic anisotropy of these grains is controlled by crystallography. For grains with very high intrinsic magnetic susceptibility ($k_i \gg 1$) – such as for ferrimagnetic magnetite particles – k_o approaches the “self-demagnetization” limit value of $1/N$, and the magnetic anisotropy is controlled by particle shape.

The degree of magnetic susceptibility shape anisotropy in a grain is defined as

$$P = (1 + k_i N_c) / (1 + k_i N_a).$$

Therefore, the net effect of the shape anisotropy is to favor the orientation of the induced magnetization along the long

axis of a magnetic grain, where the internal demagnetizing field is minimized.

Magnetostriction: Stress Anisotropy

The application of an external stress to a rock may cause reversible as well as irreversible changes to its magnetic anisotropy. In fact, straining a ferromagnetic crystal can rotate its spontaneous magnetization away from the easy axis given by magnetocrystalline anisotropy. Moreover, the application of a magnetic field which rotates the spontaneous magnetization of a ferromagnetic crystal away from its intrinsic preferred direction can strain the lattice and modify the crystal shape, producing changes in its demagnetization factors and shape anisotropy. Both processes are expression of the phenomenon known as magnetostriction and are due to changes in the exchange energy when modifications in the magnetization or in the crystal shape alter the spin-orbit coupling in neighboring atoms.

Magnetostriction is defined as the spontaneous change in crystal dimensions that accompanies the process of magnetization or as the change in the magnetization of a crystal as a result of the application of stress. The linear saturation magnetostriction constant λ is defined as the fractional change in the length $\Delta L/L$ of a demagnetized ferromagnetic crystal as its magnetization increases from zero to saturation. Magnetostriction is positive when the crystal expands in the direction of magnetization. In magnetite, at room temperature, the magnetostriction depends crucially on the direction, being positive for a magnetization along a $\langle 111 \rangle$ easy axis but negative for magnetization along a $\langle 100 \rangle$ hard axis.

In a cubic crystal, the anisotropy of λ is given by (e.g., Dunlop and Özdemir 1997)

$$\lambda = 3/2 \lambda_{100} (\alpha_1^2 \gamma_1^2 + \alpha_2^2 \gamma_2^2 + \alpha_3^2 \gamma_3^2 - 2/3) + 3 \lambda_{111} (\alpha_1 \alpha_2 \gamma_1 \gamma_2 + \alpha_2 \alpha_3 \gamma_2 \gamma_3 + \alpha_3 \alpha_1 \gamma_3 \gamma_1),$$

where α_i ($i = 1, 2, 3$) are the direction cosines of saturation magnetization ($\mathbf{Ms}$) with respect to the $\langle 100 \rangle$, $\langle 010 \rangle$, and $\langle 001 \rangle$ crystallographic axes, respectively, and γ_i are the equivalent direction cosines of the direction along which the magnetostriction is being measured. λ_{100} and λ_{111} are the crystal's magnetostriction constants along the $\langle 100 \rangle$ and $\langle 111 \rangle$ crystallographic axes for $\mathbf{M}$ aligned in the same direction.

In the case of isotropic magnetostriction (i.e., $\lambda_s = \lambda_{100} = \lambda_{111}$), the magnetoelastic energy caused by the application of a stress σ to a magnetic crystal is given by

$$E = 3/2 \lambda_s \sigma \sin^2 \theta$$

where θ is the angle between $\mathbf{Ms}$ and σ .

In this simple case, the energy is minimum when $\mathbf{Ms}$ is parallel to σ and $\lambda_s \sigma > 0$ or when $\mathbf{Ms}$ is perpendicular to σ and $\lambda_s \sigma < 0$. Thus, the application of a uniaxial stress will cause $\mathbf{Ms}$ to rotate toward the stress axis if $\lambda_s \sigma > 0$ (i.e., under the combination of a compression and a negative λ_s or of a tension and positive λ_s).

In general, the relationship between the change in the magnetic anisotropy and the increase of strain induced by a stress field is not simple, reflecting the complex behavior due to the superposition of the strain effects upon an already magnetically anisotropic specimen.

Exchange Anisotropy

Exchange anisotropy results from superexchange interaction across the interface of two different magnetically ordered phases. In the simplest model, the origin of exchange anisotropy is considered to be the coupling of a ferromagnetic spin system to an antiferromagnetic spin system, separated by a planar interface. To obtain a preferred direction of the coupling, it is necessary that Curie temperature T_C of the ferromagnet is greater than the Néel temperature T_N of the antiferromagnet. When a magnetic field is applied at a temperature $T_N < T < T_C$, the spins of the ferromagnet will orient parallel to the applied field. Then, as the material is cooled through T_N , the spins of the antiferromagnetic lattice closest to the ferromagnet will align in the same direction as the ferromagnet, and subsequent spin planes in the antiferromagnetic lattice will orient antiparallel to each other. When the antiferromagnet is fully ordered, with high magnetocrystalline anisotropy, it holds the magnetization of the ferromagnetic material in the direction of the applied field, giving rise to a unidirectional magnetic anisotropy.

Exchange anisotropy can be detected from a shifted hysteresis loop when the material is cooled in a magnetic field and from the presence of a $\sin \theta$ term in torque curves measured in high fields due to a rotational hysteresis loss.

Exchange anisotropy in natural minerals has been found in titanomagnetite or in the intergrowth of maghemite with hematite and has been invoked to explain self-reversals in the direction of magnetization of natural rocks. In the last decades, there has been an increase of interest about this magnetic phenomenon because of the importance of its technological applications in spin-valve devices, such as hard disk read heads, and in magnetic thin layers.

Anisotropy of Magnetic Susceptibility (AMS)

Magnetic anisotropy of rocks is often determined by means of the analysis of the anisotropy of the magnetic susceptibility (AMS), which is the property that has found the most applications in geophysical studies so far. The anisotropy of magnetic susceptibility of rocks depends on mineral content,

their relative abundance, intrinsic susceptibility and anisotropy, preferential crystallographic orientation, and grain shape alignment. Therefore, knowing the composition of the rock-forming minerals and their magnetic anisotropy characteristics, it is possible to determine the spatial distribution of the grains and infer the geological processes that originated it. A preferential orientation-distribution of mineral grains is in fact typical of almost all rock types, and it develops during various geological processes, such as water flow in sediments, magma flow in igneous rocks, ductile deformation in metamorphic rocks, and even incipient strain in the paramagnetic clay matrix of apparently undeformed fine-grained sediments.

The magnetic susceptibility is the capability of a material to be magnetized under the effect of an external magnetic field. Being all the materials “susceptible” to become magnetized in the presence of an applied magnetic field, the magnetic susceptibility describes this transient magnetism within a material sample. If the magnetic field is relatively weak, the magnetization of a rock is a linear function of the intensity of this field. The low-field magnetic susceptibility is defined as the ratio of the induced magnetization ($\mathbf{M}$ dipole moment per unit volume or $\mathbf{J}$ dipole moment per unit mass) to the applied low-intensity magnetic field ($\mathbf{H}$). Only for isotropic substances, the induced magnetization is strictly parallel to the applied field, and the magnetic susceptibility is a scalar. In anisotropic media, like minerals and rocks, the induced magnetization is not parallel to the applied field, and the magnetization induced along the direction i is related to the magnetic field acting along the direction j by

$$\mathbf{J}_i = \chi_{ij} \mathbf{H}_j (\text{mass specific}) \quad \chi \text{ is given in SI units of m}^3/\text{kg}$$

or

$$\mathbf{M}_i = k_{ij} \mathbf{H}_j (\text{volume specific}) \quad k \text{ is dimensionless in SI units}$$

On the basis of the magnetic susceptibility, all substances may be classified as diamagnetic, paramagnetic, and ferromagnetic.

In diamagnetic minerals, the induced magnetization increases linearly with the increasing field, but in the opposite direction, the magnetic susceptibility is therefore negative and typically very low ($k \sim 10^{-5}$ SI). In paramagnetic minerals, the magnetization increases linearly with increasing field, along the same direction, and the susceptibility is positive ($k \sim 10^{-4} - 10^{-5}$ SI). In ferromagnetic minerals, the induced magnetization increases nonlinearly with increasing field, and the magnetic susceptibility is typically positive and much higher (k ranges from 10^{-3} to 10 SI) than in diamagnetic and paramagnetic minerals.

If the applied field is so low to approximate as linear the magnetic response of the body, the magnetic susceptibility

can be represented by a set of coefficients (k_{ij}) that form a second-order symmetric tensor. Each coefficient quantifies how much a body can be magnetized along a certain Cartesian direction according to a Cartesian component of the applied field. The relationship between the components of the magnetization $\mathbf{M}_i$ ($i = 1, 2, 3$) and the components of the magnetic field $\mathbf{H}_j$ ($j = 1, 2, 3$) is expressed by the equations

$$M_1 = k_{11}H_1 + k_{12}H_2 + k_{13}H_3$$

$$M_2 = k_{21}H_1 + k_{22}H_2 + k_{23}H_3$$

$$M_3 = k_{31}H_1 + k_{32}H_2 + k_{33}H_3$$

which can be rewritten in subscript notation as

$$\mathbf{M}_i = k_{ij} \mathbf{H}_j$$

where the coefficient k_{ij} defines a matrix that has six independent elements, since the second-order tensor must be symmetric to guarantee real eigenvalues, and k_{ij} is imposed equal to k_{ji} . Among all possible Cartesian reference systems, there exists one in which the non-diagonal terms of the tensor are zero so that the above equations simplify to

$$M_1 = k_{11}H_1$$

$$M_2 = k_{22}H_2$$

$$M_3 = k_{33}H_3$$

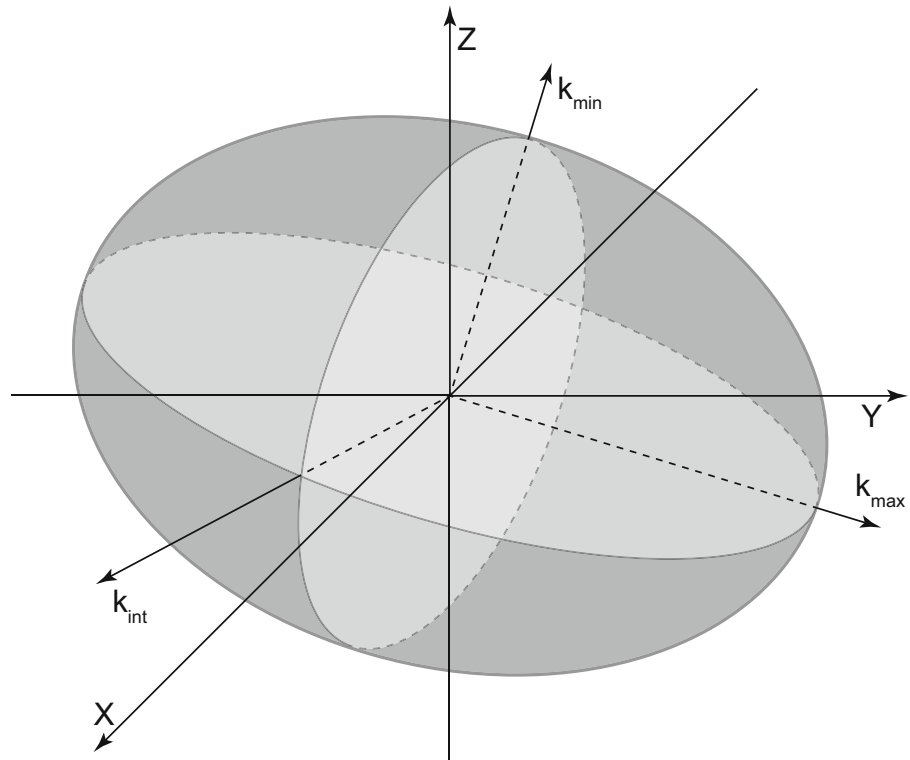
where in this case the three coefficients k_{11} , k_{22} , and k_{33} are the eigenvalues of the magnetic susceptibility tensor: they are called the principal susceptibilities and are generally indicated as $k_{\max} \geq k_{\text{int}} \geq k_{\min}$ (or $k_1 \geq k_2 \geq k_3$), the maximum, intermediate, and minimum susceptibilities.

The corresponding eigenvectors are the maximum, intermediate, minimum principal susceptibility directions that are the directions along which the induced magnetization is strictly parallel to the direction of the applied field. The magnetic susceptibility tensor may be represented geometrically by a triaxial ellipsoid, which is termed the magnitude ellipsoid, whose axes are parallel to the AMS tensor eigenvectors and whose semi-axes lengths are proportional to the AMS tensor eigenvalues, (Fig. 2). We will refer to this ellipsoid in the following as the AMS ellipsoid. Since for diamagnetic minerals it is possible to have negative AMS eigenvalues making the magnitude ellipsoid difficult to visualize, the magnetic susceptibility ellipsoid may also be expressed by the representation quadric, in which the length of the radius vector in any direction from the origin is equal to the inverse square root of the susceptibility in that direction. It has however a less direct relationship to the eigenvalues.

In practical measurements, the AMS eigenvectors are determined with two uncertainty angles, which define the regions where each principal susceptibility direction lies

Magnetic Anisotropy,

Fig. 2 The magnetic susceptibility ellipsoid. The AMS tensor may be geometrically represented by a triaxial ellipsoid, in which the three orthogonal axes correspond to the AMS eigenvectors, $k_{\max}$, k_{int} , and $k_{\min}$, respectively. The orientation of the AMS eigenvectors is then defined in a reference system of Cartesian coordinates (X, Y, and Z)



with a probability of 95% (Fig. 3). The definition of a magnetic susceptibility ellipsoid identifies two important elements: a “magnetic foliation” (the plane orthogonal to the direction of minimum magnetic susceptibility) and a “magnetic lineation” (the direction of maximum magnetic susceptibility).

AMS ellipsoid shapes are classified according to the relationships between the magnetic susceptibility eigenvalues:

$k_1 \approx k_2 \approx k_3$; isotropic susceptibility, the AMS ellipsoid is a sphere (Fig. 4a).

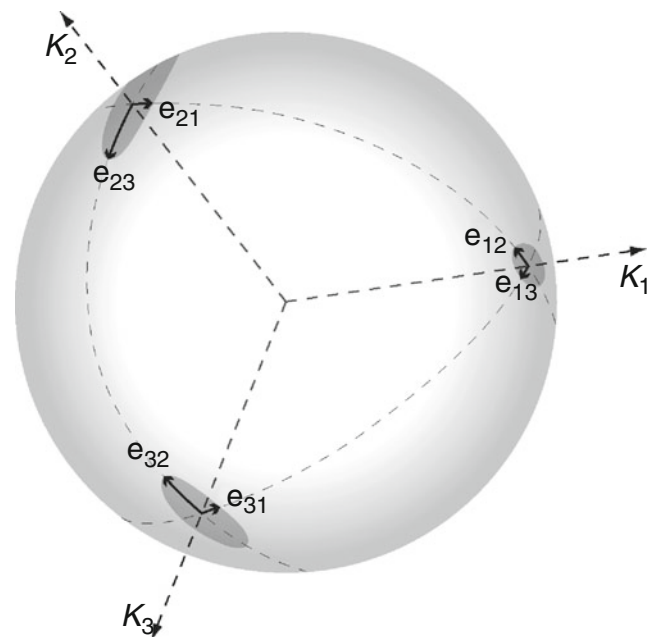
$k_3 \ll k_2 \approx k_1$; the AMS ellipsoid has an oblate shape (i.e., the magnetic fabric is planar, Fig. 4b).

$k_1 \gg k_2 \approx k_3$; the AMS ellipsoid has a prolate shape (i.e., the magnetic fabric is linear, Fig. 4c).

$k_1 > k_2 > k_3$; the AMS ellipsoid is triaxial (Fig. 4d).

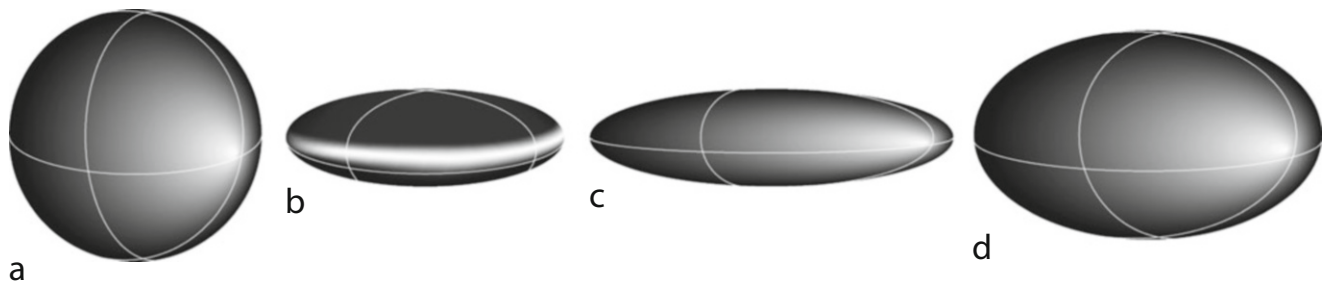
The orientations of principal susceptibilities for a set of samples are conventionally visualized in a Schmidt equal-area projection, lower hemisphere (Fig. 5).

The bulk low-field magnetic susceptibility of a rock is a summation of the contribution from all mineral species composing the rock, weighted according to their relative abundance and susceptibilities. Obviously, the contribution of ferromagnetic minerals to the overall magnetic susceptibility of a rock specimen is overwhelming upon those of the paramagnetic and diamagnetic matrix. The contribution of the matrix will be negligible when the percentage of ferromagnetic minerals exceeds 0.1% ($k > 3000\text{--}4000 \mu\text{SI}$). Conversely, for rocks in which $k < 200\text{--}300 \mu\text{SI}$, the magnetic

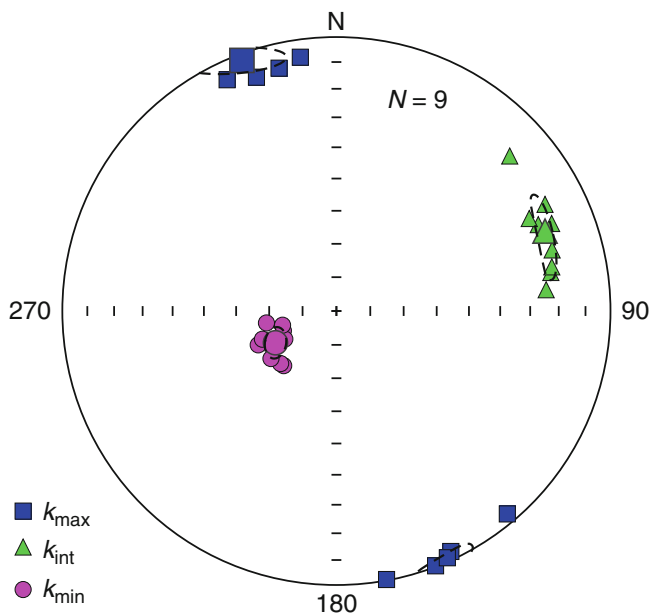


Magnetic Anisotropy, Fig. 3 Relationship of the 95% confidence ellipses (e_{12} , e_{23} , e_{13}) to the AMS eigenvectors. Around each principal susceptibility axis, the major and minor semi-axes of 95% confidence ellipse lay within the planes defined by the eigenvectors

susceptibility is almost completely controlled by the contribution of the paramagnetic matrix. For very small or even negative k values (i.e., $k < 50 \mu\text{SI}$), the contribution of the



Magnetic Anisotropy, Fig. 4 Shape of AMS ellipsoids. (a) Spherical; (b) oblate; (c) prolate, (d) triaxial



Magnetic Anisotropy, Fig. 5 Equal area projection of the principal AMS directions and the 95% confidence ellipses (dashed lines) for a single sampling site. Squares, $k_{\max}$; triangles, k_{int} ; circles, $k_{\min}$. Small symbols indicate individual specimens, and large symbols indicate the mean tensor

diamagnetic matrix to the bulk susceptibility of a rock specimen is not negligible.

All the rock's constituent mineral fractions contribute to the development of an AMS fabric depending on their intrinsic susceptibility, crystalline and shape anisotropy at the grain scale, and degree of preferential orientation-distribution.

Among ferromagnetic minerals, hematite and pyrrhotite are the most anisotropic, both with anisotropy degree $P > 100$. Magnetite shows a typical range of P variation between 1.1 for nearly isometric grains and 3 for highly anisometric grains (Table 1). Paramagnetic minerals are characterized by a much smaller AMS degree. In biotite and most phyllosilicates, the $k_{\max}$ and $k_{\min}$ directions are, respectively, parallel and orthogonal to the basal plane, and their AMS ellipsoid is typically oblate. Finally, the main diamagnetic minerals typically show low magnetic susceptibilities and

low anisotropy (Table 1). A recent review of single crystal magnetic anisotropy for silicate and carbonate rock-forming minerals, as well as for accessory magnetic minerals (i.e., iron oxides and sulfides), is provided by Biedermann (2018).

AMS is determined from a number of susceptibility measurements taken accordingly to specifically designed sets of directions in magnetic susceptibility bridges. Commonly used susceptibility bridges operate with alternating current (AC). AMS can be measured manually or utilizing an automated sample handler. The choice of measurement axes for the determination of AMS is critical. A minimum of six measurements is necessary to solve for the six unknown components of the AMS tensor. A 15-measurement position scheme is typically utilized for AMS measurements in manual mode. New 3D automatic specimen rotators instead require the user to insert the specimen in one position only, and the specimen is then automatically rotated around two axes, taking a total of 320 measurements.

With the aim to enhance the paramagnetic contribution to the magnetic susceptibility and fabric, AMS measurements may be carried out also on samples kept at low temperatures (e.g., at liquid nitrogen temperature, ~ 77 K). This enhancement occurs because diamagnetic and, to the first order, ferromagnetic susceptibilities are temperature independent, whereas paramagnetic susceptibility increases by decreasing absolute temperature T according to the Curie–Weiss law, $k_{\text{para}} = C/(T - \Theta)$, where C is a constant value and Θ is the paramagnetic Curie temperature.

Recent instrumental implementation in susceptibility bridges allows the measurement of magnetic susceptibility and AMS in variable fields and frequency, as well as the simultaneous measurement of both the in-phase and out-of-phase susceptibilities and their anisotropies (Hrouda et al. 2017). Consequently, analytical techniques have been developed for direct determination of the magnetic subfabrics of magnetic minerals that exhibit low-field variation of susceptibility, frequency-dependent susceptibility, or measurable out-of-phase susceptibility, such as pyrrhotite, titanomagnetite, and magnetically viscous ultrafine superparamagnetic grains of magnetite or maghemite (Hrouda et al. 2017).

Magnetic Anisotropy, Table 1 Grain AMS determinations reported for various common minerals

Mineral	k (SI)	P	T	AMS type
<i>Ferromagnetic (s.l.)</i>				
Magnetite	2.0–5.7	1.1–3.0	Variable	Shape
Hematite	$1 \times 10^{-3} \div 5 \times 10^{-2}$	> 100	~ 1	Magnetocrystalline
Pyrrhotite	$5 \times 10^{-2} \div 3 \times 10^{-1}$	100–10000	~ 1	Magnetocrystalline
<i>Paramagnetic</i>				
Biotite	$1.0 \div 2.8 (\times 10^{-3})$	$1.2 \div 1.6$	$0.9 \div 1$	Magnetocrystalline
Muscovite	$0.12 \div 0.17 (\times 10^{-3})$	$1.3 \div 1.4$	$0.4 \div 0.7$	Magnetocrystalline
Phlogopite	$0.3 \div 1.2 (\times 10^{-3})$	$1.3 \div 1.5$	$0.78 \div 0.95$	Magnetocrystalline
Chlorite	$0.07 \div 1.55 (\times 10^{-3})$	$1.2 \div 1.7$	$0.3 \div 0.7$	Magnetocrystalline
Hornblende	$8.92 (\times 10^{-3})$	1.67	–0.51	Magnetocrystalline
Tourmaline	$1.69 (\times 10^{-3})$	1.005	~ 1	Magnetocrystalline
Siderite	$3.98 (\times 10^{-3})$	1.56	–0.9	Magnetocrystalline
<i>Diamagnetic</i>				
Quartz	$-13.4 \div -15.4 (\times 10^{-6})$	1.01	–	Magnetocrystalline
Calcite	$-13.8 (\times 10^{-6})$	1.11	1	Magnetocrystalline
Aragonite	$-19 (\times 10^{-6})$	1.15	0.8	Magnetocrystalline

Adapted from Hrouda (1993, 2007); Tarling and Hrouda (1993), and Borradaile and Jackson (2004)

In comparison to other methods of petrofabric analysis, the study of AMS has the advantage to be quick, cost-effective, and highly sensitive, and above all, it can be performed systematically on all rock types. AMS data are usually employed to obtain information about the geological processes responsible of the fabric in rocks, both during lithogenesis and during any subsequent deformational event.

AMS: Applications in Geology and Geophysics

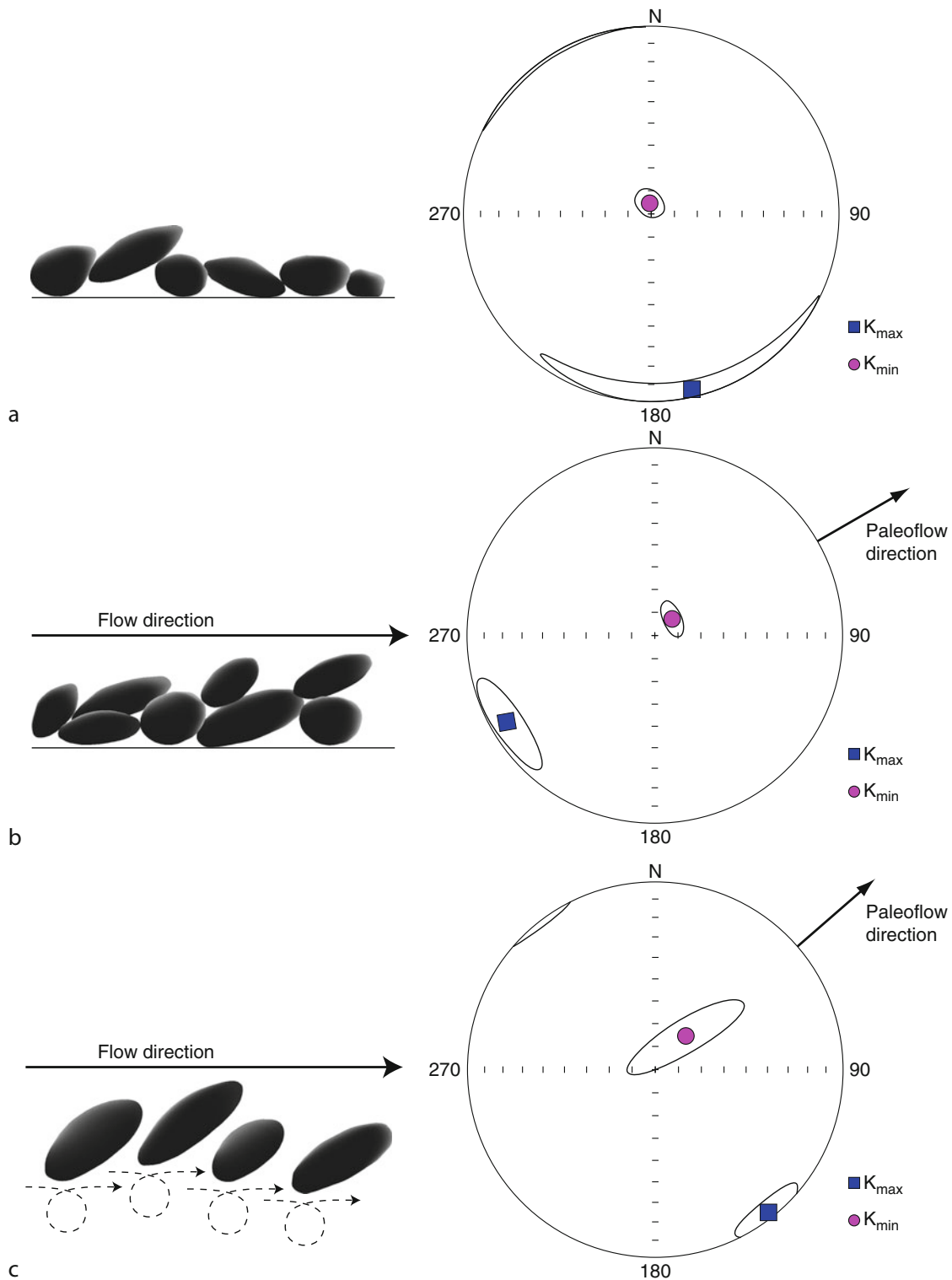
Magnetic anisotropy is an essential petrostructural phenomenon which can be used to study rock fabric, deformation, and various other processes that took place during the geologic history of a rock. The AMS of rocks can be determined with high accuracy, and the method is so sensitive that in rocks with a very weak preferred orientation-distribution of minerals, it is the only approach that gives reasonable results. The study of AMS has been therefore used as a tool of structural analysis for almost all rock types, and AMS analyses were applied to investigate a wide variety of geological processes capable to produce a preferential orientation-distribution of minerals in rocks. Moreover, AMS studies are an important complement to paleomagnetic researches, as a way to assess the reliability of paleomagnetic data and to determine possible deflections in the natural remanent magnetization components induced by deformation and compaction. For these reasons, the analysis of the AMS has experienced broad use in many branches of the Earth Sciences.

A brief overview of the wide spectrum of potential applications of the AMS research is provided below.

In undeformed **sediments**, the magnetic anisotropy may indicate the orientation of the sedimentation paleo-horizontal and eventual paleo-flow directions. The primary magnetic

fabric of sediments is strongly affected by the processes which govern deposition and compaction. In quiet water deposition, such as in a lagoon or lake environment, gravity-driven sedimentation brings platy grains to lay with their longer dimensions statistically parallel to the bedding-compaction plane. Then, with further sediment burial, the effect of diagenetic compaction on platy minerals by pressure and water expulsion reinforces the parallelism between the magnetic foliation and the bedding plane.

Thus, if the magnetic fabric is purely depositional, or related only to compaction loading, the AMS ellipsoid shows an oblate shape, with the $k_{\min}$ axis perpendicular to the bedding plane (which can be determined through AMS analyses with high accuracy) and the k_{int} and $k_{\max}$ axes mostly scattered in the bedding plane itself resulting in a wide uncertainty ellipse (Fig. 6a). The action of flow currents may cause the maximum susceptibility axis to be aligned either parallel or perpendicular to the paleocurrent direction (e.g., Hamilton and Rees 1970). In moderate currents, platy particles tend to imbricate, and elongated particles tend to align with their greatest axis parallel to the flow. On the AMS results obtained from a group of specimens, the imbrication induces a slightly vertical offset of the mean $k_{\min}$ and the clustering of the $k_{\max}$ axes in a direction antiparallel to the paleo-flow (Fig. 6b); the overall AMS fabric is still characterized by an oblate ellipsoid. In such a case, imbrication yields a better estimate of the flow direction than magnetic lineation, which may wander within the foliation plane and even be orthogonal to the flow direction. In fact, when deposition occurs under high current flow (traction transport mode of sedimentation), the distribution of $k_{\min}$ axes from various samples is streaked, the $k_{\max}$ axes cluster in a direction perpendicular to the flow (Fig. 6c),



Magnetic Anisotropy, Fig. 6 Schematic representations of the AMS development for undeformed sediments deposited in quiet conditions (a) under the action of weak or moderate water currents (b) and under the

action of strong currents, sufficient to entrain particles and to induce the rolling of elongated grains with their long axis perpendicular to the water flow (c). (Modified from Tauxe 2005)

and the fabric is characterized by prolate or triaxial AMS ellipsoids. The study of AMS was also applied in aeolian deposits of Chinese loess to reconstruct paleo-wind directions

and in recent and contemporaneous marine sediments to reconstruct the dynamics of deposition following tsunami events.

In general, sedimentary depositional fabrics are comparatively low both in susceptibility ($<10^{-4}$ SI) and in anisotropy ($P < 1.05$).

AMS measurements have been also combined with measurements of the anisotropy of P-waves velocity with the aim to analyze the 3D geometry of the pore space and to infer the anisotropy of permeability in porous rocks.

In **igneous rocks**, AMS analyses are mostly used to determine the orientation of the plane and flux direction of the magma flows. In these rocks, the viscosity of the magma plays a fundamental role for the genesis of a rock fabric. The movement/rotation of magnetic minerals is hindered when viscosity reaches a level which impedes any further ductile flow.

It is useful to distinguish the case of extrusive and shallow intrusive rocks (lava flows, ignimbrites, and dykes) from plutonic rocks. In volcanic rocks, the processes which induce a preferential orientation of the magnetic minerals (mostly titanomagnetites) are connected to the plane and direction of magma flow, both for lavas from effusive eruptions, for pyroclastic flows from explosive volcanism, and for dykes and other shallow tabular intrusive rocks (e.g., Cañón-Tapia 2004). Titanomagnetite grains are oriented in the flow plane with their greatest dimension parallel to the magma flow, so that for lava flows the magnetic foliation matches the (paleo) topographic surface, whereas for dykes or sills it develops parallel to the walls of the hosting rocks. The magnetic lineation is mostly parallel to the magma flow directions, even though perpendicular or oblique relationships can also be rarely found, especially in dykes. The efficiency of the volcanic flow to impart a preferential orientation on magnetic grains is usually relatively low, causing rather low anisotropy degrees, comparable to those typical of undeformed sediments.

A random fabric may be an indication of turbulent motion, which for dykes and sills occurs in the case of injection along wide enough fractures and for ignimbrites and surges occurs at a short distance from the vent where the eruption energy causes a chaotic motion of the transported solid particles. A coherent fabric may develop however at farther distance from the eruption centers where the kinetic energy decreases and the particles deposit on the morphological paleo-surfaces.

As for undeformed sediments, an eventual imbrication of the magnetic grains may be used to reconstruct the sense of the flow. In fact, the normal fabric of pyroclastic flows is considered the one with the foliation plunging upflow, so that the flow direction is given by the azimuth of the foliation pole and the lineation is close to it. AMS analyses in volcanic rocks have been used to constrain the location of the eruption centers.

In plutonic rocks, the analysis of the AMS may provide information on the processes of magma injection, emplacement, differentiation, and crystallization. The AMS fabric may therefore be affected by the flow of the magma, the

changes in its viscosity and composition, and the finite deformation it undergoes under regional stress before complete crystallization. Generally, the AMS of plutonic rocks whose magnetic fabric was produced through magma flow are characterized by a low anisotropy degree which overlaps with that of volcanic rocks and undeformed sediments. The shape of their AMS ellipsoid ranges from oblate to prolate according to the local character of the magma flow; the magnetic foliation is parallel to the flow plane, and the magnetic lineation is parallel to the flow direction. AMS fabrics of plutonic rocks have the potential to determine syn-magmatic deformation and to distinguish between compressional and extensional tectonic settings acting during emplacement. Conversely, the AMS of plutonic rocks which cooled under high temperature and pressures in the lower levels of the crust and are associated with metamorphic rocks may have a high anisotropy degree mostly determined by ductile deformation, which is a relatively efficient mechanism for the reorientation of magnetic minerals. The magnetic fabric in these rocks usually indicates a preferred orientation of magnetite minerals that nucleated at a late stage of the geological history of the pluton, and are therefore younger than the surrounding silicates. These magnetite grains grew in highly anisometric intergranular spaces and show therefore a high degree of shape anisotropy. This magnetic fabric is usually termed as mimetic magnetic fabric.

In **deformed rocks** AMS analyses are used to evaluate the orientation of the strain ellipsoid (e.g., Borradaile and Jackson 2010). In fact, AMS analysis is a very sensitive tool for the detection of rock strain, even in sediments that seem to have undergone weak or no deformation and where other strain markers are scarcely developed or even totally absent.

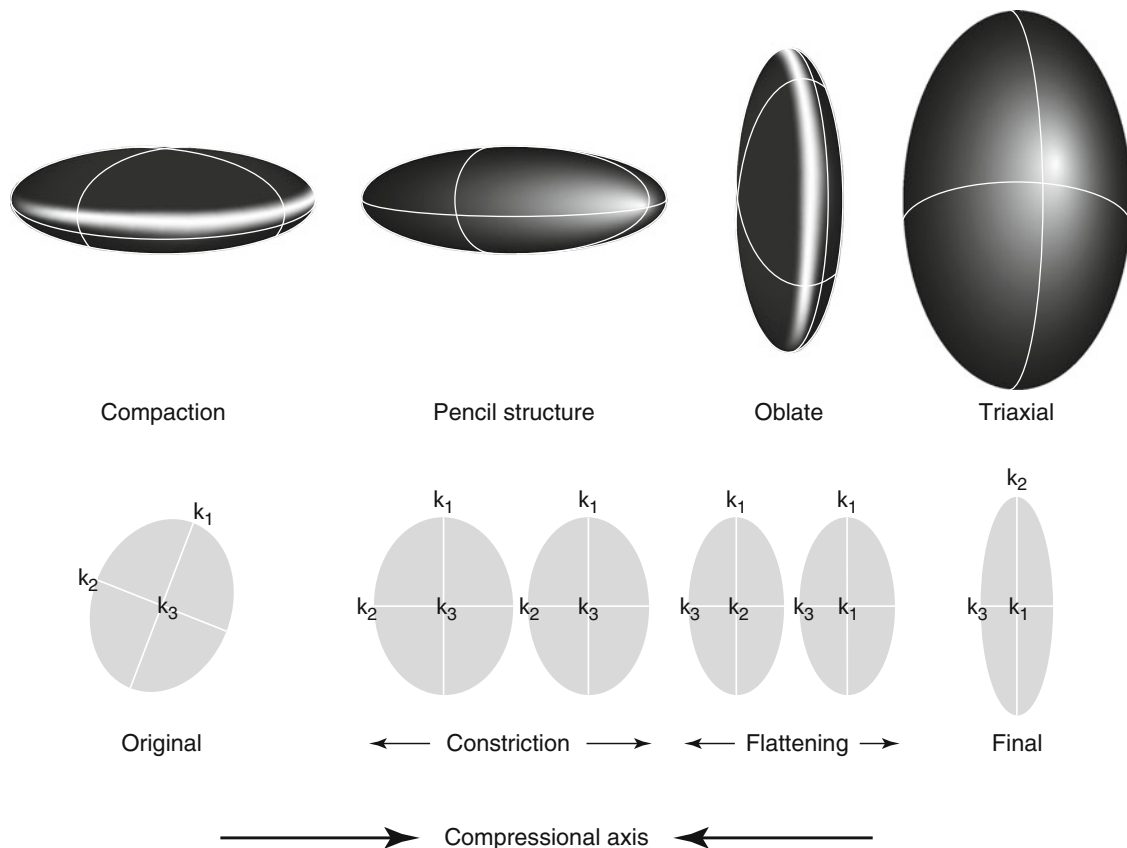
AMS studies in nearly undeformed sediments have also been used as a tool for recognizing intervals affected by significant coring-induced deformation and assess the reliability of paleomagnetic data derived from such sediments.

AMS is extremely sensitive to incipient strain, and it develops well before other macro- and mesoscopic strain features, such as cleavage, can be observed. As a matter of fact, the preferred orientation of phyllosilicate grains in mudstones may reflect the strain produced by compaction and tectonic processes. If there is a stress field acting on a sediment, causing progressive strain, the primary magnetic fabric is modified according to the nature and the extent of the deformation (e.g., Graham 1966; Parés 2015; Bilardello 2016) (Fig. 7). The first effect of strain upon such undeformed-compaction AMS fabric is to induce the clustering of the $k_{\max}$ axes in the direction of maximum stretching, that is, roughly perpendicular to the maximum compressive stress. The rotation of platy grains around the intermediate strain axis makes the planar bedding fabric less pronounced. Anyway, at such very first stage of deformation, the $k_{\min}$ axes are still normal to the bedding. AMS studies carried in clayey and weakly deformed sediments pointed out a

significant relationship between the magnetic fabric of apparently undeformed clays and the tectonic history of the sedimentary basins in which they were deposited, including compressional as well as extensional tectonic settings (e.g., Mattei et al. 1997). In such sediments, the magnetic foliation generally indicates the sedimentation-compaction plane, whereas the magnetic lineation is perpendicular to the shortening direction in the compressional settings and parallel to the stretching direction in the extensional settings. The magnetic lineation found in macroscopically undeformed fine-grained sediments from extensional settings develops from the spatial distribution of the basal planes of the paramagnetic phyllosilicates of the clay matrix and is parallel to the common axis of differently oriented basal planes (e.g., Cifelli et al. 2005). A further increase of strain under compression causes the dispersion (scattering) of the $k_{\min}$ axes in such a way that they tend to describe a girdle normal to the $k_{\max}$ cluster: the AMS ellipsoid becomes prolate, and the primary depositional fabric is progressively obliterated and overprinted by a tectonic fabric (Fig. 7). Mineral rotations are sufficient at this stage to produce a fabric strong enough to impart the well-known “pencil structure.” Further increasing strain occurs entering in the field of metamorphism and leads to the

progressive return of the AMS ellipsoid into the flattening field. Mineral rotation leads to the production of an imperfectly formed cleavage crossing bedding, called embryonic cleavage. A cleavage stage is achieved when the tectonic strain imprint is sufficiently strong to form a dominant planar cleavage fabric, where platy and acicular minerals are oriented in this cleavage plane. The magnetic foliation plane which develops at this stage is parallel to the plane of cleavage or schistosity visible in the rock. With increasing strain, the planar cleavage and lineation are intensified, this last one very close to or parallel to the strain maximum stretching direction (Fig. 7). Recent studies extended the range of magnetic fabric investigations in deformed rocks to address a wide variety of other geological problems, such as the mechanism of emplacement of large salt bodies, the cumulative strain due to glacier flow in tills and glacioteconites, the estimation of the shock pressure in weakly shocked rocks from an impact crater, the mechanism of recrystallization, and penetrative deformation along major faults planes and cataclastic fault rocks in shallow faults exposed at the earth surface.

The AMS of **metamorphic rocks** is generally considerably higher than the AMS in undeformed sedimentary and volcanic rocks. As a matter of fact, the mechanisms of ductile



Magnetic Anisotropy, Fig. 7 Schematic representations of the sequence of AMS development in progressively deformed sediments in a compressional setting. (Redrawn after Graham 1966)

deformation and recrystallization operating during metamorphism (e.g., diffusion, crystal plasticity, particulate flow, neomineralization) may be very effective to determine a preferential orientation-distribution of magnetic grains. AMS has been investigated also in the thermometamorphic aureole rocks surrounding igneous plutons with the aim to reconstruct the finite strain patterns acquired at the peak of contact metamorphism and to constrain mechanisms of pluton emplacement. In rocks with strong tectonic fabrics, the principal directions of susceptibility and of finite strain coincide, and a quantitative genetic relationship between strain and AMS has been pointed out in some cases (see review by Borradaile and Jackson 2010). Nevertheless such relationships are not universal, and strain-AMS relationship or correlations found under certain circumstances and lithologies cannot be extrapolated to different situations.

A magnetic fabric is defined normal when the principal magnetic axes correspond one by one to the petrofabric axes. An inverse or intermediate magnetic fabric occurs when the magnetic susceptibility principal axes exchange their position. Inverse AMS fabrics are associated with the prevalence of certain magnetic minerals (see Rochette et al. 1992). Specifically, some minerals possess an inverse AMS fabric related to their shape (e.g., goethite and tourmaline) or to intracrystalline gliding during ductile deformation (e.g., siderite and other iron-bearing carbonates). Moreover, SD ferrimagnetic particles with prolate shape anisotropy may originate an inverse magnetic fabric as well. As a matter of fact, pseudo single domain (PSD) and multidomain (MD) particles of a ferrimagnetic mineral such as magnetite have a maximum susceptibility parallel to their long (easy) morphological axis, while a stable single domain (SD) magnetite particle, whose magnetization is always saturated along its long axis, has a maximum susceptibility value perpendicular to it. Even if inverse and anomalous AMS fabric are relatively rare in nature, in all cases when AMS directions cannot be correlated with obvious structural markers, the mineralogical source of the AMS should be closely investigated and the AMS compared with other types of magnetic anisotropy (i.e., the anisotropy of artificially induced magnetic remanences).

Anisotropy of the Magnetic Remanence (AMR)

AMS represents the sum of the susceptibility anisotropies of all the mineral components in a rock, including the diamagnetic and paramagnetic fractions. Studies on the anisotropy of the magnetic remanence (AMR) are directly relevant to paleomagnetic studies, since they only measure the anisotropy of the remanence carrying particles. AMR studies also have a large spectrum of application in geological sciences (e.g., see Jackson 1991).

In fact, AMS may be inappropriate for the purposes of correcting paleomagnetic data, since it depends critically on the size of the remanence carrying particles. One of the advantages of AMR is that it precludes the effect of inverse AMS fabric due to single domain magnetite, since the grain long axis is the axis of maximum remanence irrespective of particle size. The most common ARMs are anisotropy of anhysteretic remanent magnetization (AARM) and anisotropy of isothermal remanent magnetization (AIRM) (see review by Bilardello 2016).

AARM: Anisotropy of Anhysteretic Remanent Magnetization

The anisotropy of anhysteretic remanent magnetization (AARM) is the most commonly studied anisotropy of remanence. An anhysteretic remanent magnetization (ARM) is acquired by a sample when it is simultaneously exposed to an alternating magnetic field with decreasing amplitude (usually with a maximum peak in the range 10–100 mT) and a steady DC magnetic field (typically values close to the intensity of the geomagnetic field, e.g., about 50–100 μ T). Prior to acquisition of the ARM, the specimen should be demagnetized in an AF peak higher than that used for the ARM acquisition and its residual remanence measured as a baseline. Then the ARM is imparted and computed as the difference between the measured remanence and the baseline.

There are two principal advantages of using ARM for the characterization of the anisotropy of the remanence: (1) the acquisition curve of ARM versus DC field is expected to be linear in the range of low DC magnetic field commonly employed, and (2) a window of AC field can be used to impart ARM. In fact, by measuring different windows of coercivity (partial AARM), one can expect to characterize different subpopulations (size and shape) of ferromagnetic grains.

The ferromagnetic particles that define the AARM tensor may have a broad range of grain sizes.

Since there is a strong empirical correlation between partial anhysteretic remanent magnetization (pARM) and magnetite grain size (e.g., Dunlop and Özdemir 1997), it is possible to isolate the AARM tensor associated with distinct populations of grains by using partial rather than total anhysteretic remanence anisotropy. This can be done by imparting a magnetization to a specific particle fraction during anisotropy determinations.

Observed AARM indicates possible anisotropy in acquiring NRM. This provides a warning that the rock might not be an accurate paleomagnetic recorder. Measuring the AARM thus provides a fast and reliable method to correct paleomagnetic deviation in rocks. In sedimentary rocks, AARM has been proposed to provide a quantitative correction for inclination error and to correctly evaluate the amplitude of paleosecular variation recorded in volcanic rocks.

AIRM: Anisotropy of Isothermal Remanent Magnetization

Another type of magnetic anisotropy is the anisotropy of isothermal remanent magnetization (AIRM). In order to study the IRM anisotropy, a sample is successively subjected to a pulse magnetic field (with intensity generally of the order of 5 mT to 60 mT) in several appropriate directions, and the acquired IRM is measured. As for the study of the anisotropy of ARM, before the measuring process and after each magnetization step, the sample is AF demagnetized. Stronger magnetic fields (e.g., higher than 1.3 T) were used to produce IRMs in red beds where multiple magnetic minerals coexist, like hematite and magnetite: the IRMs magnetize all the phases present, and the subsequent AF treatment will only demagnetize the low-coercivity magnetite, allowing determination of the remanence anisotropy of high-coercivity hematite only. The determination of the IRM anisotropy makes it possible to study the ferromagnetic fraction of a rock separately from the contribution of the paramagnetic and diamagnetic matrix.

Although usually successful, AIRM risks nonlinear effects where strong magnetic field is required.

HFAMS: High-Field Anisotropy of Magnetic Susceptibility

High-field methods are used to separate the AMS due to ferromagnetic minerals from the AMS due to diamagnetic and paramagnetic minerals. These methods are based on the field dependence of the magnetic susceptibility and rely on the different responses of ferromagnetic and paramagnetic minerals to large applied fields, i.e., large enough to saturate the ferromagnetic mineral fraction. For rocks that have both ferromagnetic and paramagnetic minerals, changes in magnetization at fields greater than ferromagnetic saturation are due primarily to the paramagnetic susceptibility.

High-field magnetic anisotropy measurements have been made using torque magnetometers or a high-field, rotating sample, cryogenic magnetometer. Although these techniques have great potential to separate paramagnetic and ferromagnetic fabrics, torque magnetometers and high-field cryogenic magnetometers are not available in most magnetic laboratories, and HFAMS is not routinely applied.

Summary

Magnetic anisotropy is an essential property of rocks and minerals.

The anisotropy of magnetic susceptibility (AMS) of rock specimens can be determined with a high degree of precision,

and the method is sensitive, rapid, and nondestructive. The study of AMS is a fundamental tool of petrofabric analysis, which has the potential to recognize even subtle preferred orientations of mineral lattices or shapes and to infer the causative processes, ranging from the effects of paleocurrents in undeformed sediments to pervasive strain in deformed rocks. As such, it has a wide range of applicability in the Earth Sciences.

The anisotropy of the magnetic remanence (AMR) instead refers to the preferential orientation-distribution of the ferromagnetic grains only. The techniques to determine AMR are time-consuming because artificial remanences must be applied and then removed along each measurement axis. The study of AMR is more directly relevant to paleomagnetic researches since it may provide a tool to correct paleomagnetic deviations in all cases when the direction of the magnetic remanence is deflected from the direction of the magnetic field in which it was acquired.

Cross-References

- Curie Temperature
- Magnetic Domains
- Magnetic Methods, Principles
- Paleomagnetism, Measurement Techniques and Instrumentation
- Paleomagnetism, Principles
- Remanent Magnetism

Bibliography

- Biedermann AR (2018) Magnetic anisotropy in single crystals: a review. *Geosciences* 8:302. <https://doi.org/10.3390/geosciences8080302>
- Bilardello D (2016) Magnetic anisotropy: theory, instrumentation, and techniques. Reference module in Earth systems and environmental sciences, Elsevier, 08-Mar-2016. <https://doi.org/10.1016/B978-0-12-409548-9.09516-6>
- Borradaile GJ, Jackson M (2004) Anisotropy of magnetic susceptibility (AMS): magnetic petrofabrics of deformed rocks. In: Martín-Hernández F, Lüneburg C, Aubourg C, and Jackson M (eds) *Magnetic Fabric Methods and Applications*. Geological Society, London, Special Publications, 238:299–360
- Borradaile GJ, Jackson M (2010) Structural geology, petrofabrics and magnetic fabrics. *J Struct Geol* 32:1519–1551
- Cañón-Tapia E (2004) Anisotropy of magnetic susceptibility of lava flows and dykes: an historical account. In: Martín-Hernández F, Lüneburg C, Aubourg C, Jackson M (eds) *Magnetic fabric methods and applications*. Special Publications/Geological Society, 238. Geological Society, London, pp 205–225
- Cifelli F, Mattei M, Chadima M, Hirt AM, Hansen A (2005) The origin of the tectonic lineation in extensional basins: combined neutron texture and magnetic analysis on “undeformed” clays. *Earth Planet Sci Lett* 235:62–78

- Dunlop DJ, Özdemir Ö (1997) Rock magnetism. Fundamentals and frontiers. Cambridge University Press, Cambridge, UK, 573pp
- Graham JW (1954) Magnetic susceptibility anisotropy, an unexploited petrofabric element. *Bull Geol Soc Am* 65:1257–1258
- Graham JW (1966) Significance of magnetic anisotropy in Appalachian sedimentary rocks. In: Steinhardt JS, Smith TJ (eds) *The Earth beneath the continents*. Geophysical monograph series 10. American Geophysical Union, Washington, DC, pp 627–648
- Hamilton N, Rees AI (1970) The use of magnetic fabric in palaeocurrent estimation. In: Runcorn SK (ed) *Palaeogeophysics*. Academic, London, pp 445–463
- Hrouda F (1982) Magnetic anisotropy of rocks and its application in geology and geophysics. *Geophys Surv* 5:37–82
- Hrouda F (1993) Theoretical models of magnetic anisotropy to strain relationship revisited. *Phys Earth Planet Inter* 77:237–249
- Hrouda F (2007) Magnetic susceptibility, anisotropy. In: Gubbins D, Herrero-Bervera E (eds) *Encyclopedia of geomagnetism and paleomagnetism*. Springer, Dordrecht, 1054pp, pp 546–560
- Hrouda F, Chadima M, Ježek J, Pokorný J (2017) Anisotropy of out-of-phase magnetic susceptibility of rocks as a tool for direct determination of magnetic subfabrics of some minerals: an introductory study. *Geophys J Int* 208:385–402
- Ising E (1942) On the magnetic properties of varved clays. *Arkiv för Matematik, Astronomi och Fysik* 29A:1–37
- Jackson M (1991) Anisotropy of magnetic remanence: a brief review of mineralogical sources, physical origins, and geological applications, and comparison with susceptibility anisotropy. *Pure Appl Geophys* 136:1–28
- Mattei M, Sagnotti L, Faccenna C, Funiciello R (1997) Magnetic fabric of weakly deformed clay-rich sediments in the Italian peninsula: relationship with compressional and extensional tectonics. *Tectonophysics* 271:107–122
- Parés JM (2015) Sixty years of anisotropy of magnetic susceptibility in deformed sedimentary rock. *Front Earth Sci* 3(4):1–13
- Rochette P, Jackson MJ, Aubourg C (1992) Rock magnetism and the interpretation of anisotropy of magnetic susceptibility. *Rev Geophys* 30:209–226
- Tarling DH, Hrouda F (1993) *The magnetic anisotropy of rocks*. Chapman & Hall, London, 217pp
- Tauxe L (2005) Lectures in paleomagnetism. Available on line at <http://earthref.org/MAGIC/books/Tauxe/2005/>

Magnetic Anomalies: Interpretation

E. Thébault

Laboratoire Magma et Volcans, Université Clermont
Auvergne (UCA), UMR CNRS 6524, Campus des Cézeaux,
6, Avenue Blaise Pascal, 63178 AUBIERE, FRANCE

Definition

Magnetic anomaly. On Earth, magnetic anomalies are defined as the vector and/or scalar differences between the measured magnetic field and estimations of the primary Earth's core and external magnetic fields. Magnetic anomalies are commonly assumed to reflect the magnetic structures of the rocks lying in the Earth's crust and the upper part of the lithosphere.

Introduction

The Earth's magnetic field is the sum of fields of internal and external origins that operate on a wide range of spatial and temporal scales (Hulot et al. 2007). The three major magnetic field sources are the core field, which is generated by electrical currents in the liquid outer core, the external fields produced by currents in the ionosphere and the magnetosphere, and the crustal field, which results from magnetized materials in the outer layer of the Earth. An anomaly field after correction for core and external fields is assumed to reflect the magnetic field of the Earth's crust. This field is one of the weakest one detectable at the altitudes of low Earth orbit (LEO) satellites. Inferring it with accuracy depends on the operator's ability to model and then to correct for core and external fields considered as perturbation fields. When the total magnetic field is measured globally in the geocentric reference frame, along the orbits of satellites for example, it may be expressed in current-free regions as the vector $\mathbf{B} = -\nabla V(\mathbf{r}, \theta, \phi, t)$ of the scalar magnetic potential V that is expanded in terms of spherical harmonics (SH):

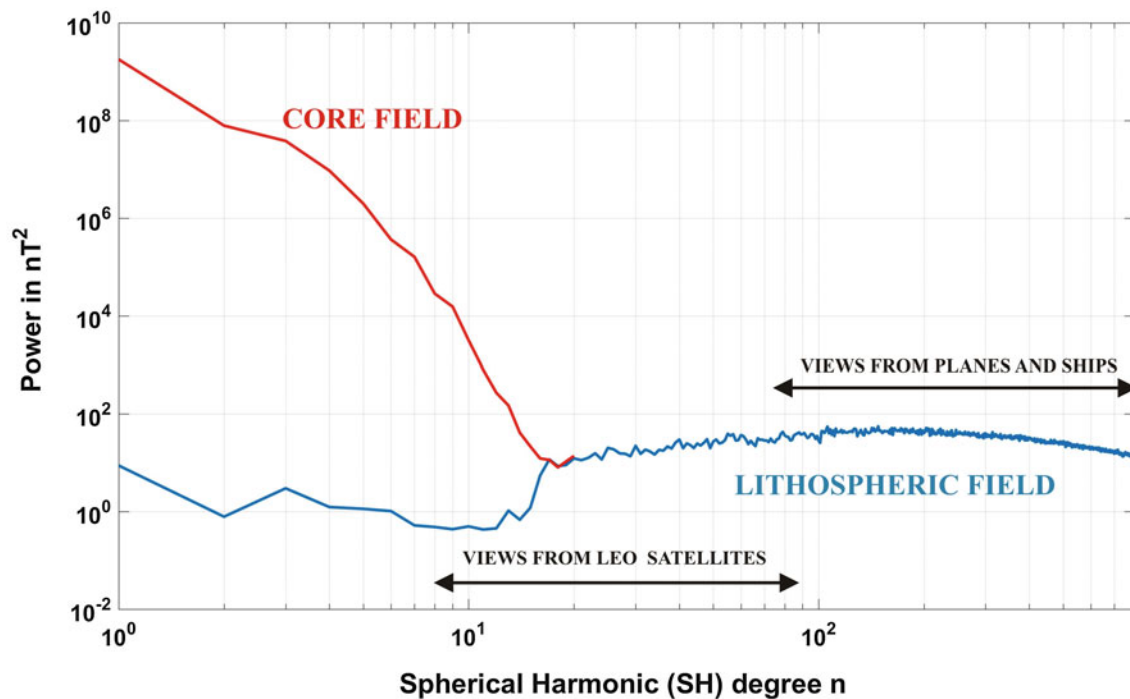
$$\begin{aligned} V(r, \theta, \phi, t) &= a \sum_{n=1}^{\infty} \sum_{m=0}^n \left(\frac{a}{r}\right)^{n+1} (g_n^m(t) \cos(m\varphi) + h_n^m(t) \sin(m\varphi)) P_n^m(\cos \theta) \\ &+ a \sum_{n=1}^{\infty} \sum_{m=0}^n \left(\frac{r}{a}\right)^n (q_n^m(t) \cos(m\varphi) + s_n^m(t) \sin(m\varphi)) P_n^m(\cos \theta). \end{aligned} \quad (1)$$

The parameters a , r , θ , and ϕ are, respectively, the Earth's reference radius (6371.2 km), the radius, the colatitude, and the longitude. The internal (g_n^m, h_n^m) and external (q_n^m, s_n^m) coefficients may be time varying and are expressed in the unit of nT with respect to the shell of radius a . In geomagnetism, the functions $P_n^m(\cos \theta)$ are the Schmidt semi-normalized-associated Legendre functions of integer degree n and order m . The spatial power spectrum in SH for the internal field writes (Lowes 1974)

$$R_n = (n+1) \left(\frac{a}{r}\right)^{2n+4} \sum_{m=0}^n [(g_n^m)^2 + (h_n^m)^2], \quad (2)$$

with R_n expressed in the unit of nT^2 . The power spectrum provides an estimate of the contribution of each SH degree n to the square amplitude of the observed magnetic field of internal origin at the radius r . The horizontal spatial wavelength λ in kilometers [km] at the Earth's mean radius a associated with each degree n is approximated by the relationship to the SH degree n (Backus et al. 1996, p. 101):

$$\lambda = \frac{2\pi a}{n}. \quad (3)$$



Magnetic Anomalies: Interpretation, Fig. 1 Geomagnetic field power spectrum to SH degree 800 at the mean Earth's radius. The core magnetic field (in red) is the IGRF predicted at epoch 2020.0 (Thébault

et al. 2015), and the lithospheric field (in blue) is computed from the WDMAM2.0. (After Lesur et al. 2016)

Figure 1 displays the power spectrum of an internal SH model expanded to SH degree $n = 800$ (Lesur et al. 2016). It includes contributions from the core and the crustal internal fields only. It is derived from the CHAMP and the Swarm trio of satellites, from airborne and ship measurements (► [Geomagnetic Field, Measurement Techniques](#)). The power spectrum shows a break around the SH degrees 13–15 that indicates the primary contribution of the long-term dynamo processes (the core field and its secular variation; see ► [Geomagnetic Field, Secular Variation](#)) at lower SH degrees. This contribution hides the crustal field structures from SH degrees 1 to 13–15. The magnetic field of the Earth's core is dynamic and more complex than a simple dipole (a simple dipole would concentrate all the energy on the SH degree $n = 1$). Many core field models based on different data, processing, and modeling techniques are available. The disagreements between the core field models for the same epoch lead to different pictures of the anomaly field. The most widely used core field model is the International Geomagnetic Reference Field (IGRF) model that was derived to SH degree 10 prior to epoch 2000.0 and to degree 13 thereafter (► [Geomagnetic Field, IGRF](#)).

The geometrical attenuation of SH harmonics with increasing altitudes is a major impediment to the detection of the lithospheric field at all scales, particularly the smallest ones. The fast coalescence of small-scale anomalies implies that magnetic measurements at different heights are not

sensitive to the same horizontal length scales. This calls for accurate core field models as well as techniques for removing the best estimates of external fields that could be quasiperiodic but are in general transient. At LEO satellite altitudes, a common practice is to select subsets of quiet-time data and to accumulate a large number of data over time in order to increase the signal-to-noise ratio and then to highlight the finest permanent crustal field structures. The spatial resolution of a magnetic anomaly thus depends on altitude and on the care taken to remove all field contributions not originating from the lithosphere. Its robustness is strongly related to the advances made toward analyzing and understanding the Earth's magnetic field as a whole. Correcting accurately for perturbation and transient external fields is the main cause of differences between lithospheric field models derived by independent groups and at different epochs, particularly those corresponding to periods of strong solar activity.

The Sources of Crustal and Lithospheric Fields

When the errors incurred in correcting the data for core and external fields are minimized, the anomaly field is assumed to have its sources buried in the crust (► [Continental Crustal Structure](#)) or deeper in the upper lithosphere although this remains a scientific debate (► [Lithosphere, Continental: Thermal Structure](#)).

The subject of rock magnetism is complex (Dunlop and Özdemir 2007), but the general mechanisms are well-known, thanks to laboratory experimental measurements made on rock samples. The amount of magnetization carried by a rock depends on several parameters such as its size, shape, mineralogy, grain domains, crustal thickness, and temperature (► [Magnetic Domains](#); ► [Remanent Magnetism](#)). Most magnetic minerals present on Earth are titanomagnetite and titanohematite, for which the Curie temperature is between 580 °C and 670 °C (► [Curie Temperature](#)). Considering the geothermal gradient of the Earth, the Curie temperature is met at about 30 km depth in stable continental regions and at 7 km depth in the oceans. These mean estimates vary significantly with the geographical location. Deeper rocks are thus considered nonmagnetic.

The minerals found in the Earth carry both permanent, $\mathbf{M}_r$ (► [Remanent Magnetism](#)), and induced magnetizations, $\mathbf{M}_i$. The magnetization, $\mathbf{M}_r$, arises when the magnetized grains of a hot material are locked in the direction of the Earth's global field when the temperature of an igneous rock decreases below the Curie temperature of its magnetic phase. The remanent magnetic field represents a small fraction of the ancient magnetic field, whose direction may be different from the contemporary field both because of rock reorientation to their present positions and because of changes of the main field directions over geological times (► [Geomagnetic Field, Polarity Reversals](#)). The ease with which a material can acquire induced magnetization depends on the magnetic susceptibility k such that $\mu_0 \mathbf{M}_i = k \mathbf{B}$, where $\mathbf{M}_i$ is the vector of induced magnetization, $\mathbf{B}$ the inducing ambient field, and μ_0 the magnetic permeability of free space. Susceptibility values vary with rock types, but the mean bulk susceptibility of the Earth's crust is about 0.02 SI. Therefore, the induced magnetization produces a field much less intense than the inducing field. Since the magnitude of the Earth's core magnetic field does not vary significantly on a decadal time scale (► [Geomagnetic Field, Secular Variation](#)), the induced lithospheric field may be considered as quasi-static over historical times so that accumulation of measurements over years is made possible. Nevertheless, some predictions indicate that the time variations of this field may become significant over decades (Thébault et al. 2009; Hulot et al. 2009) so that data accumulation over long periods of time or between distant epochs may introduce errors. The total magnetization $\mathbf{M} = \mathbf{M}_i + \mathbf{M}_r$ responsible for the observed crustal field points toward a different direction from both the present and the ancient fields. The Königsberger ratio M_r/M_i measures the relative importance of remanent to induced magnetization in the laboratory. Without collecting rock specimens and measuring their magnetic properties in the laboratory, it is difficult to resolve accurately the magnetization into its induced and remanent components. However, the Königsberger ratio is generally large for oceanic rocks. The rather homogeneous

nature and structure of the oceanic crust (► [Lithosphere, Oceanic](#)) also allows inferring the remanent magnetization over the extent of oceanic crusts by interpolating the magnetic properties between geographically distant samples (see ► [Paleomagnetism, Principles](#)). In contrast, it is almost impossible to predict the remanent magnetization over continents and to separate it from the induced one. A common practice is to disregard remanent magnetization on continental areas because rock materials contain coarse magnetite grains. However, these first-order approximations are questionable for detailed analyses as their validity depends on the length scales of the magnetized structures under consideration.

Measuring Magnetic Anomalies from Planes and Ships

Near-surface surveys are routinely undertaken for mineral and hydrocarbon exploration and sometimes for the delineation of structures buried at depth. Airborne magnetic surveys are generally carried out to support mineral and petroleum exploration and aid in subsurface geological mapping (► [Magnetic Methods, Airborne](#)). These surveys are commonly executed on a regular pattern at some hundred meters altitude above mean terrain clearance, with a horizontal distance about twice as much as the aircraft altitude (Reeves 2005). The flight lines are oriented perpendicular to the geological strike directions to detect better the magnetic field contrasts. For reconnaissance mapping of large regions (e.g., states or countries), typical line spacing is 1 km between 1 and 5 km altitude above the sea level or above the topography. More widely spaced segments crossing the flight lines at right angles are then used to correct the measurements for offsets caused, either by external magnetic field time variations or by measurement and attitude errors.

The local and regional surveys do not detect the same structures of magnetic anomalies. Low-altitude surveys provide locally detailed profiles of short-scale magnetic field contrasts but little information about the magnetic structures exceeding the survey dimension. Likewise, high-altitude surveys provide a synoptic view of the magnetic anomalies but generally are not sensitive to the features with length scales smaller than the altitude of measurements. Theoretically, it is possible to highlight all magnetic field structures from one single altitude measurements by continuing the map downward close to the source layer. However, this procedure enhances the noise, and unless the measurements are accurate enough, the finest structures are distorted or inaccessible.

The anomaly field is defined by correcting the raw data for a core field model, either with complex magnetic field models or more generally with the International Geomagnetic Reference Field model (► [Geomagnetic Field, IGRF](#)), or

simply with *ad hoc* regional trend corrections. External field variations that occurred during the survey are sometimes explicitly corrected for. In this latter case, current practice is to use the magnetic records of nearest magnetic observatories or a nearby ground station purposefully set up during the survey acquisition. This procedure allows monitoring the transient disturbances and correcting for them. In most cases, the external field is further minimized by statistical or manual analyses of crossovers between the flight lines and by applying line-leveling algorithms. The final product is a map of anomaly values built by interpolation between data points and leveled to a common altitude onto a regular grid. An anomaly map built upon the compilation of various surveys displays crustal features with typical spatial scales from kilometers to hundred kilometers. However, because of their limited lateral extension, the stitching of different panels, and sometimes differences in the data processing protocol applied, they are more sensitive to magnetic field structures lying in the upper part of the crust.

Satellite Measurements

Large geologic structures with dimensions of a few hundred kilometers can be inferred from LEO satellites. Satellites provide scalar and vector homogeneous measurements at the Earth's scale (► [Magnetic Methods, Principles](#)). At the LEO satellite altitudes, the lithospheric field is one of the weakest detectable sources. It represents less than 0.1% of the full signal, and magnetic anomalies rarely exceed 20 nT in amplitude (► [Magnetic Methods, Principles](#)). Satellites are also close to (and sometimes fly through) the electrically conductive ionosphere. At satellite altitudes, ionospheric fields external to the Earth's surface are seen as internal ones, and the separation between internal and external source fields is more challenging than for airborne measurements. Satellite measurements may be exploited globally to co-estimate the core and the lithospheric field in SH, an operation difficult to perform with airborne and marine measurements. This co-estimation considerably limits the error introduced by correcting the data with an independent core field model such as the IGRF ones. Data selections during nighttimes permit to mitigate the contamination of the ionospheric magnetic field variations more prominent in the day-side. Magnetic field proxies developed from ground surface magnetic field observatory measurements are also extensively used to select the measurements during magnetospheric quiet times only. The lithospheric field models derived from space are then more conveniently continued downward to the Earth's surface even if this procedure amplifies the noise.

The motion of a satellite represents another source of ambiguity. As it is constantly moving along its orbit, its displacements create signals that can be interpreted in terms

of genuine lithospheric field spatial variations or in terms of time-varying perturbation magnetic fields. Altitude uncertainties of the spacecraft, time errors, misalignment of magnetometers, and the absence of baseline control introduce a total error whose magnitude is enhanced by the large spacecraft velocity. At about 7.5 km.s^{-1} at 400 km altitude for the CHAMP and Swarm satellites (Reigber et al. 2002; Olsen et al. 2013), such uncertainties can convert into a few nT error. The absence of traverse lines and the low number of crossovers also introduce North-South artificial stripes that may easily be confused with genuine anomalies when the satellite measurements are analyzed orbit by orbit (Thébault et al. 2012). Therefore, contrary to aeromagnetic data that need to be acquired once, only long periods of observations and data accumulation over the same area allow improving the signal-to-noise ratio at satellite altitudes. After almost two decades of nearly continuous measurements, the lithospheric field signal still drops into the noise somewhere between SH degree 90 and 100. CHAMP satellite measurements down to 250 km altitude and the innovative use of magnetic gradients measurements from the ongoing Swarm mission (Friis-Christensen et al. 2006) improved the view of the Earth's lithospheric magnetic field up to SH degree 130 (about 350 km wavelengths). Further, a priori assumption and regularization are also applied to extend models to higher SH degrees (Olsen et al. 2017). However, at these wavelengths close to the satellite resolution, lithospheric field models derived from space differ and have a typical standard deviation of about 0.5 nT at the mean altitude of 400 km. Lower-altitude measurements would enhance the signal-to-noise ratio but would reduce the lifetime of satellites. Finding the best trade-off between signal detection at short length scales, lifetime, and data accumulation remains challenging.

When merged together, aeromagnetic and satellite data provide a complementary view of the lithospheric field. Yet, it is suspected that these combined models have a loss of gain between 200 and 400 km wavelength because of the lack of intermediate altitude data between airplanes and satellites. Current efforts focus on ways to improve the resolution achievable from space by considering gradient measurements and constellation of satellites such as the Swarm trio of satellites. As a follow-on to the Swarm mission, the community investigates the feasibility of sending at lower costs a constellation of cube satellites in order to perform simultaneous multipoint magnetic field measurements on different orbits.

World Magnetic Anomalies: WDMAM

Homogeneous global lithospheric magnetic field models are essential for evaluating the crustal structures, understanding the geological processes, for establishing tectonic models of

continental and oceanic crust, for reconstructing the geometry of paleocontinents, and for contributing to geodynamic models. The merging of ground, near-surface, and satellite measurements is the only way to access this multi-scale view of the Earth's lithospheric magnetic field. The World Digital Magnetic Anomaly Map (WDMAM) project is an international effort directed toward compiling magnetic anomaly data available in the public domain at the worldwide scale (Korhonen et al. 2007). The updated version of the WDMAM is based upon more than 50 years of aeromagnetic and marine surveys and uses recent satellite lithospheric field models. It is available as a 3 arcmin spacing grid at 5 km altitude (Lesur et al. 2016). It relies on published continental-scale compilations derived by independent groups (e.g., the North Atlantic Magnetic Anomaly Group (NAMAG) or the Antarctic Digital Anomaly Map (ADMAP)), upon independent contributions from countries and individuals, and most recent lithospheric field model derived from the Swarm mission (Fig. 2). Continuous efforts are being made to improve the quality and the data coverage from airplane, marine, and satellite platforms to release updated grid maps of the WDMAM (www.wdmam.org).

Qualitative Interpretations

Most susceptibility contrasts may give rise to a magnetic field detectable by a magnetometer (► [Geomagnetic Field, Measurement Techniques](#)). Magnetic surveys are traditionally acquired for resource exploration in order to detect ore bodies, but also oil and natural gases, which are trapped in sedimentary basins. Magnetic maps are also useful to track anthropogenic buried materials, like unexploded bombs,

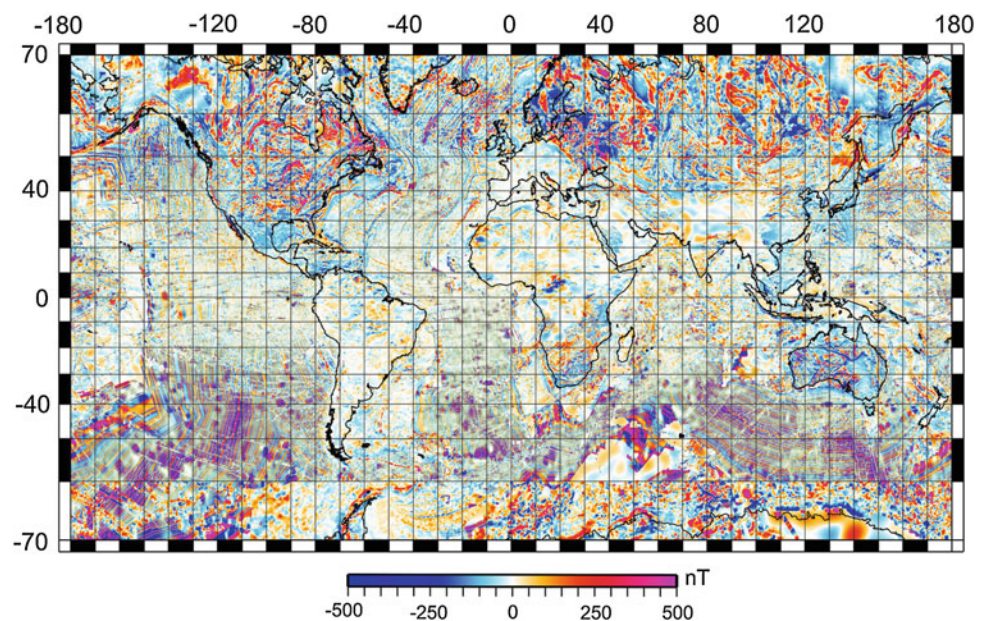
hidden weapons, and abandoned oil wells, but may also be used for guiding archaeological excavations.

At a regional scale, mapping the patterns of magnetic anomalies is a very effective way of reconnoitering large areas of geology (Purucker and Whaler 2007) in remote regions where rock outcrop is absent and bedrock covered by glacial, sand, vegetation, or water overburdens. In particular, the interpretation of magnetic anomalies over oceans was crucial to the development of the plate tectonic paradigm (Vine and Matthew 1963). Over continents, tectonic processes are detailed in a magnetic anomaly map because they are responsible for the creation, destruction, and movement of magnetic materials within the lithosphere that modify pre-existing magnetic signatures. New magnetic contrasts are created by volcanism and related igneous processes such as dike emplacement, rifting, and faulting.

At a continental scale, magnetic anomalies are particularly prominent above the main old and cold continental masses over Africa, Antarctica, Asia, Australia, Europe, Greenland, India, and North and South America (Fig. 2) and comparatively weaker over younger and hotter oceans (at least from space; see ► [Magnetic Methods, Principles](#)). The largest magnetic anomaly features correlate well with cratons, Precambrian and Paleozoic provinces. Within each structure, magnetic anomalies of smaller extension like mobile belts, collision zones, and volcanic areas can be characterized. Crustal thickness variations at continent ocean boundaries and large subduction zones are also sometimes outlined with a strong correlation between the seismic Mohorovicic (Moho) discontinuity and the magnetic crustal thickness. Large magnetic anomalies often correlate with plate boundaries and seismotectonic activity around the world. Reconstructions of wide separated geological units or rotated blocks make

Magnetic Anomalies: Interpretation,

Fig. 2 Lithospheric magnetic field model derived from CHAMP, Swarm, and the WDMAM grid. (After Lesur et al. 2016. Projection is Mercator)



extensive use of the seafloor stripes and transform faults' magnetic signatures.

The picture of a magnetic anomaly is generally more understandable after a procedure of pole reduction that transforms an observed magnetic anomaly into the anomaly that would be measured at the north magnetic pole. The pole reduction method corrects for the induction effect by relocating the magnetic anomalies so that its maximum lies in the middle above the magnetized body. The flank of anomalies may also be clarified by upward and downward continuation, and depth estimations are obtained by the application of numerical filters under simplifying approximations (see Baranov (1975) for an overview of these techniques).

Quantitative Geological Interpretations

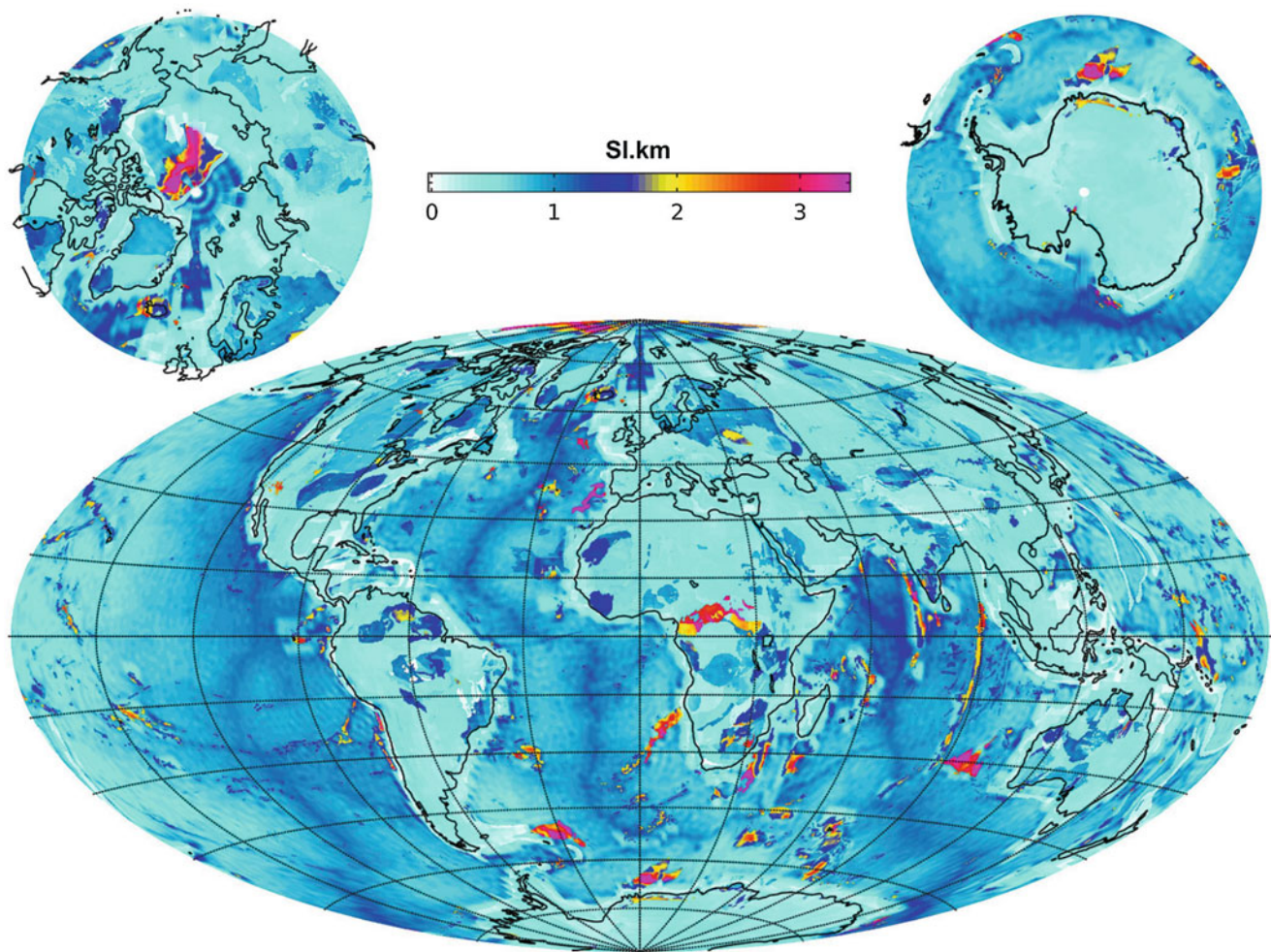
The quantitative interpretation of magnetic maps directly in terms of true magnetization is difficult. A single magnetic anomaly can be caused by an infinite variety of magnetized bodies. Moreover, not all magnetization distributions produce magnetic fields detectable above the Earth's surface (Runcorn 1975). The spatial variations in the measured magnetic field reflect lateral contrasts of magnetization rather than the magnetization itself. Two adjacent geological units possessing both strong and comparable magnetization may produce a magnetic field comparable to the field of two bodies with smaller magnetization. Outcropping geological obvious features that do not necessarily have a detectable magnetic signature as part of the magnetization is hidden in the crust (Gubbins et al. 2011; Vervelidou and Lesur 2018). The interpretations of magnetic maps in terms of magnetization distribution of rock in the crust therefore generally require additional geophysical information. For this reason, the inferred magnetization from field models is referred to as the apparent bulk magnetization.

Locally, on continents, a quantitative description of a magnetized body is performed by using processing techniques inherited from Fourier analyses. It can also be achieved by setting a distribution of equivalent magnetic dipoles in the subsurface and interpreting the observed anomaly in terms of inducing field (► [Magnetic Anomalies: Interpretation](#)). When the geometrical parameters of the geological source are bounded, the magnetization of a body can be found by trials and errors by a comparison with the magnetic measurements (► [Magnetic Anomalies: Interpretation](#)), and its rock type can be subsequently inferred from a table of rock magnetization (► [Magnetic Domains](#); ► [Remanent Magnetism](#)). In most areas, interpreting individually each magnetic anomaly is not straightforward if the geophysical body has a complex geometry. In such cases, it is more convenient to analyze the magnetic signal statistically in the spectral rather than in

the geographical domain. This allows estimating the average source depth on various length scales (Thébault and Vervelidou 2015) or estimating the statistical contribution of the induced and remanent fields by comparisons with theoretical power spectra. Maus and Haak (2002) verified, for example, that the lithospheric magnetization was statistically mostly induced over continent and remanent over oceans at the wavelengths detected by the LEO satellites. The spectral analysis justifies that in the geographical domain, the anomaly field over continent can be approximated by the anomaly field produced by an array of equivalent dipole with moments aligned along the main core field.

The ambiguities in interpreting the magnetic anomaly maps can be further reduced by requesting guidance from other disciplines like, gravity, seismology, heat flux, geology, etc. The geophysical approach that consists in analyzing crustal field models in large integrated geophysical system requires an extensive work for linking analytically or numerically the large number of geophysical parameters responsible for the magnetic field signal. So far, simplifying assumptions are still necessary. Such approximations are more easily justified from space as the magnetization can be assumed to vary laterally and to be contained in a crustal layer of constant thickness as at 400 km altitude, the Earth's crust appears as a thin layer.

The consideration of the bare thickness contrast between continental (~30 km) and oceanic (~7 km) crusts of uniform, but different, susceptibility values generates magnetic anomaly field that is seen in magnetic field measurements along the boundaries of oceans and continents (Counil et al. 1991). This indicates that magnetic crustal thickness plays a significant role in the dichotomy between oceanic and continental magnetic field signals observed from space (see ► [Magnetic Methods, Principles](#), Fig. 1). Cohen and Achache (1994) further introduced a model of oceanic topography to estimate a more realistic crustal thickness in some oceanic areas; they could explain the magnetic anomalies observed over oceanic plateaus. On ocean bottoms, remanent magnetism predicted by Dyment and Arkani-Hamed (1998) and compared to measurements could explain well large magnetic signatures over superchrons in the North Atlantic and Indian oceans. This modeling approach could in turn be also used to infer the magnetization distribution of the ocean floors with interpolation/extrapolation techniques (Masterton et al. 2012; Dyment et al. 2015). Following these studies, Hemant and Maus (2005) merged these different approaches in an attempt to predict all structures of the lithospheric field seen at satellite altitudes and to estimate the worldwide distribution of the rock susceptibility integrated within the entire crustal thickness (Fig. 3). They relied on a priori geophysical data about the nature of the magnetization, the susceptibility of continental rocks, the stratification of the crust, the depth to the Moho, and the boundaries of the geological units at a global



Magnetic Anomalies: Interpretation, Fig. 3 Map of the equivalent susceptibility of the Earth's magnetic crust in SI.km. (After Hemant and Maus (2005) updated with the Swarm lithospheric magnetic field models. Projection is Mollweide)

scale. Their forward model was subsequently compared to the lithospheric field model derived from the CHAMP measurements. They could show that our understanding of the Earth's magnetic lithosphere was to a first order statistically reasonable at large wavelengths but mostly failed at the smaller wavelengths. Moreover, the long-standing question concerning the presence of magnetization within the upper lithosphere could not be addressed. Some geophysical arguments suggest that magnetization could persist below the Moho seismic discontinuity (Ferré et al. 2014). However, estimation of the magnetic crustal thickness by trials and errors from iterative comparisons between the forward-predicted crustal fields with observed satellite magnetic anomalies (Purucker et al. 2002) and independent regional statistical analyses of models does not support this conclusion (Vervelidou and Thébault 2015). The presence of magnetization in the upper lithosphere remains a crucial question for inferring the thermal state and the composition of the crust/lithosphere (Fox Maule et al. 2005).

Summary

On Earth, a magnetic anomaly is defined with respect to the core and the external fields that are considered as perturbation magnetic fields. Characterizing the anomaly field and depicting all spatial dimensions are challenging as data processing leading to anomaly maps differs with the altitudes and the epochs of measurements. Nowadays, the main sources of discrepancies are the rapid temporal changes of the external magnetic field that remain poorly sampled. Magnetic anomaly maps give the essence of the crustal magnetic field that in turn provides useful geological inference and information about the magnetization of rocks in the subsurface. However, the magnetization depends on a large variety of parameters such as the composition of the crust, the heat flux, the age, the crustal thickness, the geological units, the stratification of the crust, etc. Some parameters of primary importance at large scale may be of secondary importance at smaller scales and conversely. Current efforts are being made to link these

geophysical parameters to the magnetization and to derive numerical simulators in order to model the magnetic crust in three dimensions. A better appreciation of the order of importance of each geophysical parameter involved in the observed lithospheric magnetic field at different spatial scales is the key to understand the magnetic structure and the thermal state of the Earth's crust, the uppermost part of the lithosphere, and its interactions with the deep mantle and surface processes.

Cross-References

- Curie Temperature
- Geomagnetic Field, Global Pattern
- Geomagnetic Field, IGRF
- Geomagnetic Field, Measurement Techniques
- Geomagnetic Field, Secular Variation
- Geomagnetic Field, Theory
- Magnetic Data Enhancements and Depth Estimation
- Magnetic Domains
- Magnetic Methods, Airborne
- Magnetic Methods, Principles
- Magnetic Methods, Surface
- Paleomagnetism, Principles
- Remanent Magnetism

Bibliography

- Backus G, Parker R, Constable C (1996) Foundations of geomagnetism. Cambridge University Press, New York
- Baranov V (1975) Potential fields and their transformations in applied geophysics. Borntraeger, Berlin/Stuttgart. ISBN 3443130089
- Cohen Y, Achache J (1994) Contribution of induced and remanent magnetisation to long-wavelength oceanic magnetic anomalies. *J Geophys Res* 99:2943–2954
- Council J, Cohen Y, Achache J (1991) The global continent–ocean magnetisation contrast: spherical harmonic analysis. *Earth Planet Sci Lett* 103:354–364
- Dunlop DJ, Özdemir O (2007) Magnetisations in rocks and minerals, chapter 8. In: Kono M (ed) Treatise on geophysics, vol 5. Elsevier, Amsterdam, pp 278–331
- Dyment J, Arkani-Hamed J (1998) Contribution of lithospheric remanent magnetisation to satellite magnetic anomalies over the world's oceans. *J Geophys Res* 103:15423–15441
- Dyment J, Choi Y, Hamoudi M, Lesur V, Thébault E (2015) Global equivalent magnetization of the oceanic lithosphere. *Earth Planet Sci Lett* 430:54–65
- Ferré EC, Friedman SA, Martin-Hernandez F, Feinberg JM, Till JL, Ionov DA, Conder JA (2014) Eight good reasons why the uppermost mantle could be magnetic. *Tectonophysics* 624:3–14
- Fox Maule C, Purucker M, Olsen N, Mosegaard K (2005) Heat flux anomalies in Antarctica revealed by satellite magnetic data. *Science* 309:464–467
- Friis-Christensen E, Lühr H, Hulot G (2006) SWARM: a constellation to study the Earth's magnetic field. *Earth Planets Space* 58(4):351–358
- Gubbins D, Ivers D, Masterton SM, Winch DE (2011) Analysis of lithospheric magnetization in vector spherical harmonics. *Geophys J Int* 187(1):99–117
- Hemant K, Maus S (2005) Geological modeling of the new CHAMP magnetic anomaly maps using a Geographical Information System (GIS) technique. *J Geophys Res B* 110:B12103. <https://doi.org/10.1029/2005JB003837>
- Hulot G, Olsen N, Sabaka TJ (2007) The present field, chapter 6. In: Kono M (ed) Geomagnetism, treatise on geophysics, vol 5. Elsevier, Amsterdam, pp 33–72
- Hulot G, Olsen N, Thébault E, Hemant K (2009) Crustal concealing of small-scale core-field secular variation. *Geophys J Int* 177(2):361–366
- Korhonen JK, Fairhead JD, Hamoudi M, Hemant K, Lesur V, Manda M, Maus S, Purucker ME, Ravat D, Sazonova T, Thébault E (2007) Magnetic anomaly map of the world – Carte des anomalies magnétiques du monde, scale: 1:50,000,000, 1st edn. Commission for the Geological Map of the World, Paris
- Langel RA, Hinze WJ (1998) The magnetic field of the Earth's lithosphere: the satellite perspective. Cambridge University Press, New York. 429 pp
- Lesur V, Hamoudi M, Choi Y, Dyment J, Thébault E (2016) Building the second version of the world digital magnetic anomaly map (WDMAM). *Earth, Planets and Space* 68(1):27
- Lowes FJ (1974) Spatial power spectrum of the main geomagnetic field and extrapolation to the core. *Geophys J R Astron Soc* 36:717–730
- Masterton SM, Gubbins D, Müller RD, Singh KH (2012) Forward modelling of oceanic lithospheric magnetization. *Geophys J Int* 192(3):951–962
- Maus S (2010) An ellipsoidal harmonic representation of Earth's lithospheric magnetic field to degree and order 720. *Geochem Geophys Geosyst* 11:Q06015
- Maus S, Haak V (2002) Is the long wavelength crustal magnetic field dominated by induced or by remanent magnetisation? *J Ind Geophys Union* 6(1):1–5
- Maus S, Yin F, Lühr H, Manoj C, Rother M, Rauberg J, Michaelis I, Stolle C, Müller RD (2008) Resolution of direction of oceanic magnetic lineations by the sixth-generation lithospheric magnetic field model from CHAMP satellite magnetic measurements. *Geochem Geophys Geosyst* 9:Q07021. <https://doi.org/10.1029/2008GC001949>
- Purucker M, Whaler W (2007) Crustal magnetism, chapter 6. In: Kono M (ed) Treatise on geophysics, vol 5. Elsevier, Amsterdam, pp 195–237
- Olsen N, Eigil F-C, Rune F, Patrick A, Ciaran DB, Arnaud C, Eelco D et al. (2013) The Swarm satellite constellation application and research facility (SCARF) and Swarm data products. *Earth, Planets and Space* 65(11):1
- Olsen N, Ravat D, Finlay CC, Kother LK (2017) LCS-1: A high-resolution global model of the lithospheric magnetic field derived from CHAMP and Swarm satellite observations. *Geophysical Journal International* 211(3):1461–1477
- Purucker M, Langlais B, Olsen N, Hulot G, Manda M (2002) The southern edge of cratonic North America: evidence from new satellite magnetometer observations. *Geophys Res Lett* 29(15):8000. <https://doi.org/10.1029/2001GL013645>
- Reeves C (2005) Aeromagnetic surveys: principles, practice and interpretation. Published by Geosoft, 155 pp. <http://www.geosoft.com/resources/news/2009/090723.asp>
- Reeves CV, De Wit M (2000) Making ends meet in Gondwana: retracing the transforms of the Indian Ocean and reconnecting continental shear zones. *Terra Nova* 12:272–280. <https://doi.org/10.1046/j.1365-3121.2000.00309.x>
- Regan RD, Cain JC, Davis WM (1975) Global magnetic anomaly map. *J Geophys Res* 80:794–802
- Reigber C, Lühr H, Schwintzer P, (2002) CHAMP mission status. *Adv Space Res* 30. [https://doi.org/10.1016/S0273-1177\(02\)00276-4](https://doi.org/10.1016/S0273-1177(02)00276-4)
- Runcom SK (1975) On the interpretation of lunar magnetism. *Phys Earth Planet Inter* 10:327–335

- Thébault E, Vervelidou F (2015) A statistical spatial power spectrum of the Earth's lithospheric magnetic field. *Geophys J Int* 201(2):605–620
- Thébault E, Hemant K, Hulot G, Olsen N (2009) On the geographical distribution of induced time-varying crustal magnetic fields. *Geophys Res Lett* 36(1):L01307
- Thébault E, Vervelidou F, Lesur V, Hamoudi M (2012) The satellite along-track analysis in planetary magnetism. *Geophysical Journal International* 188(3):891–907
- Thébault E, Christopher CF, Ciarán DB, Patrick A, Julien A, Olivier B, Francois B et al. (2015) International geomagnetic reference field: the 12th generation. *Earth, Planets and Space* 67(1):79
- Vervelidou F, Lesur V (2018) Unveiling Earth's hidden magnetization. *Geophys Res Lett* 45(22):12–283
- Vervelidou F, Thébault E (2015) Global maps of the magnetic thickness and magnetization of the Earth's lithosphere. *Earth Planets Space* 67(1):173
- Vine FJ, Matthew DH (1963) Magnetic anomalies over oceanic ridges. *Nature* 199:947–949

Magnetic Anomaly Map, Global

Kumar Hemant Singh

Department of Earth Sciences, Indian Institute of Technology Bombay, Mumbai, India

Definition

Global Magnetic Anomaly Map is an international effort to integrate near-surface and downward-continued spaceborne maps of the magnetic anomaly field. These global maps are available as grids and find use in geologic and tectonic interpretation.

Introduction

Magnetic data have been collected from magnetometers towed by aircraft and ships for many decades now. Magnetic anomaly maps derived from regional airborne (*Magnetic methods, airborne*) and marine surveys are beneficial to our understanding of the subsurface structure, chemical composition, and geodynamics of the Earth's crust (*Continental crustal structure*) and lithosphere (Purucker and Whaler 2007). Maps derived from marine surveys were instrumental in the discovery of seafloor spreading and in the development of the theory of plate tectonics. Continental anomaly maps derived from aeromagnetic surveys interpreted with other geophysical information are used to delineate geologic and tectonic provinces and structures such as faults, dykes, and lineations (*Magnetic anomalies, interpretation*). These maps are also widely used, for example, in natural resource exploration, in estimating the thermal state of the Earth's lithosphere (*continental lithosphere: thermal structure*), and the

magnetic characterization (*Remanent magnetization*) of volcanic surfaces and deep interiors on both continents and oceans.

Satellite missions (*Magnetic methods, satellite*) have also made significant contributions in enhancing our knowledge of the Earth's magnetic field (c.f. Langel and Hinze 1998). Satellite data are acquired and processed in three-dimensional spherical coordinates; hence, the derived magnetic anomaly maps are on a global scale. The satellite data are measured typically 400–750 km away from the causative sources in the Earth's lithosphere, lack sufficient resolution of geological features smaller than 400 km in wavelengths, but find extensive use in studying the long-wavelength anomalies. In comparison, magnetic anomaly maps prepared from data recorded near the surface (~ few kms) and therefore near to the sources resolve short-wavelength anomaly features better (*Magnetic anomalies, interpretation*). There are, however, many geologic problems that could be better solved with the use of the self-consistent global magnetic anomaly maps that take advantage of the strengths of each data set while bypassing its limitation. For example, repeated rifting (*Continental rifting*) and movement of tectonic plates (*Plates and paleoreconstructions*) over hundreds of millions of years has made once adjacent landmasses drift (*Continental drift*) to a completely different region of the world. Reassembly requires consistently prepared magnetic anomaly maps over the entire globe consisting of both short and long wavelengths. The world digital magnetic anomaly map (WDMAM) derived from merging near-surface magnetic data and a global satellite model will aid evaluation of hypotheses about geologic origins of tectonic rifts, fracture zones that run across the continents, and tectonic history of many regions.

The WDMAM is the magnetic field defined on a grid of 3 arc-min angular grid resolution at an altitude of 5 km above the World Geodetic System 1984 (WGS84) reference ellipsoid (Fig. 1) (Korhonen et al. 2007). The world magnetic anomaly map is the first international collaborative effort to compile magnetic anomaly information on a global scale. The map is derived from more than 50 years of aeromagnetic, shipborne surveys over land and sea and supplemented by observations from earth-orbiting satellites. The global map shows the variation in the strength of the magnetic field (*Geomagnetic field, global pattern*) after the Earth's dipole field, and external field sources such as magnetosphere and ionosphere have been subtracted from the measured field. The remaining variations in the field (few hundreds of nanoteslas (nT)) are due to changes in the magnetic properties of the crustal rocks. In most places, the magnetic anomaly field is less than 1% of the main magnetic field of the Earth.

The magnetic fields derived in this map are internally consistent within the measurement domain, extending from the Earth's surface to satellite altitude (<750 km). The map



Magnetic Anomaly Map, Global, Fig. 1 The World Digital Magnetic Anomaly Map published by Commission for the Geological Map of the World (CGMW) in July 2007. The map is shown in Mercator Projection and is at an equatorial scale 1:50,000,000. (Courtesy: Korhonen et al. 2007)

comprises both the short and long wavelength anomalies from different sources. Short wavelength anomalies are derived from merging continental compilations and marine track line data. Long-wavelength features in this global map are based on a downward continued grid of the MF5 model, a model derived from CHAMP satellite magnetometer measurements. In oceanic regions where the data gaps exist, the digital age map of the ocean floor (Müller et al. 1997) was combined with the geomagnetic polarity time scale (Gee and Kent 2007; Kent and Gradstein 1986) (*Geomagnetic field, polarity reversals*) to estimate the total field anomaly at 5 km altitude.

The global magnetic anomaly map is available in both digital and print form. Two versions (A and B) of the map are available on DVD. The B version is shown in Fig. 1. The map is displayed in natural colors (red = positive values, blue = negative values) and produced from Generic Mapping Tools (GMT) (Wessel and Smith 1998). The A version differs in its handling of areas with data gaps, which are filled with the downward continued MF5 model. The map images and the DVD version of the WDMAM 2007 are available at <http://ftp.gtk.fi/WDMAM2007/>.

Early Developments

Magnetic surveys have been carried out on the land, sea, and from the air over all continents and oceans. The patchwork of aeromagnetic surveys over the continents and the tracks of oceanographic survey vessels across the oceans, together make up an enormous but distinct body of data. This incomplete and nonuniform coverage of near-surface data has recently been supplemented by reliable maps from Earth-orbiting satellites such as CHAMP (<http://www-app2.gfz-potsdam.de/pb1/op/champ/>). These satellite-derived models (*Magnetic method, satellite*) provide uniform coverage over the globe that can help resolve uncertainties in near-surface compilations produced by stitching together a disparate set of data. A new initiative was therefore required to combine these near-surface anomaly maps into a reliable global compilation and make it more accessible for geological mapping, educational purposes, and future research.

The original data is held by a large number of individual organizations around the world and, in some cases, is still commercially used and politically sensitive. In 1977, the International Association of Geomagnetism and Aeronomy (IAGA) (<http://www.ngdc.noaa.gov/IAGA/vmod/>) realized the importance of magnetic anomaly mapping for geologic and tectonic interpretation (*Magnetic anomalies, interpretation*) and noted the existence of numerous airborne and marine magnetic data. They began a concerted international effort to bring the global magnetic anomaly data set into place alongside other worldwide databases to maximize its use. Through a series of resolutions in 1979, 1995, and 1997, the

WDMAM task force was constituted by IAGA at the International Union of Geodesy and Geophysics General (IUGG) Assembly in 2003 to prepare a roadmap to complete the work. The job of the task force included (1) to catalogue existing data compilations that cover significant areas of the Earth; (2) to negotiate access to such data at a resolution sufficient for a global compilation; (3) to produce a preliminary paper map at the IAGA Assembly in Toulouse, France, July 2005; and (4) while addressing the main scientific and practical issues, to produce a digital magnetic anomaly map of the world for the 24th IUGG General Assembly, Perugia, Italy, July 2007.

The WDMAM task force urged various organizations and custodians of data to release nonsensitive versions of the classified or confidential aeromagnetic and marine track line digital data to the public domain for the WDMAM project. The organizations holding the data responded to the call and made near-surface data available for the first version of the WDMAM map. In June 2006, a call for submission of WDMAM candidate models was issued. In November 2006, five models were submitted for evaluation. Five teams at the National Geophysical Data Center (NGDC), NASA – Goddard Space Flight Center (GSFC), Geological Survey of Finland (GTK), GeoForschungsZentrum, Potsdam, Germany, and the University of Leeds, UK, produced candidate models for the WDMAM project (<http://projects.gtk.fi/WDMAM/project/>). After careful evaluation, the candidate model developed by NGDC was selected as a base map for representing anomalies over continents while the GAMMA model was selected for oceans. The first version of the official WDMAM was unveiled at the 24th IUGG General Assembly in Perugia, Italy, in July 2007 by the Commission for the Geological Map of the World (CGMW). United Nations Educational, Scientific and Cultural Organization (UNESCO) kindly supported and provided the funding for the printing.

Aeromagnetic and Marine Compilations

The first airborne surveys were done over the former Soviet Union in 1936, and the first maps of the total intensity were published at 1:2.5 million scale in 1974. These maps were later digitized and made available at one and three arc-minute grid resolution from the NGDC magnetic data archive (<http://ols.ndbc.noaa.gov/plolstore/plsql/olstore.prodspecific?prodnum=G01284-CDR-A0001>). More recently, another magnetic compilation for Russia compiled by the A. R. Karpinsky Russian Geological Research Institute is available as a 5-km resolution of the data grid (VSEGEI, <http://www.vsegei.ru/WAY/247038/locale/EN>).

The first ever North American continental-scale compilation was derived by merging magnetic anomaly maps of USA, Canada, and surrounding waters. The map was released in 1987 as a part of the Decade of North American geology.

The map had erroneous long-wavelength features, which revealed shortcomings and pitfalls in stitching together residual intensity maps reduced from measured data using different IGRF models (*Magnetic data, corrections and enhancements*). After correcting for these erroneous long wavelengths using Comprehensive Model (CM) (Sabaka et al. 2002), the improved maps for USA and Canada, in addition to the aeromagnetic compilation for Mexico, were combined to produce the revised North American magnetic anomaly map (Bankey et al. 2002). The Arctic and North Atlantic magnetic compilations including nearby landmasses were produced from a patchwork of shipborne and airborne surveys (Verhoef et al. 1996). Because the compilation had incorrect long wavelengths greater than 400 km, it was filtered before releasing the grid to the public domain.

Across the Atlantic, Wonik et al. (2001) prepared the first European compilation by merging magnetic data from north, west, and east Europe. The map retains the long-wavelength features but shows a poor correlation when compared with satellite data. However, there have been many developments in the last decade in preparation of high-resolution magnetic compilations from marine and airborne surveys. The Geological Survey of Japan compiled the grid for East Asia in 2002. The fourth edition, but a significant improvement over its previous version, the high-resolution magnetic anomaly grid for Australia and its surrounding oceans, was released (Minty et al. 2003). The noticeable feature of this map is its control over the long-wavelength anomalies which was achieved by flying long aeromagnetic traverses around Australia. The maps for China have also been recently made available in digital form by the Chinese National Aerogeophysics Surveys. Another map that was digitized for WDMAM project is a combined map for China, Mongolia, and Russia produced in 1995 by the Geological Survey of Canada. This map also had long wavelengths greater than 400 km corrected. In addition, the Antarctic Digital Magnetic Anomaly Project (ADMAP) provided the continental-scale compilation over Antarctica (Golynsky et al. 2001). The map was derived from merging aeromagnetic and marine data with the satellite magnetic data.

In addition to these data in the public domain, there have been efforts by industry-led consortia to produce maps for Africa, South America, and Middle East (http://www.getech.com/data/global_mag_margins.htm). Geological Survey of India also contributed the Indian magnetic compilation. Over oceans, NGDC's archives of track line data sets (<http://www.ngdc.noaa.gov/mgg/geodas/geodas.html>) were made available. In addition, the Project Magnet ocean traverse data were also available at http://www.ngdc.noaa.gov/geomag/proj_mag.shtml.

The details of the marine and airborne magnetic compilations that were used by the WDMAM task force to produce the candidate models can be found in Singh (2011) and at <http://projects.gtk.fi/WDMAM/project/contribution/>. The data density greatly varies between the northern and the southern

hemispheres. The coverage is especially sparse over oceans in the southern hemisphere where some of the marine track line data were not available for the project. Missing data are noted over parts of the Arabian Peninsula, India, East Asia, and many countries within Africa and South America. For processing the compilations, the datasets were grouped in three sets. The NGDC and Project Magnet marine and aeromagnetic track line data comprised the first set. Due to large errors and inconsistencies to be discussed later, the Project Magnet data was not used in the preparation of GAMMA candidate model. The second set included the gridded aeromagnetic and marine magnetic anomaly grids. The downward-continued MF5 model derived from the CHAMP satellite data made the third set. Of the track line profile data, the NGDC marine data and Project Magnet measurements together make 27 million points on 31,000 tracks. With scales ranging from country to continents, a total of 19 gridded data sets were made available.

The metadata information from the compilations when available shows the coverage to be in various projections and datums depending on their geographical representation on the globe. Hence, prior to applying any filtering or merging procedures, all grids were preprocessed and transformed from their local reference system to a common representation on the WGS84 global reference ellipsoid, using conversion formulas (Snyder 1987). The software, Oasis Montaj, Geosoft[®] was used in the preparation of the GAMMA model for upward-continuation of the data, while the Leeds model was produced on the ESRI's ArcGIS (www.esri.com).

Satellite Compilations

Several global models of the lithospheric magnetic field, based on satellite missions, have been produced in the last five decades (*Magnetic methods, satellite*). Although spherical harmonic analyses of data gathered at satellite altitude are typically used to compute the magnetic potential, they follow two different philosophies. The comprehensive approach (CM4), proposed by Sabaka et al. (2004), uses POGO, Magsat, Ørsted, CHAMP, and SAC-C measurements (*Geomagnetic fields, measurement techniques*) to derive a model that accounts for magnetic sources (*Geomagnetic field, theory*) from the Earth's main field to the external field sources in the ionosphere and magnetosphere and their temporal variations, including the crustal magnetic field component up to spherical harmonic degree 60. In contrast to the comprehensive approach, the second philosophy strictly focuses on the lithospheric field representation. This approach requires careful data selection and processing to clean the data from nonlithospheric sources (*Magnetic methods, satellites*). MF series of models are derived from CHAMP scalar and vector magnetometer measurements (<http://geomag.org/models/index.html>). The MF4 model used 4 years of

CHAMP data and is derived up to spherical harmonic degree 90. The map showed inconsistencies in the Polar Regions due to high data noise levels. The use of localized basis functions by Lesur and Maus (2006) led to the derivation of MF4x model which showed improvement in the lithospheric field at high latitudes. In 2007, the fifth generation of CHAMP models, MF5, was released (Maus et al. 2007a). MF5 extends to spherical harmonic degree 100 (400 km resolution) and showed a significant improvement, especially between 60° and 80° latitude, compared to previous models. After comparing the MF series of satellite compilations, the WDMAM committee recommended using the MF5 model for the preparation of candidate models. Recently, using only 3 years of CHAMP measurements, a high-resolution MF6 model was derived (Maus et al. 2008). The MF6 model is able to resolve the lithospheric magnetic field to degree 120, corresponding to 333 km wavelength. The MF6 model is also the first satellite-data-based global magnetic anomaly map to resolve the direction of oceanic magnetic lineations, which makes a helpful addition to the study of oceanic lithosphere (*Lithosphere, oceanic*) from space. The EMAG2 model uses the MF6 model to substitute the long-wavelength components. The last map produced by Maus and his group in the MF series of global lithospheric magnetic field models is MF7 (Maus 2010a). The model resolves crustal field up to spherical harmonic degree $n = 133$ which corresponds to 300 km wavelength on the Earth's surface. It is still the most widely used models, and the modeling strategy has remained largely unchanged from the previous version MF6. The model uses data from May 2007 to April 2010 when CHAMP satellite was in low-Earth orbit, which gives it the edge over its predecessors. In a recent development, LCS-1 model derived from CHAMP and Swarm data is able to resolve the lithospheric field up to spherical harmonic degree 185 (corresponding to 216 km horizontal wavelength) (Olsen et al. 2017).

Preparation of the Global Magnetic Anomaly Map

The NGDC candidate model (Maus et al. 2007b) uses the method of Least Squares Collocation (LSC) and line-leveling methods to generate a regular grid from irregularly spaced marine and aeromagnetic measurements. Other approaches for computing the candidate model include removal of a regional polynomial; iteratively adjusting a low-degree main field model for different epochs from each of the compilations; upward-continuation of the data to 5 km above the WGS84 ellipsoid; and finally knitting, merging, and filtering them together (Hamoudi et al. 2007). Long wavelengths (>400 km or spherical harmonic degrees ≤ 100) were removed from the individual compilations and replaced with the MF5 anomaly map of the total field

downward continued to 5 km altitude (Hemant et al. 2007). The following sections describe the preprocessing techniques and methods followed to derive the NGDC candidate model for WDMAM.

The WDMAM requires an accurate estimate of the crustal magnetic field on a 5-km angular resolution grid. Gridded data sets were mostly available in 5-km grid or even had a coarser resolution. In the case of track line data over the oceans, the resolution was variable (Singh 2011). Therefore, to prevent further loss of information during processing, all grids were decimated to 2.5 km angular resolution using the minimum curvature algorithm of GMT. Most of the compilations, in particular over the Northern Hemisphere, had overlapping regions. The compilations known to have higher accuracy or finer resolution were retained within overlapping parts by rejecting the coarser ones. In the process, the 15 arc-min GETECH data of South America and Africa were discarded whenever a higher resolution grid for the same region was available. Generally, the aeromagnetic surveys over the land are draped, meaning that they follow the surface topography at a constant altitude above the terrain. Thus, the derived compilations from these surveys follow topography. Therefore, to accurately process the altitude information with respect to a geoid, a topographic model of the Earth was used. The NGDC ETOPO-2 (<http://www.ngdc.noaa.gov/mgg/fliers/01mgg04.html>) surface elevation grid was linearly interpolated to the measurement positions and added to the terrain clearance of all draped surveys, to compute the measurement altitude above the geoid.

The continental-scale compilations used in WDMAM (Singh 2011) were produced after stitching together many regional aeromagnetic anomaly grids (Ravat et al. 2002). These regional surveys were flown at different times and, hence, used different International Geomagnetic Reference Fields (IGRF) models (*Geomagnetic field, IGRF*) to reduce the measured data. The patchwork grids were therefore prone to offsets and edge effects and mismatch in anomaly shapes and strengths across overlapping regions. They also had the potential to introduce spurious anomalies. These edge effects were corrected by subtracting a linear 2D trend from all smaller size grids.

It is well known that the genuine crustal anomalies of strengths ~ 500 nT can easily be contaminated by external field disturbances of amplitudes 100 nT. Thus, it has been a challenge to filter out short-wavelength noise from aeromagnetic and marine track line data. To filter these spurious signals, spatiotemporal variations of the magnetic field of strong but fast external field disturbances and those recorded by instruments onboard relatively slow-moving ships and aircraft are compared by computing the along-track derivative of the magnetic field residual. As a selection criterion, if the derivative exceeded 100 nT/km, 20 preceding and following measurement points on a track were removed. This selection

procedure removed about 3% of the marine track line profile data and 9% of the Project Magnet Data, which possibly were affected by external field disturbances.

Methods

Line Leveling of Data

In aeromagnetic surveys, inaccuracy in the measured positions of crossover points of acquisition, line-to-line differences in flying heights along adjacent tracks, and inadequate compensation for the magnetic field resulting from the aircraft lead to errors covering a region that extends beyond the line spacing in each direction (*Magnetic data, corrections and enhancements; Magnetic methods, airborne*). These errors are visible as linear anomalies parallel to the flight lines. If uncorrected, the computation of derivatives enhances these noise effects along the lines. This problem is common in acquiring aeromagnetic data. A common processing method to remove these errors is to line-level the data. A set of tie-lines, perpendicular to but spaced greater than that of the main survey lines, is normally acquired for line leveling. The line leveling procedure computes differences in the field values at the crossover points of the survey and tie lines, and corrections are applied to minimize these differences in the least-square sense (Reeves 2005).

The track line data, both marine and aeromagnetic, used for the WDMAM project is also likely to have this kind of problem. These track line data are first corrected for the main field and external fields. Additional navigational errors, instrument biases and disturbances from the unfiltered external field remain in the data causing uncertainty in the longer wavelength part of the magnetic field. The line leveling technique addresses these remaining errors in the track line data by minimizing the differences between adjacent or crossing tracks (Maus et al. 2007b).

The technique defines a correction function for each track given by

$$f_j(d) = \sum_{k=0}^{M_j} b_{j,k} \cos(\pi kd/X_j). \quad (1)$$

Here, X_j is the great circle distance from the first to the last point of the j th track and M_j is the number of parameters required for the estimation of the correction function of this track. The algorithm assumes $M_j = \text{trunc}(X_j/400 \text{ km}) + 1$ to correct for offsets with half-wavelength greater than 400 km. The choice of 400-km wavelength is based on the minimum wavelength the MF5 model is able to resolve. To minimize the misfit between all adjacent tracks and the differences of the track line profiles to the gridded data, the parameters $b_{j,k}$ are estimated from inversion. The line-leveling algorithm then minimizes the distance-weighted misfit between all

measurement locations. The weight function $W(r_{1,2}, z_1, z_2)$ used here is given by

$$W(r_{1,2}, z_1, z_2) = \max \left(0, 1 - \frac{1}{R_s} \sqrt{(r_{1,2}^2 + 2(z_1 - z_2)^2)} \right). \quad (2)$$

where R_s is the search radius, $r_{1,2} = ((x_2 - x_1)^2 + (y_2 - y_1)^2)^{1/2}$ is the horizontal separation, and z_1 and z_2 are the vertical coordinates of two locations 1 and 2 in the Cartesian coordinate system. In Eq. (2), the vertical separation $(z_1 - z_2)$ is up-weighted by a factor 2 to account for the greater variability of a potential field in the vertical than in the horizontal direction. The weights proportional to the distance from the margins of the grids were chosen. A value of 0.01 on the margin and 1.0 at 200 km inland were used. Although the line-leveling technique was effective in minimizing the misfit between adjacent tracks, it also introduced spurious anomalies with half-wavelengths larger than 200 km. However, these spurious long-wavelength anomalies were later replaced by those from MF5 model.

The implementation of the technique first required the decimation of 31,000 tracks of NGDC and Project Magnet data into 98,000 segments. These track line data were then line-leveled against each other by estimating 110,000 correction coefficients ($b_{j,k}$) from one grand inversion. This procedure also minimized offsets where tracks overlapped with gridded data. The segments that were affected by strong external field disturbances were removed when differences between two adjacent segments exceeded 550 nT and misfit against the gridded data exceeded 500 nT. The line-leveling and selection procedure further reduced the rms misfit between two adjacent segments from 380 to 130 nT and against the gridded data from 400 to 90 nT.

Least Squares Collocation

LSC is a scheme for interpolation by the weighted average of the measurements similar to the method called Kriging. The weighted averages are obtained in a minimum-variance sense. The technique of LSC has been used to solve a wide variety of problems in physical geodesy (Moritz 1980). It has been especially useful in certain satellite missions applications, such as satellite altimetry and satellite-to-satellite tracking. The LSC method is discussed in Langel and Hinze (1998) and has been applied by Maus et al. (2007b) to compute the NGDC global magnetic anomaly. Since locations of the measurement points in both track line and gridded data are not on a regular grid, one requires an interpolation procedure to estimate the field at regular grid nodes.

Following Moritz (1980), the LSC method estimates vector components of magnetic anomaly field values $\mathbf{b}$ on a predefined grid from a given set of vector magnetic measurements $\mathbf{c}$ at nearby locations. The vector of measurements, $\mathbf{c}$, is

defined as the sums of the true anomaly vector $\mathbf{a}$ plus an associated noise vector $\mathbf{e}$ (i.e., $\mathbf{c} = \mathbf{a} + \mathbf{e}$). To estimate the value of $\underline{b}$, covariance matrices are first computed as follows:

$$V_a = E(aa^T), \quad V_b = E(bb^T)$$

$$V_{b,a} = E(ba^T), \quad V_d = E(ee^T)$$

where E indicates the expected value. The collocation requires the estimation of the anomaly field $\underline{b}$ in the least square sense from

$$\underline{b} = V_{a,b}^T V_c^{-1} c, \quad (3)$$

where $V_{b,a} = (V_{a,b})^T$ and

$$V_c = E(cc^T) = V_a + V_d. \quad (4)$$

Here, it is assumed that $\mathbf{e}$ and $\mathbf{a}$ are uncorrelated, so that $V_{a,e} = V_{e,a} = 0$; and V_d is the error covariance matrix of measurement noise. The efficiency of the LSC method then lies in the computation of the covariance matrices V_a , V_b , and $V_{a,b}$ from properties of the true signal ($\mathbf{a}$) and the estimation of V_d from the properties of the noise.

For a Cartesian approximation of the data volume, Langel and Hinze (1998) suggest approximate correlation functions V_2 and V_3 between two locations at (x_1, y_1, z_1) and (x_2, y_2, z_2) of the form

$$V_2(r_{1,2}, z_1, z_2) = V_0 \beta \left(r_{1,2}^2 + (z_1 + z_2 + \beta)^2 \right)^{-1/2}, \quad (5)$$

where $\beta = r_c / \sqrt{3}$, and

$$V_3(r_{1,2}, z_1, z_2) = V_0 \beta^2 (z_1 + z_2 + \beta) \left(r_{1,2}^2 + (z_1 + z_2 + \beta)^2 \right)^{-3/2}, \quad (6)$$

where $\beta = r_c / 0.766$.

Here, the variance V_0 is a linear scale factor for the horizontal axis, while the correlation length r_c is a scale factor on the vertical axis. For magnetic anomaly fields, the use of one of these correlation functions for collocation depends on the variation of the function V_2 and V_3 with distance. Based on correlation functions for NGDC magnetic data, Australia and North America, empirically, it was determined that V_3 gives a better fit than V_2 . From V_3 , the values of V_0 and r_c are estimated to be 40,000 nT² and 15 km, respectively.

The value of $\underline{b}$ from Eq. (3) can then be estimated by computing the matrix $V_{a,b}$ given by (Maus et al. 2007b)

$$V_{a,b} = V_3(r_{a,b}, z_a, z_b), \quad (7)$$

and the matrix V_c from Eq. (4)

$$V_c = V_3(r_{1,2}, z_1, z_2) + E(ee^T), \quad (8)$$

are evaluated for all pairs of estimation (b) and measurement (a) locations. The error covariance matrix of measurement noise $E(ee^T)$ arises due to differences between adjacent tracks and between a track and a gridded data which undermine all other sources of errors in the data. After line-leveling, it was assumed the uncorrected difference is 25 nT and the error covariance matrix was set to 625 nT² for data pairs from the same set, and 0 nT² for data pairs from different sets. The line-leveled data were finally merged with the gridded data using the LSC method. This grid is called the final merged grid.

Merging Satellite Models

Finally, the NGDC candidate model was derived by combining the final merged grid with MF5 by computing the magnetic potential. As a first step, the magnetic potential of the line-leveled data was computed to spherical harmonic degree 120. The degree correlation (c.f., Langel and Hinze 1998) between this model and MF5 was shown to be about 0.6 at degree 100. This high correlation offered some validation of high-degree coefficients of MF5. The correlation suggests that CHAMP satellite data is able to resolve many small-scale features beyond degree 100 which could be crustal in origin. The power spectra of line-leveled data show excess power at low-degree (Singh 2011). This could be due to the remaining main field in the continental-scale grids and errors that are introduced in the long-wavelengths during stitching together the small-scale aeromagnetic surveys. At high degrees, the line-leveled data has 10% more power than MF5. Following a good spectral agreement between the line-leveled and MF5 model at degree 100, a sharp cutoff was used to substitute the long wavelengths of the near-surface data with the MF5 model. The substitution is carried out by subtracting the near-surface model to degree 100 from each grid cell and then adding the prediction from MF5. The final NGDC candidate model for WDMAM is a 3 arc-min angular grid of the total intensity anomaly at 5 km above the WGS84 geoid. The digital grid of the NGDC candidate for WDMAM also referred to as EMAG3 model is available at <http://geomag.org/models/wdmam.html>, and additional plugins can be downloaded from <http://www.getech.com/downloads/WDMAM> for visualization in NASA World Wind.

Other Candidate Models for WDMAM

There were four other candidate models for WDMAM. Two of those models, the GAMMA model produced by GeoForschungsZentrum, Potsdam (Hamoudi et al. 2007), and the Leeds model derived by the team at University of

Leeds, Leeds (Hemant et al. 2007), are published. These models are described next.

The Gamma Model

Most techniques were designed to improve the compatibility of the overlapping areas of neighboring grids. Hamoudi et al. (2007) applied different procedures to improve this incompatibility. These procedures are useful and easy to implement if only a few grids are used, for they require only first order estimation of improved long-wavelength values.

Prior to processing, the gridded and track line data were statistically analyzed to check for their self-consistency. For a pure anomaly field, one expects a symmetric distribution with zero mean. A skewed anomaly distribution not centered on zero possibly indicates some more or less severe contamination of other magnetic field contributions. In this respect, Russian, East Asian, and European compilations showed large anomaly intensity means. The standard deviation of anomaly distribution was usually between 100 and 200 nT for continents but could be as large as 930 nT on oceans. Oceanic magnetic anomalies of remanent origin have amplitudes larger than on continents, but such a wide distribution can only be explained by noise, bad tracks, and outliers in the data. Another criterion, based on the arithmetic sum, highlighted some bias in the marine, the Russian, Australian, and the East Asian compilations. These statistics were used to provide the available data and compilation with a quality flag.

The next procedure devoted to the analysis of data compatibility in the overlapping regions was based on Visual Inspection. By binning the data on a coarse grid, the spurious isolated points or track lines could be easily identified and manually removed. Other track lines, having consistent values but suspicious behavior, such as a few Project Magnet data crossing the African continent or possibly biased data offshore Senegal were removed. Some GETECH South American and African compilations showed unusual shapes like over Bangui. Considering that these grids were coarser than any others, it was decided to allow them a low weight in the final process. Similarly, the Indian grid based on repeat stations showed poor resolution, so magnetic map boundaries of Qureshy (1982) on the eastern coast were applied. The European grid showed unexpected geographical shifts by a few kilometers, in particular over Germany, so the grids were used for analysis only when no other data were available in that region. In general, most grids suffered from edge effects, possibly resulting from some kind of filtering at the last stage of their production. Some of these errors were identified but could not be systematically corrected due to lack of complete metadata information.

The third procedure was to carry out an arguably homogeneous correction for the core magnetic field. It has been customary to correct all aeromagnetic surveys with IGRF/DGRF core field models (*International Geomagnetic*

Reference Field, Magnetic Anomalies, Interpretations). Older surveys were corrected for more inaccurate core field models, particularly for those carried out at epochs when rapid core field variations were not yet identified (*Geomagnetic field, secular variation*). The core field correction was thus often followed by two-dimensional polynomial fitting or a Cartesian Fourier filtering. The DGRF models or polynomial fitting parameters were added to such old surveys whenever they were known, and a more recent core field model correction such as CM4 was used (Sabaka et al. 2004). When the core field models were unknown, a better solution was found by trials and errors. A DGRF core field model for different epochs was added to each of these surveys, and the CM4 was removed. When it improved the statistics (arithmetic mean, histogram of residuals), this correction was deemed fruitful. This procedure was not applied only for the Russian grid because it was outside of the CM4 time span (1960–2002). This procedure was shown to improve the data statistics by lowering the mean intensity anomaly toward zero and the continuity in overlapping areas between adjacent compilations and other artifacts caused by the different model biases. For countrywide compilations with a clear knowledge of the acquisition date, in France, Italy, and Spain, this procedure provided a better way to correct for secular variation between different compilations. For some other surveys, however, for which the exact epoch and corrections were poorly known, a regional polynomial adjustment was necessary. In order to avoid Cartesian distortion due to the Earth's curvature, a regional second-order polynomial from all grids in the WGS84 geographic reference was removed. This polynomial fitting was not applied to the Australian grid, where the long-wavelength control was good, or to compilations over geographic poles (Arctic and Antarctic grids), where the processing procedure was based on using Cartesian reference frame instead. After this correction, the grids were shown to have the expected properties characterizing the anomaly field with the average anomaly intensity and an arithmetic sum zero (Singh 2011).

The fourth step consisted of upward continuing of the data to the desired altitude of 5 km. Marine track line data sets were also upward continued to mitigate effects of the sharp transition at the ocean-continent boundaries. Grids smaller than 2000 km width, such as Argentina, Austria, Fennoscandia, Italy, France, Mexico, and Spain, were upward continued to 5 km altitude above the geoid without any further processing. However, larger compilations (>2000 km) were first split into 2000 × 2000 sq. km and upward continued individually. This dimension corresponds to the maximum size for which the Earth's curvature can be neglected (Nakagawa et al. 1985). After the upward continuation, the edge effects between adjacent grids were not significant.

The preliminary processing discussed earlier reduced the large discontinuities but did not fully remove them. The Grid-

knitting tool of GeoSoft was used to smooth the transition between adjacent grids. Some spurious wavelengths were created but were filtered out later. Merging large compilations with small compilations includes a risk because the adjustment is better constrained by the large compilation even if the small one has apparently a better quality. Before merging the grids, compilations with comparable grid sizes were generated and more weight was given to the grid showing the best statistical properties. Therefore, French, Italian, and Spanish grids were first merged together, followed by building a second grid, the Finland, Fennoscandian, European, and Austrian grids. These two new compilations were merged together and the grids from Russia, Eurasia, Middle East, Antarctica, and North America were successively added to the compilation. The remaining grids that did not overlap with the grids listed above were simply added in the corresponding geographical areas. This generated a $3' \times 3'$ global grid using GMT (<https://www.soest.hawaii.edu/gmt/>).

Finally, the spurious intermediate wavelengths introduced by the merging were filtered out by a spherical harmonic transformation. The wavelengths larger than those corresponding to spherical harmonic degree 100 were filtered out of the global grid and were replaced with those from the downward-continued MF5 model. The candidate model along with accompanying documentation is available at <http://www.earthref.org>.

The Leeds Model

To check for inconsistencies in the gridded data, the strength of total field intensities across the two overlaying grids is compared. Within the overlapping regions of Eurasia and Europe, total field intensities disagree in strength by 10–200 nT. At the grid boundaries, these intensities vary by 100–150 nT. These differences indicate possible errors in the long-wavelength content of the two grids. The discrepancies could also be attributable to the relative difference in the anomaly strength and resolution of the original grid. These differences are also observed between the Arctic and the North America compilations. Continental and oceanic compilations also show discontinuities in many regions.

For minimizing differences in the anomalies between various grids, a low-pass Gaussian filter with a cutoff wavelength of 400 km is applied in the 2-D space domain using GMT software. The cutoff wavelength is set to 400 km because the longer wavelengths are later substituted from the MF5 model, which is robust up to spherical harmonic degree 100 (~400 km). The filtered output is a long wavelength grid (>400 km) and strongly resembles a satellite-derived magnetic anomaly map. The short wavelength grid is computed by subtracting the long wavelengths from the original grid and shows finer details of the regional geology and structure of the crust (Singh 2011).

To analyze the systematic errors in the long wavelengths between adjacent overlapping grids, the difference in the long wavelengths is computed. A visual inspection shows differences in the anomalies exceeding hundreds of nT and often shows a lack of continuity across two or more grids. Because it is difficult to judge the trustworthiness of an overlapping long-wavelength grid, it was best suited to replace the filtered long wavelength with a reliable satellite model.

Techniques that merge two adjacent compilations often employ a weighted average of different-order surfaces such that the two adjoining maps are seamless within the merging area. These processes modify the grids within the overlaying regions and near the edges and alter the strength and pattern of anomaly features. In this process, anomalies are modified, or at times genuine crustal anomaly features are even entirely eliminated. The merging technique applied to merge short-wavelength magnetic anomaly data minimizes abrupt changes along boundaries of overlapping grids. Using the MOSAIC toolbox (<http://www.esri.com/>), the grid values within the overlapping regions were determined from Hermite Cubic functions (Franke 1982) and the short wavelength grids merged.

The short wavelength anomaly differences between grids vary substantially; hence, it is difficult to prescribe the order of precedence for merging. However, the quality of an individual compilation is also reflected in the long-wavelength differences. Therefore, the order of precedence of the grid merging was based on the difference between long wavelengths and the quality of the parent grid. Thus, the short-wavelength grids were merged in the following order: Australia, North America, Austria, Canary Island, East Asia, Eurasia, Europe, Arctic, Spain, Fennoscandia, France, Russia, Argentina margin and inland, South Africa, Antarctica, Italy, India, Middle East, Africa, and South America, NGDC track line and Project Magnet data. The grids were merged beginning with the highest precedence. The accuracy and reliability of the merged grid is assessed by investigating the effect of merging on the continuity of anomaly patterns across two or more grids and the effects on the anomaly shapes and strengths within and near the grid boundary. The differences for the merged European and Eurasian grid ranged from ± 100 to 200 nT and $< \pm 100$ nT for the remaining grid. A small difference of 8 nT was noted between Australian and East Asian grids indicating consistent anomaly patterns between the pair.

Finally, the total field anomalies of the short-wavelength merged magnetic anomaly map were added to the downward-continued total field anomaly of the MF5 model to produce the candidate model of the LEEDS-WDMAM (Singh 2011).

Earth Magnetic Anomaly Grid (EMAG2)

As an improvement to the WDMAM, track line data were added in the southern oceans, where the data gaps were significant. For WDMAM, these gaps in the ocean were filled with magnetic anomalies computed from ocean age model and geomagnetic polarity time scale. EMAG2 (Maus et al. 2009) was extrapolated into the unsurveyed areas using directional gridding based on the ocean age model of Müller et al. (2008). To further enhance the accuracy of the model over oceans, instead of using precompiled ocean magnetic anomaly grids, original track line data were used where available. Line leveling algorithm was the primary method to minimize crossover errors. The new Arctic compilation (Circum-Arctic Mapping Project) was used instead of the old Arctic compilation. A new crustal magnetic field model derived from CHAMP data, the MF6 model (Maus et al. 2008), is used for EMAG2 to correct the long wavelength anomaly field (wavelengths >330 km) replacing the MF5 model used for WDMAM model.

EMAG2 is 2 arc min grid defined at a height of 4 km above the geoid. The ocean data includes NGDC's GEODAS marine data archive, Antarctic Digital Magnetic Anomaly Project data (Golynsky et al. 2001), and the Project Magnetic Airborne data of the Naval Research Lab (NRL).

The line-leveling algorithm remains the same as described earlier. However, the search radius was reduced from $R_s = 100$ to 8 km in order to minimize the crossover effects in processing the track line data. In addition, the number of correction coefficients per track is now $N_i = \text{trunc}(X_i/300 \text{ km}) + 1$. These efforts reduced the crossover errors from 92 to 70 nT and reduced the RMS misfit to the merged grid from 121 to 97 nT.

Least squares collocation method requires estimation of the correlation function. No new correlation functions was computed for land areas as the gridded compilation remained the same. Notable changes for oceans, however, were made. For instance, for Australia, 100 m terrain clearance was used. Due to the inclusion of more accurate grids and original marine and airborne surveys, the assumed variance V_0 was reduced from 40,000 to 33,000 nT and the correlation length r_c from 15 to 14 km.

The methodology of LSC was slightly modified for EMAG2. The correlation analysis now considered the anisotropy of oceanic magnetic anomalies. Track line data, just the way it was processed for NGDC candidate model, were first divided into small windows but along the constant heading direction of the track. Next, for each track, the azimuth for every pair of measurement points was calculated and compared with the direction of the oceanic isochrons at the locations of both points. The pair was removed from further processing if they met one of the following criteria: (1) the azimuth of the isochrons was not well defined, (2) the azimuth

differed by more than 5° between the points, or (3) the topographic gradient between the two points exceeded 3%. The NGDC ETOPO-2 bathymetry values were used to compute the topographic gradient. This gradient could detect and remove the magnetic anomalies due to seamounts, which otherwise would appear as noise. The anisotropic correlation function was then computed in 10° directional bins; from 0° (parallel) to 90° (perpendicular, i.e., in the spreading direction) to the isochrons. This anisotropic correlation model allowed the LSC method to be used as a directional gridding algorithm.

From the analysis of covariance function in the isochron direction, it was noted that the function value reduced too rapidly. Furthermore, extrapolating the field far from a track line made the function reduce to zero. This produced erroneous stripes parallel to the track lines. To stabilize the covariance function, the correlation length was doubled when gridding the ocean areas. This made the anisotropy of the oceanic field in EMAG2 to be smoother than it perhaps is.

EMAG2 also suffers inaccuracies from the oceanic crustal age model due to errors in the estimated ocean isochrons. This error could easily have contaminated those estimated oceanic magnetic anomalies of EMAG2, where the uncertainties in isochron age are large. To reduce these uncertainties, empirically, the anisotropy factor was increased fourfold for all oceanic isochrons of age <140 Ma and a linear reduction to a factor 2.25 for ages >150 Ma was used. This procedure helped reduce the errors in EMAG2 within the oceans, in particular over the older crust.

For compiling the final EMAG2 grid, the portion of the original total field anomaly up to degree 120 is replaced by that of the MF6 model. A spectral comparison between the original and modified grids shows a slight discrepancy above degree 120. This is perhaps due to the leakage of long-wavelength power into shorter wavelengths. But the overall comparison generally validates the power content in EMAG2. The EMAG2 grid can be downloaded from <http://earthref.org/cgi-bin/er.cgi?s=erda.cgi;n=970> or from <http://geomag.org/models/EMAG2>.

New shipborne and trackline measurements were made available for the study of oceanic crust which led to the derivation of a significantly improved version of EMAG2, called the EMAG2V3. It represents magnetic anomalies at sea level in the ocean regions, and the wavelengths greater than 300 km were replaced with the CHAMP lithospheric field model, MF7 (Maus 2010a). The EMAG2V3 grid can be downloaded from <https://www.ngdc.noaa.gov/geomag/emag2.html>.

NGDC-720 Lithospheric Magnetic Model

In the last decade, successful attempts have been made to improve the resolution of the global magnetic anomaly

maps. The expansion of scalar magnetic potential into ellipsoidal harmonics in the least squares sense (Maus 2010b) led to the development of NGDC-720 model that represented anomaly of the total intensity of the field from degrees 16 to 719 corresponding to the variation of the wavelength from 2500 to 56 km. Newer versions of the models are available at <https://geomag.us/models/ngdc720.html>. The wavelengths larger than 330 km were

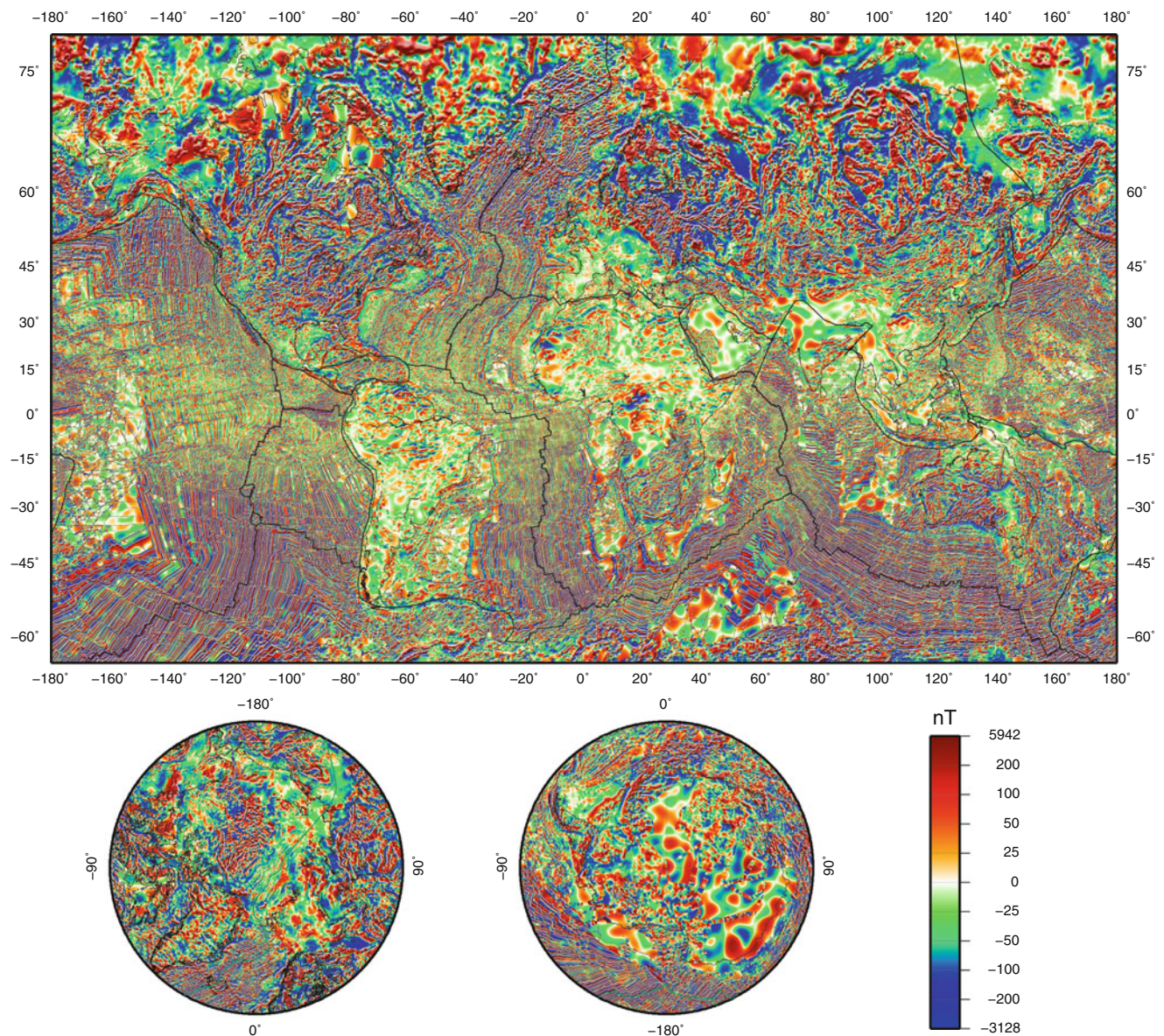
from the MF6 lithospheric field model (Maus et al. 2008).

Enhanced Magnetic Model

In the last half-a-decade, the Enhanced Magnetic Model (EMM) started replacing the NGDC-720 models. The

Magnetic Anomaly Map, Global, Table 1 Data sets, their statistics and sources, used for deriving the WDMAM2.0 model. (Courtesy: Lesur et al. 2016)

Name	N	Mean	Std dev.	Min/max	Index	Sources
Marine data	7,193,664	−4	110	−2057/2627	11	Quesnel et al. (2009)
France	1,488,585	0.54	32.00	−156/396	12	Le Mouél, Ann. Geophys., 26(2), 1970; Courtillot et al., Earth Planet Sci. Lett., 47, 1980
Italy	439,320	11.8	59.02	−435/588	13	Chiappini et al., Ann Geophys, 45(5), 2000
Spain	494,171	0.02	16.66	−62/184	14	Socias, Earth Planet Sci. Lett., 105, 1991
Germany	20,815	−2.62	36.70	−96/171	15	Gerald.Gabriel@liag-hannover.de
Austria	5294	12.6	26.12	−35/137	16	GSA (http://www.geologie.ac.at/)
Fennoscandia	3,505,857	−30	154	−580/1500	17	J. Korhonen, GTK (http://en.gtk.fi/)
United Kingdom	119,509	−2.71	121	−962/1342	18	British Geological Survey, NERC, IPR/132-01CT
Portugal	151,142	−10	33	−111/219	19	Miranda et al., Earth Planet Sci. Lett., 9, 1989
Romania	100,092	19	70	−144/418	20	Dr. L. Besutiu, Institute of Geodynamics, Romania (http://www.geodin.ro/)
Japan (2007)	91,094	−42	73	−401/416	21	http://www.ccop.or.th/
East Asia	958,040	−28	68	−509/763	22	http://www.ccop.or.th/
Russia (VSEGEI)	168,8415	0.9	158	−2274/5658	23	T. Litvinova, VSEGEI (http://www.vsegei.ru/en)
Djibouti	266,171	−2.8	85	−293/241	24	Courtillot et al., Earth Planet Sci. Lett., 47, 1980
Middle-East (2007)	1,812,464	0.56	73	−904/690	25	http://members.casema.nl/errenwijlens/itc/aaime/
India	92,784	18.5	82	−341/427	26	GSI (http://www.gsi.gov.in/)
Circum-Arctic	3,258,710	−4	99	−699/2754	27	Gaina, C., Werner, S. C. and the CAMP-GM group, NGU Report 2009.010
Guadeloupe	353,672	24	46	−76/190	28	IPG, Paris, France (http://www.ipgp.fr/)
French Guiana	99,066	−3.9	72	−283/143	29	BRGM, France (http://infoterre.brgm.fr/)
Bahama-Cuba	329,975	0.9	75	−410/681	30	Batista-Rodriguez et al., Geophysics, 72(3), https://doi.org/10.1190/1.2712425
Nure-Namag	2,012,238	−11	123	−3667/3998	31	Ravat et al., USGS open-file report 2009-1258
Algeria	2,126,820	0.2	48	−189/274	32	Ministry of Energy and Mines, Algiers, Algeria
Morocco	394,179	−1.4	43	−194/207	33	Ministry of Energy and Mines, Rabat, Morocco
Nigeria	132,407	26.3	40	−163/177	34	Nigerian Geological Survey, Abuja, Nigeria
Argentina-Shelf	1,497,947	−9	45	−150/318	35	Max et al., Marine Petroleum Geology, 16, 1999
Argentina-Inland	34,134	−1.5	104	−3152/1173	36	SEGEMAR (http://www.segemar.gov.ar/)
Australia	7,089,164	−41.2	137	−1674/3360	37	Geosciences Australia (http://www.agso.gov.au/map/magnetics_info.html/)
South African DC	427,270	−3.2	120	−1788/2030	38	SADC (http://www.sadc.int/)
Uganda	123,020	−83	116	673/301	39	Ministry of Energy and Mineral Development, Uganda
Mozambique	419,248	159	171	−409/1007	40	Geologian Tutkimuskeskus, Espoo, Finland
SaNaBoZi	83,380	77	109	−203/756	41	SADC (http://www.sadc.int/)
Acad Vernadsky	69,123	170	166	−180/426	42	http://wdmam.org/pdf/WDMAM_data_contrib.pdf
Admap	56,125,472	15	104	−776/1371	43	http://www.antarctica.ac.uk/about_antarctica/geography/maps/admap.php/
Marine Model	901,539	−5	152	−1662/1639	44	Dyment et al., EPSL, 430, 2015
Eurasia (2007)	540,898	−2	102	−891/819	45	GSC (http://gsc.nrcan.gc.ca/)
Getech (2007)	1,038,080	1	59	−1024/536	46	GETECH (http://www.getech.com/world_data.php/)
GRIMM (2007)	3,645,256	8	55	−235/580	47	Lesur et al., Solid Earth, 4, 2013



Magnetic Anomaly Map, Global, Fig. 2 The World Digital Magnetic Anomaly Map version 2.0 (WDMAM2.0) publicly released by 26th IUGG General Assembly in June 2015. (Courtesy: Lesur et al. 2016)

EMM2015 which extended to SH degrees 790 was able to resolve magnetic anomalies up to 51 km wavelengths, at least theoretically. The EMM2015 model was derived by merging ground magnetic measurements with the near-surface compilations and satellite data. An updated version, the EMM2017, was released on July 05, 2017, which included the European Space Agency's (ESA) three-satellite Swarm mission data. The EMM2017 is also an improved version of EMAG2V3 model and represents anomalies to Spherical Harmonic degree 790. The two models EMM2015 and EMM2017 differ primarily in oceanic areas owing to the different methodologies used in the derivation of their respective parent, EMAG2 and EMAG2V3, models. The EMM2017 is valid from 2000.0

to 2022.0 and has been developed primarily for the purpose of navigation. The model can be downloaded from <https://www.ngdc.noaa.gov/geomag/EMM/>.

World Digital Magnetic Anomaly MAP Version 2.0 (WDMAM2.0)

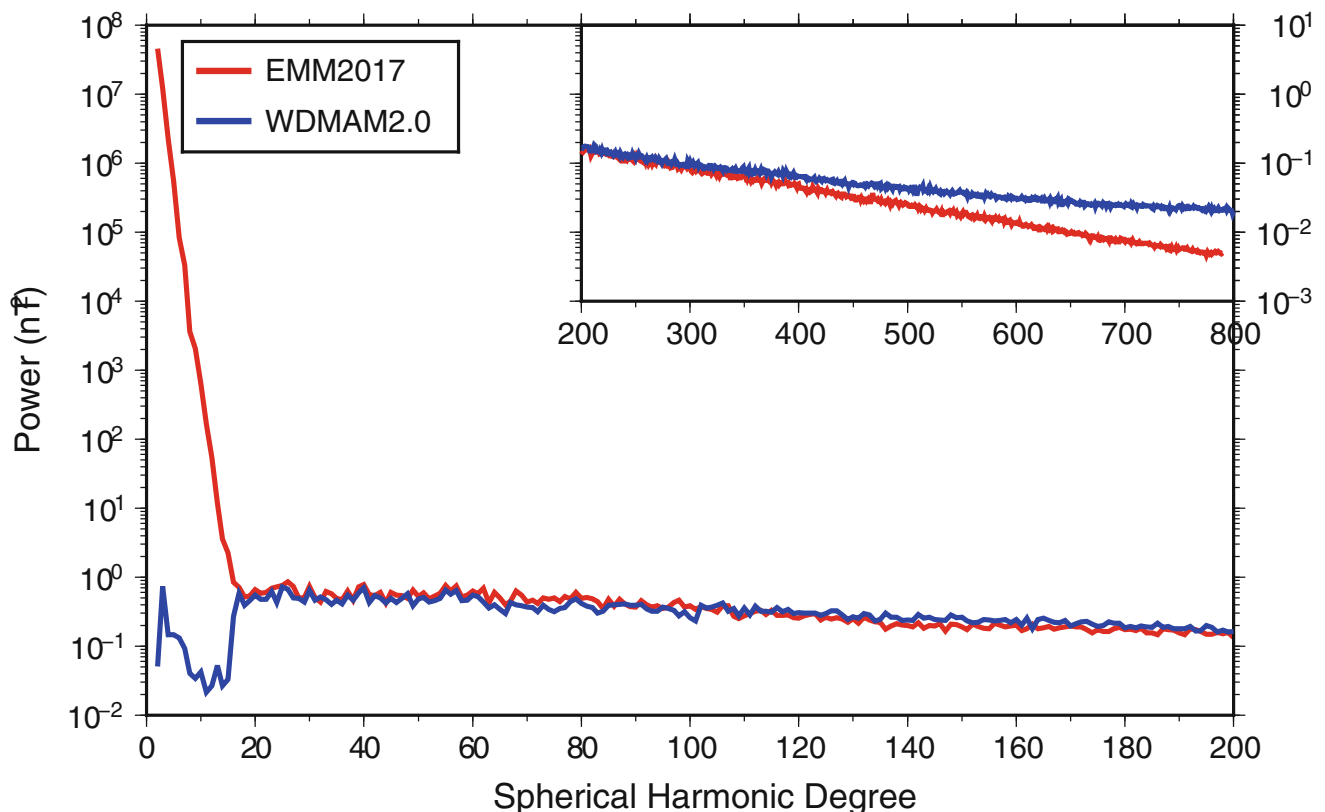
When the WDMAM project was conceptualized, it was envisaged to have WDMAM to be an ongoing project, with plans to produce updates of the map and grid every 4 years. As part of these updates, it was envisioned that more organizations will make additional data available for new versions of

WDMAM, in particular over regions where large data gaps exist. Toward this, Maus et al. continued to collect and process the data and produced their own maps, e.g., NGDC-720, EMAG2, and higher versions of the same. Thus, a need to derive the candidate models for WDMAM2.0 with the inclusion of data with prior agreement of the data owners was agreed upon.

After the release of the first version of the WDMAM (Korhonen et al. 2007), new aeromagnetic compilations over the continents and trackline data in the oceanic regions were released for the preparation of the second version of the WDMAM, or referred to as WDMAM2.0 (Table 1). Only one candidate model developed jointly by two independent groups, one at the GeoForschungsZentrum in Potsdam Germany, and the other at the Institut de Physique du Globe in Paris, France, was submitted for evaluation. The processing technique for the derivation of WDMAM2.0 (Lesur et al. 2016) was same as that followed for the derivation of the candidate GAMMA model (Hamoudi et al. 2007). All continental compilations were upward continued to 5 km above the WGS84 datum before merging them with the marine magnetic anomalies derived at the sea surface (0 km). The data gap over the marine areas has been filled up with the magnetization model of the ocean floor (Dyment et al. 2015). The

final merged grid is of resolution 0.05×0.05 degrees which are roughly equivalent to 5×5 km cell size on the Earth's surface. The resulting WDMAM2.0 grid (Fig. 2) represents 25,927,200 total intensity values computed relative to the CM4 field calculated for epoch 1990 and is derived up to SH degrees 16–800, of which the long wavelengths corresponding to SH degrees 16–100 are corrected using lithospheric field model GRIMM_L120 model (Lesur et al. 2013). The WDMAM2.0 was approved by IAGA during the 26th IUGG assembly in Prague, June 2015, and was officially released for its use by the scientific community (<http://wdmam.org/>).

The power spectra of EMM2017 and WDMAM2.0 models for spherical harmonic degrees 1–800 (degrees 1–790 in the case of EMM2017) are plotted in Fig. 3. A prominent low power is noticed in WDMAM2.0 model from degrees 1 to 15 compared to EMM2017. This is because the main field (*Geomagnetic field*, IGRF) has been removed in WDMAM2.0 model while EMM2017 retains it. Instead, the long wavelength anomalies from the Earth's lithosphere can be seen for degrees <15 , which otherwise remain hidden by the dominating main field of the Earth. The power from degrees 15 to 105 shows marginally higher power in EMM2017 compared to those of WDMAM2.0, which is



Magnetic Anomaly Map, Global, Fig. 3 Power spectra of the Enhanced Magnetic Model (EMM2017) and WDMAM2.0 models computed from spherical harmonic. (Courtesy: Lesur et al. 2016)

possibly due to unremoved main field in continental-scale grids and erroneous long-wavelength anomalies introduced in stitching together small-scale grids. Beyond the spherical harmonic degree 105, WDMAM2.0 shows consistently higher power than EMM2017, which becomes more pronounced beyond degree 250. This could perhaps be due to the addition of ocean magnetization model (Dyment et al. 2015) to fill data gaps in the marine region during the preparation of WDMAM2.0, something that is not a part of EMM2017 model. Again, considering the different data sources and their uncertainties and various data processing steps and corrections, the disagreement in power ($<0.1 \text{ nT}^2$) is judged small.

Summary

In summary, with methods discussed above, it will be possible to regularly update WDMAM model without excessive effort to reliably synthesize a map that accurately represents the world's magnetic anomalies. This would also require information on the long-wavelength field in order to continuously improve the representation of the world magnetic anomalies. The data from Swarm satellite mission of the European Space Agency (<http://www.esa.int/esaLP/LPswarm.html>) has already considerably improved the lithospheric magnetic anomaly maps by providing highly accurate measurements of magnetic field gradients and will likely further improve the long-wavelength anomalies in next generations of WDMAMs and EMMs. It is our hope that these global magnetic anomaly maps will enjoy widespread usage, and inspire further aeromagnetic and marine data holders to release the data, thus helping to produce maps having truly global coverage. Finally, the WDMAM and WDMAM2.0 would not have existed without the willingness of institutions to constantly revise and release their national scale compilation.

Cross-References

- Continental Crustal Structure
- Continental Drift
- Continental Rifts
- Core Dynamo
- Geomagnetic Field, Global Pattern
- Geomagnetic Field, IGRF
- Geomagnetic Field, Measurement Techniques
- Geomagnetic Field, Polarity Reversals
- Geomagnetic Field, Secular Variation
- Geomagnetic Field, Theory
- Lithosphere, Continental: Thermal Structure
- Lithosphere, Oceanic
- Magnetic Anomalies: Interpretation

- Magnetic Data Enhancements and Depth Estimation
- Magnetic Methods, Airborne
- Magnetic Methods, Principles
- Plates and Paleoreconstructions
- Remanent Magnetism

Bibliography

- Bankey VA, Cuevas D, Daniels AA, Finn I (2002) Hernandez and project members, digital data grids for the magnetic anomaly map of North America, USGS (2002), open-file report 02-414
- Batista-Rodríguez JA, Pérez-Flores MA, Quiroga-Goode G, Gallardo LA (2007) Geometry of ophiolites in eastern Cuba from 3D inversion of aeromagnetic data, constrained by surface geology. *Geophys* 72: B81–B91 <https://doi.org/10.1190/1.2712425>
- Chiappini M, Meloni A, Boschi E, Faggioni O, Beverini N, Carmisciano C, Marson I (2000) Shaded relief magnetic anomaly map of Italy and surrounding marine areas. *Ann Di Geofis* 43 (5):983–989
- Courtillot V, Galdeano A, Le Mouél JL (1980) Propagation of an accreting plate boundary: discussion of a new aeromagnetic survey of the Republic of Djibouti. *Earth Planet Sci Lett* 47:144–160
- Dyment J, Choi Y, Hamoudi M, Lesur V, Thébaud E (2015) Global equivalent magnetization of the oceanic lithosphere. *Earth Planet Sci Lett* 430:54–65. <https://doi.org/10.1016/j.epsl.2015.08.002>
- Franke R (1982) Smooth interpolation of scattered data by local thin plate splines. *Comput Math Appl* 8(4):273–281
- Gaina C, Werner SC, CAMP-GM group (2009) Circum-Arctic Mapping Project – Gravity and Magnetic Maps (CAMP-GM) Report no.: 2009.010 https://www.ngu.no/upload/Gaina_et_al_CAMP_Report_2009_010.pdf. Last Accessed 01 July 2020
- Gee JS, Kent DV (2007) Chapter 12: Source of oceanic magnetic anomalies and the geomagnetic polarity time scale. In: Kono M (ed) *Geomagnetism: treatise on geophysics*, vol 5. Elsevier, Amsterdam. (in press)
- Golynsky A et al (2001) ADMAP – magnetic anomaly map of the Antarctic, vol. 10, scale 1:10,000,000. British Antarctic Survey, Cambridge
- Hamoudi M, Thebaud E, Lesur V, Manda M (2007) GeoForschungsZentrum Anomaly Magnetic Map (GAMMA): a candidate model for the World Digital Magnetic Anomaly Map. *Geochem Geophys Geosyst*. <https://doi.org/10.1029/2007GC001638>
- Hemant K, Thebaud E, Manda M, Ravat D, Maus S (2007) Magnetic anomaly map of the world: merging airborne, marine and ground-based magnetic data sets. *Earth Planet Sci Lett* 260:56–71. <https://doi.org/10.1016/j.epsl.2007.05.040>
- Kent DV, Gradstein FM (1986) Jurassic to recent chronology. In: Vogt PR, Tucholke BE (eds) *The western North Atlantic region, Geology of North America*, vol M. Geological Society America, Boulder, pp 45–50
- Korhonen JK, Fairhead JD, Hamoudi M, Hemant K, Lesur V, Manda M, Maus S, Purucker ME, Ravat D, Sazonova T, Thébaud E (2007) Magnetic anomaly map of the world – carte des anomalies magnétiques du monde, scale: 1:50,000,000, 1st edn. Commission for the Geological Map of the World, Paris
- Langel RA, Hinze WJ (1998) *The magnetic field of the Earth's lithosphere: the satellite perspective*. Cambridge University Press, Cambridge
- Le Mouél JL (1970) Le levé aéromagnétique de la France. Calcul des composantes du champ à partir des mesures de l'intensité. *Ann Geophys* 26(2):229–258
- Lesur V, Maus S (2006) A global lithospheric magnetic field model with reduced noise level in polar regions. *Geophys Res Lett* 33:L13304

- Lesur V, Rother M, Vervelidou F, Hamoudi M, Thébault E (2013) Post-processing scheme for modeling the lithospheric magnetic field. *Solid Earth* 4:105–118. <https://doi.org/10.5194/sed-4-105-2013>
- Lesur V, Hamoudi M, Choi Y, Dyment J, Thébault E (2016) Building the second version of the World Digital Magnetic Anomaly Map (WDMAM). *Earth Planets Space* 68:27. <https://doi.org/10.1186/s40623-016-0404-6>
- Maus S (2010a) Magnetic field model MF7. Available at: www.geomag.us/models/MF7.html. Last accessed 21 Feb 2020
- Maus S (2010b) An ellipsoidal harmonic representation of Earth's lithospheric magnetic field to degree and order 720. *Geochem Geophys Geosyst* 11:Q06015. <https://doi.org/10.1029/2010GC003026>
- Maus S, Luehr H, Martin R, Hemant K, Balasis G, Ritter P, Claudia S (2007a) Fifth-generation lithospheric magnetic field model from CHAMP satellite measurements. *Geochem Geophys Geosyst* 8:Q05013. <https://doi.org/10.1029/2006GC001521>
- Maus S, Sazonova T, Hemant K, Fairhead JD, Ravat D (2007b) National Geophysical Data Center candidate for the World Digital Magnetic Anomaly Map. *Geochem Geophys Geosyst*. <https://doi.org/10.1029/2007GC001643>
- Maus S, Yin F, Luehr H, Manoj C, Rother M, Rauberg J, Michaelis I, Stolle C, Müller RD (2008) Resolution of direction of oceanic magnetic lineations by the sixth-generation lithospheric magnetic field model from CHAMP satellite magnetic measurements. *Geochem Geophys Geosyst* 9:Q07021. <https://doi.org/10.1029/2008GC001949>
- Maus S, Barckhausen U, Berkenbosch H, Bournas N, Brozena J, Childers V, Dostaler F, Fairhead JD, Finn C, von Frese RRB, Gaina C, Golynsky S, Kucks R, Lühr H, Milligan P, Mogren S, Müller RD, Olesen O, Pilkington M, Saltus R, Schreckenberger B, Thébault E, Caratori Tontini F (2009) EMAG2: a 2-arc min resolution Earth Magnetic Anomaly Grid compiled from satellite, airborne, and marine magnetic measurements. *Geochem Geophys Geosyst* 10:Q08005. <https://doi.org/10.1029/2009GC002471>
- Max MD, Ghidella M, Kovacs L, Paterlini M, Valladares JA (1999) Geology of the Argentine continental shelf and margin from aeromagnetic survey. *Mar Petrol Geol* 16(1):41–64. [https://doi.org/10.1016/S0264-8172\(98\)00063-4](https://doi.org/10.1016/S0264-8172(98)00063-4)
- Minty BRS, Milligan PR, Luyendyk API, Mackey T (2003) Merging airborne magnetic surveys into continental-scale compilations. *Geophysics* 68:988–995
- Miranda JM, Galdeano A, Rossignol JC, Mendes LA, Victor (1989) Aeromagnetic anomalies in mainland Portugal and their tectonic implications. *Earth Planet Sci Lett* 95(1–2):161–172. [https://doi.org/10.1016/0012-821X\(89\)90174-X](https://doi.org/10.1016/0012-821X(89)90174-X)
- Moritz H (1980) Advanced physical geodesy. Herbert Wichmann Verlag, Karlsruhe
- Müller RD, Roest WR, Royer J-Y, Gahagan LM, Sclater JG (1997) Digital isochrons of the world's ocean floor. *J Geophys Res* 102:3211–3214
- Müller RD, Sdrolias M, Gaina C, Roest WR (2008) Age, spreading rates, and spreading asymmetry of the world's ocean crust. *Geochem Geophys Geosyst* 9:Q04006. <https://doi.org/10.1029/2007GC001743>
- Nakagawa I, Yukutake T, Fukushima N (1985) Extraction of magnetic anomalies of crustal origin from Magsat data over the area of Japanese Islands. *J Geophys Res* 90:2609–2615
- Olsen N, Ravat D, Finlay C, Kother LK (2017) LCS-1: a high-resolution global model of the lithospheric magnetic field derived from CHAMP and Swarm satellite observations. *Geophys J Int* 211(3):1461–1477. <https://doi.org/10.1093/gji/ggx381>
- Purucker M, Whaler W (2007) Chapter 6: Crustal magnetism. In: Kono M (ed) *Treatise on geophysics*, vol 5. Elsevier, Amsterdam, pp 195–237
- Quesnel Y, Catalàn M, Ishihara T (2009) A new global marine magnetic anomaly data set. *J Geophys Res* 114(B04106). <https://doi.org/10.1029/2008JB006144>
- Qureshy MN (1982) Geophysical and Landsat lineament mapping – an approach illustrated from West-Central and South India. *Photogrammetria* 37:161–184
- Ravat D, Finn C, Hill P, Kucks R, Phillips J, Blakely R, Bouligand C, Sabaka T, Elshayat A, Aref A, Elawadi E (2009) A preliminary, full spectrum, magnetic anomaly grid of the United States with improved long wavelengths for studying continental dynamics—A website for distribution of data: U.S. Geological Survey Open-File Report 2009–1258, 2 p. <https://pubs.usgs.gov/of/2009/1258/downloads/OF09-1258.pdf>. Last Accessed 01 July 2020
- Ravat D, Whaler KA, Pilkington M, Sabaka T, Purucker ME (2002) Compatibility of the high-altitude aeromagnetic and satellite-altitude magnetic anomalies over Canada. *Geophysics* 67:546–554
- Reeves C (2005) Aeromagnetic surveys: principles, practice and interpretation. GeoSoft, Washington, DC, p 155. <http://www.geosoft.com/resources/news/2009/090723.asp>
- Sabaka TJ, Olsen N, Langel RA (2002) A comprehensive model of the quiet-time near-Earth magnetic field: Phase 3. *Geophys J Int* 151:32–68
- Sabaka TJ, Olsen N, Purucker ME (2004) Extending comprehensive models of the Earth's magnetic field with Oersted and CHAMP data. *Geophys J Int* 159:521–547. <https://doi.org/10.1111/j.1365-246X.2004.02421.x>
- Singh KH (2011) Magnetic, global anomaly map. In: Gupta HK (ed) *Encyclopedia of solid Earth geophysics*, Encyclopedia of Earth sciences series. Springer, Dordrecht
- Snyder JP (1987) Map projections – a working manual, US Geological Survey. Professional Paper, vol. 1395
- Socias I, Mezcua J (1996) Levantamiento aeromagnético del archipiélago canario. Publ. Tec. 35. Inst. Geogr. Nacl., Madrid, p 28
- Socias I, Mezcua J, Lynam J, Del Potro R (1991) Interpretation of an aeromagnetic survey of the Spanish mainland. *Earth Planet Sci Lett* 105(1–3):55–64
- Verhoef JR, Macnab R, Roest W, Arkani-Hamed, J (1996) Geological Survey of Canada, Magnetic anomalies of the Arctic and North Atlantic Oceans and Adjacent Land Areas. Open file 3125a
- Wessel P, Smith W (1998) New, improved version of Generic Mapping Tools released. *EOS Trans Am Geophys Union* 79:579
- Wonik T, Trippler K, Geipel H, Greinwald S, Pashkevitch I (2001) Magnetic anomaly map for Northern, Western and Eastern Europe. *Terra Nova* 13:203–213

Magnetic Data Enhancements and Depth Estimation

Clive Foss

Mineral Resources, CSIRO, Lindfield, NSW, Australia

Synonyms

Magnetic data processing

Definition

Magnetic data are measurements of the magnetic field (generally of its intensity). Corrections of magnetic data are processing operations designed to remove unwanted features in the data. Enhancements of magnetic field data are

processing operations designed to preferentially accentuate the expression of a selected magnetization at the expense of others.

Introduction

Measurement of the earth's magnetic field is used in exploration geophysics as a remote-sensing method to investigate subsurface distributions of magnetization (see ► [“Magnetic Methods, Surface”](#)). The measurements require correction to eliminate artifacts and to remove the primary core field of the earth, thereby improving the expression of local field variations due to upper crustal magnetizations. These corrections are applied to produce clean and robust output suitable for interpretation (see ► [“Magnetic Anomalies: Interpretation”](#)). For specific interpretation objectives, enhancements are also applied to the data to increase the expression of selected sources at the expense of others.

The strong linkage between electricity and magnetic fields allows the development and calibration of magnetometers that supply absolute measurements of the magnetic field (see ► [“Magnetometers”](#)). Most processing corrections are designed to safeguard this integrity of data values and provide data suitable for input to modeling and inversion. Conversely, some enhancement operations distort the data and care must be taken to ensure that enhanced data is not used inappropriately.

Most magnetic field measurements are of the strength of the field, termed “total magnetic intensity” or TMI. Although these measurements are commonly treated as a scalar, it is important to remember that they are the amplitude of a vector, that the direction of that vector is also significant, and that in regions of strong fields the direction may vary across the survey area. The second most common form of magnetic field measurements are gradients of TMI, acquired by simultaneous measurement with two or more TMI sensors at fixed offsets (see ► [“Magnetic Gradiometry”](#)). Other forms of magnetic field data include measurements of the Cartesian components of the field and of the gradient tensor. Details of corrections and enhancement change only slightly between these different types of magnetic field measurements. The mathematical derivation of many correction and enhancement operators is presented in “Potential Theory in Gravity and Magnetic Application” (Blakely 1996) which provides an excellent basis to understand this topic. A brief review by Milligan and Gunn (1997) (http://www.ga.gov.au/image_cache/GA2486.pdf) is also useful reading.

Standard Magnetic Data Processing Corrections

Measurement of the earth's magnetic field provides a spot value at a specific time and location. Magnetic field data is

recorded together with ancillary information such as horizontal position, elevation, and time, as required for processing and interpretation. Corrections of data that are specific to instrumentation are an integral part of data acquisition and are not discussed in this topic. Common corrections in the processing of magnetic data include diurnal correction, leveling, IGRF correction, gridding, and reduction to pole.

Diurnal Correction

The earth's magnetic field is continuously changing (see ► [“Geomagnetic Field, Secular Variation”](#)), and this clearly presents a challenge in mapping spatial variations. Temporal changes of the magnetic field during the progress of a survey are generally termed “diurnal” variation, although this name more correctly describes field variations specifically at the 24 h rotation cycle of the earth. On normal “quiet” days the strength of the geomagnetic field varies across a range of approximately 30–100 nT. The amplitude and pattern of diurnal variation are functions of latitude, but vary sufficiently day to day that they must be measured at a fixed base station for subtraction from the field data. Surveys presenting the greatest challenges for diurnal correction are marine surveys away from suitable locations for a base station or that cover areas of substantial lateral variation in bathymetry or crustal conductivity (which influence diurnal variation through secondary induction fields). If a simple diurnal correction is insufficient, a spatial function can be generated from a distributed network of base stations incorporating time shifts according to the longitude of the base stations and the survey measurements.

There are also periods of rapid, high-amplitude variations in the magnetic field termed “magnetic storms” which follow periods of high sun spot activity. It is not feasible to accurately correct for field variations during magnetic storms and surveys are generally suspended during these periods.

Network Adjustment (Leveling)

Most magnetic field measurements are made sequentially from platforms moving along a profile. In particular, airborne magnetic surveys are generally flown on parallel, equally spaced flight lines (see ► [“Magnetic Methods, Airborne”](#)). To reduce errors arising from imperfections in diurnal correction and to compensate for any residual platform heading effects, a subset of lines, termed “tie lines,” is flown perpendicular to the main flight line direction at a spacing typically five or ten times larger than for the flight lines. The purpose of the tie lines is to provide a set of intersections with the flight lines that can be used for network adjustment or “leveling” of the data. Network adjustment progressively changes base levels, gradients, and curvatures of individual lines and line segments to minimize the misfit of flight line and tie line measurements across the network of intersection points.

Leveling errors are primarily manifest as sharp field variations along the axes of the flight lines.

Microleveling

The standard network adjustment of data is an empirical, least-squares optimization which leaves residual errors. Microleveling or “de-correlation” is a grid-based filtering process designed to attenuate residual level shifts between adjacent lines. These features are recognized as being parallel to the flight line direction and having a spacing related to the flight line spacing. Any valid magnetic field variations with similar characteristics will also be attenuated by microleveling (although any features parallel to the flight line direction are always of uncertain validity and difficult to interpret). Microleveling commonly has the objective of attenuating noise which becomes evident in the sun-shading of images, or to smooth “herring-bone” patterns in contour lines. Microleveling is also used to precondition data for subsequent enhancements that would otherwise include strong expressions of these residual leveling errors. Microleveling improves the quality of images of the magnetic field, but it also unpredictably attenuates field variations, and the output should not be used for modeling, inversion, or depth estimation studies.

Elevation Correction

Absolute elevation values are much less significant in magnetic field data processing than in gravity data processing (see ► “Gravity, Data to Anomalies”). The IGRF definition (see ► “Geomagnetic Field, IGRF”) includes elevation terms, but these only become significant over ranges much more substantial than those encountered in most surveys. More significant to the information content of magnetic field data is the elevation separation between the sensor and the shallowest magnetic sources. For this reason, airborne surveys are commonly flown on drape surfaces at a nominal terrain clearance. Rugged terrain clearly increases the problem of drape flying, and rugged areas are best flown on a compromise, smooth drape surface to reduce abrupt elevation steps from line to line. Rugged terrain also poses problems for leveling of data because of difficulties in achieving identical elevations on flight lines and tie lines at each intersection point. The most effective way to correct for these elevation differences is to re-project data from the sensor elevation onto a suitable smooth drape surface by using a set of equivalent sources (Emilia 1973) to invert the data, and then forward computing the field from those sources onto the reference surface.

Terrain correction, used extensively in processing gravity data, is rarely applied to magnetic field data. It is unusual for terrain to have a consistent magnetization such that the magnetic field of the terrain can be derived by any simple integration. Also, the ground surface is only an effective magnetization contrast over the top of strongly magnetized

bodies, and in situations of deep weathering or transported cover, the top of any strong magnetizations may be substantially beneath the ground surface. In areas of rugged terrain, the influence of terrain magnetization can sometimes be recognized, but these variations are generally less than those due to variations in separation between the sensor and the terrain. Typical elevation clearance variations include lower clearance over peaks and crests than over the flanking slopes, and increased clearance over deeply incised valleys. It is generally this variation in terrain clearance which carries the danger of incorrect geological interpretation. For instance, an increase in terrain clearance causes attenuation of magnetic field expressions of shallow sources which might be misinterpreted as due to zones of alteration and partial magnetite destruction.

Each magnetic field data channel should be linked to an elevation channel. For the primary field measurements, this is generally the sensor elevation determined from GPS altitude measurements. The smooth drape surface derived or implicit in leveling the magnetic field data is the effective elevation of that data for any modeling or inversion. With some airborne magnetic data, no elevation channel is supplied, and in these cases, the best option is to derive a synthetic elevation channel by interpolating the ground elevation (e.g., from SRTM data) onto the lines, adding the nominal terrain clearance reported for the survey, and then low-pass filtering. Another situation sometimes faced by an interpreter is to have only radar altimeter data available. The radar altimeter can be added to the terrain channel to recover sensor elevation, but before doing this, the radar data should be carefully inspected, as it may contain short-wavelength noise which needs to be smoothed, and possibly spikes or abrupt steps which must be eliminated.

De-culturing of Magnetic Field Data

The expressions of man-made features in magnetic field data used for geological studies are termed “cultural” anomalies, and the process of removing these anomalies is termed “de-culturing” the data. Anomalies may be recognized as being man-made by their sharpness, indicating that they are distinctly shallower than other field variations in the area, by distinctive characteristics, or by correlation with mapped man-made features. Automated search algorithms can be used to scan data for possible cultural anomalies, but generally removing cultural anomalies is an interpretive, labor-intensive effort. Removal of cultural features is best done from the primary data. More general removal of sharp variations in the data is known as “despiking.”

Gridding of Magnetic Field Data

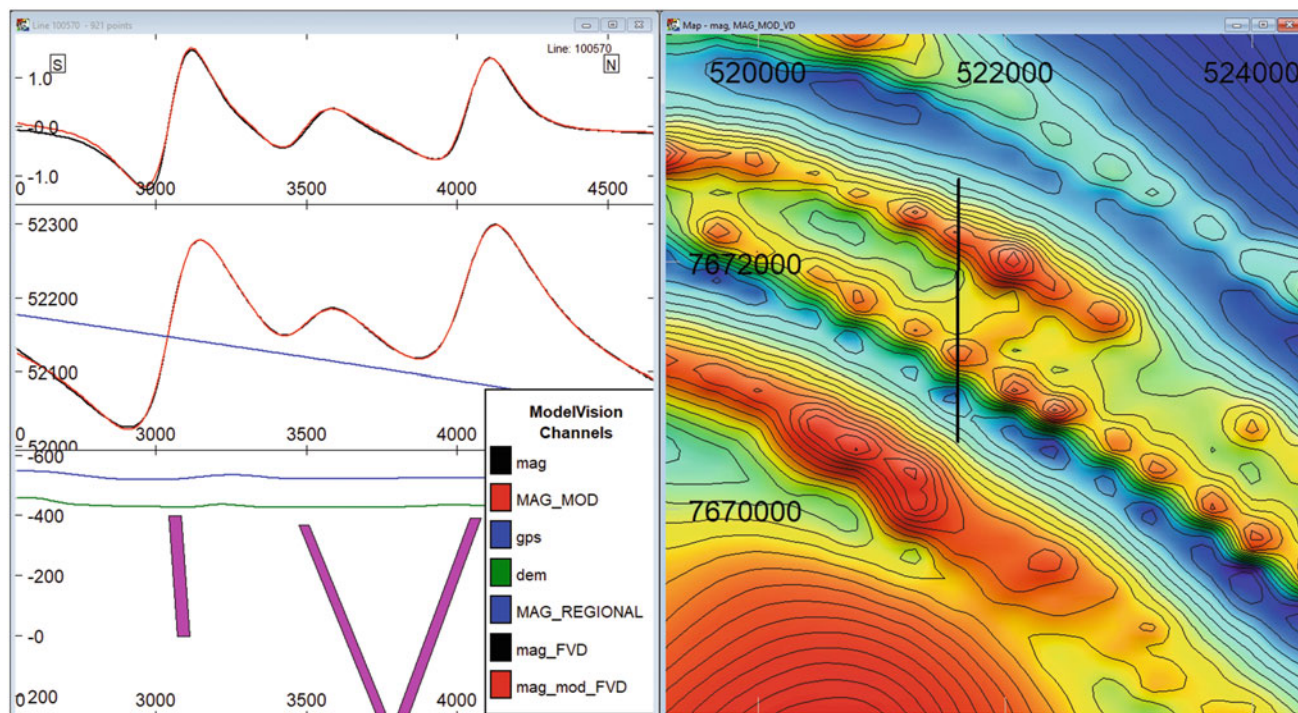
Operations through to leveling of the magnetic data are performed on the primary located data, and key channels are appended to the data so that changes can be tracked through

the processing sequence. Once data has been leveled, the next process is generally to grid the data. Gridding is not a correction process, but is crucial to all subsequent processing, enhancement, display, and interpretation. Gridding interpolates the data from the measurement locations to nodes of a regular mesh, creating a new and fundamentally different construct of the data. Magnetic field data is generally gridded using minimum curvature operators to cell sizes of 1/4 or 1/5 of the line spacing. Minimum curvature operators (Briggs 1974) are generally implemented as multipass functions with control parameters which can be modified to adjust to different data sets. Minimum curvature is an appropriate imposition on magnetic field data as the magnetic field is continuous with continuous derivatives. More sophisticated gridding algorithms, such as anisotropic diffusion, can also be justified as supplying superior interpolation of the field, but minimum curvature is the current industry standard and is adequate in most situations. Gridding breaks down most readily for sharp anomalies due to small, shallow sources. This can occur either where the magnetized unit is particularly shallow or where it trends oblique to the flight line direction. In these situations, the anomaly due to a continuous magnetic unit may break down to a “string of pearls” or “bicycle chain” pattern (Fig. 1) because of the insufficiency of the gridding algorithm to interpolate that sharp field curvature from one

flight line intersection to the next. These artifacts restrict the use of sun-shading to highlight subtle features in the data and must be removed prior to any attempt to highlight those details by subsequent enhancement processing. These gridding artifacts can be attenuated by re-gridding at a coarser cell-size, or by using a redesigned gridding operator to produce a smoother, lower curvature grid. Interpretation of the depth to the top of a magnetic source is based primarily on the sharpness of curvature of magnetic field variations, and limitations in preserving the sharpest gradients through the gridding process mean that depths estimated from inversion or modeling of grid data may be overestimated, particularly for shallow sources.

IGRF Correction

Across survey areas of length several tens of kilometers or more, variations in the earth’s primary field may contribute significantly to the measured field variations. In order to emphasize more local field variations due to crustal sources, the International Geomagnetic Reference Field (IGRF), a global model of the earth’s magnetic field, is subtracted from the measured data. The IGRF is not a separation of field components from specific sources, but the long wavelength IGRF terms represent mostly core components and broad, lower crustal magnetizations. The IGRF is defined



Magnetic Data Enhancements and Depth Estimation, Fig. 1 Example aeromagnetic profile over a synclinal fold in the Frew River area of the Northern Territory of Australia (data courtesy of the Northern Territory Geological Survey). The TMI image on the right shows a “string of pearls” pattern. The top track of the profile data on the left

shows a vertical derivative enhancement of the measured TMI (black) and model computed TMI (red). The model was generated by inversion to match this filter enhancement. Profile modeling and inversion avoids the problem of grid artifacts

and published by the International Union of Geodesy and Geophysics (IUGG) (See entry ► [“Geomagnetic Field, IGRF”](#)).

Reduction to Pole (RTP) Transform

The external field of a magnetic body varies substantially as a function of the local geomagnetic inclination (see ► [“Magnetic Anomalies: Interpretation”](#)). The simplest relationship between the magnetic field and source magnetizations occurs at the geomagnetic poles where the field orientation is vertical (see ► [“Geomagnetic Field, Global Pattern”](#)). At the geomagnetic poles, induced anomalies due to steeply dipping bodies are essentially positive and are centered above the magnetization. The reduction to pole transform is a fast Fourier transform (FFT) filter applied to replicate the field variations expected in a vertical field. The RTP transform is unstable when applied to low geomagnetic inclination fields (this is generally perceived to be a problem at inclinations of less than 30°). The severity of the problems depends on data quality and sufficiency of sampling, and these problems arise in particular with any apparent North–South trending features in the data. Several methods can be used to overcome low inclination RTP problems. In one approach, the RTP filter can be distorted to reduce the instability. The output of the transform is not strictly valid, but may have the desired characteristics of a steep inclination field. Another approach is to use an equivalent source technique. The reduction to equator (RTE) transform also provides a partial alternative to RTP at low geomagnetic inclinations. The RTE transform shifts anomalies North–South to be directly above the source magnetizations, but the East–West elongation which is a natural characteristic of anomalies at low geomagnetic inclinations is slightly amplified. Note that RTP or RTE transforms should not be used in modeling or inversion studies, as forward modeling algorithms readily incorporate the local geomagnetic field direction (although RTP images may help to approximately locate bodies in starting models).

The commonly used RTP transform requires that the direction of magnetization be in the local direction of the geomagnetic field, and for bodies with substantial remanent magnetization (see ► [“Remanent Magnetism”](#)) in a different direction this transform is invalidated (see recovery of source magnetization direction from magnetic field data). If however the resultant magnetization direction (the vector addition of induced and remanent magnetizations) is known from palaeomagnetic studies, magnetic moment analysis, or inversions of the magnetic field data, then an alternative RTP operator can be used which allows for separate specification of the geomagnetic field and the magnetization direction (Fig. 2). All FFT filters have potential pitfalls, in part because of the fundamental assumption that the fields repeat infinitely beyond the sampled area. The FFT transform requires some processing to address level shifts between opposite edges of

the sample area and to pad the sample area with a smooth extrapolation. However well this is done, the output around the margins of the data is less reliable than in the center. This problem is of particular concern with the RTP transform, especially at the northern and southern edges of the data, and for longer wavelength field variations that are migrated furthest by the RTP transform.

The Pseudo-Gravity Transform

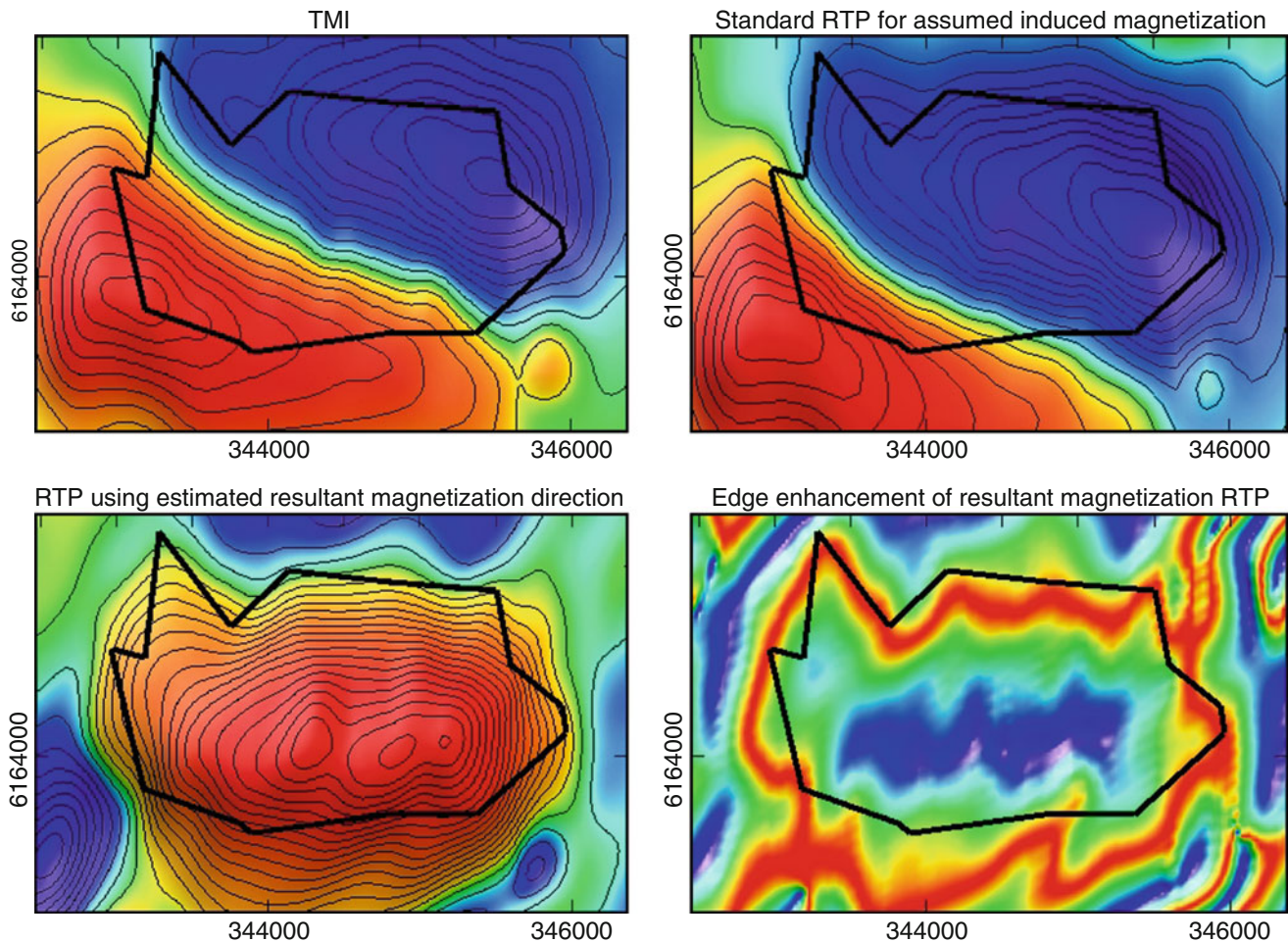
Gravity and magnetic fields are linked by the common fundamental theory of potential fields (see ► [“Magnetic Methods, Principles”](#)), as described by Poisson’s relationship (Baranov 1957). If the direction of magnetization is consistent, and if there is a constant ratio between density and magnetization, then the RTP magnetic field and the vertical gradient of the gravity field are identical. For any but the simplest geological system, there is no such reliable linkage between magnetization and density. Nevertheless, Poisson’s relationship provides a basis for converting a magnetic field to an equivalent gravity field expression. This transform includes an inverse vertical derivative operation which amplifies the expression of longer wavelength components of the field. The pseudo-gravity transform is sometimes used to apply analysis and processing techniques developed for gravity data to the magnetic field data but this must be done with care. For instance, the pseudo-gravity transform accentuates the longer wavelengths of the field but if the data from a small survey area is transformed those longer wavelengths are insufficiently sampled.

Enhancements of Magnetic Data

Enhancement of magnetic field data applies to any operation designed to highlight field variations due to a selected source of magnetization. These operations include application of various filters and also selection of display options such as histogram normalization and sun-shading.

Image Display of Magnetic Field Data

The display of magnetic field data is not an operation in the sense that it changes values of the data, but nevertheless it is highly significant in determining how the data is perceived and how easily geological information can be recovered from it. Magnetic field data generally has much sharper variations than gravity data across the same area, with much of the amplitude range in only a small subset of the data. For this reason, magnetic field data is often displayed using a histogram equalized color stretch devoting equal areas of the image to each color band. This form of imaging provides a much improved representation of detail across the complete range of the data. Histogram equalization does however have the drawback that it is difficult to appreciate the true



Magnetic Data Enhancements and Depth Estimation, Fig. 2 The top right image is of a TMI anomaly over the Black Hill Norite in South Australia which has strong remanent magnetization (data courtesy of the Department of primary Industries and Resources South Australia). The contour interval is 200 nT, and the outline is of the source body derived from a staged inversion of the TMI data using a resultant magnetization direction determined from magnetic moment analysis. The top right

image is of a standard RTP transform using an incorrect assumption that magnetization is in the geomagnetic field direction. The bottom left image is an RTP which correctly incorporates the resultant magnetization direction to produce a compact, positive anomaly centered over the source body. The bottom right image is an edge enhancement based on the tilt filter, applied to the resultant magnetization RTP. The maximum in this filter approximately matches the edge of the computed model

amplitude range of features or to compare the relative amplitudes of different features.

Another important enhancement tool in image display is the use of sun shading. This process involves the addition of shadows and highlights to an image to mimic the illumination of the data surface by a light source. These highlights and shadows reveal features of shorter wavelength and more subtle amplitude variation than can be easily recognized in the color mapping of the amplitude values. With a vertical light source, the highlights appear along the axes of minima and maxima. However, light sources are generally set to moderate or low inclinations to produce a stronger, directional enhancement of features trending perpendicular to the illumination direction. Illumination applied parallel to the flight line direction minimizes the expression of any residual leveling errors.

Vertical Continuation of the Field

The amplitude of a magnetic field above a source varies with elevation as an exponential function of wavelength. This relationship can be readily exploited with FFT filters to recompute the field at a higher elevation (“upward continuation”) or lower elevation (“downward continuation”). Upward continuation attenuates the field through an increased separation from the magnetization. Conversely, downward continuation towards the source or sources amplifies field variations. Upward continuation is inherently stable and acts as a smoothing filter, whereas the amplification of field variations due to downward continuation becomes unstable at depths approaching the source depth or shallower if there is any shorter wavelength contamination of the data. Methods including equivalent source computations can be used for continuations in which the input and/or output surfaces are

irregular. Upward continuation is useful for attenuation of fields from shallow magnetizations to improve focus on modeling of deeper sources as the effect of the filter can be directly compensated by changing the elevation at which the model field is computed.

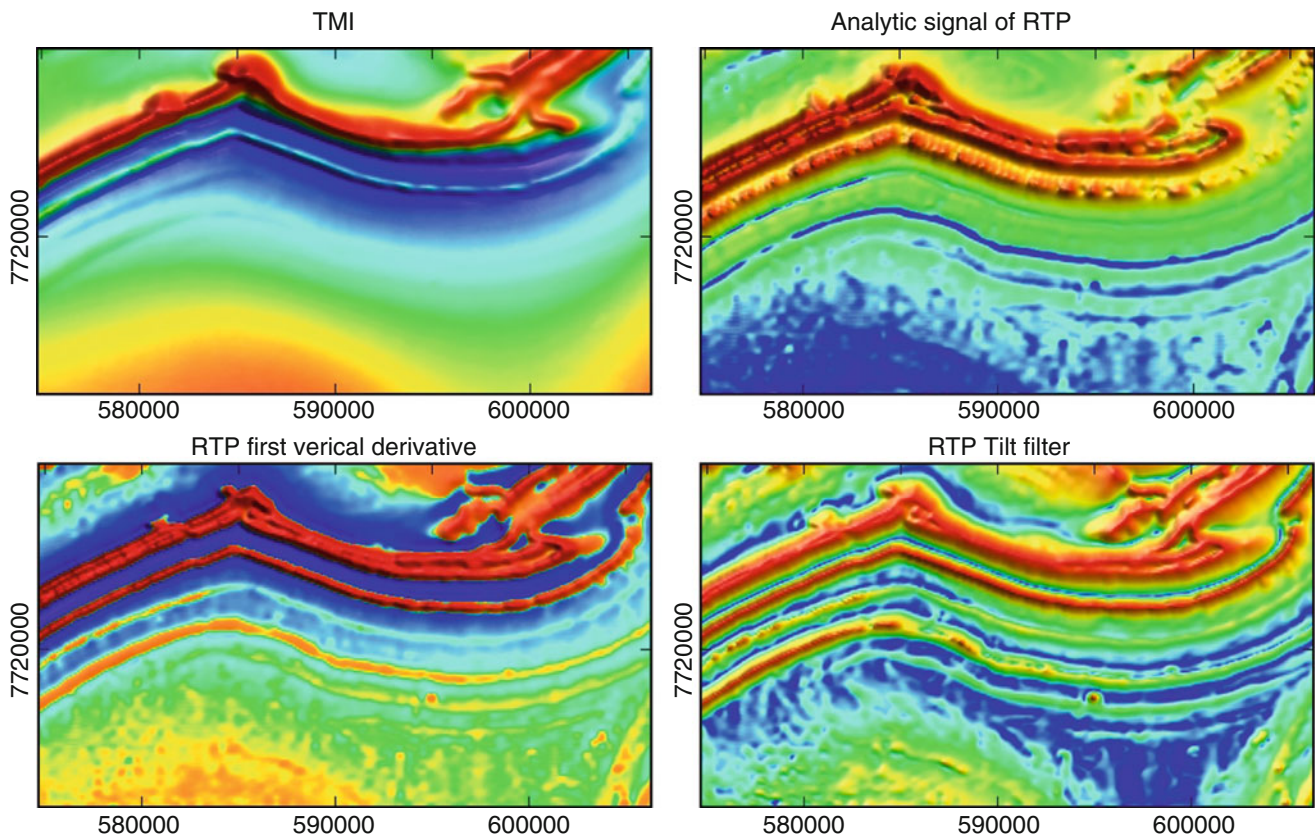
Vertical Gradients of the Field

Vertical gradient (or alternatively named “vertical derivative”) filters preferentially amplify short-wavelength components of the field at the expense of longer wavelengths (Fig. 3). Vertical gradient filters are generally applied to gridded data using FFT filters and are therefore influenced by the characteristics of the gridding operator. Vertical gradient filters can also be applied to profile data on the assumption that the horizontal gradient measured along the profile is the only horizontal gradient. In some cases, this is a poor approximation but for highly elongate anomalies and with profiles approximately perpendicular to those anomalies the approximation may be acceptable. If, furthermore, the magnetic sources are shallow compared to the line spacing of the survey then there may be considerably more short-wavelength detail

in the profile data than is preserved through the gridding process. Vertical gradient profile filtering can be used to retain and enhance the shortest measured wavelengths during modeling and inversion by matching the outputs of the filter applied identically to the measured and computed fields (Fig. 1). The second vertical derivative is a more extreme enhancement of a magnetic field. As such, it provides a more distinctive amplification of fields from shallow sources, but simultaneously it is more sensitive to any short-wavelength artifacts in the data. In the event that a derivative of the field unacceptably accentuates noise or shallow magnetizations, the data can be preconditioned with a mild upward continuation before applying the derivative filter.

Bandpass and Matched Filters

The general relationship between depth of a source and the wavelength of its magnetic field can also be exploited to enhance the expression of sources from particular depths through bandpass filtering. Short-wavelength field variations can only be generated by shallow sources, so high pass filtering attenuates the expression of deep sources and



Magnetic Data Enhancements and Depth Estimation, Fig. 3 (All images are histogram equalized). The top left image shows TMI over the Frew River area of the Northern Territory of Australia (data courtesy of the Northern Territory Geological Survey). The bottom left image is of the vertical gradient of TMI. This enhancement reveals weak anomalies across the center of the image which are poorly imaged in the TMI data.

The top right image is the modulus of the analytic signal, which also enhances the expression of shallow magnetizations. The tilt filter image in the lower right is the most extreme enhancement and provides the best expression of weak, near-surface magnetizations across the lower gradient southern half of the area

conversely low pass filtering tends to attenuate the expression of shallow sources. However, bandpass filters rarely, if ever, provide a comprehensive separation of magnetic fields from different sources so their outputs are not suitable for modeling or inversion studies (unless the identical filter is applied to the forward computed model fields).

The power spectrum display of a magnetic field commonly consists of segments which approximate to straight lines, the slope of which is a function of the depth to the assemblage of sources whose field contributions dominate that part of the spectrum (Spector and Grant 1970). The inspection of a power spectrum can support selection of cutoff wavelengths for bandpass filtering, with the taper of the filter also very important. Filters should not truncate abruptly nor should they target too narrow a band of wavelengths. Size, shape, and depth extent of bodies also strongly influence their magnetic field power spectrum, and the analysis by Spector and Grant is valid only statistically and for an assemblage of bodies. With some bandpass filtering, the transform back into the space domain has been claimed to localize fields from sources at particular depths, with terms such as “depth slice” and “pseudo-depth slice.” These filters may produce images which are useful for interpretation but any claims that they effectively invert the magnetic field data to supply a 3D mapping of magnetization are untenable. Nor does fitting a straight line to a segment of a power spectrum prove the effective wave-number separation of field contributions from sources at any particular depth, as implied by classifying these filters as matched filters. The combination of spectral characteristics and spatial localization supplied by wavelet processing (see ► “Wavelet Analysis”) promises considerable advantages over FFT methods, and some wavelet processing methods are now being used, particularly for localizing edges of magnetization in multiscale edge detection.

Regional-Residual Separation

Potential fields are additive, and the magnetic field near the surface of the earth generally contains contributions from multiple crustal magnetizations. Separation of field variations from different sources is interpretive and nonunique, with no general analytic solution. One class of field separations is to remove local features of sharp curvature in the field (“anomalies” or “residuals”) from broader, lower curvature variations due to deeper or more distant sources (the “regional”), a process termed “regional-residual” separation. Regional-residual separation is most reliably performed in cases for which the fields to be separated have quite different characteristics; if the characteristics of the fields are similar, their separation becomes more interpretive and less certain. Regional fields are also due to geological structures, but which are too deep or distant to be of immediate interest in a specific project. For modeling

studies, the regional separation should be adjustable during the modeling, because a better realization of this separation emerges as the model is developed. The objectives and methods of regional-residual separation for magnetic fields are identical to those for gravity fields (see Gravity Data, Regional-Residual Separation) but the separations tend to be more easily made for magnetic data because curvature is higher for magnetic fields than for gravity fields and because small, shallow magnetizations create sharp magnetic anomalies distinctively different to the regional field while any gravity anomalies of those bodies may be too weak to be easily detected.

The Total Gradient (Analytic Signal)

The total gradient or analytic signal (more correctly the modulus of the analytic signal) (Roest et al. 1992) is a popular gradient enhancement used to highlight the distribution of shallow magnetization.

For compact or thin bodies, the modulus of the analytic signal peaks approximately over the center of the body and for wider bodies over their edges. It is only weakly sensitive to magnetization direction and indeed the two-dimensional case with only one horizontal gradient is independent of magnetization direction. The modulus of the analytic signal of TMI has much lower sensitivity to the inclination of the geomagnetic field than the original TMI data and provides a means to analyze low latitude magnetic fields without the concerns of the RTP operator. Source parameters such as depth can be estimated from statistics of the filter output but these advanced applications are more appropriately classified as interpretation rather than enhancement processing.

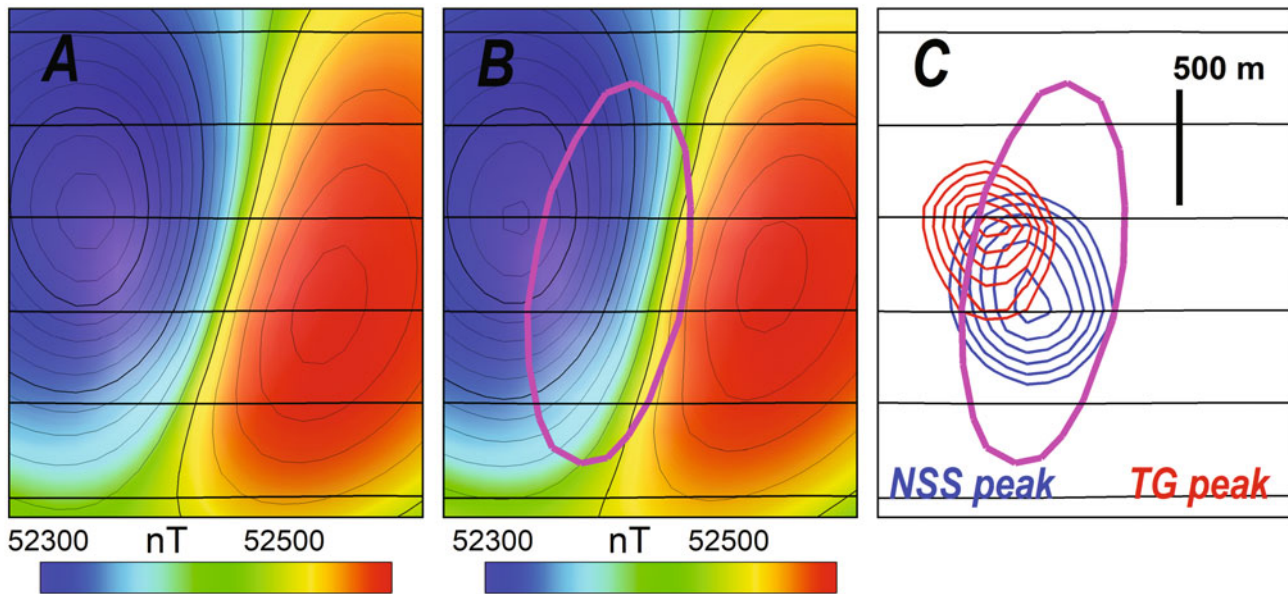
The Tilt Filter

The tilt filter (Miller and Singh 1994) has also become a popular enhancement.

The filter is computed from ratios of gradients and thereby retains no amplitude information. The filter output is defined at all places, but in areas of low gradient there is no signal to enhance so it is important that tilt images be interpreted with reference to the original data (Fig. 3). The tilt filter and analytic signal are both expressions of gradient ratios, and as such, there are clearly linkages between them in their tendency to amplify field variations from shallow sources, highlight the location of subvertical edges, and to indicate depth to magnetization contrasts according to their sharpness of variation.

The Normalized Source Strength (NSS)

The normalized source strength (Beiki et al. 2012) is derived from the eigenvalues of the magnetic gradient tensor. It highlights the distribution of shallow magnetization in much the same way as does the total gradient and in many situations the



Magnetic Data Enhancements and Depth Estimation, Fig. 4 (a) measured TMI anomaly at Ethabuka (data courtesy of the Geological Survey of Queensland), (b) TMI computed from elliptic pipe inversion model, (c) the NSS (blue contours) peaks over the center of the model but

the total gradient (red contours) peaks towards the edge of the model and is a less reliable estimation of the location of magnetization. (See recovery of source magnetization direction from magnetic field data)

two outputs provide similar results. The computational overheads of computing the NSS are at least an order of magnitude greater than required to compute the total gradient, but in some cases (according to the shape of distribution of the magnetization and the magnetization direction) the improved performance of the NSS may justify the additional effort it requires (Fig. 4).

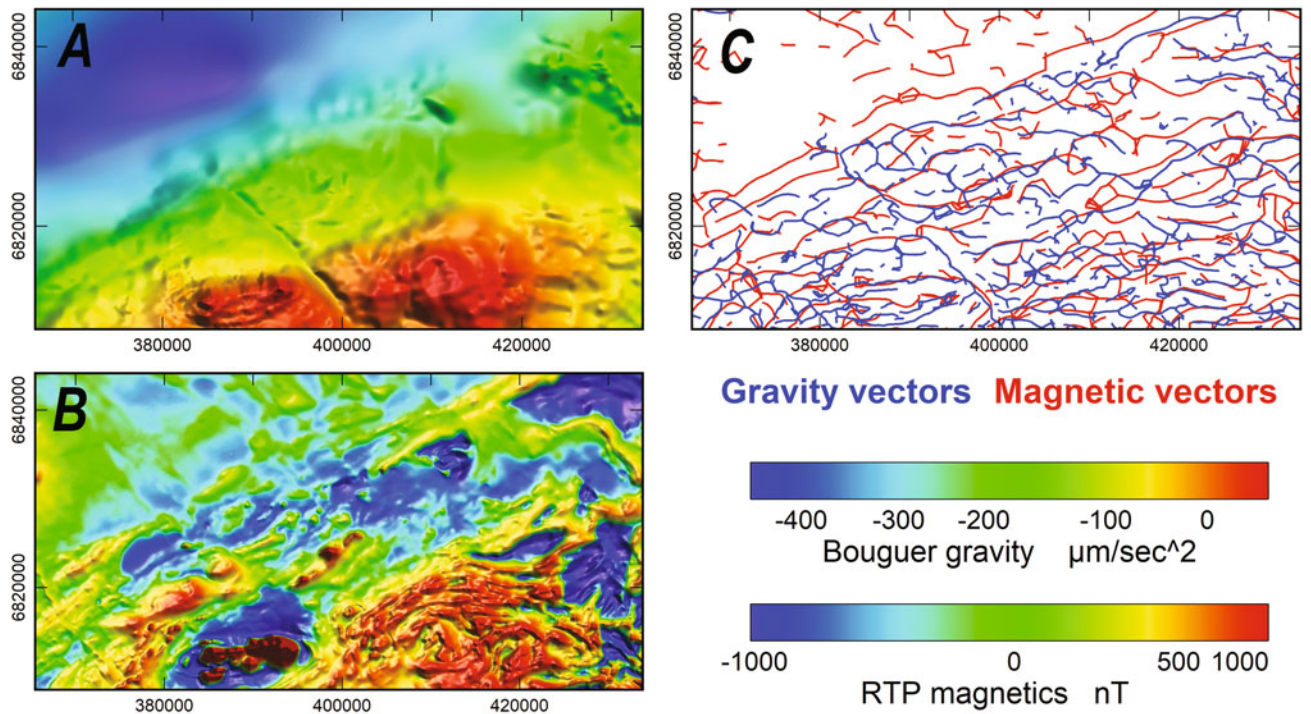
Multiscale Edge Detection (“Worms”)

Ridge vectors of local maxima in the horizontal gradients of a magnetic field (worms) approximately map the horizontal location of steeply plunging magnetization contrasts. Applied following successive upward continuations of the field, these edge vectors are progressively more sensitive to magnetization contrasts at greater depths and sets of the vectors computed across a range of continuation heights provide an automated method of magnetic field analysis (Hornby et al. 1999). Multiscale edges defined by the wavelets have some affinities to different wavelength components separated in Fourier analysis, with the detail from shallow sources mostly carried in the small-scale edges, and broader, large-scale edges providing more information about deeper sources. Multiscale edge vectors are intrinsically linked to the analytic signal and tilt filter and to other curvature functions used to map source edges (Blakely and Simpson 1986; Grauch and Cordell 1987). Recovery of source information from potential fields as line vectors allows those sources and their spacing and orientation patterns to be investigated using vector analysis tools. Line vectors also provide advantage in

investigating similarities and differences between potential fields such as gravity and magnetics which are less easily summarized in comparison between images of those fields (Fig. 5).

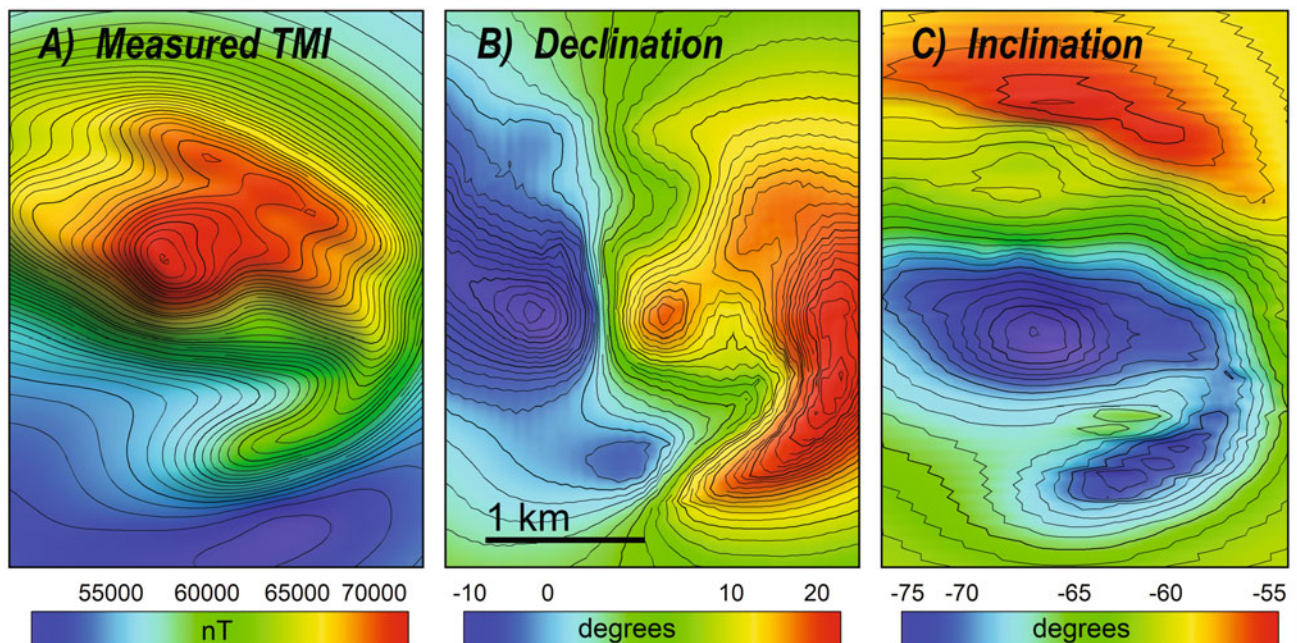
Mapping Extreme Field Variations

Most magnetic field variations investigated in geophysical studies have amplitudes of less than a thousand nT and are measured in background fields of between 25,000 to 70,000 nT. Two consequences are that any misorientation of vector measurements is heavily penalized by accompanying error in the background field contributions and that the total field direction is consistent across anomalies permitting it to be treated as a potential field. In off-earth, low-field environments orientation error of vector measurements is greatly reduced and the total field direction varies across anomalies, disqualifying it for use with the enhancement and analysis tools described above. Similar considerations apply to extreme amplitude anomalies on earth. An iterative correction scheme can be used to convert measured total field data to vector-consistent magnetic field data for use with FFT enhancement and analysis tools (Clark 2013). The analysis produces grids of vector components of the field from which field declination and inclination grids can be derived (Fig. 6). Forward modeling algorithms used to compute magnetic fields provide the measured field directly and the field declination and inclination can be independently estimated from vector components computed from those models (Fig. 7).



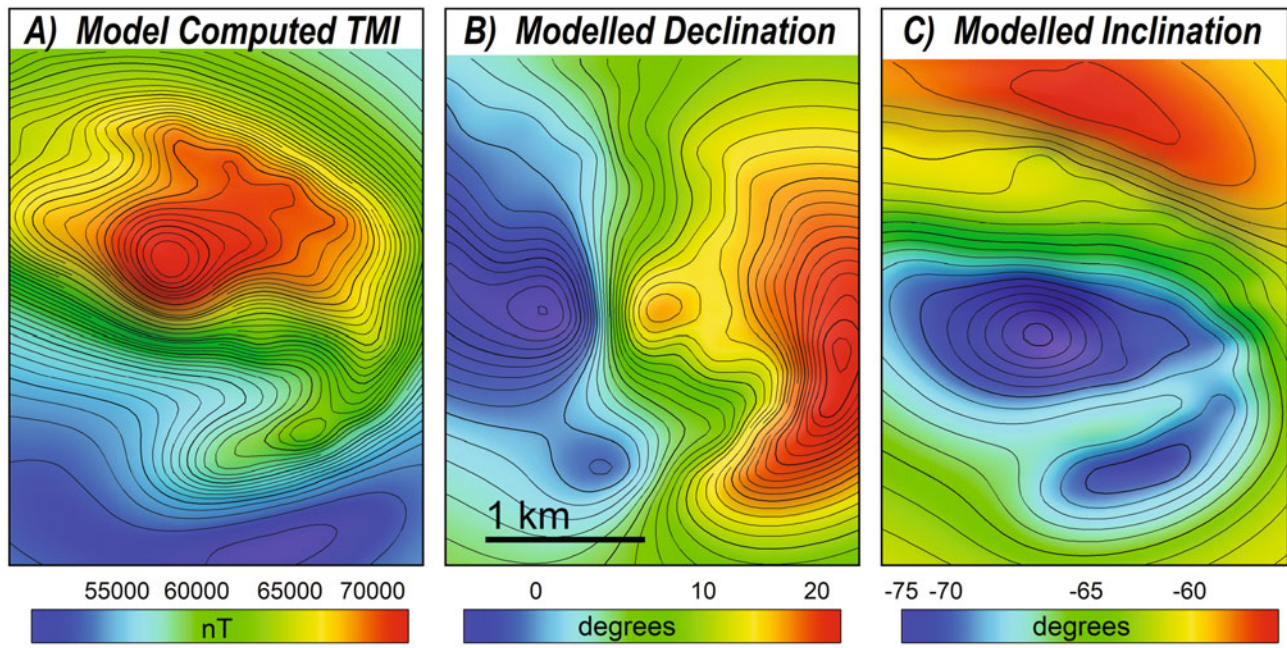
Magnetic Data Enhancements and Depth Estimation, Fig. 5 (a) Bouguer gravity and (b) TMI over the Murloocoppie Area (data courtesy of the Geological Survey of South Australia), (c) gravity (blue) and

magnetic (red) worms highlight relationships between the fields more clearly than through comparison of the grid images



Magnetic Data Enhancements and Depth Estimation, Fig. 6 (a) Measured TMI at Paragon Bore (data courtesy of the Geological Survey of South Australia), (b) field declination, and (c) inclination derived from

east and north components generated in iterative transform of measured TMI to vector consistent TMI (coded by Blair McKenzie) for the 20,000 nT anomaly in a 56,000 nT field



Magnetic Data Enhancements and Depth Estimation, Fig. 7 (a) Computed TMI at Paragon Bore from an inversion of the measured TMI in Fig. 6, (b) field declination, and (c) inclination forward computed from

the inversion model and which match the equivalent outputs shown in Fig. 6 (b) and (c) derived independently from the measured TMI to vector-consistent TMI transform

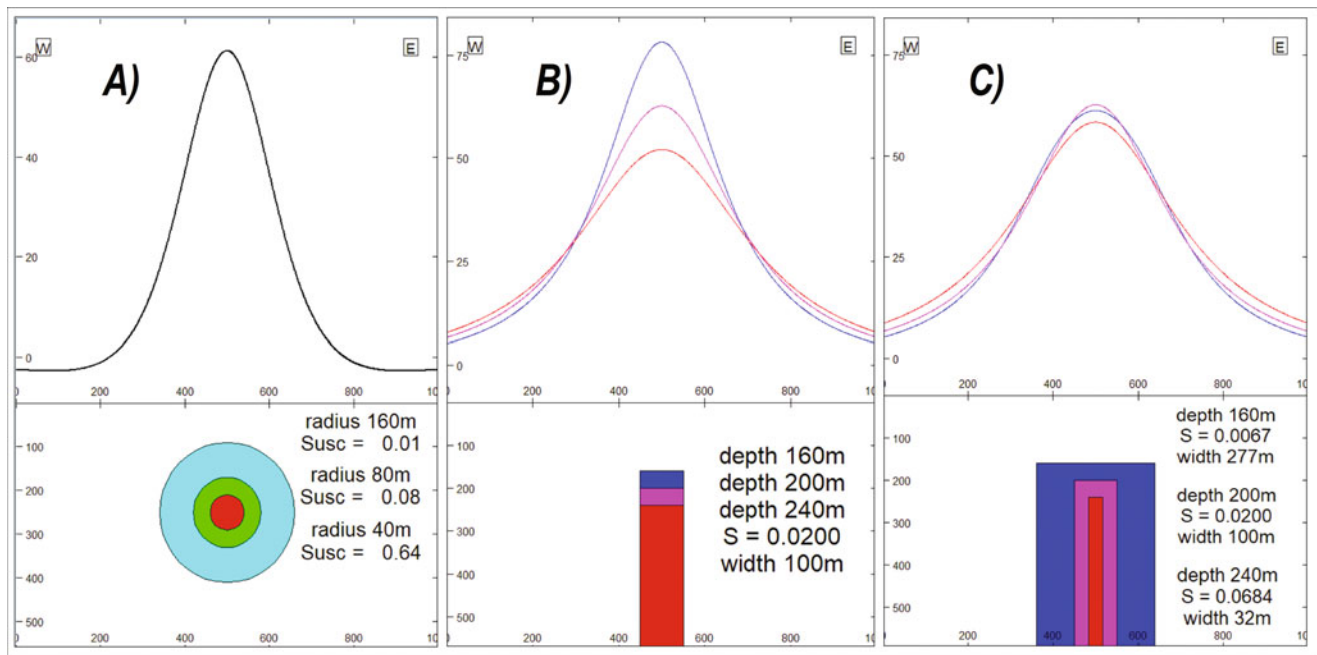
Analytic Derivation of Magnetic Source Parameters

Nonuniqueness in interpreting a distribution of magnetization from study of a measured magnetic field precludes any definitive analytic solution. All source depth estimation is dependent on an assumed source geometry model or equivalent statistic such as the structural index. Voxel models without shape constraint are unsuited to source depth estimation and generally require an imposed bias such as a depth weighting function to select a preferred model. Parametric models are better suited to recovery of source statistics but are dependent on the model characteristics providing a valid representation of the magnetization. Parametric inversion provides a powerful analysis of the sensitivity with which various model characteristics can be recovered.

Compact bodies have intrinsically low sensitivity in estimation of depth to top. Figure 8 shows three co-centered magnetic spheres which produce an identical magnetic anomaly. Anomaly amplitude is a function of the magnetic moment (the product of volume and magnetization) but not of the individual magnetization and volume values. Although the three spheres produce an identical anomaly, there are significant differences in depth to their tops. This is an extreme example of the ability of one parameter to compensate for variation of another. Similar but inexact relationships are also shown in Fig. 8 for narrow vertical bodies of large depth and

strike extent (“dykes”). The magnetic anomalies of these bodies depend on only three parameters: width, depth to top, and magnetic susceptibility. There is considerable sensitivity of the magnetic anomaly to variation in depth to top where that is the only parameter changed (Fig. 8b), but if thickness and magnetic susceptibility are also allowed to change the sensitivity to depth is severely reduced (Fig. 8c).

Automatic mapping of source magnetization distributions has been one of the most prolific subjects of geophysical publication over many years. Those methods that have achieved widespread acceptance and use can mostly be classified into a small number of groupings, namely, Wener deconvolution, Euler deconvolution, the Naudy method, and the source parameter imaging (“SPI”) method. The fundamental assumptions of the various methods are necessarily similar, and in many cases identical, but nevertheless there are some practical differences which might lead to advantages in applying certain methods in certain circumstances. In all cases, the key factor for successful application of automated source detection and parameter estimation is the suitability of the geology. If the geology provides discrete, homogeneous, well-separated, moderate to strong magnetizations in simple and predictable geometries then any of the automated source estimation methods should work well. If however the geology results in irregular and indistinct distributions of magnetization, then there is no basis for the successful application of any of the methods and they can only lead to fictitious and



Magnetic Data Enhancements and Depth Estimation, Fig. 8 (a) Co-centered spherical magnetizations of identical moment but different radius and depth to top generate identical anomalies, (b) TMI anomalies over identical tabular bodies of large depth and strike extent with -20%

and $+20\%$ depth offset, (c) TMI anomalies of the depth-offset bodies with susceptibility and thickness adjusted to minimize least-squares difference between the anomalies

misleading results. This gives rise to a common dilemma in source depth estimation; whether to set strict quality conditions and accept just a few solutions of apparently high validity, or whether to relax the quality conditions to generate a more comprehensive set of solutions of less certain validity. In practice, validity of solutions is commonly evaluated from their quality coefficients, by clustering of solutions and their internal consistency, or simply by visual inspection and judgment as to whether the magnetic field anomaly the solutions are derived from is reliable and suitable. The ability to rapidly generate source parameter estimates from automatic or semi-automatic methods is in part offset by the time and discretion required to evaluate and interpret the reliability and usefulness of those results. All automated source depth estimators also have in common that they are dependent on the quality and sufficiency of the magnetic field data. A concise review of magnetic depth estimation methods and practical advice on their application is given by Li (2003).

The Werner Method

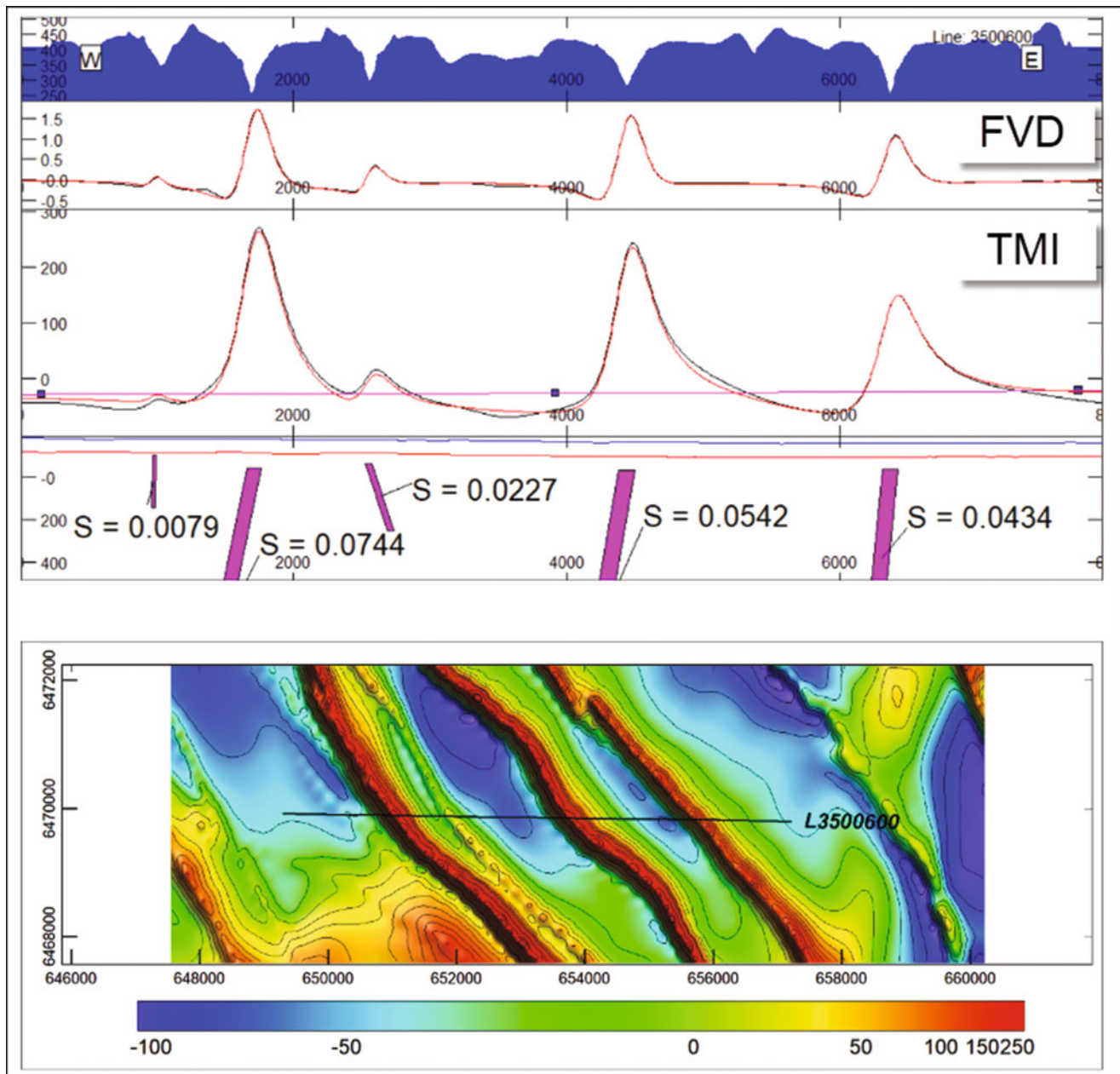
The Werner method of automatically generating magnetic source solutions from profile magnetic field measurements (Werner 1953) solves for simple geological structures by solution of linear equations applied to profile data sampled within a moving window. The equations are specified for bodies of nominated structural index and can be solved either for single profile data or can be extended to a three-dimensional multiple source analysis.

The Naudy Method

The Naudy method (Naudy 1971) is an elegant approach to finding and solving for magnetic sources from analysis of profiles of magnetic field data. As a window is moved along a magnetic field profile, the data is split into symmetric and asymmetric components about the center of the window. Firstly, the reference anomaly is computed for a target body of the nominated type and specified properties and a correlation coefficient is derived between that target curve and the processed field data curves. A coefficient value beyond a set threshold flags a possible solution, which is then fed to a second stage inversion process which searches for the optimum source parameters. A key advantage of the Naudy method is that it supplies bodies which can then be delivered into a modeling package for more detailed testing against the data they were derived from, and as starting models for more interpretive studies. The Naudy method has been extended by Shi (1991) to run on the vertical derivative (Fig. 9) with advantages that the derivative is more sensitive to the top of the body and does not require definition of the background field.

Euler Deconvolution

Euler deconvolution solves for Euler's homogeneity equation within a moving window assuming that the field variation within the window is due only to a single source body with a defined structural index. Reid et al. (1990) extended the analysis to a two-dimensional window of grid data. Sources of error are that the structural index value used may be

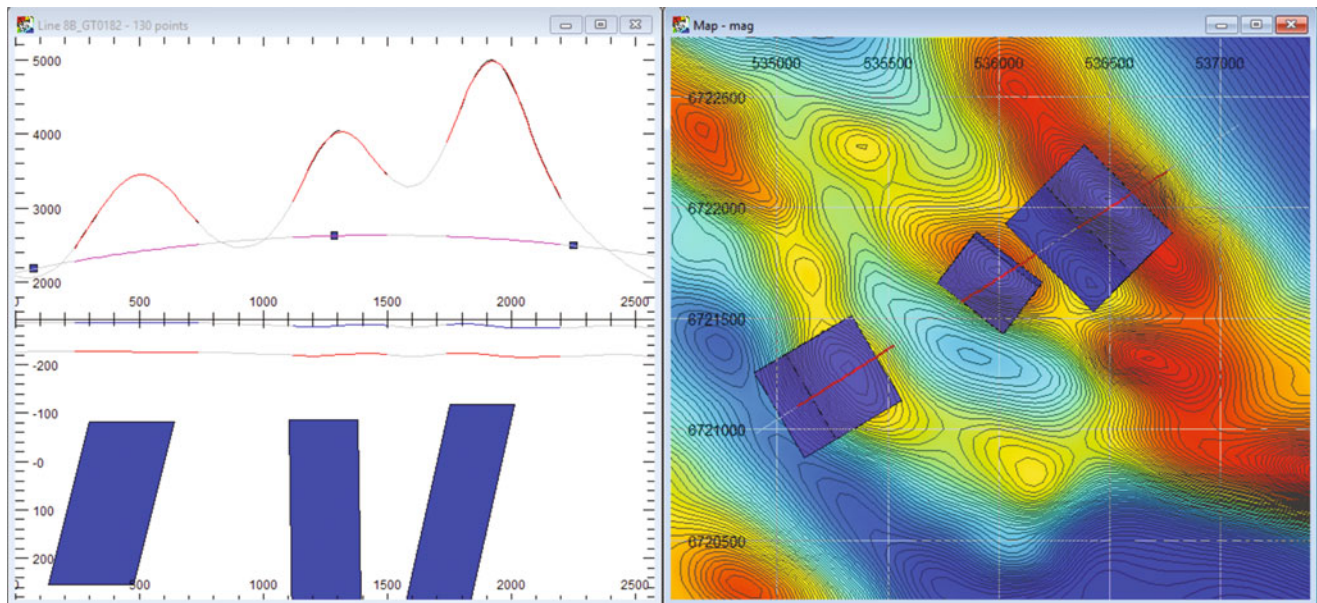


Magnetic Data Enhancements and Depth Estimation, Fig. 9 Naudy AutoMag analysis tuned to the vertical derivative over Gairdner Dykes (lower track). Data is from the Gawler Craton Airborne Survey (courtesy of the Geological Survey of South Australia) with a

profile of computed (red) and measured (black) TMI and vertical derivative channels and (top track) the similarity coefficient for which minima below a set threshold trigger source solutions as seeds for the bodies (shown in magenta) that are subsequently fine-tuned with inversion

inappropriate and that the field variation within the window may be due to more than one source. As a consequence, the solutions themselves require interpretation. There have been numerous adaptations of the method to try to estimate the appropriate structural index for each solution and to analyze clusters of solutions. Euler deconvolution has also been extended to application with the modulus of the analytic signal and to solution for multiple sources. Other implementations of the Euler homogeneity equation to derivatives of the magnetic

field which supply estimates of structural index as an integral part of the analysis are termed “extended” Euler methods. Nabighian and Hansen (2001) have shown that the extended Euler method is a unification of Euler deconvolution and the Werner method. Thurston and Smith (1997) applied Euler’s homogeneity equation to expressions of the complex analytic signal in a method termed source parameter imaging (SPI) which recovers source estimates for dipping 2-D contacts and dipping 2-D thin sheets from gridded magnetic field data. This



Magnetic Data Enhancements and Depth Estimation, Fig. 10 Example sweet-spot depth inversion from the Gawler Craton Airborne Survey (courtesy of the Geological Survey of South Australia). The grid traverse (left) was selected in map view to sample 3 TMI

anomalies (right). Three flat-topped source bodies located beneath the anomalies are strike and azimuth adjusted to match the anomaly and the profile data segments are inverted to generate three source depth estimates

method requires recovery of higher derivatives of the magnetic field, and ultimately it is the validity of those higher derivatives in the presence of noise and processing artifacts that determines the usefulness of the method.

Targeted (“Sweet-Spot”) Inversion

Automatic source analysis does not recover the full value of magnetic survey data and is a poor justification for substantial decisions such as the location and design of boreholes. The distribution of geological information in magnetic field data is highly irregular and can only be recovered above locations of separated moderate or strong magnetization contrasts. These locations can be selected manually by an interpreter or can be generated automatically as profiles through anomaly peaks perpendicular to the local trend of the field. Computation-intensive inversions of the resulting small data sets find optimum models of selected type. For estimation of depth to the top of magnetization, the most appropriate models have horizontal top faces (Fig. 10) because unconstrained curvature of the top face substantially increases the depth range across which solutions can be found to fit the data. The sweet-spot method supplies high quality magnetic source depth estimates but is still bound by fundamental nonuniqueness constraints.

Summary

A substantial toolbox of methods exists to process and enhance magnetic field data, and these capabilities are

continuously being extended as computing power and speed increase. Most of the methods require careful inspection of the data to select optimum parameters and to detect any undesirable characteristic in the outputs. Enhancement operations are best selected and designed for a specific interpretational objective, and different enhancements can be applied to the same data for different objectives. Operations such as gridding and visualization are an integral part of processing and enhancement and also require careful attention to avoid misrepresentation of data. There is an arbitrary delineation between methods such as multiscale edge detection described here as an advanced enhancement and automatic depth estimation such as Euler deconvolution which attempts a marginally greater level of discrimination and represents transition into interpretation. The current trend of automatic source depth estimation is towards ever greater sophistication using high-order derivatives of the field, but these methods place ever greater stress on the validity of the data such that the derived parameters require careful evaluation. For situations requiring greater confidence in the results, automated depth analyses can be replaced with interpreter-controlled processes such as sweet-spot inversion.

Cross-References

- [Geomagnetic Field, Global Pattern](#)
- [Geomagnetic Field, IGRF](#)
- [Geomagnetic Field, Secular Variation](#)

- Gravity Data, Regional: Residual Separation
- Gravity, Data to Anomalies
- Magnetic Anomalies: Interpretation
- Magnetic Gradiometry
- Magnetic Methods, Airborne
- Magnetic Methods, Principles
- Magnetic Methods, Surface
- Magnetometers
- Recovery of Source Magnetization Direction from Magnetic Field Data
- Remanent Magnetism
- Wavelet Analysis

Bibliography

- Baranov V (1957) A new method for interpretation of aeromagnetic maps: pseudo-gravimetric anomalies. *Geophysics* 22:359–383
- Beiki M, Clark DA, Austin JR, Foss CA (2012) Estimating source location utilizing normalized magnetic source strength calculated from magnetic gradient tensor data. *Geophysics* 77:J23–J37
- Blakely RJ (1996) *Potential theory in gravity and magnetic applications*. Cambridge University Press, Cambridge
- Blakely RJ, Simpson RW (1986) Locating edges of source bodies from magnetic or gravity anomalies. *Geophysics* 51:1494–1498
- Briggs IC (1974) Machine contouring using minimum curvature. *Geophysics* 39:39–48
- Clark DA (2013) New methods for interpretation of magnetic vector and tensor gradient data 2: application to the Mount Leyshon Anomaly, Queensland, Australia. *Explor Geophys* 44:114–127
- Emilia DA (1973) Equivalent sources used as an analytic base for processing total magnetic field profiles. *Geophysics* 38:339–348
- Grauch VJS, Cordell L (1987) Limitations of determining density or magnetic boundaries from the horizontal gradient of gravity or pseudogravity data. *Geophysics* 52:118–121
- Hornby P, Boschetti F, Horowitz FG (1999) Analysis of potential field data in the wavelt domain. *Geophys J Int* 137:175–196
- Li X (2003) On the use of different methods for estimating magnetic depth. *Lead Edge* 22:1090–1099
- Miller HG, Singh VJ (1994) Potential field tilt – a new concept for location of potential field sources. *Appl Geophys* 32:213–217
- Milligan PR, Gunn PJ (1997) Enhancement and presentation of airborne geophysical data. *AGSO J Aust Geol Geophys* 17:63–75
- Nabigian MN, Hansen RO (2001) Unification of Euler and Werner deconvolution in three dimensions via the generalized Hilbert transform. *Geophysics* 66:1805–1810
- Naudy H (1971) Automatic determination of depth on aeromagnetic profiles. *Geophysics* 36:717–722
- Reid AB, Allsop JM, Granser H, Millett AJ, Somerton IW (1990) Magnetic interpretation in three dimensions using Euler deconvolution. *Geophysics* 55:80–91
- Roest WR, Verhoef J, Pilkington M (1992) Magnetic interpretation using the 3-D analytic signal. *Geophysics* 57:116–125
- Shi Z (1991) An improved Naudy-based technique for estimating depth from magnetic profiles. *Explor Geophys* 22:357–362
- Spector A, Grant FS (1970) Statistical models for interpreting aeromagnetic data. *Geophysics* 35:293–302
- Thurston JB, Smith RS (1997) Automatic conversion of magnetic data to depth, dip, and susceptibility contrast using the SPI™ method. *Geophysics* 62:807–813
- Werner S (1953) Interpretation of magnetic anomalies at sheet-like bodies. *Sveriges Geologiska Undersok Ser C Arsbok* 43(6):413

Magnetic Domains

Susan L. Halgedahl

Department of Geology and Geophysics,
University of Utah, Salt Lake City, UT, USA

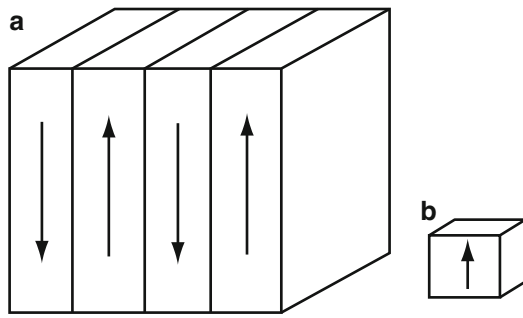
Definition

Magnetic domain	A magnetic domain is a region in which the magnetic spins of atoms or molecules are aligned. Many magnetic particles consist of two or more magnetic domains, each of uniform magnetization separated by narrow zones called domain walls, in which the spins change orientation from one domain to the next.
NRM (natural remanent magnetization)	The magnetic moment (per unit volume or per unit mass) acquired by a rock in nature.
TRM (thermal remanent magnetization)	The magnetic moment acquired by a rock as it cools from the Curie temperature in the Earth's field.

Introduction

The existence of a paleomagnetic record attests to the ability of magnetic minerals in rocks to preserve their natural remanent magnetizations over geologic time. The building block of *Remanent Magnetism* is the magnetic domain, a region where the magnetic spins are aligned in the same direction. In the early days of paleomagnetism, it was thought that the stable components of natural remanent magnetization (NRM) mainly resided in extremely small magnetic mineral grains that occupied the single-domain (SD) state. However, it is now recognized that, due to their very small size and scarcity, SD particles may not be the major carriers of NRM in many rocks. Instead, it is more likely that much of NRM is carried by grains that, by virtue of their larger sizes, are subdivided into several domains (Fig. 1). Adjacent domains are separated by domain walls, the transition regions where magnetic spins change direction from one domain to the next.

Remanences carried by SD assemblages can be highly stable, because destroying the remanence requires rotation of the particles' magnetic moments across strong energy barriers. In contrast, remanences carried by particles with two or more domains usually owe their stability to pinning of walls at crystal defects; these energy barriers are overcome more easily than those controlling SD grains. But another necessary criterion for paleomagnetic stability over geologic time is that a grain's overall domain state resists drastic



Magnetic Domains, Fig. 1 Illustrations of (a) a multidomain cube containing four domains and (b) a single-domain cube of uniform magnetization

reorganization caused by the changing physical/chemical conditions affecting the rock, such as moderately elevated temperature, moderate pressure, and some chemical alteration. For this reason, we discuss both theories and experiments to investigate how domain structure depends on grain size, applied field, and temperature in natural magnetic minerals and their synthetic analogues. (See detailed treatments of these subjects in standard texts such as Chikazumi (1964), Cullity (1972), Stacey and Banerjee (1974), and Dunlop and Ozdemir (1997)).

Classical Models of Magnetic Domain Structure

Exchange Energy

Models of magnetic domain structure are based on three energies: exchange energy, anisotropy energy, and magnetostatic energy. Magnetostatic energy is the fundamental reason why particles subdivide into two or more domains because, in so doing, this energy is reduced. However, this subdivision carries an energy “price”: the energy of domain walls. In most classical theories, the first two energies above make the largest contributions to the surface energy of the wall. From seminal calculations, Heisenberg (1928) predicted the existence of the exchange interaction, which determines whether adjacent magnetic spins are aligned parallel or antiparallel with respect to each other in a specific substance. In a relatively few materials, this interaction leads to a configuration of lowest energy when neighboring spin vectors are aligned. This gives rise to the rare phenomenon of spontaneous magnetization (M_s) and magnetic remanence. Such materials possess a Curie temperature, above which spontaneous magnetism vanishes, because (a) thermal agitation disrupts the alignment of adjacent spins, and (b) thermal expansion of the lattice increases spin–spin distances and thus weakens the exchange interaction.

Magnetic Anisotropy Energy

Magnetic anisotropy causes magnetic behavior to depend on the direction in which it is measured. In magnetite, for

example, magnetocrystalline anisotropy results in a lower external field being required to saturate a crystal along the $\langle 111 \rangle$ directions than along $\langle 100 \rangle$. Magnetocrystalline anisotropy arises from the coupling between spins and orbits of the electrons, so that work is required to rotate spins out of the directions of lowest energy (“easy” directions). “Easy” and “hard” directions depend on the specific material and its crystal structure. In cubic materials such as magnetite, the easy directions are along $\langle 111 \rangle$, whereas in iron $\langle 100 \rangle$ are easy directions (e.g., for more comprehensive discussions, see Chikazumi (1964); Cullity (1972)).

Also due to spin-orbit coupling, the lattice of a ferromagnetic material will spontaneously strain below the Curie temperature. The inverse effect is that external stress and/or variable internal strain can rotate the spontaneous magnetization away from the easy direction given by crystalline anisotropy. Depending on the material and the stress direction, applied stress may cause M_s to rotate either toward the stress axis or toward the direction perpendicular to the axis of stress.

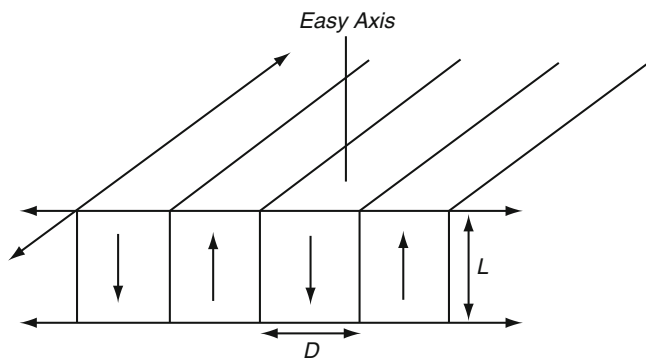
Magnetostatic Energy

Magnetostatic energy is the work, per unit volume, required to assemble a population of magnetic “free poles” into a particular geometric configuration. A simple but highly relevant case was addressed by Kittel (1949), who analyzed a semi-infinite plate of thickness L containing lamellar domains of uniform width D , with spontaneous magnetizations normal to the plate surface (Fig. 2). Kittel obtained $E_m = 1.705 M_s^2 D$, where E_m is the magnetostatic energy, per unit area of plate surface (erg/cm², in cgs). Kittel’s calculations were refined by Rhodes and Rowlands (1954), who addressed finite, rectangular grains. Their results produced “Rhodes and Rowlands” functions, which yield the total magnetostatic energy of rectangular grains of specified axial ratios.

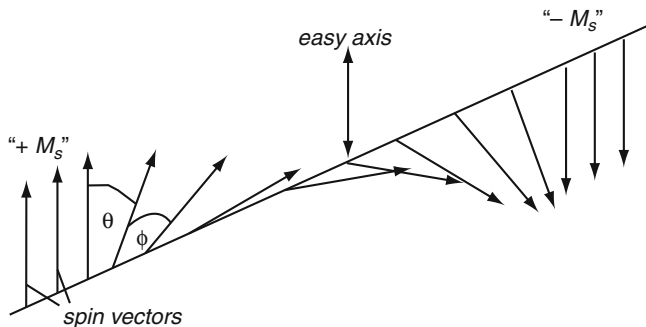
Energy and Width of the Domain Wall

A domain wall is the transition region where the spontaneous magnetization changes direction from one domain to the next, so that each domain is distinct from a wall. (Micromagnetic theories relax this assumption and will be discussed in later.) The two most important energies contributing to the wall’s energy and width are exchange energy and anisotropy energy, the latter due either to magnetocrystalline anisotropy or stress. The magnetostatic energy of the wall itself also plays a role; this energy was first analyzed by Amar (1958) and grows especially important when the wall width approaches that of the particle.

In the simplest models, spins in a wall reach an equilibrium configuration when magnetic anisotropy and exchange energies are balanced. For the 180° “Bloch” wall shown in Fig. 3 in which spins rotate through 180°, the energy per unit area of wall surface is $E_w = 2\pi (AK)^{1/2}$, where K is the



Magnetic Domains, Fig. 2 Illustration of the semi-infinite, magnetized plate of thickness L , whose magnetostatic energy was calculated by Kittel (1949). The plate contains domains of identical width D , with magnetizations perpendicular to the plate's surface. Walls are assumed to be infinitely thin



Magnetic Domains, Fig. 3 Illustration of spins in a 180° Bloch wall. θ is the angle between a spin and the easy axis of magnetization ϕ is the angle between adjacent spins

magnetic anisotropy constant due to magnetocrystalline anisotropy and/or stress and A is the material's exchange constant. Note that a submicroscopic change in either A or K can change the local wall energy. Such local changes constitute "wall pinning" sites, where wall motion is impeded by either an energy barrier or an energy trough. The width of a 180° wall is $\delta_w = \pi (A/K)^{1/2}$. In magnetite, values predicted for E_w and δ_w are approximately 0.9 erg/cm² and 0.3 μm , respectively (e.g., see Dunlop and Ozdemir 1997).

Domain Width Versus Grain Size

For the magnetized plate illustrated in Fig. 2, domain width D can be calculated for the lowest energy state by minimizing the sum of magnetostatic and wall energies per unit volume of material: this gives the half-power law $D = (1/M_s) (E_w L / 1.705)^{1/2}$ (Kittel 1949).

Magnetic minerals in rocks, however, rarely grow as thin platelets. Moskowitz and Halgedahl (1987) calculated the number of domains versus grain size in rectangular particles of $x = 0.6$ titanomagnetite ("TM60": $\text{Fe}_{2.4}\text{Ti}_{0.4}\text{O}_4$) containing planar domains separated by 180° walls. Magnetostatic

energy, including that from the walls' moments, was accounted for (Rhodes and Rowlands 1954; Amar 1958), and it was assumed that grains occupied states of absolute minimum energy. Their calculations yielded two principal results: (1) particles encompassing a wide range of grain sizes could contain the same number of domains, and (2) a plot of N (number of domains) versus L (grain thickness) could be fitted well by a power law $N \propto L^{1/2}$. It follows that domain width D also follows a half-power law in L . Thus, the general type of functional dependence of D on L is the same for both finite, rectangular grains and semi-infinite plates.

Single-Domain/Two-Domain Transition Size d_0

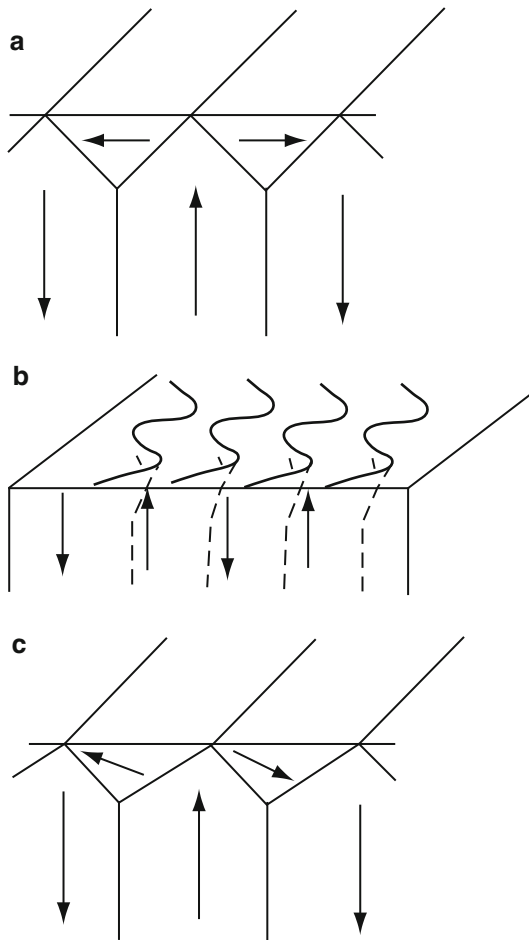
Grains containing only two or three domains can rival the remanences and coercivities of SD grains, and such particles are often referred to as being "pseudosingle-domain" (PSD). A particle may favor a two-domain over a SD state because the magnetostatic energy associated with two domains is much lower than that for saturation. However, below d_0 the energy price of adding a wall is too great to result in a state of minimum energy. At d_0 , the total energy of the two-domain state and the SD state are equal. PSD grains can be common in many rocks and, in terms of interpreting rock magnetic behavior, it is important to know the grain sizes corresponding to the onset of PSD behavior in various magnetic minerals.

This transition size depends on the material's magnetic properties, its state of stress, the particle shape, and temperature. Moskowitz and Banerjee (1979) calculated total energies of SD, two-domain, and three-domain cubes of stress-free magnetite containing lamellar domains at room temperature. Magnetostatic energy of the walls was included in their calculations (Amar 1958). They obtained $d_0 \approx 0.08 \mu\text{m}$. Similar calculations for TM60 by Moskowitz and Halgedahl (1987) yielded $d_0 \approx 0.5 \mu\text{m}$ for unstressed particles at room temperature. Raising the stress level to 100 MPa shifted d_0 upward to 1 μm .

Domains and Domain Walls Near Crystal Surfaces

In relatively thick crystals, domains and domain walls may change their geometric styles near and at crystal surfaces, in order to lower the total magnetostatic energy with respect to that of the "open" structure shown in Fig. 2. The particular style depends largely on the dominant kind of anisotropy, as well as on the relative strengths of magnetostatic and anisotropy energies. When $2\pi M_s^2/K \gg 1$ in a uniaxial material, prism-shaped closure domains bounded by 90° walls may completely close off magnetic flux at the crystal surface (Fig. 4a). Because closure domains greatly reduce magnetostatic energy, the "body" domains can be several times broader than predicted for the "open" structure.

When $2\pi M_s^2/K \ll 1$ in a uniaxial substance, a large amount of anisotropy energy results when M_s is perpendicular



Magnetic Domains, Fig. 4 Three styles in which domains and domain walls may terminate at a crystal surface. (a) Prismshaped surface closure domains at the surface of a material which is either uniaxial, with $2\pi M_s^2/K \gg 1.0$, or cubic, such as iron, whose easy axes are along $\langle 100 \rangle$. Here, 90° walls separate closure domains from the principal “body” domains that fill most of the crystal. Arrows indicate the sense of spontaneous magnetization within the domains. (b) Wavy walls at the surface of a uniaxial material with $2\pi M_s^2/K < 1.0$. Waviness dies out with increasing distance from the surface. (c) Prism-shaped closure domains at the surface of a cubic material, such as magnetite, whose easy directions of magnetization are along $\langle 111 \rangle$. Closure domains and body domains are separated by 71° and 109° walls

to the “easy” axis, making the style of surface closure shown in Fig. 4a energetically unfavorable. Instead, walls that are planar within the body of the crystal can become wavy at the surface (Fig. 4b). In extremely thick crystals, wavy walls can alternate with rows of reverse spikes. These elaborate surface domain structures lower magnetostatic energy by achieving a closer mixture of “positive” and “negative” free magnetic poles (e.g., see Szymczak 1968).

Large crystals governed by cubic magnetocrystalline anisotropy also reduce surface flux through prism-shaped closure domains at the surface. When $\langle 100 \rangle$ are easy directions, as in iron, closure domains are bounded by 90° walls (Fig. 4a). When $\langle 111 \rangle$ are easy directions, as in

magnetite, closure domains are bounded by 71° and 109° walls (Fig. 4c).

Temperature Dependence of Domain Structure

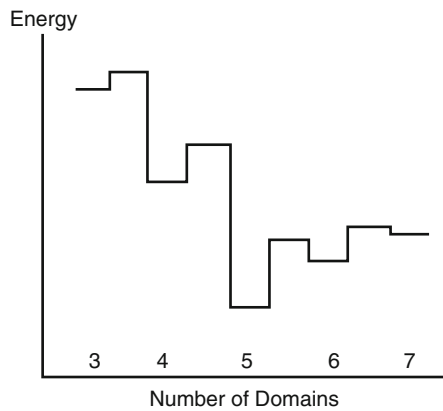
Understanding how domain structure evolves during both heating to, and cooling from, the Curie point is crucial to understanding the acquisition and thermal stability of thermal remanent magnetization (TRM). If the number of domains changes significantly during cooling from the Curie point in a weak field, it is reasonable to hypothesize that TRM will not become “frozen in,” or blocked, until the overall domain structure reaches a stable configuration.

According to Kittel’s original model (Fig. 2), grains will nucleate (add) domain walls and domains with heating in zero field, if the wall energy drops more rapidly with increasing temperature than does the magnetostatic term. Energy-wise, in this first case, a particle can “afford” to add domains with heating. Conversely, during cooling from the Curie point, a grain will denucleate (lose) domains and domain walls if wall energy rises more quickly than does the magnetostatic energy with decreasing temperature. If wall energy drops less rapidly with increasing temperature than does the magnetostatic term, then the opposite scenarios apply. This predicted behavior relies on the assumptions that the particle is able to maintain a global energy minimum (GEM) domain state at all temperatures and that the total magnetostatic energy of the walls themselves can be ignored.

Using Amar’s (1958) model, Moskowitz and Halgedahl (1987) calculated the number of domains between room temperature and the Curie point in parallel-epipeds of TM60. As discussed earlier, they investigated two cases: dominant crystalline anisotropy (zero stress) and high stress ($\sigma = 100$ MPa). In TM60 grains larger than a few micrometers, most of their results gave an increase in the number of domains with heating. Exceptions to this overall pattern were cases in which walls broadened so dramatically with increasing temperature that they nearly filled the particle and rendered nucleation unfavorable. During cooling from the Curie point in zero field, the domain “blocking temperature” – the temperature below which the number of domains remained constant – increased both with decreasing grain size and with internal stress.

Micromagnetic Models

In contrast to classical models of magnetic domain structure, micromagnetic models do not assume the presence of discrete domains. Instead, they allow the orientations of M_s -vectors to vary among extremely small subvolumes, into which a grain is divided. A stable configuration is obtained numerically when the grand sum of exchange, anisotropy, and magnetostatic energies over all subvolumes is minimum.

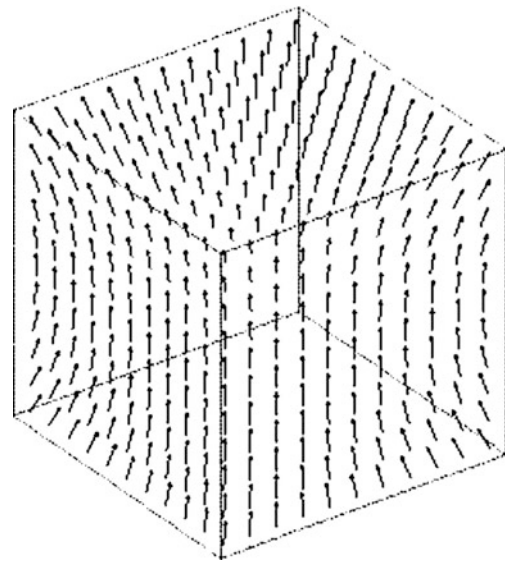


Magnetic Domains, Fig. 5 Diagram illustrating the relative energies and energy barriers associated with local energy minimum (LEM) domain states. Each LEM state is characterized by a unique number of domains and is separated from adjacent states by nucleation and denucleation energy barriers. In this diagram, the global energy minimum, or GEM, domain state has five domains

In rock magnetism, micromagnetic models have focused on magnetite, and they assume unstressed, defect-free crystals.

Moon and Merrill (1984, 1985) were the first in rock magnetism to construct one-dimensional (1D) micromagnetic models for defect-free magnetite cubes. Not only did they calculate the energies resulting from different numbers of domains within a grain of given size, but they also calculated the energy barriers associated with nucleation and denucleation of domains and domain walls. Their major breakthrough was the discovery that a particle can occupy a range of local energy minimum (LEM) domain states. Each LEM state is characterized by a unique number of domains and is separated from adjacent states by energy barriers. As illustrated in Fig. 5, the GEM state is the configuration of lowest energy but, owing to the energy barriers between states, a LEM state can be quite stable as well.

Several authors have extended micromagnetic calculations for magnetite to two dimensions (2D) and three dimensions (3D) (e.g., Williams and Dunlop 1989, 1990, 1995; Newell et al. 1993; Xu et al. 1994; Fabian et al. 1996; Fukuma and Dunlop 1998; Williams and Wright 1998). These models yield a variety of exotic, nonuniform configurations of magnetization, such as “flower” and “vortex” states. Analogous to Moon and Merrill’s results, these models produce both LEM and GEM states, although with very different spin structures from those of 1D models. For example, Fig. 6 illustrates a “flower” state in a cube magnetized parallel to the z -axis. The flower state is reminiscent of a classical SD state of uniform magnetization, except that the M_s -vectors are canted at and near the crystal surface. In relatively large cubes of magnetite – for example, 4 or 5 μm – both 2D and 3D models yield configurations closely



Magnetic Domains, Fig. 6 Illustration of a “flower” state obtained through three-dimensional micromagnetic modeling of a cube largely magnetized along the cube’s z -axis

approaching those of classical domain structures expected for magnetite, such as “body” domains and closure domains at the surface (e.g., Xu et al. 1994; Williams and Wright 1998). In submicron magnetite undergoing hysteresis, magnetization reversal can occur through LEM–LEM transitions (e.g., flower to vortex state) (e.g., Williams and Dunlop 1995).

Dunlop et al. (1994) used 1D micromagnetic models to investigate LEM states in stress-free magnetite particles during TRM acquisition. Their study was motivated by experimental results reported earlier by Halgedahl (1991) (see discussions below). Dunlop et al. (1994) calculated the energy barriers for all combinations among SD/two-domain/three-domain transitions during cooling from the Curie point of magnetite in a weak external field. According to their results, after acquiring TRM under conditions of thermal equilibrium, populations would be overwhelmingly biased toward GEM domain states, and an individual particle should not exhibit a range of LEM states after several identical TRM runs.

Using renormalization group theory, Ye and Merrill (1995) arrived at a different conclusion. According to their calculations, short-range ordering of spins just below the Curie point could give rise to a variety of LEM states in the same particle after replicate coolings.

Experimental Studies of Magnetic Domains

Methods of Imaging Domains and Domain Walls

Rock magnetists have mainly used three methods to image domains and domain walls: the Bitter pattern method,

which uses liquid magnetic colloid to image walls, with a resolution limit of about 1 mm (e.g., see details in Halgedahl and Fuller 1983); the magneto-optical Kerr effect (MOKE), which images entire domains from the rotation of the polarization plane of light reflected by a magnetized specimen (e.g., Hoffmann et al. 1987; Worm et al. 1991; Heider and Hoffmann 1992; Ambatiello et al. 1999); and magnetic force microscopy (MFM), which can image magnetic features as small as 0.01 mm through the voltage induced in a vibrating magnetic needle situated very close to a particle's surface (Williams et al. 1992; Proksch et al. 1994; Moloni et al. 1996; Pokhil and Moskowitz 1996, 1997; Frandson et al. 2004). Two other methods – transmission electron microscopy (TEM) and off-axis electron holography – have been used in some studies of magnetite, most notably magnetotactic bacteria and magnetite intergrowths in ulvospinel (Dunin-Borowsky et al. 2001; McCartney et al. 2001; Harrison et al. 2002).

Styles of Domains Observed in Magnetic Minerals of Paleomagnetic Significance

In rock magnetism, the majority of domain observation studies have focused on four magnetic minerals, all important to paleomagnetism: pyrrhotite (Fe_7S_8), titanomagnetite of roughly intermediate composition (near $\text{Fe}_{2.4}\text{Ti}_6\text{O}_4$, or TM60), magnetite (Fe_3O_4), and hematite (Fe_2O_3).

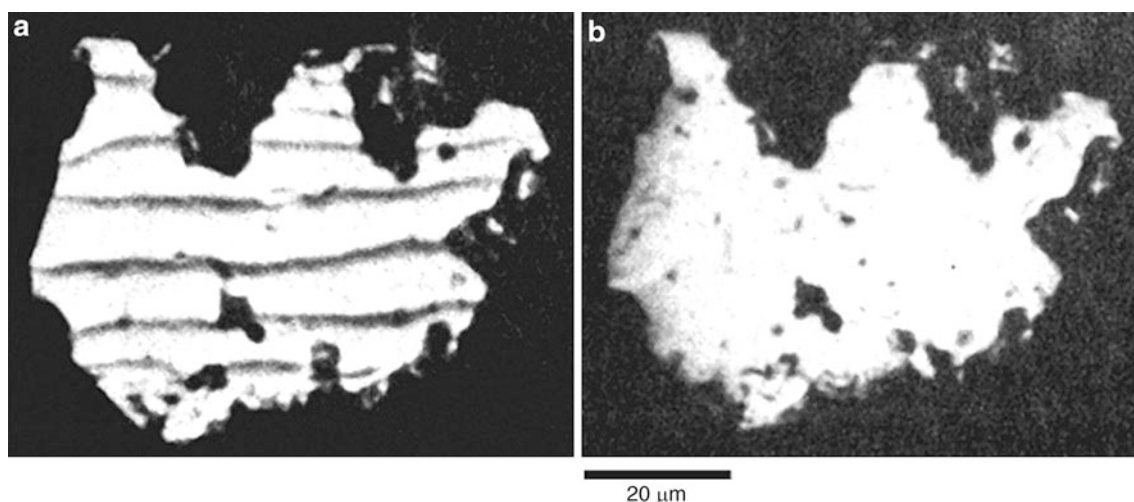
Owing to its high magnetocrystalline anisotropy constant and relative insensitivity to stress, pyrrhotite behaves magnetically as a uniaxial material. When studied with the Bitter colloid method, pyrrhotite often exhibits fairly simple domain patterns which suggest lamellar domains separated by 180° walls (Fig. 7a) (Soffel 1977b; Halgedahl and Fuller 1983).

Despite being cubic, intermediate titanomagnetites rarely, if ever, exhibit the arrays of closure domains, 71° , and 109° walls predicted from theories. Instead, these minerals often display very complex patterns of densely spaced, curved walls (Appel and Soffel 1984, 1985). However, some grains do exhibit a simple array of parallel walls, such as that shown in Fig. 8a (e.g., Halgedahl and Fuller 1980, 1981), but it is not unusual to observe wavy walls alternating with rows of reverse spikes (Halgedahl 1987; Moskowitz et al. 1988). Both simple and wavy patterns suggest a dominant, internal stress that yields a uniaxial anisotropy, although the origin of this stress is still unclear.

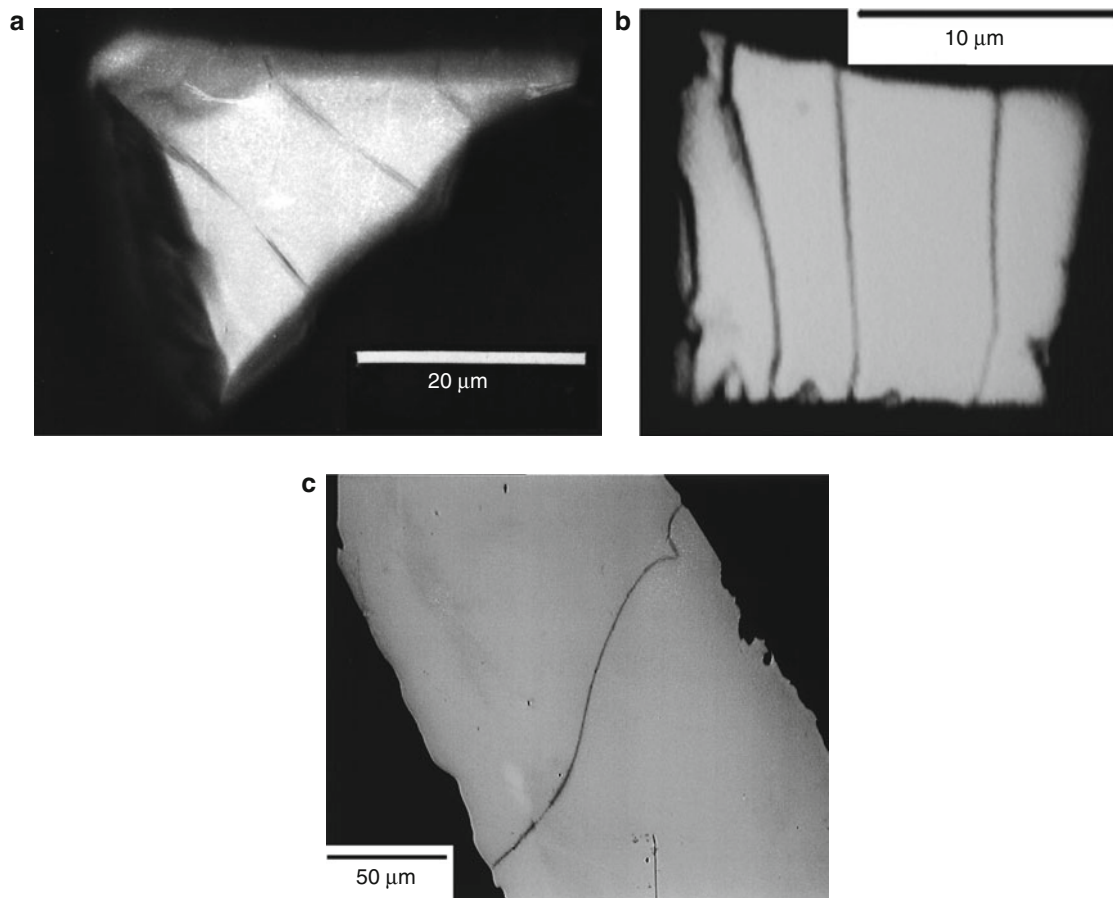
Small magnetite grains grown in a synthetic rocklike matrix with the glass-ceramic method (Worm and Markert 1987) generally display simple arrays of straight walls (e.g., Worm et al. 1991; Geiß et al. 1996) (Fig. 8b). In such samples, there appears to be a paucity of closure domains, perhaps the result of high internal stress generated by quenching during synthesis.

In contrast, domain images obtained from large crystals of natural magnetite and small magnetite grains in a variety of rocks reveal 180° , 71° , and 109° walls and closure domains, in accordance with theory (Bogdanov and Vlasov 1965, 1966; Smith 1980; Boyd et al. 1984; Ozdemir and Dunlop 1993, 1997, 2006; Ozdemir et al. 1995; Ambatiello et al. 1999). Evidently, internal stress is low in these natural magnetites, so that magnetocrystalline anisotropy is dominant.

To date, domain studies on hematite have been limited to large (e.g., 100 μm to 1 mm) platelets. Even large crystals such as these contain very few walls and a fairly simple domain structure, owing to hematite's weak spontaneous



Magnetic Domains, Fig. 7 Bitter patterns on a particle of natural pyrrhotite (a) after demagnetization in an alternating field of 1000 Oe and (b) in an apparently SD-like state after acquiring saturation remanence in 15 kOe



Magnetic Domains, Fig. 8 (a) Bitter pattern on a grain of intermediate titanomagnetite ($x \approx 0.6$) in oceanic basalt drilled near the Mid-Atlantic Ridge. (b) Bitter pattern on a grain of magnetite synthesized with the glass-ceramic method. In this particular state of magnetization, the grain contains four walls, whose lengths are commensurate with the particle's length. Note that one wall is pinned very near the particle's extreme left-hand edge. The small triangular patterns at the lower edge of the particle

represent walls which enclose small reverse spike domains. (c) Bitter pattern of a wall on a very large (approximately 150 μm width, 1 mm length) platelet of natural hematite from Elba, Italy. At most this platelet exhibits only 1–2 principal walls, although small edge domains are often observed. Apparently, the wall is bowing around a defect near the upper right-hand edge of the photograph. Only a small part of the crystal surface is shown

magnetization (about 2 emu/cm^3) and low magnetostatic energy (Fig. 8c) (Halgedahl 1995, 1998).

Number of Domains Versus Grain Size

Both Kittel's original model of domains in a semi-infinite platelet and calculations for finite grains by (e.g., Moskowitz and Halgedahl 1987) lead to the prediction that domain width $D \propto L^{1/2}$, where L is plate or particle thickness. These predictions are supported by domain studies of natural magnetic minerals which generally yield a power-law dependence of D on L , although the power may differ somewhat from 0.5.

Soffel (1971) was the first to study the grain-size dependence of the number of domains in a paleomagnetically important magnetic mineral. From Bitter patterns on grains of natural, intermediate ($x \approx 0.55$) titanomagnetite in a basalt, he determined that, on average, both D and N (number of domains = L/D) $\propto L^{1/2}$, where L was average particle size. His results yielded a single domain/two domain

transition size of 0.6 μm and a wall energy density of about 1 erg/cm^2 for this composition.

Similarly, from Bitter patterns on natural pyrrhotite in a Bavarian diabase, both Soffel (1977b) and Halgedahl and Fuller (1983) found that $D \propto L^{0.40}$ to $D \propto L^{0.45}$, depending on the magnetization state. For this pyrrhotite sample, the SD boundary fell between 1.5 and 2 μm . Geiß et al. (1996) obtained $D \propto L^{0.45}$ from Bitter patterns on synthetic magnetite particles grown in a glass-ceramic matrix (Worm and Markert 1987). They obtained a SD/two-domain transition size for magnetite of approximately 0.25 μm .

Domain Wall Widths

When studied in the scanning electron microscope (SEM), patterns of dried magnetic colloid afford high-resolution views of domain walls. Moskowitz et al. (1988) applied this method to polycrystalline pellets of synthetic titanomagnetite substituted with aluminum and magnesium

($\text{Fe}_{2.2}\text{Al}_{0.1}\text{Mg}_{0.1}\text{Ti}_{0.6}\text{O}_4$: “AMTM60”). Dried Bitter patterns on unpolished surfaces were virtually identical in style to those common to materials controlled by strong uniaxial anisotropy. Patterns indicated closely spaced stripe domains, sinusoidally wavy walls, and nested arrays of reverse spikes. Measurements from SEM images yielded wall widths between 0.170 and 0.400 μm .

Very high resolution images of domains and domain walls in magnetite have been obtained with MFM. The first images were reported by Williams et al. (1992), who recorded a magnetic profile across a 180° wall on a $\{110\}$ surface of a large natural magnetite crystal. Spins within the wall reversed their polarity of rotation along the length of the wall, demonstrating that walls in real materials can be much more complex than those portrayed by simple models. Using MFM, Pokhil and Moskowitz (1996, 1997) made a similar finding in glass-ceramic magnetite. Segments of opposite polarity were separated by Bloch lines, the transition regions where polarity changes sense. The number of Bloch lines within a wall varied among repeated alternating field (AF) demagnetization treatments. Thus walls, like particles, can occupy LEM states.

Owing to its high-resolution capabilities, the MFM can provide estimates of wall width. Proksch et al. (1994) obtained MFM profiles across a 180° wall on a $\{110\}$ surface of natural magnetite, thus obtaining a wall width of about 0.21 μm .

Experimental Evidence for Local Energy Minimum (LEM) Domain States

In their study of Bitter patterns on natural pyrrhotite, Halgedahl and Fuller (1983) noted that grains of virtually identical size could contain very different numbers of domains, despite these same particles having undergone the same magnetic treatments. Moreover, they found that an individual particle could arrive in very different domain states – that is, with different numbers of walls – after different cycles of minor hysteresis.

These observations led Halgedahl and Fuller (1983) to the conclusion that a particle could occupy domain states other than the ground state. Subsequent domain observation studies by Halgedahl on AMTM60 and by Geiß et al. (1996) on magnetite provided strong evidence for LEM states in other magnetic minerals.

LEM States and Thermomagnetic Treatments

A particularly unexpected type of LEM state in pyrrhotite and intermediate titanomagnetite is a SD-like state that certain particles can occupy after being saturated in a strong external field, even though these same particles readily accommodate walls in other states of magnetization. Bitter patterns on intermediate ($x \approx 0.6$) titanomagnetite in oceanic basalt and on natural pyrrhotite in diabase were studied during

hysteresis by Halgedahl and Fuller (1980, 1983). In states of saturation remanence, most particles in a large population contained one or more walls, as expected on the basis of previous theories (Fig. 7a). However, several tens of percent of the finer (5–15 μm) particles appeared saturated after the maximum field was shut off (Fig. 7b). Halgedahl and Fuller (1980, 1983) proposed that grains that failed to nucleate walls could make a substantial contribution to saturation remanence. Particles in SD-like states remained saturated, until nucleation was accomplished by applying a back field of sufficient magnitude, and this nucleation field dropped off with increasing grain size L according to a power law in $L^{-1/2}$. Similarly, Boyd et al. (1984) reported Bitter patterns on natural magnetite grains carrying saturation remanence, and these patterns suggested SD-like states. These results were surprising, in view of magnetite’s strong tendency to self-demagnetize.

Subsequent domain observations and hysteresis measurements of hematite platelets from Elba, Italy demonstrated that, after exposure to strong fields, even large crystals could arrive in states of near-saturation. Back fields were necessary to nucleate principal domains, and, similar to pyrrhotite, these nucleation fields also followed a power law in $L^{-1/2}$ (Halgedahl 1995, 1998).

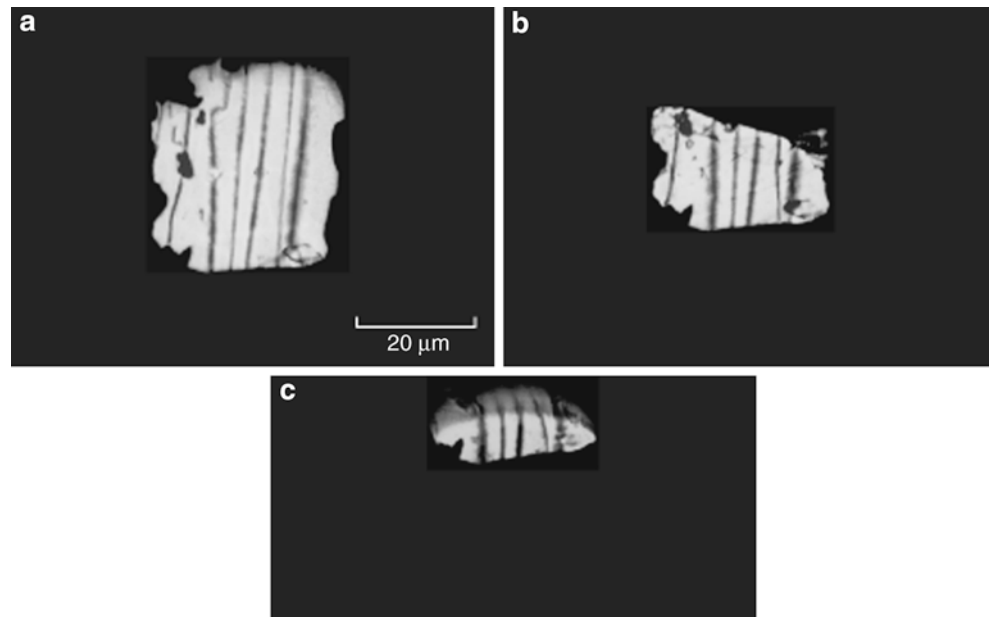
Results from titanomagnetite and pyrrhotite suggesting SD-like states were interpreted by Halgedahl and Fuller (1980, 1983) in light of previous theoretical and experimental work on high-anisotropy, industrial magnets. In such materials, nucleation of walls is an energetically difficult process, because the internal demagnetizing field, which triggers nucleation, may be too weak to overcome the anisotropy field, which hinders nucleation (Brown 1963; Becker 1969, 1971a, b, 1976). Such particles may require an external back field, which aids the demagnetizing field in order to nucleate walls and accompanying domains.

The style of domain structure and the range of LEM states that a particle can occupy can depend on thermomagnetic history. From Bitter patterns on natural, polycrystalline pyrrhotite, Halgedahl and Fuller (1981) discovered that crystallites displayed arrays of undulating walls on their surfaces after acquiring TRM in a weak external field. By contrast, after AF demagnetization in a strong peak field the same crystallites exhibited planar walls. Evidently, cooling locked in a high-temperature configuration of walls whose curved shapes promoted lower magnetostatic energy than would a planar geometry.

As indicated by Moon and Merrill’s theoretical results for magnetite, a particular LEM state can be stable across a broad range of grain size. This prediction is born out by experiments by Halgedahl and Ye (2000), who investigated the effects of mechanical thinning on domain states in natural pyrrhotite. Individual pyrrhotite grains in a diabase were mechanically thinned and their Bitter patterns observed after

Magnetic Domains,

Fig. 9 Bitter patterns on a grain of natural pyrrhotite in a diabase before and after mechanical thinning. (a) Initial state, before thinning. (b) After thinning the particle to about one-half of its original length along a direction parallel to the trends of the Bitter lines. (c) Final state, after the particle has been thinned to about one-fourth of its original length



each thinning step. Despite some grains being thinned to about one-fourth of their initial diameter, the widths of surviving domains and positions of surviving walls remained unaffected (Fig. 9). Neither nucleations nor denucleations were observed, although calculations indicated that thinning would cause significant changes in GEM domain states.

Bitter patterns on crystallites of polycrystalline AMTM60 were studied after many replicate TRM acquisition and AF demagnetization experiments (Halgedahl 1991). In each particle, the number of domains varied from one experiment to the next, defining a distribution of LEM states. For states of weak-field TRM, this distribution could be broad and could include single-domain-like states (Fig. 10). After replicate AF demagnetizations, however, typical distributions were narrow and clustered about a most probable state, possibly the GEM state.

Evolution of Magnetic Domain Structures at Elevated Temperatures

The manner in which domain structure evolves with temperature carries clear implications for TRM. Cooling-induced nucleations and denucleations of walls/domains could trigger sudden changes of the internal demagnetizing field. As a result, preexisting walls which survive a domain transition could be dislodged from imperfections where they were pinned initially at higher temperatures.

The first Bitter patterns observed on magnetite above room temperature in the Earth's field were made by Heider et al. (1988), using hydrothermally recrystallized crystals. Depending on the particle, Bitter patterns could be followed to approximately 200 °C, above which temperature the

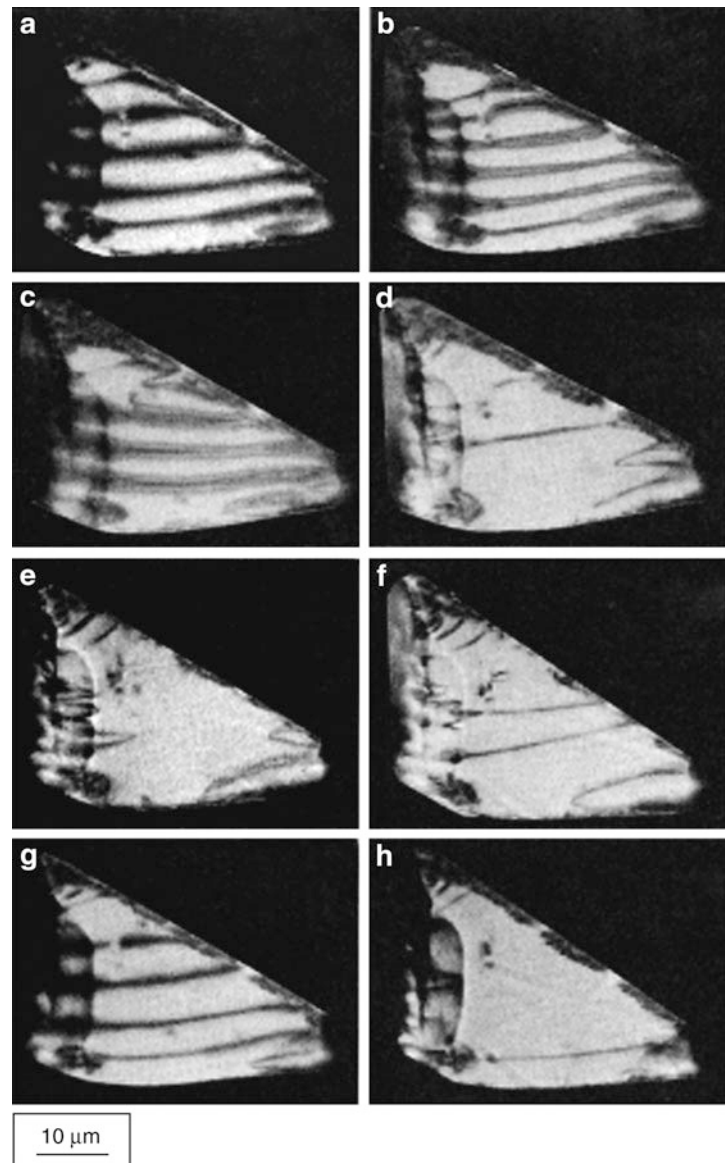
patterns grew too faint to discern. Surprisingly, heating to very moderate temperatures drove certain walls across much of a particle; in some cases, denucleation occurred. Upon cooling, walls reassembled in a similar, though not identical, arrangement to that observed initially at room temperature. In some cases, repeated thermal cycling between room temperature and about 200 °C produced different numbers of domains in the same particle.

Ambatiello et al. (1999) used the MOKE to study domain widths versus temperature in several large (several mm) crystals of natural magnetite. At room temperature, {110} planes were dominated by broad lamellar domains that terminated in closure domains at crystal edges. On {111} planes, they found complex, nested arrays of very small closure domains, which finely subdivided the main closure structure. Domain widths generally increased with heating, and such changes were thermally reversible. To explain their results, these authors hypothesized that heating promoted an increasingly finer subdivision into small domains near the crystal surface; this would reduce magnetostatic energy with heating and thus cause broadening of body domains (also see Muxworthy and Williams 2006).

Bitter patterns observed at elevated temperatures on natural, intermediate titanomagnetites were reported by Soffel (1977a), Metcalf and Fuller (1987a, b, 1988), and Halgedahl (1987). During heating, patterns gradually faded to obscurity, with few significant changes in domain structure. Remarkably, some titanomagnetite particles studied by Metcalf and Fuller (1987a, b, 1988) displayed no Bitter lines after cooling from the Curie point in the Earth's field. This observation provided evidence for Halgedahl

Magnetic Domains,

Fig. 10 Bitter patterns on a crystallite of titanomagnetite (“AMTM60”: see text) after each of eight replicate TRM acquisition runs in the Earth’s field



and Fuller’s (1983) earlier proposal that weak-field TRM acquired by populations of PSD grains could, in part, be attributable to particles which failed to nucleate walls during cooling.

The possible importance of LEM states to TRM acquisition was raised by work on synthetic, polycrystalline AMTM60 with a Curie point near 75 °C (Halgedahl 1991). As shown in Fig. 10, replicate TRM experiments often produced different numbers of domains in the same particle. Some of these states suggested that the particle was entirely saturated, with no visible Bitter lines, or nearly saturated, with only small spike domains at grain boundaries. Patterns that evolved during cooling revealed that denucleation was one mechanism by which a particle arrived in a final LEM state. Denucleation occurred either

through contraction of large, preexisting spike domains, or by straight walls moving together and coalescing into spikes. Often, spikes would collapse altogether. In some cases, denucleation left behind large volumes which, apparently, contained no walls, although walls were present elsewhere in the grain. In other cases, denucleation left behind a grain that appeared nearly saturated, but for small edge domains. These results led to the idea of trans-domain TRM (Halgedahl 1991).

Magnetic Domain Structure in Magnetite at Low Temperatures

Magnetite undergoes two types of transitions at low temperatures, which can profoundly affect domain structure and remanence (e.g., see detailed discussions in Stacey and

Banerjee (1974), Dunlop and Ozdemir (1997)). At the isotropic point (approximately 130 K), the first magnetocrystalline anisotropy constant, K_1 , passes through zero as it changes sign from negative at temperatures above the transition to positive below it. At the Verwey transition, T_v (approximately 120 K), magnetite undergoes a crystallographic transition from cubic to monoclinic.

Thus at 130 K domain walls should broaden dramatically, because wall width is proportional to $K_1^{-1/2}$. It follows that, by passage through 130 K, walls may break free of narrow defects which pinned them at other temperatures. At the Verwey transition, the easy axis of magnetization changes direction. In multidomain and pseudosingle-domain magnetite, these thermal passages may cause demagnetization and complete reorganization of domain structure. Low-temperature demagnetization has proved useful in removing certain spurious components of NRM which are surprisingly resistant to thermal demagnetization above room temperature. Thus it is important to determine how these transitions affect domains.

Using an MFM specially adapted to operate at low temperature, Moloni et al. (1996) were the first to image domains on {110} magnetite planes at temperatures near the two transitions. As the crystal was warmed from 77 K to a few degrees just below T_v , the domain structure disappeared entirely below the instrument's noise level, evidently undergoing a complete reorganization near the crystallographic transition.

Cross-References

- [Paleomagnetic Field Intensity](#)
- [Paleomagnetism, Principles](#)
- [Remanent Magnetism](#)

Bibliography

- Amar H (1958) Magnetization mechanism and domain structure of multidomain particles. *Phys Rev* 111:149–153
- Ambatiello A, Fabian K, Hoffmann V (1999) Magnetic domain structure of multidomain magnetite as a function of temperature: observation by Kerr microscopy. *Phys Earth Planet Inter* 112:55–80
- Appel E, Soffel HC (1984) Model for the domain state of Ti-rich titanomagnetites. *Geophys Res Lett* 11:189–192
- Appel E, Soffel HC (1985) Domain state of Ti-rich titanomagnetites deduced from domain structure observations and susceptibility measurements. *J Geophys* 56:121–132
- Becker JJ (1969) Observations of magnetization reversal in cobalt-rare-earth particles. *IEEE Trans Magn MAG-5*:211–214
- Becker JJ (1971a) Magnetization discontinuities in cobalt-rare-earth particles. *J Appl Phys* 42:1537–1538
- Becker JJ (1971b) Interpretation of hysteresis loops of cobalt-rare-earth particles. *IEEE Trans Magn MAG-7*:644–647
- Becker JJ (1976) Reversal mechanism in copper-modified cobalt-rare-earth particles. *IEEE Trans Magn MAG-12*:965–967
- Bogdanov AK, Vlasov AY (1965) Domain structure in a single crystal of magnetite, (English trans.), *Izv. Earth Phys* 1:28–32
- Bogdanov AA, Vlasov AY (1966) The domain structure of magnetite particles, (English trans.), *Izv. Phys Solid Earth* 9:577–581
- Boyd JR, Fuller M, Halgedahl S (1984) Domain wall nucleation as a controlling factor in the behaviour of fine magnetic particles in rocks. *Geophys Res Lett* 11:193–196
- Brown WF (1963) *Micromagnetics*. Wiley, New York, p 143
- Chikazumi S (1964) *Physics of Magnetism*. Wiley, New York, p 664
- Cullity BD (1972) *Introduction to Magnetic Materials*. Addison-Wesley, Reading, MA, p 666
- Dunin-Borowsky RE, McCartney MR, Posfai M, Frankel RB, Bazylinski DA, Buseck PR (2001) Off-axis electron holography of magnetotactic bacteria: magnetic microstructure of strains MV-1 and MS-1. *Eur J Mineral* 13(4):671–684
- Dunlop DJ, Ozdemir O (1997) *Rock Magnetism: Fundamentals and Frontiers*. Cambridge University Press, Cambridge, UK, 573 pp
- Dunlop DJ, Newell AJ, Enkin RJ (1994) Transdomain thermoremanent magnetization. *J Geophys Res* 99:19741–19755
- Fabian K, Kirchner A, Williams W, Heider F, Leibl T (1996) Three-dimensional micromagnetic calculations for magnetite using FFT. *Geophys J Int* 124:89–104
- Foss S, Moskowitz B, Walsh B (1996) Localized micromagnetic perturbation of domain walls in magnetite using a magnetic force microscope. *Appl Phys Lett* 69:3426–3428
- Foss S, Moskowitz BM, Proksch R, Dahlberg ED (1998) Domain wall structures in single-crystal magnetite investigated by magnetic force microscopy. *J Geophys Res* 103:30551–30560
- Fransson C, Stipp SLS, McEnroe SA, Madsen MB, Knudsen JM (2004) Magnetic domain structures and stray fields of individual elongated magnetite grains revealed by magnetic force microscopy (MFM). *Phys Earth Planet Inter* 141:121–129
- Fukuma K, Dunlop DJ (1998) Grain-size dependence of two-dimensional micromagnetic structures for pseudo-single-domain magnetite (0.2–2.5 μm). *Geophys J Int* 134:843–848
- Geiß CE, Heider F, Soffel HC (1996) Magnetic domain observations on magnetite and titanomagnetite grains (0.5–10 μm). *Geophys J Int* 124:75–88
- Halgedahl SL (1987) Domain pattern observations in rock magnetism: progress and problems. *Phys Earth Planet Inter* 46:127–163
- Halgedahl SL (1991) Magnetic domain patterns observed on synthetic Ti-rich titanomagnetite as a function of temperature and in states of thermoremanent magnetization. *J Geophys Res* 96:3943–3972
- Halgedahl SL (1995) Bitter patterns versus hysteresis behavior in small single particles of hematite. *J Geophys Res* 100:353–364
- Halgedahl SL (1998) Barkhausen jumps in larger versus small platelets of natural hematite. *J Geophys Res* 103:30575–30589
- Halgedahl S, Fuller M (1980) Magnetic domain observations of nucleation processes in fine particles of intermediate titanomagnetite. *Nature* 288:70–72
- Halgedahl SL, Fuller M (1981) The dependence of magnetic domain structure upon magnetization state in polycrystalline pyrrhotite. *Phys Earth Planet Inter* 26:93–97
- Halgedahl S, Fuller M (1983) The dependence of magnetic domain structure upon magnetization state with emphasis upon nucleation as a mechanism for pseudo-single domain behavior. *J Geophys Res* 88:6505–6522
- Halgedahl SL, Ye J (2000) Observed effects of mechanical grain-size reduction on the domain structure of pyrrhotite. *Earth Planet Sci Lett* 178:457–467

- Harrison TJ, Dunin-Borkowski RE, Putnis A (2002) Direct imaging of nanoscale magnetic interactions in minerals. *Proc Natl Acad Sci* 99:16556–16561
- Heider F (1990) Temperature dependence of domain structure in natural magnetite and its significance for multi-domain TRM models. *Phys Earth Planet Inter* 65:54–61
- Heider F, Hoffmann V (1992) Magneto-optical Kerr effect on magnetite crystals with externally applied fields. *Earth Planet Sci Lett* 108:131–138
- Heider F, Halgedahl SL, Dunlop DJ (1988) Temperature dependence of magnetic domains in magnetite crystals. *Geophys Res Lett* 15:499–502
- Heisenberg W (1928) Zur theorie de Ferromagnetismus. *Z Phys* 49:619–636
- Hoffmann V, Schafer R, Appel E, Hubert A, Soffel H (1987) First domain observations with the magneto-optical Kerr effect on Ti-ferrites in rocks and their synthetic equivalents. *J Magn Magn Mater* 71:90–94
- Kittel C (1949) Physical theory of ferromagnetic domains. *Rev Mod Phys* 21:541–583
- McCartney MR, Lins U, Farina M, Buseck PR, Frankel RB (2001) Magnetic microstructure of bacterial magnetite by electron holography. *Eur J Mineral* 13(4):685–689
- Metcalf M, Fuller M (1987a) Magnetic remanence measurements of single particles and the nature of domain patterns in titanomagnetites. *Geophys Res Lett* 14:1207–1210
- Metcalf M, Fuller M (1987b) Domain observations of titanomagnetites during hysteresis at elevated temperatures and thermal cycling. *Phys Earth Planet Inter* 46:120–126
- Metcalf M, Fuller M (1988) A synthetic TRM induction curve for fine particles generated from domain observations. *Geophys Res Lett* 15:503–506
- Moloni K, Moskowitz BM, Dahlberg ED (1996) Domain structures in single crystal magnetite below the Verwey transition as observed with a low-temperature magnetic force microscope. *Geophys Res Lett* 23:2851–2854
- Moon T, Merrill RT (1984) The magnetic moments of non-uniformly magnetized grains. *Phys Earth Planet Inter* 34:186–194
- Moon TS, Merrill RT (1985) Nucleation theory and domain states in multidomain magnetic material. *Phys Earth Planet Inter* 37:214–222
- Moskowitz BM, Banerjee SK (1979) Grain size limits for pseudosingle domain behavior in magnetite: implications for paleomagnetism. *IEEE Trans Magn MAG-15*:1241–1246
- Moskowitz BM, Halgedahl SL (1987) Theoretical temperature and grain-size dependence of domain state in $x = 0.6$ titanomagnetite. *J Geophys Res* 92:10667–10682
- Moskowitz BM, Halgedahl SL, Lawson CA (1988) Magnetic domains on unpolished and polished surfaces of titanium-rich titanomagnetite. *J Geophys Res* 93:3372–3386
- Muxworthy AR, Williams W (1999) Micromagnetic calculations of hysteresis as a function of temperature in pseudo-single domain magnetite. *Geophys Res Lett* 26:1065–1068
- Muxworthy AR, Williams W (2006) Observations of viscous magnetization in multidomain magnetite. *J Geophys Res* 111:B01103. <https://doi.org/10.1029/2005JB003902>
- Newell AJ, Dunlop DJ, Williams W (1993) A twodimensional micromagnetic model of magnetization and fields in magnetite. *J Geophys Res* 98:9533–9549
- Ozdemir O, Dunlop DJ (1993) Magnetic domain structures on a natural single crystal of magnetite. *Geophys Res Lett* 20:1835–1838
- Ozdemir O, Dunlop DJ (1997) Effect of crystal defects and internal stress on the domain structure and magnetic properties of magnetite. *J Geophys Res* 102:20211–20224
- Ozdemir O, Dunlop DJ (2006) Magnetic domain observations on magnetite crystals in biotite and hornblende grains. *J Geophys Res* 111: B06103. <https://doi.org/10.1029/2005JB004090>
- Ozdemir O, Xu S, Dunlop DJ (1995) Closure domains in magnetite. *J Geophys Res* 100:2193–2209
- Pokhil TG, Moskowitz BM (1996) Magnetic force microscope study of domain wall structures in magnetite. *J Appl Phys* 79:6064–6066
- Pokhil TG, Moskowitz BM (1997) Magnetic domains and domain walls in pseudo-single-domain magnetite studied with magnetic force microscopy. *J Geophys Res* 102:22681–22694
- Proksch RB, Foss S, Dahlberg ED (1994) High resolution magnetic force microscopy of domain wall fine structures. *IEEE Trans Magn* 30:4467–4472
- Rhodes P, Rowlands G (1954) Demagnetizing energies of uniformly magnetised rectangular blocks. *Proc Leeds Philos Liter Soc Sci Sect 6*:191–210
- Sahu A, Moskowitz BM (1995) Thermal dependence of magnetocrystalline anisotropy and magnetostriction constants of single crystal $\text{Fe}_{2.4}\text{Ti}_{0.6}\text{O}_4$. *Geophys Res Lett* 22:449–452
- Smith PPK (1980) The application of Lorentz electron microscopy to the study of rock magnetism. *Inst Phys Conf Ser* 52:125–128
- Soffel H (1971) The single-domain-multidomain transition in natural intermediate titanomagnetites. *Z Geophys* 37:451–470
- Soffel HC (1977a) Domain structure of titanomagnetites and its variation with temperature. *J Geomagn Geoelectr* 29:277–284
- Soffel H (1977b) Pseudo-single-domain effects and single-domain multidomain transition in natural pyrrhotite deduced from domain structure observations. *J Geophys* 42:351–359
- Soffel HC, Aumüller C, Hoffmann V, Appel E (1990) Three-dimensional domain observations of magnetite and titanomagnetites using the dried colloid SEM method. *Phys Earth Planet Inter* 65:43–53
- Stacey FD, Banerjee SK (1974) The physical principles of rock magnetism. Elsevier, Amsterdam, p 195
- Szymczak R (1968) The magnetic structure of ferromagnetic materials of uniaxial structure. *Electron Technol* 1:5–43
- Williams W, Dunlop DJ (1989) Three-dimensional micromagnetic modelling of ferromagnetic domain structure. *Nature* 337:634–637
- Williams W, Dunlop DJ (1990) Some effects of grain shape and varying external magnetic fields on the magnetic structure of small grains of magnetite. *Phys Earth Planet Inter* 65:1–14
- Williams W, Dunlop DJ (1995) Simulation of magnetic hysteresis in pseudo-single-domain grains of magnetite. *J Geophys Res* 100:3859–3871
- Williams W, Wright TM (1998) High-resolution micromagnetic models of fine grains of magnetite. *J Geophys Res* 103:30537–30550
- Williams W, Hoffmann V, Heider F, Goddenhenrich T, Heiden C (1992) Magnetic force microscopy imaging of domain walls in magnetite. *Geophys J Int* 111:417–423
- Worm H-U, Markert H (1987) The preparation of dispersed titanomagnetite particles by the glass-ceramic method. *Phys Earth Planet Inter* 46:263–270
- Worm H-U, Ryan PJ, Banerjee SK (1991) Domain size, closure domains, and the importance of magnetostriction in magnetite. *Earth Planet Sci Lett* 102:71–78
- Xu S, Dunlop DJ, Newell AJ (1994) Micromagnetic modelling of two-dimensional domain structures in magnetite. *J Geophys Res* 99:9035–9044
- Ye J, Halgedahl SL (2000) Theoretical effects of mechanical grain-size reduction on GEM domain states in pyrrhotite. *Earth Planet Sci Lett* 178:73–85
- Ye J, Merrill RT (1995) The use of renormalization group theory to explain the large variation of domain states observed in titanomagnetites and implications for paleo-magnetism. *J Geophys Res* 100:17899–17907

Magnetic Gradiometry

Harald von der Osten-Woldenburg
National Heritage Department, Regional Government of
Baden-Wuerttemberg, Esslingen am Neckar, Germany

Definition

Magnetometer	Instrument with a single sensor that measures magnetic flux density.
Magnetic gradiometer	Pairs of magnetometers with sensors separated by a fixed distance.

The Earth's magnetic field at any point on or near the Earth's surface is the vector sum of the contributions from the primary field due to the dynamo in the Earth's liquid core and the crustal field from the magnetic mineral content of local rocks. This vector has both an orientation and an amplitude. Slight changes in any direction influence the orientation and amplitude. Introduce a highly magnetic rock formation into an otherwise homogenous host, and the local magnetic vector will change. In a three-dimensional world, there are nine (3×3) spatial gradients forming a tensor which defines the anomalous field.

On the Conceptual Design of Gradiometers

A gradiometer is a special measuring instrument that consists of more than one (usually two, rarely three or four) magnetometers. The differences between the readings of the sensors are seen as an approximation to the gradient of the magnetic field along their alignment. The (usually) two magnetometers of a gradiometer are arranged either vertically or horizontally. The distance of the two sensors can vary between several decimeters (for archaeological or environmental mapping) and several kilometers (for research of the deeper-lying geological structures of the Earth's crust, see ► [“Magnetic Methods, Airborne”](#)) depending on the field of application.

Magnetometers

Quantum magnetometers (see also ► [“Magnetometers”](#)) and saturable-core magnetometers are generally suitable for use in gradiometers. Alongside the fluxgate saturable-core magnetometers, cesium vapor magnetometers have become a standard in many areas (aeromagnetic survey, marine geophysics, environmental geophysics, and archaeological prospection) due to the improved noise signal levels against the

Overhauser magnetometers or the helium-cooled sensors (Nabighan and Asten 2002).

Thanks to new technological advances, for example, in the further development of magnetic gradiometer systems based on the superconducting gradiometric sensors SQUID (superconducting quantum interference device), systems are now available that enable very precise measurements of extremely small magnetic field changes. Due to this development, the importance of magnetic gradient tensor and field vector measurements has increased significantly in many applications (Clark 2012). Thus, a significantly higher magnetic and spatial resolution of magnetic maps is achieved than with conventional magnetometers.

Fluxgate gradiometers measure a single component of the Earth's magnetic field (usually the vertical component), while gradiometers consisting of quantum magnetometers (i.e., proton precession and cesium vapor sensors) measure the component of an anomaly in the direction of the Earth's ambient magnetic field, as this is much larger than the anomaly itself (Blakely 1996). Additionally, fluxgate magnetometers are directionally sensitive and subject to a temperature-related drift, and therefore often have to be recalibrated during measurements.

Instead of using a pair of different magnetic field sensors and differentiating their outputs in order to derive a magnetic gradient value, a single sensitive element can be used for measuring directly a magnetic gradient. This is done by the use of a stiff metallic string clamped at both ends and pumped with an AC current, the frequency of which is tuned to the second eigenmode of the string. The string is excited at that eigenmode in the presence of a quasi-static magnetic gradient. Then, the corresponding mechanical displacements of the string can be measured by the use of an inductive technique with an instrumental noise envelope of less than $\pm 10^{-12}$ m, per 1 s measurement interval (Veryaskin 2001). Sheng et al. (2017) describe a new magnetic gradiometer based on optical spectroscopy of alkali atoms in the vapor phase. This type of magnetic gradiometer can be extremely interesting for applications in archeology and environmental sciences.

In the field of environmental geophysics, there has been a trend towards the use of multi-channel gradiometer systems in recent years (Fig. 1). On a trailer, usually 10 or up to 20 probes are pulled over the measuring surface at a distance of 0.50 m (or 0.25 m if 20 probes are used) by an off-road vehicle. The georeferencing is done in real time by means of a differential GPS system.

Advantages and Limitations

In measurements with a gradiometer, the measured signal amplitude of the magnetic anomaly of an interfering body decreases by the power of three or four depending on the distance between the two sensors. A gradiometer is therefore



Magnetic Gradiometry, Fig. 1 The 10-channel fluxgate gradiometer system, manufactured by Eastern Atlas GmbH, Berlin, using the Fluxgate sensors of the Institute Dr. Foerster, Reutlingen, Germany. Magnetic field data are georeferenced in real time using a differential GPS system

somewhat less sensitive to deeper-lying features than a total-field magnetometer, where the signal amplitude of an anomaly decreases by the power of three depending on the distance of the sensor to the interfering body. However, gradiometers react very sensitively to interfering features close to the surface and possess a higher spatial resolution in small-scale structures than total-field magnetometers. The variability in time of the Earth's magnetic field and other sources of interference that are far enough away from the gradiometer, and whose influence on the sensors is equal or at least of a similar level during gradiometer measurements, can be eliminated – or at least significantly reduced – by subtracting the simultaneous readings of the two sensors from one another. An advantage of gradient measurements, arising from the ambiguity present in all potential field measurements, is the suppression of broad regional changes in the magnitude of the magnetic field – the long wavelength component. With gradiometer measurements, the local variations are enhanced, making small and weakly magnetic targets recognizable. Gradiometer surveys are particularly useful in areas that are geologically complex. Due to the resulting difference, the magnitude of the anomalies is, however, smaller than the amplitude that one would achieve with a total-field

magnetometer. This means that a total-field magnetometer can detect anomalies of smaller and deeper-lying features which cannot be detected by a gradiometer. Fluxgate gradiometers are very sensitive to their position in relation to the magnetic vector field: an ever so slight change in the inclination of the instrument during measurements can lead to incorrect results (Schmidt and Clark 2000).

The use of vertical gradiometers for magnetic mapping effectively removes the time variations in the magnetic field map without having to establish base stations, networks, or tie lines and is the only insurance against elimination of the high-frequency component that contains the small or weakly magnetic targets. Gradiometers have shown that they can offer a high degree of immunity from diurnal and minor magnetic storm activity in the ambient magnetic field; they can enhance near-surface, small, or weakly magnetic anomalies.

Areas of Application

Vertical axial gradiometers are used for measuring the vertical differences (the vertical gradient) in the vertical component of

Magnetic Gradiometry,

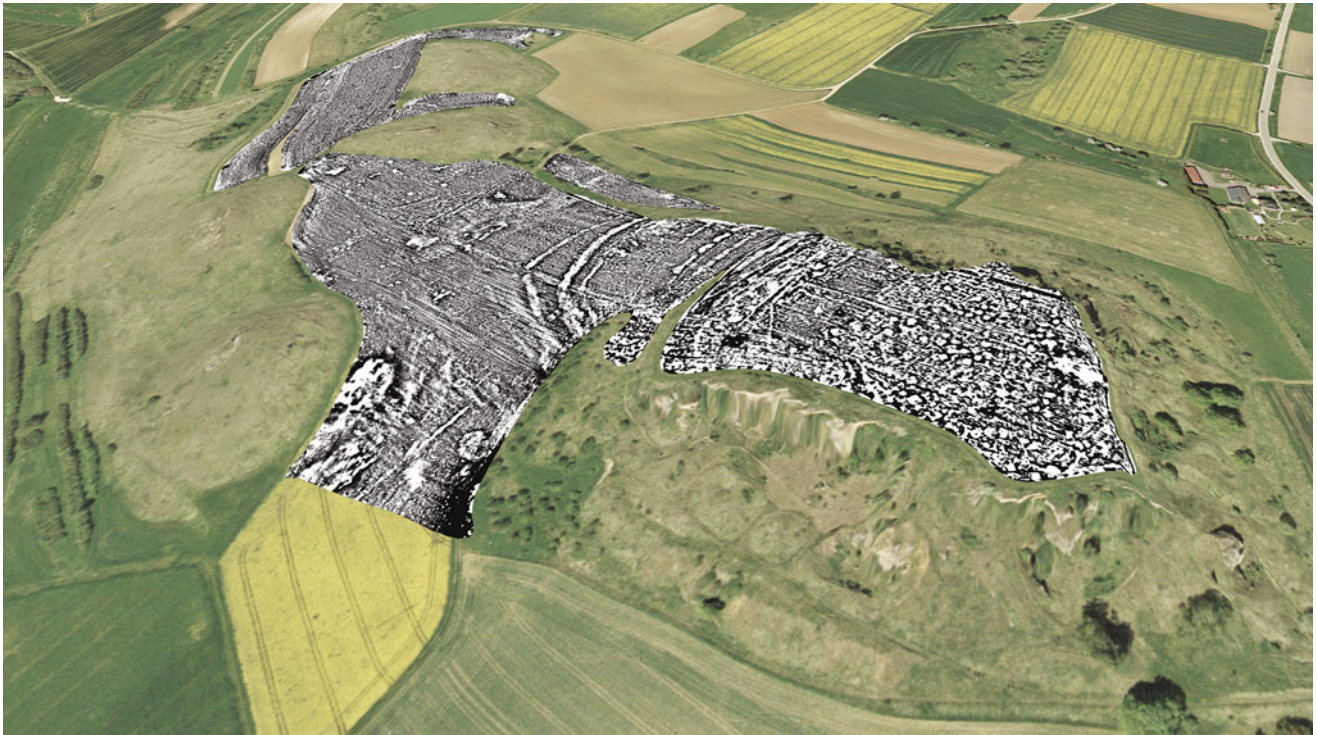
Fig. 2 The four-channel fluxgate gradiometer of the Institute Dr. Foerster, Reutlingen, Germany. Measurements in a grid of 0.05 by 0.06 m are possible with this device. The two fluxgate sensors are 0.65 m apart



the Earth's magnetic field. This configuration is commonly used for ground-based geophysical surveys in the field of environmental geophysics, in areas with complex geology and archaeological prospecting, as the influences of deeper-lying interferences can be suppressed by this setup of the sensors. For practical reasons, the distance between the sensors of a gradiometer used for ground-based surveys lies somewhere between 0.5 and 1 m. As a rule, the ground penetration and detection ranges to about one to two meters below the surface. Deeper detection can be achieved if the interfering body is accordingly large and strongly magnetized. Fluxgate gradiometers are commonly found in environmental geophysics and in archaeological prospecting (Fig. 2). Surveys can be conducted at faster rates, and the highest requirements for spatial resolution of the smallest archaeologically relevant features can be met with measurement grids of 0.05 by 0.06 m without the fluxgate sensors influencing each other. Sometimes, these very small measuring intervals are necessary in archaeological prospecting in order to achieve the best possible spatial resolution of archaeological features (von der Osten-Woldenburg et al. 2006). The magnetic gradiometer map of a Roman fort and its adjacent settlement (Fig. 3) that was obtained with a measurement grid of 0.25 by 0.25 m shows a large number of diverse, often very characteristic anomalies which allow conclusions to be drawn on the position, shape, and size of defensive trenches, centrally located staff buildings with stone foundations, wooden barracks with their hearths, as well as buildings (stone-foundation buildings, street layouts, cellars, and pits) of the settlement adjacent to the fort.

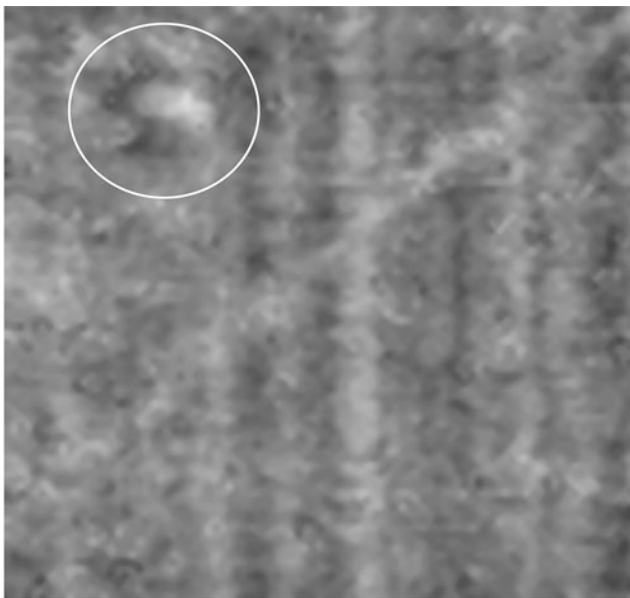
Magnetic gradiometry is also employed in forensic geophysics, even if the radar method has been established as the standard procedure (Cheetham 2005; Hansen and Pringle 2013). However, certain general conditions such as a sufficient distance to the iron cores of reinforced concrete, fences, and other metal constructions must be provided, and it must be ensured that there is no metal refuse in the ground itself. The procedure is identical to that of archaeological prospecting: research is undertaken in small-scale, near-surface survey as the burials themselves are usually close to the surface. The human body only has a small measure of magnetic susceptibility, so a direct detection of the remains by magnetic gradiometry is not possible. But the excavation of a grave and the disruption of strata leads to disturbances of the layers of soil close to the surface that are usually magnetically homogenous, and the refilling of the grave with the excavated material – in which the magnetic particles are now statistically aligned – usually generates weak geomagnetic anomalies, often smaller than 2–3 nT (Fig. 4). The displacement of soil with a significantly higher susceptibility by the human body can intensify this effect.

Axial gradiometers are also used in aeromagnetic surveys (see ► “Magnetic Methods, Airborne”), that is, by helicopter for the detection of ordnance remains close to the surface (Doll et al. 2006), in the prospection of mineral deposits (see ► “Magnetic Methods, Surface”) and for research of anomalies in the Earth's magnetic field and deeper-lying geological structures of the Earth's crust (see ► “Magnetovariation Studies”) from stratospheric balloons with a base length of sensors as far as 6 km apart (Webers et al. 2009).



Magnetic Gradiometry, Fig. 3 Archaeological survey: Magnetic gradiometer map of a neolithic (5000 BC) to the La-Tène-period (sixth/seventh century BC) sequence of settlements on the mountain Goldberg, part of the Nördlinger Ries, Germany. On the plateau of the mountain

floor plans of the houses are clearly visible. The defensive structures in the west of this settlement date back to 4500 BC. West-East extension of the magnetogram: about 650 m



Magnetic Gradiometry, Fig. 4 Forensic geophysics: Detection of a burial at the edge of a field (marked by a circle). The linear anomalies are produced by drainage pipes and geological structures in the substratum

Transverse and longitudinal gradient magnetic measurements are applied in aeromagnetic surveys (Hogg 2004), in marine geophysics, and space exploration (see ► “[Magnetic Methods, Principles](#)”). The distance of the magnetometers from one to the other can range from 1.5 m (e.g., in cable and pipeline survey, or for detecting smaller metallic objects) to more than 500 m (for the detection of magnetic sediments close to the surface, shipwrecks, and the prospecting of mineral deposits), depending on the required area of application.

Magnetic gradient tensor data are used, for example, for a better determination of the small-scale structure of the Earth’s lithospheric field (Kotsiaros and Olsen 2012). They can also be used to clean up areas which are contaminated with various types of unexploded ordonances (UXO) and landmines (Zuo et al. 2017).

Summary

Magnetics belong to the most commonly used methods for prospecting in applied geophysics when attempting to infer the position, geometry, and physical parameters of a body of interference or geological structure through the measurement of magnetic anomalies (Schmidt and Clark 2006).

Gradiometers with two vertically arranged magnetometers that are positioned between 0.5 and 1 m apart, and which measure the vertical gradient and vertical component of the Earth's magnetic field, have distinct advantages over total-field magnetometers, especially in ground-based magnetic measurements in environmental geophysics (surveying of contaminated sites, examination of building construction sites, archaeological research of former settlements): they offer a high resolution in smaller features which is, however, gained at the cost of a lesser penetration of the ground in comparison to measurements made with total-field magnetometers. Additionally, they are very sensitive to close-by sources of interference. Background interference (such as the variability in time of the Earth's magnetic field and other nearby sources of interference) can be eliminated or significantly reduced by subtracting the simultaneous readings of the two sensors in the gradiometer arrangement.

Cross-References

- [Magnetic Gradiometry in Archaeo-geophysics](#)
- [Magnetic Methods, Airborne](#)
- [Magnetic Methods, Principles](#)
- [Magnetic Methods, Surface](#)
- [Magnetometers](#)
- [Magnetovariation Studies](#)

Bibliography

- Blakely RJ (1996) Potential theory in gravity and magnetic applications. Cambridge University Press, Cambridge
- Cheetham P (2005) Forensic geophysical survey. In: Hunter J, Cox M (eds) Forensic archaeology: advances in theory and practice. Routledge, London/New York, pp 62–92
- Clark DA (2012) New methods for interpretation of magnetic vector and gradient tensor data I: eigenvector analysis and the normalized source strength. *Explor Geophys* 43:267–282. <https://doi.org/10.1071/EG12020>
- Doll WE, Gamey TJ, Beard LP, Bell DT (2006) Airborne vertical magnetic gradient for near-surface applications. *Lead Edge* 25:50–53
- Hansen JD, Pringle JK (2013) Comparison of magnetic, electrical and ground penetrating radar surveys to detect buried forensic objects in semi-urban and domestic patio environments. In: Pirrie D, Ruffeli A, Dawson LA (eds) Environmental and criminal geoforensics. Special publications, vol 384. Geological Society, London. <https://doi.org/10.1144/SP384.13>
- Hogg S (2004) Practicalities, pitfalls and new developments in airborne magnetic gradiometry. *First Break* 22:59–66
- Kotsiaros S, Olsen N (2012) The geomagnetic field gradient tensor. *Int J Geomath* 3:297–314. <https://doi.org/10.1007/s13137-012-0041-6>
- Nabighan MN, Asten MW (2002) Metalliferous mining geophysics – state of the art in the last decade of the 20th century and the beginning of the new millennium. *Geophysics* 67:964–978
- Schmidt PW, Clark DA (2000) Advantages of measuring the magnetic gradient tensor. *Preview* 85:26–30
- Schmidt PW, Clark DA (2006) The magnetic gradient tensor: its properties and uses in source characterization. *Lead Edge* 25:75–78
- Sheng J, Wan S, Sun Y, Dou R, Guo Y, Wie K, He K, Qin J, Gao J-H (2017) Magnetoencephalography with a Cs-based high-sensitivity compact atomic magnetometer. *Rev Sci Instrum* 88:094304
- Veryaskin AV (2001) Magnetic gradiometry: a new method for magnetic gradient measurements. *Sensors Actuators A Phys* 91(1–2):233–235
- von der Osten-Woldenburg H, Chaume B, Reinhard W (2006) New archaeological discoveries through magnetic gradiometry: the early Celtic settlement on Mont Lassois, France. *Lead Edge* 25:46–48
- Webers W, Tsvetkov Yu, Brekhov O, Kraoivny A, Nikolaev A, Filippov S, Pchelkin A (2009) Complex satellite (“Swarm”) and stratospheric balloons geomagnetic researches. In: ESA's second swarm international science meeting, Potsdam, 2009
- Zuo B, Wang L, Chen W (2017) Full tensor eigenvector analysis on airborne magnetic gradiometer data for the detection of dipole-like magnetic sources. *Sensors* 17:1976. <https://doi.org/10.3390/s17091976>

Magnetic Gradiometry in Archaeo-geophysics

Harald von der Osten-Woldenburg

National Heritage Department, Regional Government of Baden-Wuerttemberg, Esslingen am Neckar, Germany

Definition

Geophysical measurement procedures have become an important and integral part of surveying methods in archaeological disciplines. Unlike hard prospecting, where archaeological remains are analyzed during an excavation which ultimately leads to their destruction, geophysical measurement procedures are commonly referred to as soft prospecting. They are nondestructive and leave the ground undisturbed. Geophysical methods are therefore often used in the targeted preparation of rescue or research excavations (von der Osten-Woldenburg 2005). They can save time and money by helping to define the location and size, apparent state of preservation, and probable age of the surveyed archaeological remains. These measurement procedures are also widely used when documenting the sites of archaeological settlements: sites that can be preserved for coming generations as they remain undisturbed in the ground.

Magnetic gradiometry is one of the most important methods when it comes to choosing from the range of geophysical measurement procedures that are suitable for archaeological surveys, and it is normally used as the initial prospecting technique. The reasons for this can be found in the usually very conclusive results, quick mapping of the site with a high level of data density, simple usage of the respective measuring instruments, and the relatively low

requirements for data processing. The reasons for the existence of magnetic and therefore archaeologically relevant anomalies may include: buried artefacts containing ferromagnetic metals; long-term changes in magnetic properties as a result of the effects of fire, for example in the case of fires; changes in the distribution of magnetic properties in the ground due to the excavation of pits or defence trenches (Ernenwein and Hargrave 2009).

Indications on the presence of yet undiscovered buried settlement sites can be provided by scanning the earth's magnetic field in close proximity to the ground and recording the weakly defined, archaeologically caused geomagnetic anomalies. This geomagnetic mapping is an important tool for the discovery of archaeological settlement traces in unmapped terrain where there are no stray material finds, no indications in historical sources and no aerial photography findings. Additional methods that complement geomagnetic gradiometry – such as electrical resistivity tomography or ground penetrating radar – must be employed where necessary to verify the subareas in which the contrast of the magnetic susceptibilities, or the residual magnetization between the archaeological objects and the soil in which they are embedded, is faint and therefore appears in the magnetogram only as areas of weakly defined anomalies.

Advantages of Magnetic Gradiometers

Gradiometers used in archaeological prospecting normally consist of an array of two sensors that are aligned at a vertical separation of 0.5 m, more commonly 0.65 m, and in some instances just over 1 m. They can detect magnetic anomalies that are weak, just beneath the surface, or of a limited size – all typical circumstances for archaeological objects – with more clarity than conventional measurements that are taken of the total intensity of the earth's magnetic field in ground proximity while using only a single sensor. They also have a much greater spatial resolution than the single-sensor, total-field measuring magnetometer (GEMSystems 2019). For that reason alone, the gradiometer is the ideal instrument for detecting smaller anomalies that lie closer to the surface and extremely practical when it comes to archaeological mapping.

It has also been shown that gradiometers have a high degree of insusceptibility toward the daily smaller magnetic storm activities that occur in the earth's magnetic field (GEMSystems 2019). Measurements with an additional base station magnetometer that eliminates the variations in the earth's magnetic field from the recorded data are therefore unnecessary.

How deep down a gradiometer can detect archaeological structures depends on the one hand on the vertical separation between the two sensors, and on the other on the intensity of the contrast of the magnetic susceptibility, or the residual

magnetization between the archaeological object and the soil in which it lies. Usually depths of 1–2 m can be reached under ideal conditions.

Instruments

Magnetic gradiometry is carried out by using magnetometers that possess extremely high susceptibility and stability. A number of magnetometer sensors and configurations are available. Cesium vapor sensors or fluxgate sensors, used in pairs in a gradiometer configuration, are a common choice. In practice there are almost no differences in the capacity between the two types of sensor when they are used in a gradiometer configuration, apart from a higher sensitivity of cesium vapor sensors over the fluxgate variant. Cost and weight factors are the main reasons that fluxgate gradiometers are the preferred choice in archaeological field surveys.

Standard gradiometers measure the vertical gradient of the earth's magnetic field locally and in close proximity to the ground. This can be overlaid by faintly magnetic subsurface archaeological structures from former settlement sites. Composite maps, so-called magnetograms, which contain the magnetic image of subsurface archaeological structures, are then created from the measured variations of this magnetic gradient.

As a rule, fluxgate gradiometry is carried out using a portable single or double gradiometer, or with up to four gradiometers that are mounted to a nonmetallic cart or carrying structure (Fig. 1). More recently, a growing number of vehicle-based measuring configurations have been used (Linford et al. 2015; Schneidhofer et al. 2016) that can carry up to 10 (Fig. 2) or even 20 fluxgate gradiometers (Eastern Atlas 2019; SENSYS Sensorik, & Systemtechnology 2019). These sensor carts are towed behind an off-road vehicle or an ATV and record the earth's magnetic field in close proximity to the ground along a profile that can be up to five meters wide while the separation between the gradiometers can range from 0.25 m to 0.5 m.

Procedure in the Field

Maps resulting from surveys that are made with an adequately high spatial resolution can provide quite recognizable and detailed contours of the buried structures. This allows a clear archaeological interpretation to be made. The necessary spatial resolution is, of course, dependent on the size of the structures that are under investigation. But as the actual size is usually not something that is known before the survey, the highest possible resolution must be chosen from the outset. Experience has shown that measurements that are taken with a maximum separation of 0.25 m along the profile line and at



Magnetic Gradiometry in Archaeo-geophysics, Fig. 1 Using a four-channel fluxgate gradiometer (FEREX 4.032 by Institut Dr Foerster, Germany) on the site of the Lauchheim Roman fort (Fig. 5). (Photo Credit: Harald von der Osten-Woldenburg)



Magnetic Gradiometry in Archaeo-geophysics, Fig. 2 A geomagnetic survey of up to 25 ha per day can be achieved – subject to soil conditions – with a ten-channel fluxgate gradiometer array (LEA D2 by Eastern Atlas, Germany) and a differential GPS system. (Photo Credit: Harald von der Osten-Woldenburg)

an interval of 0.25 m are absolutely necessary to create and improve the composite image and therefore to address and interpret smaller magnetic anomalies.

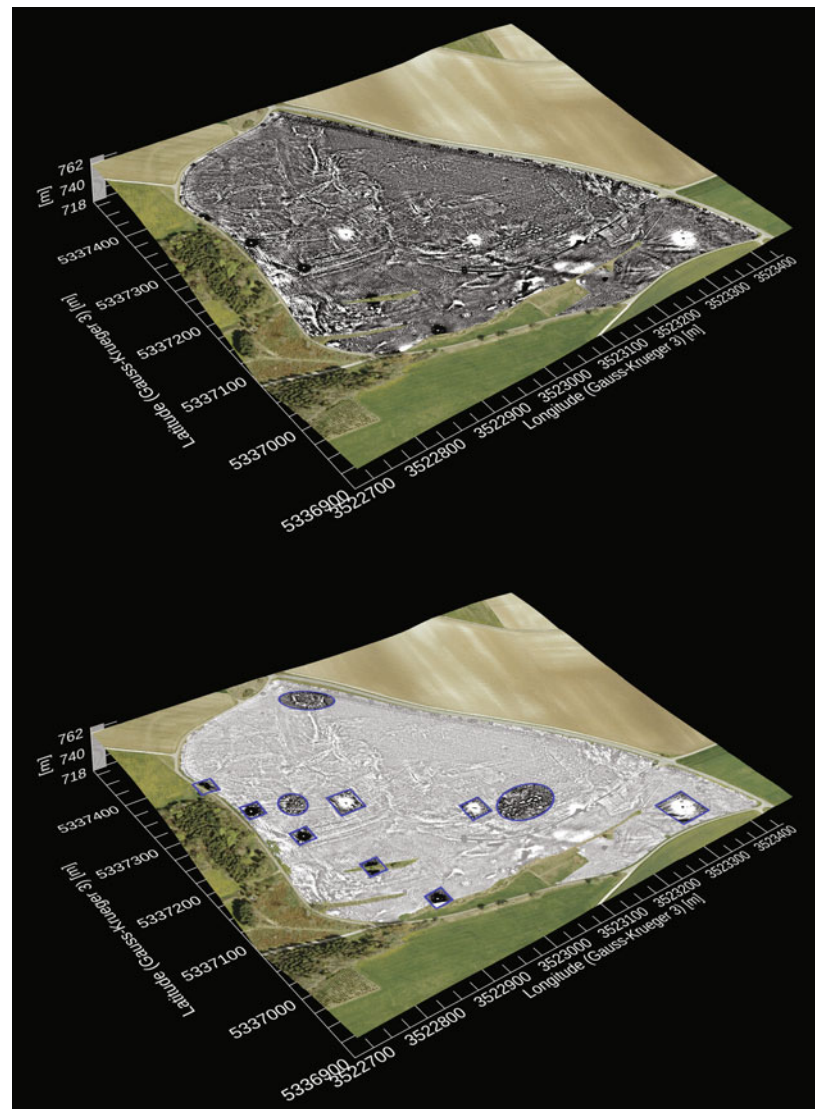
Surveys with gradiometer arrays, such as the single channel (viz. an instrument consisting of one gradiometer) FM36 or FM256 (both with a vertical sensor separation of 0.5 m) by Geoscan Research, Bradford/UK (Geoscan Research 2019), the two-channel Grad601 (vertical sensor separation: 1 m) by Bartington, Witney/UK (Bartington 2019), the FM256 Dual Gradiometer System by Geoscan Research (horizontal separation between the two gradiometers: 1 m), or the four-channel FEREX 4.034 by Institut Dr Foerster, Reutlingen/Germany (Foerster 2019), are usually performed within a regular grid and without the use of a GPS receiver attached to the gradiometer. In order to accomplish this, several grids must first be pegged with sizes that can range – depending on the instrument that will be used – from 20 m × 20 m to 40 m × 40 m. Guidelines are strung within these grids that

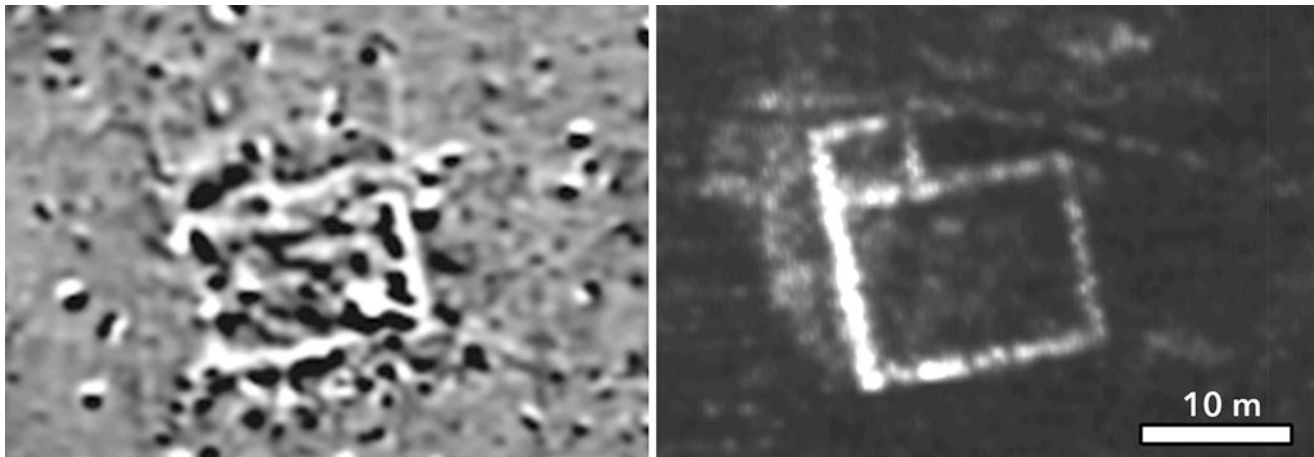
are one or two meters apart and allow the precise following of the parallel-aligned profiles. Gradiometers equipped with an audio signal (metronome) in combination with respective markings along the lines make it easier to follow a predefined speed when taking the measurements, which makes the magnetic recording of the buried structures as distortion-free as possible.

Grids are not necessary in cases where a multichannel measurement system with at least ten fluxgate gradiometers is towed behind an off-road vehicle or ATV. Moreover, these surveys can also be seamlessly performed in irregularly shaped areas. This is made possible by using an RTK-DGPS system: The base station of a differential GPS system is set up at the edge of the survey area. The GPS-rover station is firmly attached to the measurement cart (Fig. 2). Such an array enables – depending on the soil conditions (e.g., meadow and ploughed field) – the mapping of relatively large areas of up to 25 ha per day with a high measurement density. In

Magnetic Gradiometry in Archaeo-geophysics,

Fig. 3 Discovery of a prehistoric settlement site in Emerfeld (district of Biberach, Germany) with the ten-channel fluxgate gradiometer array LEA D2. (Photo Credit: Harald von der Osten-Woldenburg)





Magnetic Gradiometry in Archaeo-geophysics, Fig. 4 The example of a Roman building: Archaeological structures discovered with the help of magnetic gradiometry can be verified in more detail through

additional geophysical surveying methods (e.g., ground penetrating radar). (Photo Credit: Harald von der Osten-Woldenburg)

addition, there is also the advantage that the georeferencing of the measurements, which takes place at the same time as the gradiometer survey, allows for on-site processing of the data within minutes and the creation of the magnetogram. This provides important information already during the measurement process that can help to decide if the survey area needs to be extended or not.

Case Studies

Magnetic gradiometry is virtually predestined for the discovery of former settlement sites in until then archaeologically unidentified locations. The availability of multichannel gradiometers has enabled the geomagnetic surveys of larger suspected sites to be completed within a couple of hours.

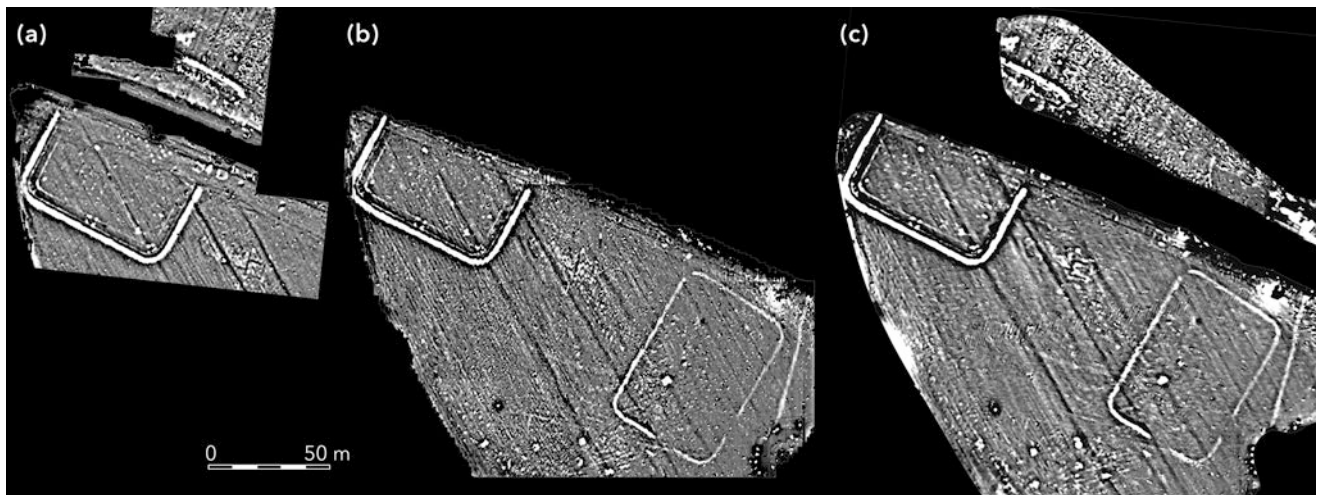
In the case of Emerfeld (in the district of Biberach, Germany), there were no indications whatsoever from either stray material finds, historical sources, or aerial archaeology. Measurements that were taken in a single day resulted in at least three locations being recorded that indicate toward the presence of a prehistoric settlement. They are marked in Fig. 3 by a circle and two ellipses. These are the characteristic geomagnetic anomalies associated with pit houses. The magnetogram also contains information on traces of ploughing (fine, parallel aligned lines), geological structures such as erosion gullies, and modern disturbances such as the location of utility poles for the power lines (marked as bright anomalies in squares) as well as a disused power line (the dark anomalies in squares).

There are situations in which gradiometers may only provide an indication that archeologically relevant structures might be present. A more detailed documentation through the targeted use of other methods becomes necessary in

cases where the initial discovery and localization is made in this manner. One such example is the discovery of a building from the Roman period (Fig. 4, left) that was recorded in just 2 min during the course of a geomagnetic gradiometry survey with a vehicle-based ten-channel fluxgate gradiometer. In a subsequent ground penetrating radar survey (SIR-4000 by GSSI, USA, 400 MHz antenna, 0.5 m profile interval), the floor plan became much clearer (Fig. 4, right). The expenditure of time for this latter survey, excluding the pegging of the grid, was higher by a factor of 10. Conversely, the initial use of the ground penetrating radar in this area would have been highly inefficient and the survey could have taken several days before the building would have been documented.

Do Different Measurement Instruments Deliver Different Results?

During road works, the existence of a small, until then unknown Roman fort, was recorded in a magnetic gradiometer survey (Fig. 5a). The survey was done in October 1992, using a British Geoscan Research FM36. Profile separation and measurement-point interval were both 0.25 m. The outer ditch, foundations of the watchtowers, the fort's main road, and post holes belonging to the fort's wooden buildings are clearly recognizable in the magnetogram, as well as a Roman estate to the east of the fort and a medieval gallows in the northern part of the fort. In July 2009, the fort, which is located in an area of meadows and therefore lies protected in the ground, was reexamined with magnetic gradiometry (Fig. 5b). This time a four-channel German fluxgate gradiometer system FEREX 4.032 by Institut Dr Foerster was used (Fig. 1). The system was equipped with CON650 sensors (vertical separation between sensors: 0.65 m). The interval



Magnetic Gradiometry in Archaeo-geophysics, Fig. 5 Comparison of surveys with different gradiometers in the example of a Roman earth and timber fort in Lauchheim, Ostalb district, Germany. Left: FM36, Geoscan Research, UK. Centre: FEREX 4.032, Institut Dr Foerster,

Germany. Right: LEA D2, Eastern Atlas, Germany. Multichannel gradiometers allow for larger areas to be surveyed in less time. The single-channel gradiometer FM36 provided more detailed information on this small fort. (Photo Credit: Harald von der Osten-Woldenburg)

between the measuring points was 0.05 m, the profile separation 0.25 m. The data was resampled based on a grid of $0.25 \text{ m} \times 0.25 \text{ m}$. The advantage of this instrument lies in the much higher prospecting speed over that of the FM36. This allowed a geomagnetic mapping of the entire meadows, and during the survey, a further Roman ditch system was discovered, the use of which is so far unknown (horse paddock?). A third survey of the site was undertaken in March 2015, using a ten-channel German Eastern Atlas LEA D2 system (Fig. 2). The sensors used in this survey were the same as those in the four-channel FEREX-system (CON650). The separation between the sensors was 0.5 m, and measurements were taken on average at an interval of 3 cm and a speed of 10 km/h. This data was also resampled based on a grid of $0.25 \text{ m} \times 0.25 \text{ m}$ (Fig. 5c).

Conclusion

Even if a comparison between the FM36 sensors and the Foerster-CON650 sensors is not strictly allowed – since the separation of the fluxgate sensors of the FM35 is 0.5 m, and that of the CON650 is 0.65 m – it still shows that the measurements taken by foot were slightly more exact and that they contain more detail than the multichannel instrument surveys. The vehicle-based ten-channel Foerster gradiometer measurements, for example, do not show the fort's eastern road, whereas they are clearly recognizable in the FM36 survey.

However, the benefits of vehicle-based surveys – covering larger areas in less time – probably outweigh the disadvantages. Only for the actual documentation of suspected

archaeological objects would it be advisable to resort to targeted resurveying by foot.

Cross-References

- [Electromagnetic Pulsations and Magnetic Storms](#)
- [Magnetic Gradiometry](#)
- [Magnetic Methods, Surface](#)
- [Magnetometers](#)

Bibliography

- Bartington (2019). https://www.bartington.com/wp-content/uploads/pdfs/datasheets/Grad-601_DS1800.pdf
- Eastern Atlas (2019). http://www.eastern-atlas.com/start/index_eng.php
- Ernenwein EG, Hargrave ML (2009) Archaeological geophysics for DoD field use: a guide for new and novice users report submitted to the environmental security technology certification program for project RC-200611: streamlined archaeo-geophysical data processing and integration for DoD field use. Center for Advanced Spatial Technologies, University of Arkansas, Fayetteville. <https://wayback.archive-it.org/6471/20150825215604/http://cast.uark.edu/assets/files/PDF/ArchaeologicalGeophysicsforDoDFieldUse.pdf>. Accessed 5 Mar 2014
- Foerster (2019). <https://www.fluxgate-magnetometer.com/de/produkte/ferex-4034/>
- GEMSystems (2019). <http://www.gemsys.ca/advantages-of-magnetic-gradiometers/>
- Geoscan Research (2019). <http://www.geoscan-research.co.uk>
- Linford N, Linford P, Payne A (2015) Chasing aeroplanes: developing a vehicle-towed caesium magnetometer array to complement aerial photography over three recently surveyed sites in the UK. Near Surf Geophys 13(6):623–631. <https://doi.org/10.3997/1873-0604.2015044>

- Schneidhofer P, Nau E, Hinterleitner A, Lugmayr A, Bill J, Gansum T, Paasche K, Seren S, Neubauer W, Draganits E, Trinks I (2016) Palaeoenvironmental analysis of large-scale, high-resolution GPR and magnetometry data sets: the Viking Age site of Gokstad in Norway. *Archaeol Anthropol Sci* 9(6):1187–1213. <https://doi.org/10.1007/s12520-015-0312-x>
- SENSYS, Sensorik & Systemtechnology GmbH (2019). <https://sensysmagnetometer.com>
- Von der Osten-Woldenburg H (2005) Applications of ground penetrating radar, magnetic and electrical mapping, and electromagnetic induction methods in archaeological investigations. In: Butler DK (ed), *Near-surface geophysics: SEG, Society of Exploration Geophysicists*, Tulsa, pp 621–626. <https://doi.org/10.1190/1.9781560801719.ch23>

Magnetic Methods, Airborne

Mike Dentith

School of Earth Sciences (M004), The University of Western Australia, Crawley, WA, Australia

Synonyms

Aeromagnetics

Definition

Airborne magnetic surveys comprise measurements of the Earth's magnetic field made with a magnetometer mounted on an aircraft.

Introduction

Making measurements from the air of the Earth's magnetic field is a well-established geophysical survey method. The normal practice is to measure the scalar amplitude of the magnetic field in the survey area, this being referred to as the total magnetic intensity (TMI), the name reflecting the fact that the measurements represent the resultant of all the magnetic fields, of whatever source, in the vicinity of the measurement. The major contributions come from the geomagnetic field and the fields due to magnetic rocks in the upper crust, the latter being the fields of interest.

Initially used to infer the depth to magnetic rocks beneath nonmagnetic cover in sedimentary basin studies and for identification of major lineaments, the products from aeromagnetic surveys have evolved to now be a reliable means of creating detailed maps of the near-surface geology. Current developments in aeromagnetism are mostly related to the increasingly sophisticated inference of the 3D distribution of

magnetism in the Earth's crust using inverse modeling methods.

Comprehensive accounts of aeromagnetic data, acquisition, processing, and interpretation are provided by Gunn (1997), Reeves (2005), Isles and Rankin (2013), and Dentith and Mudge (2014).

Making Magnetic Measurements from the Air

Airborne magnetic data are acquired as a series of parallel “survey” or “flight” lines and more widely spaced orthogonal “tie” lines (Fig. 1). The purpose of the tie lines is to create repeat measurements at the locations where they cross the flight lines, which are required during processing of the data (see below).

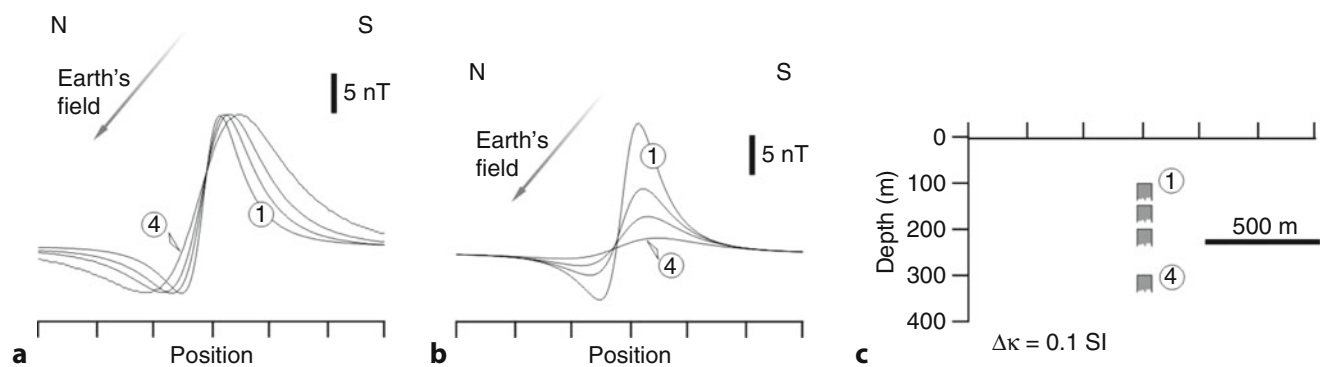
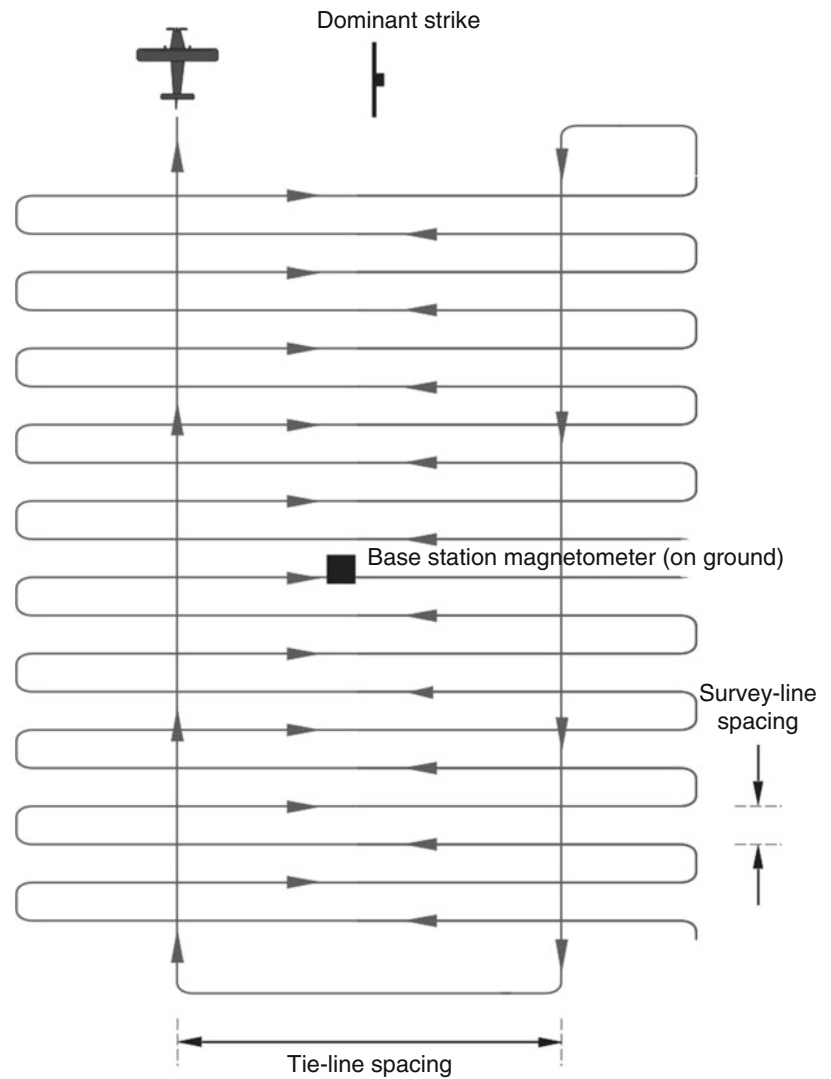
An advantage of making magnetic measurements from the air is that unless the survey area is small, typically less than a few square kilometers, it is much faster and cheaper to acquire the data than for a ground-based survey. The main disadvantage is that turbulence and variations in topography, especially if the terrain is rugged, may seriously affect the ability to position the aircraft as would be preferred and also affect the individual measurements in a hard to predict way.

Equipment

Airborne measurements are routinely made using fixed-wing aircraft and helicopters. Surveys using fixed-wing aircraft are significantly cheaper than helicopter surveys, but helicopters have the advantage of being able to fly lower and better maintain a consistent ground clearance in rugged terrain, but have the disadvantage of being less stable survey platforms.

To reduce effects of magnetic fields originating in the aircraft the magnetic sensor, usually a cesium vapor magnetometer (see ► “Magnetometers”), is located as far as possible from the main body of the aircraft and hence the magnetic fields associated with ferrous components and electric currents. For fixed-wing aircraft the magnetic sensor is mounted in a “stinger” located aft of the aircraft. In helicopters the sensor may be in a “boom” that positions the magnetometer in front of the aircraft or it may be within a towed “bird.” The bird is a bomb-like device suspended on a cable beneath the aircraft. A “compensation” system within the aircraft is used to further reduce the effects of magnetic fields originating from the aircraft.

In addition to carrying one or more magnetometers, the survey aircraft also measures terrain clearance using an altimeter and GPS position and time. One or more “base station” magnetometers are also deployed on the ground to measure temporal changes in the magnetic field in the survey area to enable such variations to be removed from the data (see below).

Magnetic Methods, Airborne,**Fig. 1** Schematic illustration of the flying pattern used in aeromagnetic surveys**Magnetic Methods, Airborne, Fig. 2** Profile of total magnetic intensity across a dyke-like magnetic body. The four profiles correspond with the bodies at different depths in (c). The profiles in (b) represent

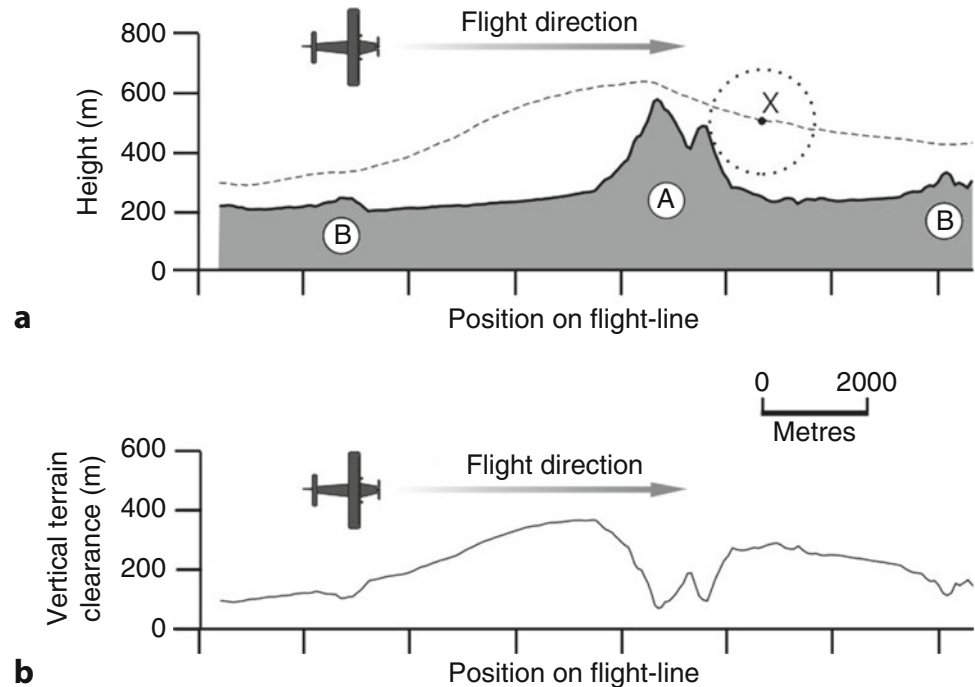
variations in TMI. In part (a) the amplitudes have been normalized to the profile with maximum amplitude (shallowest source)

The Importance of Source-Sensor Separation

An important source of noise in aeromagnetic data is associated with changes in distance between the magnetic sensor

and the magnetic materials. Any changes in this distance will create variations in TMI, which could easily be confused with those caused by changes in the local geology. As the distance

Magnetic Methods, Airborne,
Fig. 3 Example of the flight path
 of a fixed-wing survey aircraft
 across rugged topography. (a)
 Topographic and altimetry profile,
 (b) vertical terrain clearance.
 (Based on data in Flis and Cowan
 2000)



increases the measured magnetic field variation decreases in amplitude and increases in wavelength (Fig. 2b). The wavelength increase is most easily seen when the amplitude decrease is removed by rescaling the individual profiles so they have the same peak-to-peak amplitude (Fig. 2a).

Ideally the survey aircraft maintains a constant ground clearance during the acquisition of the data, referred to as “drape” flying. In practice, aircraft are subject to the effects of wind and turbulence and the inaccuracies of human control. In addition, all aircraft have a limited ability to maneuver, especially climb. Figure 3 shows the flight path of a fixed-wing aircraft across a ridge and adjacent areas. When short wavelength variations in topography (B on the diagram) are encountered, the aircraft cannot respond quickly enough to maintain the separation and positive relief will bring the magnetic sources closer to the sensor, with a corresponding increase in the amplitude of the magnetic field (c.f. Fig. 2), and vice versa. When large topographic features are encountered (A on the diagram) the aircraft will inevitably vary its ground clearance because of its limited ability to climb and to a lesser extent descend. If, as is the normal case, alternate flight lines are flown in opposite directions, adjacent parts of survey lines will have markedly different ground clearance adjacent to topographic features. This will introduce spurious variations into the data.

Current practice is to “program” a preplanned flight path for the aircraft, which the pilot follows using a real-time display. This program accounts for the terrain and the aircraft’s performance. The aim is to achieve the best “loose drape,” whereby terrain clearance effects are

minimized as far as is practical. Note that even if a perfect drape was achieved, the effects of terrain would not be totally removed because even in the ideal scenario there are still the effects of variations in the distances to magnetic materials because it is not just the materials directly below the aircraft that are influencing the measurements. The measurement at X in Fig. 3 will be affected more by the magnetic materials to the side than beneath because of the lesser distance to these materials, although this scenario only occurs in severe terrain.

Survey Design

Sampling theory requires individual measurements of TMI to be spaced at a maximum of half the wavelength of the shortest wavelength of variation. Economic and safety factors mean this is rarely practical and some aliasing of responses is the norm. This is not necessarily a major problem since during qualitative interpretation of the data it is relative changes in amplitude and texture that are used, and in fact accurate definition of the variations is only required when specific anomalies need to be analyzed quantitatively (see ► “[Magnetic Anomalies: Interpretation](#)”) and often there is a follow-up, more detailed, survey of the area of interest to improve anomaly characterization.

Survey Specifications

The survey geometry, illustrated in Fig. 1, requires definition of three key parameters: the survey line spacing and

Magnetic Methods, Airborne, Table 1 Typical survey specifications for aeromagnetic surveys

Survey line spacing (m)	Tie-line spacing (m)	Flight height (m)	Sampling ratio: Survey/Tie line ^a	Shortest along-strike wavelength sufficiently sampled
50	500	40	7.5	100
100	1000	50	15	200
200	2000	60	30	400
400	4000	80	60	800

^aAssumed flight speed is 240 km/h (67 m/s) and 10 Hz sampling resulting in a 6.7 m along survey line sampling interval. These parameters are typical for a survey with a fixed-wing aircraft

orientation and the flight height. Typical values of two of these variables for different survey types are given in Table 1. Note that the tie-line spacing is a dependent variable, typically being set to ten times the survey line spacing, although this may be reduced to five times or less for high resolution surveys.

The survey line spacing controls the cost of the survey, which for fixed-wing aircraft is based on the total length of lines flown. The total line length of a survey in terms of the survey area and line spacing can be estimated from the equation below (Brodie 2002):

$$\text{Total line length} = \frac{1000 \left(1 + \frac{\Delta_{\text{Survey lines}}}{\Delta_{\text{Tie lines}}} \right) \text{Survey area}}{\Delta_{\text{Survey lines}}}$$

where $\Delta_{\text{Survey lines}}$ and $\Delta_{\text{Tie lines}}$ are the survey- and tie-line spacings in meters, respectively, the total line length is in kilometers, and the survey area is in square kilometers.

To this cost must be added nonproduction costs such as mobilization and “stand by” costs associated with factors outside the acquisition company’s control, for example, bad weather, magnetic storms. With helicopter surveys the time spent in the air is also taken into account, and may be significant in mountainous areas where weather conditions severely restrict data acquisition both in terms of flying time and location (Mudge 1996).

It is clear from Fig. 2 that the lower the flight height, the greater will be the amplitude of the targeted responses and also the shorter the wavelength of variation, and hence the greater the ability to resolve adjacent features. Figure 2 implies the lowest possible flight height should be used, but safety considerations mean flight heights are some tens of meters. “Crop-duster” fixed-wing aircraft or heli-borne surveys may fly lower but ultra-low flying is not necessarily an advantage from a technical point of view since it means maintaining consistent terrain clearance is more difficult. Besides, there may be very shallow magnetic sources whose anomalies mask those of deeper, more significant variations (Doyle and Lindeman 1985).

Reid (1980) showed that accurate definition of a TMI anomaly requires the sample spacing to be a maximum of half the source-sensor separation. The combination of the

rate at which the aircraft’s magnetometers make a measurement (typically 0.1 s) and the velocity of the aircraft (typically 220–280 km/h for a fixed-wing aircraft) result in an along flight line sample spacing of about 7 m for fixed-wing aircraft. Helicopter surveys with a 0.1 s sampling typically have a 4–5 m spacing, but the trend is to employ faster sampling magnetometers, reducing this spacing to even less. The important outcome from the above is that the along-line sampling exceeds the requirements of sampling theory for practical flight heights, and the variations in TMI in the direction of the flight path can be considered as more than adequately characterized. The same is not true for variations in the directions perpendicular to the survey line direction because, as shown in Table 1, the spacing of these lines is very much greater than the along-line sampling interval. Note that since anomaly wavelength decreases with source-sensor separation, a low flight height combined with wide survey lines will lead to greater aliasing in the direction perpendicular to the survey lines. This is reflected in the parallel increase in flight height with line spacing in Table 1.

Conventionally airborne geophysical surveys are flown with the flight lines oriented perpendicular to the dominant geological strike direction. The rationale for this is that there will be shorter wavelength variations across strike than parallel to strike, and hence the greatest sampling density should be in the strike-perpendicular direction. For elongate anomalies the short wavelength variation parallel to strike is small, so significant aliasing does not occur. In practice economic considerations dictate the survey line spacing, meaning there is inevitably some aliasing of TMI variations in the direction perpendicular to the survey lines.

Even if there is some variation in the strike across the survey area, and there is also a conflict caused by variations in stratigraphic and structural strike, it is best to use a single-line orientation. If this is not the case the data are very difficult to process into an easily interpreted form.

For surveys in equatorial regions the east–west elongation of magnetic responses associated with low declinations (<30°) means it is always preferable to fly surveys with north–south survey lines.

Data Processing

The term data processing refers here to the removal of noise prior to enhancement (see ► [“Magnetic Data Enhancements and Depth Estimation”](#)) and interpretation of the data (see ► [“Magnetic Anomalies: Interpretation”](#)). The aim of the processing is to create an instantaneous “snap shot” of TMI variations across the survey area with the contribution from the geomagnetic field removed. It is normal practice to complete at least preliminary processing in the field to ensure that the data are of acceptable quality.

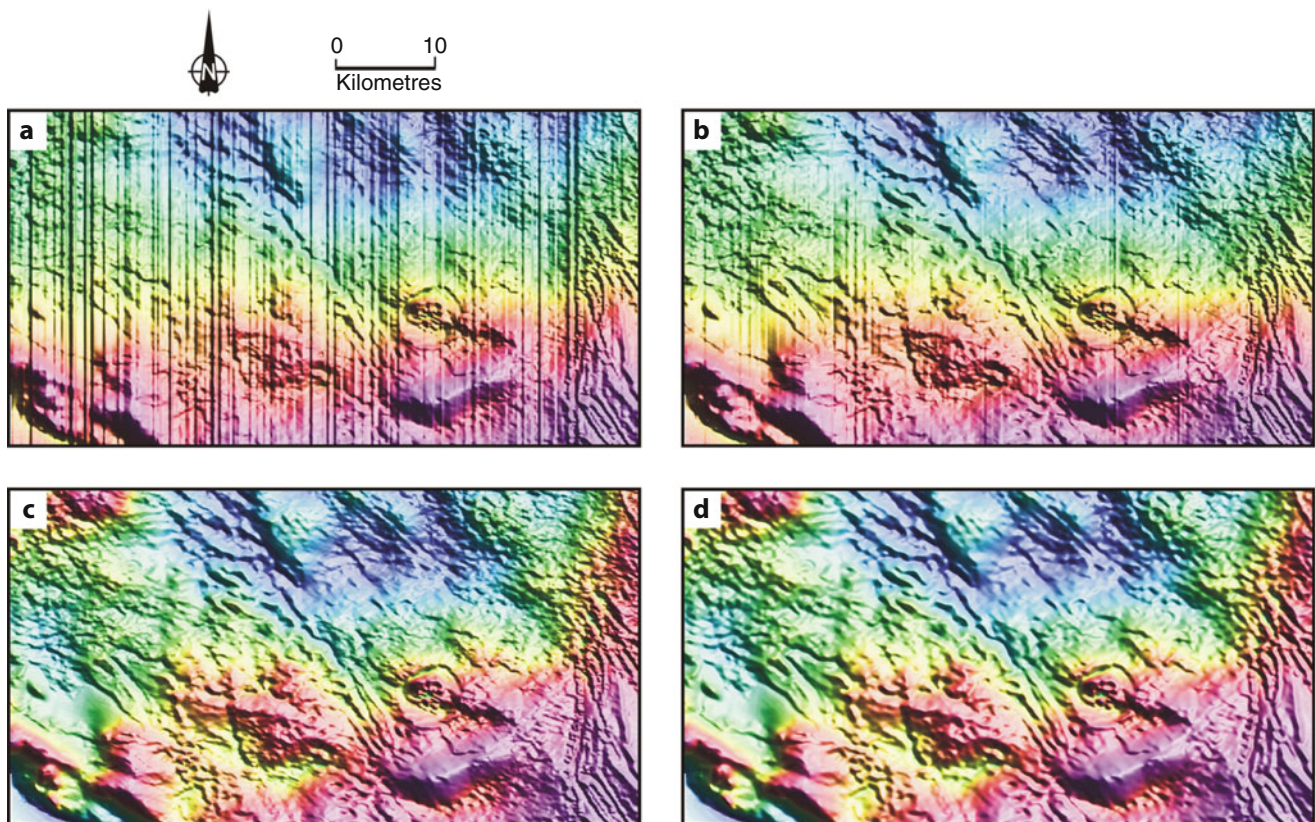
The main source of noise in aeromagnetic data is temporal changes in the Earth’s magnetic field, which occur during the course of the survey and hence from the mobile sensor’s perspective are indistinguishable from spatial variations in TMI due to the geology of the survey area. It is to determine such variations that the repeat measurements inherent in the survey/tie-line geometry are required (Fig. 1), along with the TMI time series recorded by each base station.

The removal of the noise is achieved in a series of steps. Figure 4 shows a raster display of an aeromagnetic dataset

after each of the processing steps described below. Figure 4a shows the data after compensation for aircraft-related effects. Note the dominant characteristic is a north–south striping (corrugations) with a wavelength approximately equal to twice the line spacing; this being primarily the result of the temporal changes in the geomagnetic field during the survey and deviations of the aircraft from the ideal flight path.

Correction for Temporal Changes in the Geomagnetic Field

The Earth’s magnetic field varies in intensity at a range of time scales; see ► [“Geomagnetic Field, Theory”](#). Of significance to aeromagnetic surveys are magnetic storms (during which surveying must be discontinued) and micropulsation and diurnal variations (see ► [“Electromagnetic Pulsations and Magnetic Storms”](#)). Micropulsations are variations in the geomagnetic field that last from a few seconds to a few minutes. They change in both phase and amplitude over distances comparable to survey dimensions. Diurnal variations have periods of 24 h and magnitudes of up to about 50 nT.



Magnetic Methods, Airborne, Fig. 4 Gridded TMI data from an aeromagnetic survey in Australia. The colors represent changes in amplitude (hot colors high, cold colors low) and the intensity represents a shaded relief presentation designed to highlight north–south trending features. (a) Data after compensation for magnetic fields originating in

the aircraft. (b) As for (a) but with correction for temporal variations in the geomagnetic field as measured by a base station magnetometer. (c) As for (b) but with international geomagnetic reference field removed and tie-line leveling corrections applied. (d) As for (c) except for micro-leveling. Images created by Cowan Geodata Services

The principal means of removing micropulsation and diurnal variations is the base station magnetometer readings (the base station readings also allow the recognition of the occurrence of a magnetic storm). The readings in the aircraft and base station are time synchronized. In theory the variation at the (stationary) base station should represent geomagnetic field variations that occur during the course of the survey, and subtracting the base station readings from the time equivalent survey reading should leave a residual dataset free of the temporal variations. However, this is an imperfect process since the amplitude and phases of the changes vary with location. To counter this multiple base stations may be used for large surveys. The subtraction causes a base line shift in the data but this is unimportant since it is relative variations in TMI that are required for interpretation of the data (see ► [“Magnetic Anomalies: Interpretation”](#)). Often an average intensity of the magnetic field in the survey area is added to each measurement to return the data to closer to likely absolute values. Figure 4b shows the magnetic dataset after corrections for temporal changes in the geomagnetic field derived from a base station dataset. The degree of striping in the data is significantly reduced, although it is not eliminated.

Removal of Geomagnetic Field

It is normal practice to remove spatial variations in TMI caused by the geomagnetic field as defined by the International Geomagnetic Reference Field (IGRF), see ► [“Geomagnetic Field, IGRF”](#). The wavelengths of variation of the field are very large compared to survey dimensions, and subtraction from the observed data will not usually significantly affect variations of interest. Nevertheless it is standard practice to remove such variations.

Tie-Line Leveling

Ideally the processing described above will mean the TMI measured along the survey- and tie-lines at their cross-over points will be identical. In practice, there will always be mismatches because of the imperfect aircraft compensation, spatial variations in diurnal and micropulsation-related changes, but also because there are uncertainties in the aircraft's position, that is, where the cross-over occurs within the datasets. Differences in elevation between the survey- and tie-line measurements are the main source of error.

The minimization of the differences at the cross-over points is referred to as “leveling.” Various approaches are used but normally there is an initial estimate of the extent to which location errors are responsible for the differences. Once these are accounted for the remaining differences at each line intersections are used to adjust the amplitudes

along the survey lines. The most common method involves fitting a smooth function to the differences and using this to estimate leveling “corrections” at each measurement point along the survey lines. Figure 4c shows the magnetic data after removal of the IGRF (minimal effect) and tie-line based leveling. Corrugations in the data are now all but invisible.

Microleveling

Microleveling is a loosely defined term for making adjustments to the data when the processing stream outlined above is complete and artifacts appear to remain in the data. The details of these methods are mostly proprietary but are all various forms of low-pass filtering. The filter characteristics are often defined by the flight and survey line spacing, for example, Minty (1991). Microleveling is usually applied after the data have been gridded (see below), and the correction so derived used to adjust the individual survey measurements prior to re-gridding. Figure 4d shows the example magnetic data after application of a commercial microleveling process. There is some improvement in the lateral continuity of anomalies.

Gridding

By far the most common form of display and input to enhancement oriented processing operations are data projected on to a regular grid. Gridding the data is a compromise between honoring detailed sampling along the survey lines and producing smooth interpolation into areas between lines where no readings have been made. The grid cell size, that is, distance between interpolated locations, is conventionally between 1/3 and 1/5 of the survey line spacing. The loss of information in the survey line direction is balanced by an acceptable degree of reliability of interpolated points located between the survey lines. Nevertheless, especially with greater survey line spacing, aliasing frequently causes laterally continuous features to appear as a series of adjacent subcircular features, referred to as “beads.” This effect becomes worse as the trend of the anomalies approaches that of the survey lines.

Discussion of Data Processing

The processing described above is a combination of processes with a basis in field measurements and processes requiring the judgment of an expert processor, because it is not practical to collect the necessary data to process the data perfectly. There is no way to reliably judge the accuracy of the final measurements, but an accuracy of ± 1 nT is often quoted for individual points in the gridded datasets.

Summary

Mapping variations in TMI allows variations in rock magnetism in the upper crust to be mapped, from which geology may be inferred. Measurements of TMI made by sensors mounted on helicopter or fixed wing crafts allow rapid and efficient acquisition of data. The main sources of noise in magnetic data from airborne platforms are temporal changes in the geomagnetic field and variations in the distance between magnetic sources and the magnetometer caused by changes in aircraft ground clearance. Temporal changes are accounted for by comparing survey data with readings contemporaneously recorded by a magnetometer on the ground. Discrepancies in repeat readings at the same location allow any remaining temporal change noise to be removed along with corrections for variations in survey altitude. The resulting data are projected on to a regular grid for display as raster images.

Cross-References

- ▶ [Electromagnetic Pulsations and Magnetic Storms](#)
- ▶ [Geomagnetic Field, IGRF](#)
- ▶ [Geomagnetic Field, Theory](#)
- ▶ [Magnetic Anomalies: Interpretation](#)
- ▶ [Magnetic Data Enhancements and Depth Estimation](#)
- ▶ [Magnetic Methods, Principles](#)
- ▶ [Magnetometers](#)

Bibliography

- Brodie RC (2002) Airborne and ground magnetics. In: Papp E (ed) Geophysical and remote sensing for regolith exploration. CRCLEME Open file report 144. CRCLEME, Bentley, Western Australia, pp 33–45
- Dentith M, Mudge ST (2014) Geophysics for the mineral exploration geoscientist. Cambridge University Press, Cambridge, p 438
- Doyle HA, Lindeman FW (1985) The effect of deep weathering on geophysical exploration in Australia – a review. *J Geol Soc Aust* 32:125–135
- Flis MF, Cowan DR (2000) Aeromagnetic drape corrections applied to the turner syncline, Hamersley Basin. *Explor Geophys* 31:84–88
- Gunn P (1997) Airborne magnetic and radiometric surveys. *AGSO J Aust Geol Geophys* 17(2):216
- Isles DJ, Rankin LR (2013) Geological interpretation of aeromagnetic data. Australian Society of Exploration Geophysicists, Melbourne, Australia, 357pp
- Minty BRS (1991) Simple micro-leveling for aeromagnetic data. *Explor Geophys* 22:591–592
- Mudge S (1996) Helicopter magnetic surveys in rugged terrains: from planning to processing. *Lead Edge* 15:305–308
- Reeves C (2005) Aeromagnetic surveys: principles, practice and interpretation. Geosoft, Toronto, Canada, p 155
- Reid AB (1980) Aeromagnetic survey design. *Geophysics* 45:973–976

Magnetic Methods, Principles

Kusumita Arora

National Geophysical Research Institute, Council of Scientific and Industrial Research (CSIR), Hyderabad, India

Definition

Magnetic Methods, Principles provides an overview of the various approaches, which generate knowledge of the different aspects of the Earth's magnetic field.

Introduction: Magnetic Field on Surface of Earth

Magnetism makes us think of a flat iron needle, which aligns itself in the north-south direction. Or it makes us think of a bar of iron, which is able to attract other pieces of iron. Essentially magnetism is a force, which can attract (or repel) another magnetic object. The property of magnetism as manifested by certain types of rocks was known to the ancient Greeks and Chinese and was used by early travelers for direction finding when stars were not visible.

Systematic study of the Earth's magnetic field and its variations started in the thirteenth century in Europe, largely based on the behavior of the compass needle, which responds to the complete magnetic force at each location of measurement. This was followed by observations at different times of the day, year and so on. The behavior of magnetic needles at different locations on the Earth's surface led to the proposal by William Gilbert around 1600 that the Earth itself was a giant magnet, which was the beginning of studies of the geomagnetic field. More than two centuries later Hans Christian Oersted observed that flow of current in a wire caused a magnetic needle to deflect. Andre Marie Ampere established that magnetism was a force generated between two electric currents; two parallel currents in same direction attract and in opposite direction repel. Later it was demonstrated that properties of bar magnets were very special cases of the same phenomenon: within magnetic materials, bulk magnetism is the result of the motion of electrons in an atom, which sets up a magnetic field around itself.

The first magnetometer was the compass, which was used to measure the direction of the field; an instrument to measure the intensity was developed in 1833 by Gauss. Vector magnetometers, which can measure the intensity of the field in three orthogonal directions came into existence during the nineteenth century. There was several fold increase of sensitivity in measurements of the strength, changes of the three

components, and the absolute directions respectively (more details can be found in article ► “[Magnetometers](#)” in this volume). With data recorded over time, it was observed that the magnitude and direction of the magnetic field changes at different locations of measurements; the magnitude and direction also change at the same location with time, sometimes irreversibly, at other times regaining the original values. It is now known that from local to global, the changes of Earth’s magnetic field cover about 7 orders of magnitude in space and about 17 orders of magnitude in time, from under a second to close to the age of the Earth (Courtillot and Le Mouél 1988).

The Earth’s steady magnetic field is produced by many sources, both above and below the planet’s surface. These include the geomagnetic dynamo in the outer core, crustal magnetization, the ionospheric dynamo, the ring current, the magnetopause current, the tail current, field-aligned currents, and convective, electrojets. Comprehensive information can be found in Mande et al. (2019). Thus, the magnetic field measured at a point on the Earth’s surface is a combination of several components (Amory-Mazaudier 2008):

$$B = B_c + B_e + B_{in} + B_a$$

Where B_c = main/core field; B_e = external field (due to ionospheric and magnetospheric sources); B_{in} = induced field generated by the external field B_e ; B_a = magnetization field (due to the magnetization of the rocks in the lithosphere).

To study each of the above, a variety of observational and corresponding computational techniques are employed. Prevalent directions of research are based on four types of data:

- The time series of variations in scalar intensity and the vector magnitude and directions of three perpendicular components, at different locations on the surface of the Earth.
- The spatial variations of scalar intensities at closely located intervals in land and oceans
- Variations of three components of magnetic (commonly along with electric) field at stations of regular intervals
- Measurements of magnetic memory of rocks of different ages, which preserve signatures of the geomagnetic field, generated 3.45 billion years ago

Major Aspects of Probing the Geomagnetic Field

Core/Main Field and Its Changes

Developments in Observational Methods

First observations of the main field characteristics based on the compass needle can be traced to the concept of declination, which began to take shape from the thirteenth century (Chapman and Bartels 1940), while that of inclination was

formalized in the sixteenth century (Norman 1581). The idea of Earth’s magnetic poles took shape and were formalized by William Gilbert and his student Mark Ridley, who established the relationship of inclination and latitude (Jonkers 2000). Le Nautonier documented the variations of declination with longitude (Mande et al. 2019).

The quality and quantity of measurements of the full field vector at specific locations on the Earth’s surface, evolved through time from early nineteenth century to present day sustained precision measurements of vector and scalar components of the field at fixed locations all over the globe, called *Magnetic Observatories*. Observatory measurement systems consist of continuous measurements (quite commonly at 1 Hz) of the variations of the three components of the magnetic field along with continuous measurements of the scalar measurements (again at 1 Hz) of the total field. Daily/bi-weekly/weekly measurements of the absolute values of inclination and declination are further measured manually. The three datasets are combined together to produce continuous absolute values of the three components of the geomagnetic field at the particular location. Measurements at Observatories are sustained typically over decades and even centuries; the first Observatories were setup in the middle of the nineteenth century and at present we have around 200 networked Observatories (<https://www.intermagnet.org/>), which record the transient and long-term changes of the field. By combining data from observatories all over the world, it is possible to map the Earth’s magnetic field and study its spatiotemporal changes.

The first discoveries of changes of the Main field (called Secular variations) happened during the seventeenth century (Jonkers 2000). Changes of declination with time at a fixed location was noted as early as 1622, while that of inclination was found some years later. Data from Magnetic Observatories showed that all components of the Earth’s magnetic field change with time, from years to millennia. From the early twentieth century networks of *magnetic repeat stations* were established, in addition to Magnetic Observatories. Repeat stations are precisely located points on the Earth’s surface, where the three components of the geomagnetic field are regularly (yearly to once in 5 years) measured to the highest possible accuracy (Newitt et al. 1996). Most countries have their own networks of repeat stations and combined with the data from observatories monitor the annual/decadal/centennial changes of each component of the magnetic field in the region.

European merchant and naval ships, later followed by research vessels sailing through the Atlantic, Pacific, Indian Oceans provided a rich source of magnetic field measurements on the surface (Jackson 1992) (more details can be found in articles ► “[Geomagnetic Field, Measurement Techniques](#)” and ► “[Magnetic Methods, Surface](#)” in this volume). Edmond Halley published the world’s first magnetic

map of the Atlantic Ocean in 1701 showing lines of equal magnetic declination, collected during two sea voyages. From these beginning, global maps of the main geomagnetic field were generated using decades of data from Magnetic Observatories (more details can be found in article ► [“Geomagnetic Field, Global Pattern”](#) in this volume). The last few decades of global measurements by satellite have enabled major leaps in mapping of this field and its short-term and long-term changes; it is now possible to calculate geomagnetic secular variation models of unprecedented spatial resolution without using repeat station measurements (more details can be found in article ► [“Magnetic Methods, Surface”](#) in this volume).

Resulting Knowledge

In 1833, the studies of Carl Friedrich Gauss led to the idea that the geomagnetic field could be approximated by a magnetic dipole, which has its origin inside the Earth. Currently it is tilted at an angle of about 11° with respect to Earth's rotational axis. The magnitude of the Earth's magnetic field at its surface ranges from 25,000 to 65,000 nanoteslas from Equator to Poles. Gauss also demonstrated that the dipolar component was inversely proportional to the square of the distance. Observatory data along with archaeomagnetic measurements indicate that the strength of the dipole field has been decreasing since the last 2000 years and it now is only about half of what it was two millennia ago. At its present rate of decrease, the strength would reach zero in approximately 1600 years or it might start to increase again sometime in the future.

Secular Variation is not constant in time, it also varies from place to place. Analyses of the variations show that most of it can be described in terms of three processes:

- A decrease in the strength of the dipole part of the magnetic field
- A westward drift in the non-dipole part of the magnetic field
- Changes in the non-drifting part of the non-dipole field

First suggested in the early twentieth century (Larmor 1919), it is generally recognized that the Main/Core field is created by the “dynamo action” of convective fluid motions in the Earth's electrically conducting outer core, subjected to rapid rotation. Various theories based on computations of rotating fluid dynamics and magnetohydrodynamics have evolved over past decades, which attempts to explain the nature of the geomagnetic field and its changes; concept of buoyancy gradients (thermal or compositional) is used to drive and sustain the geodynamo (Hollerbach 1996 and references therein). Kono and Roberts (2002) presented an exhaustive comparative analysis of different geodynamo models propounded by researchers; the dominance of the axial field; drift of the field, secular variations, and even polarity reversals have been simulated in most of the models

and broadly match the known data. In recent developments, Blanc et al. (2020) have used thermochemical modeling to suggest that crystallizing magma ocean in the Earth's mantle is a plausible mechanism to sustain an early magnetic field.

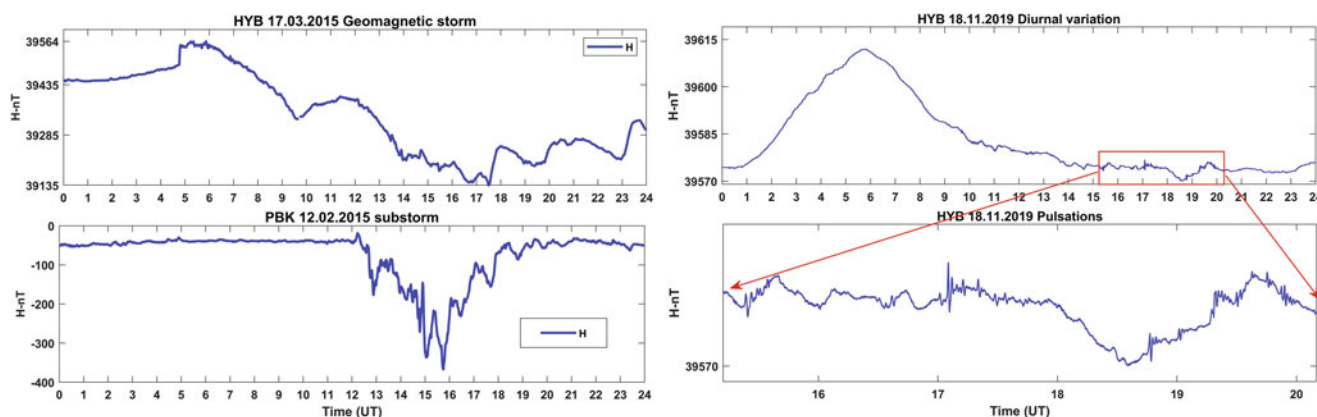
External Field

Developments in Observational Methods

Short term variations of the field, were discovered in the first quarter of the eighteenth century. In 1722, George Graham discovered the diurnal variation of declination. In 1716, Halley connected disturbances of the Earth's magnetic field with the occurrence of an aurora over London, in 1741, Graham in London and Andreas Celsius in Uppsala observed simultaneous perturbations, which would later be identified as the polar aurora (Canton 1759). Superimposed on the regular daily variation of the magnetic field are irregular variations during disturbed magnetic days. These have many different signatures on the magnetic records, also depending on the latitude of the location of observation. Examples of such signatures in daily magnetograms from the low latitude Magnetic Observatory of Hyderabad (HYB) can be found in the figure below. On examination of data from magnetic observatories, it was evident that changes in amplitudes of the geomagnetic field, which recovered to original values could span over a few minutes to 11 and 22 year cycles (Cliver 2015). With the availability of the vast dataset from dedicated satellite missions, it became possible to analyze the global patterns of external field changes (Friis-Christensen et al. 2006; Xiong et al. 2016; Fathy et al. 2019; Kim et al. 2019).

Resulting Knowledge

The Earth's magnetic field, which is the strongest one among the rocky planets, extends 36,000 miles into space, created by the action of the geomagnetic dynamo, without which neither the geomagnetic field nor the other sources would exist. Not far above Earth's surface the effect of other sources becomes as strong as or stronger than that of the Main/Core field. The Earth's magnetosphere is the spatial domain of the magnetic field lines that connect to the Earth and is the region of space where the earth's magnetic field plays a controlling role on various particles and radiations found there. Details of its structure started emerging from the early twentieth century with the advent of satellite observations. The magnetosphere would be a magnetic dipole but is distorted under the influence of magnetized solar wind, compressing it on the dayside. Solar activity controls the changes of the external field through two main kinds of interaction processes: energy transfer between the solar wind and the magnetosphere, and ionization of the E-Layer in the ionosphere (Gavoret et al. 1986). The interactions of the magnetosphere with the electromagnetic forces and emissions from the Sun's magnetic field, produces the ionospheric dynamo, the ring current, the

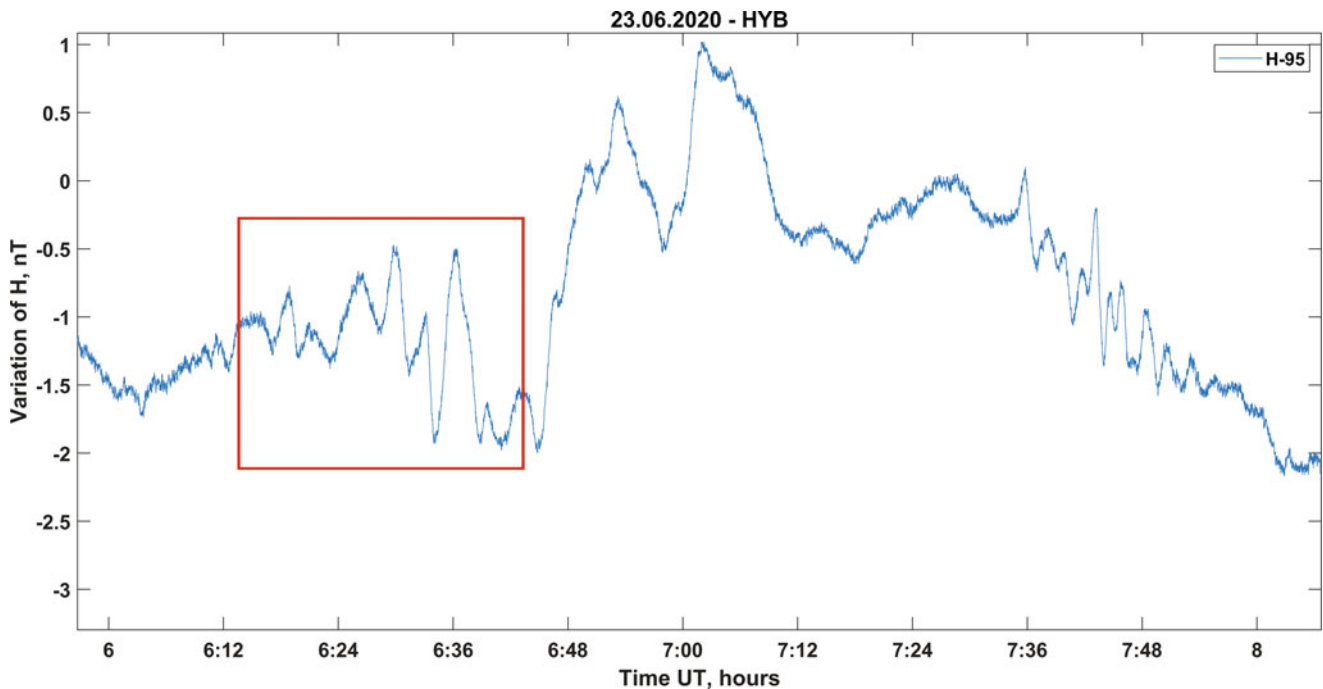


Magnetic Methods, Principles, Fig. 1 HYB daily magnetograms showing signatures irregular variations associated with magnetic storms

magnetopause current, the tail current, field-aligned currents and the electrojets. Each of these current systems are associated with a corresponding magnetic field, which are superposed on the Main field.

Explanations of the observations of the regular variations of the Earth's magnetic field on all the days at the sunlit side of the Earth started in the late nineteenth century. In situ measurements of neutral atmosphere motions and ionospheric electric currents began during the twentieth century, using on the measurements of rockets, meteoric radars, and Incoherent scatter sounders. The regular variations are attributed to the current systems of the ionosphere. These currents are generated by the motion of the atmosphere which drags preferentially the ions through the Earth's magnetic field lines in the E region dynamo ($100 \text{ km} < h < 150 \text{ km}$) and cause a differential drift between ions and electrons. The effect of this daily regular *ionospheric dynamo* is manifested as a diurnal pattern in the daily magnetograms of Observatories (Fig. 1). Electrojet currents, which travel around the E region of the ionosphere from 100–150 km altitudes are Hall currents primarily carried by electrons, which occur in the Equatorial and Auroral regions of the Earth's magnetic field. The auroral electrojets, which have typical amplitudes of 100–1000 nT, may reach upto 3000 nT during intense geomagnetic storms and also shift equatorward (e.g., Feldstein et al. 1997; Ahn et al. 2000). The eastward component of the electrojet flows in the afternoon sector and terminates in the premidnight region where it joins the westward component. The westward electrojet flows through the morning and midnight sectors and typically extends into the evening sector along the poleward border of the auroral oval where it diverges as upward field-aligned currents. The Equatorial Electrojet causes a magnetic field disturbance of the horizontal component of the terrestrial magnetic field at the Earth's surface over a distance of about 600 km across the dip equator with maximum amplitudes of the order of 50–100 nT being superposed on the Sq (Hamid et al. 2014).

Works of many scientists during the twentieth century resulted in the mapping of the magnetospheric currents and understanding of their imprints on the magnetic field (Ganushkina et al. 2018). Signatures of ionospheric and magnetospheric currents on the daily magnetograms of a typical magnetic observatory are shown in Fig. 1; seasonal and solar cycle patterns of such signatures during geomagnetic quiet conditions have been discussed in details by Olsen and Stolle (2017). Energization of the ring current by storms or substorms causes a worldwide depression of the regular daily field over several hours to days, which later recovers (Williams 1985; Zhao et al. 2015) its original trends. Characteristics of field-aligned currents (FAC) associated with substorms were studied from data of Traid and subsequent satellites. During storms and substorms the average latitude width of these currents increases by 20–30% and the centers of these regions shift equatorward by 2° – 3° with respect to the quiet time values, with corresponding reflections in the magnetic records. The FAC is the primary source for the eastward-westward magnetic field perturbations at high latitude (Adhikari et al. 2017). Ultralow frequency magnetic pulsations originate from magnetospheric processes, playing a major role in magnetospheric dynamics and energy transport; they appear as near sinusoidal or irregular fluctuations in the magnetograms at ground recording stations/Observatories, as shown in red box in Fig. 2. Satellite based measurements of the magnetic field perturbations, associated electric field and particle density have allowed study of spatial distribution and propagation mechanisms of pulsations (Anderson 1990). Waves of different frequencies and polarizations originate in different regions of the magnetosphere and are controlled by the solar wind dynamic pressure and magnetic field; changes in orientation of the interplanetary magnetic field or an increase in solar wind velocity have strong influence on the type of waves (McPherron 2005).



Magnetic Methods, Principles, Fig. 2 Pulsation observed at low latitude Observatory

Lithospheric Field

Developments in Observational Methods

The property of magnetism exhibited by some rocks have been known since the time of ancient Greeks and Chinese; compasses were observed to deflect from their original orientation, when placed near such matter. Such rocks found in the crust and upper mantle produce the Lithospheric field. The attraction of compass needles (dipping needles) to natural iron formations and their association with base metal deposits emerged during the nineteenth century, pioneered by efforts in Sweden. By the early twentieth century, it was understood that rocks that have magnetically susceptible minerals, like magnetite, acquires magnetic properties of their own under the influence of the ambient field (the phenomenon is called ferromagnetism). Regions hosting such magnetic rocks/minerals within the crust produce a positive magnetic anomaly, viz., enhanced values compared to that of surroundings. With advances in instrumentation and data acquisition techniques, airborne magnetic measurements were made from a balloon in 1910. There was intensification of efforts for airborne submarine detection during World War II (Reford and Sumner 1964; Hanna 1990), which brought in a new era of airborne magnetic surveys, both for the exploration industry and for countries to map regional geology at national scales (Hanna 1990; Hood 1990). Some of the first aeromagnetic maps were published in 1946. The techniques were soon adapted for marine use. In 1955 the first 2D marine magnetic survey

was conducted off the coast of southern California in the eastern Pacific Ocean (Mason 1958). This marine magnetic survey revealed the pattern of magnetic stripes offset by a fracture zone, which were later attributed to seafloor spreading during periods of geomagnetic reversals (Vine and Matthews 1963; Morley and Larochelle 1964). More details may be found in ► “Magnetic Methods, Surface” in this volume.

The proton precession magnetometer was developed in 1955, followed by the optically pumped and SQUID magnetometers during the late 1990s; later magnetic gradiometers were designed. Parallely major computational techniques were developed to remove the effects of the ambient geomagnetic and solar magnetic fields and isolate and study the local magnetic anomaly maps for the depths and dimensions of the subsurface source bodies (Nabighian et al. 2005). Efforts over the last few decades have been focused on producing a global, near-surface map of the crustal magnetic field of the Earth. This map includes data from airborne, shipborne, terrestrial as well as satellite (MAGSAT, OERSTED, CHAMP, SWARM) measurements (Lesur et al. 2016).

Resulting Knowledge

Anomaly maps depicting crustal magnetization represent the cumulative effects of the earth’s magnetic field from the time of formation of the rocks within the crust, through various stages of metamorphism and finally the present ambient field. As the main field of the earth changes its strength and direction over millions of years, these influences are imprinted

cumulatively on the magnetization patterns of the minerals in the rocks. Metamorphism, petrology and redox state all have important effects on the magnetism of crustal materials. Mapping of the crustal magnetic field provides a third dimension to surface observations of the Earth's composition and geologic structure, aided by the suite of mathematical tools developed to assist with their interpretation. Studies of crustal magnetism have contributed to geodynamic models of the lithosphere, geologic mapping and petroleum and mineral exploration.

The crustal/lithospheric field originates in iron-bearing rocks near the Earth's surface (upper 20–30 km) where temperatures are sufficiently low, i.e., less than about 580 °C (the Curie temperature of magnetite). Hence it provides insight into the distribution of magnetic minerals within the crust, which represent the magnetism at great pressure and temperature for plutonic rocks, the metamorphic environment for altered igneous and sedimentary formations, past geomagnetic fields for volcanic rocks. On a regional scale, magnetic anomalies illustrate terrane boundaries by a contrast in magnetic fabric at the contact. Crustal magnetization consists of two types: induced and remanent (more details available in ► [“Remanent Magnetism”](#) in this volume). Induced magnetization is the signature of the present geomagnetic field; it is proportional to the strength of the inducing field and vanishes when the primary field vanishes. Remanent magnetization is produced by a primary field, but once created it persists after the primary field has disappeared, depending on the concentration of ferromagnetic minerals in the rocks.

Detailed magnetic surveys continue to be a primary tool for locating favorable host rocks for potential mineralized zones, faults, hydrothermal sources and for general geologic mapping. It is frequently used to study basin structure and map basement features below, which may be favorable for hydrocarbon accumulation. Magnetic methods are also used for environmental contamination issues, seismic hazards, landslide hazards, archaeological mapping and so on.

Induced Field

Methods to Study Subsurface Conductivity Structure

The phenomenon of electromagnetic induction was first discovered in the 1830's by Michael Faraday, who observed that when he moved a permanent magnet in and out of a single loop of wire it induced a voltage and a current was produced i.e., by changing the magnetic field an electric current can be generated, the strength of which is proportional to the rate of change of flux. This led to the law linking electricity with magnetism: Faraday's Law of Electromagnetic Induction.

During space weather events, the electric currents in ionosphere and magnetosphere experience rapid fluctuations causing corresponding variations in the magnetic field. The time varying magnetic field induces telluric currents in the

ground to create a secondary magnetic field, which is location dependent; the electrical conductivities of the local rock formations will determine the current strength, which tends to reduce the amplitude of the Z component and enhance the magnitude of the H component. The generation of induced fields is in effect the influence of external fields on lithospheric fields, as most commonly the geological entities having higher magnetism tend also to be more conductive.

Early researchers applied this principle to investigate phenomenon of electromagnetic induction within the Earth, generated by the variability of the geomagnetic field, starting from the second half of nineteenth century. This is called the magnetotelluric (MT) method. The Earth's naturally varying electric and magnetic fields are measured over a wide range of frequencies from 10,000 to 0.0001 Hz. The ratio of the electric field to magnetic field provides simple information about subsurface conductivity. Due to the skin effect, the ratio at higher frequency ranges gives information on the shallow Earth, whereas deeper information is provided by the low frequency range. The ratio is usually represented as both apparent resistivity and phase as functions of frequency. MT measurements can investigate depths from about 300 m down to hundreds of kilometers. The vertical resolution of MT mainly depends on the frequency being measured whereas horizontal resolution of MT mainly depends on the distance between sounding locations.

Geomagnetic depth sounding (GDS) or magnetovariational studies (MVS) are also conducted at large periods of magnetic variation data, whereby different kinds of transfer functions are used to calculate the subsurface conductivity structure. GDS has much longer periods than MVS; the calculations for the former use a spherical Earth model. Large scale conductors at upper mantle depths can be detected, which contribute to the study of tectonics. More will be found in ► [“Magnetotellurics, Crustal Imaging”](#) of this volume.

The oceans play a special role in this induction due to their relatively high conductivity, which leads to large lateral variability in surface conductance. Electric currents that generate secondary fields are induced in the oceans by two different processes: (a) by time varying external magnetic fields, and (b) by the motion of the conducting ocean water through the Earth's main magnetic field. As the electrical conductivity of both seawater and material under the oceans differs considerably from that of the continents, an enhanced effect usually referred to as the “geomagnetic coast effect” (GCE) is observed at data from such locations.

Methods to Study Geomagnetically Induced Currents (GIC) in the Earth's Surface

The same space weather events also induce currents (GIC) in conductors on the surface of the Earth, most commonly electric transmission grids, buried pipelines and cables, railway

tracks. GIC were first observed on the emerging electric telegraph network in 1847–1848 during storm events of Solar cycle 9 (Ronalds 2016). Explosion in various conducting networks have made the significance of GIC greater in modern society. The effect is stronger at higher latitudes and have been studied for necessary mitigation measures in Canadian, Finnish and Scandinavian power grids and pipelines since the 1970s (Thomson et al. 2005). GIC of tens to hundreds of amperes have been recorded. GIC have also been recorded at mid-latitudes during major storms. There may even be a risk to low latitude areas, especially during a storm commencing suddenly because of the high, short-period rate of change of the field that occurs on the day side of the Earth.

Paleo Field

Developments in Observational Methods

It was discovered in the nineteenth century that minerals in ancient rocks can become magnetized parallel to the Earth's magnetic field; this property enabled researchers determine the orientation and intensity of Earth's magnetic field back through time. As rocks form and cool, the electrons within individual grains can shift in the direction of the surrounding magnetic field. Once the rock cools past a certain temperature, known as the Curie temperature, the orientations of the electrons become permanent. Scientists can determine their age and use standard magnetometers to measure their orientation, to derive the strength and orientation of the Earth's magnetic field at the point in time when this characteristic was acquired. Details of this approach are provided in “► [Paleomagnetism, Principles](#)” of this volume.

Direct measurements of magnetic properties of basalts as well as other igneous, metamorphic and sedimentary rocks revealed the differences in their magnetic properties (Faujas de St-Fonds 1788; Fournet 1848). Laboratory experiments of de-magnetizing and remagnetizing materials like iron, led to the concepts of remanent and induced magnetization (Mellon 1853a, b) and also established that direction of magnetization coincides with that of the applied field and not of the crystallographic axes of the rock material (Sidot 1868). Observations of magnetic properties of man-made items like bricks could produce information about magnetic field of more recent times (Folgheraiter 1899), an approach which led to the discipline of Archaeomagnetism.

Resulting Knowledge

Based on measurements on selected quartz crystals from the Greenstone Belt of the Kaapvaal craton of South Africa, it has been inferred that a magnetic field half as strong as today's existed 3.2 billion years ago and a field of strength about 70% of the present existed 3.45 billion years ago (Tarduno et al. 2010). More recently, analyses of zircons from the Jack Hills

(Western Australia) provide evidence for a geomagnetic field 4.2 billion year old, just a few hundred million years after the lunar-forming giant impact (Tarduno et al. 2020), although other scientists still debate this issue.

The geomagnetic field played a vital role in protecting the planet from damaging solar radiation, which may have stripped it of its atmosphere and water, making it completely unsuitable for life as we know it. Knowledge of its changes with time allows a window into the changes of the Earth's core, which in turn tell us about the thermal evolution of the planet. Knowledge of the early field could also inform us about the conditions in which the first life forms appeared.

Early in the twentieth century, Bernard Brunhes (1906) found that some rocks are magnetized in the opposite orientation to the Earth's present-day magnetic field. This led him to propose that the Earth's magnetic field had reversed its polarity in the past. These reversals have subsequently been shown to be non-periodic and the Earth's magnetic field reversal history is now well known for the past 175 million years and more approximately understood to the beginning of the Paleozoic (ca. 545 Ma), based on evidences from rocks, which preserved these signatures. The most recent Matuyama-Brunhes reversal, which marks the Lower-Middle Pleistocene boundary, has been identified in dozens of marine sediment cores, exposed sedimentary successions and a few lava flow sequences (Singer et al. 2019 and references therein). Details can be found in ► “[Paleomagnetism, Measurement Techniques and Instrumentation](#)” of this volume.

Summary

The physical phenomenon of magnetism is one of the oldest subjects of study by humans. This article has tried to provide an overview of the major approaches which investigate the various aspects of Earth's magnetism. They include investigations of the present field in all its aspects: Core/Main field, Lithospheric field, External field, Induced field, as well the studies of the origin and evolution of the field since the time of formation of the Earth. Details of each aspect have been dealt with in other articles of this volume.

Cross-References

- [Geomagnetic Field, Global Pattern](#)
- [Geomagnetic Field, IGRF](#)
- [Geomagnetic Field, Measurement Techniques](#)
- [Geomagnetic Field, Secular Variation](#)
- [Geomagnetic Field, Theory](#)
- [Magnetic Anisotropy](#)
- [Magnetic Data Enhancements and Depth Estimation](#)

- [Magnetic Methods, Surface](#)
- [Magnetometers](#)
- [Paleomagnetism, Measurement Techniques and Instrumentation](#)
- [Paleomagnetism, Principles](#)
- [Remanent Magnetism](#)

Acknowledgments This is contribution numbered NGRI/Lib/2020/Pub-118 from CSIR-National Geophysical Research Institute, Hyderabad, India. Ms L. Manjula has provided the figures.

Bibliography

- Adhikari B, Dahal S, Chapagain NP (2017) Study of field-aligned current (FAC), interplanetary electric field component (E_y), interplanetary magnetic field component (B_z), and northward (x) and eastward (y) components of geomagnetic field during supersubstorm. *Earth Space Sci.* <https://doi.org/10.1002/2017EA000258>
- Ahn B-H, Chen GX, Sun W, Gjerloev JW, Kamide Y, Sigwarth JB, Frank LA (2000) The Lowest Possible Latitude of the Westward Electrojet during Severely Disturbed Periods; Multiscale Coupling of Sun-Earth Processes, Pages 145–156, <https://doi.org/10.1016/B978-044451881-1/50012-5>, Elsevier Publications, ISBN 978-0-444-51881-1
- Amory-Mazaudier C (2008) Sun Earth's system: the transient variations of the Earth's magnetic field. In: Hady A, Wanas MI (eds) First Middle East and Africa IAU-regional meeting proceedings MEARIM No. 1, 2008 © 2008 International Astronomical Union. <https://doi.org/10.1017/97740330200173>
- Anderson BJ, Engebretson MJ, Rounds SP, Zanetti LJ, Potemra TA (1990) A statistical study of Pc 3–5 pulsations observed by the AMPTE/CCE Magnetic Fields Experiment, 1. Occurrence distributions; <https://doi.org/10.1029/JA095iA07p10495>
- Blakely RJ (1995) Potential theory in gravity and magnetic applications. Cambridge University Press, Cambridge
- Blanc NA, Stegman DR, Ziegler LB (2020) Thermal and magnetic evolution of a crystallizing basal magma ocean in Earth's mantle. *Earth Planet Sci Lett* 534:116085. <https://doi.org/10.1016/j.epsl.2020.116085>
- Brunhes B (1906) Recherches sur la direction de l'aimantation des roches volcaniques. *J Phys* 5:705–724
- Canton J (1759) An attempt to account for the regular diurnal variation of the horizontal magnetic needle, and also for its irregular variation at the time of an Aurora Borealis. *R Soc Philos Trans.* <https://doi.org/10.1098/rstl.1759.0040>
- Chapman S, Bartels J (1940) *Geomagnetism*. Oxford University Press, New York. 1049 pp
- Cliwer EW (2015) The extended cycle of solar activity and the Sun's 22-year magnetic cycle. In: Balogh A, Hudson H, Petrovay K, von Steiger R (eds) *The solar activity cycle*. Space sciences series of ISSI, vol 53. Springer, New York
- Courtillot V, Le Mouél JL (1988) Time variations of the Earth's magnetic field: from daily to secular. *Annu Rev Earth Planet Sci* 16:389–476
- Courtillot V, Le Mouél JL (2007) The study of Earth's magnetism (1269–1950): a foundation by Peregrinus and subsequent development of geomagnetism and paleomagnetism. *Rev Geophys.* <https://doi.org/10.1029/2006RG000198>
- Fathy A, Ghamry E, Arora K (2019) Mid and low-latitude ionospheric field-aligned currents derived from the Swarm satellite constellation and their variations with local time, longitude, and season. *Adv Space Res.* <https://doi.org/10.1016/j.asr.2019.07.022>
- Faujas de St-Fonds B (1788) *Essai sur l'Histoire Naturelle des Roches de Trapp*. Rue et Hôtel Serpente, Paris
- Feldstein YI, Grafe A, Gromova LI, Popov VA (1997) Auroral electrojets during geomagnetic storms, *Journal of Geophys. Res Space Phys.* <https://doi.org/10.1029/97JA00577>
- Folgheraiter G (1899) Sur les variations séculaires de l'inclinaison magnétique dans l'antiquité. *J Phys* 8:5–16
- Fournet J (1848) Aperçus sur le magnétisme des minerais et des roches, et sur les causes de quelques anomalies du magnétisme terrestre. *Ann Soc Agric Lyon* 1–53
- Friis-Christensen E, Lühr H, Hulot G (2006) SWARM: a constellation to study the Earth's magnetic field. *Earth Planets Space* 58(4):351–358. <https://doi.org/10.1186/BF03351933>
- Ganushkina NY, Liemohn MW, Dubyagin S (2018) Current systems in the Earth's magnetosphere. *Rev Geophys.* <https://doi.org/10.1002/2017RG000590>
- Gavoret J, Gilbert D, Menvielle M, Le Moué! JL (1986) Long-term variations of the external and internal components of the Earth's magnetic field. *J Geophys Res* 91:4787–4796
- Hamid NSA, Liu H, Uozumi T, Yumoto K, Bhaskara V, Yoshikawa A, Sanchez JA (2014) Relationship between the equatorial electrojet and global Sq currents at the dip equator region. *Earth Planets Space* 66:146
- Hanna WF (1990) Some historical notes on early magnetic surveying in the U. S. Geological Survey. In WF Hanna (ed) *Geologic applications of modern aeromagnetic surveys*: U. S. Geological Survey Bulletin 1924, 63–73
- Helman DS (2013) Earth electricity: a review of mechanisms which cause telluric currents in the lithosphere. *Ann Geophys* 56(5): G0564. <https://doi.org/10.4401/ag-6184>
- Hood PJ (1990) Aeromagnetic survey program of Canada, mineral applications, and vertical gradiometry. In WF Hanna (ed) *Geologic applications of modern aeromagnetic surveys*: U. S. Geological Survey Bulletin 1924, 7–23
- Hollerbach R (1996) On the theory of the geodynamo. *Phys Earth Planet Inter* 98:163–185
- Jackson A (1992) Historical data for geomagnetic field modelling. In: Langel RA, Baldwin RT (eds) *Types and characteristics of data for geomagnetic field modelling*. NASA conference publication 3153. National Aeronautics and Space Administration (NASA), Washington, pp 1–30
- Jonkers ART (2000) *North by Northwest*, 2 vols. Cuvillier, Göttingen, 1002 pp
- Juusola L, Vanhamäki H, Viljanen A, Smirnov M (2020) Induced telluric currents play a major role in the interpretation of geomagnetic variations. *Ann Geophys Discuss.* <https://doi.org/10.5194/angeo-2020-21>. in review
- Kim KH, Park JH, Lee DH, Lysak R, Kwon HJ, Hwang J (2019) Magnetic field oscillations observed by Swarm satellites in the Nightside upper ionosphere during low-latitude Pi2 pulsations. *J Geophys Res Space Phys.* <https://doi.org/10.1029/2019JA026608>
- Kono M, Roberts PH (2002) Recent geodynamo simulations and observations of the geomagnetic field. *Rev Geophys* 40(4):1013. <https://doi.org/10.1029/2000RG000102>
- Larmor J (1919) Possible rotational origin of magnetic fields of sun and earth. *Electrical Review* 85:412ff. Reprinted in *Engineering* vol. 108, pages 461ff
- Lesur V, Hamoudi M, Choi Y, Dymont J, Thébaud E (2016) Building the second version of the world digital magnetic anomaly map (WDMAM). *Earth Planets Space* 68:27
- Mandea M, Mayaud PN (2004) Guillaume le Nautonier, un précurseur dans l'histoire du géomagnétisme. *Rev Hist Sci* 57:161–174
- Mandea M, Korte M, Yau A, Petrovsky E (eds) (2019) *Geomagnetism, aeronomy and space weather: a journey from the Earth's core to the sun*, Part of special Publications of the International Union of Geodesy and Geophysics. Cambridge University Press, Cambridge, ISBN 9781108418485

- Mason RG (1958) A magnetic survey off the west coast of the United States between latitudes 30° and 36° N, longitudes 121° and 128° W: *Geophysical Journal of the Royal Astronomical Society* 1:320–329
- McPherron RL (2005) Magnetic pulsations: their sources and relation to solar wind and geomagnetic activity. *Surv Geophys* 26:545–592. <https://doi.org/10.1007/s10712-005-1758-7>
- Melloni M (1853a) Sur l'aimantation des roches volcaniques. *C R Hebd Seances Acad Sci* 37:229–231
- Melloni M (1853b) Du magnétisme des roches. *C R Hebd Seances Acad Sci* 37:966–968
- Morley LW, Larochelle A (1964) Paleomagnetism as a means of dating geological events. *Royal Society of Canada Special Publication* 8:39–50
- Nabighian MN, Grauch VJS, Hansen RO, LaFehr TR, Li Y, Peirce JW, Phillips JD, Ruder ME (2005) The historical development of the magnetic method in exploration. *Geophysics* 70(6):33ND–61ND. <https://doi.org/10.1190/1.2133784>. 75th Anniversary
- Newitt LR, Barton CE, Bitterly J (1996) *Guide for Magnetic Repeat Station Surveys*; International Association of Geomagnetism and Aeronomy, Working Group V-8: Analysis of the Global and Regional Geomagnetic Field and its Secular Variation, ISBN: 0-9650686-1-7
- Norman R (1581) *The Newe Attractiue*, Containyng a short discourse of the magnes or lodestone, and amongst other his vertues, of a newe discovered secret and subtill propertie, concernyng the declinyng of the needle, touched therewith under the plaine of the horizon. J. Kyngston, London
- Olsen N, Stolle C (2017) Magnetic signatures of ionospheric and magnetospheric current systems during geomagnetic quiet conditions – an overview. *Space Sci Rev* 206:5–25. <https://doi.org/10.1007/s11214-016-0279-7>
- Reford MS, Sumner JS (1964) *Aeromagnetics*: *Geophysics* 29:482–516
- Ronalds BF (2016) *Sir Francis Ronalds: father of the electric telegraph*. Imperial College Press, London. ISBN 978-1-78326-917-4
- Sidot T (1868) Recherches sur la polarité magnétique de la pyrite de fer et de l'oxyde correspondant préparés artificiellement. *C R Hebd Seances Acad Sci* 67:175–176
- Singer BS, Jicha BR, Mochizuki N, Coe RS (2019) Synchronizing volcanic, sedimentary, and ice core records of Earth's last magnetic polarity reversal. *Sci Adv* 5(8):eaaw4621. <https://doi.org/10.1126/sciadv.aaw4621>
- Tarduno JA et al (2010) Geodynamo, solar wind, and magnetopause 3.4 to 3.45 billion years ago. *Science* 327:1238
- Tarduno JA, Cottrella RD, Bono RK, Oda H, Davis WJ, Fayeke M, Erve OV, Nimmo F, Huang W, Thern ER, Fearn S, Mitra G, Smirnov AV, Blackman EG (2020) Paleomagnetism indicates that primary magnetite in zircon records a strong Hadean geodynamo. *PNAS* 117(5):2309–2318. <https://doi.org/10.1073/pnas.1916553117>
- Thomson AWP, McKay AJ, Clarke E, Reay SJ (2005) Surface electric fields and geomagnetically induced currents in the Scottish Power grid during the 30 October 2003 geomagnetic storm. *AGU Space Weather* 3:S11002. <https://doi.org/10.1029/2005SW000156>
- Vine FJ, Matthews DH (1963) Magnetic anomalies over oceanic ridges: *Nature* 199:947–949
- Vine FJ, Wilson JT (1965) Magnetic anomalies over a young Oceanic ridge off Vancouver Island. *Science* 150(3695):485–489
- Williams DJ (1985) Dynamics of the earth's ring current – theory and observation. *Space Sci Rev* 42:375–396. NASA-supported research. ISSN 0038-6308
- Xiong C, Zhou YL, Lühr H, Ma SY (2016) Diurnal evolution of the F region electron density local time gradient at low and middle latitudes resolved by the Swarm constellation. *J Geophys Res Space Phys*. <https://doi.org/10.1002/2016JA023034>
- Zhao H et al (2015) The evolution of ring current ion energy density and energy content during geomagnetic storms based on Van Allen Probes measurements. *J Geophys Res Space Phys* 120:7493–7511. <https://doi.org/10.1002/2015JA021533>

Magnetic Methods, Surface

Nandini Nagarajan

National Geophysical Research Institute, Hyderabad, India

Definition

This entry describes the techniques, instrumentation, and uses of ground-based magnetic measurements on land and at sea. It includes sections on the network of Magnetic Observatories on ground, where long-term measurements of the field and its changes are observed. It also includes the techniques of land and marine magnetic surveys, which are aimed at extracting local or regional information of subsurface structure.

Magnetic Observatory

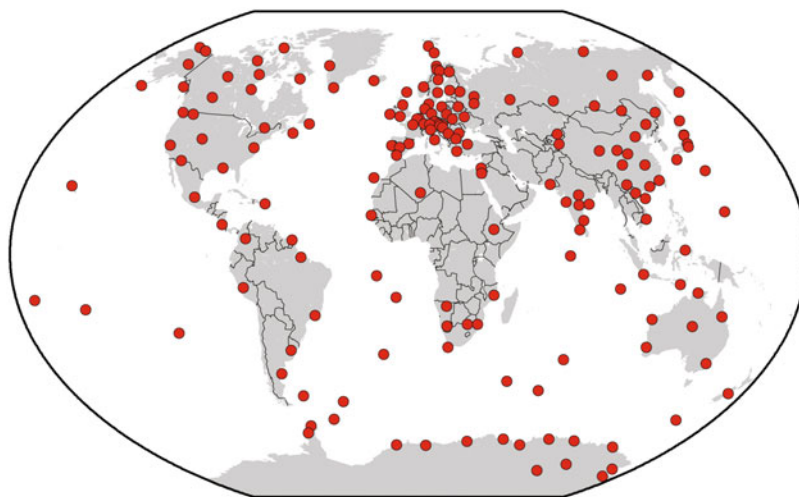
Magnetic observatories continuously measure the Earth's magnetic field at fixed locations at the Earth's surface. They are designed to operate for several decades to stably record rapid and slow variations of the earth's magnetic field and report time-averaged values that estimate the form of the Earth's magnetic field. Since observatories started operating in the nineteenth century, the methods of measurement, reporting, and analysis have varied with region and time. After the International Geophysical Year (IGY –1957–58), the quality and organization of the data have been standardized, making spatial and temporal studies easier, more refined, and consequent models more accurate. Classical torsion magnetometers, with analogue-photographic systems, were replaced in 1980s, by improved fluxgate magnetometers and digital data acquisition systems. The International Association for Geomagnetism and Aeronomy (IAGA) is the coordinating body for the magnetic observatories of the world, standardization of measurement, exchange of data, and presentation of models of the Earth's magnetic field. There are about 200 magnetic observatories reporting standardized data, and several more making continuous variation measurements with absolute control (Fig. 1).

INTERMAGNET is network (International Real-time Magnetic Observatory Network) arising out of continuing improvements in instrumentation and upgradation of magnetic standards evolved by IAGA that provides 1-min values of the three components of the geomagnetic field, in near-real time. With developments in instrumentation, data acquisition, and data dissemination, an increasing number of observatories are producing a filtered 1-s data product in addition to

Nandini Nagarajan: was retired.

Magnetic Methods, Surface,

Fig. 1 Currently operating observatories (updated in www.geomag.bgs.ac.uk)



traditional 1-min data, hourly means, daily means, monthly means, and annual means. Presently there are over 200 observatories, worldwide.

210MM is a chain of variation stations and observatories, that cover Russia, Japan, South east Asia, and Oceania along the magnetic meridian 210° . This chain has reported data from remote regions of the Southern Indian and Pacific Oceans and made seminal contributions in research. The project has also revived observatories in Africa, providing technical support.

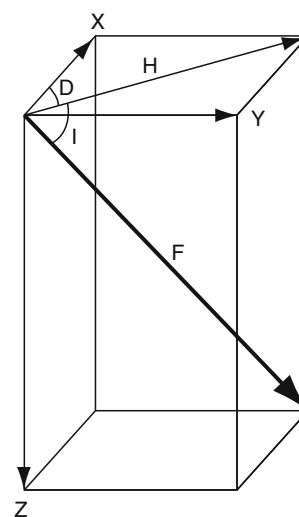
Measurements at Magnetic Observatories

Vector and scalar measurements: Magnetic measurements can be scalar (total field) or vector. To define the vector, 3 out of 7 elements need to be measured simultaneously (Fig. 2). Commonly any three of these elements are measured in a magnetic observatory as variations, to fully define variations of the total magnetic field.

Absolute measurements: Most vector magnetometers used in magnetic observatories (earlier torsion, and recently fluxgates), usually drift in time due to several factors such as temperature variations, magnetic moment, and ageing of the electronics. In order to periodically recalibrate the vector measurements, absolute measurements fully defining the magnetic field vector are performed, to complement continuous measurements, on regular basis (typically once a week). Such measurements are made by a trained observer using a single-axis fluxgate magnetometer mounted onto a non-magnetic theodolite and a scalar magnetometer. From which, the continuous measurements can be calibrated to a baseline defined by absolute measurements. Two types of data are available from observatories: preliminary data, which are made available in quasi-real time (less than 72 h), and definitive data, which are produced only once a year, typically a

Magnetic Methods, Surface,

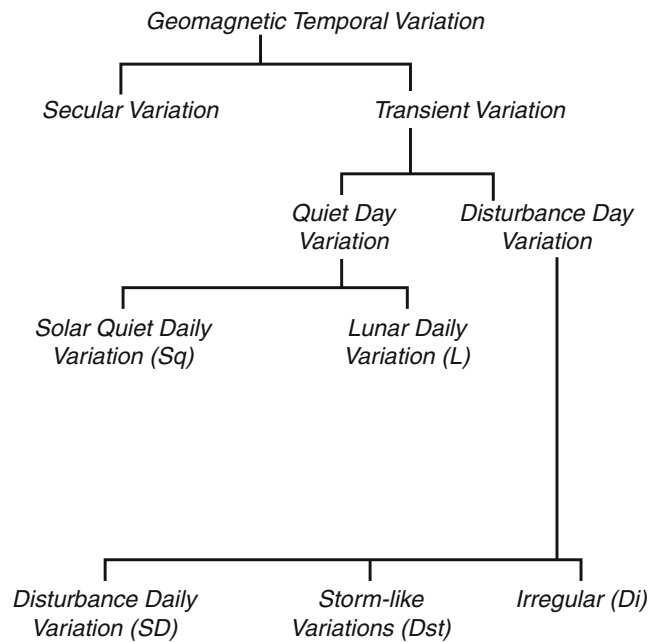
Fig. 2 Magnetic elements defined in geographic and geomagnetic coordinates (Jankowski and Sucksdorf 1996)



few months after the end of the calendar year. After reducing time variations to absolute baseline at a location, continuous observations of the vector magnetic field are produced at desired sampled intervals: minute, hour, daily, month, year, etc. These data are available for all years going back to the start of observations at each observatory. The format for easy assimilation and distribution of this data are formulated by IAGA working groups – updated periodically since 1957–58. This data is also available at World Data Center for Geomagnetism, established in different countries, partly by their research institutions and partly funded by IAGA.

Observations from this Data

Magnetic observatory data are used to report and study periodic, episodic, and transient variations that are classified by time scale and mechanism: Long-term secular variation,



Magnetic Methods, Surface, Fig. 3 Diagram showing the different types of geomagnetic variations (Okeke et al., <http://www.sao.ac.za/~wgssa/as4/fokeke.html>)

11-year solar cycle, annual, semi-annual, seasonal, and daily variation and variations over minutes to days, classed as disturbances (Fig. 3).

Observatory yearbooks often give the average solar daily variation for different months of the year, and even separately for quiet, disturbed, and all days. Based on an analysis of magnetic activity by trained observers, and estimation of the activity indices (most commonly K), five quiet and disturbed days are selected for each month by the International Service of Geomagnetic Indices.

Magnetic variations have characteristics, varying over latitude and over a wide range of periods attributed to different sources – solar activity – from 11 year to 27 days, diurnal harmonics of ionospheric currents from 24 to 4 h, shorter events like sudden commencements, sudden impulses, substorms (bays) -2 h to 1 min, magnetic pulsations, with small amplitude and short periods, from 1 to 60 s. Hourly and daily variations are a measure of solar interaction with the Earth's magnetic field. This can be both solar radiation interacting daily with the ionosphere or corpuscular streams interacting with the magnetosphere, manifested as disturbances. Regular solar daily variation is also a function of season, solar activity and latitude. The variation of the magnetic field on magnetically quiet days is called the solar quiet day variation or Sq variation. Magnetic storms (~day), substorms (~hour), pulsations (~min), sudden impulses, and solar flares are some of the transient disturbances. Modelling the fast and slow variations of the Earth's magnetic field has made rapid strides. Currently, it is possible to access models, based on studies of

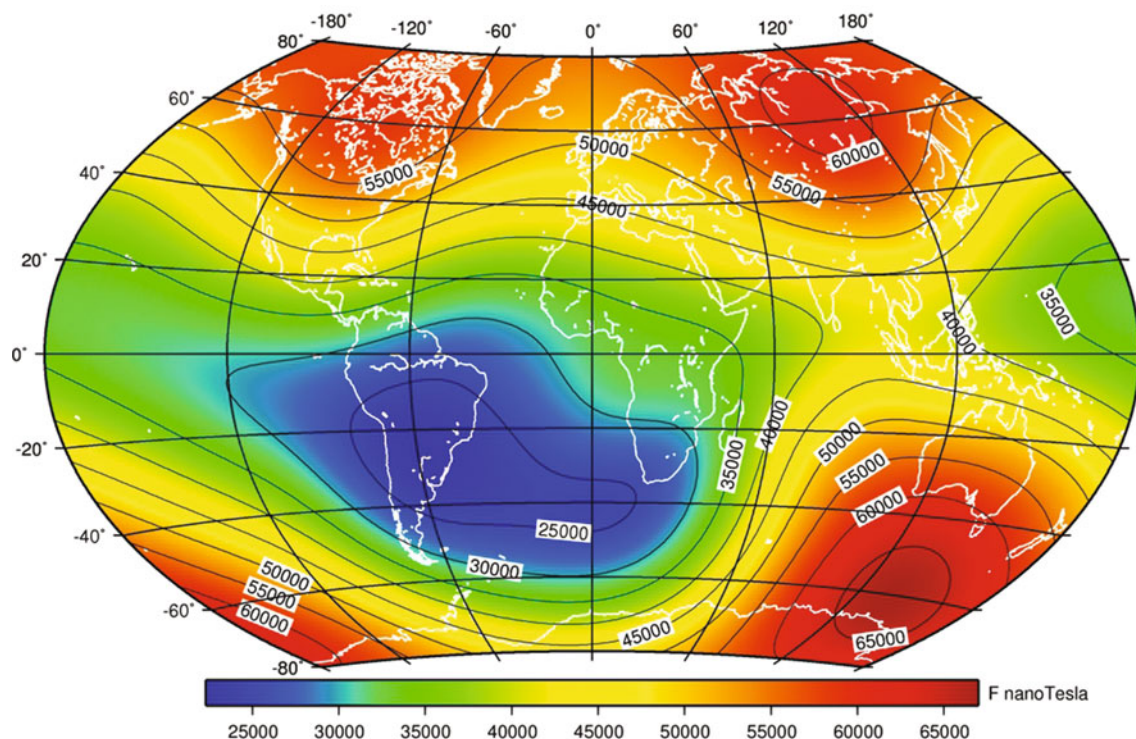
observatory variations, solar parameters that provide an averaged trend of quiet or storm time variations at any latitude or longitude. These have been compared with observatory data, providing robust estimates, in Realtime of magnetic disturbances. The longer, slower variations are obtained from monthly and annual means of observatory data. The trends seen in the data compiled over the last 150 years have been used to produce spherical harmonic representation of the Earth's magnetic field which is internally generated and its secular variation which are indicative of processes within the Earth's mantle and outer core. Definition of the main magnetic field and secular variation are currently supplemented by low-earth orbiting satellites (OERSTED, CHAMP, SWARM missions). This is one of the fastest growing areas of study.

Products and Uses of Observatory Data

Magnetic field values at different points on the Earth, their variation on different timescales are obtained. The Earth's magnetic field has been approximated as a dipole and more accurately expressed as a spherical harmonic series, based on time-averaged means of observatory data, by Gauss, demonstrating that 95% of the field intensity is derived from inside the Earth and 5% due to external currents in the atmosphere (now known as magnetosphere). From variations of the magnetic field at regularly sampled intervals, much information can be extracted. The time averaged values over months and years provide a data set approximating the potential field of the internally generated magnetic field. These values define the shape of the Earth's magnetic potential field (Fig. 4). This is termed the International Geomagnetic Reference Field, for a stated epoch (instant in time). Other data sets from magnetic surveys and satellites are added to refine the IGRF.

Daily variation at each observatory has contributed to the definition of solar-induced daily ionospheric currents (Sq). Variation with latitude and season has also been defined (Fig. 5). Monthly, seasonal, and annual variations are largely controlled by radiation (F10.7) from the sun. The effects of solar radiation on the ionosphere drive electric current loops in the ionosphere (height ~100 km). Study of these regular variations from both observatory data and in recent decades, from satellite data, has resulted in better models of the current loops and their variability. In addition, solar induced disturbances affect the earth's magnetic field.

Study and classification of magnetic disturbances, the influence of sunspot activity, and a theory explaining these phenomena, are evolved from a study of observatory data, during 1920–1940. This theory has since been corroborated from a direct observation of corpuscular radiation streams from the Sun and magnetospheric currents, by satellites. Each day's variations are analyzed and classified from quiet to disturbed, as a measure of magnetic activity in the earth's



Magnetic Methods, Surface, Fig. 4 Spatial variation of the Earth Magnetic field scalar (F), approximating a dipole field, over the Earth (www.geomag.bgs.ac.uk)

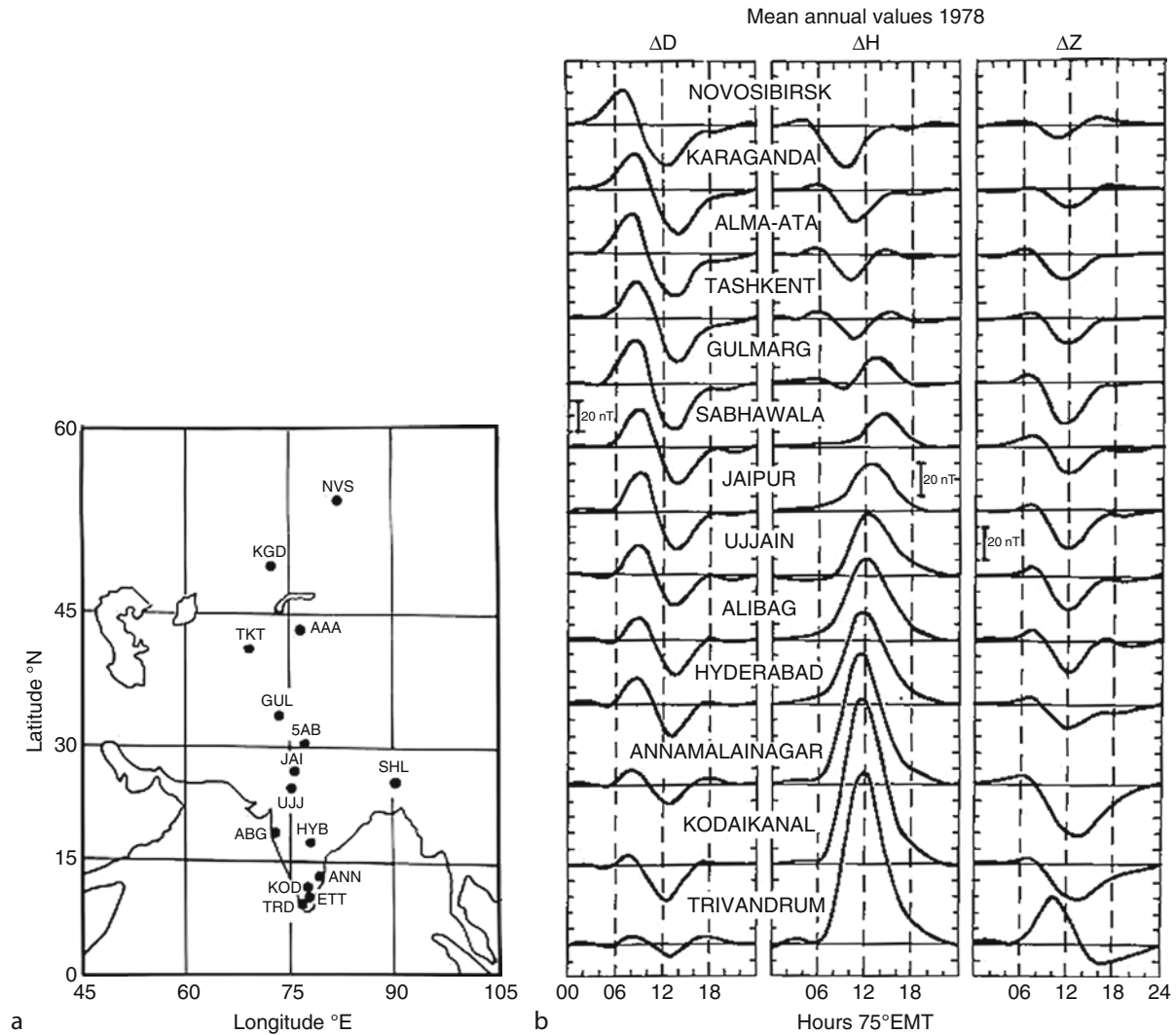
magnetosphere, in a manner that eliminates geographic difference. Geomagnetic indices were formulated as a composite of reported variations from observatories at different locations. These indices and several new ones continue to be reported and are updated, revised in accordance with discussion and resolutions passed by IAGA.

Observatory data are useful in determining rapid, medium term as well as slow variations in Earth's magnetic field. Each range of variations, studied for specific signals, contributes to the understanding of different processes (Fig. 6). Some of the phenomena being currently studied are:

- (i) Secular variation of the magnetic field at different places on the Earth. It is then expressed in spherical harmonic coefficients and included in the IGRF by adding time derivatives of coefficients. Variation in monthly and annual means, define the secular, slow change of the Earth's internal magnetic field. This has been mapped going back to about 300 years. The pattern of secular variation places constraints on the theory of generation of the Earth's magnetic field and maps fluid flow in the outer core. It is also the only means of studying electromagnetic and thermal properties of the Earth's outer core (depth: 3000 km). Knowledge of these variations is important in theoretical modeling of the geomagnetic dynamo. It is the only means of studying the motions

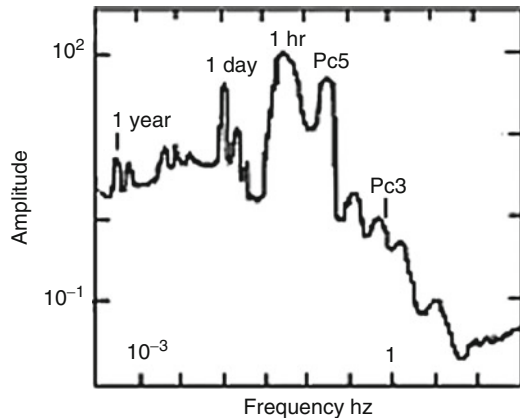
of the electrically conducting fluid in the Earth's outer core, where the main geomagnetic field is generated. These data are also widely used to calculate time-varying core field models, often in combination with satellite magnetic data or historical data.

- (ii) The regular variations of the magnetic field are related to rotation and/or orbital movements of the Earth, Sun, and Moon. The most prominent is the diurnal variation or solar daily variation, having amplitudes of the order of 10–100 nT (nanotesla). Solar radiation ionizes the higher atmosphere during the daylight hours, and the gravitational forces of the Sun and the Moon force the ionospheric layers in a tidal motion. So, the ionized gas in the ionosphere moves in the magnetic field of the Earth, creating electric currents which are seen as daily variations in magnetic recordings. There are two well-known periodic variations, the solar daily variation and the lunar daily variation. Studies of such variations at observatories around the world have resulted in models of Sq variations that incorporate ionospheric, atmospheric, and inter planetary effects. It is now possible to separate these effects, using averaged models.
- (iii) A global network is essential in monitoring the pattern of geomagnetic storms, local disturbances, mapping of currents generated in the magnetosphere, and thereby, inferences about presence of charged particles, models



Magnetic Methods, Surface, Fig. 5 Diurnal variation of the magnetic field, showing changes with latitude for solar quiet days. (Parkinson 1983, reproduced with permission by Jankowski and Sucksdorf

(1996)) (www.wessexarch.co.uk marine survey, www.geologyscience.com magnetic survey)



Magnetic Methods, Surface, Fig. 6 A illustration of the wide spectrum geomagnetic variations (~ 1 s to 1 year) – (Serson 1983)

of magnetic field lines. These networks complement the intensive studies of satellite-borne magnetic data of the past two decades and with the ongoing SWARM mission

(iv) Induced currents in the crust and mantle are caused by external variations of the earth's magnetic field. For each periodic and episodic event described above, a corresponding current is induced in the finitely conducting layers of the Earth. An estimation of these currents provide a measure of the electrical conductivity distribution within the Earth's crust and mantle. This has been applied as a tool to study the Earth's interior and a large volume of electromagnetic variation data has been collected solely to this purpose. In the past decade, induced signals from satellite magnetic data have been used to model electrical conductivity distribution, over continents and oceans.

- (v) Ground observation networks provide calibration and control of magnetic survey satellites and complement detailed ground surveys of the magnetic field.
- (vi) Magnetic pulsations: 10 Hz – 600 s – are small amplitude variations generated at the boundary between magnetosphere and solar wind, under various dynamic conditions and transmitted through the magnetosphere and ionosphere and carry signatures of the properties of these regions as well.

Land and Marine Surveys

Principles of Survey Measurements

The measurement of spatial variation of magnetic field depends on determining the value at a spot, obtaining the position, time of measurement, and eliminating temporal variation, during the period of the survey, at each location. For this data have to be reduced to a time epoch in contrast to observatory measurements, which provide time variations at a fixed point. The objectives of a survey govern the use of specific magnetic instruments, surveying parameters (for example, observation elevation and data spacing), and the coordinate system used for the analysis of the data. For geological interpretation either total field scalar or vector measurements are suitable. But mapping of internally induced currents, in the crust – for tectonic studies, or in the mantle and outer core – for studies of the deep interior, require vector measurements. Instruments used vary from torsion magnetometer, fluxgate, proton precession, SQUID, etc. The precision of most modern instruments is between 0.1 and 1 nT. The residual, after removing temporal variation, is the variation in potential caused by deflection of field lines, through a volume of subsurface, that has non-uniform magnetic properties

Land Surveys

Magnetic surveys on land are conducted to find changes in physical properties of rocks in the subsurface (magnetic susceptibility). Local variations in the Earth's magnetic field are most often caused by ferromagnetic objects or other magnetic objects located in the vicinity of the studied area. Through measuring and recording of this variation, it is possible to detect objects submerged beneath the water (and sediments). An accurate map of the object's position can be created by recording the sensor's position along with the magnetic variation.

Vector or scalar field values are recorded, with precise location (GPS or surveying tools) and time. Time stamping of the data is critical, as the diurnal variations (quiet) as well as disturbances (storms) of the Earth's magnetic field are recorded at a base station in the survey area and eliminated

from the measurements over the survey area. The data is reduced to a time epoch, removing these external variations. IGRF values for that region are also be deducted. These corrections are made to eliminate temporal variations that are described above, from the data. Contributions of the Earth's main field at any particular location are removed from either vector or scalar measurements. Residual values obtained from the survey indicate changes in magnetic susceptibility in the subsurface, from which we infer, that there are changes in composition (Fig. 7a). This information is used to interpret the survey findings for a range of applications. Developments in Unmanned Aerial Vehicles (UAV) technology has made them suitable for geophysical surveys. Concurrently, magnetometers have also been adapted for this use, made robust and lightweight. The successful use of UAV-based surveys has increased the scope and capability of magnetic surveys over rough, forested and inaccessible terrains.

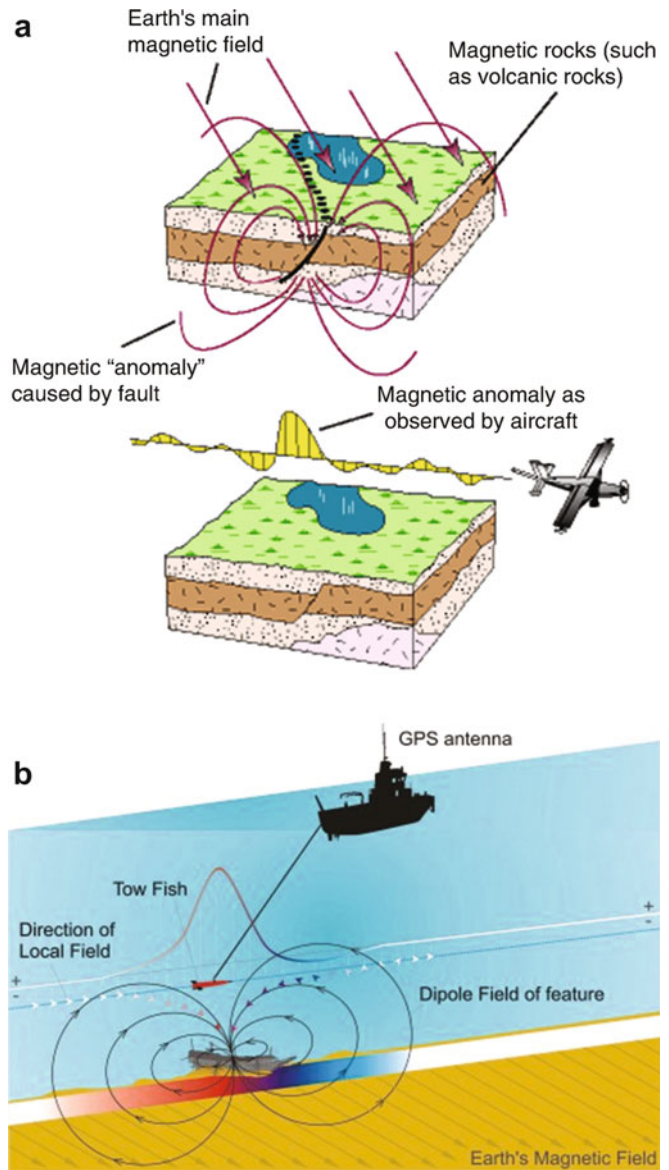
Typical applications are found in archeology, engineering studies, search and recovery of buried objects, mineral exploration, and geological structure mapping. The most significant change in magnetic properties occurs at Curie point, the depth varies considerably in different tectonic settings. Therefore, large-scale magnetic surveys provide significant inputs to deep crustal studies. Magnetic surveys are a viable method to obtain information regarding the crystalline basement over areas of sedimentary cover. The method has also been used to provide information on upper crustal structure and thermal state, along major geotranssects. Mapping of spatial magnetic variation provides information about the thickness of the basement rock beneath sedimentary basins/supracrustal rocks and thermal state of the crust of that region. For these surveys, satellite magnetic data has also been effectively used.

Marine Magnetic Surveys

Marine magnetometry surveys are carried out using a magnetometer specially modified for marine operations (Fig. 7b). The magnetometer is installed in a water-tight tow-fish, which is towed behind the vessel using a tow cable. The sensor used most often is a quantum magnetometer (proton precession or Overhauser). Magnetometers are somewhat more difficult to use than SONAR; however, they do provide accurate and reliable data. Since magnetometers do not directly measure the physical properties of the areas or objects of interest, the output signals of these sensors require some signal processing for translation into usable data. Methods of marine magnetic surveys are more elaborate, since instruments have to be towed behind a vessel and data collection is challenging. However, principles of analysis are similar. Similar corrections to marine survey data are applied to eliminate temporal variation and the contribution of the main magnetic field. However, with marine surveys, location of the towed magnetometer is also a critical factor. The speed of towing, movement of the magnetometer due to ocean swell, pitch, and roll

Magnetic Methods, Surface,

Fig. 7 (a) Sketch of land survey https://link.springer.com/referenceworkentry/10.1007/978-3-319-73568-9_8. (b) Schematic of a marine survey <http://ets.wessexarch.co.uk/recs/how-we-study-the-seafloor/geophysical-survey/04-magsurvey/>



reduce the accuracy of measurements. For measurements made in optimal sea conditions, time, trace, and main field corrections are adequate.

Magnetic surveys have a very diverse range of applications, including locating objects such as submarines, sunken ships, hazards for tunnel boring machines, hazards in coal mines, unexploded ordnance, toxic waste drums, as well as a wide range of mineral deposits and geological structures. Marine magnetic methods also map the variation in thickness of the oceanic crust – which consists of uniform magnetic material (basalt). Significant variations in marine magnetic measurements have been attributed to pulses of ocean-floor formation. These pulses belonging to different epochs have been identified by adjacent stripes of differing magnetization, of the ocean floor – a significant discovery of the phenomenon of “sea-floor spreading.”

Instrumentation Technologies

The principles of magnetometers are:

- (i) Torsion magnetometers: These measure the opposing force to the magnetic field – as in torsion fiber suspended magnetometers and deflection measurements. This was the oldest and mostly accepted method for nearly 200 years. Examples: Quartz horizontal magnetometer (QHM), Balance magnetometer Zero (BMZ), Kew Magnetometer, La Cour/Eschenhagen variometers and others.
- (ii) Flux measurement: It amplifies and measures in a ferrite sensor, with appropriate hysteresis – fluxgate magnetometers. Accuracy of fluxgate magnetometers can be adapted to different ranges of measurement. When

mounted in triaxial mode, they are given reliable values of the field. The Magnetometers are sensitive to temperature variation and internal noise, but these have been minimized by improved design. They are suitable for observatories, satellite, and unmanned vector measurements.

- (iii) **Magnetic Theodolite:** Estimating opposing force to the magnetic field as applied to a sensor of finite magnetic moment – using an oscillating magnet and theodolite, an old, classical method for absolute values – several instruments are still in use as secondary standards. A modern version of this is the Declination Inclination magnetometer. Thus, uses a singly fluxgate counted on a theodolite. This is adjusted to a null position to find the inclination of the magnetic field. In conjunction with total field measurement, this would result in vector components. The theodolite also gives azimuth and declination measurements. Since it is a null observation, accuracy is good. This has been adopted as standard for modern magnetic observatories and is increasingly used for surveys also.
- (iv) **Precession:** Based on the frequency of precession of hydrogen atoms (protons) after a strong external field is applied. Precession frequency is proportional to the ambient field. This is the most stable accurate method at present. Counting of frequency is very accurate with improved digital time stamping. The proton precession magnetometer measures only the scalar of total field intensity. An improvement of the same principle is the Overhauser magnetometer, which stimulates precession almost continuously and thereby produces a steady signal that is a measure of ambient field. This magnetometer is now chosen for most remote, continuous, and surveying operations.
- (v) **Optically pumped magnetometers:** The principle is of optical excitation of cesium (or other) gas, with a focused beam of coherent light. Zeeman splitting effect produces splitting of the lines emitted by the sample gas. The frequency of split is proportional to the ambient magnetic field.
- (vi) **Induction coil magnetometers:** This operates on the simple principle of measuring the output of a coil of many windings around a ferrite core. The signal is proportional to the derivative of the magnetic field. The induction coil has good response of a wide frequency range of measurement. This is used for rapid variation, e.g., pulsations, Magnetotellurics, etc.
- (vii) A recent development is the use of unmanned aerial vehicles (UAV), fitted with magnetometers. The magnetometers used are fluxgate, as well as quantum magnetometers and optically pumped magnetometers.

Summary

The shape, strength, and variations of the Earth's magnetic field have been studied over 500 years. Starting from measurements, repeated in time, at different places on earth, a model of the Earth's magnetic field in space and time has been built up. Magnetic observatories are places where long-time series of accurate measurements of the magnetic field are made, analyzed, and disseminated to institutions and persons seeking this information. Variations in different period ranges provide information about diverse processes, affecting the magnetic field. Short period variations (~min, hr) are indicative of disturbances in the magnetosphere caused by the Sun's corpuscular streams and magnetic fields, medium-term variations (~day, year) describe effects of the sun's radiation, and long-term variations (~year, decade, century) provide insights into the nature of the Earth's internal magnetic field, its origin, and modification in the Earth's outer core (secular variation). Closely spaced measurements on the Earth's surface are used in studies for resource exploration, tectonics, and mapping the formation of the oceanic crust.

The form of the Earth's magnetic field is redefined every 5 years, collating all available measurements as the IGRF. This also provides estimates of secular variation. Progress in satellite-based magnetic measurements (both vector and scalar) since 1999 have improved accuracy and precision of modelling spatial and temporal variations of the Earth's magnetic field. This has resulted in increasingly accurate models of processes on the Sun, in space, and on and within the Earth.

Cross-References

- ▶ [Electromagnetic Pulsations and Magnetic Storms](#)
- ▶ [Equatorial Electrojet](#)
- ▶ [Geoelectromagnetism](#)
- ▶ [Geomagnetic Field, Global Pattern](#)
- ▶ [Geomagnetic Field, IGRF](#)
- ▶ [Geomagnetic Field, Measurement Techniques](#)
- ▶ [Geomagnetic Field, Secular Variation](#)
- ▶ [International Geophysical Year](#)
- ▶ [Magnetic Anomalies: Interpretation](#)
- ▶ [Magnetic Anomaly Map, Global](#)
- ▶ [Magnetic Methods, Airborne](#)
- ▶ [Magnetic Methods, Principles](#)
- ▶ [Magnetometers](#)
- ▶ [Magnetovariation Studies](#)

Bibliography

- Campbell WH (1997) Introduction to geomagnetic fields. Cambridge University Press, Cambridge

- Chapman S, Bartels J (1940) *Geomagnetism*, vol I, II. Clarendon, Oxford
<http://www.geomag.bgs.ac.uk/earthmag.html>
<http://www.icsu-fags.org/ps06isgi.htm>
<http://wdc.kugi.kyoto-u.ac.jp/-WDC>
- Jankowski J, Sucksdorf C (1996) Guide for magnetic measurements and observatory practice. IAGA, Warsaw, p 238
- Kearey P, Brooks M, Hill I (2002) *An introduction to geophysical exploration*, 3rd edn. Wiley, Oxford
- Merrill RT, McElhinny MW, McFadden PL (1996) *The magnetic field of the Earth*, International geophysics series, vol 63. Academic Press, San Diego, California, 531 pp
- Moskowitz BM, Chandler V (2015) Resources in the near-surface earth in treatise on geophysics, 2nd edn. In: Schubert G Editor-in-Chief, editor. *Treatise on Geophysics*. 2. Elsevier; The Netherlands: p. 5604
- Parkinson WD (1983) *Introduction to geomagnetism*. Scottish Academic Press, Edinburgh
- Parshin AV, Morozov VA, Blinov AV, Kosterev AN, Budyak AE (2018) Low-altitude geophysical magnetic prospecting based on multi-rotor UAV as a promising replacement for traditional ground survey. *Geo-spatial Inf Sci* 21(1). Special Issue: Information Processing for Unmanned Aerial Vehicles (UAVs) in Surveying, Mapping, and Navigation. Guest Editors: Gui-Song Xia, Mihai Datcu, Wen Yang, Xiang Bai
- Shimizu H, Utada H (1999) Ocean hemisphere geomagnetic network: its instrumental design and perspective for long-term geomagnetic observations in the Pacific. *Earth Planets Space* 51:917–932
- Thébault E, Finlay CC, Zvereva T (2015) International geomagnetic reference field: the 12th generation. *Earth Planets Space* 67:79
- Thomson AWP (2014) Geomagnetic observatories: monitoring the Earth's magnetic and space weather environment. *R Meteorol Soc.* <https://doi.org/10.1002/wea.2329>. Special Issue www.intermagnet.org
- Yumoto K, The 210° MM Magnetic Observation Group (1996) The STEP 210° magnetic meridian network project. *J Geomagn Geoelectr* 48:1297–1309

Magnetic Modeling, Theory, and Computation

Mioara Manda¹, Carmen Gaina² and Vincent Lesur³

¹CNES – Centre National d'Etudes Spatiales, Paris, France

²Department of Geosciences, CEED – Centre for Earth Evolution and Dynamics, University of Oslo, Oslo, Norway

³Université de Paris, Institut de physique du globe de Paris, CNRS, Paris, France

Definition

The geomagnetic field has a complex pattern, both in shape and strength, and changes continuously over time. Mathematical models need to be computed to describe the strength and direction at any point on or above the Earth's surface.

Introduction

When either a ground observatory or a satellite takes a geomagnetic field measurement, it represents the superposition of

many sources. The greatest contribution is generated by the dynamo action within the fluid, iron-rich core of the Earth, known as the core field. Although depending upon the altitude and location of the measurement, the data may have sizeable contributions from the static lithospheric field being generated largely from rocks and remanent magnetism within the Earth's crust. Internal in origin are also small magnetic signals produced by the electrical currents associated to the flow in oceans. On the other side, the external field sources originate in the ionosphere and magnetosphere. These sources include the daily solar quiet variations, the ring current variations, and many other current systems (e.g., the field-aligned currents and the auroral and equatorial electrojets). We also note that currents are induced in the Earth from these external fields. Sporadic magnetic storms and pulses, which have short-term effects, also play a role in the measured magnetic field values.

The knowledge of the Earth's magnetic field relies on various data sources, ranging from rock magnetization measurements to direct magnetic field measurements, from historic magnetic measurements to recent high-quality data provided by magnetic observatories and the magnetic satellites as MAGSAT, Ørsted, CHAMP, SAC-C, and recent Swarm constellation (Manda and Purucker 2018). A combination of ground and continuous satellite measurements allows the core magnetic field and its time variation to be described with a very high resolution in space and in time, on a global scale (e.g., Finlay et al. 2017). On the other hand, recent dynamo simulations have also contributed to the understanding of the core field geometry and its dynamics, and temporal changes: some are dipole dominated, and can show polarity reversals (Jones 2015), others more towards Earth-like, do not show reversals anymore (Schaeffer et al. 2017).

The geomagnetic field changes in space and time, therefore magnetic observations must continually be acquired, on ground and near-Earth's space. The magnetic observations are performed in different ways and using various platforms. Geomagnetic observatories have provided the highest quality data, however with a very uneven spatial distribution. To increase the spatial density regionally, repeat station surveys have been conducted by many teams, but at a national scale only. Over the last decades, magnetic satellites have offered a uniform, global coverage, but face problems of attitude control and ambient current in the surrounding plasma.

This very brief description of the sources contributing into magnetic measurements and of the variety of magnetic observations shows how challenging is to obtain a reliable mathematical description of the magnetic field to be used in the modeling process. New mathematical tools need to be developed to accurately describe magnetic data. Depending on the available data and the base functions, these models are global or regional, static or time dependent. The mathematical

description of the different kinds of field models, their physical backgrounds, and applicability are vast topics subjects, and cannot be covered in detail here. Therefore, the main aim of this contribution is to give the reader a basic support and information for more detailed references.

Global Field Modeling

Spatial and Temporal Description

The geomagnetic field is global in nature, so the most obvious approach is to describe it on the whole Earth's surface. The magnetic field, following the Maxwell equations, is so it can be written as:

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (1)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (2)$$

where $\mathbf{H}$ is the magnetic field, $\mathbf{B}$ is the magnetic induction, $\mathbf{J}$ is the current density, and $\partial \mathbf{D} / \partial t$ is the electric displacement current density. In a region without magnetic field sources, for example, from the Earth's surface up to about 50 km, it is reasonable to assume that $\mathbf{J} = 0$ and $\partial \mathbf{D} / \partial t = 0$, so $\nabla \times \mathbf{H} = 0$, meaning that the vector field is conservative in the region of interest, and the magnetic field $\mathbf{H}$ can be expressed as $\mathbf{H} = -\nabla V$, where V is a scalar potential. Because $\mathbf{B} = \mu_0 \mathbf{H}$ above the Earth's surface (where $\mu_0 = 4\pi \times 10^{-7} \text{ Hm}^{-1}$), it follows that for the geomagnetic field $\nabla \cdot \mathbf{H} = 0$ and that V has to satisfy Laplace's equation:

$$\nabla^2 V(\theta, \phi, r) = 0. \quad (3)$$

Here, we underline that in geomagnetism the magnetic induction $\mathbf{B}$ is traditionally called "magnetic field," and we use this notation in the following. In spherical coordinates (θ, ϕ, r) , which are the geocentric colatitude, longitude, and radial distance, Laplace's equation takes the following form:

$$\begin{aligned} \frac{1}{r} \frac{\partial^2}{\partial r^2} (rV) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial V}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 V}{\partial \phi^2} \\ = 0. \end{aligned} \quad (4)$$

Equation 4 solutions are products of three expressions (the first is to be only a function of r , the second only a function of θ , and the third only a function of ϕ) and can be solved by separation of variables. A completely general solution is provided by spherical harmonics, and can be written in such a form to express two important contributions in the potential field: a part V^{int} describing the internal (core and lithospheric) sources, and a part, V^{ext} , describing the external (mainly magnetospheric) sources:

$$\begin{aligned} V &= V^{\text{int}} + V^{\text{ext}} \\ &= a \sum_{n=1}^{N_{\text{int}}} \sum_{m=0}^n \\ &\quad \times (g_n^m \cos m\phi + h_n^m \sin m\phi) \left(\frac{a}{r} \right)^{n+1} P_n^m(\cos \theta) \\ &\quad + a \sum_{n=1}^{N_{\text{ext}}} \sum_{m=0}^n \\ &\quad \times (q_n^m \cos m\phi + s_n^m \sin m\phi) \left(\frac{r}{a} \right)^n P_n^m(\cos \theta), \end{aligned} \quad (5)$$

where $a = 6371.2 \text{ km}$ is a reference radius, (θ, ϕ, r) are geocentric coordinates, P_n^m are the associated Schmidt-normalized Legendre functions, N_{int} is the maximum degree and order of the internal potential coefficients g_n^m, h_n^m , and N_{ext} is that of the external potential coefficients q_n^m, s_n^m .

Some of the base functions show special characteristics and are named individually:

- Zonal functions are characterized by $m = 0$ and oscillate only in the latitudinal direction
- Sectorial functions are defined by $n = m$ and oscillate only in the longitudinal direction
- Tesseral functions oscillate in both the longitudinal and latitudinal direction

The Gauss coefficients are used to define the magnetic field in terms of "sources." The first term, g_1^0 , is associated with the geocentric dipole oriented along the vertical axis, i.e., axis of the Earth's rotation, with dipole moment $g_1^0 4\pi a^3 / \mu_0$. The next two terms, g_1^1, h_1^1 , characterize the geocentric dipoles oriented along the two last orthogonal axis. The magnitude of the *geocentric dipole* is given by $M = (4\pi a^3 / \mu_0) \left((g_1^0)^2 + (g_1^1)^2 + (h_1^1)^2 \right)^{1/2}$, and it is tilted at roughly 11° from the rotation axis. The potential for a dipole falls off as r^{-2} , and the strength of the field components as r^{-3} . Similarly, for $n = 2$ the terms represent the *geocentric quadrupole* (potential falls off as r^{-3} and the strength as r^{-4}), for $n = 3$ the terms represent the *geocentric octopole* (potential falls off as r^{-4} and the strength as r^{-5}), and so forth. An accurate description of the internal field to high degrees and orders involves a weighting scheme of many parameters of this inverse problem and on the way to handle the ionospheric and magnetospheric signatures in spherical harmonics (e.g., Finlay et al. 2016; Finlay et al. 2017; Lesur et al. 2010; Sabaka et al. 2015).

The first three Gauss coefficients (g_0^1, g_1^1, h_1^1) represent the *dipole* field, while the remaining terms represent the *nondipole field*. The lowest degree terms correspond to the largest wavelength features of the field, and are dominated by zonal harmonics. The first Gauss coefficient

has a $\cos\theta$ dependence, so at distance r the “wavelength” of the associated magnetic feature is $2\pi r$. Generally speaking, the term g_n^m has a wavelength of $2\pi r/(n + \frac{1}{2})$ (Backus et al. 1996). To describe only the part of the geomagnetic field with sources beneath the Earth’s surface, only the part V^{int} of Eq. 5 is used. However, this first term of Eq. 5 describes both the core and lithospheric fields. The usual way to separate sources originating in the core from those originating in the lithosphere (Loves 1966) is by computing the mean square value of the field over the Earth’s surface produced by harmonics of a given n :

$$W_n = (n+1) \sum_m \left((g_n^m)^2 + (h_n^m)^2 \right). \quad (6)$$

Variation of W_n depends on the degree n , which is known as the power or Lowes–Mauersberger spectrum, and Fig. 1 shows its behavior. The steep part of the spectrum ($n \leq 13$) indicates the signature of the long wavelengths of the core field; the transition degrees ($n = 14-15$) can be attributed to the signals from both core and lithospheric sources, and the higher degrees ($n \geq 16$) are dominantly lithospheric in origin.

The above description in terms of spherical harmonics is used to characterize static or “snapshot” fields. Considering the definition of the magnetic field components X , Y , and Z (in the northward, eastward, and radially inwards directions) and the definition of the spherical polar coordinates (Merrill et al. 1996), it is apparent that $X = -B_\theta$, $Y = B_\phi$ and $Z = -B_r$. For a given epoch, the components of the geomagnetic field are obtained from the model

coefficients using the internal part of Eq. 5, as derivatives of potential.

$$\begin{aligned} B_\theta &= -\frac{1}{r} \frac{\partial}{\partial \theta} (V^{\text{int}}) \\ &= -\sum_{n=1}^{N_{\text{int}}} \sum_{m=0}^n (g_n^m \cos m\phi + h_n^m \sin m\phi) \left(\frac{a}{r}\right)^{n+2} \frac{\partial}{\partial \theta} P_n^m(\cos \theta) \end{aligned} \quad (7)$$

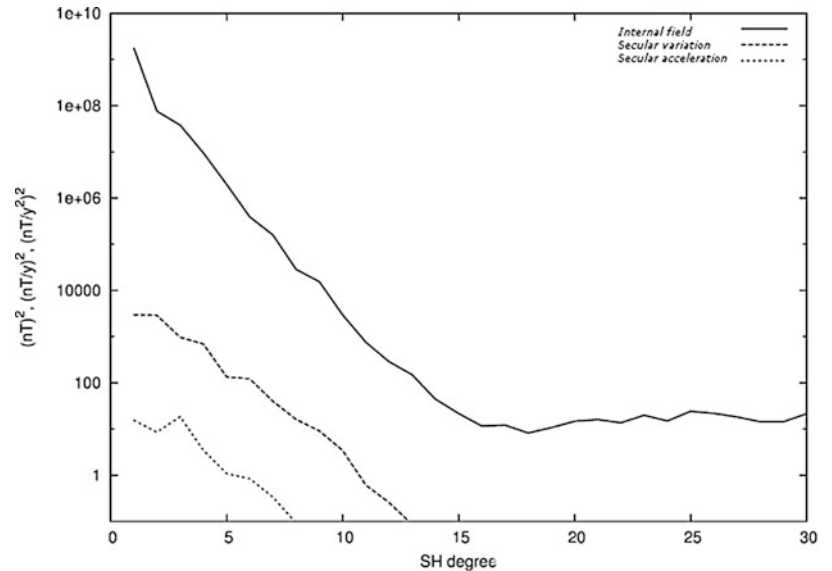
$$\begin{aligned} B_\phi &= -\frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} (V^{\text{int}}) \\ &= \frac{1}{\sin \theta} \sum_{n=1}^{N_{\text{int}}} \sum_{m=0}^n m (g_n^m \sin m\phi - h_n^m \cos m\phi) \left(\frac{a}{r}\right)^{n+2} P_n^m(\cos \theta) \end{aligned} \quad (8)$$

$$\begin{aligned} B_r &= -\frac{\partial}{\partial r} (V^{\text{int}}) \\ &= \sum_{n=1}^{N_{\text{int}}} \sum_{m=0}^n (n+1) (g_n^m \cos m\phi + h_n^m \sin m\phi) \left(\frac{a}{r}\right)^{n+2} P_n^m(\cos \theta) \end{aligned} \quad (9)$$

Traditionally, the system of coordinates used in geomagnetism is the North, East, Center (sometimes denoted X , Y , Z).

Since the geomagnetic field changes in space and time, the Gauss coefficients are also time dependent. In order to model time variations of internal origin, the *secular variation*, about 1 year of continuous observations is needed. This variation is often assumed to be constant over short time scales, and can be introduced by adding a secular variation potential, V_{sv} , truncated to N_{sv}^{max} .

Magnetic Modeling, Theory, and Computation, Fig. 1 The power spectrum for an internal part of a magnetic field model: the steep part of the spectrum ($n \leq 13$) indicates the core field contribution, the transition degrees ($n = 14-15$) can be attributed to both core and lithospheric contributions, and the higher degrees ($n \geq 16$) describe the lithospheric contribution. Units are in nT^2



$$V_{sv} = a \sum_{n=1}^{N_{sv}^{max}} \left(\frac{a}{r} \right)^{n+1} \sum_{m=0}^n (t - T_0) \left(\dot{g}_n^m \cos(m\phi) + \dot{h}_n^m \sin(m\phi) \right) P_n^m(\cos(\theta)) \quad (10)$$

where $\dot{g}_n^m$ and $\dot{h}_n^m$ denote the time derivatives of the internal Gauss coefficients, T_0 the reference time (i.e., the epoch of the core field model), and t is the considered time. Recently, due to a few decades of high-resolution, high-precision satellite magnetic data, the secular acceleration can also be described, by coefficients $\ddot{g}_n^m$ and $\ddot{h}_n^m$, denoting the second time derivative of the Gauss coefficients.

The time dependence of the core field coefficients $\{g_n^m(t), h_n^m(t)\}$ up to a given degree n can also be described by temporally more complex functions, e.g., B-splines. Splines of order four result in a piecewise linear representation of $\partial_t^2 \mathbf{B}$, which is not favorable for studying sudden changes in the rate of core field changes (geomagnetic jerks). Higher-order splines result in a smoother representation of $\partial_t^2 \mathbf{B}$ and are therefore preferable. In order to be able to describe rapid field temporal variations of the core field, a 6-month knot separation is recommended.

Inverse Problem and Implementation

To obtain a geomagnetic model and thereafter the field component description, an inverse problem needs to be solved, for which the standard model estimation technique is summarized below.

The aim is to describe a given dataset in terms of a series of parameters. The data can be assembled in a data vector $\mathbf{d}$ of length N

$$\mathbf{d}^T = (\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_N). \quad (11)$$

The model is defined by a set of M parameters assembled in a model vector of length M

$$\mathbf{m}^T = (\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_M). \quad (12)$$

The basic assumption is that the data are related to the model parameters through a function $\mathbf{F}(\mathbf{m})$, leading to the relation:

$$\mathbf{d} = \mathbf{F}(\mathbf{m}), \quad (13)$$

which, for linear (or linearized) problems, reduces in a matrix form:

$$\mathbf{d} = \mathbf{A}\mathbf{m}. \quad (14)$$

where $\mathbf{A}$ is a $N \times M$ matrix of coefficients, independent of data values and model parameters. In the case of spherical

harmonics analysis, the data are often the three components of the magnetic field vector. The Gauss coefficients define the parameter vector $\mathbf{m}$. The $\mathbf{A}$ matrix elements contains M base functions evaluated at N data positions – i.e., spherical harmonics (or their gradients) and their radial dependence.

Generally, more data are available than parameters to be resolved, so $N \geq M$. In this case, the above equation cannot be satisfied for every data: the inverse problem is overdetermined and has no exact solution. Hence, the model parameters are sought such that the above equation is solved approximately. The least square method minimizes the sum of the squares of the errors in order to find a solution. The usual solution to this problem is given by:

$$\mathbf{m} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \mathbf{d} \quad (15)$$

as long as the inverse matrix $\mathbf{G}^{-1} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1}$ exists. To define the weight matrix $\mathbf{W}$, we need to consider that magnetic data from different platforms, and at different times, can be affected by different error sources and these error sources can even be correlated. The covariance matrix $\mathbf{C}_d$ associated with these errors must be estimated, inverted, and then used as weight matrix, i.e., $\mathbf{W} = \mathbf{C}_d^{-1}$. For uncorrelated data, the covariance matrix is diagonal. In practice, this is rarely the case, but for computational simplicity diagonal covariance matrices are often used.

Inverse problems are sometimes ill-posed – i.e., some parameters or combination of parameters are not resolved by data. In that case the inverse of $\mathbf{G} = (\mathbf{A}^T \mathbf{W} \mathbf{A})$ does not exist, or is numerically difficult to estimate. It shows that small errors in data values can lead to considerable changes in the parameters. The error amplification must be controlled by appropriate measures in the solution process: inverse problems can be regularized by introducing regularization matrices in space and time. A solution in the least square sense writes now as:

$$\mathbf{m} = (\mathbf{A}^T \mathbf{W} \mathbf{A} + \lambda_t \Lambda_t + \lambda_s \Lambda_s)^{-1} \mathbf{A}^T \mathbf{W} \mathbf{d}. \quad (16)$$

It corresponds to a minimum of the functional as:

$$\Phi = (\mathbf{d} - \mathbf{A}\mathbf{m})^T \mathbf{W} (\mathbf{d} - \mathbf{A}\mathbf{m}) + \lambda_t \mathbf{m}^T \Lambda_t \mathbf{m} + \lambda_s \mathbf{m}^T \Lambda_s \mathbf{m} \quad (17)$$

The choice of λ_t and λ_s damping parameter values is difficult, and often a computationally intensive task. If criteria exist to determined these parameters, they are often unpractical to establish, and therefore the parameters choice often depends on the modeler skills, interests, and a priori knowledge on the solution. The same formalism is used to impose constraints on the model solution.

The models developed are more and more complex to satisfy the high quality of the recent satellite data due to the overall progress in data acquisition techniques and quality. This has led to the development of complex codes, over several years, that use up-to-date numerical techniques.

From a computational point of view, when large datasets are used, the most time-consuming step in building a model is to compute the matrix $\mathbf{G} = (\mathbf{A}^T \mathbf{W} \mathbf{A})$. Typically, a parallel computing environment is necessary to compute this matrix in a reasonable amount of time – e.g., less than a few days. Often, a re-weighted least square approach is needed to reach the best solution, requiring the evaluation of $\mathbf{G}$ numerous times. In comparison, inverting the matrix takes only few minutes for models with 10 to 20 thousands of parameters. Overall, several months are needed by a trained modeler to derived a robust model of the Earth's interior magnetic field.

Recent Developments in Magnetic Field Modeling

Although magnetic field models obtained through the inversion process described in the previous section provide an accurate view of the present main geomagnetic field, they still strongly rely on the modeler skills and choices. Further, as the combined datasets from different satellite missions accumulate, it becomes more and more difficult to handle the large amount of data and the complexity of the models required to fit them. As a consequence, new approaches are needed. These are typically iterative, handle a small amount of data at a time, and rely mostly on Kalman filters to build their temporal evolution. In their current implementation to geomagnetic field modeling these approaches do not introduce heavy computational costs, even for the most complex models. Of course such approaches are very close to data assimilation methods. They have been introduced in geomagnetism mainly to study the core dynamics from magnetic field models (Aubert 2015; Baerenzung et al. 2018; Barrois et al. 2017; Gillet et al. 2015).

These approaches require three steps: analysis, prediction, and smoothing. The analysis step is nothing else than a Bayesian-based modeling (e.g., Gillet et al. 2013; Holschneider et al. 2016; McLeod 1996) where Gaussian distributions are used to describe the model and data error distributions. Technically, this step shows strong similarities with what has been described in the previous section. It consists of correcting a known magnetic field model such that it fits a small set of data. It writes:

$$\mathbf{m} = \tilde{\mathbf{m}} + \left(\mathbf{A}^T \mathbf{C}_d^{-1} \mathbf{A} + \mathbf{C}_m^{-1} \right)^{-1} \mathbf{A}^T \mathbf{C}_d^{-1} (\mathbf{d} - \mathbf{A} \tilde{\mathbf{m}}) \quad (18)$$

where $\tilde{\mathbf{m}}$ is the known magnetic field – i.e., the prior. It can also be a null vector. $\mathbf{C}_m$ is its covariance matrix and $\mathbf{C}_d$ is, as before, the covariance matrix of the data errors. Outside the usual difficulty in estimating $\mathbf{C}_d$, this step also requires

knowing the matrix $\mathbf{C}_m$, and inverting it. In Holschneider et al. (2016) simple correlation structures, and associated correlation matrices, for all possible magnetic sources of internal and external origins are proposed. More sophisticated matrices can be derived from theoretical model of the sources – e.g., from numerical dynamo experiment (Aubert 2015). The posterior covariance matrix of the model $\mathbf{m}$ can be estimated using:

$$\mathbf{C}_m = \left(\mathbf{A}^T \mathbf{C}_d^{-1} \mathbf{A} + \mathbf{C}_m^{-1} \right)^{-1}. \quad (19)$$

We note that when working with satellite data, the process described above is often preferred to an approach using classic Kalman gain matrices, because the number of accumulated data after a few days easily exceed the dimension of the models.

The prediction step requires introducing a multivariate random variable $\mathbf{y}$ normally distributed with a mean $\bar{\mathbf{y}}$ and covariance matrix $\mathbf{C}_y$ – i. e. $\mathbf{y} \sim \mathcal{N}(\bar{\mathbf{y}}, \mathbf{C}_y)$. Assuming that there is a known magnetic field model $\mathbf{m}_k$ and its associated covariance matrix $\mathbf{C}_{m_k}$ at time t_k , then at time t_{k+1} a prediction of the model $\tilde{\mathbf{m}}_{k+1}$ and its covariance matrix $\mathbf{C}_{\tilde{m}_{k+1}}$ are obtained through:

$$\begin{aligned} \tilde{\mathbf{m}}_{k+1} &= \mathbf{P}_k \mathbf{m}_k + \bar{\mathbf{y}} \\ \mathbf{C}_{\tilde{m}_{k+1}} &= \mathbf{P}_k \mathbf{C}_{m_k} \mathbf{P}_k^T + \mathbf{C}_y. \end{aligned} \quad (20)$$

The operator $\mathbf{P}_k$ predicts the evolution of the magnetic field model (the description is limited here to a simple linear case). This operator generally depends on the time step $t_{k+1} - t_k$ and on the characteristic time scales of the model. It can be as simple as a scaled identity matrix. The multivariate random variable $\mathbf{y}$ describes the error associated with the prediction process. So far, in most applications of this method to geomagnetic field modeling it has zero mean – i.e., $\bar{\mathbf{y}} = 0$. As $\mathbf{P}_k$, the covariance matrix $\mathbf{C}_y$ necessarily depends on the time step $t_{k+1} - t_k$ and characteristic time scales of the model, but it must also describe the correlation structure of the model. Examples of prediction process, although not necessarily linear as presented here, can be found, e.g., in Baerenzung et al. (2018), Beggan and Whaler (2009), Fournier et al. (2015), and Gillet et al. (2015).

A succession of analysis and prediction steps allows to build a sequence of magnetic field models for the required time span. The only difficulty of this sequential approach is that at a given time the model is built without knowing how the observed geomagnetic field will evolve in the future. This limitation is resolved by using the smoothing step. There are several ways to handle the smoothing and we present here a simple method that works for linear Gaussian problems (Anderson and Moore 1979). Let assume a sequence of models $\mathbf{m}_k$ and their associated covariance matrices $\mathbf{C}_{m_k}$, for

$k = 1, \dots, N$ with $t_{k+1} = t_k + \delta k$. The sequence of smoothed models $\mathbf{m}_k^s$ is derived iteratively, backward in time, starting from $\mathbf{m}_N^s = \mathbf{m}_N$. We introduce first the matrix $\mathbf{G}_k$ defined by

$$\mathbf{G}_k = \mathbf{C}_{m_k} \mathbf{P}_k' (\mathbf{P}_k \mathbf{C}_{m_k} \mathbf{P}_k' + \mathbf{C}_y)^{-1}. \quad (21)$$

The sequence of models is built with

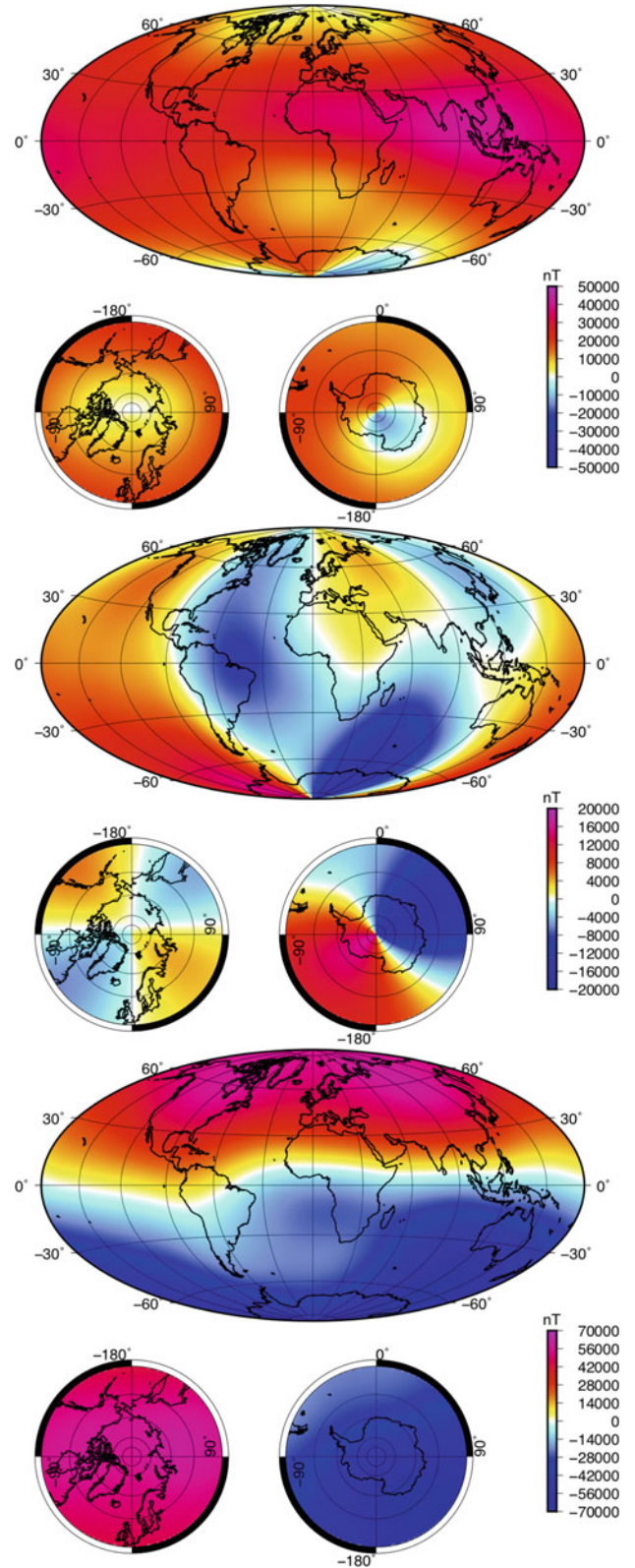
$$\begin{aligned} \mathbf{m}_k^s &= \mathbf{m}_k + \mathbf{G}_k (\mathbf{m}_{k+1}^s - \mathbf{P}_k \mathbf{m}_k) \\ \mathbf{C}_{m_k^s} &= \mathbf{C}_{m_k} - \mathbf{G}_k (\mathbf{P}_k \mathbf{C}_{m_k} \mathbf{P}_k' + \mathbf{C}_y - \mathbf{C}_{m_{k+1}^s}) \mathbf{G}_k'. \end{aligned} \quad (22)$$

To our knowledge, currently, the only magnetic field model released using specifically the technique described here is (Lesur et al. 2017), although no smoothing was applied. Several other authors have shown progress in this direction. There is no doubt that such models will be released soon, in particular for the publication of the next version of the IGRF that is due in 2020. As a new release, the model used in this study to map features of the core field and its variations at the Earth's surface and at the core–mantle boundary (CMB) is the one proposed as candidate for IGRF-13 (Ropp et al. 2020). We note that (Barrois et al. 2018) released a model established using a very similar technique, where the covariances are estimated through ensemble methods. The algorithm presented in their paper does not include a smoothing step. The setup is generally rather complex as the geomagnetic main field and the flow are co-estimated at the top of the liquid outer core.

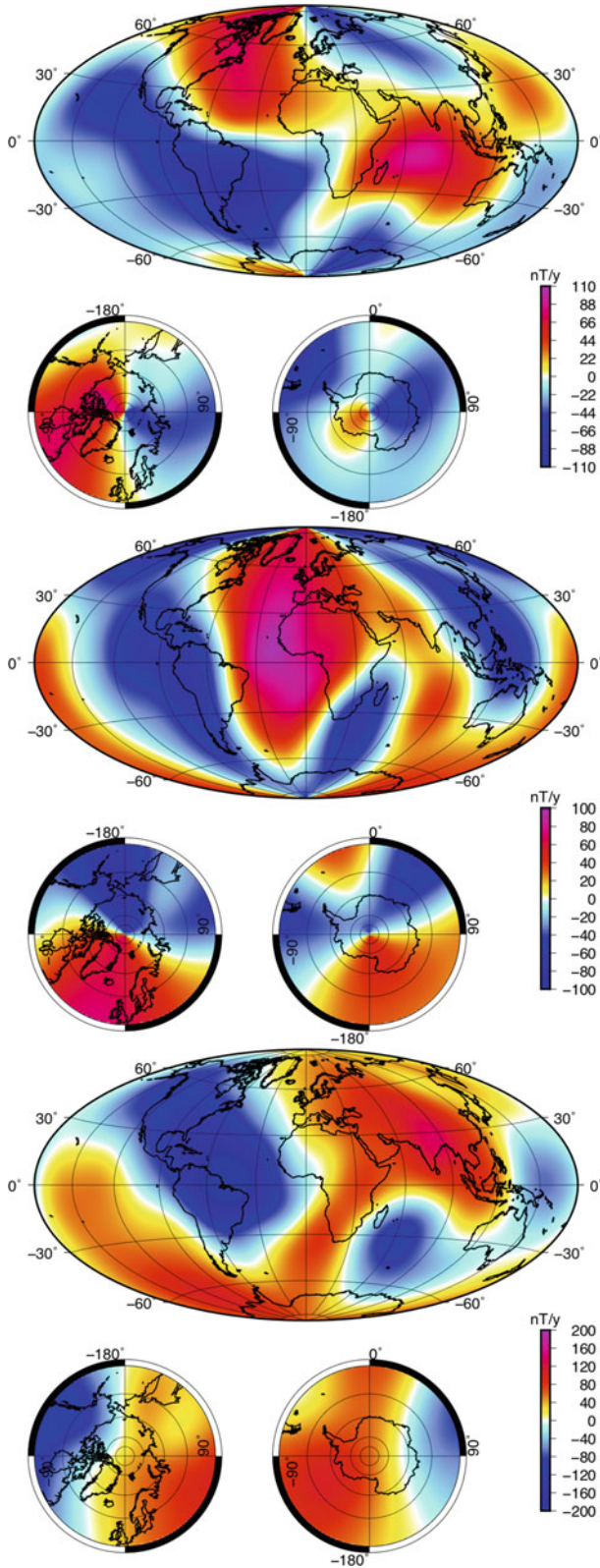
Maps of the core field and its secular variations, at the Earth's surface are plotted in Figs. 2 and 3.

Regional Field Modeling

To describe the magnetic field at regional scales, different methods have been proposed during the last decades, starting with the spherical cap harmonic analysis proposed by Haines (1985) and the related translated origin spherical cap harmonic analysis (Korte and Holme 2003). The spherical cap harmonic analysis permits the use of data from only a region of the Earth while satisfying the constraints of potential field theory. Indeed, the region to be studied has to be defined by a spherical cap and satisfies the zero curl and divergence conditions. The method is claimed to be valid over any spherical cap at any altitude above the Earth's surface. On the basis of these assertions, the method has been widely used for constructing regional magnetic maps. However, a few issues are to be taken into account when using this approach. The main one appears as a failure to correctly model the radial dependence (i.e., that the models cannot be upward and downward continued and data acquired at different altitudes



Magnetic Modeling, Theory, and Computation, Fig. 2 Maps of the core field components (units in nT): northward component (top), eastward component (middle), and radially inwards direction (bottom) at the Earth's surface in 2019. Mollweide projection is used



Magnetic Modeling, Theory, and Computation, Fig. 3 Maps of the core field secular variation (units in nT year^{-1}): northward component (top), eastward component (middle), and radially inwards direction (bottom) at the Earth's surface in 2019. Mollweide projection is used

cannot be simultaneously inverted). In order to solve this difficulty, (Thebault et al. 2004) have reposed this method as a boundary value problem within a cone extending above the reference surface, thereby allowing satellite data to be downward continued to the Earth's surface.

Some different basic functions, such as harmonic splines (Shure et al. 1982) or with more local support, wavelet-like functions (Holschneider et al. 2003; Lesur 2006) are also suited to describe datasets with variable resolution over the area of interest. Only these two approaches are summarized in the following.

We also note that maps of magnetic field at regional scales can be used together with gravity field anomalies to get information about physical properties of the Earth's crust and upper mantle (see also next section), helpful in understanding geological settings and tectonic structures. Depending on data availability, whether from the ground, airborne, or from satellites, potential field anomaly maps contain information on different ranges of spatial wavelengths, roughly corresponding to sources at different depths. One example is provided by (Korte and Mandea 2016).

Harmonic Splines

A detailed description of harmonic splines can be found in Shure et al. (1982), and a recent application for the Southern African continent magnetic data in Nahayo et al. (2018). The main interests of this representation is that it offers a smoothness property to the derived field model.

The method considers the magnetic field as expressed as the gradient of a scalar potential, as shown in the previous section. Using the spherical harmonics Y_n^m of degree n and order m , the internal part of the global magnetic field can be written as

$$B = -\nabla a \sum_{n,m}^N g_n^m \left(\frac{a}{r}\right)^{n+1} Y_n^m(\theta, \phi) \quad (23)$$

with the same notations as in 5. An alternative representation of the field is given by:

$$B(\theta, \phi, r) = -\sum_i \alpha_i \nabla F_i^{Nr}(\theta, \phi, r) \quad (24)$$

$$F_i^{Nr} = a \sum_{n,m}^N f_n(n+1) \left(\frac{a}{r}\right)^{n+1} \left(\frac{a}{r_i}\right)^{n+2} Y_n^m(\theta_i, \phi_i) Y_n^m(\theta, \phi) \quad (25)$$

$$f_n = \frac{2n+1}{4\pi n^2(n+1)^4} \quad (26)$$

with this choice of parameter f_n values, the modeled magnetic field minimizes the quantity

$$\int_{\Omega} |\nabla_h^2 B_r|^2 d\omega. \quad (27)$$

In the framework of regional analysis, in order to insure a fit to the data, it is necessary to introduce for the representation of the magnetic field two further sets of functions similar to 25. This gives:

$$B(\theta, \phi, r) = -\nabla \left(\sum_i \alpha_i^r F_i^{Nr}(\theta, \phi, r) + \sum_i \alpha_i^\theta F_i^{N\theta}(\theta, \phi, r) + \sum_i \alpha_i^\phi F_i^{N\phi}(\theta, \phi, r) \right) \\ F_i^{N\theta} = a \sum_{n,m} f_n \left(\frac{a}{r} \right)^{n+1} \left(\frac{a}{r_i} \right)^{n+2} \frac{\partial}{\partial \theta} Y_n^m(\theta_i, \phi_i) Y_n^m(\theta, \phi) \quad (28)$$

$$F_i^{N\phi} = a \sum_{n,m} f_n \left(\frac{a}{r} \right)^{n+1} \left(\frac{a}{r_i} \right)^{n+2} \frac{1}{\sin \theta} \\ \times \frac{\partial}{\partial \phi} Y_n^m(\theta_i, \phi_i) Y_n^m(\theta, \phi). \quad (29)$$

As a recent application of this approach we note the study made by Nahayo et al. (2018). The authors investigated the Southern African continent region where the geomagnetic field is weaker and changes more rapidly compared to other regions of the Earth. The CHAMP satellite data and relatively dense spaced ground-based data obtained from the repeat stations network together with permanent geomagnetic observatories are used. For the period 2005–2010 these datasets are combined in a new regional, harmonic spline-based, core field model, used to investigate the evolution of the South Atlantic anomaly in this region.

Wavelet Analysis

The wavelet frames and the work by Chambodut et al. (2005) and Holschneider et al. (2003) have pioneered the use of wavelet frames for constructing regional potential field models. In particular on the sphere, the used functions should satisfy the following properties: (i) harmonic prolongation must be easily computable; (ii) function itself has to be numerically easy to compute; and (iii) functions must be localized on the sphere.

The indicated requirements lead to constructions based on the Poisson kernel of spherical functions. A Poisson wavelet at a given point on the unit sphere is expressed as:

$$\psi_a^k(x) = N_a^k \sum_n (an)^k e^{-an} Q_n(\cos \mu) \quad (30)$$

where N_a^k is a normalization constant and μ is the angle between the wavelet position and the data point position. The Q_n function is defined by

$$Q_n(\cos \mu) = (2n+1)P_n(\cos \mu), \quad (31)$$

which can be expressed easily in terms of products of spherical harmonics.

The wavelet is defined by three parameters: scale a , position, and order k . The scale a is a measure of the wavelength of the wavelet without dimension, and order k characterizes the global shape of the wavelets (number of oscillations). The wavelets on the sphere may also be considered as a restriction on the sphere of a function that is harmonic outside and that has singularities inside it. The main characteristic of a wavelet is its relatively well-localized amplitude on the sphere, hence well adapted to regional modeling, while spanning a relatively limited range of wavelengths.

One application of this approach concerns the separation of the internal and external contributions in vector magnetic data, as those provided by CHAMP and Swarm satellites. A pioneering work in this domain is found in Mayer and Maier (2006) where it is shown that wavelets technique is especially applicable for modeling and analyzing small spatial effects like the geomagnetic field originating in the lithosphere.

Interpretation

The geomagnetic field models, at global and regional scales, play an important role in our modern world. The Earth's magnetic field can be used as a tool in screening our planet, and for understanding its deep interior processes. Moreover, these models can be used for some specific applications, as in navigation, to protect the technological systems affected by large geomagnetically induced currents, in satellite operations, or exploration geophysics. From this list, which is far for being complete, three case studies are given here, one linked to interpretation of the magnetic field and its secular variation at the core–mantle boundary, and two linked to the lithospheric field.

Core Magnetic Field at Core–Mantle Boundary

A better description of the short temporal and spatial variability of the internal field would provide insights on the processes generating the internal field and the dynamical regime in Earth's liquid outer core. The core field changes on temporal scales from months to millennia, even longer. For temporal scales covering the satellite era the lithospheric field is

considered static. Indeed, recently, (Sebera et al. 2019) shows that the effect of the time-varying lithospheric field over 17 years (since the launch of CHAMP) is at the edge of the detectability reaching peak-to-peak values about 0.5 nT depending on the area and the altitude.

As described above, models of the magnetic field generated inside the Earth are dominated for the long wavelengths by the core field, and for the shorter wavelengths by the lithospheric field. The longest wavelengths of these models – roughly up to spherical harmonics of degree 14 – can therefore be used to try to understand the generation process of the core field. Results from seismological observations have shown that the core can be separated in two regions; a solid inner core of radius estimated at 1220 km and a liquid outer core in the spherical shell extending up to 3485 km. This liquid outer core is very conductive and made mainly of liquid iron in convection, which in presence of a magnetic field, can regenerate this field, balancing its natural decay due to ohmic dissipation. This process is known as the geodynamo. The best way to retrieve information about the dynamics of the core and the fluid motions at the top of the core is through magnetic data acquired at, or above, the Earth's surface. The magnetic field models are therefore downward continued, through the mantle, to the core–mantle boundary (CMB) assuming that the mantle is an insulator. In principle, only the radial component of the magnetic field can be derived just below the CMB, because of the unknown jump of the electrical conductivity across the CMB, associated with a discontinuity of the magnetic field components tangential to the boundary.

An image of this vertical down component of the core magnetic field just below the CMB is shown in Fig. 4, together with its expected time derivative. These images correspond to a magnetic field model truncated to spherical harmonic degree 13. Contributions from shorter wavelengths may considerably modify these images, but these contributions are unknown, being hidden by the lithospheric field or the noise, at the Earth's surface. However, considerable information can be readily extracted from these maps. One can see that the core field has a much more complex structure at the CMB than on the Earth's surface (Fig. 2). In particular, the presence in the southern hemisphere of a large patch of reverse orientation of the magnetic field – the field there is oriented towards the core interior, whereas in a simple dipole approximation one would expect the field to be oriented outward. This patch and its evolution has attracted scientist's attention over the last decades because it is associated with the weakness of the Earth's field in the South-Atlantic region, it contributes significantly in the decrease in the strength of the Earth's dipole field component, and it has sometimes been thought as an early stage of a reversal process (Hulot et al. 2002; Olsen and Manda 2008). Regarding the secular variation, it is obviously concentrated at mid latitudes and high

latitudes of the north hemisphere. It is extremely weak under the Pacific ocean and southern part of the south hemisphere.

The combination of the Maxwell equations leads to the so-called induction equation. Under the frozen-flux approximation, that neglects the diffusion in front of induction, the radial component of this equation writes:

$$\frac{\partial}{\partial t} B_r = -\nabla_h \cdot (B_r U_h) \quad (32)$$

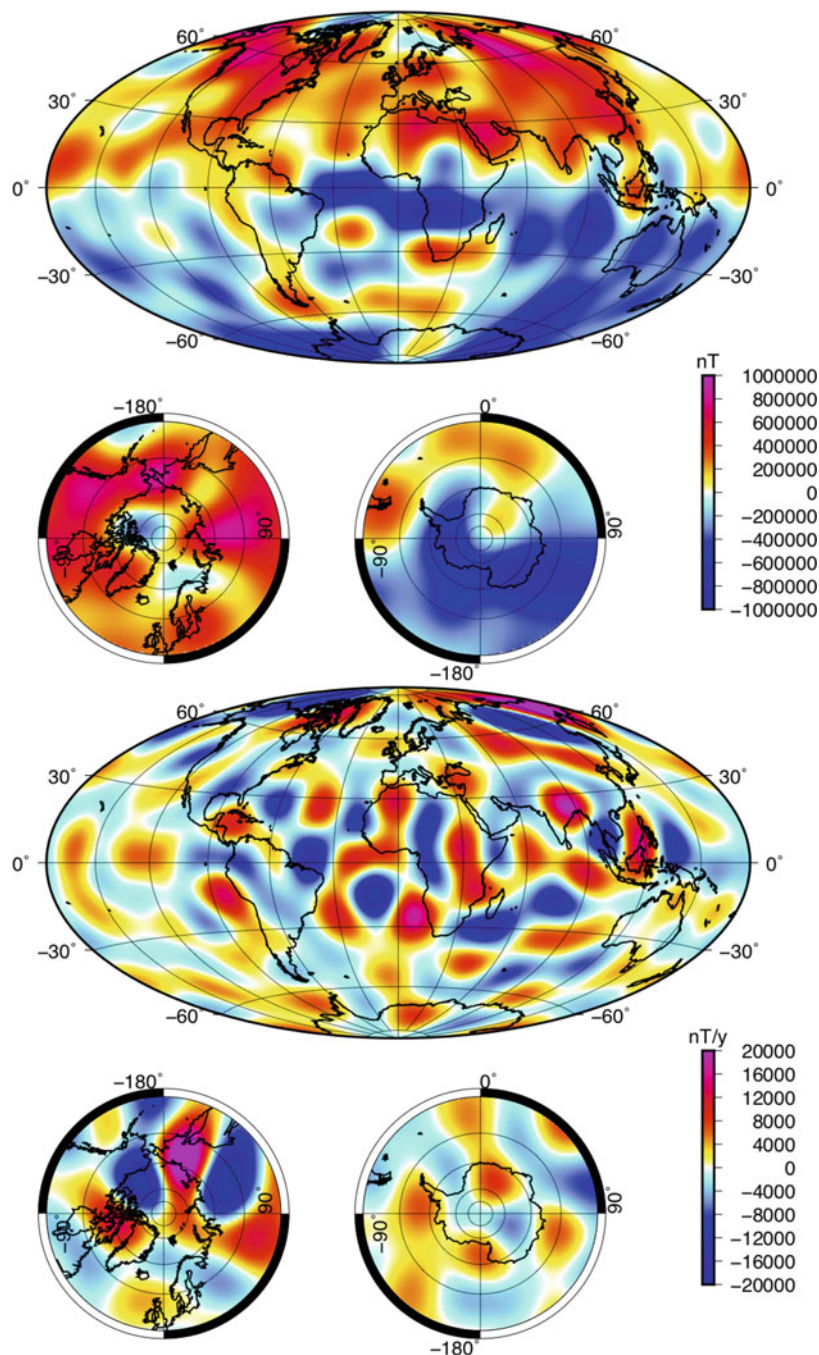
where B_r is the radial component of the magnetic field, U_h is the horizontal component of the flow under the CMB, and ∇_h is the tangential gradient on the sphere. This equation simply shows that the main source of secular variation is the advection of the field lines by the flow. Since radial component of both the field and its secular variation can be estimated at the CMB, this equation provides a way to get some information on the flow at the top of the liquid outer core, for large spatial wavelength and relatively short time scales. By combining this with constraints derived from the equations of fluid dynamics and/or magnetohydrodynamics, it is possible to derive physical properties of the outer core and shed light on unexplained phenomenon as geomagnetic jerks, (Manda et al. 2010) or contributions to the variation of the length of the day.

Lithospheric Magnetic Field over the Arctic and African Regions

The magnetic signature of the lithosphere is related to all the processes that affected and shaped the figure of our planet. To first order, the remanent magnetic contributions are canceled out at large wavelength and the induced part dominates above continental domains. The exact contribution of the remanent magnetization is actually difficult to estimate, unless the physical properties of the magnetic sources are known through rock sampling and analysis. It is possible to estimate what is their relative contribution by using global forward approaches, mostly relying on the distribution of the vertically integrated magnetization. A simple estimate is based on the lateral magnetization contrast associated with the boundary between oceanic and continental domains. Indeed, the abrupt boundaries between the thin oceanic crust and the thicker continental crust translate into a large-scale magnetic signal which is present above the continental margins, and also above continental and oceanic basins. This does not suffice to explain all the observed magnetic anomalies. Those interested in the crustal magnetic field are referred to some reviews, like Langlais et al. 2009.

Generally speaking, the lithospheric magnetic field reflects the magnetization associated with minerals cooling down below their Curie temperature in the presence of a magnetic field. Above the Curie temperature, the magnetization vanishes, and minerals exhibit then paramagnetic susceptibility,

Magnetic Modeling, Theory, and Computation, Fig. 4 Map of the core field vertical down component (units in nT, top) and its secular variation (units in nT year⁻¹, bottom) at the core–mantle boundary, for 2019. Mollweide projection is used



with little effect compared to magnetization. Therefore, rocks are essentially nonmagnetic at temperatures greater than the Curie temperature characterizing their main magnetic minerals constituents. Taking into account a normal geothermal gradient, sources of the measured magnetic signal are then restricted to some maximum 30–40 km depth, excepting the old cratonic areas where the Curie depth – depth at which the temperature reach the Curie temperature – may be deeper. Different methods are proposed to estimating the depth to the

bottom of magnetic sources from magnetic data and based on the spectral analysis of magnetic anomalies (Blakely 1995; Bouligand et al. 2009; Spector and Grant 1970). A recent study (Audet and Gosselin 2019) concluded that the estimated Curie depth values are rather unreliable unless prior information on one or all model parameters is included in calculations. Curie depth estimates where the lithosphere is thinner or the magnetization is likely to be uniform are more reliable than in regions with thick and heterogeneous lithosphere.

Satellite data is of particular importance for unraveling the origin and structure of large lithospheric domains, which are difficult to access and map, like the polar regions or the African continent.

The Arctic region. In the last two decades, the Arctic region became a frontier for scientific discoveries and exploration studies. All circum-Arctic nations embarked on a quest for collecting new data and deciphering the structure and evolution of the Arctic crust and lithosphere. Along with bathymetric, gravimetric, and seismological studies, acquisition, processing, and interpretation of magnetic data are vital components for the understanding of lithospheric architecture. Modern magnetic data was collected and used in regional studies (e.g., Dossing et al. 2013), and to date, the circum-Arctic magnetic anomaly map (CAMP-GM) (Gaina et al. 2010, 2011), map based on airborne magnetic data, is still the only high-resolution magnetic data compilation available for the Arctic region (defined here as the region north of 60°). Updated World Digital Magnetic Anomaly maps incorporate an upward continuation (to 5 km) of the CAMP-GM compilation, or are using vintage, lower-resolution North Atlantic–Arctic magnetic anomaly maps (Lesur et al. 2016). Meanwhile, new global lithospheric models have been constructed based on satellite data. Due to the inaccessibility and harsh climate of this area, satellite data plays a crucial role for the Arctic science and examples of lithospheric models and their geological interpretation are shown below.

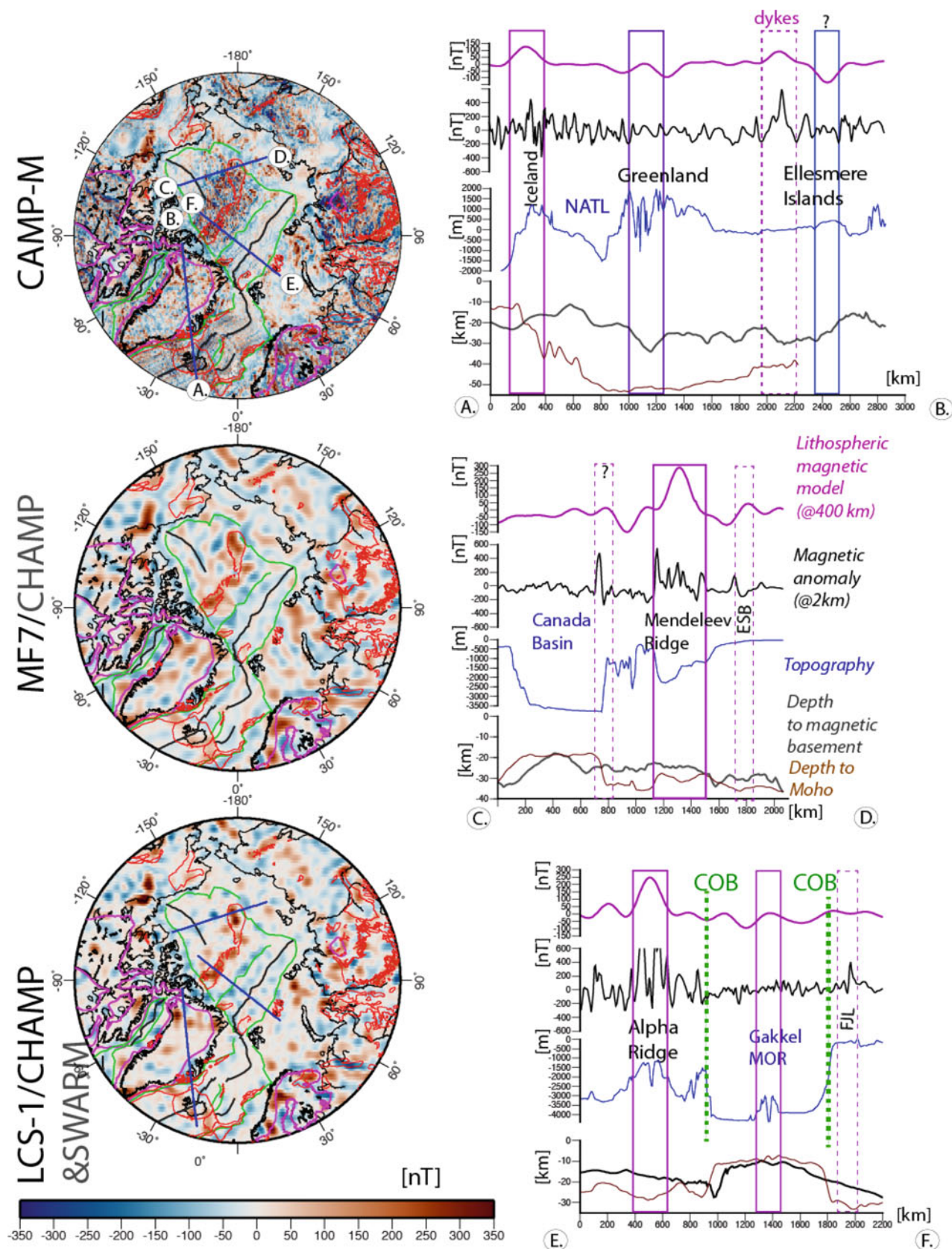
Figure 5 shows maps of magnetic anomalies recorded closer to Earth's surface (CAMP-GM compilation, Gaina et al. 2011), which display the magnetic signature of bodies in the Earth's crust, and magnetic anomalies measured from satellites (at 350 km from surface, as illustrated by the MF7 model (Maus et al. 2007), based on CHAMP satellite data, and at 460 km, shown by the LCS-1 model, based on CHAMP and Swarm data, (Olsen et al. 2017)), which image thicker than normal and/or deeply buried magnetic bodies. By upward continuing the magnetic data collected closer to the surface (either shipborne or airborne), a first-order estimation of shallow versus deep-magnetized bodies can be made.

Three profiles across Arctic continental and oceanic domains show how airborne and satellite magnetic data may help in identifying the nature and to estimate the thickness of main tectonic bodies (Fig. 5). Short wavelengths of airborne magnetic data detect magnetic anomalies due to seafloor spreading and dyke intrusions, and can be used together with other geophysical data (e.g., gravity or seismic data) to identify the transition between continental and oceanic lithosphere (Fig. 5). Seafloor spreading anomalies were previously identified in satellite magnetic

lithospheric models (see Maus et al. 2008), but the ultraslow spreading ridge in the High Arctic, the Gakkel Mid-Ocean Ridge, is imaged for the first time by the combined CHAMP and Swarm magnetic data (model LCS-1, Olsen et al. 2017), (Fig. 5). The oldest continental crust, usually found in the cratonic areas and as Proterozoic accreted crust, generates large positive magnetic anomalies, well visible in satellite lithospheric magnetic data. This crust contains large and deep volcanic bodies in the North American shield, Greenland, the Baltic shield in Eurasia, and the Siberian platform in NE Asia, which are now imaged in greater detail by the LCS-1 model (Fig. 5). One of the most striking features in the Arctic domains is the strong magnetic anomaly close to the North Pole that correlates with a large, igneous oceanic plateau called the Alpha Mendeleev ridge. The age and exact origin of this large igneous province is not known, although numerous studies speculate on an episode of Cretaceous volcanism that affected an old oceanic crust or a partially continental crust. This composite plateau generates the highest magnetic anomaly in the Arctic region (Fig. 5). The intensity and extent of the magnetic anomalies recorded by aircrafts or satellites point toward a very thick, volcanic crust. Magnetic anomaly data can be used to compute the depth to the magnetic basement, and in the case of Arctic large igneous provinces (LIPS) this can be as deep as 30 km (Li et al. 2017) (Fig. 5). Note that the results depend on the quality of input data and methodology. Although the new lithospheric magnetic model LCS-1 (Olsen et al. 2017) helps to better determine the overall location and extent of large magnetic bodies, only deep-sea drilling will be able to reveal the true nature of the underlying crust at the core of the Arctic.

The African continent. The African continent preserves a diversity of tectonic provinces, from ancient Archaean cratons and numerous orogenic belts to recent rifted areas like the East African rift system (e.g., Bouysse et al. 2010). However, comprehensive regional tectonic interpretations are hindered by the fact that large areas of the African continent are not accessible or simply not mapped in detail; considerable data gaps still persist in recent global compilations, including ground and airborne magnetic data (e.g., EMAG2v3 by Meyer et al. (2017), Fig. 6).

Higher-resolution lithospheric models based on satellite data (e.g., LCS-1 model, based on CHAMP and Swarm data, (Olsen et al. 2017)), shown in Fig. 6, provide continuous coverage and can image the regional tectonic domains with magmatic and metamorphic histories. High values of magnetic anomaly intensity displayed by the LCS-1 model are mostly associated with magmatic provinces (including LIPS, red outlines in Fig. 6 both on land and in oceanic domains). They also characterize the areas



Magnetic Modeling, Theory, and Computation,
Fig. 5 Lithospheric magnetic grids for the circum-Arctic region:

CAMP-GM ((Gaina et al. 2011) based on airborne data compilation upward continued at 2 km (upper left map)); MF7 ((Maus et al. 2007)

where old orogenic belts have been interpreted based on scattered geological evidences (as portrayed by the World Geological Map, e.g., Bouysse et al. 2010, dark grey polygons in Fig. 6). Some of these orogenic belts are in areas poorly or unmapped by geophysical data, like the Congo region in the central Africa (Fig. 6). The novel lithospheric models based on satellite data are therefore important tools to complement the geological models, and they increase our potential to better understand the spatial lithospheric heterogeneities and their role in tectonic evolution.

Conclusions

To model and interpret the Earth's magnetic field, two ingredients are crucial: global high-accuracy, high-resolution data and reliable mathematical tools.

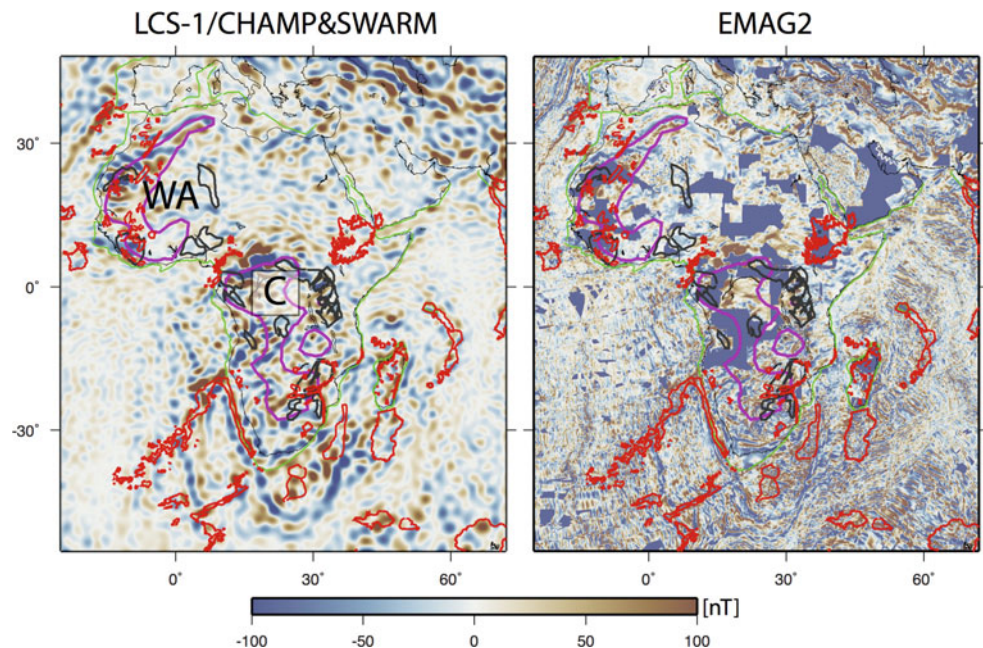
Since 2013, a trio of satellites in the frame of the ESA Swarm mission provides a unique view inside the Earth. This

mission offers the most detailed data yet on the geomagnetic field of the Earth and its temporal evolution, giving new insights into improving our knowledge of the Earth's interior, thus the core and crustal fields. Their quasi-polar orbit allows to cover nearly entire Earth's surface; however, gaps still exist over the polar regions, inducing uncertainties on zonal coefficients. Another difficulty comes from describing the slow evolution of the external fields which limits the description of rapidly variations of the core fields.

These measurements together with those provided by magnetic observatories are valuable data essential for modeling the geomagnetic field and its interaction with other physical aspects of the Earth's system. Separating the signature of the different sources in both satellite and ground-based data can be done with a subtle weighting scheme of the many parameters entering this inverse problem, as shown here. The need of both high-resolution and complete datasets and continuously improved methods will allow us to better describe the spatial and temporal variations of geomagnetic internal field.

Magnetic Modeling, Theory, and Computation,

Fig. 6 Lithospheric magnetic grids for the African continent: Left – LCS-1 (Olsen et al. 2017) based on CHAMP and Swarm satellite data, Right – EMAG2v3 (Meyer et al. 2017) based on ground and airborne data compilation. Outlines of cratonic regions are shown in magenta (modified after (Celli et al. 2020), Archean and Proterozoic magmatic and metamorphic units (modified after (Bouysse et al. 2010)) are shown in dark grey, large igneous provinces (LIPS, after (Johansson et al. 2018)) in red, and continent–ocean boundaries (after (Müller et al. 2016)) in green. Abbreviations: *WA* West Africa, *C* Congo



Magnetic Modeling, Theory, and Computation, Fig. 5 (continued) based on CHAMP satellite data (middle left map)); and LCS-1 ((Olsen et al. 2017) based on CHAMP and Swarm satellite data (lower left map)). Outlines of cratonic regions are shown in magenta (Bouysse et al. 2010), large igneous provinces (LIPS, Johansson et al. 2018) in red, and continent–ocean boundaries (Müller et al. 2016) in green. Three profiles through the circum-Arctic lithosphere (location on upper and lower left maps) display magnetic anomalies from the satellite and airborne data,

and compare the Curie depth (Li et al. 2017) estimated from magnetic data with depth to Moho based on gravity data (Lebedeva-Ivanova et al. 2019). A tentative interpretation of the most prominent magnetic anomalies observed in the satellite and airborne data is shown by vertical boxes of different colors. Abbreviations: *COB* continent–ocean boundary, *NATL* North Atlantic ocean, *ESB* East Siberian Basin, *FJL* Franz Josef Land

Acknowledgments M.M. and C.G. acknowledge support from ESA through the project “3D Earth – A Dynamic Living Planet.” C.G. acknowledges the support from the Research Council of Norway through its Centers of Excellence funding scheme, project 223272. We thank the Section Editor, Kusumita Arora, for the editorial work. MM emphasizes that the research leading to these results has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (GRACE-FUL Synergy Grant agreement No 855677).

Bibliography

- Anderson B, Moore J (1979) Optimal filtering. Information and system sciences series. Prentice-Hall, Englewood Cliffs, ISBN 0-13-638122-7
- Aubert J (2015) Geomagnetic forecasts driven by thermal wind dynamics in the Earth’s core. *Geophys J Int* 203:1738–1751
- Audet P, Gosselin JM (2019) Curie depth estimation from magnetic anomaly data: a reassessment using multitaper spectral analysis and Bayesian inference. *Geophys J Int* 218:494–507
- Backus G, Parker R, Constable C (1996) Foundations of geomagnetism. Cambridge University Press, New York
- Baerenzung J, Holschneider M, Wicht J, Sanchez S, Lesur V (2018) Modeling and predicting the short term evolution of the geomagnetic field. *J Geophys Res* 123:4539–4560
- Barrois O, Gillet N, Aubert J (2017) Contributions to the geomagnetic secular variation from a reanalysis of core surface dynamics. *Geophys J Int* 211(1):50–68
- Barrois O, Hammer MD, Finlay CC, Martin Y, Gillet N (2018) Assimilation of ground and satellite magnetic measurements: inference of core surface magnetic and velocity field changes. *Geophys J Int* 215(1):695–712
- Beggan CD, Whaler K (2009) Forecasting change of the magnetic field using core surface flows and ensemble Kalman filtering. *Geophys Res Lett* 36(18):L18303
- Blakely R (1995) Potential theory in gravity and magnetic applications. Cambridge University Press, Cambridge
- Bouligand C, Glen JMG, Blakely RJ (2009) Mapping Curie temperature depth in the western United States with a fractal model for crustal magnetization. *J Geophys Res* 114:2169–9313
- Bouysson P, Ségoufin J, Jepsen HF, Pulvertaft TCR, Gaina C, Gernigon L, Grikurov G, Leitchenkov G, Gohl K, Pubellier M, Shokalsky S, Lagabrielle Y, Jolivet L, Iturralde-Vinent M, Chamot-Rooke N, Rabaute A (2010) Geological map of the world, 3rd edn. Main map in Mercator projection and Polar regions in polar stereographic projection, scale: 1/25.000.000 (242×108.5 cm). CGMW and UNESCO – Commission for the Geological Map of the World
- Celli N, Lebedev S, Schaeffer A, Gaina C (2020) African cratonic lithosphere carved by mantle plumes. *Nat Commun*. <https://doi.org/10.1038/s41467-019-13871-2>
- Chambodut A, Panet I, Mandeau M, Diamant M, Jamet O, Holschneider M (2005) Wavelet frames: an alternative to the spherical harmonics representation of potential fields? *Geophys J Int* 163:875–899. <https://doi.org/10.1111/j.1365-246X.2005.02754.x>
- Dossing A, Jackson HR, Matzka J, Einarsson I, Rasmussen TM, Olesen AV, Brozena JM (2013) On the origin of the Amerasia Basin and the High Arctic Large Igneous Province—results of new aeromagnetic data. *Earth Planet Sci Lett* 363:219–230
- Finlay CC, Olsen N, Kotsiaros S, Gillet N, Tøffner-Clausen L (2016) Recent geomagnetic secular variation from Swarm and ground observatories as estimated in the CHAOS-6 geomagnetic field model. *Earth Planets Space* 68:112
- Finlay CC, Lesur V, Thebaud E, Vervelidou F, Morschhauser A, Shore R (2017) Challenges handling magnetospheric and ionospheric signals in internal geomagnetic field modelling. *Space Sci Rev* 206:157–189
- Fournier A, Auber J, Thébaud E (2015) A candidate secular variation model for IGRF-12 based on swarm data and inverse geodynamo modelling. *Earth Planets Space* 67:81
- Gaina C, Werner SC, Mandeau M, The CAMP-GM Group (2010) Circum-Arctic magnetic and gravity maps, Circum-Arctic geophysical maps (gravity and magnetic anomalies), scale 1:5000000. CCGM/CCWM, Paris
- Gaina C, Werner SC, Saltus R, Maus S, CAMP-GM Group (2011) Circum-Arctic mapping project: new magnetic and gravity anomaly maps of the Arctic. *Geol Soc Mem* 35. <https://doi.org/10.1144/M35.3>
- Gillet N, Jault D, Finlay CC, Olsen N (2013) Stochastic modelling of the Earth’s magnetic field: inversion for covariances over the observatory era. *Geophys Geochem Geosyst* 14:766–786
- Gillet N, Barrois O, Finlay CC (2015) Stochastic forecasting of the geomagnetic field from the cov-obs.xl geomagnetic field model, and candidate models for IGRF-12. *Earth Planets Space* 67(1):1–14
- Haines GV (1985) Spherical cap harmonic analysis. *J Geophys Res* 90:2583–2591
- Holschneider M, Chambodut A, Mandeau M (2003) From global to regional analysis of the magnetic field on the sphere using wavelet frames. *Phys Earth Planet Inter* 135(2):107–124
- Holschneider M, Lesur V, Mauerberger S, Baerenzung J (2016) Correlation based modelling and separation of geomagnetic field components. *J Geophys Res* 121:3142–3160
- Hulot G, Eymin C, Langlais B, Mandeau M, Olsen N (2002) Small-scale structure dynamics of the geodynamo inferred from Ørsted and MAGSAT satellite data. *Nature* 416:620–623
- Johansson L, Zahirovic S, Müller RD (2018) The interplay between the eruption and weathering of Large Igneous Provinces and the deep-time carbon cycle. *Geophys Res Lett* 45:5380–5389. <https://doi.org/10.1029/2017GL076691>
- Jones C (2015) Thermal and compositional convection in the outer core. In: Treatise in geophysics, 2nd edn. Core dynamics, vol 8. Elsevier, Amsterdam, pp 115–159
- Korte M, Holme R (2003) Regularization of spherical cap harmonics. *Geophys J Int* 153(1):253–262
- Korte M, Mandeau M (2016) Geopotential field anomalies and regional tectonic features – two case studies: southern Africa and Germany, geopotential field anomalies and regional tectonic features – two case studies: southern Africa and Germany. *Solid Earth* 7:751–768. <https://doi.org/10.5194/se-7-751-2016>
- Langlais B, Lesur V, Purucker ME, Connerney JEP, Mandeau M (2009) Crustal magnetic fields of terrestrial planets. *Space Sci Rev* 152:22. <https://doi.org/10.1007/s11214-009-9557-y>
- Lebedeva-Ivanova N, Gaina C, Minakov A, Kashubin S (2019) ArcCRUST: Arctic crustal thickness from 3D gravity inversion. *Geophys Geochem Geosyst*. <https://doi.org/10.1029/2018GC008098>
- Lesur V (2006) Introducing localized constraints in global geomagnetic field modelling. *Earth Planets Space* 58:477–483
- Lesur V, Wardinski I, Hamoudi M, Rother M (2010) The second generation of the GFZ reference internal magnetic model: GRIMM-2. *Earth Planets Space* 62:765–773
- Lesur V, Hamoudi M, Choi Y, Dyment J, Thebaud E (2016) Building the second version of the World Digital Magnetic Anomaly Map (WDMAM). *Earth Planets Space* 68:27. <https://doi.org/10.1186/s40623-016-0404-6>
- Lesur V, Wardinski I, Baerenzung J, Holschneider M (2017) On the frequency spectra of the core magnetic field Gauss coefficients. *Phys Earth Planet Inter*. <https://doi.org/10.1016/j.pepi.2017.05.017>
- Li CF, Lu Y, Wang YJ (2017) A global reference model of Curie-point depths based on EMAG2. *Sci Rep* 7:2045–2322
- Lowes FJ (1966) Mean-square values on sphere of spherical harmonic vector fields. *J Geophys Res* 71:2179

- Mandea M, Holme R, Pais A, Pinheiro K, Jackson A, Verbanac G (2010) Geomagnetic jerks: rapid core field variations and core dynamics. *Space Sci Rev* 155:147–175. <https://doi.org/10.1007/s11214-010-9663-x>
- Mandea M, Purucker M (2018) The varying core magnetic field from a space weather perspective. *Space Sci Rev*. <https://doi.org/10.1007/s11214-017-0443-8>
- Maus S, Luhr H, Rother M, Hemant K, Balasis G, Ritter P, Stolle C (2007) Fifth-generation lithospheric magnetic field model from CHAMP satellite measurements. *Geophys Geochem Geosyst* 8:1525–2027
- Maus S, Yin F, Luhr H, Manoj C, Rother M, Rauberg J, Michaelis I, Stolle C, Muller R (2008) Resolution of direction of oceanic magnetic lineations by the sixth-generation lithospheric magnetic field model from CHAMP satellite magnetic measurements. *Geophys Geochem Geosyst* 9:1525–2027
- Maus S, Barckhausen U, Berkenbosch H et al (2009) EMAG2: a 2-arc min resolution Earth Magnetic Anomaly Grid compiled from satellite, airborne, and marine magnetic measurements. *Geophys Geochem Geosyst* 10:1525–2027
- McLeod MG (1996) Spatial and temporal power spectra of the geomagnetic field. *J Geophys Res* 101:2745–2763
- Merrill RT, McElhinny MW, McFadden PL (1996) *The magnetic field of the Earth*. Academic Press, London
- Mayer C, Maier T (2006) Separating inner and outer Earth's magnetic field from CHAMP satellite measurements by means of vector scaling functions and wavelets. *Geophys J Int* 167(3):1188–1203
- Meyer B, Chulliat A, Saltus R (2017) Derivation and error analysis of the Earth Magnetic Anomaly Grid at 2 arc min resolution version 3 (EMAG2v3). *Geophys Geochem Geosyst* 18:4522–4537
- Müller RD, Seton M, Zahirovic S, Williams SE, Matthews KJ, Wright NM, Shephard GE, Maloney KT, Barnett-Moore N, Hosseinpour M, Bower DJ, Cannon J (2016) Ocean Basin evolution and global-scale plate reorganization events since Pangea breakup. *Annu Rev Earth Planet Sci* 44:107–138
- Nahayo E, Kotze P, Korte M, Webb S (2018) A harmonic spline magnetic main field model for Southern Africa combining ground and satellite data to describe the evolution of the South Atlantic Anomaly in this region between 2005 and 2010. *Earth Planets Space* 70:30
- Olsen N, Mandea M (2008) Rapidly changing flows in the Earth's core. *Nat Geosci*. <https://doi.org/10.1038/ngeo203>
- Olsen N, Ravat D, Finlay CC, Kother LK (2017) LCS-1: a high-resolution global model of the lithospheric magnetic field derived from CHAMP and Swarm satellite observations. *Geophys J Int* 211:1461–1477
- Ropp G, Lesur V, Baerenzung J, Holschneider M (2020) Sequential modelling of the Earth's core magnetic field. *Earth Planets Space* (in Press). <https://doi.org/10.1186/s40623-020-01230-1>
- Sabaka TJ, Olsen N, Tyler RH, Kuvshinov A (2015) CM5, a pre-Swarm comprehensive geomagnetic field model derived from over 12 yr of CHAMP, Ørsted, SAC-C and observatory data. *Geophys J Int* 200:1596–1626
- Schaeffer N, Jault D, Nataf H-C, Fournier A (2017) Turbulent geodynamo simulations: a leap towards Earth's core. *Geophys J Int* 211(1):1–29
- Sebera J, Haagmans R, Bakytiev E, Bezdek A (2019) On the observability of the time-variable lithospheric signal in satellite magnetic data. *Surv Geophys* 40:1229. <https://doi.org/10.1007/s10712-019-09546-8>
- Shure L, Parker R, Backus G (1982) Harmonic splines for geomagnetic modelling. *Phys Earth Planet Inter* 28:215–229. [https://doi.org/10.1016/0031-9201\(82\)90003-6](https://doi.org/10.1016/0031-9201(82)90003-6)
- Spector A, Grant FS (1970) Statistical models for interpreting aeromagnetic data. *Geophysics* 35:293
- Thebault E, Schott JJ, Mandea M, Hoffbeck JP (2004) A new proposal for spherical cap harmonic modelling. *Geophys J Int* 159:83–103. <https://doi.org/10.1111/j.1365-246X.2004.02361.x>

Magnetometers

Ivan Hrvoic

GEM Systems, Inc, Markham, ON, Canada

Synonyms

Geomagnetic instruments

Definition

Magnetometers are devices/instruments that measure magnetic field/magnetic flux density in particular Earth's magnetic field either vectorially or scalarly.

Introduction

First measurements of magnetic field intensity have been introduced by Carl Friedrich Gauss (Kaufman et al. 2009) in 1834. He has established a number of magnetic observatories to continuously measure magnetic field of Earth (see ► “[Geomagnetic Field, Measurement Techniques](#)” and ► “[Geomagnetic Field, Secular Variation](#)”). Prior to that, compasses were in use for several centuries mostly for navigation but with no knowledge of the causes of magnetic needle behavior. Since Gauss' times, many methods of magnetic field measurement have been developed but there are two pillars of magnetometry discovered and developed in the twentieth century:

1. Fluxgate magnetometers for measurement of components of the vector of magnetic field (Primdahl 1979; Korepanov et al. 2007).
2. Scalar, quantum magnetometers with high precision of measurement and a possibility of measurements in motion (Abraham 1961; Alexandrov and Bonch-Bruевич 1992; Hrvoic 2004, 2008)

Presently, absolute measurements of magnetic field direction by optical methods (DI-Flux or DIM) based on fluxgate technology and the scalar measurements (Proton precession, Overhauser or Potassium) of its magnitude set up the limits of precision of the magnetic field vector measurement.

Magnetometry is a mature science/technology with myriad of applications. The need for better, more precise, smaller magnetometers fuels continuous research in the field.

This review will be treating separately vector and scalar magnetometers describing their principles of operation and main features, and with the view of their application in different fields of geophysical exploration.

Scalar magnetometers with their high precision and absolute accuracy now dominate measurements in mineral and oil exploration, volcanology, ordnance detection, archeology and partially magnetic observatories while vector magnetometer's role is limited to magnetic observatories, partially space measurement (Acuña 2002) (see ► “Equatorial Electrojet” and ► “Magnetic Methods, Principles”), and in vertical gradiometers for archeological research (see ► “Archaeomagnetism”). Scalar magnetometers are used to calibrate fluxgates. Some experiments with tensor measurements also include fluxgates and/or SQUIDs (Supercooled Quantum Interference Devices) (Clark 1993).

All modern magnetometers are computerized instruments with nonvolatile memory for display, storage, and review of data.

Scalar (Quantum) Magnetometers

Advent of scalar magnetometers in the second part of the twentieth century made a substantial contribution to measurement of both vector and scalar values of magnetic field. They offer precession frequency as a measure of magnetic field. Scalar magnetometer's sensitivities are determined only by achievable quality of precession signal, value of gyromagnetic constant, width of the spectral line (or time of decay of the precession signal), and signal-to-noise ratio. Scalar magnetometers have high sensitivity and accuracy of readings, virtually no drift with temperature or time. Measurements depend very weakly on sensor orientation or movement allowing for measurement in motion.

Proton, Overhauser, and Alkali metal (optically pumped) magnetometers are now overwhelmingly used in mineral/diamond/oil exploration, weapons detection, volcanology, archeology, magnetic observatories. Most of the current research is done in variations of optical pumping, and the progress in the last 25–30 years is phenomenal. Sensitivities have been improved from some nT to fractions of pT , perhaps four orders of magnitude, far exceeding requirements for ground (see ► “Magnetic Methods, Surface”), airborne (see ► “Magnetic Methods, Airborne”), or marine surveys. Speed of readings increased from perhaps once per second to tens and hundreds of readings per second limited only by increasing noise and/or possibility to usefully store a flood of data. With the development of instrumentation, the use of gradiometers with two or more sensors has increased (see ► “Magnetic Gradiometry”). It improves the quality of surveys, determination of depth of the anomaly producing body, following the direction and depth of the buried pipelines or electrical power lines etc.

Scalar magnetometer consists of a sensor, separated by a cable from electronics (to avoid its stray magnetic fields). Electronics has an analogue part that generates precession



Magnetometers, Fig. 1 Overhauser magnetometer – sensor, cable and electronics console

signal from the sensor and digital microprocessor based on part that controls the operation; measures Larmor frequency; converts it into units of magnetic field; and displays, stores, and/or outputs the data. Review of data is often possible (Fig. 1).

Global positioning plays a big role in geophysical ground and airborne surveys. In stationary measurements, GPS provides precise timing (1 μs accurate pulses referenced to Greenwich standard time).

Background Physics

Scalar magnetometers are all based on a spin of subatomic particles – electrons and protons (Abragam 1961; Schumacher 1970).

Spinning charged particles make magnetic dipoles. The dipoles are precessing around ambient (applied) magnetic field following Quantum Physics rules (Fig. 2).

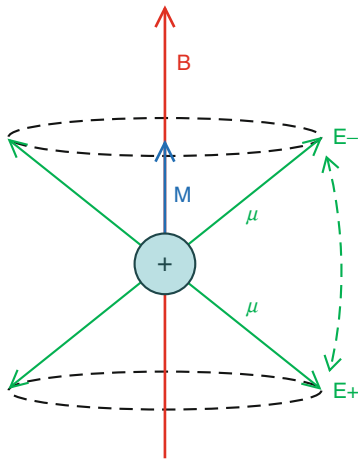
Precession frequency is proportional to magnetic flux density:

$$\omega_0 = \gamma B$$

ω_0 is an angular precession frequency (Larmor frequency), B magnetic induction (flux density), and γ gyromagnetic constant.

Spinning protons (electrons) orient themselves in magnetic field. Only two angles of orientation are allowed: acute (lower energy level) and obtuse (higher energy level).

Since individual particles precess at random phases, their dipolar magnetic field in the plane perpendicular to the magnetic field is averaged out. A projection of the dipolar magnetic fields in the direction of the applied magnetic field is static. Since orientation at the acute angle to the direction of the field is more populated than the one at the obtuse angle, a minute “polarization” of precessing particles creates a magnetization M_0 collinear with magnetic induction B :



Magnetometers, Fig. 2 Magnetic moment μ of a proton (or electron) precesses around magnetic induction B . Resulting magnetization M is collinear with B and static. Components orthogonal to B are averaged out. Coupling energy $E = -\mu B$

$$M_0 = \frac{N\gamma^2 h^2 / 4\pi^2 B}{4kT\mu_0}$$

N is the number of particles, γ gyromagnetic constant, h Planck's constant, T absolute temperature, k , μ_0 constants.

When placed in the flux density B , M_0 will reach its "thermal equilibrium" exponentially with the "longitudinal" time constant T_1 .

When deflected by about 90° , it will precess around the field by the Larmor frequency and decay exponentially by the "transversal" time constant T_2 . (Time constants T_1 and T_2 depend on the aggregate state of the assembly of spinning particles. In liquids and vapors, they may be several seconds long, while in solids, T_2 time is only in milliseconds and the precession signal disappears very quickly. This is why all scalar magnetometers have liquid or gaseous sensors.) Particles at the lower energy level can accept energy in the form of magnetic field at Larmor frequency and flip to the higher energy level. Leveling of the two energy level populations eliminates M_0 . This is saturation of the precession spectral line. Proportionality constant, the gyromagnetic constant (not always a constant), is precisely determined to better than one part per million accuracy only for protons in water (Hrvoic 1996):

$$\gamma_p = 0.2675153362 \text{ rad/nT}$$

Magnetization M_0 is too small to produce detectable precession signal when deflected in the plane of precession. All scalar magnetometers therefore need to increase polarization and magnetization M and thus increase the sensitivity of the instrument. Polarization is increased in different ways:

- By temporarily increasing B to few hundreds Gauss (Proton magnetometers)
- Transferring thermal equilibrium polarization of electrons to protons in liquid sensors (Overhauser magnetometers)
- Using light polarization in alkali metals, and ^4He magnetometers

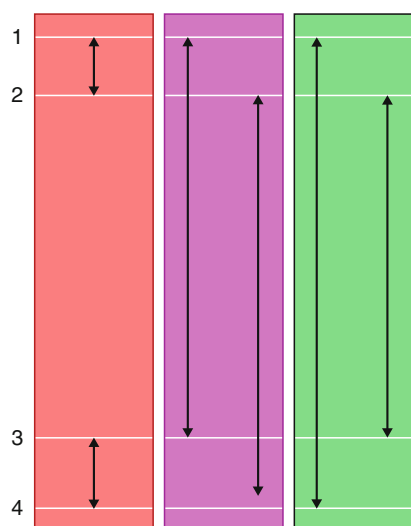
Proton Magnetometers

Proton magnetometers are the oldest scalar magnetometers. The first commercial units were produced in early 1960s as portable instruments. In continuation, airborne instruments appeared with optimized speed of readings and sensitivity, large sensors, etc. Later development of Overhauser and optically pumped magnetometers has eliminated Proton magnetometers from airborne surveys. However, they remain very popular in various ground surveys and observatories.

Proton magnetometer's sensor contains liquid rich in protons and polarization/pick-up coil. Polarization is done by increasing flux density B for a time comparable with T_1 of the sensor liquid. Electrical current passing through polarization coil creates strong magnetic field that polarizes protons of the sensor liquid. The coil is usually immersed in liquid to maximize the coupling. Alternatively, omnidirectional toroidal coil is used. Polarization field of few hundred Gauss must be roughly at the right angle to the measured field. Its removal must be fast to leave the newly formed magnetization in the plane of precession. Practical portable Proton magnetometers achieve about 0.1 nT sensitivity at a few seconds cycle, somewhat less in once per second measurements.

Overhauser Magnetometer

Overhauser effect originates from double resonance experiments in metals (Overhauser 1953). Russian and French scientists have greatly contributed to the development of modern Overhauser magnetometers based on nitroxide free radicals (Abragam 1961; Pomerantsev 1968). Practical realization of the magnetometer is as follows: Sensor liquid containing protons is placed in an RF resonator. It has small concentration of a Nitroxide free radical – stable chemical with one unpaired electron. The unpaired electron is dwelling close to the nitrogen nucleus, in its magnetic field of some 24 Gauss. Its electron paramagnetic resonance frequency in the Earth's magnetic field is about 30,000 times higher than the one of protons. Unpaired electrons couple with the protons of the sensor, creating a four energy level system (Fig. 3). Coupling can be either scalar or dipole – dipole. If one of them predominates, saturating electron resonance at some 60 MHz, transfers part of electron magnetization to protons. Increased proton polarization (magnetization) needs a 90° deflection to be turned in the plane of precession. The deflection can be either pulsed ($\pi/2$ pulse) or stationary by applying a weak rotating magnetic field of Larmor frequency in the plane of precession. Transferred magnetization is far superior to any



Magnetometers, Fig. 3 Four energy levels combined electron–proton system. Transitions 1–4 (weak scalar coupling) and 2–3 (dipole-dipole coupling) involve combined transitions of electrons and protons. 1–3, 2–4, 1–2, and 3–4 are independent transitions of electrons and protons

achievable by proton magnetometers and the measurement does not need to be interrupted for polarization; RF saturating field can be present continuously. In pulsed mode, repeated $\pi/2$ pulses interrupt measurement by about 25–30 ms. Low power consumption, no warm-up delay, higher sensitivity (10 pT for one reading per second), absolute accuracy, reasonable speed of readings (up to five readings per second), and omnidirectional sensors make Overhauser magnetometers very convenient and attractive for magnetic observatories, volcanological exploration, marine surveys, and base stations for airborne surveys. Oersted and CHAMP magnetic satellites use continuous Overhauser magnetometers as reference to vector measurement of magnetic field.

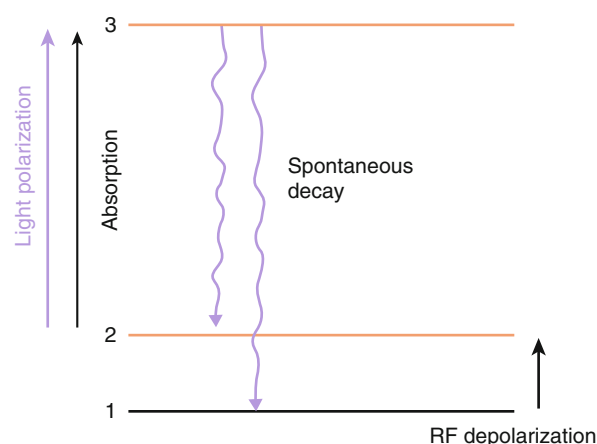
Optically Pumped Magnetometers

This is the latest to be invented and the most potent family of magnetometers. It consists of elements of the first column of the table of chemical elements. In gaseous form, alkali metals have an unpaired electron in their valence shell.

Gaseous sensor is subject to a circularly polarized light (D_1 line) (Happer 1972). Polarization process lifts electrons from a higher energy level to a metastable state from which electrons fall back to both levels.

Eventually, all electrons are in the lower energy level while the upper one is depleted. Absorption of light is then reduced and the sensor is more transparent. Rotating RF magnetic field of the Larmor frequency will depolarize the sensor and increase absorption of the light (Fig. 4).

As a result, the polarizing light will be modulated by the Larmor frequency. Its detection produces an electrical signal for the measurement of magnetic field. Helium 4 magnetometers that have two valence electrons need a weak discharge to



Magnetometers, Fig. 4 Optical polarization (lifting an electron from upper energy level to a metastable state) and depolarization (lifting electrons from lower to upper energy level) result in modulation of optical beam by Larmor frequency

lift one of them in a metastable state. Laser-pumped Helium 4 magnetometer has sub pT sensitivity. Due to the influence of nucleus (strong scalar coupling with valence electrons), all Alkali metal magnetometers have a number of spectral lines in their EPR spectrum (Breit and Rabi 1931).

Potassium spectral lines are well spaced and the operation of Potassium magnetometer can be based on a single narrow spectral line. This ensures very high sensitivity and the absolute accuracy similar to the one of Proton/Overhauser magnetometers. Cesium and rubidium lines are very close to each other, and they overlap and make a relatively wide composite single line. Operation of the magnetometer in self-oscillating mode is at the peak of the composite line. Problem is that the position of the peak depends on an angle between sensor axis and magnetic field direction. A heading error due to this effect can be reduced by a “split beam” technique, which symmetrizes the composite line. Reduction of the heading error to some nT or even a fraction of nT is possible. However, the split beam technique reduces the sensitivity by about one order of magnitude. Absolute accuracy of only few nT can be achieved by Cs and Rb magnetometers.

Measurement of Frequency

In the magnetic field of Earth (20–65 μT), precession frequency of protons is an acoustic frequency in a range of some 850 Hz–3 kHz, while Alkali metal optically pumped magnetometers produce 70–450 kHz, ^4He 280 kHz to about 2 MHz. Considering achieved precisions of measurement – one part per million or a small fraction of that for Proton and Overhauser magnetometers respectively and even few parts per billion for ^4He and ^{39}K or ^{41}K – we are resolving Larmor frequency to a millihertz precision. Times of all zero-crossings of precession frequency are taken, and an average period is measured by the least squares fit. Zero-crossing

times must be understood as a phase information too (0° , 180° , 360° , etc.), and this allows measurement to the sub Hz precisions. Measurement of phase difference to obtain the frequency results in an unusual noise dependence on the speed of readings. Most of measurements concerned with the noise bandwidth specify noise as per square root of Hertz. Doubling the bandwidth increases noise by square root of two. Not so in precession frequency measurement. When we double bandwidth by doubling the number of readings per second the noise increases by two square roots of two or $2^{3/2}$ instead of $2^{1/2}$. Defining noise as per square root of Hertz is therefore deceiving, although it may be calculated correctly for a particular number of readings per second.

Another peculiarity of precession frequency measurement is a phenomenon of “outliers” or “spikes” that occur when a zero-crossing time is either missing or modified by a spike of phase noise usually showing lower value of magnetic field. Preponderance of outliers increases dramatically as signal/noise ratio of precession signal decreases.

Absolute Accuracy and Precision

Absolute accuracy of measurement can be determined directly for Proton (Overhauser) magnetometers only. Proton gyromagnetic constant gets periodically refined by national standards institutions of a number of major countries (NIST, USA; NPL, U.K; VNIIM, Russia; NIIM, China).

Absolute accuracy of better than one part per million is possible to achieve. Besides the precise gyromagnetic constant, it requires a number of conditions to be fulfilled: sensitivity must always exceed the absolute accuracy, frequency reference precision must be adequate, phase stability of the signal and time determination of a zero-crossing must be proper, chemical shift of the precession frequency of the sensor liquid must be taken into account (Hrvoic 1996). Since we measure frequency of rotating magnetic moment, rotation of the sensor in the plane of precession introduces error (rotational Doppler). This error, insignificant for optically pumped magnetometers, may be significant for Proton and Overhauser magnetometers (about 23 nT offset for one cycle per second rotation of the sensor).

Alkali metal magnetometers have only potassium capable of high absolute accuracy (comparing its readings with Proton/Overhauser magnetometers) (Alexandrov and Bonch-Bruевич 1992). Cesium and rubidium operate on lumped spectral lines, (and ^4He has a very wide spectral line) and their absolute accuracies are in few nT range.

Great majority of nowadays measurements do not require high absolute accuracy but only sensitivity and repeatability or relative accuracy of readings. Exceptions are magnetic observatories, space measurements, and perhaps some standards measurements derived from magnetic measurements (determination of electrical current standard for example).

Sensitivities of scalar magnetometers vary. Details depend heavily on particular design, but an order of magnitude can be assessed. Proton magnetometers are nowadays used for ground exploration or calibration of vector magnetometers (fluxgates) only. Slow rate of readings can produce about 0.1 nT sensitivity and absolute accuracy. Up to two readings per second are possible, but at this rate, the noise is in nT range. Overhauser magnetometers achieve 0.01 nT at one reading per second, maximum rate five per second. High absolute accuracy Overhauser magnetometers are standard at magnetic observatories and very prominent in marine and ground surveys, and base stations for airborne surveys. Composite spectral line Alkali vapor magnetometers (Cesium, Rubidium) have similar sensitivity to Overhauser at one reading per second, but their top speed of readings is over 20/s. Their absolute accuracy is only few nT . Laser-pumped ^4He reaches sub pT sensitivities at one reading per second. Potassium has reached 0.05 pT sensitivity at one reading per second. Practical potassium assemblies for ground and airborne surveys feature sub pT once per second and about 7–10 pT rms noise at ten readings per second. Like cesium and rubidium, maximum speed of readings exceeds 20/s. Increase of noise is rather a practical limit to a number of readings per second. Cesium magnetometers are prominent in airborne surveys while potassium covers top of the field, especially in gradiometry due to its high absolute accuracy. Potassium is exclusively used in some new high sensitivity magnetic observatories and Earthquake study centers.

Vector Magnetometers

Vector magnetometers are generally used for static measurements (observatories) or in vertical suspension (archeology) or rarely in the component airborne surveys where only computed total field may reach 0.1 nT sensitivity, while components stay useless due to changes in aircraft attitude. In some recent attempt, accelerometers are used to account for attitude changes and apply corrections.

Main characteristics of most of the vector magnetometers are (except for SQUIDs):

1. Good sensitivity of measurement 0.1 nT common, 0.01 nT or somewhat lower (Korepanov et al. 2007) the best reported.
2. Very precise determination of the direction of measurement. One arc second change in direction means a fraction of one nT difference in readings, depending on an angle. In some orientations, vector magnetometers can resolve 0.01 nT field and well below one arc second angle.

3. Relatively substantial temperature dependence ($0.1 \text{ nT}/^\circ\text{C}$ or more).
4. Relatively substantial time variation – aging (2 nT or more per year).

Consequences of the above are difficulties in orienting in the magnetic field and determination of the orthogonality of the three components. Due to their temperature dependence/aging, the vector magnetometers at the magnetic observatories need calibration by Proton/Overhauser magnetometers.

Fluxgate Magnetometer

Main vector magnetometer today is the fluxgate. Its operation is based on nonlinearity of magnetic properties of steel (ferromagnetic material) (Primdahl 1979). When placed in strong magnetic field, steel will lose its high magnetic permeability in saturation, i.e., relative permeability will fall from several hundreds to close to one. A coil is wound around a layered core of steel, and alternating current is driving the core to saturation in two opposite directions alternatively. When not saturated, the core will concentrate the magnetic flux of the applied magnetic field. When saturated, the flux will deconcentrate itself. Pick-up coil wound around the core will detect changes in the flux passing through the core. With no external magnetic field, the saturations on positive and negative sides will be symmetrical i.e. the pick-up voltage will have only odd harmonics. When the external magnetic field is present, the saturations become unsymmetrical and second harmonic appears in the pick-up voltage, its amplitude proportional to the strength of magnetic field. Phase detected second harmonic is then a measure of applied magnetic field in the direction of the core. It is convenient to use two pieces of the core with excitation windings in opposition and the pick-up coil encompassing both cores as this eliminates pick-up voltage due to the original driving current. Instead of double iron core, a ring or “racetrack” shaped cores can be used (Fig. 5).

Today’s fluxgates can achieve $0.1 \text{ nT}/^\circ\text{C}$ temperature coefficient with some nT per year aging. Composition of the iron

cores plays a big role in the quality of fluxgates. Some recent new developments (Korepanov et al. 2007) claim better noise and aging characteristics.

During the Second World War, a total field – oriented single fluxgate was used to detect submarines. ASQ-10 instrument had about 0.1 nT sensitivity and the total field orientation of the sensor was kept by feedback electronics/mechanics.

Absolute Measurement

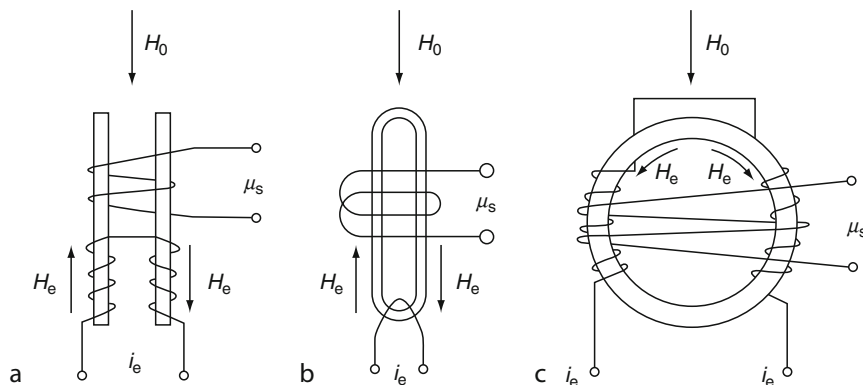
Declination/Inclination magnetometer (DI-Flux or DIM) is a variation of fluxgate measurement (Jankowski and Sucksdorff 1996). Fluxgate magnetometer sensor is mounted collinear with the optical axis of (nonmagnetic) theodolite, in horizontal plane. Theodolite is rotated, and two zero field angles are marked. The telescope is then reversed and two more zero readings determined. Magnetic north in arbitrary coordinate system is then calculated from angles of the four zero field measurements. The calculated angle is then referred to a known marker to determine true declination angle related to north direction. To determine inclination, the theodolite is oriented in just determined magnetic declination vertical plane and four zero field positions are determined in that plane. Averaging will give the angle off horizontal, i.e., the inclination. Some attempts have been made (see ► “Geomagnetic Field, Measurement Techniques”) to automate absolute measurements.

Simplified version of DIM is the Declinometer. It has a suspended magnet on a torsionless fiber, a mirror attached to it at the right angle to its magnetic axis and in front of a telescope. The theodolite is turned until the telescope becomes at right angle to the mirror. Direction of magnetic meridian is read from the base of the theodolite. This is referenced to a known reference mark to determine real declination angle.

Weakness of this measurement is an angle mirror-magnetic axis of the magnet that has to be 90° . This is mitigated by using two magnets of different magnetic moments.

Magnetometers,

Fig. 5 Excitation current drives the two cores (a) in saturation in opposite directions. Pick-up coil detects unsymmetry in saturation by measuring second harmonic. The cores can be connected at the ends to create “race track” (b) or ring core (c) shapes. Ring core allows for measurement of two orthogonal components, on one core



Magnetometers, Fig. 6 DIdD Overhauser/Potassium suspended system (a) and with external vertical coil to measure inclination directly (b)



Quantum Vector Magnetometers

SQUID is a vector magnetometer of exceptional fT range sensitivity. It operates at cryogenic temperatures (liquid helium or liquid hydrogen) (Clark 1993). It is used sporadically in geophysics for tensor calculations or short base gradiometers.

DIdD (Delta inclination delta declination) is a vector magnetometer with two orthogonal bias coils and a scalar magnetometer (Overhauser or Potassium) (Fig. 6).

Coils are oriented at right angles to magnetic field direction, one in a horizontal plane, one in vertical magnetic declination plane. A sequence of four biased fields and one unbiased allows for determination of departures of the field from the preset one. Orientation of one bias coil in the vertical direction and the other in the horizontal East-West direction (now possible with the precise GPS direction determination or a North seeking gyroscope) will convert dIdD instrument into an absolute D and I instrument. Potassium magnetometer can provide superior sensitivity and absolute accuracy. The stability of the instrument as for temperature and aging can be superior to the best fluxgates. This instrument is still in development. It is meant for use in the directional drilling for oil and minerals and possibly in a refined version for magnetic observatories.

Summary

Magnetic measurements are an essential part of geophysical exploration for minerals and oil, archeological, volcanological explorations, research of atmosphere, Sun's influence, crustal studies, earthquake research, etc. Scalar

magnetometers are dominant in most fields. They are described in some detail – their physics, principles of operation, details of design, sensitivities and absolute accuracy. Vector magnetometers are essential in magnetic observatories, space measurements, and many scientific investigations. Fluxgate magnetometers and variations of absolute determination of magnetic field direction have been described in some detail. Quantum magnetometers as vector magnetometers are briefly described. Refinement of Potassium magnetometers may improve sensitivity of presently used scalar magnetometers by two orders of magnitude to sub pT range.

Cross-References

- [Archaeomagnetism](#)
- [Equatorial Electrojet](#)
- [Geomagnetic Field, Measurement Techniques](#)
- [Geomagnetic Field, Secular Variation](#)
- [Magnetic Gradiometry](#)
- [Magnetic Methods, Airborne](#)
- [Magnetic Methods, Principles](#)
- [Magnetic Methods, Surface](#)

Bibliography

- Abragam A (1961) The principles of nuclear magnetism. Oxford University Press, New York
- Acuña MH (2002) Space-based magnetometers. Rev Sci Instrum 73:11
- Alexandrov E, Bonch-Bruевич V (1992) Optically pumped atomic magnetometers after three decades. Opt Eng 31(4):711
- Breit G, Rabi II (1931) Measurement of nuclear spin. Phys Rev 38:2082

- Clark J (1993) SQUIDS: theory and practice. In: Weinstock H, Ralston RW (eds) *The new superconducting electronics*. Kluwer Academic, Dordrecht, pp 123–180
- Happer W (1972) Optical pumping. *Rev Mod Phys* 44:169
- Hrvoic I (1996) Requirements for obtaining high accuracy with proton magnetometers. In: Rasson JL (ed) *Proceedings of the VIth workshop on geomagnetic observatory instruments data acquisition and processing*, pp 70–72. www.gemsys.ca
- Hrvoic I (2004) Short review of scalar magnetometers (unpublished). www.gemsys.ca
- Hrvoic I (2008) Development of a new high sensitivity potassium magnetometer for geophysical mapping. *First Break* 26:81–85. www.gemsys.ca
- Jankowski J, Sucksdorff C (1996) Guides for magnetic measurements and observatory practice. IAGA iaga_sg@gfz-potsdam.de
- Kaufman AA, Kleinberg RL, Hansen R (2009) *Principles of magnetic methods in geophysics*. Elsevier Press, Amsterdam/London
- Korepanov V, Klymovych Y, Kuznetsov O, Pristay A, Marusenkova A, Rasson J (2007) New INTERMAGNET fluxgate magnetometer XII IAGA workshop. *Publ Inst Geophys Pol Acad Sci C-99(398)*:291–298
- Körmendi A (2008) Private communication. alpar44@freemail.hu
- Overhauser WA (1953) Polarization of nuclei in metals. *Phys Rev* 92:411
- Pomerantsev NM (1968) Quantum magnetometry. Joint Publication Research Service JPRS 46444
- Primdahl F (1979) The fluxgate magnetometer. *J Phys E Sci Instrum* 12:241–253
- Schumacher R (1970) *Introduction to magnetic resonance*. Benjamin W. A., New York

Magnetotelluric Data Processing

Gary Egbert

College of Oceanic and Atmospheric Sciences,
Oregon State University, Corvallis, OR, USA

Definition

Magnetotelluric data processing: statistical estimation and analysis of impedance tensors and other transfer functions from measured electromagnetic time-series data.

Introduction

In the magnetotelluric (MT) method (see “► [Magnetotelluric Interpretation](#)”), *data processing* generally refers to the initial analysis steps required to use naturally occurring time variations of electromagnetic (EM) fields measured at the surface to map the distribution of electrical conductivity within the Earth (see “► [Goelectromagnetism](#)”). The data are acquired as time series, generally including three components of the vector magnetic fields, along with the two components of the horizontal electric field. Earth conductivity variations have a relatively subtle effect on the

raw EM data, which most directly reflect the temporal pattern of external source variations. Sophisticated data processing methods have been developed to extract the interpretable Earth conductivity signal from these data. Two distinct steps can be distinguished. In the first, time series processing methods are used to estimate frequency-dependent linear relations or transfer functions (TFs), relating magnetic and electric field components. In the second step, distortion and strike analysis, the full set of TFs is reduced to a simpler set, consistent with simplified one or two-dimensional models of Earth conductivity, typically allowing for shallow near-surface complications. After these data processing steps, TFs are inverted for conductivity, and the resulting images are interpreted in terms of composition, physical state, and Earth structure.

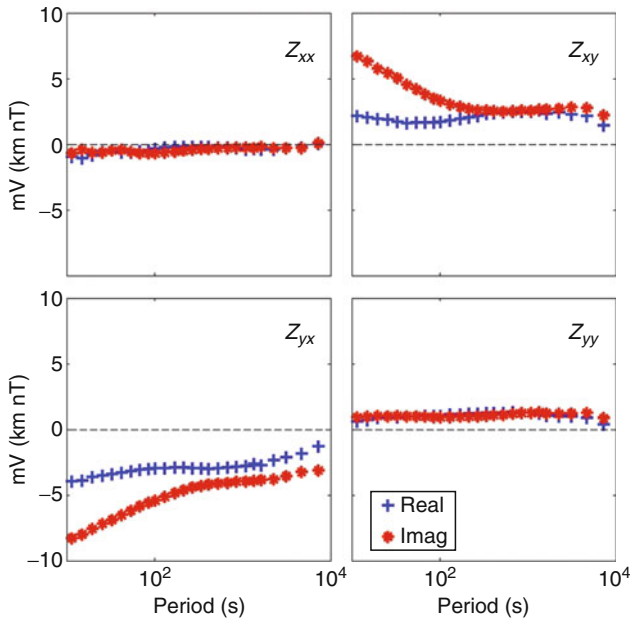
MT Transfer Functions

The basic assumption underlying the MT method is that the external source fields are spatially uniform, an approximation that can be physically justified provided the spatial scale of external magnetic fields at Earth’s surface is large compared to the depth of penetration of the EM fields in the conducting Earth. Except at very long periods ($T > 10^4$ s), for which penetration depths in the Earth can exceed several hundred kilometers, and in the auroral zone where ionospheric current systems vary over short length scales, this assumption holds quite well for natural sources. Local cultural sources of EM noise have much shorter spatial scale, and will not always satisfy the uniform source assumption. These sorts of coherent noise, which violate the most fundamental MT assumption, represent the most serious obstacle to successful MT transfer function (TF) estimation.

For periodic sources at a fixed frequency, the assumption of spatial uniformity implies that sources can be expressed as linear combinations of two simple modes: unit magnitude uniform sources linearly polarized North–South and East–West. This in turn implies a linear relationship (in the frequency domain) between the horizontal electric and magnetic field vectors measured at a single site.

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} Z_{xx}(\omega) & Z_{xy}(\omega) \\ Z_{yx}(\omega) & Z_{yy}(\omega) \end{pmatrix} \begin{pmatrix} H_x \\ H_y \end{pmatrix} \quad (1)$$

The frequency-dependent 2×2 TF $Z(\omega)$ is referred to as the impedance tensor; see Fig. 1 for an example. Equation 1 can be justified by the linearity of Maxwell’s equations, and the assumption that all sources can be expressed as linear combinations of two linearly independent polarizations (Egbert and Booker 1989). Under these circumstances, two independent field components (e.g., H_x and H_y at one



Magnetotelluric Data Processing, Fig. 1 Example of an impedance tensor, the 2×2 TF relating horizontal electric and magnetic field components. Real and imaginary parts of the complex impedance components are plotted as a function of period. Note that the diagonal components of the impedance are relatively small; for the case of a 2D conductivity distribution, these components would vanish identically, if the impedance were expressed in the proper coordinate system

location) uniquely determine all other EM field components at any nearby site. The uniform source assumption thus also justifies other sorts of TFs. In the other most important case, the vertical magnetic field component is related to the two horizontal components at the local site

$$H_z = \begin{pmatrix} T_x(\omega) & T_y(\omega) \end{pmatrix} \begin{pmatrix} H_x \\ H_y \end{pmatrix}. \quad (2)$$

Vertical field TFs, which are commonly estimated along with the impedance in most modern MT surveys, are referred to variously as the Tipper, Parkinson vector, or Wiese vector. Inter-station TFs, which relate horizontal magnetic fields at two sites, are also sometimes considered (e.g., Egbert 2002).

Least Squares Transfer Function Estimation

The earliest approaches to MT TF estimation were based on classical time series methods (e.g., Bendat and Piersol 1971), applying a simple linear least squares (LS) fitting procedure in the frequency domain. To be explicit, consider estimating the MT TF between one component of the electric field (e.g., E_x) and the two horizontal magnetic field components H_x and H_y at a single fixed frequency ω . This TF corresponds to the first row of the usual MT impedance tensor. All of the discussion here and in subsequent

sections applies equally to the second row of the impedance, and to the vertical field TF. Because all of the TFs are most succinctly described in the frequency domain as in Eqs. (1 and 2), the first step in data processing is generally to Fourier transform the data. After possibly de-spiking and/or pre-whitening, time series for each component are divided into M short time windows, tapered, and Fourier transformed (Fig. 2). Fourier coefficients for N frequencies in a band centered around ω are used for the TF estimate, for a total of $I = MN$ complex data. Estimates $\hat{\mathbf{Z}}$ of the TF (i.e., impedance elements) are then obtained for frequency ω by least squares (LS) fitting of the linear model

$$\begin{pmatrix} E_1 \\ \vdots \\ E_I \end{pmatrix} = \begin{pmatrix} H_{x1} & H_{y1} \\ \vdots & \vdots \\ H_{xI} & H_{yI} \end{pmatrix} \begin{pmatrix} Z_{xx} \\ Z_{xy} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_I \end{pmatrix}, \quad (3)$$

or in matrix notation

$$\mathbf{E} = \mathbf{H}\mathbf{Z} + \boldsymbol{\varepsilon}. \quad (4)$$

With standard least squares (LS), this is accomplished by minimizing the sum of the squares of the residuals:

$$\sum_i \left| E_{xi} - (H_{xi}\hat{Z}_{xx} + H_{yi}\hat{Z}_{xy}) \right|^2 = \sum_i |r_i|^2 \rightarrow \min \quad (5)$$

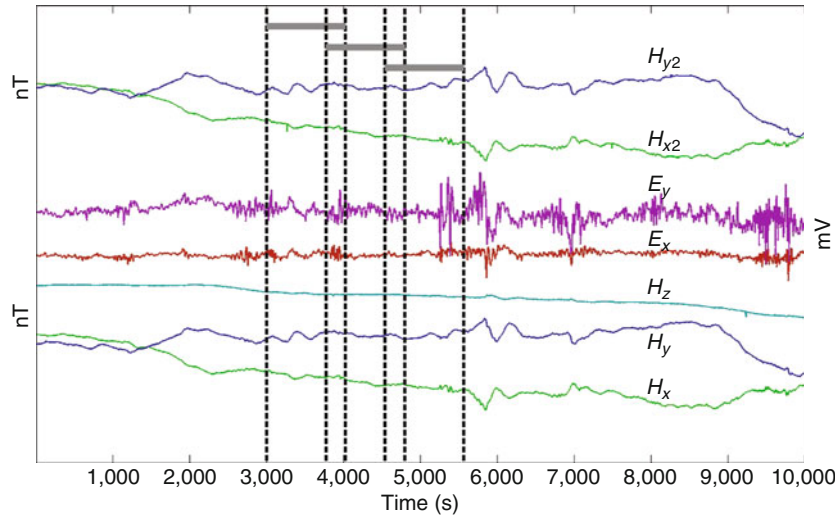
yielding

$$\hat{\mathbf{Z}} = (\mathbf{H}^* \mathbf{H})^{-1} (\mathbf{H}^* \mathbf{E}) \quad (6)$$

where the superscript asterisk denotes the conjugate transpose of the complex matrix.

Robust Estimation

The simple LS estimator implicitly assumes a Gaussian distribution for the errors ε_i in Eq. (3). This assumption often fails for MT data due to the non-stationarity of both signal and noise. Both can be highly variable in time, resulting in a marginal error distribution in the frequency domain that is heavy tailed and contaminated by outliers. As a result, the simple LS estimate all too frequently leads to very poor TF estimates with large error bars (Fig. 3a). A number of MT processing methods have been proposed to overcome these difficulties, generally using some sort of automated screening or weighting of the data. Early efforts in this direction used ad hoc schemes, for example weighting data segments based on broad-band coherence between input and output channels. A more rigorously justifiable approach is based on the regression M-estimate (RME; Huber 1981), a variant on LS that is



Magnetotelluric Data Processing, Fig. 2 Example of time series collected with a long-period MT system. From bottom, three vector components of the magnetic field (measured with a fluxgate magnetometer), two colocated orthogonal horizontal electric field components (measured as the potential difference between pairs of electrodes separated by 100 m), and two horizontal components of the magnetic field at a distant remote site. Three overlapping data segments are indicated by

the vertical dashed lines. Segments such as these are Fourier transformed to yield a series of complex Fourier coefficients for each data channel, which are then used for the various transfer function estimates discussed in the text. Longer segments are processed to get Fourier coefficients for longer period data, typically after low-pass filtering and decimating the time series

robust to violations of distributional assumptions and resistant to outliers (Egbert and Booker 1986; Chave et al. 1987). For the RME, the quadratic loss functional of (5) is replaced by

$$\sum_i \rho \left(|E_{xi} - (H_{xi} \hat{Z}_{xx} + H_{yi} \hat{Z}_{xy})| / \hat{\sigma} \right) = \sum_i \rho(r_i / \hat{\sigma}) \rightarrow \min \quad (7)$$

where $\hat{\sigma}$ is some estimate of the scale of typical residuals. LS is a special case of (7), with $\rho(r) = r^2$. By choosing the functional form for $\rho(r)$ so that large residuals are penalized less heavily than with the quadratic used for LS, the influence of outliers on the estimate can be substantially reduced. For robust estimation of MT TFs, the so-called *Huber* loss function (Huber 1981)

$$\rho(r) = \begin{cases} r^2/2 & |r| < r_0 \\ r_0 |r| - r_0^2/2 & |r| \geq r_0 \end{cases} \quad (8)$$

is commonly used with $r_0 = 1.5$.

The more general loss function of (7) can be minimized with an iterative scheme, which can be viewed as weighted LS, with weights determined from the data. Define the *influence function* $\psi(r) = \rho'(r)$ (the derivative of the loss function), and set $w(r) = \psi(r)/r$. Then, it is easily shown that the minimizer of (7) satisfies

$$\hat{\mathbf{Z}} = (\mathbf{H}^* \mathbf{W} \mathbf{H})^{-1} (\mathbf{H}^* \mathbf{W} \mathbf{E}), \quad (9)$$

where $\mathbf{W} = \text{diag}[w_1 \dots w_I] = \text{diag}[w(|r_1|) \dots w(|r_I|)]$ is a diagonal matrix of weights. The RME thus corresponds approximately to the weighted LS problem $\sum_i w(|r_i|) |r_i|^2 \rightarrow$

$\min$. The weights depend on the normalized residuals $r_i = (E_{xi} - [Z_{xx}H_{xi} - Z_{xy}H_{yi}]) / \hat{\sigma}$, and hence on the TF estimate, so an iterative procedure is required. Given an estimate of the TF, and of the error scale $\hat{\sigma}$, normalized residuals can be used to calculate weights, and the weighted LS problem can be solved for a new TF estimate. This procedure can be started from a standard LS estimate of the TF (and some robust estimate of error scale, for example, based on the median absolute deviation of residuals) and then repeated until convergence.

For convex loss functions (e.g., the Huber function of (8)), convergence of this procedure to the unique minimizer of (7) is guaranteed (Huber 1981). For the Huber loss function in (8), the weights are

$$w(r) = \begin{cases} 1 & |r| < r_0 \\ r/r_0 & |r| \geq r_0 \end{cases}, \quad (10)$$

i.e., data corresponding to large normalized residuals get smaller weights. To allow for a sharp cutoff, with data exceeding a hard threshold discarded, the loss function $\rho(r)$ must be non-convex, and convergence of the iterative minimization algorithm cannot be guaranteed. The standard

practice is to iterate with a convex loss function such as (8) to convergence, followed by a final step or two with a hard cutoff to completely discard extreme outliers. This completely automated procedure frequently results in significant improvements in TF estimates, as indicated by improvements in smoothness and physical realizability of apparent resistivity and phase curves (Fig. 3b), and reproducibility of results.

More ad hoc schemes are also frequently used for down-weighting noisy or inconsistent data. These can be viewed as special cases of the weighted LS estimate of (9) but with weights now determined by some criteria other than residual magnitude. For example, in coherence weighting, broad-band coherence of input and output channels is used to down-weight time segments with low signal-to-noise ratios. Such methods can of course be combined with the RME.

Leverage

The RME can still be excessively influenced by a small number of data sections with large signal (or more typically noise) amplitude. In the terminology of linear statistical models, these large amplitude data are *leverage points*. For the LS estimate, the predicted data are

$$\mathbf{E} = [\mathbf{H}(\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*] \mathbf{E} \quad (11)$$

with the matrix expression in brackets (which maps observed data to predicted) referred to as the “hat matrix.” It is readily shown that the sum of the diagonal elements of the hat matrix satisfy $\sum_{i=1}^I h_{ii} = 2$, and that the individual diagonal elements h_{ii} can be interpreted as the fraction of total magnetic field signal power in the i th Fourier coefficient. Large values of h_{ii} indicate data points with inordinate influence on the TF estimates. In extreme cases, leverage points are used heavily to predict themselves, making the iterative RME ineffective. To deal with leverage points in routine MT processing, it is thus useful to include an additional weighting term (a function of the magnitude of h_{ii}) to reduce the influence of any observations of unusually large amplitude (e.g., Chave and Thomson 1989).

Remote Reference

The linear statistical model (4) is strictly appropriate to the case where noise is restricted to the output, or predicted electric field channels. Violation of this assumption results in the downward bias of estimated impedance amplitudes. These biases are proportional to the ratio of noise power to signal power, and can be quite severe in the so-called MT “dead bands” centered at frequencies 1–5 kHz and 0.1–1 Hz (examples of such biases are seen in Fig. 3a, b, at periods shorter than 20 s). To avoid these bias errors, horizontal magnetic fields recorded simultaneously at a remote reference site are correlated with the EM fields at the

local site (Gamble et al. 1979). Letting R_{xi} and R_{yi} be the Fourier coefficients for the two remote site components for the i th data segment, and $\mathbf{R}$ the corresponding $I \times 2$ matrix, then the remote reference analog of the LS estimate is

$$\hat{\mathbf{Z}} = [\mathbf{R}^*\mathbf{H}]^{-1}[\mathbf{R}^*\mathbf{E}]. \quad (12)$$

It can be shown that the remote reference estimate is unbiased if noise at the remote site is not coherent with noise in magnetic and electric fields at the local site. Thus, the remote reference estimate can eliminate or significantly reduce the effects of coherent cultural EM noise, provided the reference site is sufficiently distant. Very large scale sources of cultural noise, in particular DC electric railways, still present a significant challenge to collecting high-quality MT data in many parts of the world. These sources can be coherent over large enough scales to contaminate both local and remote signals, but still have significant power at wavelengths which are short enough to violate the MT uniform source assumption.

To generalize the RME to remote reference, one can iterate the weighted analog of (12), i.e., $\hat{\mathbf{Z}} = [\mathbf{R}^*\mathbf{W}\mathbf{H}]^{-1}[\mathbf{R}^*\mathbf{W}\mathbf{E}]$, weights on the diagonal of $\mathbf{W}$ determined from the residual magnitudes, exactly as for the single site robust estimator. It is also useful to add some additional weighting to allow for outliers at the remote site, and as with the single site estimates, for leverage. When arrays of simultaneously operating EM instruments are available, more complex procedures, which use data from multiple sites to define the reference fields are possible (Egbert 2002).

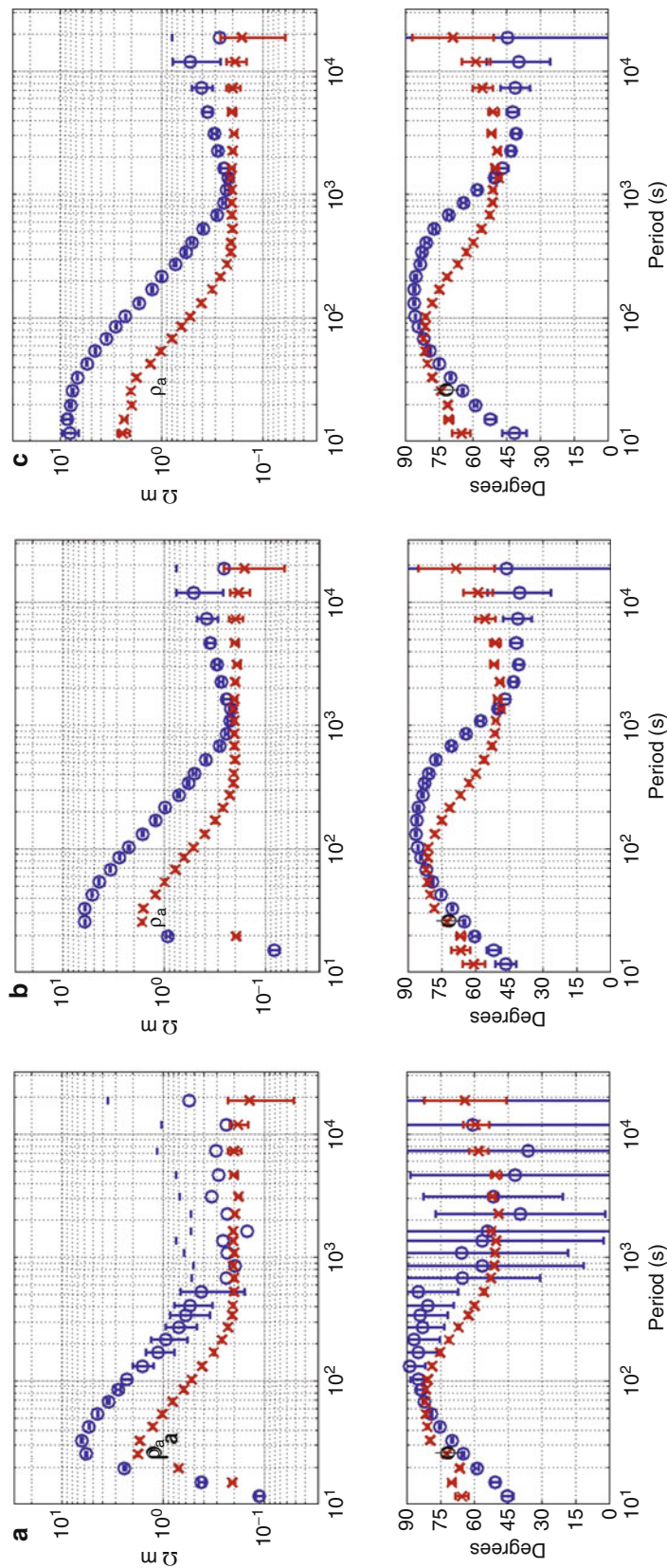
Error Estimates

With a statistical approach to TF estimation, one also obtains error variances that define the precision of the TF estimates. These error bars are required to assess the adequacy of fit of models derived by inversion, and thus play a critical role in the interpretation process. The covariance of the linear LS TF estimate of (6) is readily derived from standard theory, based on linear propagation of errors:

$$\begin{aligned} \text{Cov}[\hat{\mathbf{Z}}] &= \hat{\sigma}^2 [\mathbf{H}^*\mathbf{H}]^{-1} \\ \hat{\sigma}^2 &= (I - 2)^{-1} \sum_i |r_i|^2. \end{aligned} \quad (13)$$

where r_i are again the residuals. An analogous expression for the error covariance for the remote reference estimate is

$$\begin{aligned} \text{Cov}[\hat{\mathbf{Z}}] &= \hat{\sigma}^2 [\mathbf{R}^*\mathbf{H}]^{-1} [\mathbf{R}^*\mathbf{R}] [\mathbf{H}^*\mathbf{H}]^{-1} \\ \hat{\sigma}^2 &= (I - 2)^{-1} \sum_i |r_i|^2. \end{aligned} \quad (14)$$



Magnetotelluric Data Processing, Fig. 3 Comparison of TFs estimated from the same time series with three procedures described in the text: (a) standard unweighted LS; (b) the regression M-estimate applied to data from a single site; and (c) the robust remote reference estimate. Here, apparent resistivities $\rho_a = (\omega\mu_0)^{-1}|Z|^2$ and phases $\phi = \tan^{-1}(Z)$ are plotted, computed from the dominant off-diagonal impedances Z_{xy} (circles) and Z_{yx} (crosses). The LS estimates are too noisy

to be useful over much of the frequency range, while both of the robust estimates vary smoothly with frequency and have small error bars. However, the single site apparent resistivities of (a) and (b) are severely biased to low values for periods shorter than about 20 s. This bias is eliminated by the remote reference processing

For both (13) and (14), the diagonal components of the 2×2 covariance matrices are the estimation error variances for the two impedance elements.

For the RME, variances of the estimates are complicated by nonlinearity, but asymptotic expressions (valid in the limit of large sample sizes) can be obtained from standard theory (Huber 1981):

$$\text{Cov}[\hat{\mathbf{Z}}] = \frac{(I-2)^{-1} \sum_i w_i |r_i|^2}{[I^{-1} \sum_i \psi'(|r_i|/\hat{\sigma})]^2} [\mathbf{H}^* \mathbf{H}]^{-1} \quad (15)$$

where r_i , w_i , $i = 1, \dots, I$ and $\hat{\sigma}$ are the residuals, and weights and error scale from the final iteration, and the prime denotes the derivative of the influence function. An analogous asymptotic covariance for the robust remote reference estimates can be given with $[\mathbf{H}^* \mathbf{H}]^{-1}$ in (15) replaced by $[\mathbf{R}^* \mathbf{H}]^{-1} [\mathbf{H}^* \mathbf{H}] [\mathbf{H}^* \mathbf{R}]^{-1}$.

As an alternative to asymptotic error estimates such as (15), the nonparametric jackknife method (Efron 1982) has also been frequently used for computing TF error estimates (Thomson and Chave 1991). The jackknife approach can also be applied to compute error bars for complicated nonlinear functions of TFs, which arise, for example, in some forms of distortion analysis. The jackknife approach significantly increases computational effort required for TF estimation, unless some approximations are used (Eisel and Egbert 2001).

Strike and Distortion Analysis

As of this writing, MT data are most often acquired in profiles across a geologically appropriate strike, and the data are interpreted with two-dimensional (2D) Earth conductivity models that vary only with depth, and across strike, a reasonable approximation for many geologic structures. If the 2D assumption holds exactly, and the coordinate axes (used to give the components of the horizontal electric and magnetic fields) are chosen so that one of x or y aligns with the geoelectric strike (i.e., the direction along which there is no variation in conductivity), the diagonal components of the impedance tensor vanish.

$$\mathbf{Z}(\omega) = \begin{bmatrix} 0 & Z_{xy}(\omega) \\ Z_{yx}(\omega) & 0 \end{bmatrix}. \quad (16)$$

This special form only holds when the impedance is expressed in the proper coordinate system. Thus, once impedances are estimated, a key step in 2D MT interpretation is to determine an appropriate geoelectric strike where (16) holds approximately. The two off-diagonal components in this coordinate system (the transverse electric or TE mode,

in which electric currents flow along the geologic strike, and transverse magnetic or TM mode, in which electric currents flow across the geologic strike; *Magnetotelluric Interpretation*), are then fit using 2D modeling and inversion programs.

Even if the 2D approximation is reasonably justified for large-scale deep structure, more complex (3D) small-scale near-surface complications in conductivity will occur. These near-surface features result in distortion of the impedance tensor, so that even when the correct regional strike is used to define the coordinate system, the impedance tensor does not have the simple form of (16). Most modern approaches to 2D MT interpretation allow for these distortion effects in strike determination, and subsequent inversion. Early approaches to strike determination, such as choosing the strike angle to minimize the magnitude of the diagonal components, without accounting for distortion effects are seldom used anymore.

In the limit where the penetration depth of the EM fields greatly exceeds the depth to the bottom of the near-surface distorting structures, the electric fields (only) are distorted in a static (frequency-independent) manner. Allowing for rotation of the measurement coordinate system by an angle θ , the distorted impedance tensor for a 2D Earth can be decomposed as

$$\begin{aligned} \mathbf{Z}(\omega) \mathbf{z} &= \mathbf{U} \mathbf{D} \tilde{\mathbf{Z}}(\omega) \mathbf{U}' \\ &= \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix} \\ &\times \begin{bmatrix} 0 & \tilde{Z}_{xy}(\omega) \\ \tilde{Z}_{yx}(\omega) & 0 \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}, \end{aligned} \quad (17)$$

where $\mathbf{D}$ is the distortion matrix, which is real, and $\tilde{\mathbf{Z}}(\omega)$ is the undistorted 2D impedance, which carries all of the frequency dependence. It is readily verified that not all of the parameters in Eq. (17) are uniquely identifiable – e.g., the distortion parameter D_{xx} cannot be determined independent of the absolute level of the impedance $\tilde{Z}_{xy}$; only the product $D_{xx} \tilde{Z}_{xy}$ is uniquely determined. However, (17) provides a suitable model to determine an optimal geoelectric strike (defined by the rotation angle θ), and this has ultimately become the most common use for strike and distortion analysis in most 2D MT interpretations. Note that in principal the same $\mathbf{D}$ and θ should apply to all frequencies at one site, while the strike angle should be the same over all sites.

Nonlinear least-squares estimation schemes can be used to fit the parameters in (17) (e.g., Smith 1995; McNeice and Jones 2001). These are often applied site by site, for various subsets of frequencies, but impedances from multiple sites can also be fit simultaneously. Finding a consistent strike provides justification, and an optimal coordinate system, for 2D interpretation. Sometimes, the analysis is also used to

eliminate sites, or frequency ranges, which are inconsistent with the generally preferred strike. Once the strike is determined and the impedance tensors are rotated, the off-diagonal components can be used as input for 2D inversion. Because the overall level of the impedance curves are still potentially effected by near-surface distortion, but the phases are not, it is common to fit phases more tightly, and relax fit to amplitudes (i.e., apparent resistivities). It is also possible to directly allow for distortion by introducing additional parameters at the inversion stage.

Summary

Data processing represents the first step in MT data interpretation, resulting in a reduction of the raw time-series data (approximately 10^6 observations) to frequency domain transfer functions such as impedances (a few hundred numbers). These transfer functions contain information about the response of the Earth to simple plane-wave sources, and provide input data for subsequent inversion and interpretation steps. Robust remote reference processing substantially improves reliability of transfer functions estimates. Except in areas of strong cultural noise (which unfortunately are increasingly common), modern processing produces smooth unbiased response curves. Distortion and strike analysis represents a further step in MT processing, essential for 2D interpretation of MT profile data. Such analysis is used to test for consistency with the assumed 2D geoelectric geometry, to estimate the strike, and to reduce the full impedance to the TE and TM mode responses appropriate to 2D interpretation.

Cross-References

- [Electrical Properties of Rocks](#)
- [Electrical Resistivity Surveys and Data Interpretation](#)
- [Geoelectromagnetism](#)
- [Instrumentation, EM](#)
- [Magnetotelluric Interpretation](#)

Bibliography

- Bendat JS, Piersol AG (1971) Random data: analysis and measurement procedures. Wiley, New York
- Chave AD, Thomson DJ (1989) Some comments on magnetotelluric response function estimation. *J Geophys Res* 94(B10). <https://doi.org/10.1029/JB094iB10p14215>
- Chave AD, Thomson DJ, Ander M (1987) On the robust estimation of power spectra, coherence and transfer functions. *J Geophys Res* 92:633–648
- Efron B (1982) The jackknife, the bootstrap and other resampling methods. Society for Industrial and Applied Mathematics, Philadelphia

- Egbert GD (2002) Processing and interpretation of electromagnetic induction array data: a review. *Surv Geophys* 23:207–249
- Egbert GD, Booker JR (1986) Robust estimation of geomagnetic transfer functions. *Geophys J R Astron Soc* 87:173–194
- Egbert GD, Booker JR (1989) Multivariate analysis of geomagnetic array data I: the response space. *J Geophys Res* 94:14227–14248
- Eisel M, Egbert GD (2001) On the stability of magnetotelluric transfer function estimates and the reliability of their variances. *Geophys J Int* 144:65–82
- Gamble T, Goubau W, Clarke J (1979) Magnetotellurics with a remote reference. *Geophysics* 44:53–68
- Huber P (1981) Robust statistics. Wiley, New York
- McNeice GW, Jones AG (2001) Multisite, multi-frequency tensor decomposition of magnetotelluric data. *Geophysics* 66:158–173
- Smith JT (1995) Understanding telluric distortion matrices. *Geophys J Int* 122:219–226
- Thomson D, Chave A (1991) Jackknifed error estimates for spectra, coherence, and transfer functions. In: Haykin S (ed) *Advances in spectrum analysis and array processing*. Prentice Hall, Englewood Cliffs

Magnetotelluric Interpretation

John F. Hermance

Department of Geological Sciences, Brown University,
Providence, RI, USA

What Is Magnetotellurics?

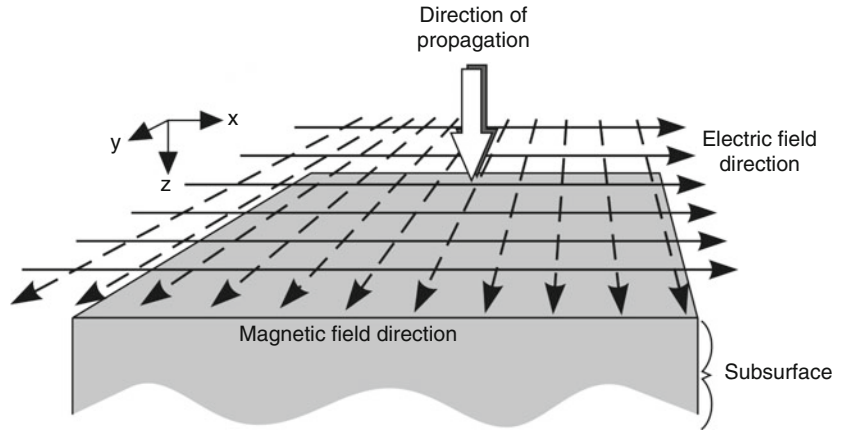
The magnetotelluric (MT) method is a geophysical technique in which the time-varying electric and magnetic fields from distant natural or artificial sources are measured at the Earth's surface, and analyzed to determine the electrical properties of the subsurface.

Workers have known from the 1800s that naturally occurring transient voltages of significant magnitude observed on grounded telegraph wires were associated with rapid fluctuations of the geomagnetic field following *solar flares* and *auroral displays*. These natural voltage fluctuations – so-called telluric fields – have since been explained as the product of eddy currents induced in the finite conducting earth by transient magnetic field disturbances in the ionosphere and magnetosphere, even the global-wide propagation of radiated signals from lightning strokes in the distant atmosphere (see ► [“Electromagnetic Pulsations and Magnetic Storms”](#)).

Early work by Kato and Kikuchi, and Rikitake in Japan, and Tikhonov in the then, Soviet Union, laid the foundation for using these natural electric and magnetic field variations to study the electrical structure of the solid earth (see the collection of classic reprints by Vozoff 1985; see also ► [“Geoelectromagnetism”](#)). However, it was the seminal paper by Louis Cagniard (1953) that first described the

Magnetotelluric Interpretation,

Fig. 1 Schematic of a uniform plane wave vertically incident on the earth's surface



comprehensive procedure for using the ratio of the horizontal electric – that is, the *telluric* – field to the horizontal *magnetic* field, for which he coined the name the magnetotelluric method.

Fundamental Concepts: Surface Impedance and Apparent Resistivity

The magnetotelluric method, as generally employed by geophysicists, invokes several fundamental concepts from the theory of the propagation of electromagnetic waves. The first concept is the physical attribute known as the *electromagnetic wave impedance* of the medium (see ► “Magnetotelluric Data Processing”), provided by computing the ratio of the electric field intensity to the orthogonal magnetic field intensity. For example, Fig. 1 shows an electromagnetic plane wave vertically incident on the earth's surface; the electric field is polarized in the x -direction, and, to be consistent with the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$, the magnetic field is polarized in the y -direction. Alternatively, for the case where the electric field might be polarized in the y -direction, the magnetic field will be polarized in the $-x$ -direction. The ideal natural field for magnetotellurics is one which is randomly polarized in *all* horizontal directions, so that usually both sets of electric and magnetic data, (E_x, H_y) and (E_y, H_x) , will be employed, leading to two estimates of the *surface impedance*:

$$Z_{xy}^{(\text{surf})} = \frac{E_x}{H_y} \text{ and } Z_{yx}^{(\text{surf})} = -\frac{E_y}{H_x} \quad (1)$$

The impedance terms in Eq. 1 can be looked upon as the coefficients of linear coupling relations between the electric and respective magnetic field components having the forms:

$$E_x = Z_{xy}^{(\text{surf})} H_y \text{ and } E_y = -Z_{yx}^{(\text{surf})} H_x \quad (2)$$

The subscripts on the impedance coefficients denote the ordering of the respective electric and corresponding orthogonal magnetic field components.

Under appropriate conditions, the *resistivity* of the uniform half-space in Fig. 1 may be determined from the measured surface impedance, since the *characteristic impedance* that a wave encounters traveling through a homogeneous medium is given by $Z_{\text{char}} = E_x/H_y = -E_y/H_x = \sqrt{i\omega\mu\rho_{\text{Tot}}}$ where $i = \sqrt{-1}$, ω is the radian frequency ($\omega = 2\pi f$, and $f = 1/T$ is the frequency in Hz, where T is the period in seconds), μ is the free-space magnetic permeability ($\mu = 4\pi \times 10^{-7}$ henries/m), and ρ_{Tot} is the total resistivity that accounts for both ohmic and dielectric effects in the medium.

The latter expressions for the ratios of the electric field to its orthogonal magnetic field counterpart, can be rearranged to estimate the resistivity of the medium according to $\rho_{\text{obs}} = (1/\omega\mu)|E_x/H_y|^2 = (1/\omega\mu)|E_y/H_x|^2$.

In general, however, the earth is *not* uniform, and as proposed by Cagniard (1953) in his classic paper, it is best to refer to the observed resistivity as an *apparent resistivity*, which might be determined using either combination of the respective field components, (E_x, H_y) or (E_y, H_x) . The corresponding apparent resistivities are commonly referred to as *Cagniard pairs*, and are given by

$$\rho_{xy}^{(\text{app})} = \frac{1}{\omega\mu} \left| \frac{E_x}{H_y} \right|^2 \text{ and } \rho_{yx}^{(\text{app})} = \frac{1}{\omega\mu} \left| \frac{E_y}{H_x} \right|^2 \quad (3)$$

For a uniform earth as well as for a plane-layered earth, as discussed below, one should expect that $\rho_{xy}^{(\text{app})} = \rho_{yx}^{(\text{app})}$. Exceptions to the latter denote the presence of lateral heterogeneities, also discussed below.

Principles of Depth Sounding

The *skin depth* or *depth of penetration* of an electromagnetic wave impinging on a uniform conductor (the earth) is given by $\delta \cong 500\sqrt{\rho \cdot T}$ meters, where ρ is the electrical resistivity of the medium in Ohm.m, and T is the period (inverse of frequency) of the signal in seconds. The consequence of the *skin-depth effect* is that low frequency (long period) waves penetrate to greater depth in a given conductor than high frequency (short period) waves (see ► “[Geoelectromagnetism](#)”).

The intrinsic resistivity of typical earth materials varies over many orders of magnitude, depending on mineralogy, water content, and temperature (see ► “[Electrical Properties of Rocks](#)”). Figure 2 summarizes the range of characteristic depths of penetration over the typical range of earth resistivities and frequencies employed in magnetotellurics. As an aside, note in the upper right corner of the figure, the roll-over of the linear isolines for high frequencies and high resistivities due to dielectric effects coming into play.

The *apparent resistivity* as a function of increasing period can be determined using spectral analysis of the measured electric and magnetic time series (see Vozoff 1985; see ► “[Magnetotelluric Data Processing](#)”). Then, from measurements of the apparent resistivity as a function of period, one can use the variable depth of penetration to electrically “probe” the earth to greater depth by progressively “tuning” the observed signals to progressively longer period.

For conventional magnetotelluric applications, at the highest frequencies ($\cong 1$ MHz) and the lowest resistivities (close to that of seawater, $1/3 \text{ } \Omega\text{-m}$, or lower), the signal (often from commercial or government radio transmitters, or, in the audio band, from distant lightning strikes) would penetrate to a few meters or few tens of meters depth. This range is useful for shallow geotechnical and groundwater investigations, and perhaps shallow geothermal work.

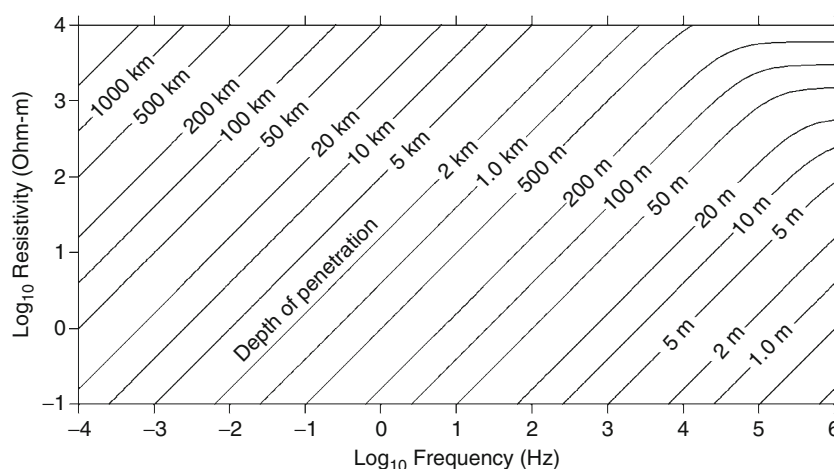
At the other end of the spectrum, at frequencies of a Hertz or lower ($T \geq 1$ s), workers usually rely on natural sources in the magnetosphere and ionosphere (such as micropulsations, magnetic substorms, and diurnal variations), (see ► “[Electromagnetic Pulsations and Magnetic Storms](#)”) to probe to depths of tens to hundreds of kilometers.

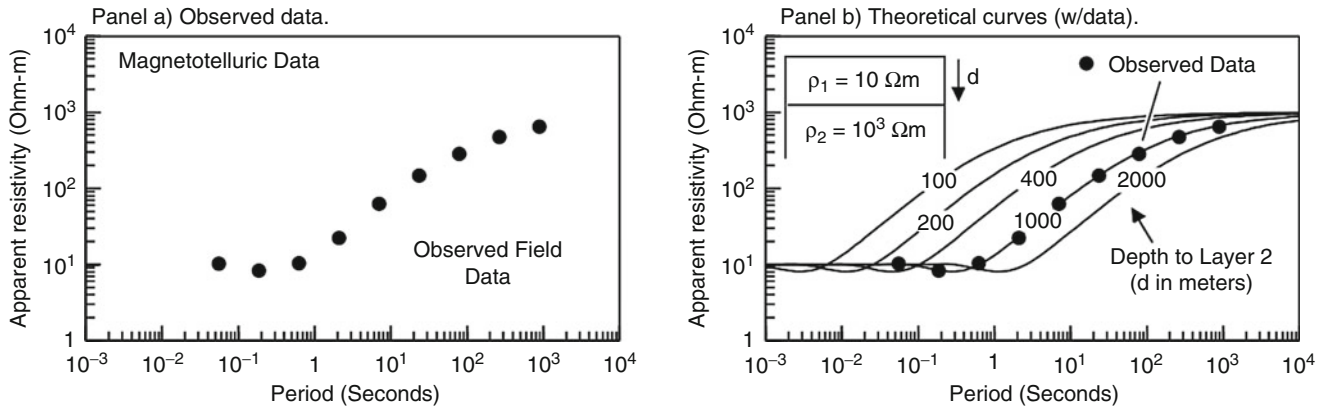
An example of the apparent resistivity that might be measured on the surface of a broad sedimentary basin as a function of period, as envisioned by Louis Cagniard in his 1953 paper, is shown in Fig. 3. The exploration geophysicist would have in mind that the relatively conductive sediments are underlain at some unknown depth by more resistive crystalline or metamorphosed basement (see ► “[Electrical Properties of Rocks](#)”). At an early stage, in the exploration of such a basin, the interpreter would like to determine three types of information from the observed MT data: *first*, the physical properties of the sedimentary basin (ρ_1 : the resistivity of layer 1); *second*, the depth d to the base of the sediments; *third*, the physical properties of the underlying basement (ρ_2 : the resistivity of layer 2). From the various models seen in Fig. 3b, a theoretical model with a surface resistivity of $10 \text{ } \Omega\text{-m}$, underlain by a more resistive $1,000 \text{ } \Omega\text{-m}$ layer at a depth of $1,000$ m, “fits” the observed data quite well. The mathematical framework through which one determines a model that best fits the set of observed data is known as *inverse theory* (see ► “[Magnetotelluric Data Processing](#),” ► “[Inverse Theory, Global Optimization](#),” ► “[Inverse Theory, Linear](#),” ► “[Inverse Theory, Monte Carlo Method](#),” and ► “[Inverse Theory, Singular Value Decomposition](#)”).

Effects of Lateral Heterogeneities

Early workers discovered the effects of lateral heterogeneities on MT data when they found that the Cagniard pairs were unequal: $\rho_{xy}^{(app)} \neq \rho_{yx}^{(app)}$ in Eq. 3. That is to say, the observed

Magnetotelluric Interpretation, Fig. 2 Depth of penetration as a function of frequency and resistivity of the host medium for geophysical applications

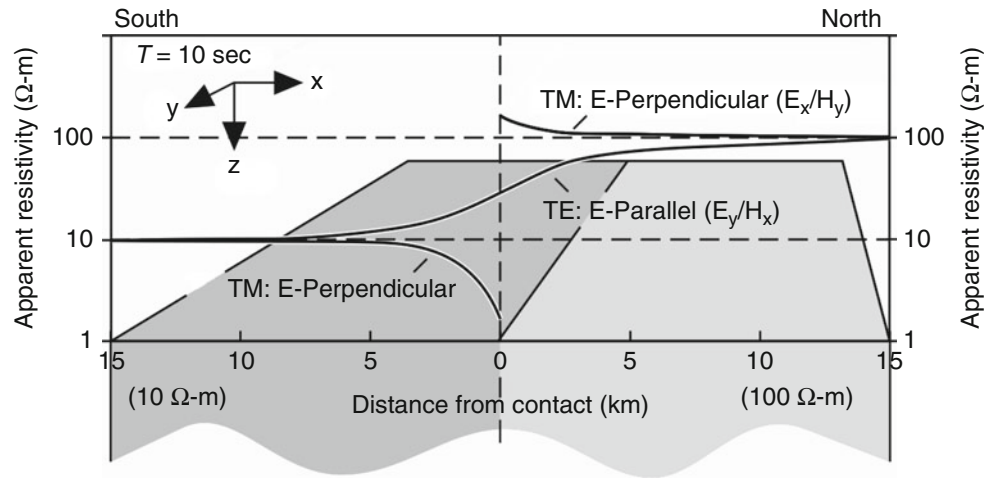




Magnetotelluric Interpretation, Fig. 3 (Panel a) An example of field data from magnetotelluric measurements. (Panel b) Field data overlaid on a set of theoretical curves representing various combinations

of model parameters, in this case the resistivity of each of two layers and the depth of layer 2

Magnetotelluric Interpretation, Fig. 4 Apparent resistivities along profiles transecting a vertical contact for two polarizations – TM and TE – of the source field relative to strike



apparent resistivity often depended on the direction in which the respective E and H fields were measured, as well as on the location of the measurements relative to local geologic features. Consider, for example, the case of the simple 2D vertical contact shown in Fig. 4. The observed apparent resistivities can be resolved into two principal modes: the incident electric source field may be polarized *perpendicular* to the strike of the contact, or it may be *parallel* to the strike of the contact in the y -direction.

Computing the corresponding apparent resistivities from the electromagnetic impedances for both cases, we obtain for the *TM* (Transverse Magnetic), or E-perpendicular mode:

$$Z_{xy}^{(\text{surf})} = \frac{E_x}{H_y}, \text{ leading to } \rho_{yx}^{\text{app}} = \frac{1}{\omega\mu} \left| \frac{E_x}{H_y} \right|^2 \quad (4)$$

and for the *TE* (Transverse Electric), or E-parallel mode

$$Z_{xy}^{(\text{surf})} = -\frac{E_y}{H_x}, \text{ leading to } \rho_{yx}^{\text{app}} = \frac{1}{\omega\mu} \left| \frac{E_y}{H_x} \right|^2 \quad (5)$$

Impedance matrix. Of course in actual field surveys one usually does not know ahead of time the trend or principal directions of local subsurface geologic features, so the electrode lines and magnetic sensors are usually deployed in standard geographic or geomagnetic coordinates. This leads to a “mixing” or “cross-coupling” of the field components. With reference to the linear coupling forms in Eq. 2, it is convenient to express these relations in the compact matrix form

$$\begin{bmatrix} E_x^{(p)} \\ E_y^{(p)} \end{bmatrix} = \begin{bmatrix} 0 & Z_{xy}^{(p)} \\ Z_{yx}^{(p)} & 0 \end{bmatrix} \begin{bmatrix} H_x^{(p)} \\ H_y^{(p)} \end{bmatrix} \quad (6)$$

where the superscript “p” denotes measurements in the principal coordinates of the feature (such as in the x - and the y -directions in Fig. 4). Rotating relation (6) through an arbitrary angle θ clockwise from the principal coordinates in Fig. 4 leads to a relation between the measured electric and magnetic field components of the form

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{bmatrix} \begin{bmatrix} H_x \\ H_y \end{bmatrix} \quad (7)$$

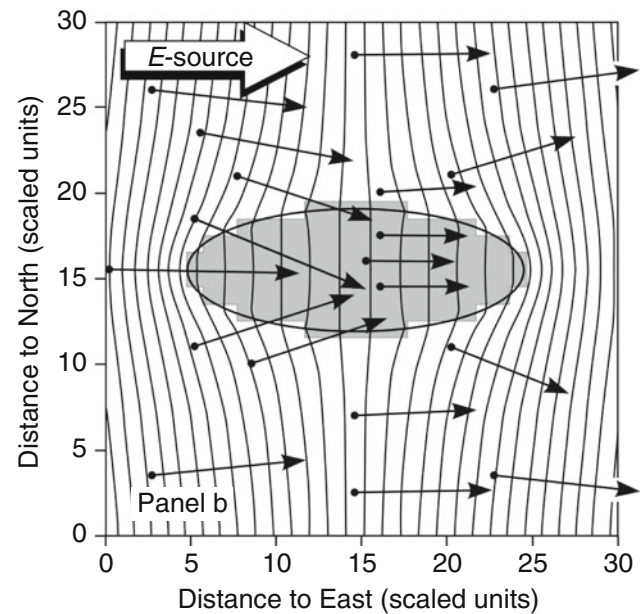
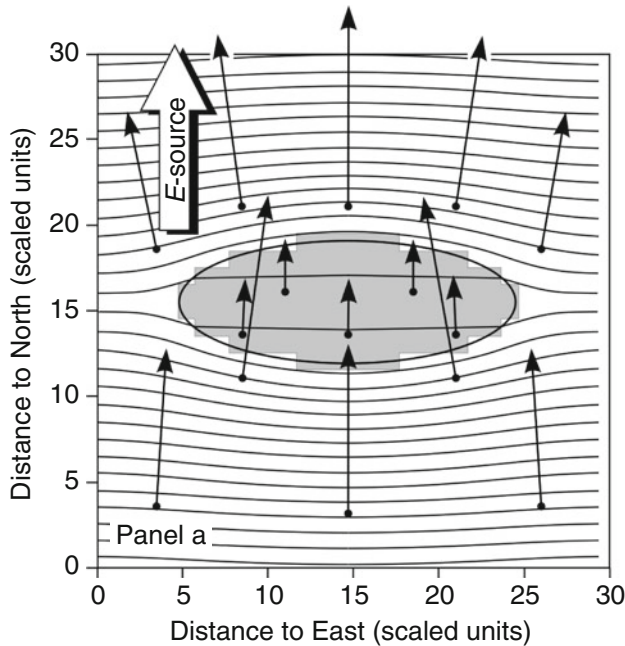
where the *impedance matrix* is often referred to as a *tensor impedance*.

Three-dimensional (3D) heterogeneities. Recognizing that the earth is, in general, 3D, electromagnetic workers have developed various approximate methods to dealing with it since the 1930s (see Keller and Frischknecht 1966). Moreover, the theory of electromagnetic induction in 3D laterally heterogeneous material has evolved significantly over the decades since 1980, largely through implementing physically representative numerical techniques (see the work cited for Stodt, Wannamaker, Mackie and Madden among others, in Simpson and Bahr 2005; see also Siripunvaraporn et al. 2005). Several of the principal effects of three dimensionality are illustrated in the example in Fig. 5 which is a plan view map of telluric (i.e., electric) field distortions in the vicinity of

a small idealized, elliptically shaped conductivity anomaly in the surface layer of an otherwise uniform or plane-layered earth. The embedded “elliptical” feature in the figure is a modest four times more conductive than the host material (although it is not uncommon for lateral contrasts to be an order of magnitude or more). The important point is that the electric field (and the associated magnetic field) can change dramatically in magnitude and direction over very small distances.

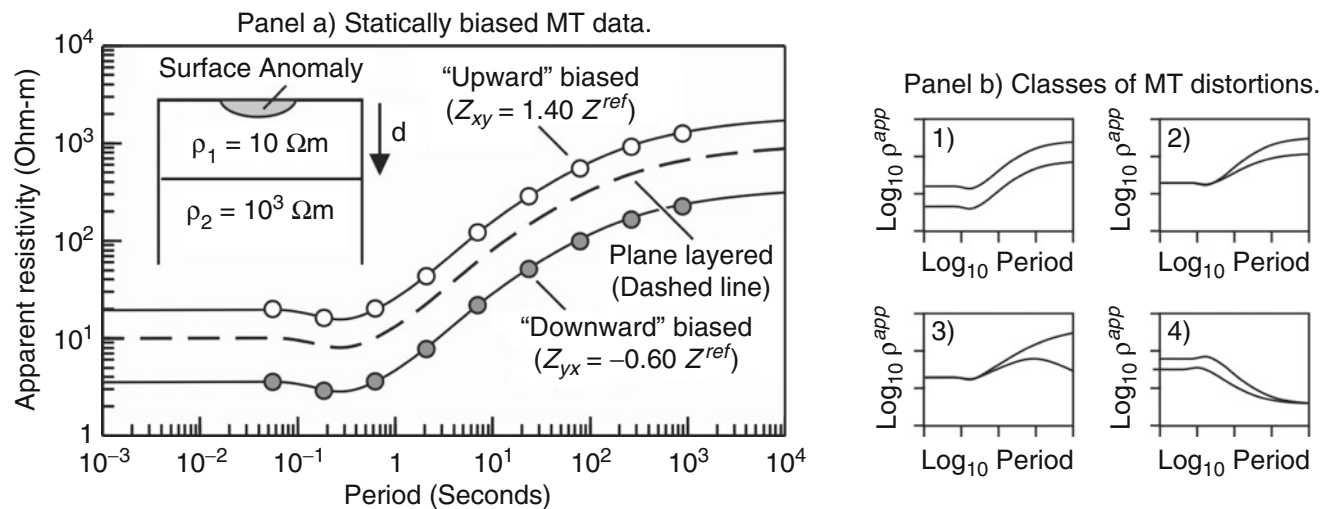
Effects of bias of MT data from lateral heterogeneities. Workers realized early in the development of magnetotellurics that linear coupling relations like Eq. 7 are most often the rule in field data (see the collection of reprints assembled by Vozoff 1985). The first step, therefore, in interpreting MT data often involves decomposing multiple independent sets of observations of both of the electric field components ($E_x^{(obs)}, E_y^{(obs)}$) and the associated magnetic field components ($H_x^{(obs)}, H_y^{(obs)}$) to extricate estimates of the principal values of apparent resistivity, ρ_{xy}^{app} and ρ_{yx}^{app} , in the principal directions. It is common practice to plot up both resistivity estimates as a function of period, as was done for the plane-layered case in Fig. 3. However, the effects of even a small heterogeneity on the MT response of an otherwise plane-layered earth is illustrated in Fig. 6, using as the background the plane-layered example in Fig. 3 (shown as the dashed line in Fig. 6).

The dashed curve in Fig. 6a represents the original plane-layered response of Fig. 3, along with the hypothetical effects



Magnetotelluric Interpretation, Fig. 5 Plan view map of telluric (or electric) field distortions in the vicinity of an idealized embedded conductivity anomaly. (Panel a) Voltage contours and selected electric

field directions for a north-directed electric source field. (Panel b) The same parameters as panel a, but for an east-directed source field



Magnetotelluric Interpretation, Fig. 6 Examples of static bias. (Panel a) Example to demonstrate the mechanism of "quasi-static shift" (referred to the dashed plane-layered case). (Panel b) Selected

classes of "distortions" of MT data due to various combinations of shallow and deep lateral heterogeneities, with static and inductive coupling effects

that might be encountered in MT field data acquired at a site within, or in the vicinity of, a small, local embedded surficial anomaly. For one of the principal polarizations, the affected apparent resistivity relative to the original plane-layered curve is "upward-biased" by almost a factor of 2, and for the orthogonal polarization the apparent resistivity is "downward-biased" to 36% of its undistorted value. (There is no reason, of course, that for some cases both $\rho_{xy}^{app}(T)$ and $\rho_{yx}^{app}(T)$ might be biased to higher values, or both biased to lower values.)

Such bias effects are endemic to magnetotelluric surveys, and are particularly challenging to those investigations that are intended to be "deep-sounding" of the crust and upper mantle (and the mantle transition zones), since it is so easy for an undetected static offset to bias the long period asymptotic values to larger, deeper, and more resistive – or smaller, shallower, and more conductive – values. If, along with the "static" offsets described above, there are 3D induction effects as well, the field data become increasingly more complicated to untangle. Figure 6b shows several selected classes of MT responses often observed in field data due to quasi-static and inductive coupling effects for various combinations of shallow and deep (or distant) lateral heterogeneities. While one must be cautious in drawing too much from a single set of sounding curves, these panels might illustrate the following generic phenomena. Panel b1 illustrates the type of static shift from a shallow surface feature discussed above. Panel b2, on the other hand, suggests a fairly uniform layer at shallow depth, but as the period and depth of penetration increases, the principal resistivities begin to separate indicating a deeper-seated inhomogeneity. One needs to be careful here, as the inhomogeneity may lay off to the side, and be a lateral effect, rather than a simple

depth effect. Panel b3 suggests a similar, but more complicated phenomenon in that the two resistivity modes do not simply separate at longer period, but their respective shapes are frequency dependent, indicating that there are significant induction effects as well as static shifts coming into play. Panel b4 is an idealized version of situations commonly encountered in real data that might be theoretically explained by a true anisotropic resistivity layer at the surface, with a more conductive layer at depth. Often, however, for real world data, the merging of the principal resistivity curves with increasing period is simply a precursor to their crossing and then shifting apart again at still longer periods. In all of these cases, drawing 1D inferences from data that are clearly affected by 2D and 3D structure is questionable. For other examples, see Vozoff (1985) and Simpson and Bahr (2005).

Current Strategies for MT Interpretation

Much of the progress on MT interpretation over the last 3 decades is due to a progressive improvement in data acquisition, robust processing of multi-site data, and careful 2D and 3D numerical modeling. Simultaneously, recording multi-station MT arrays have become the standard practice for industry, academia, and government surveys. Recognizing the importance of high-density coverage using simultaneous observations of MT fields at multiple sites, several national programs are developing multi-station MT arrays. Typical instrumentation for such facilities is described by Smirnov et al. (2008), with some of the preliminary results of the Earthscope MT array in the United States reported by Patro and Egbert (2008) (See also ► "Instrumentation, EM").

There is an increasing emphasis, when interpreting MT data, on either explicitly using 3D models, or at least on using 3D models as guides when using the more efficient, less computationally intensive, 2D inverse techniques. The earth, however, *is* three-dimensional, and most workers agree that any attempt to simplify its inherent complexity with 2D or even 1D plane-layered models, needs to be assessed at some level with effective and representative 3D models. Such models have progressively evolved from the initial contributions reported in the reprint collection of Vozoff (1985). A number of notable reports on 3D numerical models – modeling, inversion, and applications – are contained in the compendium of Oristaglio and Spies (1995), with additional developments by Mackie et al. (1993), Siripunvaraporn et al. (2005), and Tong et al. (2009).

In spite of the promise of full 3D MT interpretations, most recognize that the inversion of data from a large regional array of broadband stations in terms of generalized 3D models tends to be computationally intensive (Newman and Alumbaugh 2000; Zhdanov and Tolstaya 2004). Thus, most surveys interpret array data in terms of 2D cross-sections. In many cases, the response of a 3D feature is locally quite 2D in nature. In other cases, it may be possible that the type of localized static shift discussed in an earlier section can be compensated for, and the observed coupling matrix can be decomposed into regionalized principal elements (Groom and Bailey 1989). Berdichevsky et al. (1998) summarize several of the major concepts that many interpreters employ when analyzing MT data from 2D (and 3D) environments. First, while the *TM* (*E*-perpendicular) mode is often quite sensitive to local near-surface features, the *TE* (*E*-parallel) mode is often more indicative of deeper, more regional features. Second, while the *TM* mode is more susceptible to the effects of static shift, it is not uncommon that the *TE* mode may be relatively undistorted. Both modes, however, are essential for a comprehensive interpretation of a study area, but must be used prudently. Ledo et al. (2002) discuss the limitations as well as the computational advantages – when appropriate – of applying 2D model inversions to actual 3D field situations.

For many investigations, the concomitant 2D and 3D effects might be considered unwanted “distortions” of deep-sounding data, and a bane to those studies intended to better understand the deep crust, or to delineate the depth to the lithosphere–asthenosphere boundary, the mantle transition zones, and so forth, on regional or sub-global scales. On the other hand, there is indeed a great deal of information in MT data that could be essential for characterizing the full 3D character of the earth, and the method, when properly executed, is proving to be a powerful means for delineating geologic features having application to a variety of fundamental geophysical and geological problems. However, it should be clear from the example in Fig. 5 that in order to identify, characterize, and account for the effects of localized surficial features, one needs to deploy a sufficient density of

sites to spatially characterize the surface field patterns. The situation is quite analogous to the “aliasing” problem that governs the strategy for properly sampling a time series; only in MT we are not dealing with simply a 1D sequence, but rather with 2D areal coverage. Berdichevsky (1999) offers a thoughtful critique of several of the underpinning paradigms and procedures of modern magnetotellurics.

Case Studies

A number of early applications of the MT method are described in the texts of Patra and Mallick (1980), and Vozoff (1985). More recent case studies are described in the text by Simpson and Bahr (2005). The most exciting of these involve multidisciplinary geophysical field efforts. Particularly, noteworthy is the series of integrated MT, seismic refraction, and gravity investigations initiated by the US Geological Survey in the Pacific Cascades volcanic belt in the Northwest US (see the review by Stanley et al. 1990). Early phases of this work led to the first major international offshore/onshore MT and geomagnetic deep-sounding (GDS) project: the EMSLAB project that extended from the offshore Juan de Fuca Rise to the east side of the Cascades volcanic system (*c.f.*, Wannamaker et al. 1989). Work in this area continues. For example, in a reanalysis of data from the EMSLAB Project, Vanyan et al. (2002) articulate the experience of many that when interpreting a profile of widely spaced, multi-site MT data in such complicated terrains, one needs to be selective in which elements of the MT impedance matrix are most consistent when developing a 2D inversion. They found that under the complex geologic conditions of the region, the strict inversion of both the *TE* and *TM* modes simultaneously at each and every site yielded an intricate, but what is likely a geophysically implausible, pattern of conductive and resistive features resulting in a large misfit of the model response to the data. The analysis improved significantly, they felt, when they were more selective in which of the principal impedance components were employed at a site. They also found that using lateral variations of the transient magnetic field, in conjunction with the MT data, significantly complemented their analysis. This point of view probably captures the spirit of most current MT investigations, and certainly underscores the importance of multiple station arrays of closely spaced sites.

In other regions, Sarma et al. (2004) report on a comprehensive MT study of an active seismic zone in western India. Solon et al. (2005) describe a large array MT study of the deep crust beneath the Himalayas in central Tibet. Munoz et al. (2008) investigate mid-crustal conductivity anomalies using 3D models of SW Europe. Wannamaker et al. (2008) describe the application of MT to understanding lithospheric deformation and magmatic processes along the Great Basin–Colorado Plateau transition zone. Continuing the work in the US Pacific

Northwest, Patro and Egbert (2008) report on their analysis of data from deployments in 2006 and 2007 of the still-evolving US Earthscope MT Transportable Array. Data were sequentially acquired in the period range from 10 to 10,000 s at 110 sites, covering the region at a nominal station spacing of 75 km. Han et al. (2009), using a 3D model to interpret MT data on the Korean peninsula, emphasize the importance of accounting for the effects of the nearby ocean and coastline. Eaton et al. (2009), in pulling together a variety of geophysical evidence, present a recent update on MT investigations of the lithosphere and asthenosphere beneath cratons, particularly for North America. Similar applications of MT to understanding significant solid earth problems are ongoing in China, Japan, Europe, Great Britain, South America, Antarctica, Africa, New Zealand, and Australia. Clearly, magnetotellurics is becoming an essential tool for a better understanding of earth systems on a variety of scales.

Summary

The magnetotelluric (MT) method is based on the physical theory of electromagnetic wave propagation in electrically heterogeneous materials. Using the “skin-depth” principle that, in a given conductor, long period signals penetrate to greater depth than short period signals, the method typically employs the rich spectrum of natural electromagnetic energy from the ionosphere, magnetosphere, and atmosphere (distant lightning storms) to “probe” the electrical structure of the earth from a few hundred meters depth to tens even hundreds of kilometers. For near-surface applications, active transmitters are used throughout the audio band (100 Hz to 20 kHz) to above a megahertz. The last decades have seen significant developments in the type of low-noise, large dynamic range instrumentation needed for such studies, along with robust real-time data processing, 3D numerical modeling, and geophysical inverse theory. The MT method involves recording and analyzing the spectral character of the complex amplitudes and phases of the electric and magnetic field components over a wide range of periods from an array of closely spaced sites on the earth’s surface. From the spatial and spectral character of such data, one can, in a sense, deconvolve the signatures of lateral and vertical variations in the bulk properties of the subsurface to develop insight on fundamental processes in the earth.

Cross-References

- [Electrical Properties of Rocks](#)
- [Electrical Resistivity Surveys and Data Interpretation](#)
- [Electromagnetic Pulsations and Magnetic Storms](#)
- [Goelectromagnetism](#)
- [Instrumentation, Electrical Resistivity](#)

- [Instrumentation, EM](#)
- [Inverse Theory, Global Optimization](#)
- [Inverse Theory, Linear](#)
- [Inverse Theory, Monte Carlo Method](#)
- [Inverse Theory, Singular Value Decomposition](#)
- [Magnetotelluric Data Processing](#)

Bibliography

- Adam A (Editor-in-Chief) (1976) Goelectric and geothermal studies (East Central Europe, Soviet Asia), KAPG geophysical monograph. Akademiai Kiado, Budapest
- Berdichevsky MN (1999) Marginal notes on magnetotellurics. *Surv Geophys* 20:341–375
- Berdichevsky MN, Dmitriev VI, Pozdnjakova EE (1998) On two-dimensional interpretation of magnetotelluric soundings. *Geophys J Int* 133(3):585–606
- Cagniard L (1953) Basic theory of the magnetotelluric method of geophysical prospecting. *Geophysics* 18:605–635
- Eaton DW, Darbyshire F, Evans RL, Gruetter H, Jones AG, Yuan X (2009) The elusive lithosphere-asthenosphere boundary (LAB) beneath cratons. *Lithos* 109(1–2):1–22
- Groom RW, Bailey RC (1989) Decomposition of magnetotelluric impedance tensors in the presence of local 3-dimensional galvanic distortion. *J Geophys Res Solid Earth Planets* 94(B2):1913–1925
- Han N, Nam MJ, Kim HJ, Lee TJ, Song Y, Suh JH (2009) Three-dimensional inversion of magnetotelluric data including sea effects obtained in Pohang, Korea. *J Appl Geophys* 68(4):533–545
- Kaufmann AA, Keller GV (1981) The magnetotelluric sounding method. Elsevier, New York
- Keller GV, Frischknecht FC (1966) Electrical methods of geophysical prospecting. Pergamon, New York
- Ledo J, Queralt P, Marti A, Jones AG (2002) Two-dimensional interpretation of three-dimensional magnetotelluric data: an example of limitations and resolution. *Geophys J Int* 150:127–139
- Mackie RL, Madden TR, Wannamaker PE (1993) 3-Dimensional magnetotelluric modeling using finite difference equations and comparisons to integral-equations solutions. *Geophysics* 58(2):215–226
- Munoz G, Mateus A, Pous J, Heise W, Santos FM, Almeida E (2008) Unraveling middle-crust conductive layers in Paleozoic Orogens through 3D modeling of magnetotelluric data: the Ossa-Morena Zone case study (SW Iberian Variscides). *J Geophys Res Solid Earth Planets* 113(B6):B06106
- Newman GA, Alumbaugh DL (2000) Three-dimensional magnetotelluric inversion using non-linear conjugate gradients. *Geophys J Int* 140(2):410–424
- Oristaglio M, Spies B (eds) (1995) Three-dimensional electromagnetics. In: Proceedings of an international symposium in Honor of Jerry Hohmann, Schlumberger-Doll Research, Cambridge, MA
- Patra HP, Mallick K (1980) Geosounding principles, 2; Time-varying geoelectric soundings. Elsevier, New York
- Patro PK, Egbert GD (2008) Regional conductivity structure of Cascadia: preliminary results from 3D inversion of U.S. Array transportable array magnetotelluric data. *Geophys Res Lett* 35:L20311. <https://doi.org/10.1029/2008GL035326>
- Sarma SVS, Prasanta B, Patro K, Harinarayana T, Veeraswamy K, Sastry RS, Sarma MVC (2004) A magnetotelluric (MT) study across the Koyna seismic zone, western India: evidence for block structure. *Phys Earth Planet Inter* 142(1–2):23–36
- Simpson F, Bahr K (2005) Practical magnetotellurics. Cambridge University Press, Cambridge
- Siripunvaraporn W, Egbert G, Lenbury Y, Uyeshima M (2005) Three-dimensional magnetotelluric inversion: data space method. *Phys Earth Planet Inter* 150:3–14

- Smirnov M, Korja T, Dynesius L, Pedersen LB, Laukkanen E (2008) Broadband magnetotelluric instruments for near-surface and lithospheric studies of electrical conductivity: a fennoscandian pool of magnetotelluric instruments. *Geophysica* 44:31–44
- Solon KD, Jones AG, Nelson KD, Unsworth MJ, Kidd WF, Wei W, Tan H, Jin S, Deng M, Booker JR, Li S, Bedrosian P (2005) Structure of the crust in the vicinity of the Banggong-Nujiang suture in central Tibet from INDEPTH magnetotelluric data. *J Geophys Res Solid Earth Planets* 110(B10):1–20
- Stanley WD, Mooney WD, Fuis GS (1990) Deep crustal structure of the Cascade Range and surrounding regions from seismic refraction and magnetotelluric data. *J Geophys Res Solid Earth Planets* 95(B12):19419–19438
- Tong XZ, Liu JX, Xie W, Xu LH, Guo RW, Cheng YT (2009) Three-dimensional forward modeling for magnetotelluric sounding by finite element method. *J Cent S Univ Technol* 16(1):136–142
- Vanyan LL, Berdichevsky MN, Pushkarev PY, Romanyuk TV (2002) A geoelectric model of the Cascadia subduction zone. *Izvestiya Phys Solid Earth* 38(10):816–845
- Vozoff K (ed) (1985) Magnetotelluric methods, Geophysics reprints series no. 5. Society of Exploration Geophysicists, Tulsa
- Wait JR (1982) *Geo-electromagnetism*. Academic, New York
- Wannamaker PE, Booker JR, Filloux JH, Jones AG, Jiracek GR, Chave AD, Tarits P, Waff HS, Egbert GD, Young CT, Stodt JA, Martinez M, Law LK, Yukutake T, Segawa JS, White A, Green AW (1989) Magnetotelluric observations across the Juan de Fuca subduction system in the EMSLAB project. *J Geophys Res Solid Earth Planets* 94(B10):14111–14125
- Wannamaker PE, Hasterok DP, Johnston JM, Stodt JA, Hall DB, Sodergren TL, Pellerin L, Maris V, Doerner WM, Groenewold KA, Unsworth MJ (2008) Lithospheric dismemberment and magmatic processes of the Great Basin-Colorado Plateau transition, Utah, implied from magnetotellurics. *Geochem Geophys Geosyst* 9:Q05019
- Zhdanov M, Tolstaya E (2004) Minimum support nonlinear parametrization in the solution of a 3D magnetotelluric inverse problem. *Inverse Probl* 20(3):937–952

Magnetotellurics, Crustal Imaging

Prasanta K. Patro

CSIR-National Geophysical Research Institute, Hyderabad, India

Definition

Magnetotelluric (MT) imaging allows electrical characterization of the deep-crust and upper mantle, and can delineate and define fundamental blocks of crust, as well as major structural features like faults, suture zones, etc. The continental crust has been formed from mantle material over the lifetime of the Earth by a series of processes indicating melting, crystallization, metamorphic, erosional, depositional, subduction, and endless reworking events. Magnetotelluric method is sensitive to the conductive phases such as partial melt, fluids, and graphite. Therefore, MT provides highly useful models on electrical conductivity of the crust which serve as

complementary to other geophysical models such as seismic and thus helps in deriving realistic crustal models. MT has indeed been applied extensively to study several of the tectonically active regions including subduction zones, continental collision, rift zones, and also cratonic regions. There are many MT/LMT studies carried out globally to understand the crustal electrical signature related to different tectonic settings; however, in the following I have presented few representative examples.

Costa Rican Subduction Zone

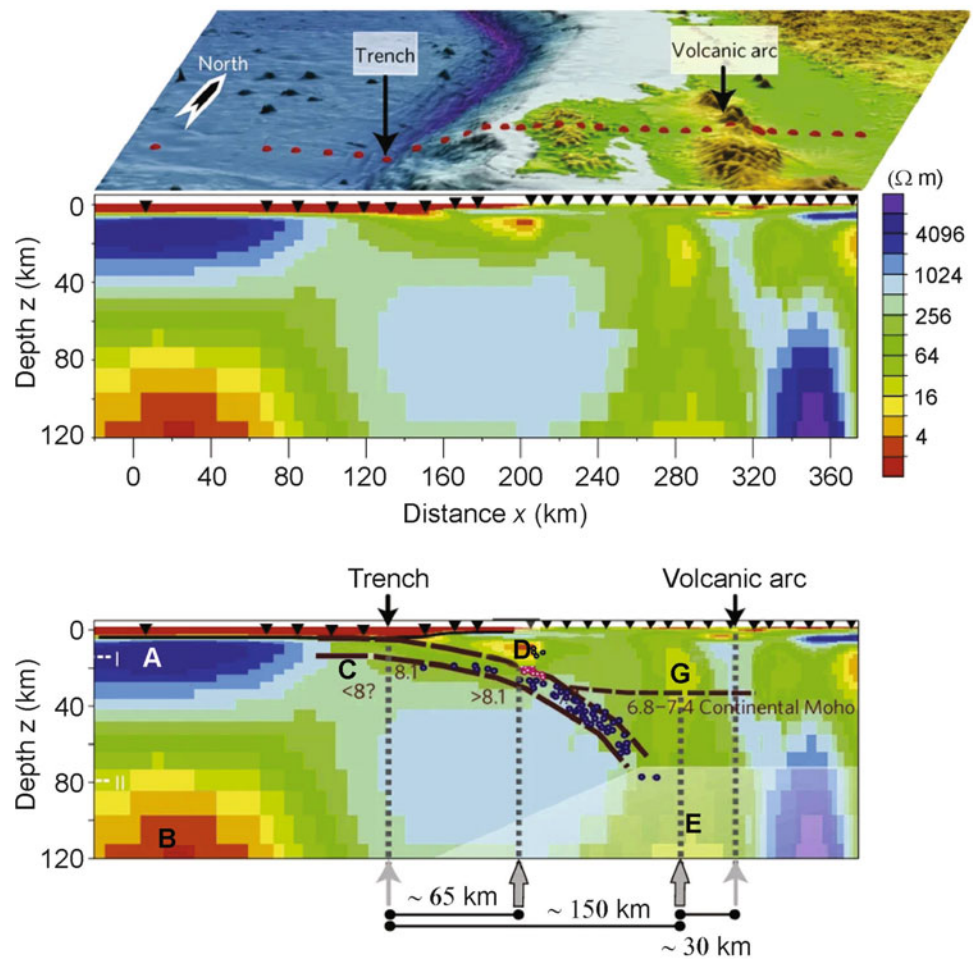
Land and Marine MT studies were carried out to image the deep Costa Rican subduction zone (Worzewski et al. 2011) focusing on the hydration–dehydration cycle. Marine MT studies revealed the oceanic lithosphere – asthenosphere configuration which extends from shallow low resistive water saturated sediments lying over high resistive pillow lavas and sheeted dykes (A) which extend up to about 40 km depth reaching the upper mantle. At still deeper depths of about 80 km, a major low-resistive anomaly (B) is found which is attributed to the presence of asthenosphere with minor melt fraction.

In these zones prior to the subduction, during the bending of the subducting slab, deep faults are formed at mantle depths and water enters into these faults resulting in the hydration. This water is stored partially in the fractures and pores in oceanic sediments, crust, and serpentinized mantle. In the present case, the observed decrease in the resistivity (C) in the bending zone (Fig. 1) is interpreted to represent the serpentinized mantle rock. In contrast, during the subduction phase, water is released from cracks and pores when plate subducts to deeper levels, and this is referred to dehydration. While water released due to shallow phase transition reactions in sediments are stored at shallow depths (D), the dehydration of crust and mantle at deeper levels results in the deep-crustal fluid accumulation leading to generation of anomalous conductive regions (see G in Fig. 1). The conductive anomaly (G) thus represents the locally existent fluid accumulation associated with the fundamental dynamic processes in subduction zones. Tectonically, the deep crustal anomalous zone G may be considered to be closely linked to the development of surface expressions such as Volcanic arcs as shown in Fig. 1.

Ethiopian Rift and Afar

Afar and Ethiopian rift regions are generally considered to provide a natural laboratory to study the process of continental rifting and formation of new oceanic basin as a result of seafloor spreading. Dabbahu magmatic segment in the Afar region, Ethiopia, has been active since 2005 and witnesses

Magnetotellurics, Crustal Imaging, Fig. 1 Electrical image of Costa-Rica subduction zone from Land and Marine MT studies (Worzewski et al. 2011)



several dyke intrusion activities in central and southern parts of the segment. Desissa et al. (2013) carried out MT imaging of the region, using the TE and TM mode joint inversion using data space Occam inversion to obtain the resistivity structure of the crust/mantle column. Major finding in this study is the imaging of a large partial melt zone associated with the magmatic activity in the crust which is responsible for the dyke injections and volcanic eruptions in this region (see Fig. 2). The conductive pathways which are linked to the deep chamber are also delineated and they are shown to be correlating with the surface geological expressions such as volcano and dyke injections. B in the figure is the projection of Badi volcano in the study area on to the profile.

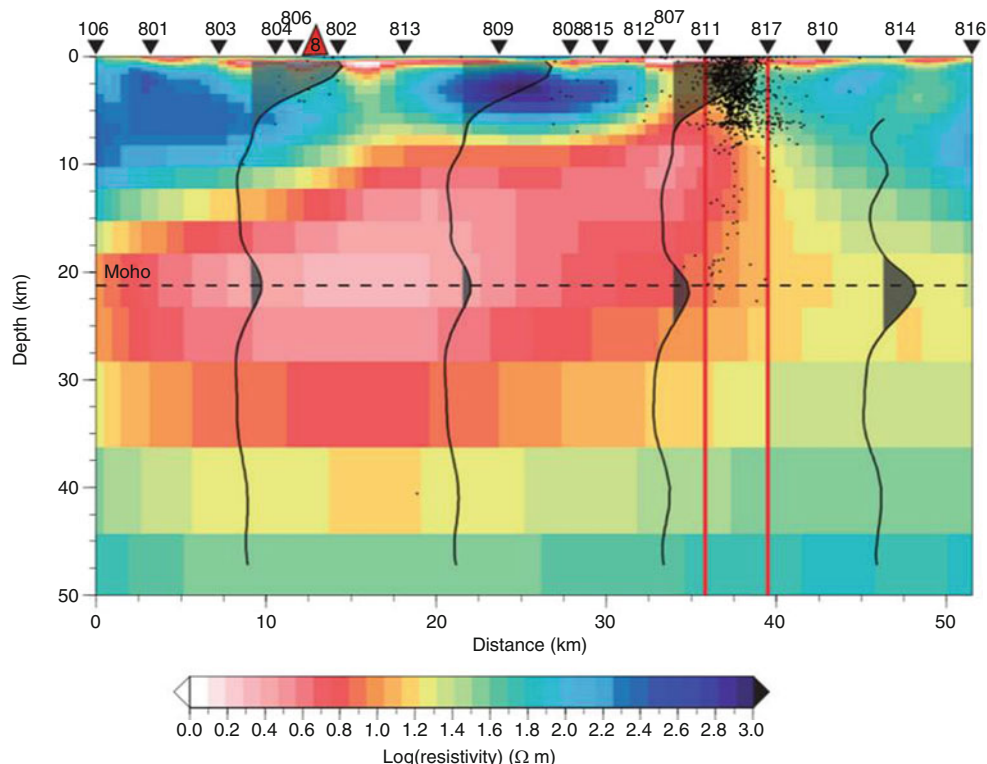
Himalayas

Himalayan mountain range is the classic example of the continent – continent collision and the resultant mountain building scenario. It formed as a result of collision between the Indian plate and the Eurasian plate during the Cenozoic

and it is still active. The Higher Himalayan mountain ranges exposing the high-grade gneisses were formed as the thrust sheet of Indian continental basement was displaced towards south along the MCT. Tethys Himalaya comprises Tethys sediments covering the Higher Himalayan crystallines (HHC). Indus suture zone comprises early tertiary granitoids of Ladakh plutonic complex. By inverting TM mode and Tipper data, Arora et al. (2007) observed a relatively resistive upper crust ($>300 \Omega m$) which represents granitic rocks of Ladakh Batholith and crystalline gneisses of Tethyan Himalayas. Besides this, the conductors were observed on both sides of the resistive Tso-Morari dome and these are interpreted as obducted ophiolite slab along the Main Zaskar back thrust.

In the MT models, the Indus suture zone is observed as a conductive zone dipping in the northeast direction and it merges with another conductive zone ($\sim 30 \Omega m$) in the mid-crustal column. But, Indus-Tsangpo Suture Zone (ITSZ) is not expressed as a conductive zone in the southern Tibet (Unsworth et al. 2005) as also in Leh region. This difference is explained as due to the weak postcollision deformation and

Magnetotellurics, Crustal Imaging, Fig. 2 Electrical signature of Ethiopian Rift system derived from magnetotellurics (Desissa et al. 2013)



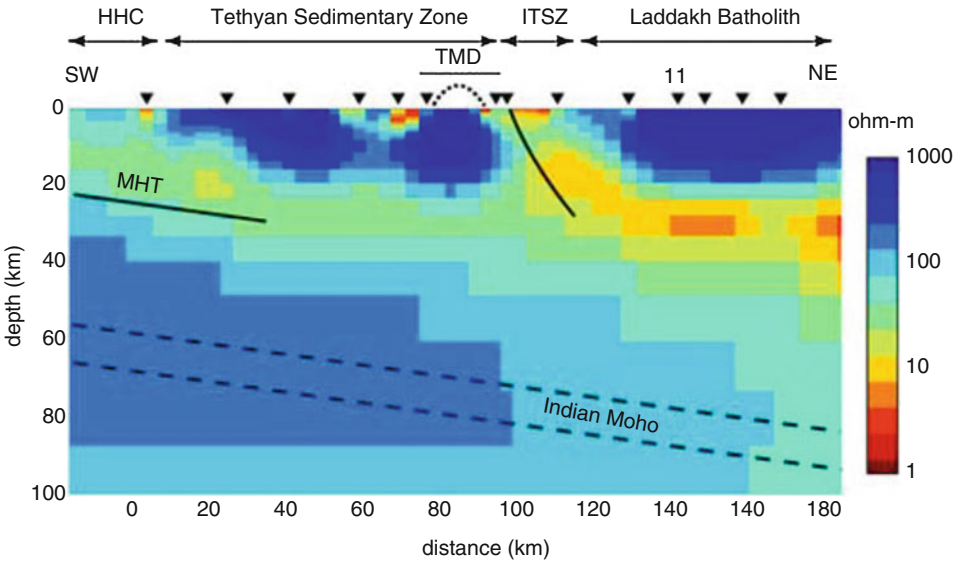
comparatively slower convergence which might have preserved this near subvertical structure and extrusion along the southern margin of Tibet and this might be the reason for the absence of ITSZ in southern Tibet. The Main Himalayan Thrust (MHT) is observed as only moderately conductive zone dipping in the northeast direction in mid crust. Low resistivity is interpreted to be due to the accreted sedimentary rocks overlying the under thrust Indian plate. North of the Indus Suture zone, resistivity decreases further to 5–10 Ωm and it is interpreted as due to the presence of fluids or presence of a few percent partial melt beneath or the combination of both. The underlying lithosphere column is observed to be resistive reflecting the signature of crystalline rocks. Indian lithosphere extends to the Indus suture zone and tends to become conductive towards further north and this is interpreted to be due to shallowing of the asthenosphere in the region (Fig. 3).

San Andreas Fault

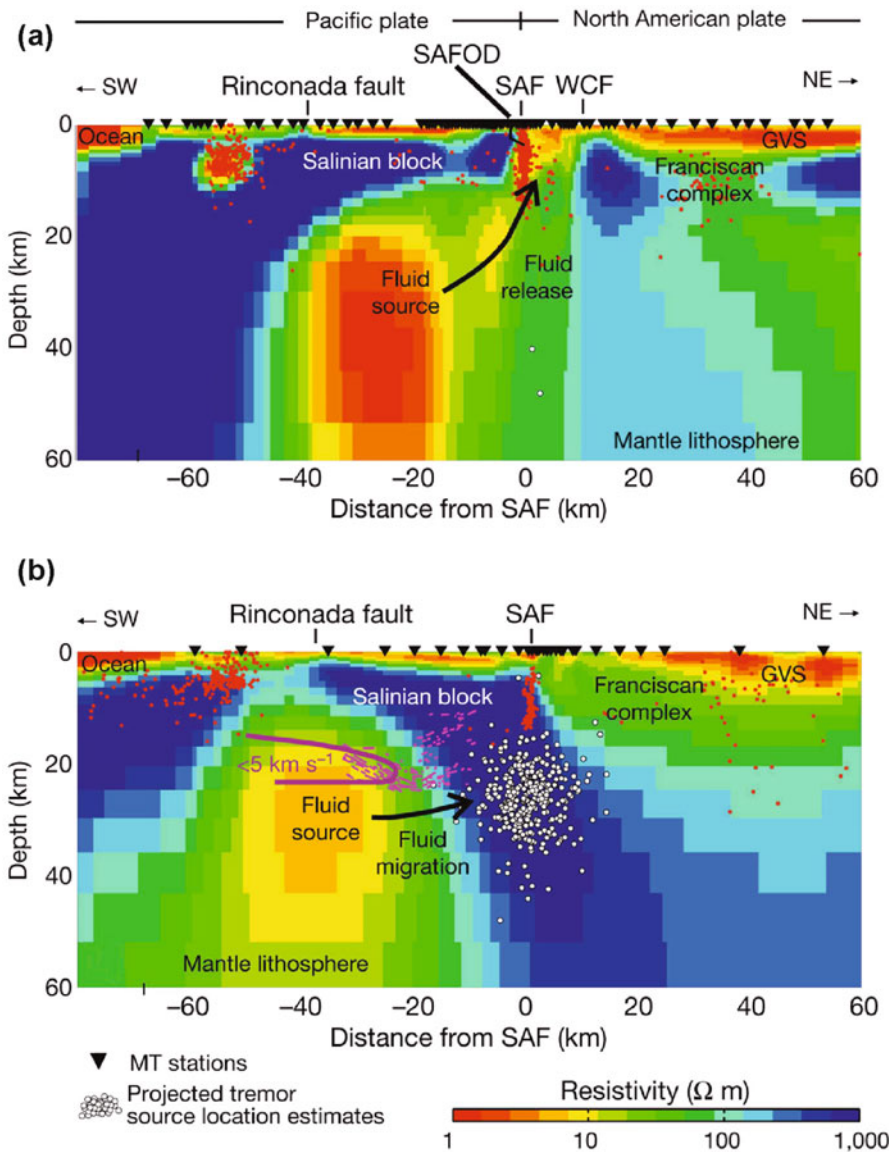
San Andreas Fault (SAF) is a continental transform fault acting as a tectonic boundary between Pacific plate and North American plate. Presence of fault-weakening fluids and frictionally weak minerals under the brittle crust are considered as the possible causes for the anomalously weak nature of the fault NW of Parkfield, California. Seismically active zones are observed near the Parkfield and the Cholame,

California. Nonvolcanic seismic tremor activities at lower crustal and upper mantle depths are observed in this region. MT studies are carried out in this region to understand the reasons behind the phenomenon and to evaluate the hypotheses (Becken et al. 2011). MT studies provided evidence for the fluid pressure induced seismicity near Cholame region. SAF near Cholame is observed to exhibit an anomalous resistivity feature extending to a depth of 20 km thus exceeding the brittle–ductile transition. Northwest of the Cholame, near SAF Observatory, an anomalously conductive zone (1–30 Ωm) at Depth was observed, extending from deeper to shallow levels in the region. The deep conductive region is interpreted as a fluid-filled zone formed as a result of dehydration of the serpentinized mantle wedge material. But near Cholame, the conductive fluid source region is bounded from above by the high resistive ($\sim 1000 \Omega\text{m}$) rocks, and interestingly the seismic tremors are concentrated in the high-resistive segments. So, the seismic activity is interpreted as due to the high fluid pressure and rock failure in mechanically strong region of high electrical resistivity as fluids migrate from a permeable, mechanically weak source to a mechanically strong region of high-electrical resistivity. Thus, hydrofracturing or reaction enhanced permeability are the suggested mechanisms for this phenomenon operating under high confining pressure in the lower crust and upper mantle. So, the periodic tremor activity observed close to the fluid source reflects the failure mechanism occurring due to the cyclic fluid pressure accumulation and release (Fig. 4).

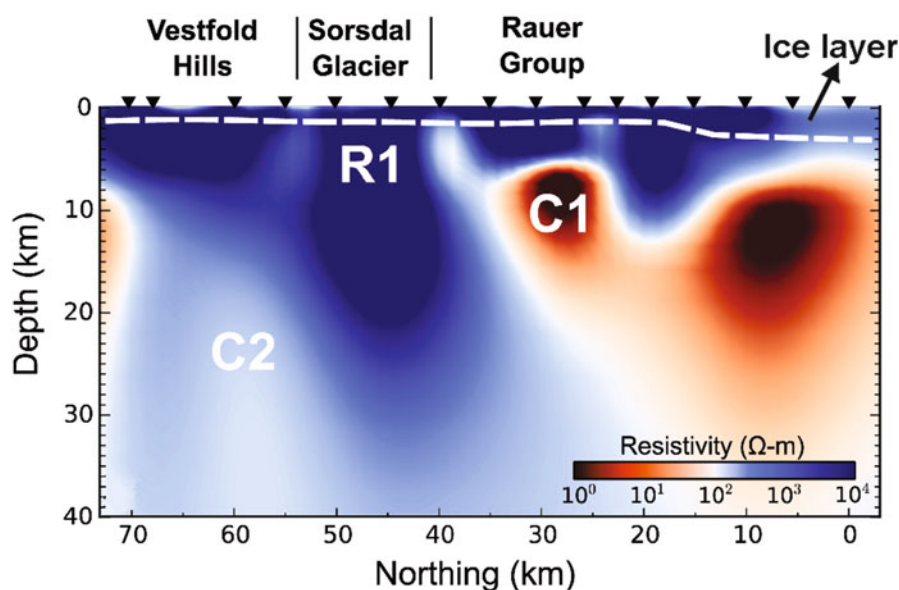
Magnetotellurics, Crustal Imaging, Fig. 3 Two-dimensional electrical image of the Himachal Himalayas covering the Tethyan sedimentary zone, ITSZ and Laddakh Batholiths (Arora et al. 2007)



Magnetotellurics, Crustal Imaging, Fig. 4 Two-dimensional resistivity cross-section at SAFOD (a) and near Cholame (b) (Becken et al. 2011)



Magnetotellurics, Crustal Imaging, Fig. 5 3D resistivity cross-section across Vestfold Hills and Rauer group, Antarctica (Modified from Peacock and Selway 2016)



Antarctica (Vestfold Hills and Rauer Group)

Archean to Neoproterozoic geological histories suggest the presence of a suture zone between Vestfold Hills and Rauer group beneath Sørsdal Glacier. MT studies were conducted to identify the exact location of the suture zone (Peacock and Selway 2016). Figure 5 shows the 3D model obtained. The top Ice cap of around 2 km thickness appears as a layer of resistivity less than the typical resistivity of pure glacier ice ($100,000 \Omega\text{m}$). The upper crust representing the crystalline rocks at crustal temperatures is moderately resistive ($<2000 \Omega\text{m}$). The conductive anomaly C1 appears at a depth of 6 km is interpreted to be due to sulfide minerals, sulfide grain-boundary film, and/or graphite grain boundary film within the dykes containing iron-rich and hydrogen-rich amphibole and clinopyroxene. The resistive zone R1, associated with suture, delineated beneath the Sørsdal Glacier is interpreted as the zone formed due to oxidation of the magnetite, sulfide, and graphite because of the high-temperature metamorphism occurred during suturing.

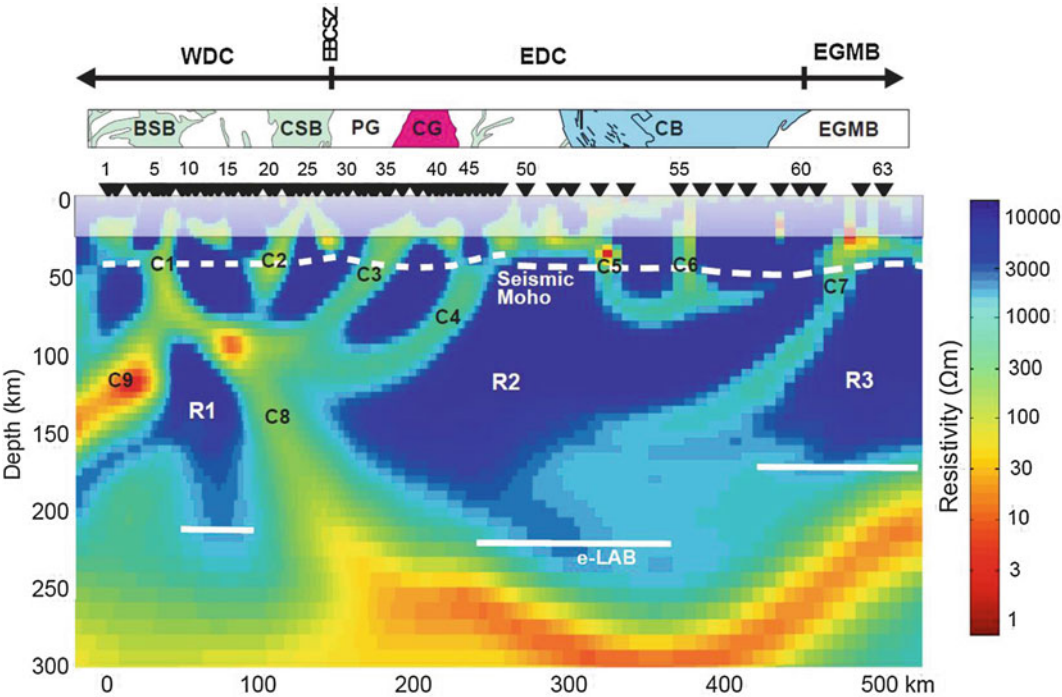
Dharwar Craton, India

MT and LMT studies were carried out across the Dharwar craton (DC) from the west coast to east coast in south India and also across the Eastern Ghat Mobile Belt (EGMB) to investigate the crustal and lithospheric structure and to identify the tectonic boundaries of different geologic domains (Malleswari et al. 2019). Modeling results brought out seven conductive features C1–C7 having resistivities between 100 and $1000 \Omega\text{m}$ within the highly resistive ($10^4 \Omega\text{m}$) background crustal column. These features are spatially

correlatable with the surface exposures of schist belts or granitic intrusions in the region and are interpreted as fossil pathways of the volcanic emplacements associated with greenstone belt and granite suite formations. The near-vertical conductive feature C8 under Chitradurga Shear Zone extending deep into the lithosphere is interpreted as mantle suture zone separating the western and eastern blocks of Dharwar Craton. Thick ($\sim 200 \text{ km}$) highly resistive lithosphere show typical Cratonic signature under Dharwar Craton, but intense lithospheric modification is also observed beneath western Dharwar craton. A major west-verging conductor identified beneath EGMB is interpreted to be the signature of Proterozoic collision process between the Indian landmass and East Antarctica. Lithosphere beneath EGMB is comparatively thinner ($\sim 120 \text{ km}$) and this might be due to Cretaceous rifting during break up of Gondwana supercontinent that resulted in the separation of India and Antarctica (Fig. 6).

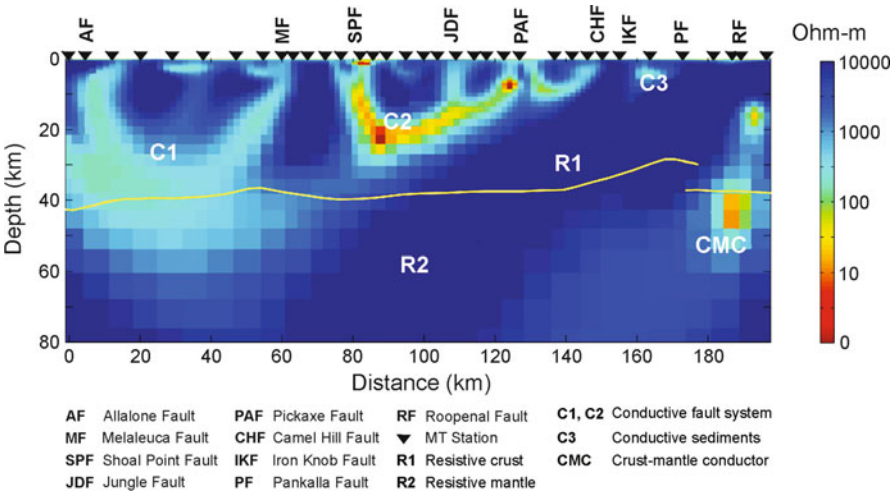
Gawler Craton, Australia

A long-period magnetotelluric data set was obtained in 2005 across the western part of the late Archaean-early Proterozoic Gawler Craton, South Australia, with an aim of delineating the electrical conductivity structure of the crust and upper mantle underneath (Curtis and Thiel 2019). The MT measurements were made along an east-west trending profile extending from the Gawler Range Volcanics in the east, crossing the Nuyts Domain and the highly prospective Meso-Proterozoic Fowler Domain, and terminating in the Eucla Basin in the west. The electrical model shows a very high resistive crust and upper mantle underneath the Nuyts



Magnetotellurics, Crustal Imaging, Fig. 6 Electrical image of the Dharwar Craton, India (Malleswari et al. 2019)

Magnetotellurics, Crustal Imaging, Fig. 7 Two-dimensional resistivity model across the Gawler Craton, Australia (modified from Curtis and Thiel 2019)



and Fowler Domain, possibly representing the cratonic root of the Gawler Craton extending to depths of ~160 km. The resistive cratonic root is closer to the surface underneath the Fowler Domain compared to that in the Nuyts Domain which supports the findings derived from the study of outcrops of meta-sediments of higher metamorphic grade in the Fowler Domain. A subvertical conductor marks the western terminus of the Fowler Domain, and the feature is imaged to upper mantle depths. On the eastern side of the Fowler Domain, another subvertical conductor extends to similar

depths. These features spatially coincide with the Tallacootra and Coorabie shear zone at the surface, respectively, and their higher conductivity is likely due to a reduction in grain size of olivine associated with an increase in influence of grain boundary diffusion thus resulting in enhanced conductivity. A comparison of the results with those from seismic studies across mobile belts worldwide shows that the more resistive response of the interpreted mobile belt for the Fowler Domain raises questions as to the nature of the domain (Fig. 7).

Summary

In the active regions such as subduction zone, rift, and collision zones, MT studies delineated the conductive regions in the crust and upper mantle indicating the presence of pathways of fluid migration, large partial melt zone associated with magmatic activity. The suture/shear zones in the cratonic regions were reflected as a conductive zone in the crust representing the grain boundary diffusion (in the case of Gawler Craton; a resistive body caused due to metamorphism associated with suturing under oxidizing condition in the case of Antarctica, between Rauer Group and Vestfold Hills; a linear vertical conductive feature extending to upper mantle in the case of Dharwar Craton. Further, MT imaged the fluid-filled zones which are the genesis for the seismicity in the San Andreas Fault region. As could be seen from the foregoing, the MT models helped providing useful clues as to the nature and configuration of different subsurface geological and structural features in a wide range of complex tectonic scenarios. From shallow levels to much deeper levels extending into the deep crust and mantle. The results were integrated with other geophysical inputs particularly seismic helped to gain insights into the tectonic processes that led to the understanding of the genesis of seismic activity.

Bibliography

- Arora BR, Unsworth MJ, Rawat G (2007) Deep resistivity structure of the northwest Indian Himalaya and its tectonic implications. *Geophys Res Lett* 34:L04307. <https://doi.org/10.1029/2006GL029165>
- Becken M, Ritter O, Bedrosian PA, Weckmann U (2011) Correlation between deep fluids, tremor and creep along the central San Andreas fault. *Nature* 480:87–90
- Curtis S, Thiel S (2019) Identifying lithospheric boundaries using magnetotellurics and Nd isotope geochemistry: an example from the Gawler Craton, Australia. *Precambrian Res* 320:403–423. <https://doi.org/10.1016/j.precamres.2018.11.013>
- Desissa M, Johnson N, Whaler K, Hautot S, Fisseha S, Dawes G (2013) A mantle magma reservoir beneath an incipient mid-ocean ridge in Afar, Ethiopia. *Nat Geosci* 6(10):861–865. <https://doi.org/10.1038/ngeo1925>
- Malleswari D, Veeraswamy K, Abdul Azeed KK, Gupta AK, Narendra Babu, Patro PK, Harinarayana T (2019) Magnetotelluric investigation of lithospheric electrical structure beneath the Dharwar Craton in South India: Evidence for mantle suture and plume-continental interaction. *Geoscience Frontiers* 10(5):1915–1930. <https://doi.org/10.1016/j.gsf.2018.10.011>
- Peacock JR, Selway K (2016) Magnetotelluric investigation of the vest fold hills and Raauer group, East Antarctica. *J Geophys Res Solid Earth* 121:2258–2273. <https://doi.org/10.1002/2015JB012677>
- Samrock F, Grayver AV, Eysteinnsson H, Saar M (2018) Magnetotelluric image of transcrustal magmatic system beneath the Tulu Moyo geothermal prospect in the Ethiopian Rift. *Geophys Res Lett* 45:12,847–12,855. <https://doi.org/10.1029/2018GL080333>
- Unsworth M, Jones A, Wei W et al (2005) Crustal rheology of the Himalaya and Southern Tibet inferred from magnetotelluric data. *Nature* 438:78–81. <https://doi.org/10.1038/nature04154>
- Worzewski T, Jegen M, Kopp H, Brasse H, Taylor Castillo W (2011) Magnetotelluric image of the fluid cycle in the Costa Rican subduction zone. *Nat Geosci* 4:108–111

Magnetovariation Studies

Nandini Nagarajan

National Geophysical Research Institute, Hyderabad, India

Definition

Magnetovariation studies identify the induced component of magnetic variations, from a spatial array of magnetic variation measurements and thereby infer the distribution of subsurface electrical conductivity in the Earth, through a linear transfer function between inducing and induced variations. These induction studies, which investigate the secondary magnetic fields induced in the solid Earth by external current systems, are undertaken to provide constraints on the electrical conductivity of the upper 1000 km of the finitely conducting earth.

The techniques involved in magnetic induction studies are as follows:

- (i) Measurement of magnetic variations over an array of stations:

Variations of the earth's magnetic field, caused by processes in the ionosphere and beyond, viz.; external variations, and the secondary fields induced by them in the Earth are recorded at observatories or at magnetometer stations explicitly set up for the purpose (► [“Geo-magnetic Field, Measurement Techniques”](#)). The sampling interval and lengths of data recorded are determined by the depth of investigation desired, limitations of instrumentation, size of the array, etc. Electric currents flow through the Earth's magnetosphere and ionosphere causing a wide spectrum of external magnetic variations (~ 1 s – 1 day) (► [“Magnetic Methods, Surface”](#)). These secondary electric currents are induced, in the conducting Earth below the source current and can be identified from spatial gradients in magnetic variations measured at adjacent or nearby stations. The normal field is defined as the sum of contributions from the external source field and that part of the internal field, which is due to the regional (1D, radial variation) electrical conductivity structure.

- (ii) Spectral analysis of variations, sufficiency of signal content, based on assumptions regarding uniform inducing fields.

Spectral content of magnetic variation are estimated for desired lengths of time windows, after which spectra are smoothed, stacked over several time windows, and significant peaks in frequency bands are estimated. To compute these from magnetic variations at a single site, certain assumptions are made about both the external sources and the underlying conductivity distribution. The external variations are generated over large-scale lengths relative to the area of the study and the incident magnetic fields are assumed to be a plane wave. For very large arrays, some modifications to the analysis are required since this assumption may not be valid. Since a large volume of data is collected, where the properties of external variations may be nonstationary, a careful selection of data intervals of suitable characteristics, conforming to the assumptions above, is made. Several techniques to eliminate effects of nonuniform, time-varying characteristics of inducing fields, have also been evolved. Further, this spectral analysis provides insights into the character of external variations and their measurement over large spatial extent on Earth. A study of spectral estimates over large arrays has resulted in classification of long, intermediate, and short wavelength variations. From this accumulated knowledge, it is possible to assess the requirements of size and density of measurements before commencing surveys. The estimation of external field characteristics has also contributed to modelling of external variations. Spectral estimates are also examined to eliminate sources of nonuniformity and man-made noise.

The geometry, intensity, and period of induced anomalous variations depend both on the source and on the distribution of electrically conducting materials within the Earth. Inducing external magnetic variations cover a wide spectrum from minutes to days, thus creating induced responses over a wide spectral range. The longer period responses derive from deeper penetration of induced currents within the crust/mantle (► [“Magnetotellurics, Crustal Imaging”](#)). The spectral content of magnetic variations is therefore a useful natural tool to probe the earth’s interior.

- (iii) Computation of transfer functions:

For magnetovariation studies, lateral gradients of the magnetic fields are used to obtain estimates of electrical conductivity contrasts in the subsurface. The induced, anomalous field is generated by telluric currents flowing in non-one-dimensional conductivity

structures (2D/3D) and, thus, entirely of internal origin. This anomalous field is not uniform and affects both the horizontal and the vertical components – evident as spatial gradients, and may be subjected to noise from man-made sources. Transfer functions, usually estimated by least squares, indicate the relation between inducing and induced variations at a single site, multiple sites, and for signals not recorded simultaneously (hypothetical event analysis). Improved methods of determining transfer functions using robust statistical approaches, using the actual properties of the data, rather than ideal distributions have refined estimation of transfer functions, improving their reliability. Simultaneous observations also provide information on the spectral space of the data, in each array study, and indicate whether new or modified interpretation parameters are to be evolved for a particular study. Fitting these with a formal inversion procedure, along with more traditional single site parameters such as MT impedances and vertical field TFs, can enhance resolution of some features (► [“Magnetotellurics, Crustal Imaging”](#)).

- (iv) Modelling to obtain induced magnetic variations from the conductivity distribution – interpret in terms of geology and structure:

Magnetovariation studies primarily delineate contrasts in electrical conductivity in the subsurface. The depth and conductivity are poorly resolved. Therefore, this method is widely used as preliminary investigation of large- and small-scale anomalies, which may later be intensively investigated using more precise electromagnetic methods (e.g., Magnetotellurics). The conductivity anomalies investigated using magnetovariation often extend over 1000s of kilometers across continents, correlate tectonic signatures to deeper crust/mantle properties, and link events of the vast geologic time-scale. Many examples of continental scale structures, as well as sharply defined shallow crustal anomalies, that are directly related to tectonic events (faulting, seismicity) are discussed in good reviews (► [“Continental Crustal Structure,”](#) ► [“Electrical Resistivity Surveys and Data Interpretation,”](#) ► [“Magnetotelluric Interpretation,”](#) ► [“GPS, Data Acquisition, and Analysis,”](#) and ► [“Electrical Properties of Rocks”](#)). The principles of electromagnetic induction have also been applied to satellite data of denser spatial coverage. Here, the forward response of conductivity distribution is directly modeled and fitted to potential data. In this manner, induced contributions to crustal magnetization of continents and ocean crust have been estimated.

Interpretation

Magnetometer array studies provide geometry of induced currents, in the Earth's subsurface. The interpretation is not confined to any predetermined shape or single conductive feature, rather delineates broadly any lateral inhomogeneities in the upper 1000 km of the Earth's crust/mantle. The conductivities are poorly constrained in this method and need to be verified using other methods (e.g., MT).

The significance of these electrical conductivity contrasts are threefold:

- (i) Indicate change in composition within depth of penetration, more graphite/conducting minerals (Artemieva and Mooney 2001)
- (ii) Changes in temperature, at appropriate depth (Gough 1989)
- (iii) Indicates accumulation of fluids, interstitial, or within the metamorphic mineral structures (Jones 1992).

The above conditions and geometry of the conductivity anomaly are evaluated or eliminated, for the region under study. The significance of any or either of these conditions can impact the tectonics and evolutionary history of the region. The complementary nature of conductivity anomalies and other geophysical studies were established in the course of several Geotransects over ~1000 km continental-scale studies. The role of fluids in the crust has been a major contribution of these studies. (► “Continental Crustal Structure,” ► “Heat Flow, Continental,” ► “Seismic Velocity and Temperature Relationships,” ► “Continental Crustal Structure,” ► “Lithosphere, Continental,” ► “Electrical Properties of Rocks,” ► “Deep Seismic Reflection and Refraction Profiling,” and ► “Lithosphere, Continental: Thermal Structure”).

Summary

At the Earth's surface magnetometers measure the composite of external (from the external source currents) and internal (from the induced currents) field components. The observed magnetic field can be considered as the sum of a normal plus an anomalous induced field. Induced currents in the finitely conducting earth manifest as gradients in magnetic variations over spatial scales of ~10–1000 km.

The interpretation of electrical conductivity anomalies in the crust provide additional insights into the nature of the Earth's crust. In the upper crust (<15 km), most conductivity anomalies were correlated with deeper extensions of tectonic elements (faults, sutures) and compositional changes. Simultaneously, the phenomenon of seismic reflectors within the lower crust were found, and combined with

this, the significant role of fluids in the lower crust/upper mantle came to be recognized. A number of Geotransects across continents were carried out in 1990s and the complexities of evolutionary history on the lower crust were revealed. The relationship between electrical conductivity, high heat flow, seismic parameters, and tectonic history were formulated. Induction studies and later magnetotelluric soundings have contributed to an understanding of processes in the Earth's crust and upper mantle. This was also possible because of the refinements in instrumentation, data processing, and computational approaches in magnetovariation studies. Specific case studies from Geotransects and regional studies are described in detail in several reviews written on this subject.

Cross-References

- Continental Crustal Structure
- Deep Seismic Reflection and Refraction Profiling
- Electrical Properties of Rocks
- Electrical Resistivity Surveys and Data Interpretation
- Electromagnetic Pulsations and Magnetic Storms
- Geoelectromagnetism
- Geomagnetic Field, Measurement Techniques
- GPS, Data Acquisition, and Analysis
- Heat Flow, Continental
- Lithosphere, Continental
- Lithosphere, Continental: Thermal Structure
- Magnetotelluric Interpretation
- Magnetovariation Studies
- Seismic Velocity and Temperature Relationships

Bibliography

- Artemieva IM, Mooney WD (2001) Thermal thickness and evolution of Precambrian lithosphere: a global study. *J Geophys Res* 106(B8):16,387–16,414
- Egbert GD (2002) Processing and interpretation of electromagnetic induction array data. *Surv Geophys* 23:207–249
- Gough DI (1989) Magnetometer array studies, earth structure, and tectonic processes. *Rev Geophys* 27(1):141–157. <https://doi.org/10.1029/RG027i001p00141>
- Gough DI, Ingham MR (1983) Interpretation methods for magnetometer arrays. *Rev Geophys* 23(4):805–827. <https://doi.org/10.1029/RG021i004p00805>
- Jones A (1992) Electrical conductivity of the lower continental crust. In: Fountain DM, Arculus R, Kay RW (eds) *Continental lower crust*. Elsevier, Amsterdam, pp 81–143
- Mooney WD, Meissner R (1992) Multi-genetic origin of crustal reflectivity: continental lower crust and Moho. In: Fountain DM, Arculus R, Kay RW (eds) *Continental lower crust*. Elsevier, Amsterdam, pp 45–79
- Schmucker U (1970) Anomalies of geomagnetic variations in the southwestern United States. *Bull. Scripps Institution of Oceanography*, 13. University of California

Mantle Convection

David Bercovici and Elvira Mulyukova
Department of Earth and Planetary Sciences, Yale University,
New Haven, CT, USA

Synonyms

Mantle circulation; Mantle dynamics

Definition

Mantle convection Thermal convection in the terrestrial planetary mantles, the rocky layer between crust and core, in which hot material rises, cold material sinks, and the induced flow governs tectonic and volcanic activity, as well as chemical segregation and cooling of the entire planet

Introduction and History

All planetary bodies retain some heat from their early formation but are inexorably cooling to space. Planetary surfaces are therefore cold relative to their hotter interiors, which thus undergo thermal convection wherein cold material is dense and sinks, while hot material is light and rises (liquid water near freezing being one of the rare exceptions to this process). Planetary atmospheres, oceans, rocky mantles, and metallic liquid cores convect and are subject to unique patterns of circulation in each domain. Silicate mantles, however, tend to be the most massive and sluggish part of terrestrial planets and therefore govern how planetary interiors evolve and cool to space (see Fig. 1).

The theory of mantle convection was originally developed to understand the thermal history of the Earth and to provide a driving mechanism for Alfred Wegener's theory of continental drift in the 1930s (see Schubert et al. 2001; Bercovici 2015). Interest in mantle convection waned for decades as Wegener's theory was criticized and apparently discredited. However, the accumulation of seafloor sounding data during World War II and refinement of paleomagnetic techniques paved the way for the discovery of seafloor spreading (Dietz 1961; Hess 1962; Vine and Matthews 1963; Vine 1966) and the birth of the grand unifying theory of plate tectonics in 1967–1968 (McKenzie and Parker 1967; Morgan 1968; Le Pichon 1968); this consequently revived interest in mantle convection as the driving mechanism for plate motions (Runcorn 1962a, 1962b) as well as non-plate tectonic volcanism such as the possible Hawaiian plume (Morgan 1971). The success of mantle convection theory in explaining plate

velocities, seafloor subsidence, volcanism, gravity anomalies, etc. leads to its further application to other terrestrial planets such as Venus and Mars, which also sustained unique forms of mantle convection, evident from volcanic activity.

Basics of Thermal or Free Convection

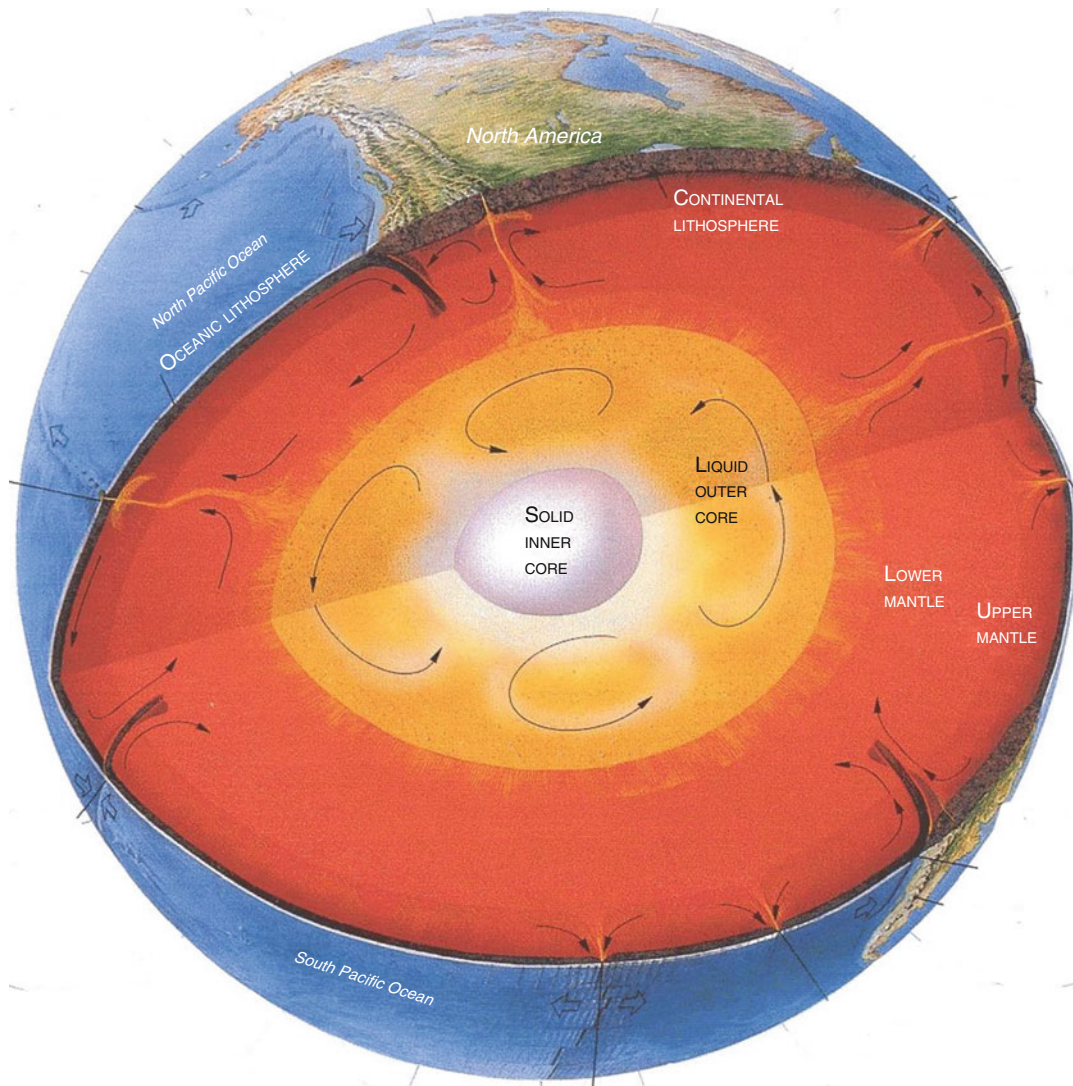
Rayleigh-Bénard Convection

The simplest form of thermal convection is referred to as Bénard convection named after the French experimentalist Henri Bénard, who in 1900 performed the first systematic experiments on convection in thin layers of oil (spermaceti) and recognized both the onset of convection from a static conductive state and the regular patterns formed in a convecting layer (Bénard 1900, 1901). Fifteen years later, the British theoretical physicist and mathematician Lord Rayleigh (John William Strutt) attempted to explain Bénard's results for the onset of convection (Rayleigh 1916) – the delay in communication between them being caused by World War I. However, the mismatch between theory and experiment was significant and not resolved until the late 1950s (Pearson 1958) when it was inferred that Bénard's experiments were influenced by surface tension or Marangoni effects not included in Rayleigh's theory (although Bénard himself was aware of these effects). Because Rayleigh's work provided the framework for nearly all thermal convection theory to follow, the simple Bénard convective system is also referred to as Rayleigh-Bénard convection.

Although Bénard's experiments were in a metal cavity, Rayleigh-Bénard convection actually refers to Rayleigh's idealized model of a thin fluid layer infinite in all horizontal directions such that the only intrinsic length scale in the system is the layer thickness. The Rayleigh-Bénard system is heated uniformly on the bottom by a heat reservoir held at a fixed temperature (i.e., the bottom boundary is everywhere isothermal), and the top is likewise held at a fixed colder temperature by another reservoir (see Fig. 2). If the layer was not fluid, heat would flow from the hot boundary to the cold one by thermal conduction. But since the fluid near the hotter base is (typically) less dense than the fluid near the colder surface, the layer is gravitationally unstable, i.e., less dense material underlies more dense material. To release gravitationally potential energy and go to a minimum energy state, the layer is induced to turnover.

Convective Onset and the Rayleigh Number

While the fluid in a Rayleigh-Bénard layer might be *gravitationally* unstable, it is not necessarily *convectively* unstable. Convective overturn of the layer is forced by heating but resisted or damped in two unique ways. Clearly, the thermal buoyancy (proportional to density contrast times gravity) of a hot fluid parcel rising from the bottom surface through colder



Mantle Convection, Fig. 1 Graphic cut away of the Earth's structure showing crust, mantle, and core. The relevant dimensions are that the Earth's average radius is 6371 km; the depth of the base of the oceanic crust is about 7 km and continental crust about 35 km; the base of the lithosphere varies from 0 at mid-ocean ridges to about 100 km near

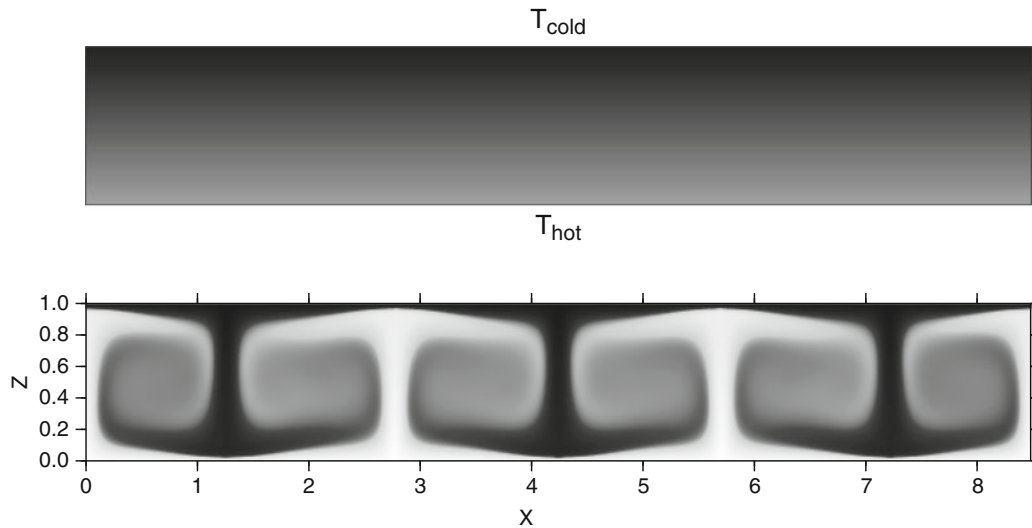
subduction zones; the base of the upper mantle is at 410 km depth, the transition zone sits between 410 and 660 km depths; the depth of the base of the mantle (the core-mantle boundary) is 2890 km; and the inner core-outer core boundary is at a depth of 5150 km. Image adapted from Lamb and Sington (1998)

surroundings acts to drive convective overturn. However, viscous drag acts to slow down this parcel, and thermal conduction, or diffusion, acts to erase its hot anomaly (i.e., it loses heat to its colder surroundings). Thus, while the fluid layer might be gravitationally unstable, hot parcels rising might move too slowly against viscous drag before being erased by thermal diffusion. Similar arguments can be made for cold material sinking from the top surface through warmer surroundings. The competition between forcing by thermal buoyancy and damping by viscosity and thermal diffusion is characterized in dimensionless ratio called the Rayleigh number

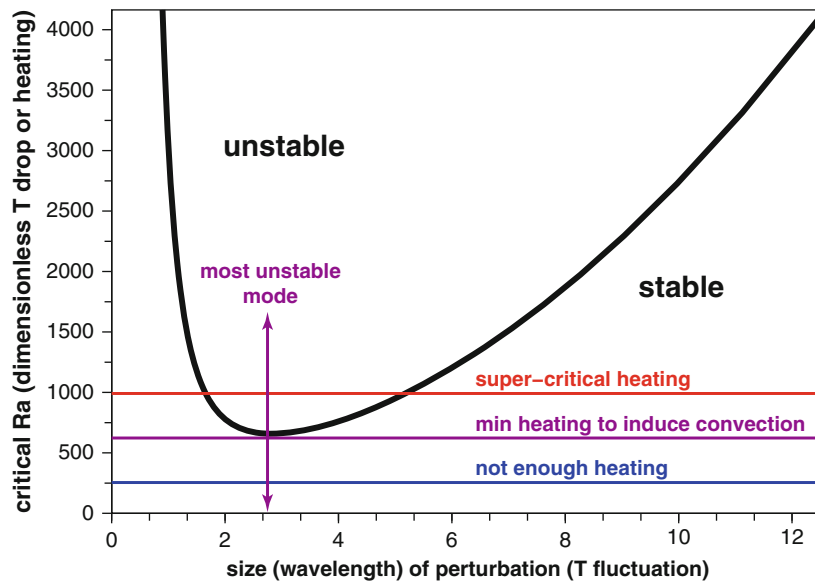
$$Ra = \frac{\rho g \alpha \Delta T d^3}{\mu \kappa} \quad (1)$$

where ρ is fluid density (units of kgm^{-3}), g is gravity (units of ms^{-2}), α is thermal expansivity (units of K^{-1}), ΔT is the difference in temperature between the bottom and top surfaces, d is the layer thickness, μ is fluid viscosity (units of Pa s), and κ is fluid thermal diffusivity (units of $\text{m}^2 \text{s}^{-1}$).

Even though $\Delta T > 0$ (i.e., heating is from below and causes gravitational instability), Ra still must exceed a certain value, called the critical Rayleigh number Ra_c for convection to occur. For $Ra < Ra_c$, the layer is stable and transports heat



Mantle Convection, Fig. 2 Rayleigh-Bénard convection: initially conducting layer (top) and numerical simulation of convection (bottom)



Mantle Convection, Fig. 3 The critical Rayleigh number Ra_c for the onset of convection is a function of wavelength or size of the thermal perturbation to a static conductive state. For a layer with isothermal and free-slip (shear-stress free) top and bottom boundaries, the relationship is $Ra_c = (k^2 + \pi^2)^3 / k^2$ where $k = 2\pi/\lambda$ and λ are wavelength (Chandrasekhar 1961). Values of Ra above the Ra_c curve are associated with the

conductive layer being convectively unstable (perturbations grow), while below the curve the layer is stable (perturbations decay). The minimum in the Ra_c curve occurs at the wavelength of the first perturbation to go unstable as heating and Ra are increased; this wavelength is often called the least stable (or most unstable) mode

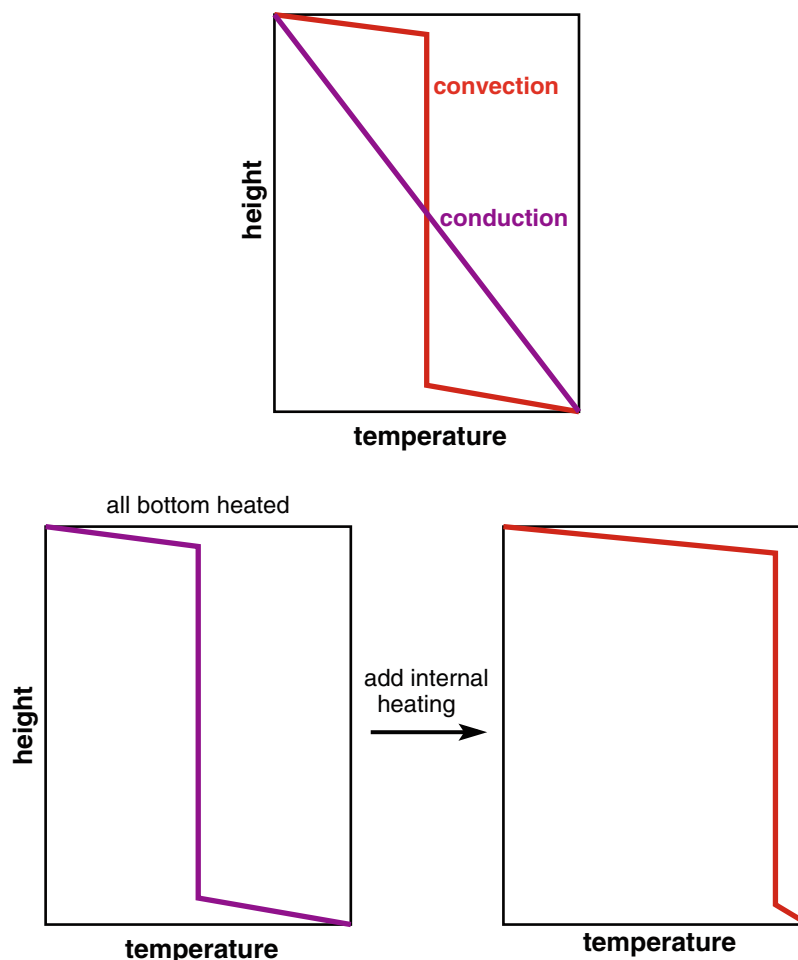
by conduction; for $Ra > Ra_c$, the layer will be convectively unstable and transport heat more rapidly via convection (see Fig. 3).

Although Ra_c varies depending on the mechanical nature of the horizontal boundaries (whether rigid or a free surface), it is typically of order 1000. This value is easily understood by considering the fate of a hot (or cold) parcel of size a and temperature anomaly ΔT . Dimensional analysis readily shows that the typical ascent rate of the parcel is $\rho g a \Delta T^2 / \mu$ (with

units of m/s). However, the rate that heat diffuses out of the parcel is κ/a (smaller parcels lose heat faster). The critical state occurs when these two rates are equal; i.e., if the buoyant ascent rate just exceeds the diffusion rate, the parcel should rise without being erased, but if the ascent rate is less than the diffusion rate, it will not rise very far before being lost. Therefore the critical state occurs if $\rho g a \Delta T^3 / (\mu \kappa) \approx 1$. Scaling purely by orders of magnitude, a small parcel of fluid can be assumed to be of order 10 times smaller than the entire

Mantle Convection,

Fig. 4 Sketch of temperature profiles, showing how convective mixing homogenizes the conductive mean temperature into a nearly isothermal state (if the fluid is incompressible) with thermal boundary layers connecting it to the cold surface and hot base (top frame). With no internal heating, the interior mean temperature is the average of the top and bottom temperatures; the effect of adding internal heating (bottom frames) is to increase the interior mean temperature and thus change the relative size and temperature drop across the top and bottom thermal boundary layers



layer, thus assuming $a \approx d/10$ leads to a critical condition for onset of convection of $\rho g a \Delta T d^3 / (\mu \kappa) \approx 1000$.

For the Earth's mantle, the typical average properties from which the Rayleigh number is constructed are $\rho \approx 4000 \text{ kg/m}^3$, $g = 10 \text{ m/s}^2$, $\alpha = 3 \times 10^{-5} \text{ K}^{-1}$, $\Delta T \approx 3000 \text{ K}$, $d = 2900 \text{ km}$, $\mu = 10^{22} \text{ Pa s}$ (dominated by the lower mantle), and $\kappa = 10^{-6} \text{ m}^2/\text{s}$ (see Schubert et al. 2001). Taken together these lead to a Rayleigh number of approximately 10^7 , which is well beyond supercritical; although the mantle viscosity is extremely high, the mantle is also very hot and very large and hence convecting vigorously.

Thermal Boundary Layers and the Nusselt Number

For a Rayleigh number Ra above the critical value Ra_c , convective circulation will mix the fluid layer, and the mixing and homogenization of the fluid will become more effective the more vigorous the convection, i.e., as Ra is further increased. With very large Ra and vigorous mixing, most of the layer is largely uniform and isothermal. (In fact, if the layer is deep enough such that pressures are comparable to fluid incompressibility, the fluid layer is not isothermal but adiabatic, wherein even without any heating or cooling, the

temperature would drop with fluid expansion on ascent and increase with fluid compression on descent.) Most of the fluid in the Rayleigh-Bénard system is at the mean temperature between the two boundary temperatures. However, the temperature still must drop from the well-mixed warm interior to the cold temperature at the top and to the hotter temperature at the bottom. The thin horizontal regions accommodating these jumps in temperature are called thermal boundary layers (Fig. 4).

Thermal boundary layers are of great importance in thermal (and mantle) convection for two reasons. First, the buoyancy of the system is primarily bound up in the thermal boundary layers since these are where most of the cold material at the top and hot material at the bottom resides and from where hot and cold thermals or currents emanate. Moreover, with regard to convection in the mantle itself, the top cold thermal boundary layer is typically associated with the Earth's lithosphere, the 100 km or so thick layer of cold and stiff mantle rock that is effectively divided into tectonic plates.

Second, since fluid in these boundary layers is near a horizontal boundary, most of the fluid motion is horizontal, and thus heat is only transported vertically across these thin

layers by conduction, but since the layers are very thin, such conductive transport is rapid. Indeed, the entire cycle of heat transport occurs by heat conducted through the bottom boundary layer, after which hot fluid in this layer becomes unstable and rises to form a convective upwelling that carries heat out of the boundary layer across the fluid layer and deposits it at the upper boundary, where the heat is then transported out of the system by conduction across the top boundary layer. The eventual heat flow (power output per unit area) out of the well-mixed layer is essentially $k\Delta T/\delta$ where k is thermal conductivity (units of $W K^{-1} m^{-1}$), $\Delta T/2$ is the temperature drop from the well-mixed interior to the surface, and we define $\delta/2$ to be the thickness of the boundary layer. By comparison, the thermal conduction across a static non-convecting layer is $k\Delta T/d$. The ratio of heat flow in the convectively well-mixed layer to the purely conductive layer is thus d/δ , which is called the Nusselt number Nu (named after the German engineer, Wilhelm Nusselt 1882–1957). The relation between Nu and convective vigor parameterized by Ra is important for understanding how convection transports heat and cools off bodies, including planets. Convective heat transport is often written as $Nu(k\Delta T/d)$, and in considering this relation, Howard (1966) argued that vigorous convective heat transport across the bulk of the well-mixed fluid layer is so fast it is not the rate limiting factor in releasing heat (only conduction across the thermal boundary layer is), and thus heat flow should be independent of fluid depth d ; this implies that since $Ra \sim d^3$ then $Nu \sim Ra^{1/3}$, which yields a convective heat flow $Nu(k\Delta T/d)$ that is independent of d . In general, since the fluid is conductive for $Ra \leq Ra_c$, one often writes that $Nu = (Ra/Ra_c)^{1/3}$ (although $Nu = 1$ for $Ra < Ra_c$), which is a reasonably accurate relationship borne out by simple experiments and computer modeling. This relationship also implies that the ratio of thermal boundary width to fluid layer depth is $\delta/d \sim Ra^{-1/3}$, which shows that the boundary layers become increasingly thin as convective mixing of the layer becomes more vigorous.

The relation of $\delta \sim Ra^{-1/3}$ applies to the horizontally averaged boundary layer thickness. However, boundary layers change with time or distance from their first formation, e.g., where an upwelling impinges on the top boundary. As the fluid in the boundary layer moves from an upwelling to a downwelling, it cools and the boundary layer thickens as more material cools next to the cold surface. The thickening depends on the thermal diffusivity κ (with units of m^2/s) and the residence time or age t near the cold boundary (i.e., time since leaving the upwelling). Simple dimensional considerations show that the boundary layer thickness goes as $\sqrt{\kappa t}$; this corresponds to the well-known $\sqrt{\text{age}}$ law for subsidence of ocean seafloor with age since formation at mid-ocean ridges, implying that seafloor gets deeper because of the cooling and thickening lithosphere.

Patterns of Convection: Upwellings, Downwellings, Plumes, and Slabs

When convection occurs, hot upwellings and cold downwellings will be separated horizontally by some optimal distance. If they are too close to one another, they induce too much viscous drag on and/or lose heat rapidly to each other; if they are too far apart, they must roll too much mass between them. The separation distance is also determined by heat transport in the thermal boundary layer between upwellings and downwelling. When hot upwelling fluid reaches the surface, it spreads laterally into the thermal boundary layer. As it travels horizontally, it cools to the surface and eventually gets cold and sinks into the downwelling; thus the upwelling-downwelling separation is also determined by the distance it takes for fluid to cool off and become heavy enough to sink.

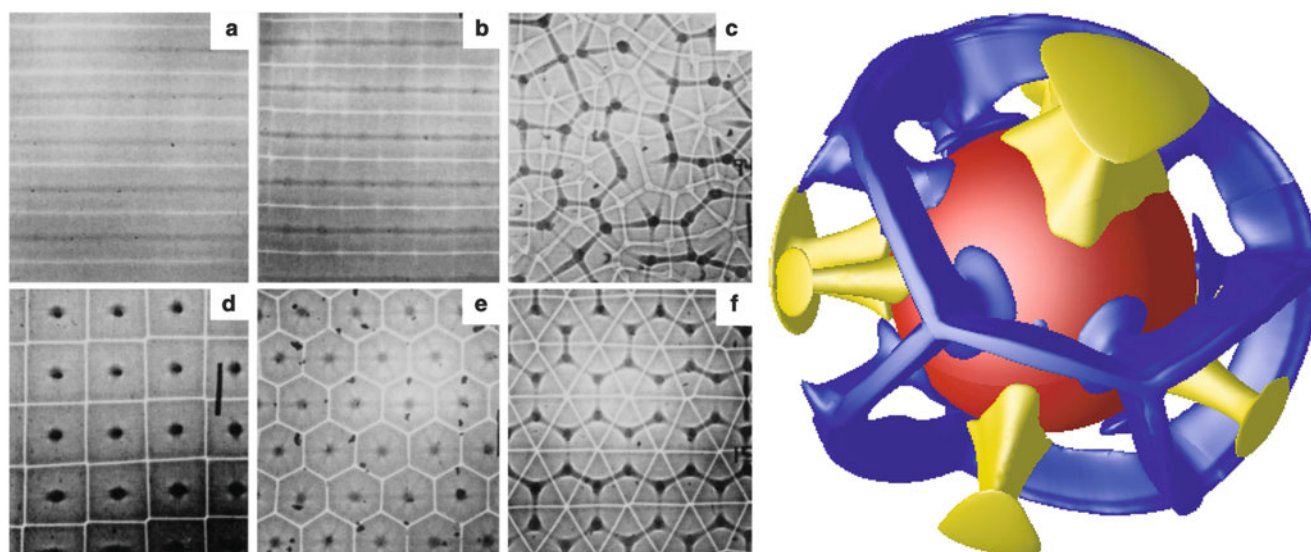
The upwelling-downwelling separation distance or convection cell size is predicted by convective stability theory to approximately equal the layer depth d (a bit larger at the onset of convection but identically d as Ra becomes very large); i.e., the cell that is either least stable or thus most likely to convect (see Fig. 3) – or equivalently the cell that optimizes gravitational potential energy (and thus heat) release – is usually the cell that is as wide as it is deep.

Viewing a convecting layer from above, the upwelling and downwellings may be separated in various patterns, such as two-dimensional rolls (sheets of upwelling rolling over into sheets of downwelling and each cell counter-rotating with its neighboring cell). In the Rayleigh-Bénard layer, which is infinite horizontally, no one location of the layer should be different than any other one, so ideally the pattern should be a regular repeating tile; as there are only so many polygons that can form a repeating tile, the patterns usually involve convection cells in the shapes of rolls (already mentioned), squares, hexagons, and triangles (Fig. 5). Of course, non-ideality and irregularities can occur due to tiny imperfections, for example, in the boundaries, leading to irregular patterns.

In many instances the 3-D pattern of convection, especially in fluids where hot fluid is less viscous than cold fluid (as is true in many materials, including the mantle), the upwelling is in the form of a cylindrical plume at the center of a canopy of sheetlike downwellings, much like a fountain (e.g., see Fig. 5, left panel, frames d and e).

Plumes and Slabs in the Mantle: Simple View

The common occurrence of sheetlike downwellings and columnar upwellings in simple convection is crudely applicable to mantle convection. Subducting slabs are where tectonic plates sink into the mantle; these are analogous to the cold sheetlike downwellings seen in 3-D convection, although much more complicated by rheology, as discussed below. Hot narrow upwelling plumes, rising from the base of the mantle to the surface, are inferred to explain intraplate



Mantle Convection, Fig. 5 Patterns of convection from laboratory experiments in a rectangular tank by White (1988) (left) and numerical simulations in a spherical shell by Zhong et al. (2000) (right). The left panel shows a top view of several experiments, using a shadowgraph

technique in which dark regions are upwellings and bright regions are downwellings (see Davaille and Limare 2015). The right panel shows three isothermal surfaces to represent cold downwellings (blue), warm upwellings (yellow), and the hot surface of the core (red)

volcanism, as in Hawaii, as well as the fixity of these volcanic hotspots relative to each other (which suggests deep anchoring in the mantle). These mantle plumes are ostensibly analogous to the pipelike upwelling in simple 3-D convection but again more complicated by unique mantle properties. (See “Mantle Plumes” essay by Farnetani and Hofmann.)

The manner of upwelling at mid-ocean ridges or spreading centers deviates from the classical picture of simple Rayleigh-Bénard convection. In particular, ridges are narrow linear features that, for the most part, involve shallow passive upwelling (inferred from their weak gravity anomalies that suggest shallow isostatic compensation, i.e., they are floating on a shallow buoyant root). Ridges are likely best explained as being pulled and rifted apart passively from a distant force (ostensibly slabs) rather than involving a deep convective upwelling that pries them open.

Energy Sources for Mantle Convection and the Earth's Thermochemical History

Heat flow coming out of the Earth is measured by heat-flow gauges (measuring conductivity of rocks first and then thermal gradients in boreholes) both in continents and oceans (see Jaupart and Mareschal 2011; Turcotte and Schubert 2014). The total heat flowing out from beneath the Earth's surface is approximately 46TW (46 trillion Watts) (Jaupart et al. 2015), which is not a large number given the surface area of the Earth and is tens of thousands of times smaller

than the heat absorbed from the Sun. Nevertheless, the terrestrial heat flow represents the source of power driving dynamics inside our planet, including mantle convection (and hence tectonic, volcanic, and other geological activity) as well as core cooling.

The source of the Earth's internal heat is a combination of primordial heat left over from planetary accretion and formation (i.e., energy release from planetesimal collisions as well as gravitational energy release from core segregation) and heat generated by unstable radioactive isotopes, especially the isotopes of uranium (^{238}U), thorium (^{232}Th), and potassium (^{40}K), although the ^{40}K half-life is much shorter than the others and so generated a large heat pulse primarily in the early Earth. Because continental crust is originally formed by partial melting and chemical segregation of early mantle material (see the discussion below in a later section), these radioactive elements tend to be concentrated in the crust (i.e., melt more readily dissolves these elements than does solid rock, so they partition toward the melt). Thus, the crust itself produces a significant fraction of the net heat output through the surface; removing the crustal component leaves approximately 38TW emanating from the mantle and core (Schubert et al. 2001; Jaupart et al. 2015).

The relative contributions of primordial cooling and radiogenic heating to the mantle (and core) heat output remain an active area of debate even today and lead to various quandaries. The most direct estimate of the concentration of radiogenic sources (U, Th, K) is by looking at the concentration in chondritic meteorites, which come from the solar systems' main asteroid belt. Because these meteorites

have been largely unaltered for 4.5 Gyrs (i.e., unmelted), they are thought to be the same as the original building blocks of the terrestrial planets. The chondritic concentrations of U, Th, and K would allow for radiogenic heating to contribute 50% or less to the total heat output (Korenaga 2008); this is often called the Urey ratio, i.e., the ratio of radiogenic heat output to the total output. With radiogenic heating this small, the only way the mantle could be as hot as it is today – while also transporting heat as it does presently – is if it was very hot and nearly molten in the geologically recent past (a few hundred million years); this is geologically untenable, since petrological and geochemical analysis demonstrate the presence of solid rock and even liquid water in the early Earth (see Halliday 2001), which therefore ceased to be molten not long after its formation 4.5 billion years ago. This paradox has led some researchers to assume that the radiogenic sources are super-chondritic, i.e., to allow heat to be produced continuously throughout the past rather than by rapid cooling from a molten state (Schubert et al. 1980). Alternatively, researchers have sought ways to keep chondritic concentrations of U, Th, and K by arguing that heat transport in the past was different than it is today; for example, higher temperatures in the past might have allowed for more melting and thus more buoyant crust and/or stronger dehydrated lithosphere that kept the top part of the convecting mantle sluggish or immobile, hence bottling up primordial heat for much later release (e.g., Korenaga 2008; Sleep 2015).

Finally, the release of both radiogenic and primordial heat, again termed collectively “internal heating” (i.e., heat coming from the bulk of the fluid), means that the idealized Rayleigh-Bénard model of convection, where heating is only along the base, is inaccurate. The effect of internal heating in addition to “basal heating” is relatively straightforward to understand. While in the Rayleigh-Bénard model the temperature of the well-mixed interior is at the average of the two boundary temperatures, the addition of internal heating acts to heat up the well-mixed interior to a higher mean temperature. This puts the interior temperature closer to the hot bottom temperature but further from the cold top temperature; the effect is to create a larger temperature drop across the top thermal boundary layer than across the bottom one; in essence, the top boundary layer must conduct out heat injected through the bottom plus heat generated or released from the interior (Fig. 4). These very different boundary layers tend to cause more negatively buoyant and stronger downwelling currents and smaller and weaker upwelling currents; this effect seems to be borne out in the Earth by the presence of many large cold slabs driving plate tectonic motion with large thermal anomalies (of order 500 K) relative to fewer upwelling plumes with weaker (200 K) thermal anomalies (c.f. Ballmer et al. 2015; Hernlund and McNamara 2015).

Effects of Mantle Material Properties

Mantle Rheology

The entire mantle of the Earth is potentially a convecting and overturning fluid. However, the mantle is almost entirely solid (with some small portions of melting near the surface delineated by volcanism and possibly much smaller areas of melting at depth) and thus flows extremely slowly. (Indeed, the term fluid does not suggest a liquid state but refers to how a medium deforms, with other modes of deformation including, e.g., elastic, plastic, or brittle behavior; these do not necessarily correlate with states of matter, i.e., gas, liquid, and solid, each of which can display either fluid or elastic behavior depending on the nature of the deformation; e.g., atmospheric sound waves are elastic behavior in a gas.)

The solid-state viscous flow of the mantle is a complex process, and there are various mechanisms that permit such “irrecoverable” deformation (one of the definitions of viscous flow). A survey of mantle deformation mechanisms or “rheology” is beyond the scope of this essay (see Ranalli 1995; Karato 2008). However, in brief, the two primary deformation mechanisms are called diffusion and dislocation creep. In any solid-state creep mechanism, mobility depends on the statistical-mechanical probability that an atom in a crystal lattice will leave the potential well of its lattice site; the potential well itself is defined by electrostatic or chemical bonds inhibiting escape and Pauli exclusion pressure preventing atoms squeezing too closely to each other. Thus, mobility depends on the Boltzmann distribution measuring the probability of having sufficient energy to overcome the lattice potential well barrier, which is often called the activation energy (or, allowing for pressure variations, the activation enthalpy). This probability depends on the Arrhenius factor $e^{-E_a/RT}$ where E_a is the activation energy (J/mol), R is the gas constant (J/K/mol), and T is temperature; RT represents the thermal excitation energy of the molecule in the well. As T goes to infinity, the probability of escaping the well goes to 1 (meaning greater mobility and hence softer material), while as T goes to 0, the probability of escape goes to 0 (meaning less mobility and stiffer material).

Stress imposed on the medium effectively changes the shape of the potential well, such that compressive stress steepens the walls of the well (squeezing atoms closer makes coulomb attraction in the chemical bonds stronger), while tension lowers the walls (separating molecules weakens the bonds); thus, the probability of escape is preferred in the direction of tension and away from compression, allowing the medium to stretch in the tensile direction by solid-state diffusion of molecules.

Simple diffusive creep works much in this way where differential stress (i.e., nonuniform stress) causes slow diffusion of molecules to allow the entire substance to deform accordingly. However, such deformation occurs at the

mineral grain level inside a rock, with either diffusion through the grains or along the grain boundaries; thus the response depends significantly on grain size.

Dislocation creep is more complicated. Dislocations are a type of crystal defect associated with misalignment in the otherwise ordered rows of atoms in a crystal lattice; these can be in the form of truncated rows of atoms or other imperfections in the order by which atomic planes are stacked in the lattice. The presence of a dislocation gives rise to internal stress within a crystal, which adds to the ambient stress acting on the rock. To relieve these stresses, dislocations move through the crystal (one atomic step at a time), interact with each other by deforming the lattice in their vicinity, and trigger the formation of new dislocations. Their motion is also limited by the mobility of atoms in the lattice, and thus they obey a similar thermally controlled Boltzmann statistics and Arrhenius factor described above for diffusion creep. The generation and motion of dislocations are an efficient way to accommodate strain, known as dislocation creep. However, the combined influences of internal and ambient stresses on dislocation dynamics are compounded into a nonlinear rheological response, wherein stress and strain rate are not linearly proportional to each other.

The viscosities for diffusion and dislocation creep mechanisms can be written as

$$\mu = \begin{cases} Ba^m e^{\frac{E_a}{RT}} & \text{for diffusion creep} \\ Aa^{1-n} e^{\frac{E_d}{RT}} & \text{for dislocation creep} \end{cases} \quad (2)$$

where A and B are proportionality constants, a is grain size, σ is stress (in fact, since stress is a tensor, σ^2 is the scalar second invariant of the stress tensor), and m and n are exponents, characteristic values of which are both 3 ± 1 . It should be emphasized that diffusion and dislocation creep occur independently of each other depending on stress and grain size: for high stress and large grains, dislocation creep dominates; for low stress and small grains, diffusion creep dominates. Dislocation creep allows for moderate softening as stress increases; diffusion creep potentially allows for significant softening if stress can reduce grain size (see the sections below on generating tectonic plates).

One of the strongest rheological effects is that of temperature; the temperature dependence of viscosity allows for many orders of magnitude variations in strength. For example, while this rheological effect allows cold subducting slabs to keep their strength and integrity to great depths as they sink, it makes hot upwelling mantle plumes more fluid, and, if they have a conduit structure, the plume material ascent rate is relatively rapid, of order 100 cm/year or more. However, this effect is most profound in the cold top thermal boundary layer or lithosphere. If viscosity is strongly temperature dependent, as it is in the Earth, the lithosphere can become so stiff that it

becomes immobile; in this case convection in the mantle would proceed beneath the lithosphere, which in turn would act like a rigid lid to the mantle (Solomatov 1995). If temperature were the only property controlling the mantle's rheological properties, then the lithosphere would be locked and immobile, and there would be no plate tectonics. While this scenario might be relevant for Venus and Mars (and Moon and Mercury), which arguably have no plate tectonics, obviously it is missing a vital ingredient to allow plate tectonics on Earth. This paradox underlies the fundamental question and mystery about why the Earth has plate tectonics at all and how it is generated on our planet but not others in our solar system (Bercovici 2003; Bercovici et al. 2015; Mulyukova and Bercovici 2019).

Compressibility, Melting, and Solid Phase Changes

Pressures deep inside the Earth's mantle are so large; they are sizable fractions of rock incompressibility or bulk modulus (e.g., mantle pressures reach 140GPa, or 1.4 million atmospheres, while bulk modulus – which has the same units as pressure – are typically a few to several 100GPa). As downwelling mantle material travels from near the surface to the base of the mantle, its density and temperature increase due to compression, called “adiabatic compression and heating” (and likewise upwelling material undergoes “adiabatic decompression and cooling”), although these increases are not large (of order several degrees Celsius). The compression and decompression of circulating material establish a weak adiabatic temperature and density increase with depth, which has a slight stabilizing effect on convection; however, because the mantle is so viscous, the thermal anomalies needed to get it to move – and in particular the temperature variations across the thermal boundary layers – are so large (of order several 100 to 1000 degrees Celsius) that the adiabatic variations are small in comparison (Bercovici et al. 1992).

Where compressibility and pressure play a dual important role is in phase changes. First, as hot upwelling mantle material approaches the surface, it actually travels along a gradually cooling adiabatic temperature profile. The upwelling does eventually melt when it gets near the surface but not because it gets hotter. Melting occurs because the melting temperature T_m drops with decreasing pressure faster than the upwelling adiabat (in essence, decreasing confining pressure makes it easier for molecules to mobilize into a melt); thus at a certain usually shallow depth of a few 10s of km to 100 km, the upwelling mantle crosses the melting temperature from solid to liquid phase and undergoes melting; however, the mantle is not a single pure substance so only partially melts. The resulting liquid melt phase is expelled to the surface where it refreezes to make basaltic crust. Such “pressure release” melting is a shallow process but is vital for chemical segregation of the mantle and development of oceanic crust. (The production of continental crust is a more

complicated process resulting from the injection of hydrated minerals into the mantle at subduction zones; this triggers flux or hydrous melting of the mantle, which then – upon melt ascent through pre-existing crust – leads to a cascade of silica-rich melt products in volcanic arc environments.)

Extreme pressures with depth can also overcome a mineral's elastic resistance to compression and cause solid-solid phase changes in which the minerals change their crystallographic structure to a more compact, denser, and incompressible state (but their chemistry remains the same). Such mineralogical phase changes have been observed in laboratory experiments in olivine, which is the major mineral component of the upper mantle (at about 60% by weight, the remainder being mostly pyroxene at shallow depths, and garnet at slightly greater depth). The pressures at which these phase changes are predicted to occur have been verified seismologically; i.e., seismic wave speeds undergo a jump at the pressures where the phase transitions and density jumps are predicted to exist. The first major phase change to occur with depth is from olivine to the same material with a wadsleyite structure, at 410 km depth. Wadsleyite changes slightly to a similar ringwoodite structure at 510 km depth. The largest phase change occurs from ringwoodite to perovskite/magnesiowüstite at 660 km depth.

The 410 km and 660 km phase changes are the two most remarkable and global phase changes in the mantle, and the region between them is called the transition zone, since it is where most of the mineralogical transitions occur, over a relatively narrow region. The mantle above the transition zone is typically identified as the upper mantle, although in some papers and books upper mantle includes the transition zone. Below the transition zone is the lower mantle, and that is universally agreed upon in the literature. The transition zone has anomalous properties due to mixing and transitions in mineral organization; for example, it is thought to be able to absorb an order of magnitude more water (per kg) than the mantle both above and below it (Wood and Corgne 2015).

Other phase changes are thought to occur with depth, although these are less well resolved, and in some instances do not appear to be global. One of the most recently discovered phase changes has been inferred to exist in the lowest part of the mantle (the bottom few 100 kilometers), called the perovskite-post-perovskite (or just the post-perovskite) transition (Murakami et al. 2004; Hernlund and McNamara 2015). This transition is still an active area of exploration.

The effect of phase transitions on mantle convection has been an active area of research since the 1960s. The discovery of the major phase change at 660 km depth coincided with the fact that deep earthquakes along subducting slabs (the Wadati-Benioff zone) go no deeper than 700 km. This seemed to imply that subducting slabs, the mantle's equivalent of cold convective downwellings (see below), did not extend into the lower mantle. Recent seismological studies using

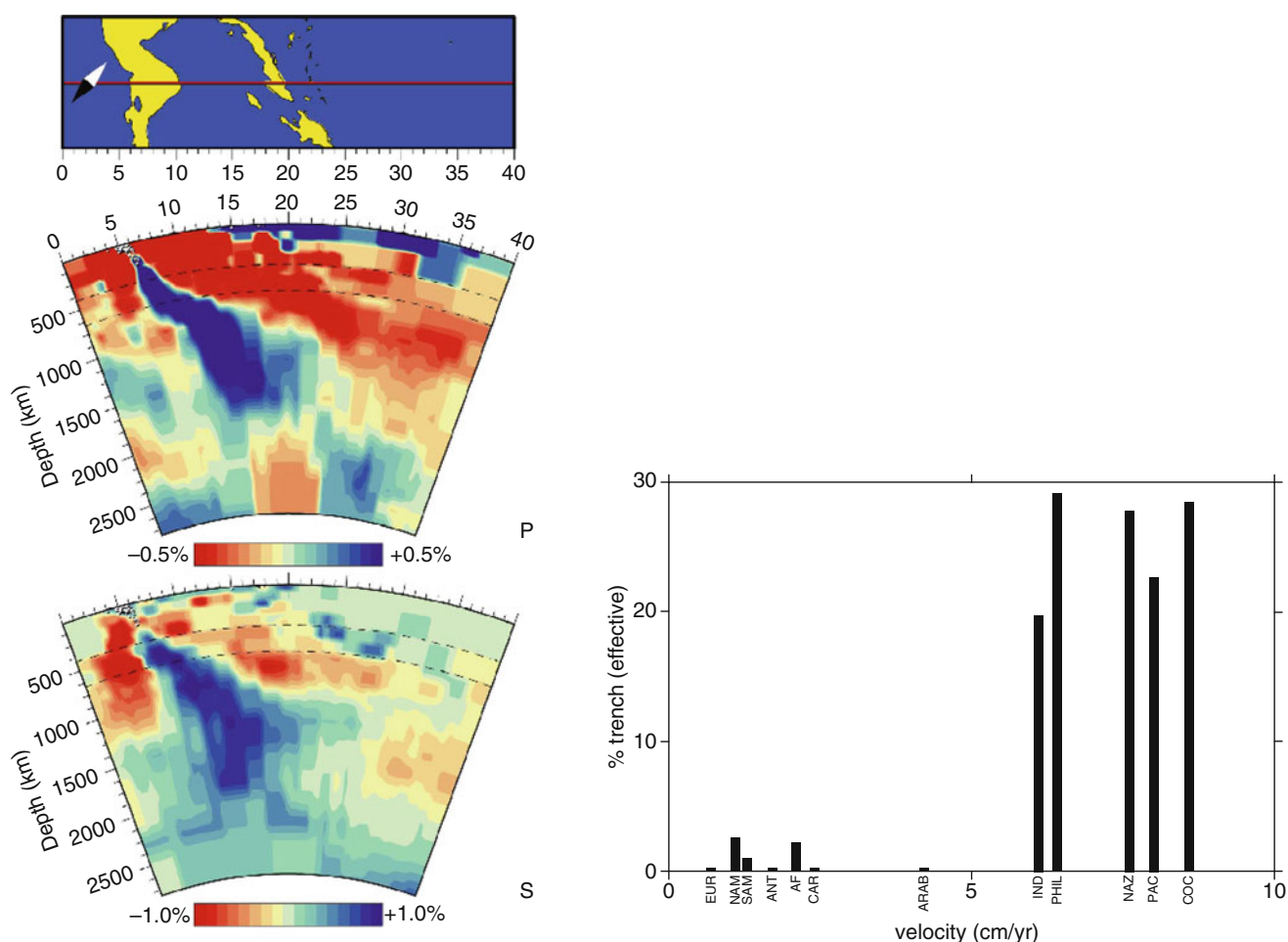
tomographic techniques to resolve ostensibly “cold” and “hot” areas of the mantle (really seismically fast and slow regions) implied that many slabs might stall or pool temporarily at the 660 km boundary, but many penetrate into the lower mantle (Fig. 6).

Whether the density jump due to a phase change could impede vertical flow has been a key question in studying the interaction of phase changes and convection. In particular, the 660 km phase change was inferred mineralogically to be “endothermic” whereby the entropy of the deeper heavier phase (below 660 km) increases, or more simply latent heat is absorbed on going down through the boundary into the denser more compact phase (while unusual in most systems, this is also true of the solid-liquid transition in water). This also means that the dependence of the transition temperature on pressure has a negative slope; thus cold material impinging on the phase boundary causes the phase change to deflect in the cold region to greater pressures; this induces a depression in the boundary that acts to buoyantly rebound upward and oppose the motion of the descending cold material. Thus the endothermic boundary at 660 km depth possibly impedes flow across that boundary (Schubert et al. 1975) (However, the actual 660 km phase change is complex and at higher temperatures might become exothermic (Weidner and Wang 1998).) Computational studies and simulations of mantle flow across this boundary demonstrated that downwellings (i.e., slabs) can indeed be impeded and stalled as they impinge on this boundary, but not permanently or globally; i.e., while some downwellings pool at the boundary, others accumulate enough “weight” to push through the boundary (Tackley et al. 1993; Christensen 1995). This picture appears to be in keeping with seismological findings, wherein some slabs appear to stagnate, pool, or deflect from their vertical descent at the 660 km boundary, while others appear to penetrate the boundary into the lower mantle (e.g., van der Hilst et al. 1997; Fukao and Obayashi 2013). Therefore, while that phase boundary impedes slabs and downwelling flow into the mantle, it is not an impermeable boundary, and there is, in the end, significant exchange between the upper and lower mantle and hence whole mantle convection.

However, while the mineralogical, seismological, and geodynamical (fluid mechanical) arguments imply that there is flow between upper and lower mantles, there are numerous data from geochemical analysis of basaltic lavas implying that the mantle is not well stirred on a large scale, i.e., it is possibly layered with poor or nonexistent communication between upper and lower mantle.

Mantle Geochemistry and Mantle Mixing

Upwelling mantle reaching the Earth's surface undergoes melting (see above under pressure-release melting), and this



Mantle Convection, Fig. 6 Left panel: tomographic images of slabs beneath Mexico (W. Spakman via van Keken et al. (2002)) extending into the lower mantle, in which the top frame shows the location of the profile, the next two frames show tomographic inversions for both P and

S wave travel time anomalies. Right frame: a global compilation of plate velocity vs trench length following Forsyth and Uyeda (1975), showing that the fastest plates are connected to slabs. Slabs are not only cold mantle downwellings but are also the primary driver for plate tectonics

melt reaches the surface in two types of volcanic settings: mid-ocean ridges, where tectonic plates spread apart and draw mantle up into the opening gap, and ocean-islands or hotspots, which are anomalously productive and localized volcanic features not necessarily associated with tectonic activity, Hawaii being the most conspicuous such feature. Melts coming from the mantle in this way are silica poor (relative to more silicic rocks such as granite) and largely basaltic; hence these volcanic regions are said to produce mid-ocean ridge basalts (MORB) and ocean-island basalts (OIB), respectively. These melts are in effect messengers from the mantle, and their petrological composition, bulk chemistry, and trace element chemistry are extensive areas of research (Hofmann 1997, 2003; van Keken et al. 2002; Tackley 2015). These two types of basalts have distinct features suggesting they come from parts of the mantle that have been isolated from each other for billions of years. That seismology, mineral physics, and fluid dynamics

(geodynamics) argue for a largely well-stirred mantle with whole-layer circulation creates a dichotomy between geophysical and geochemical observations. This paradox has been one of the most fervent areas of debate in mantle dynamics for the last 30–40 years.

Because MORB and OIB are both basalts and thus have similar bulk chemistry, geochemical measurements largely focus on trace elements, in particular incompatible elements, which dissolve more readily in a rock's melt phase than its solid phase; hence during partial melting incompatible elements partition toward the melt. Indeed, the trace element signature of elements, such as uranium, thorium, and helium, has demonstrated that MORB and OIB are very distinct. In particular, MORBs appear to be significantly depleted in such trace elements relative to OIB. Since such elements tend to be removed by melting, it implies that MORBs come from a region of the mantle that has already been melted and depleted of trace elements, while OIBs come from a region of the

mantle that has undergone little previous melting and depletion. This observation implies that MORBs come from an upper mantle that has been cycled repeatedly through the plate tectonic process of mid-ocean ridge melting and separation of crust (and trace elements) from mantle, in essence cleaning the MORB source. In contrast, OIB would appear to come from a part of the mantle that has seen little of this melt processing and hence would be isolated presumably at depth from the upper mantle and plate tectonic circulation.

There are other geochemical observations that argue for separated and isolated regions or reservoirs in the mantle. For example, the concentration of radioactive daughter isotopes (e.g., ^{206}Pb , which is the final product of the decay of ^{238}U) relative to the abundance of that element's dominant and primordial isotope (e.g., ^{204}Pb) is a metric for reservoir isolation from surface processing.

The relative accumulation of daughter products implies that the rock in which they reside has seen little processing or partial melting that would have cleaned out these elements after they formed. A small relative abundance of daughter isotopes means the sample has been recently processed and cleaned, and thus little time has passed in which to produce new daughter isotopes.

Indeed, many OIBs tend to show distinctly greater relative abundance of daughter products (e.g., the concentration ratio $^{206}\text{Pb}/^{204}\text{Pb}$) than do MORB for many isotopic ratios, implying some isolation of the OIB source. However, OIBs from various islands are also fairly different from each other suggesting that reservoirs isolated from the upper mantle (the presumed MORB source) are possibly also isolated from each other. Moreover, the OIB isotopic ratios have some variation, from low MORB-like values to much higher values, which indicates that there is some mixing of a "young" processed MORB source-like mantle and a more primitive, isolated one.

These isotopic ratios are also more easily interpreted for refractory daughter products since they do not tend to escape the system (also the reason they are used to radiometrically date rocks). Volatile products, especially isotopes of noble gases such as helium and argon, require different interpretations since they can readily escape the mantle and, for helium, escape the Earth. For example, MORB has a high daughter to primordial isotope ratio $^4\text{He}/^3\text{He}$ relative to many OIBs, which is opposite to the refractory ratios involving, for example, lead isotopes. This is often interpreted as a result from degassing and loss of primordial helium ^3He from the upper mantle through plate tectonic and mid-ocean ridge processing and the subsequent repopulating of helium with its radiogenic isotope ^4He (i.e., α -particles from most large element decay sequences); in contrast, an isolated lower mantle or OIB source reservoir would have undergone little loss of primordial helium thus maintaining a smaller isotopic ratio $^4\text{He}/^3\text{He}$.

The production of the argon isotope ^{40}Ar from the decay of the potassium isotope ^{40}K has two important arguments relative to mantle layering. First, the total amount of original ^{40}K in the Earth can be roughly estimated from chondritic abundances (and other arguments beyond the scope of this essay). However, the amount of ^{40}Ar it should have produced over the age of the Earth is far in excess (by a factor of 2) of the ^{40}Ar in the atmosphere, implying that much of this argon is still buried and isolated. Moreover, one can also estimate from the trace element composition of MORBs themselves that the MORB source region as it stands now would have been lacking primordial ^{40}K and even if the entire mantle were composed of this MORB source region, it would not have been able to produce even the atmospheric levels of ^{40}Ar ; this implies that the bulk of original ^{40}K was buried in a layer different and more enriched than the MORB source region, which then produced most of the ^{40}Ar , much of which is still buried in this layer.

An often quoted straw man argument for mantle layering is the heat-flow paradox. In this case, it is reasoned that if the entire mantle were composed of MORB source material with its depleted concentration of heat-producing elements (U, Th, K), then it would not be able to produce the total heat output through the top of the mantle (about 38TW). This suggests that the heat-producing elements allowing for the mantle heat output must be buried at depth. However, this makes the assumption that most heat output is from radiogenic heating, whereas possibly less than half of it is; if most heat output is from secular cooling (loss of primordial heat), then the heat-flow paradox argument is questionable. A similar but shakier argument is based on the fact that the volcanic flux of the helium isotope ^4He , which is produced from heat-producing radioactive decay of U and Th, seems to be very low relative to what would be expected given the heat that is emanating from the mantle, implying that ^4He is also buried at depth. Again, if heat output is more than 50% from secular cooling and not all from radiogenic heating, then the low flux of helium is to be expected. Even if that were not the case, mechanisms for the flux of helium are not the same as for flux of heat; i.e., heat always escapes to space eventually by convection, conduction, and even radiation, whereas helium only escapes from the mantle if it passes through narrow melting zones, which is not inevitable, and thus it can be hidden and buried almost anywhere and not only hidden at great depth.

Finally, the production of continental crust also argues for an isolated layer in the mantle. Continental crust represents an accumulated history of mantle melting, segregation of lighter components, and removal of incompatible elements to the surface. If the continental crust were removed uniformly from a whole mantle made of primordial "bulk silicate earth" (i.e., an Earth derived from chondritic material and only segregated into mantle and core), then the concentration

of incompatible elements would not have been reduced enough to produce a mantle made of MORB source material (i.e., removal of continental crust would not have depleted the whole mantle enough to make MORB source). However, if the crust were removed from 1/3 to 1/2 of the mantle, that resulting depleted portion would very closely match MORB source composition. This argues that the continental crust segregated only from the upper portion of the mantle, not the whole mantle, and thus there remains a deeper unsegregated mantle at depth (see van Keken et al. 2002; Tackley 2015).

Although there are numerous geochemical arguments for a layered mantle with an isolated and undepleted mantle at depth, they largely conflict with geophysical evidence for whole mantle convection. Mineral physics experiments suggest that the 660 km phase change boundary might provide an impediment to mantle flow but not an impermeable barrier. Seismic tomography consistently shows subducting slabs and apparently cold downwellings extending into the lower mantle (van der Hilst et al. 1997; Grand et al. 1997). Recent high-resolution images of a hot upwelling mantle plume beneath Hawaii (Wolfe et al. 2009) as well as seismic images of other plumes (Montelli et al. 2004; French and Romanowicz 2015) also suggest vertical upward transport across the 660 km boundary. Finally, geodynamical arguments against separated layers suggest that if a lower mantle held most of the mantle's heat-producing elements, it would be implausibly hot, and by heating up the bottom of the upper mantle, it would generate much bigger mantle plumes than would be observed (Tackley 2002).

The contradiction between geochemical and geophysical inference of layered vs whole mantle convection has been persistent problem in the study of the Earth's mantle. Attempts to reconcile these observations have been numerous. A reasonably popular approach has been to allow that the 660 km boundary is not a barrier to mantle flow, but that the barrier exists at greater depth. There are seismically observable layers at the bottom of the mantle (the D'' layer), which could store enriched material, although these are also so thin, they could possibly overheat (depending on the amount of radioactive heat sources stored there). As a compromise, it has been argued that the enriched mantle exists in an approximately 1000-km-thick layer at the base of the mantle (Kellogg et al. 1999), although a layer this thick has never been seismologically observed (see Fig. 7).

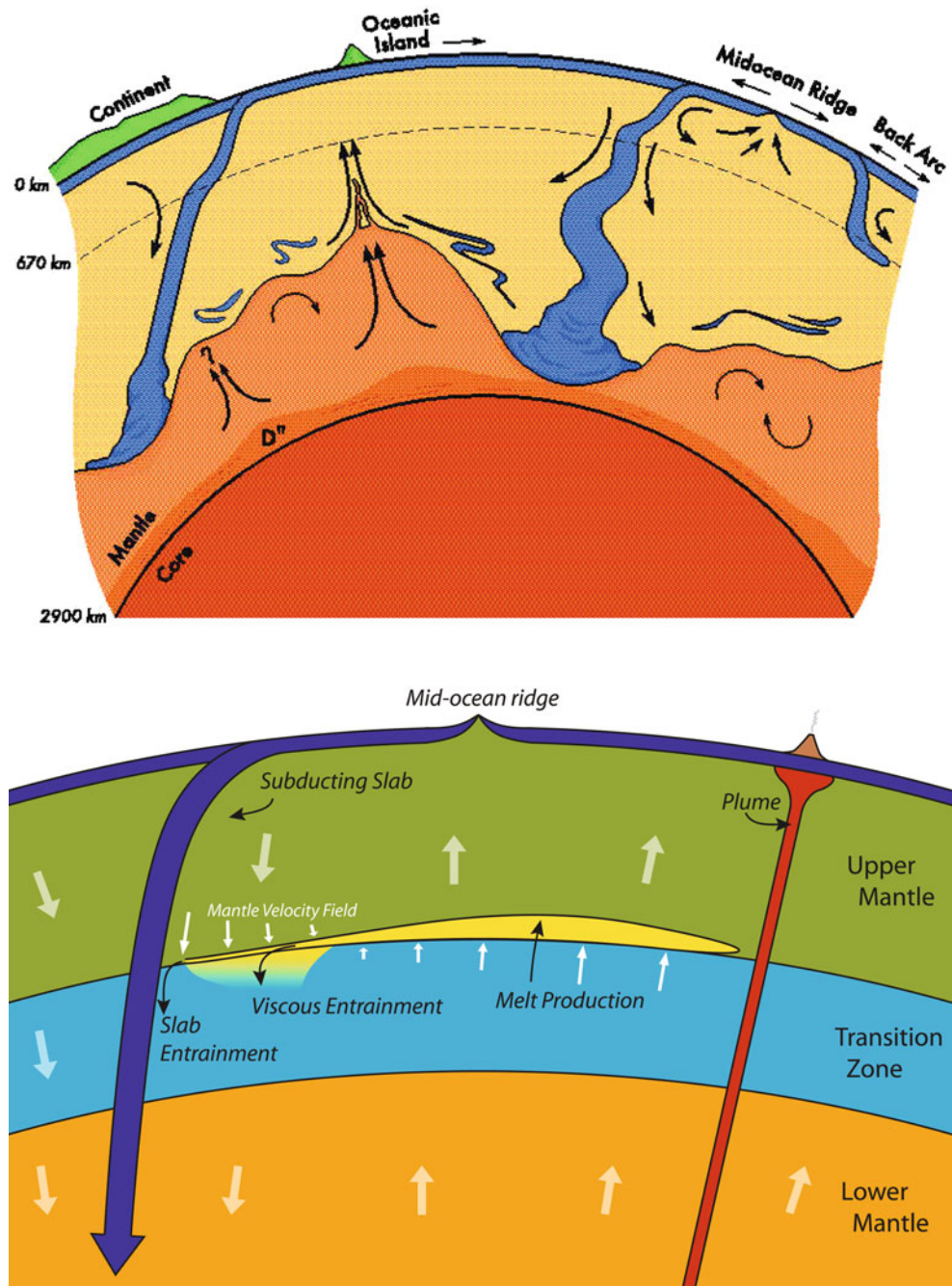
In the last decade, there has been mounting evidence for two large regions of chemical heterogeneity gathered into two piles on the core-mantle boundary beneath Africa and the South Pacific. These regions are detectable for their anomalously low seismic shear velocities and are termed large low-shear-velocity provinces, or LLSVPs (Garnero and McNamara 2008; Garnero et al. 2016; Hernlund and McNamara 2015). These piles arise naturally in convection models from the action of cold downwellings or subducting

slabs impinging on the core-mantle boundary and sweeping chemically dense material toward convergence points, much as how continents accumulate on the surface (McNamara and Zhong 2005; Zhong and Liu 2016). However, the source of chemically anomalous material that comprises LLSVPs remains unknown, with the main candidates being a primordial layer that formed early in the Earth's history (Lee et al. 2010) and/or accumulated pools of dense mineral phases (e.g., oceanic crustal basalt converted into eclogite) that were delivered to the bottom of the mantle through billions of years of the Earth's subduction history (Christensen and Hofmann 1994; Mulyukova et al. 2015). The isolation and longevity of these piles (e.g., Mulyukova et al. 2015) make them a viable source for isolated and semiprimitive source materials for OIB. Indeed, upwelling mantle plumes are hypothesized to initiate and emerge from the edges of these dense piles, thus supplying enriched OIB material (e.g., Hernlund and McNamara 2015; Dannberg and Gassmöller 2018). The fixity of LLSVPs for hundreds of millions of years has also been proposed as evidence that they are primordial features of the deep Earth (e.g., Conrad et al. 2013), although this interpretation remains debated (e.g., Rudolph and Zhong 2013; Zhong and Liu 2016). Nevertheless, LLSVPs are a viable source for chemical layering in the mantle that would give rise to distinct OIBs; to a large extent, they are similar to the abyssal layer predicted by Kellogg et al. (1999), albeit in a smaller and non-contiguous layer.

Other mechanisms for reconciling geochemical and geophysical observations, but not invoking layering, have been proposed as well. These mostly involve differential melting. For example, one model considers the whole mantle as a plum-pudding mix of volatile (e.g., water)-enriched plums in a depleted, drier, and harder to melt pudding. The pressure release melting in mantle plumes is stopped at higher pressures at the base of 100-km-thick lithosphere, so this could involve mostly melting of volatile-rich plums leading to an enriched OIB signature; upwelling mantle making it all the way to the surface at ridges would undergo more pressure drop and thus could also melt the drier mantle pudding, resulting in MORB that seems depleted relative to OIB (Ito and Mahoney 2005a, 2005b). Another model exploits the fact that transition zone minerals seem to be able to absorb water more readily than material above it (and below it). A transition zone with a little water will be dry relative to its solubility or water storage capacity. However upwelling material passing through the transition zone would carry this slightly damp material into the upper mantle at the 410 km boundary, and since the upper mantle olivine has poor water solubility, it would be closer to saturation and likely melt. A little melting at 410 km depth of the broad upwelling mantle (forced upward by the downward flux of slabs) passing through and out of the transition zone would cause it to be

Mantle Convection,

Fig. 7 Two end-member mantle mixing models. In the abyssal layered model (left), the source for enriched ocean island basalt (OIB) is in a deep primordial layer, while the source for depleted mid-ocean ridge basalt (MORB) is in the upper recycled mantle (after Kellogg et al. (1999)). In differential melting models, such as the water-filter model (right), the MORB and OIB sources undergo different styles of melting, but the mantle still undergoes whole-mantle circulation; see text for further explanation (after Bercovici and Karato (2003); Leahy and Bercovici (2007))



stripped of incompatible elements as it flows into the upper mantle, leaving a depleted MORB source region; because of the high pressures and high compressibility of melt, the partial melt that has cleaned this upwelling mantle would be dense and remain behind, eventually to be entrained by slab-driven downwelling back into the lower mantle. Upwelling mantle plumes, on the other hand, would go through the transition zone too fast to become hydrated and thus would undergo little melting and filtering at 410 km depths, leaving largely

enriched OIB. This model, called the transition-zone water filter (Bercovici and Karato 2003; Karato et al. 2006; Leahy and Bercovici 2010) (see Fig. 7), predicts that the 410 km should be the site of melting, and this has been borne out in various seismological studies (e.g., Revenaugh and Sipkin 1994; Song et al. 2004; Tauzin et al. 2010); however, the theory is still controversial given the state of knowledge of melt properties and their solubilities of incompatible elements at these depths and pressures (Wood and Corgne 2015).

Mantle Convection and the Generation of Plate Tectonics

The link between plate tectonics and mantle convection is one of the oldest and most challenging problems in the history of geodynamics. The original theories of mantle convection put forward by Holmes (1931) were developed in the context of explaining continental drift as articulated by Wegener (1924). Although the later theory of plate tectonics is the grand-unifying principle of geology, it is a kinematic theory in that it describes surface motions, but not their cause. Mantle convection is widely accepted to be the engine for plate motions since it is a fundamental mechanism for exploiting the energy sources of the Earth's interior, i.e., loss of primordial and radiogenic heat. It is now generally regarded that the plates themselves are a feature of mantle convection in that they are the mobile upper thermal boundary layer of convective cells that thicken as they cool in their migration away from ridges until they are heavy enough to sink along subduction zones.

One of the major accomplishments of mantle dynamics theory is that convective fluid velocities, calculated in any number of ways, consistently predict the measured scales of plate tectonic velocities, i.e., between 1 and 10 cm/yr. Such velocities were inferred even in the 1930s from both gravity and heat-flow measurements (Pekeris 1935; Hales 1936) and are well-known by the force balance on sinking slabs (e.g., Davies and Richards 1992), as well as global heat extraction from mid-mantle cooling by slabs (Bercovici 2003). Moreover, plate motions are well correlated with the presence of slab forcing, in particular that tectonic plates with a significant portion of subduction all have velocities roughly an order of magnitude faster than plates without substantial subduction zone (Forsyth and Uyeda 1975); see Fig. 6. Thus, because cold sinking slabs seem to be the major expression of convection, major driver of plate motions implies that the plates are convection (see Coltice et al. 2019).

The Plate Generation Problem

While convective forces (especially slab pull) are important for driving plate tectonics, understanding how plates self-consistently arise from mantle convection is a major goal in geodynamics. Up until the early 1990s, it was believed that since plate-like motion of the lithosphere was essentially discontinuous, it could not be predicted or reproduced by fluid dynamical convection theories. However, in the last 25 years or so, there has been major progress with fluid dynamical models yielding plate-like motion by incorporating more sophisticated rheological weakening mechanisms, such as brittle/plastic yielding or damage mechanics.

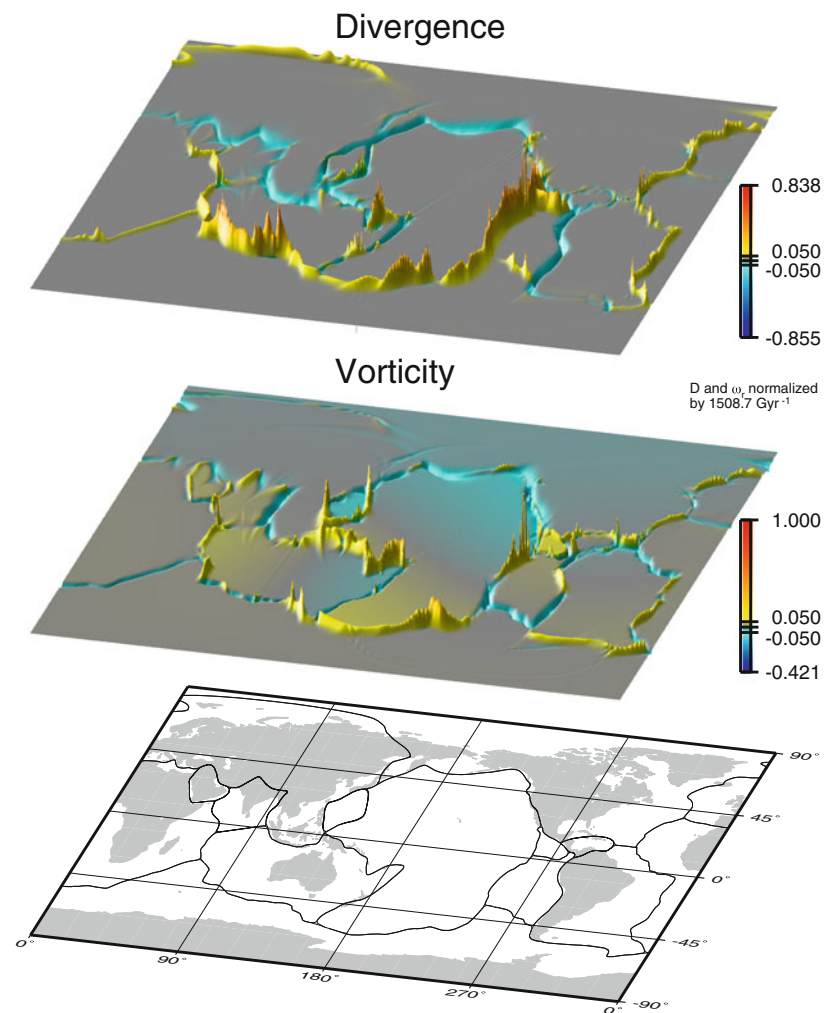
Both a clue and frustration in understanding the origin of plate tectonics are that the Earth appears to be the only known terrestrial planet that has plate tectonics (see Gregg 2015), in

addition to liquid water as well as life, which are all either causative (i.e., necessary conditions for each other) or coincidental. While our planet supports plate tectonics, our ostensible twin, Venus, does not; this remains a leading-order quandary in Earth sciences, and to solve it one must understand how and why plate tectonics is generated at all. Of course geoscience has, until the mid-1990s, only sampled the few planets of our own solar system, and thus the data was sparse; with the advent of extrasolar planet discovery (see Laughlin and Lissauer 2015), this sparseness should be mitigated, and perhaps other planets will be discovered with plate-like mantle circulation from which we will learn more about how our own planet works (e.g., Valencia et al. 2007; Korenaga 2010; van Heck and Tackley 2011; Foley et al. 2012; Tackley et al. 2013).

There are of course various qualitative and philosophical questions about how and why plate tectonics forms, evolves, and exists, but for these to be predicted or reproduced in a physical theory, one requires quantifiable questions. In short, what are the metrics of plate generation? Two fundamental features of plate-like motion exist for instantaneous motions: these are *plateness* (Weinstein and Olson 1992) and toroidal motion (e.g., Hager and O'Connell 1979; O'Connell et al. 1991; Bercovici et al. 2000, 2015). Plateness is essentially the extent to which surface motion and the strength of the lithosphere are like that for nearly rigid blocks separated by narrow or localized weak boundaries. Toroidal flow is characterized by strike-slip motion and plate spin (Fig. 8). Toroidal motion has no direct driving force in buoyant convection (which drives only vertical and divergent motion); however, it has as much energy in the present-day plate tectonic velocity field as the buoyantly driven motion (called poloidal motion). The globally averaged toroidal motion is dependent on the lithosphere's reference frame (e.g., hotspot frame), and the field in general changes through time (Cadek and Ricard 1992; Lithgow-Bertelloni et al. 1993); however, the toroidal field is a quantifiable and significant feature of global plate motions. Such measurable quantities as plateness and toroidal flow are important for testing the predictions of plate generation theories. Both phenomena rely on reasonably strong nonlinear rheological feedback effects to permit large strength variations for high plateness (i.e., rapidly deforming zones are weak, slowly deforming ones are strong), as well as coupling of buoyantly driven flow to vertical torques that drive toroidal spin and shear (Bercovici 2003; Bercovici et al. 2015).

In the last few decades, plate generation models have become increasingly sophisticated, in concert with further expansion and accessibility of high-performance computing. Instantaneous plate-like behavior has been achieved with convection models employing various forms of plastic yield and self-weakening criteria. Incorporation of these laws has led to the prediction of reasonable toroidal and poloidal flow (Bercovici 1993, 1995) and by including the rheological

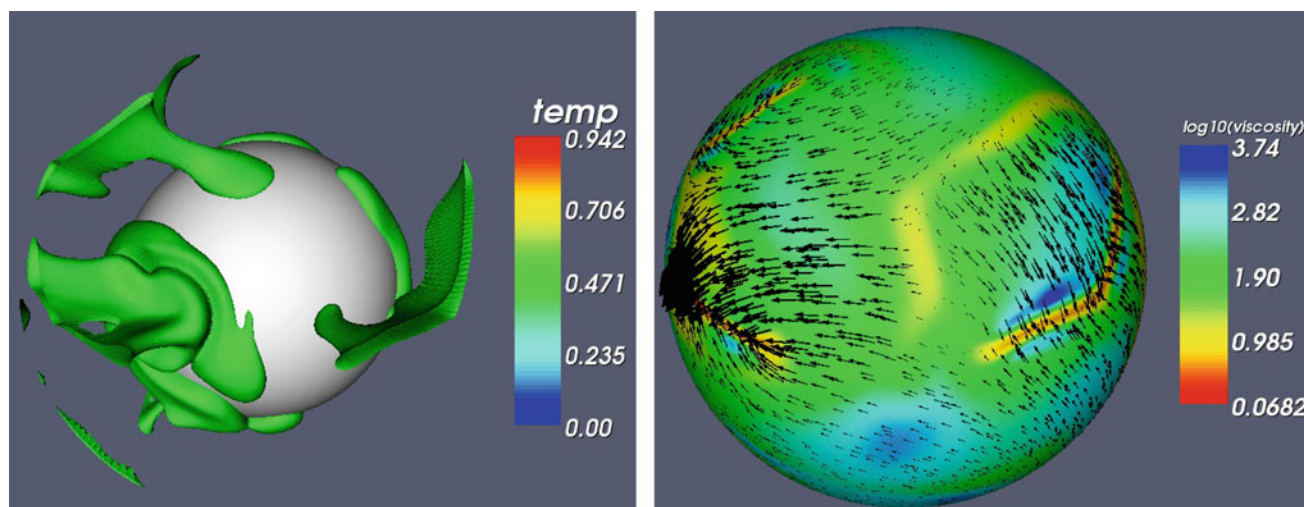
Mantle Convection, Fig. 8 The horizontal divergence (which measures spreading and convergence rates) and vertical vorticity (which measures angular velocity and horizontal shear) of present-day plate motions (from Dumoulin et al. 1998), which are associated with poloidal and toroidal flow, respectively. Poloidal flow is equivalent to basic convective motion driven directly by thermal buoyancy. Although present in the Earth's plate-mantle system, toroidal flow does not arise naturally in basic viscous convection but requires the coupling of convective motion with nonlinear rheological effects



effects of melting at ridges has attained localized passive spreading zones (Tackley 2000). Models using plastic yielding have been extended to three-dimensional spherical models, creating the first global models of plate generation from mantle convection (van Heck and Tackley 2008; Foley and Becker 2009) (Fig. 9), including the effects of continents (e.g., Coltice et al. 2019).

However, models that use plastic or instantaneous self-weakening rheologies only allow weak zones to exist while being deformed, thus cannot correctly model dormant weak zones (e.g., sutures and inactive fracture zones). Rheological mechanisms that allow weakening to persist over time have also been studied. While thermal weakening is a well-understood mechanism, thermal anomalies diffuse too fast, and it is highly unlikely that, for example, sutures are thermal remnants. Weakening by hydration, or with water as a secondary phase (i.e., by providing pore pressure), is a strong candidate as well, although the mechanism for ingesting water to depth is problematic (Korenaga 2017) and requires mechanisms such as cracking enhanced by thermal stresses (Korenaga 2007).

Damage in the form of grain-size reduction (e.g., Karato et al. 1980; Braun et al. 1999; Bercovici and Ricard 2005; Austin and Evans 2007; Rozel et al. 2011) is also a strong candidate for a weakening mechanism, especially given evidence in the form of mylonites, which are small grained and highly deformed metamorphic rocks pervasive in many plate boundary environments (e.g., White et al. 1980; Jin et al. 1998; Warren and Hirth 2006; Skemer et al. 2010; Hansen et al. 2013). The coexistence of grain-size reduction with grain-size-sensitive (e.g., diffusion creep) rheology allows a positive shear-localizing feedback mechanism to generate weak plate boundaries, and its occurrence at depths where the lithosphere is strongest makes grain-size weakening particularly appealing. However, grain-size reduction and weakening in olivine (the upper mantle's primary mineral phase) tend to occur in exclusive areas of deformation space (i.e., weakening occurs during diffusion creep, while grain-size reduction occurs in dislocation creep; see De Bresser et al. (2001)), and thus a self-softening feedback by which plate boundaries form is problematic. Moreover, grain growth in olivine is possibly relatively fast (Karato 1989), which would

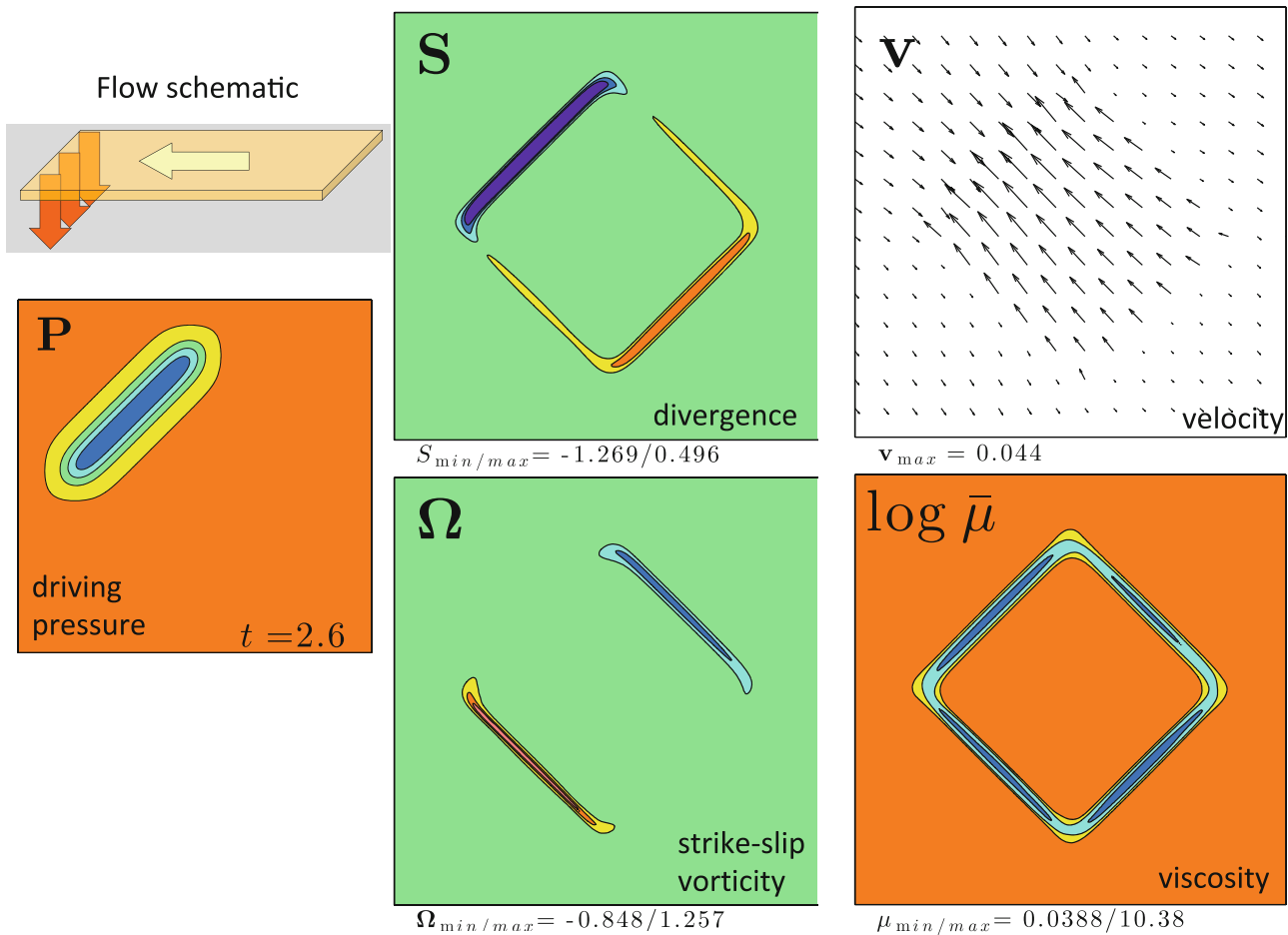


Mantle Convection, Fig. 9 Three-dimensional spherical shell convection with a plastic-type lithospheric rheology showing plate-like behavior. Left panel shows isothermal surfaces and in particular cold downwellings. The right panel shows surface viscosity with velocity

vectors superimposed. Note the passive divergent and rheological weak zone forming midway between the two major downwelling regions (Adapted from Foley and Becker (2009). American Geophysical Union)

promote rapid healing of weak zones. However, these problems are likely mitigated by the fact that mantle lithospheric rocks are polyminerallic, with two dominant mineral phases of olivine and pyroxene; indeed, plate boundary mylonites exhibit grain-size reduction associated with the intermixing of mineral phases (e.g., Olgaard 1990; Warren and Hirth 2006; Skemer et al. 2010; Herwegh et al. 2011; Linckens et al. 2015; Tasaka et al. 2017; Cross and Skemer 2017; Wiesman et al. 2018; Bercovici and Ricard 2016; Bercovici and Skemer 2017; Mulyukova and Bercovici 2017; Bercovici and Mulyukova 2018). Two-phase or polyminerallic grain-damage theory (Bercovici and Ricard 2012) has provided a physical framework for the formation of plate boundary weak zones based on laboratory and field studies of mylonites. A key component of the theory is that mutual Zener pinning between the mineral phases both facilitates damage and grain-size reduction while also impeding grain growth and healing. Applications of the theory have demonstrated the generation of toroidal motion and strike-slip deformation from poloidal flow (Bercovici and Ricard 2013) and facilitating the formation of subduction zones through the collapse of passive margins (Mulyukova and Bercovici 2018). Grain-damage theory has also been used to explain how Hadean proto-subduction and associated damage and weakening allowed for the accumulation and inheritance of weak zones, which permitted the emergence of global plate tectonics in the Archean Earth by approximately 3Ga (Bercovici and Ricard 2014) (Fig. 10) as hypothesized by Condie and Kröner (2008) (see also Korenaga 2013); for a review of grain-damage, applications to plate generation, and recent developments, see Mulyukova and Bercovici (2019).

Use of both plastic/brittle yielding and damage theories of plate generation has been used to elucidate the planetary dichotomy between the Earth and Venus and the causal link between climate, liquid water, and plate generation. The Earth and Venus are ostensible twins, but the Earth has plate tectonics and Venus does not. This is often attributed to lack of liquid water on Venus, which would otherwise lubricate plate motions; however, the Earth's lithosphere is likely to be no more hydrated than Venus's because of melting dehydration at ridges (Hirth and Kohlstedt 1996). Recent studies have hypothesized that the role of water in maintaining plate motion is not to lubricate plates but to be an agent for the carbon cycle, which thus allows for a temperate climate on the Earth. A cool surface on the Earth, according to one hypothesis (Lenardic et al. 2008), causes a larger temperature drop across the lithosphere than would occur on Venus (because of temperature-dependent viscosity), and thus lithospheric buoyant stresses are large enough on the Earth to induce plastic/brittle failure, but not on Venus. An alternative hypothesis (Landuyt and Bercovici 2009; Foley et al. 2012; Bercovici and Ricard 2014) states that plate-like motion depends on the competition between damage and healing (where, e.g., if lithospheric damage and weakening are due to grain-size reduction, then healing and stiffening are due to normal grain growth), where a high damage-to-healing ratio promotes plate-like motion, while a lower ratio yields stagnant-lid behavior. A cooler surface temperature inhibits healing, while a hot surface promotes healing, thus leading to plate-like behavior on a planet like the Earth with a temperate climate but not on a planet like Venus. Both hypotheses emphasize that water dictates conditions for



Mantle Convection, Fig. 10 A calculation of pressure-driven flow in a thin 2-D horizontal layer model of the lithosphere, where the pressure low is akin to a single subduction zone (see schematic frame). The rheology of the layer is governed by the grain-size evolution, damage, and pinning model of Bercovici and Ricard (2012, 2013). In this case the low pressure zone P is imposed and then rotated about a vertical axis by 90 degrees three times (with roughly 10Myrs dimensional time between rotations to develop damaged weak zones) as an idealization of intermittent and chaotic subduction during the early Archean. The bands of damage induced by flow and deformation induced by the pressure low from a previous orientation are long-lived, inherited, and amplified by

the lithospheric flow of the next orientation, resulting in localized but passive bands of strike-slip vorticity Ω and positive divergence S (red and yellow contours; convergence with $S < 0$, indicated by blue contours, is actively driven by the pressure low). Thus a complete plate arises with a contiguous weak plate boundary (indicated by viscosity $\bar{\mu}$) while only being driven by subduction, which is the final state shown (with final dimensionless time indicated on the P frame). Dimensionless extrema for contours or vectors are indicated below each frame, save P which is always between 0 and 1. This model is proposed to explain the emergence of plate tectonics in the Archean, from proto-subduction 4Ga to global tectonics by ~ 3 Ga. (Adapted from Bercovici and Ricard 2014)

plate generation by its modulation of climate and not on direct strength reduction.

Finally, subduction initiation continues to be an extremely challenging issue in geodynamics (Stern 2004; Wada and King 2015). The strength of thick cold pre-subduction lithosphere is such that it should never go unstable and sink, at least not on geological time scales (or cosmological ones either). Thus how and why subduction zones form remains enigmatic. Mechanisms range from weakening by rifting (Kemp and Stevenson 1996; Schubert and Zhang 1997), sediment loading and water injection (Regenauer-Lieb et al. 2001), reactivation of pre-existing fault-zones (Toth and Gurnis 1998; Hall et al. 2003; Lebrun et al. 2003), initiation

by mantle plumes, especially in the Archean Earth (Gerya et al. 2015), and grain evolution and damage in passive margins (Mulyukova and Bercovici 2018). That subduction on the Earth is one-sided (i.e., the lithosphere or thermal boundary layer on one side of the subduction zone sinks, not both sides as in simple convection) is also a significant mystery, although it is possibly caused by hydrous melting and weakening of the mantle wedge above the sinking slab, which inhibits entrainment of the overriding lithosphere (Gerya et al. 2008). Finally, an unresolved enigma concerns the age of subduction zones. The convective picture of plate motions would have plates subduct when they get old, cold, and heavy. However, the seafloor age distribution implies that

subduction rate is independent of plate age, such that the age of plates at subduction zones is distributed from nearly zero age (i.e., subducting ridges) to the oldest ages of roughly 200 Myrs (see Becker et al. 2009).

Summary

Mantle convection is the central theory for how the Earth works, i.e., what drives geological motions as well as the cooling history and chemical evolution of the planet. The theory is based on the simple notion that the Earth (and other terrestrial planets) must cool to space and in so doing release heat and gravitational potential energy through thermal convection. Thermal convection theory itself is a well-established physical theory rooted in classical mechanics and fluid dynamics. Much of the physics of basic convection goes far in describing circulation and structure in the Earth's mantle, for example, the flow velocities, establishment of thermal boundary layers, and even the prevalence of slab-like downwellings and plume-like upwellings. However, numerous quandaries and paradoxes persist because much remains to be understood regarding the many complexities of the exotic convecting "fluid" in the mantle. How the Earth appears to be stirred by deep subducting slabs but still appear to be unmixed when it comes up at either mid-ocean ridges or ocean islands remains a mystery, much of which is due to our incomplete understanding of the processes of melting, chemical segregation, and mixing in the mantle; however, recent progress in imaging and modeling chemically dense piles on the core-mantle boundary (LLSVPs) has advanced our understanding considerably. How and why the Earth's mantle convects in the form of plate tectonics, unlike other terrestrial planets, remains one of the biggest questions in geoscience; this problem is undoubtedly related to the mantle's exotic rheology in which flow depends on multiple properties including temperature, stress, chemistry, and mineral grain size. Recent work coupling geodynamics with rock physics and mineral grain evolution has provided fresh insight and a new theoretical framework with which to understand how tectonic plates are possibly generated. Thus, while mantle convection remains one of the grand unifying physical theories of how the Earth works, many of the major questions and mysteries about the mantle remain unsolved and are thus ripe for discovery by future generations of Earth scientists.

Bibliography

- Austin N, Evans B (2007) Paleowattmeters: a scaling relation for dynamically recrystallized grain size. *Geology* 35:343–346
- Ballmer M, van Keken PE, Ito G (2015) Hotspots, large igneous provinces, and melting anomalies, chap. 10. In: Bercovici D (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics: mantle dynamics*, vol 7, 2nd edn. Elsevier, New York, pp 393–459
- Becker T, Conrad C, Buffett B, Müller R (2009) Past and present seafloor age distributions and the temporal evolution of plate tectonic heat transport. *Earth Planet Sci Lett* 278:233–242
- Bénard H (1900) Les tourbillons cellulaires dans une nappe liquide. *Revue générale des Sciences pures et appliquées* 11:1261–1271, 1309–1328
- Bénard H (1901) Les tourbillons cellulaires dans une nappe liquide transportant de la chaleur par convection en régime permanent. *Annales de Chimie et de Physique* 23:62–144
- Bercovici D (1993) A simple model of plate generation from mantle flow. *Geophys J Int* 114:635–650
- Bercovici D (1995) A source-sink model of the generation of plate tectonics from non-Newtonian mantle flow. *J Geophys Res* 100:2013–2030
- Bercovici D (2003) The generation of plate tectonics from mantle convection. *Earth Planet Sci Lett* 205:107–121
- Bercovici D (2015) Mantle dynamics, past, present and future: an overview. In: Bercovici D (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics: mantle dynamics*, vol 7, 2nd edn. Elsevier, New York, pp 1–30
- Bercovici D, Karato S (2003) Whole mantle convection and the transition-zone water filter. *Nature* 425:39–44
- Bercovici D, Mulyukova E (2018) A continuum model for phase mixing and grain-damage relevant to tectonic plate boundary evolution. *Phys Earth Planet Inter* 285:23–44. <https://doi.org/10.1016/j.pepi.2018.10.005>
- Bercovici D, Ricard Y (2005) Tectonic plate generation and two-phase damage: void growth versus grainsize reduction. *J Geophys Res* 110: B03,401. <https://doi.org/10.1029/2004JB003,181>
- Bercovici D, Ricard Y (2012) Mechanisms for the generation of plate tectonics by two-phase grain-damage and pinning. *Phys Earth Planet Inter* 202–203:27–55. <https://doi.org/10.1016/j.pepi.2012.05.003>
- Bercovici D, Ricard Y (2013) Generation of plate tectonics with two-phase grain-damage and pinning: source-sink model and toroidal flow. *Earth Planet Sci Lett* 365:275–288. <https://doi.org/10.1016/j.epsl.2013.02.002>
- Bercovici D, Ricard Y (2014) Plate tectonics, damage and inheritance. *Nature* 508:513–516. <https://doi.org/10.1038/nature13072>
- Bercovici D, Ricard Y (2016) Grain-damage hysteresis and plate-tectonic states. *Phys Earth Planet Inter* 253:31–47. <https://doi.org/10.1016/j.pepi.2016.01.005>
- Bercovici D, Skemer P (2017) Grain damage, phase mixing and plate-boundary formation. *J Geodyn* 108:40–55. <https://doi.org/10.1016/j.jog.2017.05.002>
- Bercovici D, Schubert G, Glatzmaier G (1992) Three-dimensional, infinite prandtl number, compressible convection in a basally heated spherical shell. *J Fluid Mech* 239:683–719
- Bercovici D, Ricard Y, Richards M (2000) The relation between mantle dynamics and plate tectonics: a primer. In: Richards MA, Gordon R, van der Hilst R (eds) *History and dynamics of global plate motions*, Geophysical monograph, vol 121. American Geophysical Union, Washington, DC, pp 5–46
- Bercovici D, Tackley PJ, Ricard Y (2015) The generation of plate tectonics from mantle dynamics, chap. 7. In: Bercovici D (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics: mantle dynamics*, vol 7, 2nd edn. Elsevier, New York, pp 271–318. <https://doi.org/10.1016/B978-0-444-53802-4.00135-4>
- Braun J, Chéry J, Poliakov A, Mainprice D, Vauchez A, Tomassi A, Daignières M (1999) A simple parameterization of strain localization in the ductile regime due to grain size reduction: a case study for olivine. *J Geophys Res Solid Earth* 104(B11):25167–25181. <https://doi.org/10.1029/1999JB900214>
- Cadek O, Ricard Y (1992) Toroidal/poloidal energy partitioning and global lithospheric rotation during Cenozoic time. *Earth Planet Sci Lett* 109:621–632
- Chandrasekhar S (1961) *Hydrodynamic and Hydromagnetic stability*. Oxford University Press, New York

- Christensen U (1995) Effects of phase transitions on mantle convection. *Annu Rev Earth Planet Sci* 23:65–88
- Christensen U, Hofmann A (1994) Segregation of subducted oceanic crust in the convecting mantle. *J Geophys Res* 99(B10):19,867–19,819
- Coltice N, Husson L, Faccenna C, Arnould M (2019) What drives tectonic plates? *Sci Adv* 5(10). <https://doi.org/10.1126/sciadv.aax4295>
- Condie K, Kröner A (2008) When did plate tectonics begin? Evidence from the geologic record. In: Condie K, Pease V (eds) *When did plate tectonics begin on planet earth?* Geological Society of America, Boulder, special Paper 440, pp 281–294
- Conrad CP, Steinberger B, Torsvik TH (2013) Stability of active mantle upwelling revealed by net characteristics of plate tectonics. *Nature* 498(7455):479–482. <https://doi.org/10.1038/nature12203>
- Cross A, Skemer P (2017) Ultramylonite generation via phase mixing in high strain experiments. *J Geophys Res-Solid Earth* 122:1744–1759. <https://doi.org/10.1002/2016JB013801>
- Dannberg J, Gassmöller R (2018) Chemical trends in ocean islands explained by plume–slab interaction. *Proc Natl Acad Sci* 115(17):4351–4356. <https://doi.org/10.1073/pnas.1714125115>
- Davaille A, Limare A (2015) Laboratory studies of mantle convection, chap. 3. In: Bercovici D (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics: mantle dynamics*, vol 7, 2nd edn. Elsevier, Oxford, pp 73–144. <https://doi.org/10.1016/B978-0-444-53802-4.00128-7>
- Davies G, Richards M (1992) Mantle convection. *J Geol* 100:151–206
- De Bresser J, ter Heege J, Spiers C (2001) Grain size reduction by dynamic recrystallization: can it result in major rheological weakening? *Int J Earth Sci* 90:28–45
- Dietz RS (1961) Continent and ocean basin evolution by spreading of the sea floor. *Nature* 190(4779):854–857
- Dumoulin C, Bercovici D, Wessel P (1998) A continuous plate-tectonic model using geophysical data to estimate plate margin widths, with a seismicity based example. *Geophys J Int* 133:379–389
- Foley B, Becker T (2009) Generation of plate-like behavior and mantle heterogeneity from a spherical, viscoplastic convection model. *Geochem Geophys Geosyst* 10(8):Q08001. <https://doi.org/10.1029/2009GC002378>
- Foley BJ, Bercovici D, Landuyt W (2012) The conditions for plate tectonics on super-earths: inferences from convection models with damage. *Earth Planet Sci Lett* 331–332:281–290
- Forsyth D, Uyeda S (1975) On the relative importance of the driving forces of plate motion. *Geophys J R Astron Soc* 43:163–200
- French SW, Romanowicz B (2015) Broad plumes rooted at the base of the earth's mantle beneath major hotspots. *Nature* 525(7567):95–99. <https://doi.org/10.1038/nature14876>
- Fukao Y, Obayashi M (2013) Subducted slabs stagnant above, penetrating through, and trapped below the 660 km discontinuity. *J Geophys Res Solid Earth* 118(11):5920–5938. <https://doi.org/10.1002/2013JB010466>
- Garnero EJ, McNamara AK (2008) Structure and dynamics of earth's lower mantle. *Science* 320(5876):626–628. <https://doi.org/10.1126/science.1148028>
- Garnero EJ, McNamara AK, Shim S-H (2016) Continent-sized anomalous zones with low seismic velocity at the base of earth's mantle. *Nat Geosci* 9(7):481–489. <https://doi.org/10.1038/ngeo2733>
- Gerya TV, Connolly JA, Yuen DA (2008) Why is terrestrial subduction one-sided? *Geology* 36(1):43–46. <https://doi.org/10.1130/G24060A.1>
- Gerya TV, Stern RJ, Baes M, Sobolev SV, Whattam SA (2015) Plate tectonics on the earth triggered by plume-induced subduction initiation. *Nature* 527(7577):221–225
- Grand S, van der Hilst R, Widiyantoro S (1997) Global seismic tomography: a snapshot of convection in the earth. *GSA Today* 7:1–7
- Gregg T (2015) 10.09-planetary tectonics and volcanism: the inner solar system. In: Schubert G (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics*, 2nd edn. Elsevier, Oxford, pp 307–325. <https://doi.org/10.1016/B978-0-444-53802-4.00187-1>
- Hager B, O'Connell R (1979) Kinematic models of large-scale flow in the earth's mantle. *J Geophys Res* 84:1031–1048
- Hales A (1936) Convection currents in the Earth. *Mon Not R Astron Soc Geophys Suppl* 3:372–379
- Hall CE, Gurnis M, Sdrolias M, Lavier LL, Mueller RD (2003) Catastrophic initiation of subduction following forced convergence across fracture zones. *Earth Planet Sci Lett* 212:15–30
- Halliday A (2001) In the beginning. ... *Nature* 409:144–145
- Hansen LN, Cheadle MJ, John BE, Swapp SM, Dick HJB, Tucholke BE, Tivey MA (2013) Mylonitic deformation at the Kane oceanic core complex: implications for the rheological behavior of oceanic detachment faults. *Geochem Geophys Geosyst* 14(8):3085–3108. <https://doi.org/10.1002/ggge.20184>
- Hernlund J, McNamara A (2015) The core-mantle boundary region, chap. 11. In: Bercovici D (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics: mantle dynamics*, vol 7, 2nd edn. Elsevier, New York, pp 461–519
- Herwegh M, Linckens J, Ebert A, Berger A, Brodhag S (2011) The role of second phases for controlling microstructural evolution in poly-mineralic rocks: a review. *J Struct Geol* 33(12):1728–1750. <https://doi.org/10.1016/j.jsg.2011.08.011>
- Hess H (1962) History of ocean basins. In: Engeln A, James H, Leonard B (eds) *Petrologic studies - a volume in honor of a.F. Buddington*. Geological Society of America, New York, pp 599–620
- Hirth G, Kohlstedt D (1996) Water in the oceanic upper mantle: implications for rheology, melt extraction and the evolution of the lithosphere. *Earth Planet Sci Lett* 144:93–108
- Hofmann A (1997) Mantle geochemistry: the message from oceanic volcanism. *Nature* 385:219–228
- Hofmann A (2003) Sampling mantle heterogeneity through oceanic basalts: isotopes and trace elements. In: Holland RCH, Turekian KK (eds) *Treatise on geochemistry: the mantle and core*. Elsevier, New York, pp 61–101
- Holmes A (1931) Radioactivity and earth movements, *Geol. Soc Glasgow Trans* 18:559–606
- Howard LN (1966) Convection at high Rayleigh number. In: Gortler H (ed) *Proceedings of the eleventh international congress of applied mechanics*. Springer, New York, pp 1109–1115
- Ito G, Mahoney J (2005a) Flow and melting of a heterogeneous mantle: 1. Method and importance to the geochemistry of ocean island and mid-ocean ridge basalts. *Earth Planet Sci Lett* 230:29–46
- Ito G, Mahoney J (2005b) Flow and melting of a heterogeneous mantle: 2. Implications for a chemically non-layered mantle. *Earth Planet Sci Lett* 230:47–63
- Jaupart C, Mareschal J-C (2011) *Heat generation and transport in the earth*. Cambridge University Press, New York
- Jaupart C, Labrosse S, Lucazeau F, Mareschal J-C (2015) Temperatures, heat and energy in the mantle of the earth. In: Bercovici D (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics: mantle dynamics*, vol 7, 2nd edn. Elsevier, New York, pp 223–270
- Jin D, Karato S, Obata M (1998) Mechanisms of shear localization in the continental lithosphere: inference from the deformation microstructures of peridotites from the Ivrea zone, northwestern Italy. *J Struct Geol* 20:195–209
- Karato S (1989) Grain growth kinetics in olivine aggregates. *Tectonophysics* 168:255–273
- Karato S (2008) *Deformation of earth materials: an introduction to the rheology of solid earth*. Cambridge University Press, Cambridge
- Karato S, Toriumi M, Fujii T (1980) Dynamic recrystallization of olivine single crystals during high temperature creep. *Geophys Res Lett* 7:649–652
- Karato S, Bercovici D, Leahy G, Richard G, Jing Z (2006) The transition zone water filter model for global material circulation: where do we stand? In: Jacobsen S, van der Lee S (eds) *Earth's deep water cycle*, AGU monograph series, vol 168. American Geophysical Union, Washington, DC, pp 289–313

- Kellogg L, Hager B, van der Hilst R (1999) Compositional stratification in the deep mantle. *Science* 283:1881–1884
- Kemp D, Stevenson D (1996) A tensile flexural model for the initiation of subduction. *Geophys J Int* 125:73–94
- Korenaga J (2007) Thermal cracking and the deep hydration of oceanic lithosphere: a key to the generation of plate tectonics? *J Geophys Res* 112. <https://doi.org/10.1029/2006JB004502>
- Korenaga J (2008) Urey ratio and the structure and evolution of earth's mantle. *Rev Geophys* 46:RG2007. <https://doi.org/10.1029/2007RG000241>
- Korenaga J (2010) On the likelihood of plate tectonics on super-earths: does size matter? *Astrophys J Lett* 725(1):L43–L46
- Korenaga J (2013) Initiation and evolution of plate tectonics on earth: theories and observations. *Annu Rev Earth Planet Sci* 41:117–151. <https://doi.org/10.1146/annurev-earth-050212-124208>
- Korenaga J (2017) On the extent of mantle hydration caused by plate bending. *Earth Planet Sci Lett* 457:1–9. <https://doi.org/10.1016/j.epsl.2016.10.011>
- Lamb S, Sington D (1998) *Earth story*. Princeton University Press, Princeton
- Landuyt W, Bercovici D (2009) Variations in planetary convection via the effect of climate on damage. *Earth Planet Sci Lett* 277:29–37
- Laughlin G, Lissauer J (2015) 10.21 – exoplanetary geophysics: an emerging discipline. In: Schubert G (ed) *Treatise on geophysics*, vol 7, 2nd edn. Elsevier, Oxford, pp 673–694. <https://doi.org/10.1016/B978-0-444-53802-4.00186-X>
- Le Pichon X (1968) Sea-floor spreading and continental drift. *J Geophys Res* 73(12):3661–3697. <https://doi.org/10.1029/JB073i012p03661>
- Leahy G, Bercovici D (2007) On the dynamics of a hydrous melt layer above the transition zone. *J Geophys Res* 112. <https://doi.org/10.1029/2006JB004631>
- Leahy G, Bercovici D (2010) Reactive infiltration of hydrous melt above the mantle transition zone. *J Geophys Res* 115:b08406. <https://doi.org/10.1029/2009JB006757>
- Lebrun J-F, Lamarche G, Collot J-Y (2003) Subduction initiation at a strike-slip plate boundary: the Cenozoic Pacific-Australian plate boundary, south of New Zealand. *J Geophys Res* 108. <https://doi.org/10.1029/2002JB002041>
- Lee C, Luffi P, Höink T, Li J, Dasgupta R, Hernlund J (2010) Upside-down differentiation and generation of a 'primordial' lower mantle. *Nature* 463(7283):930–933
- Lenardic A, Jellinek M, Moresi L-N (2008) A climate change induced transition in the tectonic style of a terrestrial planet. *Earth Planet Sci Lett* 271:34–42
- Linckens J, Herwegh M, Müntener O (2015) Small quantity but large effect—how minor phases control strain localization in upper mantle shear zones. *Tectonophysics* 643:26–43. <https://doi.org/10.1016/j.tecto.2014.12.008>
- Lithgow-Bertelloni C, Richards M, Ricard Y, O'Connell R, Engebretson D (1993) Toroidal-poloidal partitioning of plate motions since 120 ma. *Geophys Res Lett* 20:375–378
- McKenzie D, Parker R (1967) The north pacific: an example of tectonics on a sphere. *Nature* 216:1276–1280
- McNamara A, Zhong S (2005) Thermochemical structures beneath Africa and the pacific ocean. *Nature* 437:1136–1139
- Montelli R, Nolet G, Dahlen FA, Masters G, Engdahl ER, Hung S-H (2004) Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* 303:338–343
- Morgan W (1968) Rises, trenches, great faults and crustal blocks. *J Geophys Res* 73:1959–1982
- Morgan W (1971) Convection plumes in the lower mantle. *Nature* 230:42–43
- Mulyukova E, Bercovici D (2017) Formation of lithospheric shear zones: effect of temperature on two-phase grain damage. *Phys Earth Planet Inter* 270:195–212. <https://doi.org/10.1016/j.pepi.2017.07.011>
- Mulyukova E, Bercovici D (2018) Collapse of passive margins by lithospheric damage and plunging grain size. *Earth Planet Sci Lett* 484:341–352. <https://doi.org/10.1016/j.epsl.2017.12.022>
- Mulyukova E, Bercovici D (2019) The generation of plate tectonics from grains to global scales: a brief review. *Tectonics* 38. <https://doi.org/10.1029/2018TC005447>
- Mulyukova E, Steinberger B, Dabrowski M, Sobolev SV (2015) Survival of LLSVPs for billions of years in a vigorously convecting mantle: replenishment and destruction of chemical anomaly. *J Geophys Res Solid Earth* 120(5):3824–3847. <https://doi.org/10.1002/2014JB011688>
- Murakami M, Hirose K, Kawamura K, Sata N, Ohishi Y (2004) Post-perovskite phase transition in MgSiO₃. *Science* 304:855–858
- O'Connell R, Gable C, Hager B (1991) Toroidal-poloidal partitioning of lithospheric plate motion. In: Sabadini R et al (eds) *Glacial isostasy, sea level and mantle rheology*. Kluwer Academic, Norwell, pp 535–551
- Olgaard D (1990) The role of second phase in localizing deformation. In: Knipe R, Rutter E (eds) *Deformation mechanisms, rheology and tectonics*, vol 54. Geological Society Special Publication, London, pp 175–181
- Pearson J (1958) On convection cells induced by surface tension. *J Fluid Mech* 4:489–500
- Pekeris C (1935) Thermal convection in the interior of the earth. *Monthly Notices Roy Astron Soc, Geophys Suppl* 3:343–367
- Ranalli G (1995) *Rheology of the earth*. Chapman and Hall Publishers, London
- Rayleigh L (1916) On convective currents in a horizontal layer of fluid when the higher temperature is on the under side. *Philos Mag* 32:529–546
- Regenauer-Lieb K, Yuen D, Branlund J (2001) The initiation of subduction: criticality by addition of water? *Science* 294:578–580
- Revenaugh J, Sipkin S (1994) Seismic evidence for silicate melt atop the 410-km mantle discontinuity. *Nature* 369:474–476
- Rozel A, Ricard Y, Bercovici D (2011) A thermodynamically self-consistent damage equation for grain size evolution during dynamic recrystallization. *Geophys J Int* 184(2):719–728. <https://doi.org/10.1111/j.1365-246X.2010.04875.x>
- Rudolph ML, Zhong S (2013) Does quadrupole stability imply llsvp fixity? *Nature* 503(7477):E3–E4. <https://doi.org/10.1038/nature12792>
- Runcorn S (1962a) Towards a theory of continental drift. *Nature* 193:311–314
- Runcorn S (1962b) Convection currents in the earth's mantle. *Nature* 195:1248–1249
- Schubert G, Zhang K (1997) Foundering of the lithosphere at the onset of subduction. *Geophys Res Lett* 24:1527–1529
- Schubert G, Turcotte D, Yuen D (1975) Role of phase transitions in a dynamic mantle. *Geophys J R Astron Soc* 42:705–735
- Schubert G, Stevenson D, Cassen P (1980) Whole planet cooling and the radiogenic heat source contents of the earth and moon. *J Geophys Res* 85:2531–2538
- Schubert G, Turcotte D, Olson P (2001) *Mantle convection in the earth and planets*. Cambridge University Press, Cambridge
- Skemer P, Warren JM, Kelemen PB, Hirth G (2010) Microstructural and rheological evolution of a mantle shear zone. *J Petrol* 51:43–53. <https://doi.org/10.1093/petrology/egp057>
- Sleep N (2015) Evolution of the earth: plate tectonics through time, chap. 6. In: Stevenson D (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics: evolution of the earth*, vol 9, 2nd edn. Elsevier, Oxford, pp 145–172. <https://doi.org/10.1016/B978-0-444-53802-4.00158-5>
- Solomatov V (1995) Scaling of temperature dependent and stress dependent viscosity convection. *Phys Fluids* 7:266–274
- Song T, Helmberger DV, Grand SP (2004) Low velocity zone atop the 410 seismic discontinuity in the northwestern United States. *Nature* 427:530–533
- Stern RJ (2004) Subduction initiation: spontaneous and induced. *Earth Planet Sci Lett* 226:275–292

- Tackley P (2000) Self-consistent generation of tectonic plates in time-dependent, three-dimensional mantle convection simulations. 2. Strain weakening and asthenosphere. *Geochem Geophys Geosyst* 1. <https://doi.org/10.1029/2000GC000043>
- Tackley P (2002) Strong heterogeneity caused by deep mantle layering. *Geochem Geophys Geosyst* 3. <https://doi.org/10.1029/2001GC000167>
- Tackley PJ (2015) Mantle geochemical geodynamics, chap. 12. In: Bercovici D (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics: mantle dynamics*, vol 7, 2nd edn. Elsevier, New York, pp 521–585
- Tackley P, Stevenson D, Glatzmaier GA, Schubert G (1993) Effects of an endothermic phase transition at 670 km depth in a spherical model of convection in the earth's mantle. *Nature* 361:699–704
- Tackley P, Ammann M, Brodholt J, Dobson D, Valencia D (2013) Mantle dynamics in super-earth: post-perovskite rheology and self-regulation of viscosity. *Icarus* 225(1):50–61. <https://doi.org/10.1016/j.icarus.2013.03.013>
- Tasaka M, Zimmerman ME, Kohlstedt DL (2017) Rheological weakening of olivine + orthopyroxene aggregates due to phase mixing: 1. Mechanical behavior. *J Geophys Res Solid Earth* 122(10):7584–7596. <https://doi.org/10.1002/2017JB014333>
- Tauzin B, Debayle E, Wittlinger G (2010) Seismic evidence for a global low-velocity layer within the earth's upper mantle. *Nat Geosci* 3:718–721. <https://doi.org/10.1038/ngeo969>
- Toth G, Gurnis M (1998) Dynamics of subduction initiation at pre-existing fault zones. *J Geophys Res* 103(18):053–18,067
- Turcotte D, Schubert G (2014) *Geodynamics*. Cambridge University Press, Cambridge, UK
- Valencia D, O'Connell R, Sasselov D (2007) Inevitability of plate tectonics on super-earth. *Astrophys J* 670:L45–L48
- van der Hilst R, Widiyantoro S, Engdahl E (1997) Evidence for deep mantle circulation from global tomography. *Nature* 386:578–584
- van Heck H, Tackley P (2008) Planforms of self-consistently generated plates in 3D spherical geometry. *Geophys Res Lett* 35:L19312. <https://doi.org/10.1029/2008GL035190>
- van Heck H, Tackley P (2011) Plate tectonics on super-earth: equally or more likely than on earth. *Earth Planet Sci Lett* 310(3–4):252–261. <https://doi.org/10.1016/j.epsl.2011.07.029>
- van Keken P, Hauri E, Ballentine C (2002) Mantle mixing: the generation, preservation, and destruction of chemical heterogeneity. *Annu Rev Earth Planet Sci* 30:493–525
- Vine FJ (1966) Spreading of the ocean floor: new evidence. *Science* 154(3755):1405–1415. <https://doi.org/10.1126/science.154.3755.1405>
- Vine F, Matthews D (1963) Magnetic anomalies over oceanic ridges. *Nature* 199:947–949
- Wada I, King SD (2015) Dynamics of subducting slabs: numerical modeling and constraints from seismology, geoid, topography, geochemistry and petrology, chap. 9. In: Bercovici D (ed). (Gerald Schubert editor-in-chief) *Treatise on geophysics: mantle dynamics*, vol 7, 2nd edn. Elsevier, New York, pp 339–392. <https://doi.org/10.1016/B978-0-444-53802-4.00135-4>
- Warren JM, Hirth G (2006) Grain size sensitive deformation mechanisms in naturally deformed peridotites. *Earth Planet Sci Lett* 248(1–2):438–450. <https://doi.org/10.1016/j.epsl.2006.06.006>
- Wegener A (1924) *The origin of continents and oceans*. Methuen and Co. Ltd., London
- Weidner D, Wang Y (1998) Chemical- and clapeyron-induced buoyancy at the 660 km discontinuity. *J Geophys Res* 103:7341–7441
- Weinstein S, Olson P (1992) Thermal convection with non-Newtonian plates. *Geophys J Int* 111:515–530
- White D (1988) The planforms and onset of convection with temperature-dependent viscosity. *J Fluid Mech* 191:247–286
- White S, Burrows S, Carreras J, Shaw N, Humphreys F (1980) On mylonites in ductile shear zones. *J Struct Geol* 2:175–187
- Wiesman H, Zimmerman M, Kohlstedt D (2018) Laboratory investigation of mechanisms for phase mixing in olivine+ferropiclasite aggregates. *Phil Trans R Soc A* 376. <https://doi.org/10.1098/rsta.2017.0417>
- Wolfe C, Solomon S, Laske G, Collins J, Detrick R, Orcutt J, Bercovici D, Hauri E (2009) Mantle shear-wave velocity structure beneath the Hawaiian hotspot. *Science* 326:1388–1390
- Wood B, Corgne A (2015) Mineralogy of the earth: trace elements and hydrogen in the earth's transition zone and lower mantle, chap. 4. In: Price GD, Stixrude L (eds). (Gerald Schubert editor-in-chief) *Treatise on geophysics*, vol 2, 2nd edn. Elsevier, Oxford, pp 61–84. <https://doi.org/10.1016/B978-0-444-53802-4.00031-2>
- Zhong S, Liu X (2016) The long-wavelength mantle structure and dynamics and implications for large-scale tectonics and volcanism in the Phanerozoic. *Gondwana Res* 29(1):83–104. <https://doi.org/10.1016/j.gr.2015.07.007>
- Zhong S, Zuber M, Moresi L, Gurnis M (2000) Role of temperature-dependent viscosity and surface plates in spherical shell models of mantle convection. *J Geophys Res* 105(11):063–11,082

Mantle D'' Layer

Thorne Lay

Earth and Planetary Sciences Department, University of California, Santa Cruz, CA, USA

Definition and Introduction

Earth's lower mantle extends from the base of the transition zone (commonly assigned a depth of 660 km, corresponding to the average depth of a global seismic velocity discontinuity) to the core-mantle boundary (CMB) at a depth of ~2,900 km below the surface. This thick layer of rock comprised of silicate and oxide minerals has gradual increases with depth of P- and S-wave seismic velocities and density that are generally consistent with adiabatic self-compression of a uniform composition material over most of the depth range (see ► [“Earth's Structure, Lower Mantle”](#)). Large-scale lateral heterogeneities in the seismic velocities of about $\pm 1\%$ are apparent in the central lower mantle, and these likely involve thermal and chemical signatures of upwellings and downwellings in the slowly convecting deep mantle system (see ► [“Mantle Convection”](#)). The deepest few hundred kilometers of the lower mantle have generally reduced seismic velocity gradients, localized seismic velocity discontinuities, and strong large-scale seismic velocity heterogeneities of $\pm 2\text{--}4\%$. This portion of the lower mantle is called the D'' layer (or, more commonly, the D'' region, because the onset of anomalous properties varies significantly in height above the CMB from place to place) in recognition of its distinctive properties relative to the shallower lower mantle (D' layer). Localized regions of extreme (10–30%) seismic velocity reductions in thin layers or lumps only tens of kilometers thick are found at the base of D''. The structural

heterogeneities in the D'' region are usually interpreted as manifestations of complex thermal and chemical boundary layers at the CMB, to some extent mirroring the structural heterogeneities found in the near-surface lithospheric boundary layers. The structure and dynamics of the D'' region are believed to be of fundamental importance to the mantle and core dynamical systems, prompting many seismological, geodynamical, and mineral physics investigations of D'' properties.

The D'' region may have distinct bulk composition from the rest of the lower mantle, involving residues from deep mantle melting, core formation, chemical reactions between the core and mantle, or dynamical segregation of materials over time, but this possibility has to be assessed in the context of expected behavior of major lower-mantle minerals for the extreme pressure-temperature (P-T) conditions near the CMB. The primary minerals in the lower mantle are thought to be $(\text{Mg}_x\text{Fe}_{1-x})\text{SiO}_3$ perovskite, $(\text{Mg}_x\text{Fe}_{1-x})\text{O}$ ferropericlasite, and smaller amounts of CaSiO_3 perovskite and high-pressure forms of SiO_2 (see ► [“Earth’s Structure, Lower Mantle”](#)). The relative amount of Mg versus Fe, given by x , is ~ 0.9 , but the properties of the minor Fe component are being intensely investigated because transitions from high-spin to low-spin in Fe^{3+} and Fe^{2+} are expected in lower mantle perovskite and ferropericlasite as pressure increases, with strong effects on chemical and transport properties of the material in the deep mantle. Some Al may substitute in both the Mg and Si sites in the Mg-perovskite. Mg-perovskite comprises as much as 70% of the lower mantle, making it the most abundant mineral in the Earth. A phase change in Mg-perovskite, discovered in 2004, occurs under P-T conditions existing a few hundred kilometers above the CMB, likely giving rise to some of the complexities of the D'' region (Hirose and Lay 2008). Mineralogical complexity of D'' is expected to be enhanced relative to the bulk of the lower mantle due to proximity to huge contrasts in composition and physical properties across the CMB.

The CMB separates the solid silicate- and oxide mineral mantle from the molten metallic-alloy outer core. The density increases from about $5,500 \text{ kg/m}^3$ to $9,900 \text{ kg/m}^3$ across the CMB, and the viscosity decreases by many orders of magnitude. These contrasts maintain a very sharp boundary and result in thermal and chemical boundary layers in both D'' and the outermost core, although any core structures may be very thin and hard to detect (see ► [“Earth’s Structure, Core”](#)). With any mass flux across the CMB being constrained to involve chemical diffusion, heat transports across the CMB by conduction from the hot core into the lower-temperature mantle. The temperature drop across the resulting CMB thermal boundary layer is estimated to be in the range $1,000\text{--}2,000^\circ\text{C}$, and there is a heat flow of $5\text{--}15 \text{ TW}$ across the boundary (the heat flow through Earth’s surface is $\sim 46 \text{ TW}$) (Lay et al. 2008). Over time, relatively dense

materials in the mantle (for example, any basaltic component in subducted slabs that penetrate into the lower mantle) are likely to have concentrated in D'', while light dross, expelled from the core may have accumulated on the underside of the CMB. Thermo-mechanical and electromagnetic interactions across the CMB couple the mantle and core dynamical systems, and this influences the geodynamo (see ► [“Core-Mantle Coupling”](#)). The D'' region is relatively accessible to seismic imaging, and its inferred structural and dynamical properties are discussed here. Trønnes (2009), Garnero and McNamara (2008), Lay (2007), and many papers in Hirose et al. (2007) provide technical reviews of ongoing research topics and conceptual models for structures and processes occurring in D''.

Seismic Velocity Models for D''

Seismology provides the highest resolution of D'' elastic material properties because it is possible to decipher seismic wave interactions with the lowermost mantle and CMB in recordings of ground shaking at seismic stations around the world. Seismic wave observations can resolve D'' region P- and S-wave velocities, reflections from sharp velocity jumps, shear-wave splitting caused by anisotropy of the elastic structure, and scattering properties. One-dimensional (1D) depth-varying seismic velocity models, such as the Preliminary Reference Earth Model (PREM) of Dziewonski and Anderson (1981), usually have reduced velocity gradients in the deepest $\sim 150 \text{ km}$ of the mantle, representing the global departure of D'' velocity structure from that of the overlying lower mantle. Efforts were initially made to interpret the low (or even negative) velocity gradients in such laterally averaged models of D'' structure as the result of a superadiabatic temperature increase across the CMB thermal boundary layer; strong temperature increases are required to account for seismic velocity reductions due to the low thermal expansion coefficient of deep mantle rock under high pressures. However, it is now clear that there is no meaningful “average” 1D structure for the D'' region useful for a robust interpretation of the boundary layer. It is more useful to discuss the seismic properties of D'' by emphasizing the lateral variations in structure, as is also the case for the lithosphere.

Seismic tomography has been applied to develop three-dimensional (3D) seismic velocity models for the entire mantle for over 30 years. This method extracts 3D P-wave and S-wave velocity fluctuations relative to 1D reference models by inversion of massive data sets with crossing ray paths for which seismic wave travel times are measured. 3D velocity inversions indicate that seismic velocity heterogeneity in the mid-lower mantle is relatively weak, but that in the lowermost $300\text{--}500 \text{ km}$ of the mantle heterogeneity increases with depth, becoming stronger in D'' than anywhere else in the

Earth except for the uppermost mantle where major thermal, chemical and partially molten heterogeneities are known to exist.

An unanticipated aspect of the D'' seismic velocity heterogeneity revealed by seismic tomography is that it is dominated by large-scale patterns, with huge continuous volumes of high or low-seismic-velocity material having scale-lengths of hundreds to thousands of kilometers. This predominance of large-scale variations is observed in both P-wave and S-wave velocities, with relatively high-velocity regions of D'' underlying the circum-Pacific margins and two nearly antipodal regions of pronounced low S-wave velocity beneath the southern central Pacific and southern Africa/southern Atlantic/southern Indian Ocean (Fig. 1a). High-seismic-velocity regions are likely to have lower temperature than low-seismic-velocity regions, although chemical differences and phase changes may contribute to the variations. Recent global tomographic models have $\pm 4\%$ S-wave velocity variations in D'', two or three times stronger than those found in the mid-lower mantle, and comparable to the $\pm 5\text{--}8\%$ variations at depths near 150 km in the upper mantle. While mid-mantle variations tend to have more spatially concentrated heterogeneities, there is a significant degree of radial continuity of high- and low-velocity anomaly patterns throughout the lower mantle, possibly linking subduction zones at the surface to high-velocity regions in D'' and hot-spot locations at the surface to low-velocity regions in D'' (Garnero and McNamara 2008).

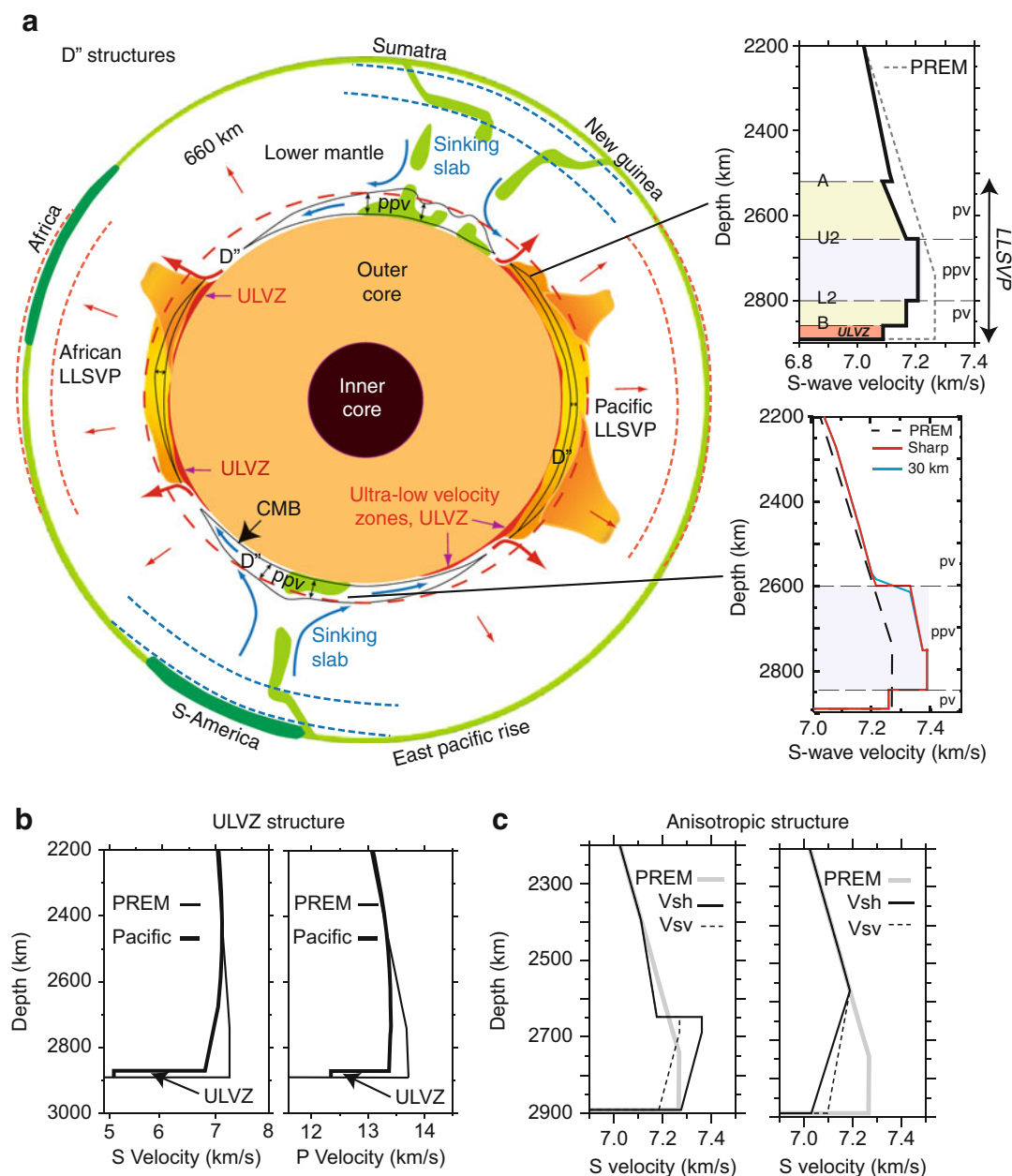
P-wave velocity variations in current tomographic models have $\pm 1.0\text{--}1.5\%$ fluctuations, with low-velocity areas beneath the southern Pacific and south Atlantic/Africa and high velocities under eastern Eurasia. There is not as strong of an increase in P-wave velocity heterogeneity in D'' relative to that in the overlying mantle as there is for S-wave velocity, which can be partially attributed to the expected greater sensitivity of S-wave velocity to temperature variations in a thermal boundary layer. There is a large-scale spatial correlation between S- and P-wave velocity patterns, as expected for temperature-induced variations, but in some regions, such as beneath the northern Pacific, this correlation breaks down. In other regions, such as beneath the central Pacific, the S-wave variations are much stronger than the P-wave variations even though both have the same sign. The decorrelation of the seismic velocities and the variation in their relative strengths provide strong evidence that thermal variations alone cannot explain the large-scale patterns of seismic velocity heterogeneity; so there is likely to be chemical heterogeneity present in D'' (Trampert et al. 2004). It is important to recognize that tomography patterns are relative to the global average velocities at D'' depths, and that these averages do not necessarily define "normal" mantle structure. This results in uncertainty in interpretations of the velocity fluctuations (for example, high velocity regions may be anomalously low temperature,

perhaps due to cool downwellings that have disrupted an overall hot boundary layer, or low-velocity regions may be anomalously hot regions, perhaps partially melted, within a chemically distinct, high velocity layer). Resolution of finer scale seismic properties in D'' is pursued in order to overcome this ambiguity in interpreting global seismic tomography images.

Seismic Velocity Discontinuities in D''

Rapid P-wave and S-wave velocity increases and/or decreases are observed several hundred kilometers above the CMB in many regions (Lay and Garnero 2007), and are sometimes used to define the top of D''. The most prominent structure is a rapid shear velocity increase of 1.5–3% 150–350 km above the CMB found over intermediate-scale (500–1,000 km) regions beneath circum-Pacific margins and some other locations (Fig. 1a). The increase in velocity may be distributed over a few tens of kilometers or it may be very abrupt (a "discontinuity") (Lay 2008), and in some regions it can vary in depth by as much as 100 km over lateral scales of just 200 km. The velocity increase is required to account for observed S-wave reflections that arrive before reflections from the CMB. Similar structures have been found in localized regions for P-waves, usually with a smaller velocity increase of 0.5–1.0%, but in some regions there is a P-wave velocity decrease at the same depth as an S-wave increase (Hutko et al. 2008). Lay and Garnero (2007) review many observations and models for this D'' seismic velocity discontinuity.

The D'' S-wave velocity discontinuity is commonly observed in regions with large volumes of high S-wave velocity material in D'', indicating that it is associated with relatively low temperature environments. An abrupt increase in velocity with depth is not expected for a thermal boundary layer structure, so most interpretations of this correlation invoke the notion of localized ponding of cool subducted lithospheric slabs that have sunk to the lowermost mantle, retaining enough thermal and chemical anomaly to account for the high seismic velocity. However, it is not clear that slab thermal and compositional anomalies can account for a sharp reflecting structure, so additional presence of a phase change may be required. Alternatively, chemically distinct high velocity material may be present in these regions, perhaps involving ancient accumulated material concentrated during core formation or segregated oceanic crustal materials that have accumulated over Earth history. Observations of an S-wave velocity increase beneath the central Pacific, a relatively low velocity region far from any historical subduction zone, complicates any attempt to interpret the velocity discontinuity solely as the result of recent slab thermal anomalies (Lay and Garnero 2007).



Mantle D'' Layer, Fig. 1 (a) Schematic cross-section through the Earth indicating large structures in the D'' region (below the red dashed line) above the core-mantle boundary (CMB) (adapted from Trønnes 2009). The two large low shear velocity provinces (LLSVPs) under Africa and the Pacific are indicated, with thin ultra-low velocity zones (ULVZs) at their base and margins. Areas of upwelling warm mantle surround and overlie the LLSVPs. Areas of relatively cool mantle downwellings are underlain by D'' material with post-perovskite (ppv) occurrence. Detailed seismic S-wave velocity models from both an LLSVP margin and a down-welling region are shown on the right, with the D'' velocity

discontinuity seen at depths of 2,600–2,650 km. Both regions have a deeper velocity decrease that may represent conversion from post-perovskite back to perovskite (pv) in the steep thermal gradient above the CMB at 2,891 km depth. (b) S-wave and P-wave velocity models from areas of intense ULVZ under the central Pacific, showing strong velocity reductions below low velocity (LLSVP) regions. (c) Representative anisotropic S-wave velocity models, indicating high Vsh velocities below circum-Pacific regions with strong D'' discontinuities and high Vsv velocities in LLSVP environments

Post-perovskite in D''

The major seismic velocity discontinuities in the transition zone are generally attributed to phase changes in upper mantle minerals that cause abrupt changes in density and elastic wave

velocities. The experimental discovery of a phase transition in MgSiO_3 perovskite (Murakami et al. 2004) for P-T conditions near the top of D'' may provide a corresponding explanation for the D'' S-wave velocity discontinuity. The high-pressure polymorph is called post-perovskite, and laboratory and

theoretical predictions of its properties indicate that it should have 1–2% higher S-wave velocity, little, or no change in P-wave velocity, and 1–1.5% higher density than perovskite, which can account for some of the seismological complexities of D'' (Hirose et al. 2007). Presence of Fe and Al in the perovskite cause the phase transition to occur over a pressure range rather than at a single pressure for a given temperature, and this may be difficult to reconcile with an efficient seismic wave reflector. The perovskite-to-post-perovskite transition occurs at lower pressure for lower temperatures, thus lateral temperature variations in a thermal boundary layer are expected to modulate the depth of the transition. If the phase transition occurs in a relatively cool region of accumulated downwelling slab material, the discontinuity should be higher above the CMB than in warmer areas of D'' with the same composition. Some calculations indicate that the pressure of the transition may decrease if Fe is present in the perovskite, thus chemical heterogeneity could also modulate the depth of the phase transition.

The P-T behavior of the post-perovskite phase change tends to enhance thermal instabilities within the thermal boundary layer, but its thermal transport properties are still uncertain due to the possibility of Fe high-spin to low-spin transition and the lack of constraint on Fe partitioning coefficients between post-perovskite and ferropericlase. One of the most important attributes of the phase transition is that it can provide an absolute temperature tie-point for D'', if an observed S-wave reflector at a specific depth (hence, pressure) is correctly attributed to the phase change and if the P-T-X behavior (where X indicates precise mineralogical composition) of the phase change is experimentally and theoretically constrained. This is important because estimates of temperature structure in the lowermost mantle and outermost core are largely based on extrapolations over tremendous depth ranges of laboratory-estimated-temperature tie-points for the olivine-spinel phase transition near 410-km depth and the iron alloy solidus at the 5,150-km deep inner core-outer core boundary.

Laboratory measurements indicate that the temperature at a depth of about 2,600 km is close to 2,500 K, if the D'' discontinuity is caused by transition to post-perovskite. With an estimated outermost core temperature of 3,500–4,000 K, this favors a 1,000–1,500° increase in temperature across the D'' thermal boundary layer. Such a strong temperature increase could cause a reversion from post-perovskite to perovskite to take place in the hottest region right above the CMB, and some evidence for such a second crossing of the phase boundary has been inferred from seismic observations of a S-wave velocity decrease 50–100 km above the CMB (Fig. 1a; Lay et al. 2006; van der Hilst et al. 2007). The effects of the phase transition on seismic velocities must be considered when inferring lateral temperature structure in D'' from seismic velocity variations, given that the volume of the high

S-wave velocity post-perovskite material should vary laterally in either a depth-modulated layer or a “lens,” depending on stability of post-perovskite right at the CMB.

Large Low Shear Velocity Provinces in D''

The two large antipodal regions with low S-wave velocity appear to be chemically distinct regions of D'' (Fig. 1a). The margins of these large low shear velocity provinces (LLSVPs) have abrupt steep-walled lateral gradients, over scales of tens of kilometers, which indicates a chemical change relative to surrounding D'' material rather than just a thermal change. The region beneath the southern Atlantic and Africa has been modeled as having margins with an abrupt –1 to –3% S-wave velocity decrease 250–300 km above the CMB (shallowing to ~800 km above the CMB under Africa), with average velocities in D'' that are 3–5% lower than for PREM (e.g., Wang and Wen 2006). The sub-Pacific LLSVP has comparable velocity contrasts and varying vertical extent, and may be two separate mounds of material that extend upward into the central lower mantle. P-wave velocity tends to be slightly low in the LLSVPs, but less than what would be expected if the S-wave velocity reductions were entirely caused by high temperatures. Thus, there is an anomalously high incompressibility in the LLSVPs, which also favors distinct chemistry. While limited in resolution, free oscillation measurements suggest that the LLSVPs are relatively high density, adding to the evidence for chemical inhomogeneity, and suggesting that these are not buoyant ‘superplumes’ as some researchers have suggested. Internal structure of LLSVPs has been detected in the form of localized velocity discontinuities (Fig. 1a) and the presence of underlying ultra-low velocity zones (see below). The velocity discontinuities may represent chemical heterogeneity or possibly post-perovskite phase transition in material with distinct Fe and Al content from surrounding high seismic velocity areas of D''.

The current position of the LLSVPs can be attributed to accumulation of chemically distinct, relatively dense material in D'' that has been displaced away from circum-Pacific areas of downwelling slab materials over the past several hundred million years (Garnero and McNamara 2008). Thermal calculations indicate that any dense chemical piles in a convecting mantle will be relatively hot as a result of inefficient heat loss, so LLSVPs would likely have combined effects of high incompressibility, high temperature, and high density. This predicts complex dynamical behavior, but the presence of the structures in D'' at this point in Earth's evolution suggests that the features are long-lived and are either remnants of much larger chemical anomalies that have slowly been entrained by mantle flow or they are being regenerated by ongoing chemical

differentiation. Aggregation of basaltic (ocean crust) components of subducted slabs is one possible mechanism for accumulation of chemically distinct material that may be sustaining the LLSVPs. The total volume of LLSVP material is several percent of the mantle, so these are significant chemical reservoirs. The possibility that the LLSVPs have been close to their present locations for hundreds of millions of years is supported by the fact that reconstructed emplacement locations of many Large Igneous Provinces (LIPs) overlie margins of the LLSVPs (Torsvik et al. 2006). If plume upwellings rise from the lateral margins of the dense piles, as suggested by numerous dynamical models, this association may be understood as a first order impact of D'' structure on geological processes at Earth's surface.

Ultra-Low Velocity Zones in D''

Thin layers or mounds of material with very strong seismic velocity reductions have been detected in D'' just above the CMB (Fig. 1b), and these are called ultralow velocity zones (ULVZs). P- and S-wave velocity models for ULVZ may have a thin layer or mound of 10–40 km thick and from hundreds to thousands of kilometers across with P-wave velocity reductions of –4 to –10% and S-wave velocity reductions of –8 to –30% (Thorne and Garnero 2004). These structures are commonly found near the margins of LLSVPs (Fig. 1a), but also occur in localized regions elsewhere. They are detected in seismic waves that reflect from or graze along the CMB.

The magnitude of the velocity reductions in the ULVZs and the factor of 2–3 ratio of S-wave velocity/P-wave velocity decrement require either the presence of a melt component or very strong chemical contrast. The temperature in the D'' thermal boundary layer will reach a peak right at the CMB (which is nearly isothermal due to the rapid convective flow occurring in the core); so ULVZs are intrinsically the hottest regions in the mantle. However, they are not globally detectable (a very thin layer < ~1 km could be present everywhere without being resolved by seismic data); so a combination of partial melting and chemical heterogeneity is implied by the patchy nature of the thicker regions of ULVZ. Seismic data indicate that the ULVZ material may be ~10% denser than surrounding D'' material, favoring high Fe content, which would contribute to the strong seismic wave velocity reductions. ULVZ affiliation with LLSVP margins may be associated with thermal convection in the LLSVPs, along with interactions with flow in the surrounding mantle (Garnero and McNamara 2008). It is not yet clear how to account for ULVZ chemical evolution, but one possibility is that they are the residue of a much more extensive lower mantle magma ocean which has largely solidified (Labrosse et al. 2007).

Seismic Velocity Anisotropy in D''

The seismic velocity structure in D'' is more anisotropic than the shallower lower mantle, with seismic velocities being dependent on the direction of propagation and the polarization of ground shaking. This results in shear-wave splitting, involving an S-wave separating into two-components with orthogonal polarization, one traveling slightly faster than the other while in the D'' region. By measuring the polarizations and travel time difference between the fast and slow S-waves, it is possible to determine the anisotropic characteristics of the medium. For most S-wave phases with ray paths grazing horizontally through the D'' region, the data can be explained by models in which horizontally polarized (SH) vibrations travel with 1–3% higher velocities (V_{sh}) than vertically polarized (SV) vibrations (Fig. 1c). This behavior is consistent with the medium having vertical transverse isotropy (VTI), which can result from hexagonally symmetric minerals with vertically oriented symmetry axes or from stacks of thin horizontal layers with periodic velocity fluctuations. The regions with the best documented cases for strong VTI in D'' tend to have higher than average S-wave velocities and strong D'' discontinuities (Fig. 1c). There are observations favoring slightly tilted (non-vertical) transverse isotropy (Maupin et al. 2005), which results in weak coupling of the SH and SV signals (the fast wave is still close to the SH polarization), as well as limited regions where SV signals are found to propagate with higher velocities (V_{sv}) than SH signals (such as in the central Pacific LLSVP, Fig. 1c).

Anisotropy in D'' is likely to be caused by either lattice-preferred orientation (LPO) or shape-preferred orientation (SPO). LPO can arise when minerals systematically orient within a flowing medium. This may exist in D'' if dislocation creep occurs, which requires relatively high stresses and low temperatures as might be found near slab downwellings. Ferropericlase has been shown to be a viable candidate for development of LPO in horizontal shear flows above the CMB, if suitable temperatures and stresses are present. It is also possible that certain slip systems can be activated such that post-perovskite can account for shear-wave splitting observed in some regions. If a region of D'' is too warm for dislocation creep, deformation will occur by diffusion processes and seismic anisotropy is only viable if heterogeneities (either chemical blobs or pockets of partial melt) are systematically aligned with the flow. Transitions from horizontal to vertical flows may account for SPO transitions from VTI to horizontal transverse isotropy, which can have V_{sv} higher than V_{sh} . Efforts continue to improve the characterization of D'' anisotropy because it has the potential to reveal ongoing deformation processes occurring in the boundary layer.

Summary

The diverse seismic properties of the D'' region appear to demonstrate the presence of heterogeneity at the base of the mantle associated with thermal and chemical boundary layers of a complex dynamical system. There is a predominance of large-scale structures in D'', and the strength of heterogeneities at intermediate and large-scales appears to be significantly greater than in the overlying lower mantle. Large low shear velocity provinces (LLSVPs), ultra-low velocity zones (ULVZs), multiple seismic discontinuities, a major phase change, and seismic anisotropy are all fundamental attributes of the D'' region. The present day configuration of D'' structures represents a "snap-shot" of an evolving system, with some aspects reflecting very long time-scale processes (LLSVPs may have been in place for at least hundreds of millions of years; ULVZs may be the last remnant of an extensive magma ocean that dates back to core formation) and much shorter time scales (post-perovskite lenses may be found in recently downwelled slab materials and seismic anisotropy may be sustained by present day dynamic shear flows). Together with the ongoing role of D'' as a thermal boundary layer that regulates cooling of the core and as the site of electromagnetic-mechanical coupling of rotation between the core and mantle, the significance of this region for Earth dynamics appears more evident than ever. While debate continues regarding the extent to which mantle material is fluxing between the shallow and deep mantle, evidence has grown to support a significant feed-back/control of D'' structures on geological processes at Earth's surface.

Cross-References

- [Core-Mantle Coupling](#)
- [Earth's Structure, Core](#)
- [Earth's Structure, Lower Mantle](#)
- [Mantle Convection](#)
- [Mantle Plumes](#)

Bibliography

- Dziewonski AM, Anderson DL (1981) Preliminary reference earth model. *Phys Earth Planet Inter* 25:297–356
- Garnero EJ, McNamara AK (2008) Structure and dynamics of Earth's lower mantle. *Science* 3230:626–628
- Hirose K, Lay T (2008) Discovery of post-perovskite and new views of the core-mantle boundary region. *Elements* 4:183–190
- Hirose K, Brodholt J, Lay T, Yuen DA (eds) (2007) *Post-Perovskite: the last mantle phase transition*. American Geophysical Union, Monograph, 174, 287 pp
- Hutko AR, Lay T, Revenaugh J, Garnero EJ (2008) Anticorrelated seismic velocity anomalies from post-perovskite in the lowermost mantle. *Science* 320:1070–1074

- Labrosse S, Hernlund JW, Coltice N (2007) A crystallizing dense magma ocean at the base of the Earth's mantle. *Nature* 450:866–869
- Lay T (2007) 1.18 deep earth structure – lower mantle and D''. In: Romanowicz B, Dziewonski A (eds) *Treatise on geophysics. Seismology and structure of the Earth*, vol 1. Elsevier, Amsterdam, pp 619–654
- Lay T (2008) Sharpness of the D'' discontinuity beneath the cocos plate: implications for the perovskite to post-perovskite phase transition. *Geophys Res Lett*:L03304. <https://doi.org/10.1029/2007GL032465>
- Lay T, Garnero EJ (2007) Reconciling the post-perovskite phase with seismological observations of lowermost mantle structure. In: Hirose K, Brodholt J, Lay T, Yuen D (eds) *Post-Perovskite: the last mantle phase transition*. American Geophysical Union, Monograph, pp 129–153
- Lay T, Hernlund J, Garnero EJ, Thorne MS (2006) A post-perovskite lens and D'' heat flux beneath the Central Pacific. *Science* 314:1272–1276
- Lay T, Hernlund J, Buffett BA (2008) Core-mantle boundary heat flow. *Nat Geosci* 1:25–32
- Maupin V, Garnero EJ, Lay T, Fouch MJ (2005) Azimuthal anisotropy in the D'' layer beneath the Caribbean. *J Geophys Res* 110. <https://doi.org/10.1029/2004JB003506>
- Murakami M, Hirose K, Kawamura K, Sata N, Ohishi Y (2004) Post-perovskite phase transition in MgSiO₃. *Science* 304:855–858
- Thorne MS, Garnero EJ (2004) Inferences on ultralow-velocity zone structure from a global analysis of SPdKS waves. *J Geophys Res* 109:B083201. <https://doi.org/10.1029/2004JB003010>
- Torsvik TH, Smethurst MA, Burke K, Steinberger B (2006) Large igneous provinces generated from the margins of the large low velocity provinces in the deep mantle. *Geophys J Int* 167:1447–1460
- Trampert J, Descamps F, Resovsky J, Yuen D (2004) Probabilistic tomography maps chemical heterogeneities throughout the lower mantle. *Science* 306:853–856
- Trønnes RG (2009) Structure, mineralogy and dynamics of the lowermost mantle. *Mineral Petrol*. <https://doi.org/10.1007/s00710-009-0068-z>
- van der Hilst RD, de Hoop MV, Wang P, Shim S-H, Ma P, Tenorio L (2007) Seismic-stratigraphy and thermal structure of Earth's core-mantle boundary region. *Science* 315:1813–1817
- Wang Y, Wen L (2006) Geometry and P- and S-velocity structures of the 'African Anomaly'. *J Geophys Res*. <https://doi.org/10.1029/2006JB004483>

Mantle Plume: Spreading Ridge Interactions

K. S. Krishna and M. Ismaiel
Centre for Earth, Ocean and Atmospheric Sciences,
University of Hyderabad, Hyderabad, India

Synonyms

Anomalous volcanic ridges and plateaus; Plate boundary tectonics; Ridge jump processes

Definition

Mantle plume	The mechanism of up-welling of buoyant mantle material through a quasi-cylindrical structure for the formation of large igneous
--------------	---

	provinces, linear aseismic ridges, plateaus and islands on surface of the Earth.
Hotspot	Place of anomalous volcanism compared with the surrounding mantle; the process melts the mantle rocks into magma. The anomalous hot places inside the mantle have expression of active volcanism on Earth's surface.
Spreading ridge system	An underwater continuous mountainous feature lies at divergent plate boundary, where new magmatic material upwells, leading to growth of the plates on either side.

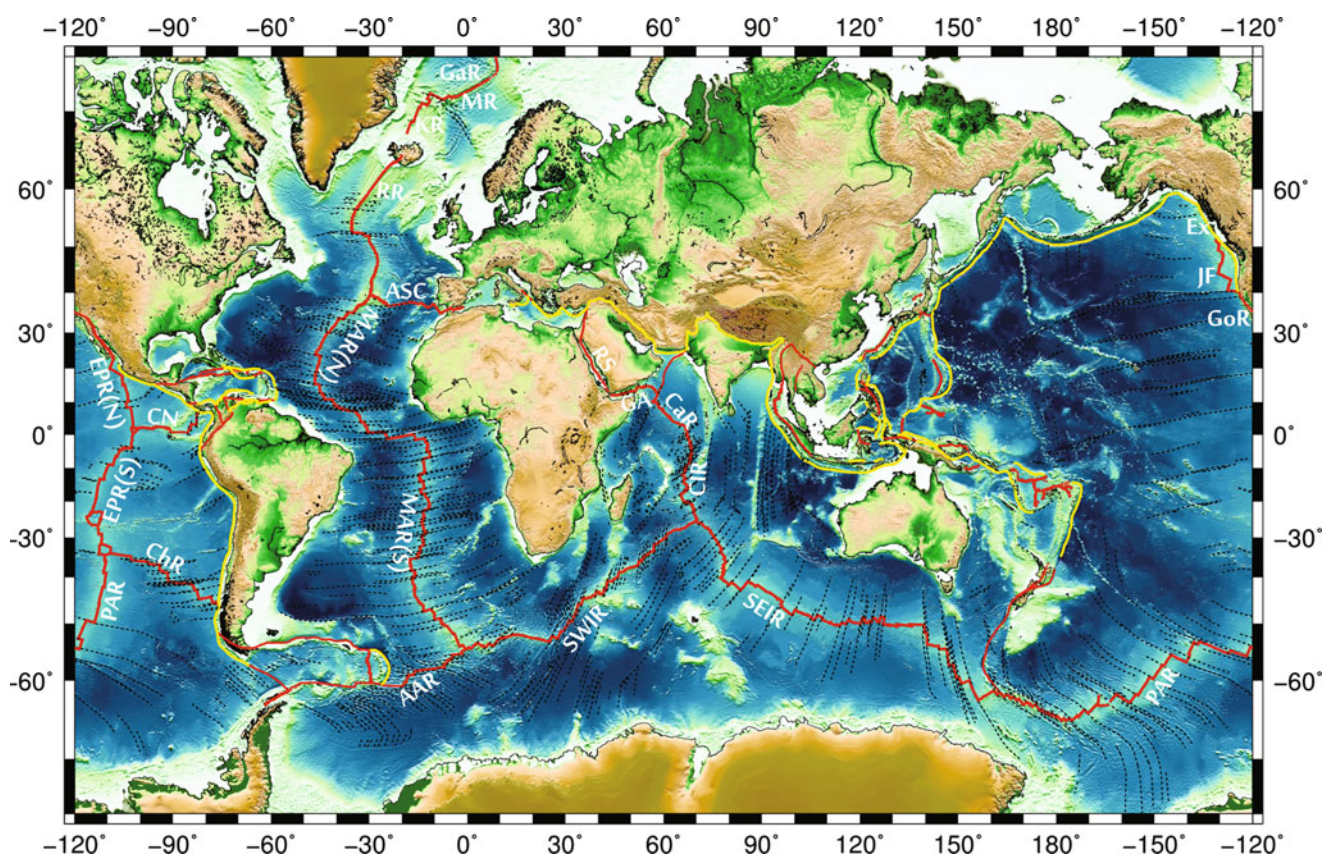
Spreading Ridge System

The rigid outer shell of the Earth is divided into a number of tectonic plates. Presently, the Earth consists of a total of 52 plates, in which 14 are large plates, viz., Africa, Antarctica, Arabia, Australia, Caribbean, Cocos, Eurasia, India, Juan de Fuca, Nazca, North America, Pacific, Philippine Sea, and

South America, and 38 are small plates (Bird 2003). The tectonic plates are bounded by one or more of the three main types of plate boundaries: (1) spreading ridges, (2) subduction zones/trenches, and (3) transform faults (Fig. 1). Plate boundaries are narrow deforming zones, characterized by significant seismicity, whereas the plate's interiors are relatively stable.

The tectonic plates are in continuous motion at a speed of up to a few centimeters per year. Each plate is in relative motion with respect to each other on the surface of the Earth. The relative motion produces new crust at spreading ridges and consumes old crust at subduction zones. Apart from these tectonic processes, plates do undergo break-ups and unifications through geologic time.

The spreading ridge, a divergent tectonic boundary between two lithospheric plates, connects with other spreading ridge system and encircles the world's oceans with a total length of about 63,000 km as a single global entity (Fig. 1). Details of all spreading ridge systems of the world oceans are placed in Table 1 for the purpose of keeping succinct



Mantle Plume: Spreading Ridge Interactions, Fig. 1 General physiographic map of the world oceans, depicting the spreading ridge systems along with major transform faults, fracture zones and subduction zones. The acronyms assigned for all ridge systems are detailed. AAR, American–Antarctic Ridge; ASC, Azores Spreading Centre; CaR, Carlsberg Ridge; ChR, Chile Rise; CIR, Central Indian Ridge; CN,

Cocos–Nazca Spreading Centre; EPR, East Pacific Rise (north and south); Ex, Explorer Ridge; GA, Gulf of Aden; GoR, Gorda Ridge; GaR, Gakkel Ridge; JF, Juan de Fuca Ridge; KR, Kolbeinsey Ridge; MAR, Mid-Atlantic Ridge (north and south); MR, Mohs Ridge; PAR, Pacific–Antarctic Ridge; RR, Reykjanes Ridge; RS, Red Sea; SEIR, Southeast Indian Ridge; SWIR, Southwest Indian Ridge

Mantle Plume: Spreading Ridge Interactions, Table 1 Details of global mid-ocean ridge systems

S. No.	Name of the spreading ridge	Name of the Ocean includes the Ridge	Total length (in km)	Adjoined tectonic plate boundaries	Present spreading rate (mm/year)	Timing of Ridge commencement (Ma)
1	Red Sea Rift	Red Sea	2,250	Africa and Arabian Plates	10	23
2	Aden-Sheba Ridge	Gulf of Aden	1,035	Soli and Arabian	15	35
3	Carlsberg Ridge	NW Indian Ocean	2,200	African and Indo-Australian plates	11–16	65
4	Central Indian Ridge	Indian Ocean	2,700	African and Indo-Australian plates	18–24	68
5	Southeast Indian Ridge	South Indian Ocean	6,000	Indo-Australian and Antarctic plates	24–40	42
6	Southwest Indian Ridge	SW Indian Ocean	7,700	African and Antarctic plates	15	165
7	Mid-Atlantic Ridge	Atlantic Ocean	16,000	Eurasia and North America, African and South America	25	150
8	Knipovich Ridge	NE- Atlantic Ocean	550	North America and Eurasia	15–17	28
9	Reykjanes Ridge	North Atlantic Ocean	900	North America and Eurasia	20	35
10	Kolbeinsey Ridge	Arctic Ocean	650	North America and Eurasia	20	26
11	Mohns Ridge	Arctic Ocean	550	North America and Eurasia	18	50
12	Gakkel Ridge	Arctic Ocean	1,800	North America and Eurasia	6	25
13	South American-Antarctic Ridge	South Atlantic Ocean	2,500	South America and Antarctic	10–15	105
14	East Scotia Ridge	South Atlantic ocean	500	Scotia and sandwich	60–90	56
15	Pacific-Antarctic Ridge	South Pacific Ocean	6,000	Pacific and Antarctic plates	45	75
16	Chile Rise	South Pacific Ocean	2,000	Nazca and Antarctic	20–30	23
17	East Pacific Rise	Pacific Ocean	8,700	Pacific and North American, Cocos, Nazca	200	180
18	Juan de Fuca Ridge	Pacific Ocean	500	Pacific and Juan de fuca	30	30
19	Explorer Ridge	North Pacific Ocean	125	Pacific and Explorer	20–30	5–7
20	Gorda Ridge	North Pacific Ocean	300	Pacific and Gorda	12–29	30

Total length of global mid-oceanic ridge systems ~ 63,000 km

information at one place. Many of the major oceanic ridge systems have been reorganized several times in the geological past that led to unification or split of the tectonic plates. One of the well-mapped ancient oceanic ridges in the Indian Ocean is the Wharton Ridge, once separated the Indian and Australian plates, and the ridge cessation at about 42 million years ago led to unification of both the aforesaid plates into a single Indo-Australian plate (Liu et al. 1983; Krishna et al. 1995, 2012).

Mantle Plume

Mantle plume is a quasi-cylindrical pipe-like structure, through which concentrated hot mantle material up-wells to the surface of the Earth. The classical hotspot model proposed by Wilson (1963) and Morgan (1971, 1972) considers the presence of magma source deep in the mantle,

therefore the chemical signatures of rising magma are distinct from those of spreading ridges. Numerous experiments have demonstrated that the plumes initiate with a leading plume head followed by a narrow conduit or plume tail connecting the plume head to the source region (Turcotte and Schubert 2002). For example, the Reunion hotspot in Indian Ocean region emplaced the Deccan flood basalts (in the form of Large Igneous Province) on western Indian Shield as a plume head (White and McKenzie 1989), Chagos-Laccadive Ridge as a plume tail, and the Reunion Island as a plume source location (Fig. 2). Initial plume volcanism is generally much more voluminous than the following volcanism in oceanic regions. In many cases, hotspots lie at the end of well-defined linear volcanic ridges, known as hotspot tracks. For example, the hotspot track associated with the Hawaiian hotspot volcanism is the Hawaiian – Emperor island – seamount chain that runs from north part of the Pacific plate to the Aleutian Islands.

Mantle Plume: Spreading Ridge Interactions, Table 2 Details of global hotspots

S. No.	Name of the hotspot	Tectonic plate	Current position of the hotspot		LIPs, Flood Basalts, if any	Linear aseismic ridge	Plateaus, if any	Timing of hotspot commencement	Interaction with ridge-systems
			Latitude	Longitude					
1	Balleny	AN	67°S	163°E	No	No	No	20 Ma	Pacific-Antarctic Ridge
2	Crozet	AN	45°S	50°E	–	Yes?	No	183 Ma	India-Antarctic Ridge Southwest Indian Ridge
3	Kerguelen	AN	49°S	69°E	Rajmahal-Sylhet, Bunbary, SKP	Ninetyeast Ridge	Kerguelen	130 Ma	Wharton Ridge and Southeast Indian Ridge
4	Marion	AN	47°S	38°E	–	Yes	Madagascar?	88 Ma	Southwest Indian Ridge
5	Eastern Australia	AU	38°S	143°E	No	Yes	No	–	No
6	Lord Howe	AU	33°S	159°E	No	Yes?	No	–	No
7	Tasmantid	AU	39°S	156°E	No	Tasmantid seamount chain	Louisiade	–	No
8	Afar	AF	10°N	43°E	Ethiopia	No	No	30 Ma	Afar Triple Junction
9	Cameroon	AF	4°N	9°E	No	Yes?	No	–	No
10	Canary	AF	28°N	20°W	No	No	Canarian Island	–	No
11	Capre Verde	AF	14°N	20°W	No	No	No	–	No
12	Comores	AF	12°S	43°E	No	No	No	–	No
13	Darfur	AF	13.0	24°E	No	Yes?	No	–	No
14	Discovery	AF	42°S	0°	No?	No	No	–	No
15	Great Meteor	AF	28°N	32°W	No?	New England		Late Cenozoic	No
16	Hogger	AF	23°N	6°E	No	No	No	–	No
17	Meteor	AF	52°S	1°E	No	Yes	No	–	
18	Réunion	AF	21°S	56°E	Deccan	Chagos-Maldives-Laccadive Ridge	Mascarene	65	Central Indian Ridge
19	St. Helena	AF	17°S	20°W	No	St. Helena-Cameroon Volcanic line	No	145 Ma	Mid-Atlantic Ridge
20	Tibesti	AF	21°N	17°E	No	Yes?	No	–	No
21	Tristan da Cunha	AF	37°S	12°W	Parana	Walvis Ridge	No	133 Ma	Mid-Atlantic Ridge
22	Vema	AF	33°S	4°E	Yes?	Yes?	Orange R.	15 Ma	No
23	Azores	EU	39°N	28°W	No	No?	Azores Plateau, Azores Archipelago	5.5 Ma	Mid-Atlantic Ridge
24	Eifel	EU	50°N	7°E	Eifel Mountain	Yes?	No	–	No
25	Iceland	EU	65°N	20°W	Yes	Yes?	Greenland	61 Ma	Mid-Atlantic Ridge
26	Bermuda	NA	33°N	67°W	No?	No	No?	33 Ma	No

(continued)

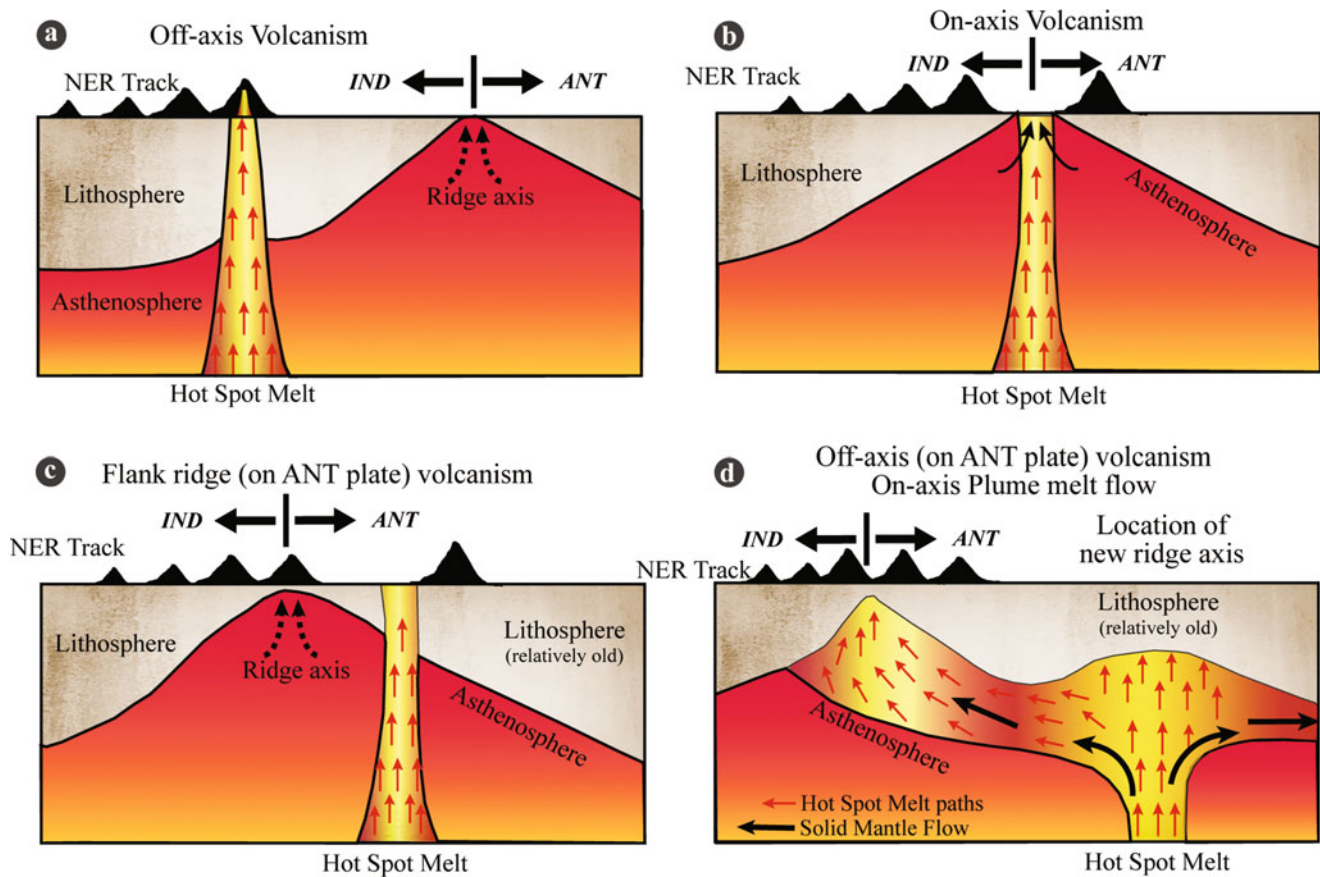
Mantle Plume: Spreading Ridge Interactions, Table 2 (continued)

S. No.	Name of the hotspot	Tectonic plate	Current position of the hotspot		LIPs, Flood Basalts, if any	Linear aseismic ridge	Plateaus, if any	Timing of hotspot commencement	Interaction with ridge-systems
			Latitude	Longitude					
27	Jan Mayen	NA	71°N	8°W	Yes?	No?	No	—	Mid-Atlantic Ridge
28	Raton	NA	37°N	104°W	No	Yes?	Colorado	—	No
29	Yellowstone	NA	44°N	111°W	Ethiopia-Yemen	Yes?	Columbia?	16 Ma	No
30	Easter	NZ	27°S	100°W	Yes	Tuamotu-Line Island chain	mid-Pac mnt?	100 Ma?	East Pacific Rise
31	Galapagos	NZ	0°	92°W	Caribbean?	Yes?	Galapagos island	90 Ma	Nazaca-Cocos Plate
32	Juan Fernández	NZ	34°S	83°W	Yes?	Juan Fernández Ridge	No	—	No
33	San Felix	NZ	26°S	80°W	No	Yes?	San Felix	—	No
34	Ascension	SA	8°S	14°W	No	No	No	—	Mid Atlantic Ridge
35	Bouvet	SA	54°S	2°E	No	No	No	—	Southwest Indian Ridge
36	Fernando do Noronha	SA	4°S	32°W	Camp?	Yes?	No	201 Ma?	Wlavis Ridge
37	Martin Vaz	SA	20°S	29°W	No	Yes?	Trindade Island	—	MOR?
38	Baja/Guadalupe	PA	27°N	113°W	No	Yes?	No	—	East Pacific Rise
39	Bowie	PA	53°N	135°W	No	Kodiak-Bowie	Queen Charlotte	1 Ma	Juan de Fuca Ridge
40	Caroline	PA	5°N	164°E	No	Yes	No	—	No
41	Cobb/Juan de Fuca	PA	46°N	128°W	No	Cobb-Eickelberg	No	—	Juan De Fuca Ridge
42	Hawaii	PA	20°N	156°W	Yes	Hawaiian-Emperor seamount chain	Subducted?	85 Ma	No
43	Louisville	PA	51°S	141°W	No	Louisville Ridge	No	122 Ma	Pacific-Antarctic Ridge
44	Macdonald	PA	30°S	140°W	Yes?	Austral-Gilbert-Marshall chain			No
45	Marquesas	PA	10°S	138°W	Marquesas	Yes	Shatski	—	No
46	Pitcairn	PA	26°S	130°W	No	Yes	Pitcairn Island	—	East Pacific Rise
47	Samoa	PA	14°S	170°W	No?	Yes	Samoa Island	14 Ma?	No
48	Socorro	PA	19°N	111°W	No	No	No	—	No
49	Society/Tahiti	PA	18°S	150°W	No	Yes		—	No

Plate names: AF, African Plate; AN, Antarctic Plate; AU, Australian Plate; EU, Eurasian Plate; NA, North American Plate; NZ, Nazca Plate; PA, Pacific Plate; SA, South American Plate; SKP, Southern Kerguelen Plateau; Hotspots positions are considered from Courtillot et al. (2003)

et al. 2002). Further, the Indian Ocean ridge systems had experienced absolute motions because of specific tectonic setting of the Antarctic plate circuit (Krishna et al. 2012).

During the volcanic emplacement, the Kerguelen mantle plume had interacted in various settings with the spreading segments of the Wharton Ridge system.

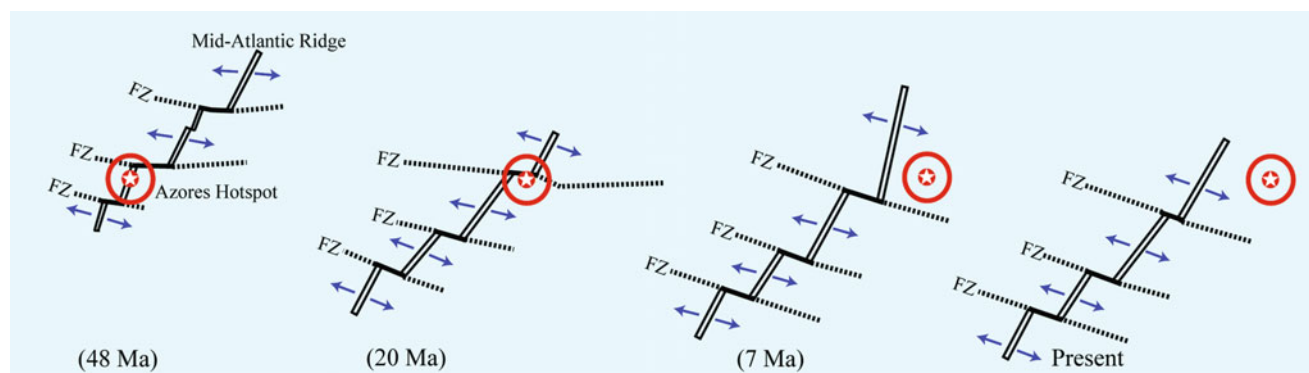


Mantle Plume: Spreading Ridge Interactions, Fig. 3 Sketch diagrams showing various dynamic settings held between the Kerguelen mantle plume and Wharton spreading centers. (After Krishna et al. 2012)

We discuss various models of interactive processes held between the Wharton spreading center and Kerguelen hotspot during the formation of the NER (Krishna et al. 1995, 1999, 2012). Initially, the hotspot was placed beneath the Indian plate, which is now located in the Bay of Bengal region, possibly at some distance from the spreading center, resulting in intraplate volcanism on the Indian plate, forming the volcanic trace approximately 1500 km distance (Fig. 3a). This model is supported by the presence of regional flexure beneath the northern part of the NER (Gopala Rao et al. 1997; Krishna et al. 2009; Sreejith and Krishna 2013). Subsequently, northward migration (absolute motion) of the Wharton spreading ridge system allowed its spreading center to coincide with the hotspot for on-axis hotspot volcanism (Fig. 3b), which may have emplaced sections of the NER simultaneously on both plates. Further northward spreading center migration placed the hotspot beneath the Antarctic plate for near-ridge or ridge-flank volcanism (Fig. 3c). Evidence of Kerguelen hotspot volcanism on Antarctic plate during the formation of the NER comes from the presence of 68 Ma ages basalts from Skiff Bank, part of the Kerguelen Plateau (Coffin et al. 2002).

With continued Wharton spreading ridge migration, the hotspot moved farther from the spreading ridge beneath the Antarctic plate. As the overlying crust was rigid, the hotspot material flowed along the base of the lithosphere to reach the spreading center for on-axis volcanism (Fig. 3d), even when the hotspot was not located in the vicinity of the spreading ridge crest. The stress field created by the hotspot may have caused uplift and weakening of the overlying oceanic lithosphere, resulting in the initiation of rifting and spreading within the Antarctic plate and leading to southward ridge jumps. The continuous process of hotspot-spreading ridge interactions may have provided necessary conditions to cause multiple ridge jumps. Observations of ridge-hotspot interactions at other locations led several researchers (Brozena and White 1990; Small 1995; Hardarson et al. 1997; Mittelstaedt et al. 2011) to conclude that ridge segments jump repeatedly toward a hotspot, even when the overall motion of the ridge may be away from the hotspot.

The Azores hotspot represented by a 200-km broad thermal anomaly and its interaction with the mid-Atlantic ridge system for four different phases are discussed (Fig. 4). The gradual migration of ridge system together with the African



Mantle Plume: Spreading Ridge Interactions, Fig. 4 The proximity of Azores hotspot and its interaction with the mid-Atlantic ridge system for the ages 48 Ma, 20 Ma, 7 Ma, and Present are shown. The

hotspot shown with a red color circle is represented by a 200-km broad thermal anomaly (McNutt et al. 1989; Wolfe et al. 1997). (The Figure is adopted from Gente et al. (2003))

plate at a rate of 13 km/Myr in SSE direction resulted in northward shift of interactive area of two melt zones during the last 48 Myr (Gente et al. 2003). The coupled melt zones, even the ridge axis is 300 km away from the hotspot location, have produced abundant volcanism, and led to emplace twice thicker oceanic crust than normal crust. The rock samples collected from shallow seafloor areas of melt zones display flat to enriched light rare earth elements signature reflecting heterogeneous mantle (Schilling et al. 1977), and well supports the ridge-hotspot interaction along the mid-Atlantic ridge system.

Crustal Structure Beneath the Spreading Ridge, Volcanic Ridge and on-Spreading-Axis Hotspot Setting

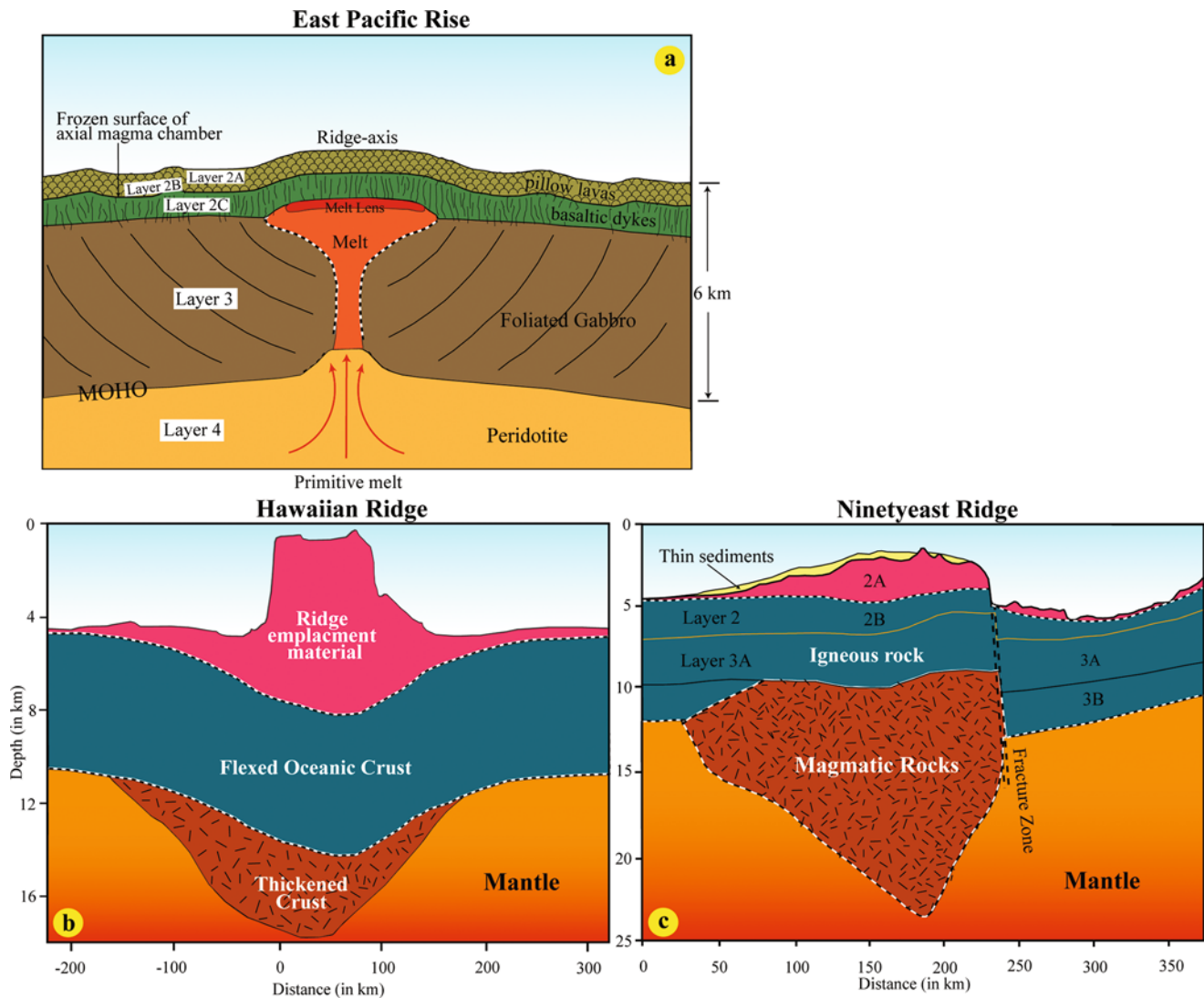
The geometry of crustal structures beneath the spreading ridge system and volcanic chain when the hotspot volcanism was in intraplate or interplate positions varies significantly. An outline of seismic structure beneath the East Pacific Rise is developed following the seismic reflection results across the Rise (Detrick et al. 1987). The structure consists of layer 2 with pillow lavas and dykes, layer 3 with gabbro cumulates, axial magma chamber with low-velocity zone on its roof beneath the ridge axis, and discontinued Moho boundary beneath the magma chamber (Fig. 5a). This is a typical structure generally found beneath the spreading ridge systems of the planet Earth. Under some spreading ridge segments, the axial magma chambers are either small or absent altogether. For example, multichannel reflection experiments carried out on the Mid-Atlantic Ridge (Detrick et al. 1990) did not find evidences for presence of such intracrustal reflections associated with an axial magma chamber.

The crustal structure of the Hawaiian Islands in the central Pacific Ocean is most credible example for a

volcanic chain built by a mantle plume activity at some distance from the spreading center (Watts and ten Brink 1989). The oceanic crust with a typical thickness of about 6 km is flexed downward with addition of extra material to the top and base of the flexed crust (Fig. 5b). Watts and ten Brink (1989) opined that the extra materials seem to have been added to the flexed crust during or following the late stages of volcanic activity. The structural model suggests that the surface loading has produced the flexure to the oceanic crust, while the deep magmatic body serves as an upward subsurface load reduced the flexure to some degree. Similar kind of large crustal thickness models are also found beneath the Ori and Tamu Massifs of the Shatsky Rise oceanic plateau, northwest Pacific Ocean (Huang et al. 2018; Sager et al. 2019), but they were evolved by massive eruptions at mid-ocean ridge with extra ordinary mantle melting, which are different from the process of plume – spreading ridge interactions.

The structural model of the Ninetyeast Ridge in the eastern Indian Ocean derived from seismic and gravity results presents an explicit geological process that underwent when the Kerguelen plume was situated right beneath the spreading center of the Wharton Ridge system (Krishna et al. 2001). The impact of surface and subsurface loads on juvenile oceanic crust significantly differs from the one associated with the off-spreading-axis volcanism (Fig. 5b, c). The flexure of the oceanic crust due to surface load is very minimal, whereas the subsurface load engulfs the lower part of the crust leading to the presence of large volume of magmatic rocks beneath the Ninetyeast Ridge.

The Iceland hotspot with ridge-centered plume in North Atlantic Ocean is a classical live example on the Earth for plume-spreading ridge interactions. The crustal structure exactly beneath the location of the Iceland hotspot, where plume-ridge interactions are continuing, shows the thickness



Mantle Plume: Spreading Ridge Interactions, Fig. 5 Crustal structure beneath the spreading ridge (East Pacific Rise) (after Detrick et al. 1987), aseismic ridge (Hawaiian-Emperor seamount chain) (after Watts

and ten Brink 1989) and intersected location of mantle plume – spreading ridge (Ninetyeast Ridge). (After Krishna et al. 2001)

of 25–30 km crust with extrusive formations in upper crust and intrusive formations in the lower crust (Kumar et al. 2007; Bjarnason 2008). The Moho boundary beneath the present location of the Iceland hotspot found to be normal and sharper (Kumar et al. 2007).

Summary

The spreading ridge, an underwater mountainous feature, extends for a length of about 63,000 km as a single continuous entity in the global oceans. It is a window to the Earth's interior, through which magma continuously up-wells, and drives the tectonic plates on either side in order to accommodate newly accreted crust. Another important dynamic

process the planet Earth experiences is mantle plume volcanism. The mechanism up-wells the deep hot mantle material to surface of the Earth for formation of large igneous provinces, linear aseismic ridges, plateaus and islands. The planet Earth presently possesses about 49 hotspots and their exterior emplacements occupy an area of 11–42% of the total Earth's surface. During the ocean expansions, the hotspots sometime or other have come close to the vicinity of spreading segments of the ridge systems due to hotspot migrations and/or absolute plate motions, leading to interactive processes between them. The mechanism of formation of internal materials and geophysical and geochemical signatures of the volcanic chains greatly vary depending upon the proximity of hotspot with spreading centers of the ridge system.

Bibliography

- Bird P (2003) An updated digital model of plate boundaries. *Geochem Geophys Geosyst* 4(3):1027. <https://doi.org/10.1029/2001GC000252>
- Bjarnason IP (2008) An Iceland hotspot saga. *Jökull* 58:3–16
- Bougault H, Joron JL, Treuil M, Maury R (1985) Local versus regional mantle heterogeneities: evidence from hygromagmaphile elements. *Initial Rep Deep Sea Drill Proj* 82:459–482
- Brozena JM, White RS (1990) Ridge jumps and propagations in the South Atlantic Ocean. *Nature* 348:149–152
- Coffin MF, Eldholm O (1992) Volcanism and continental break-up: a global compilation of large igneous provinces. In: Storey BC, Alabaster T, Pankhurst RJ (eds) *Magmatism and the causes of continental break-up*, Geological society special publication, vol 68. Geological Society of London, London, pp 21–34
- Coffin MF, Pringle MS, Duncan RA, Gladchenko TP, Storey M, Muller RD, Gahagan LA (2002) Kerguelen hotspot magma output since 130 Ma. *J Petrol* 43:1121–1137
- Courtillot V, Davaille A, Besse J, Stock J (2003) Three distinct types of hotspots in the Earth's mantle. *Earth Planet Sci Lett* 205:295–308
- Detrick RS, Buhl P, Vera E, Mutter J, Orcutt J, Madsen J, Brocher T (1987) Multi-channel seismic imaging of a crustal magma chamber along the East Pacific Rise. *Nature* 326:35–41
- Detrick RS, Mutter JC, Buhl P, Kim II (1990) No evidence from multi-channel reflection data for a crustal magma chamber in the MARK area on the Mid-Atlantic Ridge. *Nature* 347:61–64
- Gente P, Dymant J, Maia M, Goslin J (2003) Interaction between the Mid-Atlantic Ridge and the Azores hot spot during the last 85 Myr: emplacement and rifting of the hot spot-derived plateaus. *Geochem Geophys Geosyst* 4:8514
- Gopala Rao D, Krishna KS, Sar D (1997) Crustal evolution and sedimentation history of the Bay of Bengal since the Cretaceous. *J Geophys Res Solid Earth* 102:17747–17768
- Hardarson BS, Fitton JG, Ellam RM, Pringle MS (1997) Rift relocation – a geochemical and geochronological investigation of a paleo-rift in Northwest Iceland. *Earth Planet Sci Lett* 153:181–196
- Huang Y, Sager WW, Tominaga M, Greene JA, Zhang J, Nakanishi M (2018) Magnetic anomaly map of Ori Massif and its implications for oceanic plateau formation. *Earth Planet Sci Lett* 501:46–55
- Krishna KS, Gopala Rao D, Ramana MV, Subrahmanyam V, Sarma KVLNS, Pilipenko AI, Shcherbakov VS, Radhakrishna Murthy IV (1995) Tectonic model for the evolution of oceanic crust in the northeastern Indian Ocean from the Late Cretaceous to the Early Tertiary. *J Geophys Res Solid Earth* 100:20011–20024
- Krishna KS, Gopala Rao D, Subba Raju LV, Chaubey AK, Shcherbakov VS, Pilipenko AI, Radhakrishna Murthy IV (1999) Paleocene on-spreading-axis hotspot volcanism along the Ninetyeast Ridge: an interaction between the Kerguelen hotspot and the Wharton spreading center. *Proc Indian Acad Sci-Earth Planet Sci* 108:255–267
- Krishna KS, Neprochnov YP, Gopala Rao D, Grinko BN (2001) Crustal structure and tectonics of the Ninetyeast Ridge from seismic and gravity studies. *Tectonics* 20(3):416–433
- Krishna KS, Michael L, Bhattacharyya R, Majumdar TJ (2009) Geoid and gravity anomaly data of conjugate regions of Bay of Bengal and Enderby Basin: new constraints on breakup and early spreading history between India and Antarctica. *J Geophys Res Solid Earth* 114:B03102. <https://doi.org/10.1029/2008JB005808>
- Krishna KS, Abraham H, Sager WW, Pringle MS, Frey F, Gopala Rao D, Levchenko OV (2012) Tectonics of the Ninetyeast Ridge derived from spreading records in adjacent oceanic basins and age constraints of the ridge. *J Geophys Res Solid Earth* 117(B4):1–19
- Kumar P, Kind R, Priestley K, Dahl-Jensen T (2007) Crustal structure of Iceland and Greenland from receiver function studies. *J Geophys Res Solid Earth* 112:B03301. <https://doi.org/10.1029/2005JB003991>
- Liu CS, Curran JR, McDonald JM (1983) New constraints on the tectonic evolution of the Eastern Indian Ocean. *Earth Planet Sci Lett* 65:331–342
- McNutt M, Fischer K, Kruse S, Natland J (1989) The origin of the Marquesas fracture zone ridge and its implications for the nature of hot spots. *Earth Planet Sci Lett* 91:381–393
- Mittelstaedt E, Soule S, Harpp K, Fornari D, McKee C, Tivey M, Geist D, Kurz MD, Sinton C, Mello C (2011) Multiple expressions of plume-ridge interaction in the Galápagos: volcanic lineaments and ridge jumps. *J Geophys Res Solid Earth* 116(13):1–31
- Morgan WJ (1971) Convection plumes in the lower mantle. *Nature* 230:42–43
- Morgan WJ (1972) Plate motions and deep mantle convection. *Geol Soc Am Mem* 132:7–22
- Müller RD, Gaina C, Roest WR, Hansen DL (2001) A recipe for microcontinent formation. *Geology* 29:203–206
- Sager WW, Huang Y, Tominaga M, Greene JA, Nakanishi M, Zhang J (2019) Oceanic plateau formation by seafloor spreading implied by Tamu Massif magnetic anomalies. *Nat Geosci* 12:661–666. <https://doi.org/10.1038/s41561-019-0390-y>
- Schilling JG, Kingsley R, Bergeron M (1977) Rare earth abundances in DSDP sites 332, 334 and 335, and inferences on the Azores mantle blob activity with time. *Initial Rep Deep Sea Drill Proj* 37:591–598. Washington
- Small C (1995) Observations of ridge-hotspot interactions in the Southern Ocean. *J Geophys Res* 100:17931–17946
- Sreejith KM, Krishna KS (2013) Spatial variations in isostatic compensation mechanisms of the Ninetyeast Ridge and their tectonic significance. *J Geophys Res Solid Earth* 118(10):5165–5184
- Talwani M, Desa MA, Ismaiel M, Krishna KS (2016) The tectonic origin of the Bay of Bengal and Bangladesh. *J Geophys Res Solid Earth* 121:4836–4851
- Turcotte DL, Schubert G (2002) *Geodynamics*, 2nd edn. Cambridge University Press, Cambridge, pp 1–456
- Watts AB, ten Brink US (1989) Crustal structure, flexure, and subsidence history of the Hawaiian Islands. *J Geophys Res Solid Earth* 94(B8):10473–10500
- White RS, McKenzie DP (1989) Magmatism at rift zones; the generation of volcanic continental margins and flood basalts. *J Geophys Res Solid Earth* 94:7685–7729
- Wilson JT (1963) A possible origin of the Hawaiian Islands. *Can J Phys* 41(6):863–870
- Wolfé CJ, Bjarnson IT, VanDecar JC, Solomon SC (1997) Seismic structure of the Iceland mantle plume. *Nature* 385:245–247

Mantle Plumes

Cinzia G. Farnetani¹ and Albrecht W. Hofmann^{2,3}

¹Institut de Physique du Globe de Paris, Paris, France

²Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA

³Max-Planck-Institut für Chemie, Mainz, Germany

Definition

Mantle plumes Localized upwelling currents of solid rock that are hotter, thus less dense, than the surrounding mantle.

Hotspots

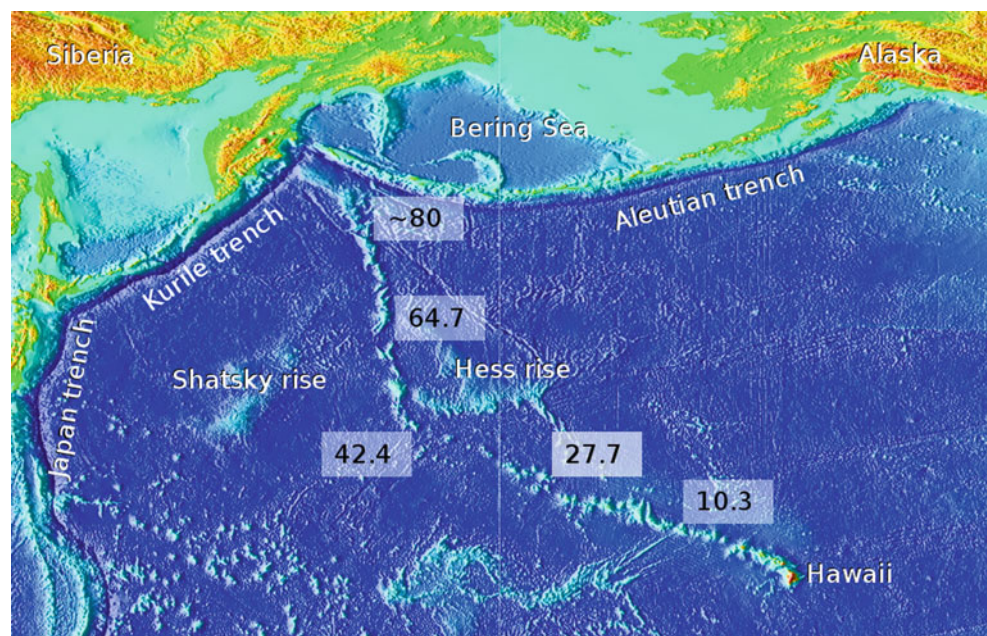
Plate tectonics provides a framework for interpreting volcanism at plate boundaries, namely, along spreading ridges (divergent boundaries) and subduction zones (convergent boundaries). However, it does not explain intraplate volcanism, such as Hawaii, nor the excess volcanism along certain sectors of a spreading ridge, as observed in Iceland. In these regions, called hot spots, the volcanic activity can last more than a hundred million years, as indicated by the resulting age-progressive volcanic chain formed during the hotspot's lifetime. For example, the 6,000 km-long Hawaiian-Emperor volcanic chain (Fig. 1) is an alignment of roughly 100 volcanoes of progressively increasing ages. The volcanic chain was created during the last 80 Ma, as the Pacific plate drifted over the Hawaiian hot spot. The vigorous and long-lasting Hawaiian volcanism readily captured the interest of geophysicists: Tuzo Wilson (1963) proposed that the Hawaiian chain was not caused by lithospheric fissures, but by convection currents in the mantle, and Jason Morgan (1971) suggested that hotspots are the surface expression of mantle plumes upwelling from the Earth's lowermost mantle. Mantle plumes can be defined as localized upwelling currents of solid rocks that are hotter, and thus less dense, than the surrounding mantle.

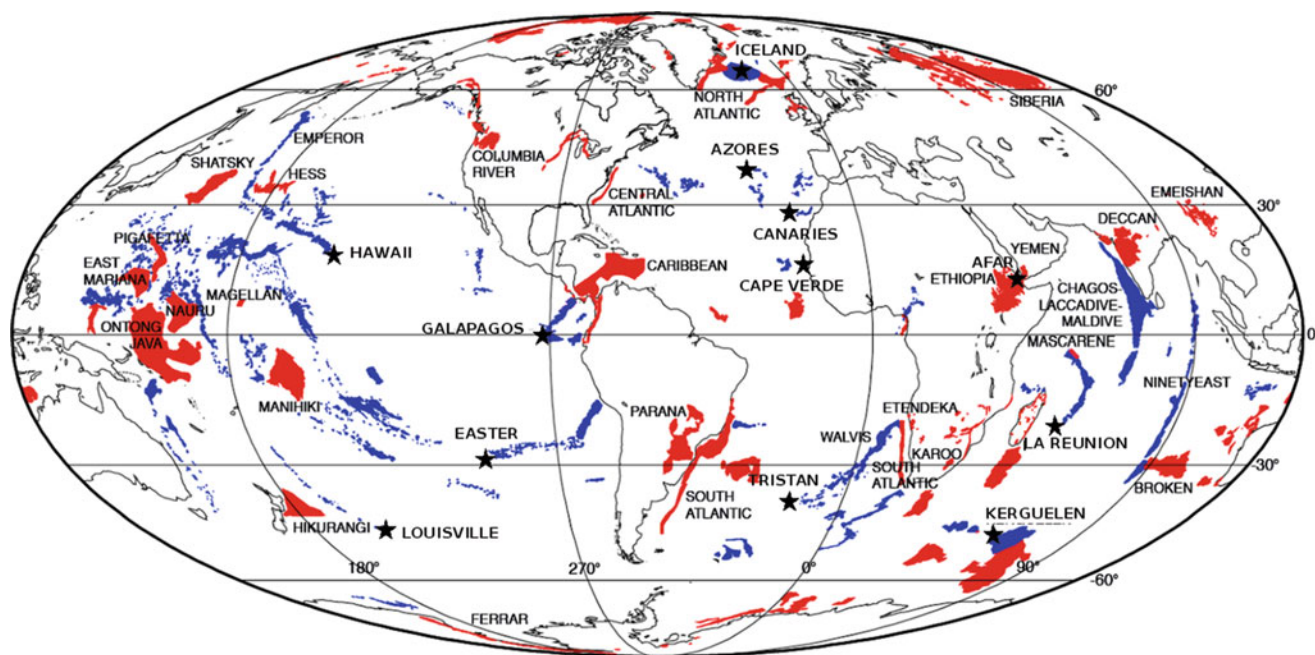
Hotspots are traditionally characterized by some or all of the following features, although exceptions do occur: (1) An age-progressive volcanic chain whose linear trend is consistent with the direction of plate motion. (2) The onset of hotspot magmatism is often marked by a Large Igneous

Province (LIP), a term including “continental flood basalts” (CFB) and “oceanic plateaus.” Estimated extruded volumes of CFBs are $1\text{--}2 \times 10^6 \text{ km}^3$, with perhaps similar (but unknown) volumes of intrusive magmatism, whereas oceanic plateaus can be one order of magnitude more voluminous (Coffin and Eldholm 1994). According to the mantle plume initiation model (Richards et al. 1989), the transient and episodic LIP magmatism corresponds to melting of a large plume “head,” whereas the subsequent hotspot activity is associated with the long-lasting and narrow plume “tail.” (3) The topographic swell is a region of anomalously high topography with a lateral width of $\sim 1,000 \text{ km}$ in the direction normal to the volcanic chain (Wessel 1993) and with an elevation of $\sim 1 \text{ km}$, which decreases along the chain. (4) Hotspot basalts are geochemically distinct and more diverse than mid-ocean ridge basalts.

Some hotspots, however, do not show all of the above features. For example, many volcanic chains lack a clear age progression and, according to Ito and van Keken (2007), there are only 13 long-lived ($>50 \text{ Ma}$) and 8 short-lived ($<20 \text{ Ma}$) age-progressive volcanic chains. Some hotspots are not associated with a LIP, whereas others such as the Hawaii and Louisville chains terminate in a subduction zone, so that the complete time record of the volcanic activity is lost. A number of LIPs do not have any volcanic track (e.g., the Shatsky rise and the Hess rise in the Pacific Ocean) and their origin is still a matter of debate. In summary, plume magmatism is often, but not always, explained by the classical thermal plume model with a voluminous spherical head followed by a narrow columnar conduit.

Mantle Plumes, Fig. 1 Seafloor topography from Smith and Sandwell (1997). Numbers indicate some ages (in Ma) along the Hawaiian-Emperor volcanic chain





Mantle Plumes, Fig. 2 Distribution of hotspots (stars) and Phanerozoic LIPs. In red: LIPs (or portions thereof) generated by a transient “plume head.” In blue: LIPs (or portions thereof) generated by a persistent “plume tail.” (Modified from Coffin and Eldholm 1994)

Global Hotspot Distribution and Hotspot Fixity

Over the years, the estimated number of hotspots has varied from 20 (Morgan 1971), to a maximum of 117 in the 1980s, whereas in recent compilations (e.g., Ito and van Keken 2007), the number ranges between 45 and 70 (see Schubert et al. 2001 and references therein). Hotspots younger than 100 Ma are generally active, although their vigor may vary considerably; older hotspots are either waning ($100 < \text{age} < 140$ Ma) or inactive ($\text{age} > 150$ Ma). The best-defined hotspots appear to be relatively stationary over time and are used as a reference frame to determine absolute plate motions. Most hotspots are situated in the oceans, one exception being Yellowstone (USA), whose volcanic activity can be traced back to the 16 Ma-old Columbia River flood basalt. The discrepancy in the number of continental vs. oceanic hotspots has three probable reasons: First, the arrival of a mantle plume can weaken the lithosphere and enhance continental breakup (Courtillot et al. 1999). For example, the magmatic activity of the Iceland plume started with the eruption of the North Atlantic Tertiary Igneous Province (62 Ma ago) and was followed by continental rifting and the appearance of an oceanic spreading ridge after a few million years. In this respect, mantle plumes may have been influential in modifying plate boundaries. The second reason concerns the different thickness (100–150 km) between oceanic and continental lithosphere: weak plumes may not melt beneath a thick continental lithosphere, but are likely to do so at lower pressure, beneath a thinner oceanic lithosphere. Third,

the lack of hotspots in continents could be due to the continents accumulating over downwellings, precisely where plumes are very unlikely to ascend.

The main oceanic hotspots and their related LIPs (Fig. 2) in the Atlantic Ocean are Iceland (0–62 Ma, LIP: the North Atlantic Tertiary Igneous Province) and Tristan da Cunha (0–125 Ma, LIP: the Paraná–Etendeka Province), whereas other hotspots like the Azores (0–20 Ma), Canaries (0–68 Ma), and Cape Verdes (0–Miocene) are not clearly associated with a LIP. In the Indian Ocean, the main hotspots are la Réunion (0–65 Ma, LIP: Deccan Traps), Kerguelen (0.1–120 Ma, LIP: the Kerguelen plateau), and Afar (0–30 Ma, LIP: Ethiopian Traps). In the Pacific Ocean, the main hotspots are Hawaii (0–76 Ma, unknown LIP, since the chain ends in a subduction zone), Louisville (1–77 Ma, possibly 120 Ma if the associated LIP is the Ontong–Java Plateau), and Galapagos (0 to possibly 90 Ma if the associated LIP is the Caribbean Plateau). There are also some hotspots (e.g., in French Polynesia) that may have a plume origin, but with an unclear age progression. Finally, the oceanic floors are “littered” by hundreds of thousands of volcanic seamounts that are certainly not created by plumes (e.g., Hieronymus and Bercovici 2000; Clouard and Gerbault 2008).

An interesting aspect of hotspots is that they are relatively stationary with respect to each other. Their respective motions of 1–2 cm/year are much less than the plate velocities, thereby approximating a fixed hotspot frame of reference. However, Tarduno et al. (2009) calculated the age and the paleolatitude of volcanoes belonging to Hawaiian–Emperor chain and

concluded that the Hawaiian hotspot moved southward at 4–5 cm/year during the period 80–47 Ma ago, whereas it remained relatively fixed (<2 cm/year) afterward. This complex behavior is probably due to the dynamical interaction between upwelling plumes and the “mantle wind” induced by large-scale mantle convection (Steinberger et al. 2004).

Evidence for Mantle Plumes

There is now widespread agreement on the existence of mantle plumes, although contrasting views do exist (see Foulger and Natland 2003 and the web site www.mantleplumes.org). Several lines of evidence support the existence of mantle plumes:

- First, the Earth’s Rayleigh number, which governs the vigor of convection, is sufficiently high (10^6 – 10^8) to insure that mantle convection is time dependent and that its thermal boundary layers (TBLs) become repeatedly unstable. ATBL is a zone characterized by a high-temperature gradient, since heat is transported dominantly by conduction. Fluid dynamical considerations indicate that Rayleigh–Taylor instabilities from a hot TBL generate thermal plumes (Loper and Stacey 1983). A prominent TBL in the Earth’s mantle is possibly the D’ zone, which extends 100–200 km above the core–mantle boundary. The existence of another TBL, for example, at 660 km depth, is still a matter of debate, but it seems unlikely, since the endothermic phase transition does not constitute a complete barrier to mantle convection. Mantle plumes are, therefore, expected to rise from the lowermost mantle, forming cylindrical conduits with a radius of 50–150 km. Such values are based on fluid dynamics, whereas conduit radius estimated by seismology (e.g., Montelli et al. 2004; Wolfe et al. 2009) are much broader.
- Second, seismic detection of narrow plume conduits is challenging, but the first exhaustive study by Montelli et al. (2004) found that at least six plumes (Ascension, Azores, Canary, Easter, Samoa, and Tahiti) extend into the lowermost mantle, whereas others are confined to the upper mantle, and in some cases the model resolution was insufficient. Recently, an extensive ocean-bottom seismological survey of Hawaii (Wolfe et al. 2009) has shown that a low seismic velocity anomaly extends into the lower mantle. Hopefully, future progress in seismic tomography will provide us with further evidence for mantle plumes.
- Third, Oceanic Island Basalts (OIBs) are geochemically distinct and more diverse than Mid-Ocean Ridge Basalts (MORBs), as reviewed by Hofmann (1997). Moreover, isotopic signatures of OIBs indicate an involvement of ancient recycled oceanic crust, as first suggested by

Hofmann and White (1982), whereas noble gases indicate that plumes may also carry primordial mantle material, possibly stored in the deep mantle.

- Fourth, Burke and Torsvik (2004) provide another set of evidence supporting the deep origin of mantle plumes. Their plate tectonic reconstruction over the last 200 Ma shows that the paleoposition of 90% of the 25 LIPs considered, were located, at the time of eruption, above lower mantle regions characterized, today, by low S-wave velocities. These broad regions, situated beneath the South-Central Pacific and Africa (see Romanowicz and Gung 2002 and references therein), are likely hotter and possibly compositionally denser than the surrounding mantle and may indeed represent a long-lived source zone of plumes.

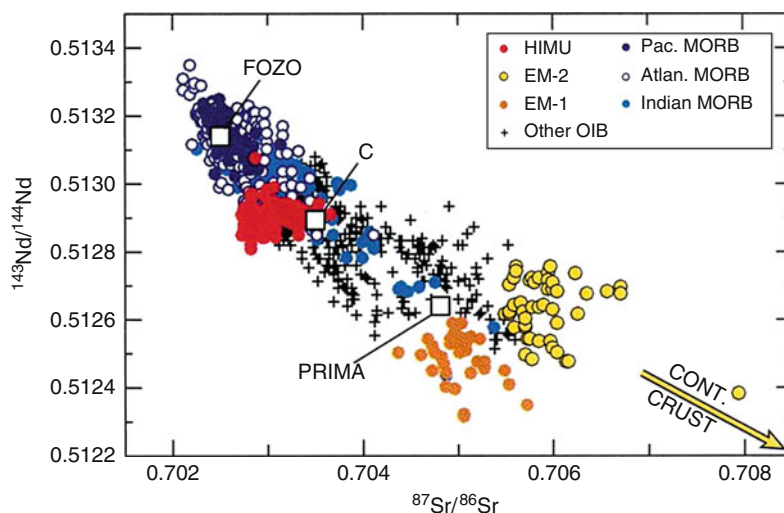
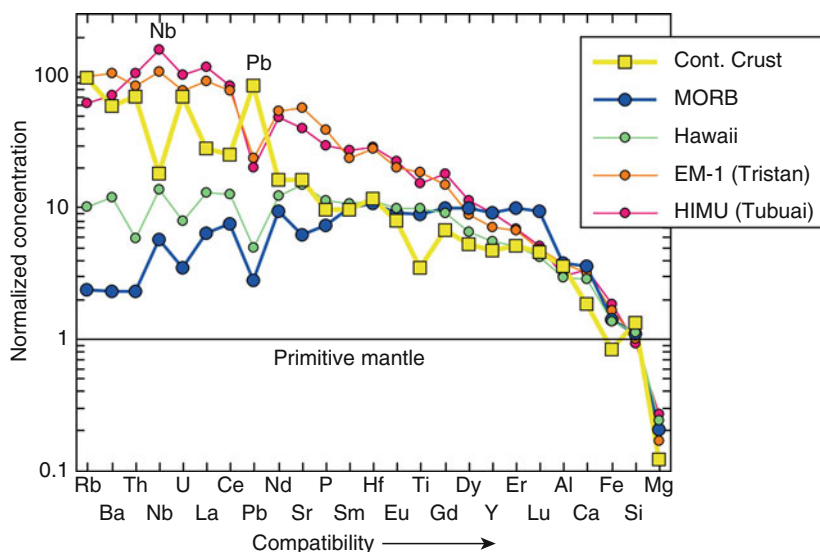
Geochemistry of Mantle Plumes

Geochemists became interested in mantle plumes when they discovered that ocean island basalts (OIBs), thought to be derived from plumes, tend to have different chemical (Schilling 1973) and isotopic (Hart et al. 1973) compositions from mid-ocean-ridge basalts (MORBs). These differences are consistent with the plume model if the deep mantle, the inferred source of plumes, is compositionally different from the upper mantle, source of MORBs. These systematic differences were borne out by many subsequent studies of samples from virtually all available ocean islands (Hofmann 2003), as illustrated by Fig. 3 for chemical compositions and Fig. 4 for isotopic compositions. Both the trace elements and the isotopic compositions indicate that the MORB source has been depleted in “mantle-incompatible” trace elements, that is, those that are scavenged from the mantle by melts that ultimately form the crust, leaving behind a “depleted” mantle. In contrast, plume-type basalts tend to be derived from less depleted mantle sources.

The above observations and conjectures led to the two-layer mantle model, in which “normal” melting producing MORBs and subduction-related magmas have extracted most of the incompatible elements from the upper mantle and sequestered them in the continental crust. By contrast, the lower mantle was thought to be largely undepleted or more “primitive.” Thus, plumes, rising from the deep mantle, would sample the lower, relatively primitive reservoir, but they might entrain more depleted mantle rocks on their way up, thus producing a mixing array between primitive and depleted reservoirs (Jacobsen and Wasserburg 1979). However, mass balance considerations for the incompatible element and isotope budget of the continental crust, the depleted upper mantle, and an undepleted lower mantle, demanded a size of the depleted reservoir of at least 50% of the mantle, which is significantly greater than the roughly 30% mass fraction of the upper mantle. This three-reservoir model

Mantle Plumes,

Fig. 3 Concentrations of selected trace and major elements, arranged in the order of ascending compatibility and normalized to primitive-mantle concentrations, for average continental crust, average MORB, average Mauna Loa, Hawaii, and OIB type representing EM-1 (Enriched mantle 1) and HIMU (high- μ , where $\mu = {}^{238}\text{U}/{}^{204}\text{Pb}$). The patterns for MORB and OIB differ by their enrichments, but show similar Nb and Pb anomalies, opposite to those of the continental crust. (From Hofmann 1997)



Mantle Plumes, Fig. 4 Nd and Sr isotopic compositions^(*) of MORBs and OIBs. EM-1 (Enriched mantle 1), EM-2 (Enriched mantle 2), and HIMU (high μ , where $\mu = {}^{238}\text{U}/{}^{204}\text{Pb}$). The squares indicate compositions of primitive mantle (PRIMA) and of mantle components FOZO and C (see Hofmann 1997 and references therein). ^(*)Some basic notions:

${}^{147}\text{Sm}$ decays to ${}^{143}\text{Nd}$. Since Sm is more compatible than Nd, progressive melting enriches the residual rock in Sm, thus MORBs have high ${}^{143}\text{Nd}/{}^{144}\text{Nd}$. ${}^{87}\text{Rb}$ decays to ${}^{87}\text{Sr}$. Since Rb is less compatible than Sr, progressive melting depletes the residual rock in Rb, thus MORBs have low ${}^{87}\text{Sr}/{}^{86}\text{Sr}$.

(often simply called the “layered mantle” model) was reinforced by the observation that many plume-derived basalts had much higher ${}^3\text{He}/{}^4\text{He}$ ratios than MORB (Farley and Neroda 1998), where ${}^3\text{He}$ is a remnant of primordial noble gases from the primitive Earth, whereas ${}^4\text{He}$ is the product of subsequent decay of uranium and thorium.

In spite of its apparent geochemical persuasiveness, the layered mantle model suffered two essentially fatal blows. During the 1980s, evidence began to accumulate that showed isotopically “enriched” (i.e., crustal-like) mantle sources in several ocean islands including Hawaii, Pitcairn, and Tristan da Cunha. This means that the endpoint of the mixing array of

the type shown in Fig. 4 cannot be an undepleted or “primitive” mantle reservoir but must represent a source component that was crust-like in that it was actually enriched in incompatible elements. And although such a mixing array may pass through the isotopic locus of a primitive mantle reservoir, it clearly does not require the involvement of a primitive reservoir. This is because, when a primitive reservoir is differentiated into enriched and depleted components, any remixing of these differentiated components will produce a compositional array passing approximately through the starting point, the locus of primitive compositions. Therefore, Hofmann and White (1982) proposed that the enriched source components

in mantle plumes actually come from subducted oceanic crust rather than a “primitive reservoir.”

The second blow to the conventional layered mantle model came from seismic evidence showing tomographic images of high seismic velocities characterizing subduction slabs that penetrate the base of the upper mantle and can be traced into the lowermost mantle (e.g., van der Hilst et al. 1997). If such deep subduction occurred during major portions of Earth’s history, convective mixing will have destroyed the chemical separation between upper and lower mantle. The first attempt to integrate numerical convection modeling with chemical differentiation at crustal levels and recycling of ocean crust to generate enriched plume sources at the base of the mantle was made by Christensen and Hofmann (1994), and this has been followed up by more elaborate simulations.

More recently, a variety of new models have sprung up. Some invoke irregularly distributed, relatively primitive plume source regions, which have been protected from convective stirring throughout Earth’s history and are thus able to preserve primitive geochemistry. Tolstikhin and Hofmann (2005) and Boyet and Carlson (2005) have proposed a new form of a two-layer mantle, one in which a relatively small, compositionally dense reservoir formed at the base of the mantle, in effect constituting the D’’ layer. This irregular layer, on average about 200 km thick, has otherwise been interpreted as a “slab graveyard.” The new geochemical model stipulates that these subducted slabs are very ancient, perhaps only a few tens of millions of years younger than the accretion of the Earth. This, in turn, requires stabilization by high intrinsic density, either because they are derived from a primordial iron-rich mafic crust, or because they were generated by downward segregation of dense partial melts in the lowermost mantle (Labrosse et al. 2007) in effect also creating an Fe-rich, dense bulk composition. Such a “new two-layer” mantle model can account for the geochemical differences between plumes and mid-ocean ridge basalts, and it resolves at least two awkward problems with the old model: (1) it does not require that plumes have particularly high helium concentrations, which are never actually observed in plume-derived basalts; (2) it explains trace element characteristics of plume-derived basalts, which are in fact inconsistent with primitive sources. The new two-layer model, by its nature, eliminates the need for any undifferentiated, primitive reservoir, but it creates a repository of primordial noble gases in the permanently sequestered D’’-layer at the base of the mantle. From there, these noble gases can easily diffuse into the overlying actual (silicate) source reservoirs of mantle plumes.

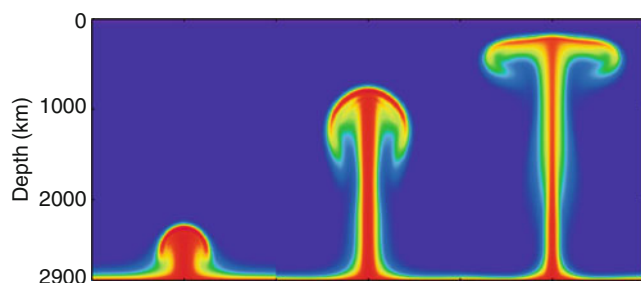
It should be emphasized that, by its very nature, geochemical evidence is extremely unlikely to either prove or disprove the existence of mantle plumes or mantle layering, because the mixing and extraction processes that deliver plume-derived or non-plume-derived melts to the surface are not known a priori. Unfortunately, in the past, it has too often

been argued that some ocean island must be derived from a deep-mantle plume because it contains high $^3\text{He}/^4\text{He}$ ratios, or near-chondritic $^{143}\text{Nd}/^{144}\text{Nd}$ ratios. Indeed, the inherent weakness of such arguments, which are clearly nonunique, has helped to discredit plume theory in some quarters. On the other hand, the plume model in combination with geochemical evidence can help to elucidate many fundamental aspects of Earth evolution, including the timescale of early Earth evolution and of crustal recycling, as well as the intensity of convective stirring in the mantle.

Overall, the “enriched” nature (i.e., enriched in highly incompatible elements relative to more compatible elements), which is evidenced by both the observed trace element and isotope abundances of the decay products of long-lived radioactive decay systems such as ^{87}Rb - ^{87}Sr , ^{147}Sm - ^{143}Nd , ^{176}Lu - ^{176}Hf in plume-derived rocks, is most easily explained by the subduction and recycling of enriched crustal rocks, rather than the involvement of any “primitive” mantle material (Hofmann 1997). Subduction of crustal materials, including ordinary ocean crust, enriched ocean crust, ocean islands and seamounts, as well as some sediments and other continental material, is based on geological and geophysical observations. Thus, there is no shortage of suitable source materials in the mantle without the need for alternative enrichment processes, such as “metasomatic” infiltration, which might nevertheless play some role. The recycled enriched materials appear to be more prevalent in mantle plumes than in mid-ocean ridge basalts, because basaltic crust, once subducted, is somewhat denser than ordinary peridotitic mantle. Because of this, it may be segregated and “stored” at the base of the mantle for geologically longer periods of time, and ultimately contribute significant portions of plume source materials (Christensen and Hofmann 1994).

Genesis and Dynamics of Plumes and Superplumes

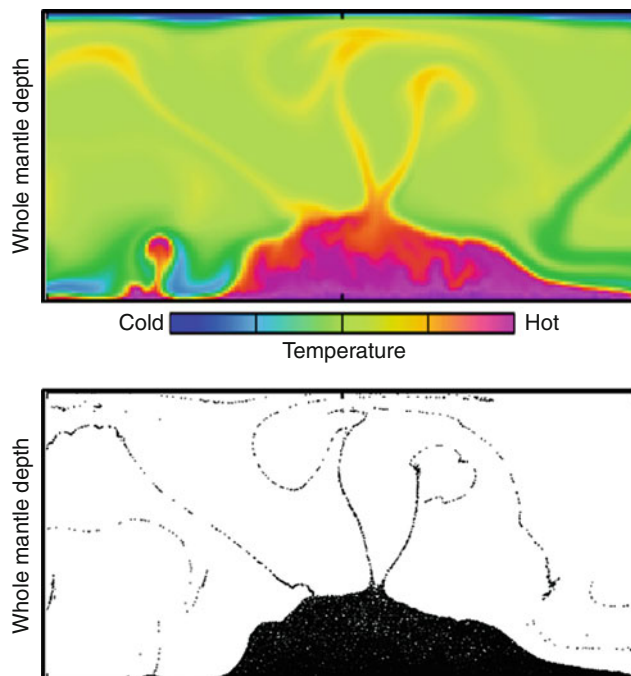
On timescales of millions of years, solid mantle rocks behave as highly viscous fluids with a viscosity of $\sim 10^{18}$ – 10^{22} Pa s (in comparison, the viscosity of a glacier is $\sim 10^{13}$ Pa s, and of a basaltic lava is ~ 10 – 10^4 Pa s). Thus, mantle dynamics is governed by a set of equations (conservation of mass, momentum, and energy) for a viscous fluid where inertial effects can be neglected. The growth rate of a Rayleigh–Taylor instability from a thermal boundary layer heated from below is controlled by the Rayleigh number: $\text{Ra} = \rho \alpha \Delta T g d^3 / k \eta$ where ΔT is the temperature contrast across a layer of thickness d , density ρ , thermal expansion coefficient α , thermal diffusivity k , and viscosity η . Instabilities are enhanced by the thermal buoyancy and inhibited by viscosity and thermal diffusivity. Using reasonable values for the above physical parameters, growing instabilities of the boundary



Mantle Plumes, Fig. 5 Numerical simulation of a thermal plume

layer form diapirs. Due to the high mantle viscosity, such a diapir will separate from the TBL only when its volume becomes sufficiently large (e.g., Whitehead and Luther 1975). The morphology of a thermal plume is controlled by the viscosity contrast between the hot fluid and the mantle above it: If the viscosity contrast is weak, plumes will have a “spout” shape, with little difference between the radius of the leading diapir (the plume head) and the following conduit (the plume tail), whereas increasing viscosity contrast leads to a larger head and a narrower conduit. This “mushroom” shape (Fig. 5) is favored, because a hot plume is likely to be 100 times less viscous than the surrounding mantle, owing to the strong temperature dependence of viscosity.

Although fluid dynamics laboratory experiments (e.g., Whitehead and Luther 1975; Griffiths and Campbell 1990) and numerical simulations (e.g., Parmentier et al. 1975; Olson et al. 1993) on purely thermal plumes enabled us to gain a quantitative understanding of plume dynamics, in the 1990s it became progressively clear that the lowermost mantle is compositionally heterogeneous. D'' is a region of preferential segregation and accumulation of denser subducted crust (Christensen and Hofmann 1994), and larger scale regions of the lower mantle may be chemically heterogeneous (Kellogg et al. 1999). This offered new and exciting avenues to explore the complex dynamics of thermochemical plumes, which can be defined as hot (positively buoyant) plumes that carry compositionally denser (negatively buoyant) material. Laboratory experiments by Davaille (1999) investigated a variety of regimes and found that instabilities may form dome-like structures with an oscillatory behavior (i.e., they rise and sink in response to a subtle balance between thermal and chemical buoyancies). Her experiments, together with numerical simulations of thermochemical convection (e.g., Tackley 1998) provide a fluid dynamically consistent framework to interpret observations that are otherwise unexplained by purely thermal convection. For example, the commonly referred “superplumes,” situated beneath the South-Central Pacific and Africa are broad (thousands of kilometers large) lower mantle zones of low seismic velocity most likely associated with active upwelling. Although there is a debate on the thermal and/or compositional origin of “superplumes,”



Mantle Plumes, Fig. 6 Numerical simulation of a thermochemical plume. Top: Temperature field. Bottom: Chemically denser material, represented with tracers. Animations are available

several lines of evidence do support their distinct composition (Ishii and Tromp 1999; Masters et al. 2000). Another issue that can be readily explained by thermochemical plumes is the discrepancy between the petrologically constrained excess temperature of plumes (100–250 °C) and the estimated temperature difference across the D'' region (~1,000 °C). Numerical simulations (Farnetani 1997) show that the deepest, hottest part of the thermal boundary layer does not upwell in mantle plumes if the compositional density contrast exceeds 2%. A denser zone in the boundary layer also tends to “anchor” the base of plume conduit, so that the flow pattern remains stable over timescales longer than the plume rise time (Jellinek and Manga 2002). Finally, thermochemical plumes present a variety of shapes and surface manifestations (Farnetani and Samuel 2005; Kumagai et al. 2007) that expand the classical plume head-tail model predictions (Fig. 6).

Plume Melting and Plume Strength

At spreading ridges, mantle rocks rise to shallow depths in response to plate spreading, and melt by adiabatic decompression (McKenzie and Bickle 1988). This is not the case for most plumes impinging at the base of a preexisting and unrifted continental or oceanic lithosphere. In order to melt at relatively high pressure (4–5 GPa), plumes must be either hotter than normal mantle, or compositionally more “fertile”

than peridotitic mantle (Condie 2001, Chap. 4). Plumes are probably both: their excess temperature is estimated between 100 and 250 °C (Putirka 2005), and they can carry 20%, or more, of recycled crust in the form of eclogite (Sobolev et al. 2007). Eclogite is more fertile than peridotite because it has a lower solidus and a greater melt productivity. According to Sobolev et al. (2005), reaction between eclogite-derived liquids and solid peridotite forms a pyroxenite, whose subsequent melting may explain some compositional characteristics of Hawaiian (and many other plume-derived) lavas. Another efficient way of lowering the solidus temperature is to have fluids (H_2O , CO_2) in the upwelling plume rocks; however, the evidence for “fluid-rich” hotspot lavas is scant. Although it is well known that melting is the most efficient way to create chemical differentiation in the mantle and that it is the unavoidable process to generate surface lavas from mantle rocks, many aspects of partial melting remain elusive (e.g., the pressure–temperature conditions during melting; the source heterogeneous composition; the mechanism of melting, of transport and of storage of the magma). This limits our quantitative understanding of the link between petrological and geochemical observations of surface lavas and the underlying mantle plume composition and dynamics.

The strength of a mantle plume can be calculated on the production rate of volcanic rocks. The method utilizes the volcano’s volume and growth time, and the estimated melt fraction. For example, Mauna Loa has a volume of $\sim 70,000 \text{ km}^3$ and it grew in $\sim 1 \text{ Ma}$, suggesting that the Hawaiian plume melt production rate M is about $0.1\text{--}0.2 \text{ km}^3/\text{year}$. Since only a fraction of the upwelling plume melts, the total plume volume flux $Q_v \gg M$. Another method, based on the rate of swell formation (Sleep 1990), enables one to calculate the plume buoyancy flux $B = Q_v \Delta\rho$ where $\Delta\rho = \rho \alpha (T_{\text{PLUME}} - T_{\text{MANTLE}})$. For Hawaii, the most vigorous hotspot, $B \sim 8,000 \text{ kg/s}$, for Iceland $B \sim 1,400 \text{ kg/s}$, for Galapagos $B \sim 1,000 \text{ kg/s}$. Finally, the plume heat flow is $Q_h = \rho_{\text{MANTLE}} C_p Q_v (T_{\text{PLUME}} - T_{\text{MANTLE}})$, where the specific heat at constant pressure is $C_p \sim 1,200 \text{ J/kg K}$. For Hawaii, the heat flow is $Q_h \sim 360 \times 10^9 \text{ W}$, and it represents $\sim 16\%$ of the global hotspot heat flow of $2.3 \times 10^{12} \text{ W}$, which, in turn is a small fraction of the Earth’s total heat flow of $44 \times 10^{12} \text{ W}$ (for further reading, see Schubert et al. 2001, Chap. 11).

Continental Flood Basalts and Continental Breakup

The eruption of a large igneous province (LIP) represents a major geologic event head. Over the Earth’s history, episodic LIP magmatism contributed to continental growth through the emplacement of CFB and through the accretion/obduction of fragments of oceanic plateaus onto continental crust (Ben-Avraham et al. 1981). Accretion may be due to the

difficulty of subducting the 20–40 km thick crust of an oceanic plateau. Three notable examples of accreted LIP fragments are found in the Solomon Island arc (from the Ontong-Java plateau), in Central America (from the Caribbean plateau), and the Wrangellia terrane outcropping in British Columbia and in SE Alaska (Fig. 7). This allochthonous terrane, locally attaining a thickness of 6 km, is likely to be a fragment of a 230 Ma-old oceanic plateau (Ben-Avraham et al. 1981; Richards et al. 1991; Greene et al. 2008).

Distinctive characteristics of LIP magmatism are the short duration (1–2 Ma) and the huge volumes ($1\text{--}10 \times 10^6 \text{ km}^3$) of eruption. Because of the high eruption rates ($> 1\text{--}2 \text{ km}^3/\text{year}$; for comparison Kilauea grows at $0.1 \text{ km}^3/\text{year}$) and the enormous surface extent ($1\text{--}2 \times 10^6 \text{ km}^2$; for comparison France’s surface is $\sim 0.5 \times 10^6 \text{ km}^2$), flood basalt eruptions are extraordinary volcanic events. CFBs consist of sub-horizontal flows of mafic (Fe- and Mg-rich) rocks, mainly tholeiitic basalts. Individual flows can extend for hundreds of kilometers, be tens to hundreds of meters thick, and have volumes of more than 10^3 km^3 . CFBs are more accessible than submerged oceanic plateaus; however they can be extensively eroded, fragmented, and dispersed on different continents by plate tectonic processes. Clearly, uncertainties regarding the original volume and surface extent increase with their age, for example, for pre-Cambrian CFB only the giant swarm of mafic dikes, feeding surface volcanism, may be left (Ernst and Buchan 2001).

The Phanerozoic CFBs are, in chronological order: The 258 Ma-old Emeishan Traps (SW China), which presently cover only $0.3 \times 10^6 \text{ km}^2$, probably a tenth of the estimated original surface; the 250 Ma-old Siberian Traps, which might have had a surface extent of $3\text{--}4 \times 10^6 \text{ km}^2$ and extrusive volumes of 10^6 km^3 ; and the 200 Ma-old lavas and dike swarms of the Central Atlantic Magmatic Province outcrop in once-contiguous parts of North America (e.g., the Palisades sill, NY), West Africa, and Brazil. This flood volcanism, with an estimated total volume of 3×10^6 to $5 \times 10^6 \text{ km}^3$, preceded the opening of the central Atlantic Ocean. The 184–182 Ma-old Karoo (Southern Africa)–Ferrar (Antarctica) traps preceded the breakup of Gondwana by 10–15 Ma. Similarly, the $\sim 135\text{--}130 \text{ Ma}$ -old magmatism of the Paraná (Brazil)–Etendeka (Namibia) province led to the opening of the South Atlantic Ocean. The Deccan Traps (India) erupted 66–65 Ma ago over an area of the subsequent rifting split apart the Seychelles from India. The North Atlantic Tertiary Igneous Province covers a surface greater than 10^6 km^2 and has an estimated volume of $6 \times 10^6 \text{ km}^3$. The earliest volcanism occurred as flood basalts in Baffin Island, West- and East-Greenland (62 Ma), and later (56–54 Ma) extended to the continental margins of Greenland, the British Isles, and Norway. These rifted edges of continents constitute the so-called volcanic passive margins of up to 8 km-thick seaward dipping basalt layers. The complete opening of the



Mantle Plumes, Fig. 7 Photograph of 1,000 m of continuous subaerial flood basalt stratigraphy in the Wrangell Mountains, Alaska. The yellow line marks the contact between Nikolai basalts and the overlying Chitstone Limestone. (From Greene et al. 2008. With permission)

North Atlantic and the appearance of truly oceanic crust occurred about 53 Ma ago. Finally, the Ethiopian–Yemen traps, ~30 Ma old, have been associated with the rifting in the Red Sea and the Gulf of Aden (Courtilot et al. 1999).

Issues concerning the timing of rifting have been hotly debated: One side claims that rifting preceded, and enhanced, flood volcanism (e.g., White and McKenzie 1989); the other side argues that rifting is not a prerequisite to flood volcanism, since it occurs only during and after volcanism (e.g., Richards et al. 1989). More geological observations are needed, as well as numerical simulations and/or laboratory experiments investigating how anomalously warm mantle can erode the lithosphere and create (or reactivate) weak zones susceptible to rifting.

LIP Magmatism and Environmental Effects

The remarkable temporal correlation between LIP magmatism and mass extinctions of terrestrial and marine organisms suggests a cause-and-effect connection (Courtilot and Renne 2003). The most notable example is the Siberian Traps at the Permian–Triassic boundary (Renne et al. 1995; Svensen et al. 2009). The emplacement of the

Deccan Traps, spanning the Cretaceous–Tertiary boundary, certainly had an environmental effect (Self et al. 2006); however, the extinction event coincided with a large asteroid impact at Chicxulub, Mexico (Schulte et al. 2010). The peak of volcanism in the North Atlantic Volcanic Province corresponds to the Paleocene–Eocene Thermal Maximum (PETM) event 55 Ma ago. Svensen et al. (2004) propose that voluminous magmatic intrusions in carbon-rich sediments induced massive release of the methane buried in marine sediments, thus suggesting that volcanic and metamorphic processes associated with the opening of the North Atlantic may explain global climate events. Self et al. (2006), Ganino and Arndt (2009), and Svensen et al. (2009) attempt to quantify the release of volatiles (CO_2 , CH_4 , SO_2) during flood volcanism and to estimate the consequences on the environment. In particular, the release of volcanic ashes and sulfuric acid aerosols can lead to darkening and cooling, SO_2 and HCl cause acid rains, whereas greenhouse gases like CO_2 and CH_4 cause global warming. Carbon has two origins: magmatic and sediment-derived. Interestingly, Ganino and Arndt (2009) show that the mass of sediment-derived CO_2 can be four to eight times larger than the mass of magmatic CO_2 , if contact metamorphism reactions occur in sedimentary rocks such as dolomite, evaporite, coal, or organic-rich shale. In

other words, the environmental effect of flood basalt magmatism depends, among other factors, on the rock type in contact with magmatic sills and lava flows (Ganino and Arndt 2009; Svensen et al. 2009).

Marine eruptions can also have an environmental effect, because they can cause variations in water chemistry and oceanic circulation. The emplacement of the (Alaska-sized) Ontong-Java plateau 120 Ma ago and of the Caribbean plateau correlate temporally with some of the major Oceanic Anoxic Events (OAEs) which occurred 183, 120, 111, and 93 Ma ago. OAEs are associated with an abrupt rise in temperature, induced by rapid influx of CO₂. The consequent increase of organic productivity causes an increase of oxygen demand in the water column, eventually leading to oxygen depletion in the oceans (for a thorough review see Jenkyns 2010 and references therein). Furthermore, the formation of black shale (marine carbon-rich sediments) may have been favored by massive hydrothermal release of trace metals, poisonous to marine life.

Plume–Lithosphere Interaction

A quantitative understanding of plume–lithosphere interaction enables us to relate geophysical observations (e.g., the spatiotemporal evolution of surface topography, variations in lithospheric thickness, etc.) to the physical parameters and dynamics of the plume. Central issues concerning plume–lithosphere interactions include, but are not limited to the following:

- First, estimating the dynamic topography (Definition of dynamic topography, extracted from Lithgow-Bertelloni and Silver (1998). “Dynamic topography is a deformation of the surface of the Earth, supported by the vertical stresses at the base of the lithosphere, which are generated by flow in the mantle below. This is in contrast to the more familiar mechanism of isostatically supported topography which is in equilibrium at the Earth’s surface and would exist even in a static mantle.”) induced by the arrival of a mantle plume head. Pioneering laboratory experiments by Olson and Nam (1986) found that a thermal diapir upwelling and spreading beneath the lithosphere induces a rapid surface uplift followed by a slower subsidence. According to Farnetani and Richards (1994), uplift should precede flood volcanism by 1–2 Ma, a prediction that is sometimes validated by geological observations (e.g., for the Emeishan Traps), but not always (e.g., for the Siberian Traps). A present-day example of dynamic topography is provided by Lithgow-Bertelloni and Silver (1998), who suggest that the anomalous elevation (~1 km) of South-Central Africa has a dynamic origin, being induced by mantle upwelling from the African “superplume,” a broad lower mantle zone of low seismic velocity anomalies.
- Second, understanding the origin of the topographic hotspot swell and the thermomechanical response of the lithosphere drifting over the plume. Two main mechanisms have been proposed for the Hawaiian swell: “thermal rejuvenation” invokes heating and thinning of the lithosphere above the hotspot, whereas “dynamic support” invokes stresses applied to the base of the lithosphere by the buoyant plume. Seismic studies by Li et al. (2004) around Hawaii found a thinning of the lithosphere from the expected 100 km (beneath the Big Island) to only 50–60 km (beneath Oahu and Kauai, 400 km downstream from the Big Island), thus suggesting a hybrid scenario, whereby “dynamic support” prevails above the plume conduit and “thermal rejuvenation” prevails downstream. Far from the conduit, thermal thinning of the lithosphere may be enhanced by small-scale convective instabilities, which take the form of rolls aligned with the plate motion, as observed in numerical simulations with strongly temperature-dependent viscosity (Moore et al. 1998).
- Third, mechanical and chemical interaction between plume magmas and the lithosphere can explain a number of geophysical and geochemical observations. For example, the occurrence of regularly spaced volcanoes in hotspot chains is probably caused by the interaction of magma transport with lithospheric flexure (Hieronymus and Bercovici 1999, 2000). Under the load of the volcanoes, the lithosphere deforms and the resulting flexural stresses have a key role in determining the locus where magma preferentially extrudes. There are also clear examples of geochemical interaction between the continental lithosphere and plume magmas “en route” to the surface. In particular, CFB lavas erupted at the onset of flood volcanism often have anomalous compositions, reflecting contamination by the lithosphere and crust. It is thus important to quantify the contribution of the continental lithosphere before inferring the plume’s geochemical fingerprint.
- Fourth, plume material may preferentially flow – and partially melt – in preexisting zones of thinned lithosphere (Sleep 1996). In other words, the thickness of the lithosphere and the slope of its base may control the spatial distribution of plume magmatism. According to Ebinger and Sleep (1998), the Cenozoic magmatism in East Africa (including the Afar, the Cameroon volcanic line to the West, and the Comoros Islands to the South) can be explained by a single plume impinging beneath the Ethiopian plateau and its subsequent lateral flow, “channeled” in preexisting rift zones characterized by a relatively shallow base of the lithosphere.

Plume–Ridge Interaction

At least 21 hotspots are situated near a spreading ridge, the clearest examples being Iceland, a ridge-centered hotspot, Galapagos, and the Azores. At Iceland, the Mid-Atlantic Ridge rises ~4 km above normal and remains anomalously shallow for ~1,000 km to the North and the South. Below Iceland, the oceanic crust is 40 km thick, whereas more normal crustal thicknesses of 8–10 km are attained only ~500 km far from the hotspot. Furthermore, compositional anomalies in trace elements and isotope ratios observed along the ridge clearly indicate plume–ridge interaction (Schilling et al. 1985). The Galapagos hotspot is situated ~200 km off-axis from the Galapagos Spreading Center, and, also in this case, geophysical and geochemical anomalies along the ridge suggest plume–ridge interaction. These observations raise several questions concerning: (a) how plume material flows toward the ridge, opposite to the plate motion, (b) how plume material, once “trapped” beneath the ridge, flows parallel to the ridge axis, and (c) how the plume–ridge flow is affected by physical parameters such as the plume volume flux, the plume–ridge distance, and plate velocity. As reviewed by Ito et al. (2003), such questions have been addressed with laboratory experiments (e.g., Kincaid et al. 1995; Feighner and Richards 1995) and theoretical and numerical models (e.g., Ribe et al. 1995; Ribe 1996; Ribe and Delattre 1998) that progressively included more realistic conditions, such as the conductive thickening of the oceanic lithosphere away from the ridge and a migrating spreading ridge. The sloping base of the oceanic lithosphere favors the flow of plume material toward the ridge (Sleep 1996), and the subsequent channeling of plume flow beneath the ridge has been calculated numerically by Albers and Christensen (2001). Such studies concluded that the width over which plume material flows along the ridge increases with higher plume flux and decreases with higher spreading rate and plume–ridge distance (see Ito et al. 2003; Ito and van Keken 2007). Another important parameter is plume viscosity, in particular, Ito et al. (2003) find that the plume flow is reduced by the process of dehydration strengthening, which increases the viscosity of the plume residuum. For a fluid dynamical analysis of plume–ridge interaction see the comprehensive paper by Ribe et al. (2007) and the review paper by Ito et al. (2003).

Summary and Conclusion

This chapter on mantle plumes has taken us from the deep Earth’s mantle where plumes originate, to sublithospheric depths where partial melting occurs, and up to the surface where LIPs have been emplaced in the geologic past, and hotspot lavas are being erupted today. We also explored the

environmental effect of massive LIP volcanism on the oceans and the atmosphere. This range of topics shows that a full understanding of mantle plumes requires a multidisciplinary approach. Important progress will be achieved by linking geophysical, geochemical, and geological observations, to models solving for the fluid dynamics of mantle plumes, melt migration, and physical volcanology, just to mention a few.

Future Directions

In our view, fundamental questions for the future include the following:

1. Quantifying the effect of heterogeneities on plume dynamics and fertility. The term heterogeneity is used here in a broad sense, from small-scale (1–10 km) lenses of recycled eclogite to the large-scale (100–1,000 km) denser superplumes that may generate short-lived instabilities rising from their top. Future studies should model heterogeneous plumes in a fully convecting mantle, with physical properties provided by mineral physics experiments at appropriate temperature and pressure conditions. Lithologic variability affects the fertility and the geochemical characteristics of lavas (Sobolev et al. 2005), but, at present, numerical simulations are unable to model the complex petrological and geochemical evolution of the solid matrix and melts.
2. Quantifying the role of the lithosphere is needed to understand how lithospheric stresses and variations in lithospheric thickness and composition may affect partial melting and the ascent of magma. Hieronymus and Bercovici (2000) provide a theory to explain the distribution of non-hotspot island and seamount chains in the South Pacific and conclude that volcanism can be activated by tensile stresses. The hypothesis that thermoelastic cracking of the oceanic lithosphere (Sandwell and Fialko 2004) can trigger magmatism should be explored quantitatively, to assess the volumes and the composition of such magmas.
3. The cause-and-effect relation between the arrival of a mantle plume and continental breakup is often invoked, but we lack a quantitative understanding of the processes involved. In other words, although there is evidence supporting the role of plumes in the creation and modification of plate boundaries, it is unclear how a large-scale system, with continents and preexisting spreading ridges responds to the arrival of a large mantle plume. Moreover, studies of plume–lithosphere interaction should explore the possibility of dynamic instabilities, such as lithospheric delamination, fingering of low-viscosity fluid, and small-scale convection (see Ito and van Keken 2007 and references therein), and their ability to generate partial melts.

4. Future studies should address in a constructive and quantitative way the current debate (see www.mantleplumes.org) on the “existence of plumes,” or better, on the “coexistence of plumes with other processes” that can produce surface volcanism. Progress will be achieved through improved seismic imaging of plumes and their depth of origin, and through a quantitative understanding of lithospheric processes, as explained above. We also note that much of the ongoing debate questioning the existence of plumes is based on a narrow view of what a plume should look like: the head-tail model is valid in a purely thermal, homogeneous mantle, but a variety of plume shapes are found if the Earth’s lower mantle is compositionally heterogeneous (Farnetani and Samuel 2005; Kumagai et al. 2007). Furthermore, when considering plumes in a convecting mantle, rather than in an unperturbed fluid, it is obvious that conduits cannot be fixed (Steinberger et al. 2004), an observation that does not preclude the existence of deep plumes.
 5. Although we have restricted our treatment to Mesozoic–Cenozoic plumes (<250 Ma), it is likely that plumes had an important role also at earlier times, although geological evidence becomes increasingly scant with geological age (Ernst and Buchan 2001). The early Earth, with its vigorous and highly time-dependent convection, probably had an intense plume magmatism. Hopefully, studies of volcanism on Mars and Venus can complement and improve our knowledge of early terrestrial magmatism (see Condie 2001, Chap. 3). Both Mars and Venus lack plate tectonics, so that old geologic features can be better preserved. On Mars, the Olympus Mons is an enormous shield volcano (volume $2 \times 10^6 \text{ km}^3$), and the Tharsis rise, which covers 20% of the Martian surface, has been volcanically active for the last 2 Ga, suggesting the presence of huge and long-lived plumes in the Martian mantle (Harderer and Christensen 1996).
- Burke K, Torsvik TH (2004) Derivation of large igneous provinces of the past 200 million years from long-term heterogeneities in the deep mantle. *Earth Planet Sci Lett* 227:531–538
- Christensen UR, Hofmann AW (1994) Segregation of subducted oceanic crust in the convecting mantle. *J Geophys Res* 99:19867–19884
- Clouard V, Gerbault M (2008) Break-up spots: could the Pacific open as a consequence of plate kinematics? *Earth Planet Sci Lett* 265:195–208
- Coffin MF, Eldholm O (1994) Large igneous provinces: crustal structure, dimensions and external consequences. *Rev Geophys* 32:1–36
- Condie KC (2001) *Mantle plumes and their record in earth history*. Cambridge University Press, Cambridge
- Courtillot VE, Renne P (2003) On the ages of flood basalt events. *Compt Rendus Geosci* 335:113–140
- Courtillot VE, Jaupart C, Manighetti I, Tapponnier P, Besse J (1999) On causal links between flood basalts and continental breakup. *Earth Planet Sci Lett* 166:177–195
- Courtillot VE, Davaille A, Besse J, Stock J (2003) Three distinct types of hot spots in the Earth’s mantle. *Earth Planet Sci Lett* 205:295–308
- Davaille A (1999) Simultaneous generation of hotspots and superswells by convection in a heterogeneous planetary mantle. *Nature* 402:756–760
- Ebinger CJ, Sleep NH (1998) Cenozoic magmatism throughout east Africa resulting from impact of a single plume. *Nature* 395:788–791
- Ernst RE, Buchan KL (2001) The use of mafic dike swarms in identifying and locating mantle plumes. In: Ernst RE, Buchan KL (eds) *Mantle plumes: their identification through time*, Geological Society of America special paper, 352. Geological Society of America, Colorado
- Farley KA, Neroda E (1998) Noble gases in the Earth’s mantle. *Annu Rev Earth Planet Sci* 26:189–218
- Farnetani CG (1997) Excess temperature of mantle plumes: the role of chemical stratification across D′. *Geophys Res Lett* 24:1583–1586
- Farnetani CG, Richards MA (1994) Numerical investigation of the mantle plume initiation model for flood basalt events. *J Geophys Res* 99:13813–13883
- Farnetani CG, Samuel H (2005) Beyond the thermal plume paradigm. *Geophys Res Lett* 32:L07311. <https://doi.org/10.1029/2005GL022360>
- Feighner MA, Richards MA (1995) The fluid dynamics of plume-ridge and plume-plate interactions: an experimental investigation. *Earth Planet Sci Lett* 129:171–182
- Foulger GR, Natland JH (2003) Is “hotspot” volcanism a consequence of plate tectonics? *Science* 300:921–922
- Ganino C, Arndt NT (2009) Climate changes caused by degassing of sediments during the emplacement of large igneous provinces. *Geology* 37:323–326. <https://doi.org/10.1130/G25325A.1>
- Greene AR, Scoates JS, Weis D (2008) Wrangellia flood basalts in Alaska: a record of plume-lithosphere interaction in a Late Triassic accreted oceanic plateau. *Geochem Geophys Geosyst* 9:Q12004. <https://doi.org/10.1029/2008GC002092>
- Griffiths RW, Campbell IH (1990) Stirring and structure in mantle starting plumes. *Earth Planet Sci Lett* 99:66–78
- Harderer H, Christensen UR (1996) A one-plume model of Martian mantle convection. *Nature* 380:507–509
- Hart SR, Schilling JG, Powell JL (1973) Basalts from Iceland and along the Reykjanes Ridge: Sr isotope geochemistry. *Nature* 246:104–107
- Hieronimus CF, Bercovici D (1999) Discrete alternating hotspot islands formed by interaction of magma transport and lithospheric flexure. *Nature* 397:604–607
- Hieronimus CF, Bercovici D (2000) Non-hotspot formation of volcanic chains: control of tectonic and flexural stresses on magma transport. *Earth Planet Sci Lett* 181:539–554
- Hofmann AW (1997) Mantle geochemistry: the message from oceanic volcanism. *Nature* 385:219–229
- Hofmann AW (2003) Sampling mantle heterogeneity through oceanic basalts: isotopes and trace elements. In: Carlson RW, Holland HD,

Cross-References

- [Core-Mantle Coupling](#)
- [Mantle Convection](#)
- [Mantle D′ Layer](#)
- [Mantle Viscosity](#)

Bibliography

- Albers M, Christensen UR (2001) Channeling of plume flow beneath mid-ocean ridges. *Earth Planet Sci Lett* 187:207–220
- Ben-Avraham Z, Nur A, Jones D, Cox A (1981) Continental accretion: from oceanic plateaus to allochthonous terranes. *Science* 213:47–54
- Boyett M, Carlson RW (2005) Nd-142 evidence for early (>4.53 Ga) global differentiation of the silicate Earth. *Science* 309:576–581

- Turekian KK (eds) *Treatise on geochemistry: the mantle and core*. Elsevier, Oxford, pp 61–101
- Hofmann AW, White WM (1982) Mantle plumes from ancient oceanic crust. *Earth Planet Sci Lett* 57:421–436
- Ishii M, Tromp J (1999) Normal-mode and free-air gravity constraints on lateral variations in velocity and density of Earth's mantle. *Science* 85:1231–1236
- Ito G, van Keken PE (2007) Hotspots and melting anomalies. In: Bercovici D (ed) *Mantle dynamics, Treatise on geophysics*, vol 7. Elsevier, Amsterdam
- Ito G, Lin J, Graham D (2003) Observational and theoretical studies of the dynamics of mantle plume–mid-ocean ridge interaction. *Rev Geophys* 41:1017. <https://doi.org/10.1029/2002RG000117>
- Jacobsen SB, Wasserburg GJ (1979) The mean age of mantle and crustal reservoirs. *J Geophys Res* 84:7411–7427
- Jellinek AM, Manga M (2002) The influence of a chemical boundary layer on the fixity, spacing and lifetime of mantle plumes. *Nature* 418:760–763
- Jellinek AM, Manga M (2004) Links between long-lived hot spots, mantle plumes, D'' and plate tectonics. *Rev Geophys* 42:RG3002. <https://doi.org/10.1029/2003RG000144>
- Jenkyns HC (2010) Geochemistry of oceanic anoxic events. *Geochem Geophys Geosyst* 11:Q03004. <https://doi.org/10.1029/2009GC002788>
- Kellogg LH, Hager BH, van der Hilst RD (1999) Compositional stratification in the deep mantle. *Science* 283:1881–1884
- Kincaid C, Ito G, Gable C (1995) Laboratory investigation of the interaction of off-axis mantle plumes and spreading centres. *Nature* 367:758–761
- Kumagai I, Davaille A, Kurita K (2007) On the fate of thermally buoyant mantle plumes at density interfaces. *Earth Planet Sci Lett* 254:180–193
- Labrosse S, Hernlund JW, Coltice N (2007) A crystallizing dense magma ocean at the base of the Earth's mantle. *Nature* 450:866–869
- Li X, Kind R, Yuan X, Wolbern I, Hanka W (2004) Rejuvenation of the lithosphere by the Hawaiian plume. *Nature* 427:827–829
- Lithgow-Bertelloni C, Silver PG (1998) Dynamic topography, plate driving forces and the African superswell. *Nature* 395:269–272
- Loper DE, Stacey FD (1983) The dynamical and thermal structure of deep mantle plumes. *Phys Earth Planet Inter* 33:305–317
- Masters G, Laske G, Bolton H, Dziewonski AM (2000) The relative behavior of shear velocity, bulk sound speed and compressional velocity in the mantle: implication for chemical and thermal structure, in Earth's deep interior. In: Karato SI, Forte A, Liebermann RC, Masters G, Stixrude L (eds) *Mineral physics and tomography from the atomic to the global scale*, vol 117. American Geophysical Union, Washington, DC, pp 63–87
- McKenzie D, Bickle MJ (1988) The volume and composition of melt generated by extension of the lithosphere. *J Petrol* 29:625–679
- Montelli R, Nolet G, Masters G, Engdahl ER, Hung S-H (2004) Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* 303:338–343
- Moore WB, Schubert G, Tackley P (1998) Three-dimensional simulations of plume-lithosphere interaction at the Hawaiian swell. *Science* 279:1008–1011
- Morgan WJ (1971) Convection plumes in the lower mantle. *Nature* 230:42–43
- Olson P, Nam I-S (1986) Formation of seafloor swells by mantle plumes. *J Geophys Res* 91:7181–7191
- Olson P, Schubert G, Anderson C (1993) Structure of axisymmetric mantle plumes. *J Geophys Res* 98:6829–6844
- Parmentier EM, Turcotte DL, Torrance KE (1975) Numerical experiments on the structure of mantle plumes. *J Geophys Res* 80:4417–4424
- Putirka KD (2005) Mantle potential temperatures at Hawaii, Iceland, and the mid-ocean ridge system, as inferred from olivine phenocrysts: evidence for thermally driven mantle plumes. *Geochem Geophys Geosyst* 6:Q05L08. <https://doi.org/10.1029/2005GC000915>
- Renne PR, Zhang ZC, Richards MA, Black MT, Basu AR (1995) Synchrony and causal relations between Permian-Triassic boundary crises and Siberian flood volcanism. *Science* 269:1413–1416
- Ribe NM (1996) The dynamics of plume-ridge interaction 2. Off-ridge plumes. *J Geophys Res* 101:16195–16204
- Ribe NM, Delattre WL (1998) The dynamics of plume-ridge interaction – III. The effects of ridge migration. *Geophys J Int* 133:511–518
- Ribe NM, Christensen UR, Theissing J (1995) The dynamics of plume-ridge interaction, 1: ridge-centered plumes. *Earth Planet Sci Lett* 134:155–168
- Ribe NM, Davaille A, Christensen UR (2007) Fluid dynamics of mantle plumes. In: Ritter JRR, Christensen UR (eds) *Mantle plumes – a multidisciplinary approach*. Springer, Berlin, pp 1–48
- Richards MA, Duncan RA, Courtillot VE (1989) Flood basalts and hotspot tracks: plume heads and tails. *Science* 246:103–107
- Richards MA, Jones DL, Duncan RA, DePaolo DJ (1991) A mantle plume initiation model for the Wrangellia flood basalt and other oceanic plateaus. *Science* 254:263–267
- Romanowicz B, Gung YC (2002) Mega-upwellings from the core–mantle boundary to the lithosphere: implications for heat flux. *Science* 296:513–516
- Sandwell DT, Fialko Y (2004) Warping and cracking of the Pacific plate by thermal contraction. *J Geophys Res* 109:B10411. <https://doi.org/10.1029/2004JB003091>
- Schilling JG (1973) Iceland mantle plume: geochemical evidence along Reykjanes Ridge. *Nature* 242:565–571
- Schilling JG, Thompson G, Kingsley R, Humphris S (1985) Hotspot-migrating ridge interaction in the South Atlantic. *Nature* 313:187–191
- Schubert G, Turcotte DL, Olson P (2001) *Mantle convection in the earth and planets*. Cambridge University Press, Cambridge
- Schulte P et al (2010) The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. *Science* 327:1214–1218
- Self S, Widdowson M, Thordarson T, Jay AE (2006) Volatile fluxes during flood basalt eruptions and potential effects on the global environment: a Deccan perspective. *Earth Planet Sci Lett* 248:518–532
- Sleep NH (1990) Hotspot and mantle plumes: some phenomenology. *J Geophys Res* 95:6715–6736
- Sleep NH (1996) Lateral flow of hot plume material ponded at sub-lithospheric depths. *J Geophys Res* 101:28065–28083
- Smith WHF, Sandwell DT (1997) Global sea floor topography from satellite altimetry and ship depth soundings. *Science* 277:1956–1962
- Sobolev AV, Hofmann AW, Sobolev SV, Nikogosian IK (2005) An olivine-free mantle source of Hawaiian shield basalts. *Nature* 434:590–597
- Sobolev AV, Hofmann AW et al (2007) The amount of recycled crust in sources of mantle-derived melts. *Science* 316:412–417
- Steinberger B, Sutherland R, O'Connell RJ (2004) Prediction of Emperor-Hawaii seamount locations from a revised model of global plate motion and mantle flow. *Nature* 430:167–173
- Svensen H, Planke S, Maltre-Sorensen A, Jamtveit B, Myklebust R, Eidem TR, Rey SS (2004) Release of methane from a volcanic basin as a mechanism for initial Eocene global warming. *Nature* 429:542–545
- Svensen H, Planke S et al (2009) Siberian gas venting and the end-Permian environmental crisis. *Earth Planet Sci Lett* 277:490–500
- Tackley PJ (1998) Three-dimensional simulations of mantle convection with a thermo-chemical basal boundary layer: D'' ? In: Gurnis M et al (eds) *The core–mantle boundary region, Geophysical monograph series*, vol 28. American Geophysical Union, Washington, DC, pp 231–253
- Tarduno J, Bunge H-P, Sleep N, Hansen U (2009) The bent Hawaiian-Emperor hotspot track: inheriting the mantle wind. *Science* 324:50–53

- Tolstikhin I, Hofmann AW (2005) Early crust on top of the Earth's core. *Phys Earth Planet Inter* 148:109–130
- van der Hilst R, Widiyantoro S, Engdahl E (1997) Evidence for deep mantle circulation from global tomography. *Nature* 386:578–584
- Wessel P (1993) Observational constraints on models of the Hawaiian hot spot swell. *J Geophys Res* 98:16095–16104
- White RS, McKenzie DP (1989) Magmatism at rift zones: the generation of volcanic continental margins and flood basalts. *J Geophys Res* 94:7685–7729
- Whitehead JA, Luther DS (1975) Dynamics of laboratory diapir and plume models. *J Geophys Res* 80:705–717
- Wilson JT (1963) A possible origin of the Hawaiian Islands. *Can J Phys* 41:863–870
- Wolfe CJ, Solomon SC, Laske G, Collins JA, Detrick RS, Orcutt JA, Bercovici D, Hauri EH (2009) Mantle shear-wave velocity structure beneath the Hawaiian hot spot. *Science* 326:1388–1390

Mantle Viscosity

W. R. Peltier

Department of Physics, University of Toronto, Toronto, ON, Canada

Definition

The viscosity of the mantle of the Earth is the physical parameter that determines the strength of viscous forces that act when polycrystalline mantle material flows like a liquid on very long timescales. The magnitude of this viscosity is on the order of 10^{21} Pa s, so high compared to water (1 Pa s) that when the mantle flows it does so without inertia.

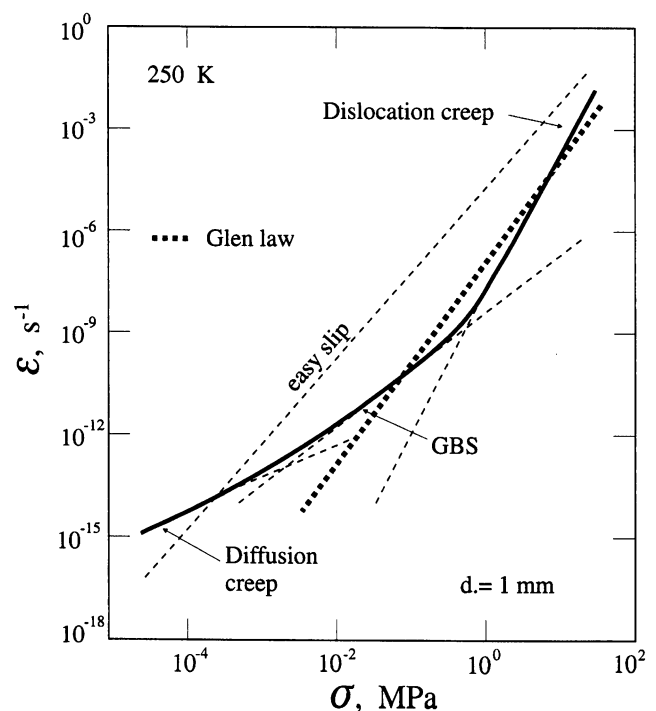
Introduction

The effective viscosity of Earth's predominantly iron-magnesium silicate mantle is a physical property of fundamental importance. In characterizing this outer shell of the planet by a viscosity at all, one is recognizing that it is capable of flowing as if it were a viscous fluid. This is in spite of the fact that on the timescales characteristic of seismic waves, it is known to be accurately describable as a Hookean elastic solid, albeit one which is endowed with a finite quality factor “Q” due to imperfections of elasticity. The idea that viscoelastic materials may behave as solids on short time scales but as viscous fluids on sufficiently long timescales is well understood in the material science literature. For Earth sciences, the fact that the mantle beneath the almost perfectly elastic surface lithosphere behaves viscously on timescales in excess of a few hundred years, the so-called Maxwell time, is vital to the understanding of plate tectonics and the origins of continental drift as involving the action of thermal convection in the mantle driven by the large contrast in temperature between

Earth's surface and the core-mantle boundary (cmb). This article provides a summary of laboratory and geophysical data that may be brought to bear as a means of measuring mantle viscosity and discusses the compatibility of these measurements with the requirements of the mantle convection process.

Methods of Measurement

One might imagine that laboratory measurements of the creep of mantle mineral assemblages at high pressure and temperature would provide all that is required to obtain the data needed to infer effective viscosity. There are several problems with this idea, however, most of which derive from the difficulty of performing the appropriate experiments under the extreme conditions characteristic of Earth's deep interior. There nevertheless does exist very detailed information on the creep of polycrystalline solids in general which may provide useful insight on the sort of behavior to be expected. Since such polycrystalline aggregates have many characteristics in common insofar as their ability to “flow” is concerned, it is instructive to consider one specific example in order to fix ideas. To this end, Fig. 1 displays a strain-rate vs. differential stress diagram for fine-grained ice (from Goldsby and Kohlstedt 1997), an important model system that has been studied in great detail because of its practical importance in glaciology. In this log-log plot a straight line is indicative of a power law relationship with the slope determining the



Mantle Viscosity, Fig. 1 The rheology of fine grained ($d = 1$ mm) ice at 250 K temperature, calculated using a particular constitutive equation (Goldsby and Kohlstedt 1997)

so-called stress exponent, the power to which the applied differential stress must be raised to predict the observed strain rate. Inspection of the figure will show that, as the differential stress is reduced for a given polycrystalline grain size, the flow law undergoes a series of discrete transitions. At the highest values of differential stress the stress exponent has a value of 4, indicating that the mechanism supporting the “flow” of the polycrystalline aggregate involves the propagation of “dislocations” through the interiors of the individual crystal grains that make up the aggregate. This process is referred to as involving “dislocation climb.” As the differential stress diminishes the stress exponent first drops to a value near 2, indicating that the mechanism supporting flow has come to involve the propagation of dislocations along the boundaries separating individual grains. This regime is referred to as involving “grain boundary sliding.” With even further reduction of the differential stress, the exponent drops to unity implying that the underlying mechanism involves the diffusion of impurities. In this regime, the stress strain relation describing the flow of the solid is now equivalent to that which holds for a simple fluid like water which we refer to as being “Newtonian” as there is a linear relationship between stress and strain rate. Further experimental data also shows that, as the temperature approaches the melting temperature, the strain-rate increases for a fixed value of the differential stress. For the mantle of the Earth, it is therefore understood that the closer the temperature to the melting temperature of the polycrystalline mantle the lower will be the effective viscosity.

Non-Newtonian Rheology from Seismic Anisotropy

It is also generally understood that at shallow depths, from the surface through the near-surface “lithosphere” in which the viscosity is extremely high, the operative rheology is significantly non-Newtonian. The extent to which this behavior persists to greater depth remains undetermined. That the shallow Earth deforms according to a non-Newtonian stress-strain relation follows from the existence of significant seismic anisotropy in this range of depth. The development of such anisotropy requires nonlinearity of the stress-strain relationship in order that individual crystal grains may be aligned by the ambient shear as occurs in the great polar ice sheets on Greenland and Antarctica. Although direct laboratory measurements on the creep of Earth materials at lithospheric depths where the pressures and temperatures are low enough provide useful confirmation of the seismic observations, they are much less illuminating for the Earth’s deep mantle because of the extreme conditions under which measurements must be made to be relevant.

Natural Experiments

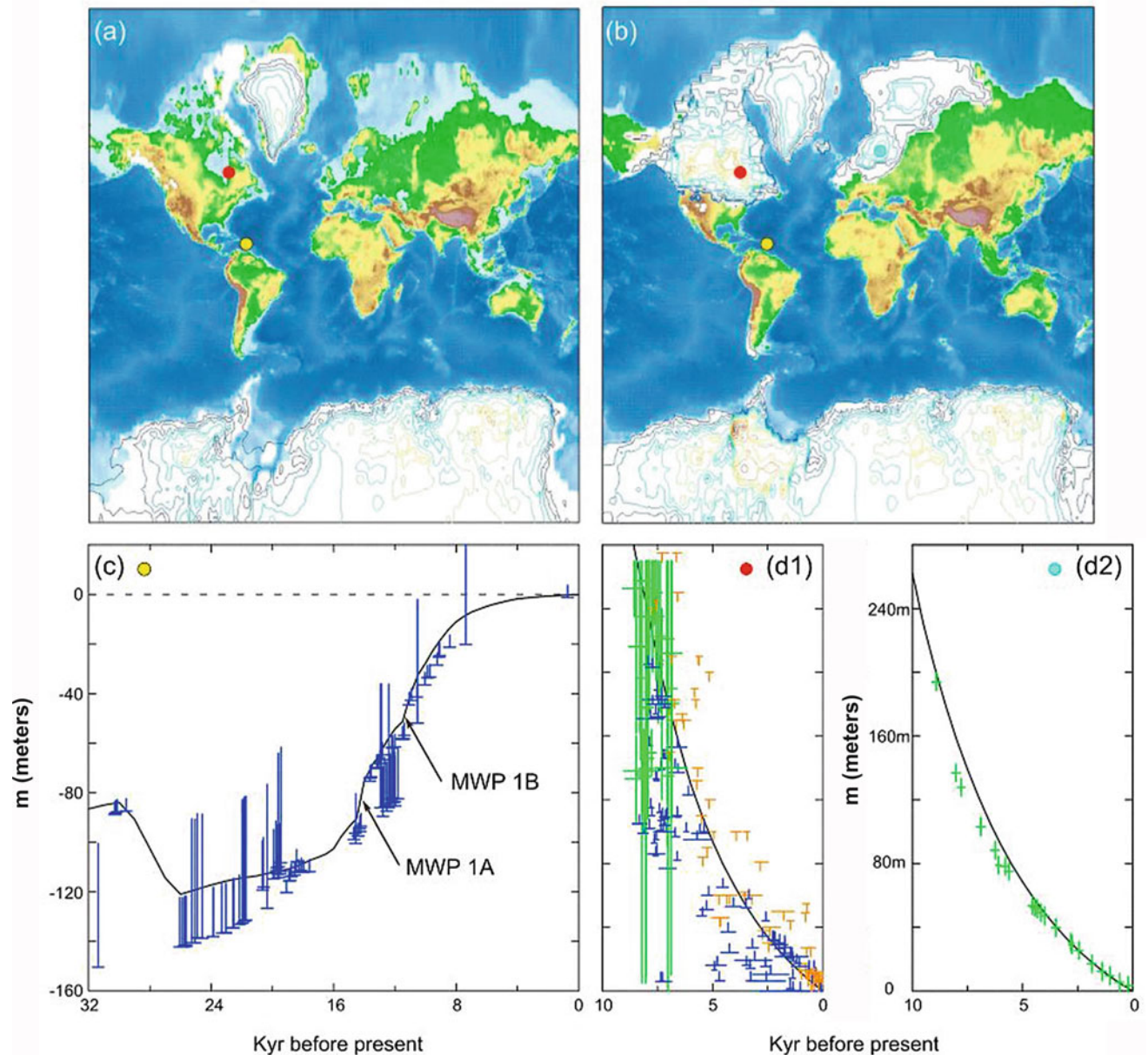
In order to obtain useful estimates of the effective viscosity at depths significantly below the lithosphere, which is on the

order of 40–100 km thick, we are obliged to employ the results of “natural experiments” that the Earth itself has performed throughout its history and for which it has also conveniently remembered the results through the fidelity of the geological recording system. The most important experiment is that associated with the Late Quaternary ice-ages, especially the quasi-periodic sequence of glaciation-deglaciation events that began about 900,000 years before present and during which massive accumulations of land ice began to appear and disappear from the continents, especially at high northern latitudes. Insofar as the inference of deep Earth viscosity is concerned, these events may be thought of as comprising an experiment during which extreme normal stresses (force per unit area) were applied to localized regions of Earth’s surface and thereafter removed when deglaciation subsequently occurred. Because the residence time of the land ice on the surface of the continents was sufficiently long compared to the Maxwell time, the response to the applied normal stresses was such as to induce a response of the loaded viscoelastic mantle that involved significant viscous flow. This suggests the possibility that the geological data that represent the Earth’s record of its response to the time-varying ice load might be interpreted to infer the value of the viscosity that governs the viscous component of the response. Also of interest is the question as to whether the viscosity so inferred is compatible with the requirements of the mantle convection explanation of the surface kinematic observations related to the process of plate tectonics.

Mantle Viscosity from Glacial Isostatic Adjustment

Observational Data Sets

Figure 2, from Peltier et al. (2015) and the earlier work of Peltier (1996), shows examples of critical data related to the Glacial Isostatic Adjustment (GIA) process that include maps of the ice-covered regions of the planet under both modern conditions (a), and the conditions that prevailed at the Last Glacial Maximum (LGM), normally thought to have occurred at approximately 21,000 years before present (b). Figure 2c shows an extremely important data set constraining the history of sea level change relative to the surface of the solid Earth at the island of Barbados in the Caribbean Sea from Peltier and Fairbanks (2006). This record covers the interval of time from LGM to the present and has been produced by the Uranium/Thorium dating of corals, the species *Acropora Palmata* being the most important as it normally dwells within 5 m of the surface of the sea while it remains an open system. These data correlate well with the record of ice-equivalent eustatic (globally averaged) sea level history and demonstrate that sea level has risen at Barbados by approximately 130 m since LGM and that the time of Last Glacial Maximum, as determined by the time at which sea level was lowest at



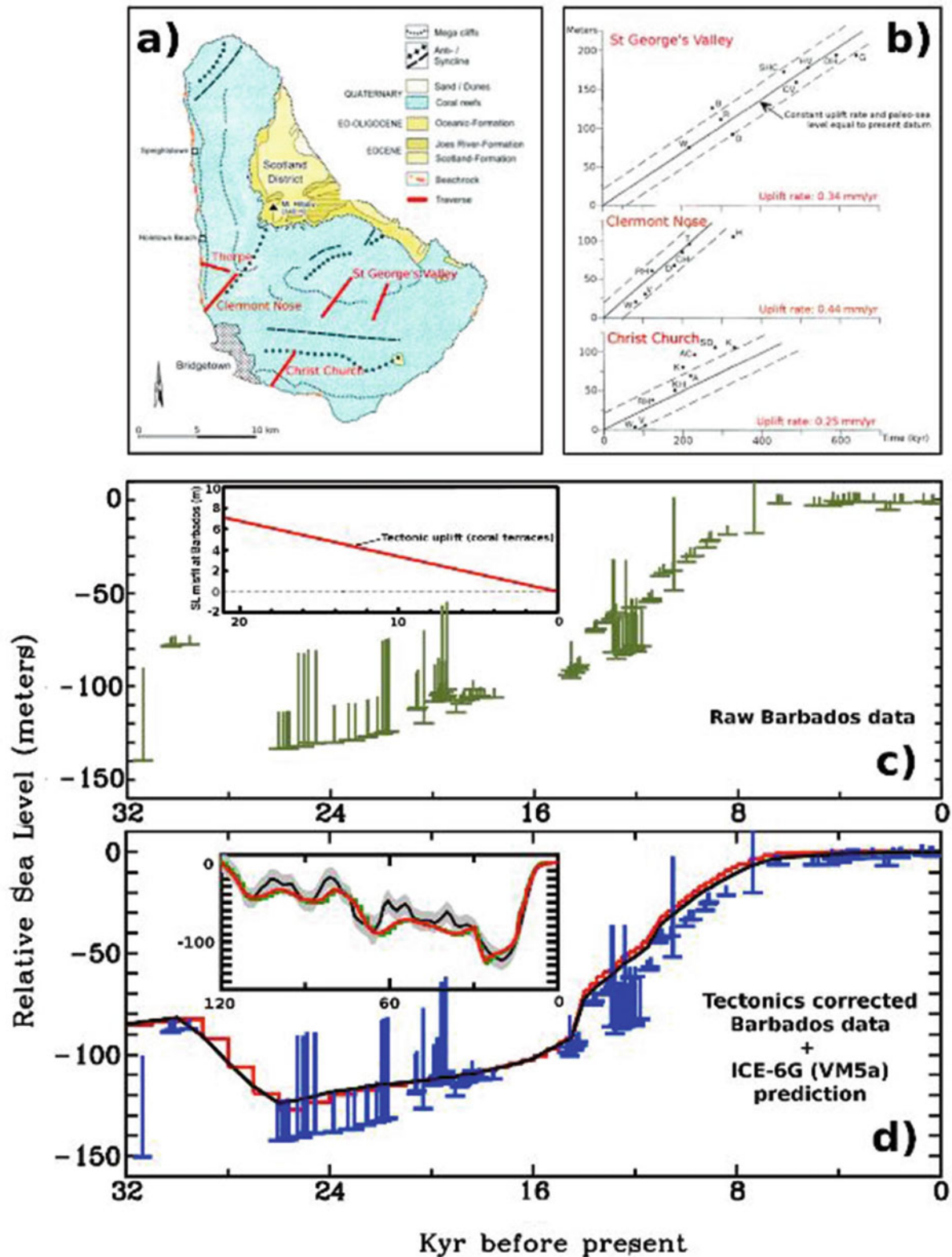
Mantle Viscosity, Fig. 2 (a) Global ice cover and surface topography and bathymetry under modern climate conditions, (b) same as (a) but for Last Glacial Maximum conditions according to the ICE-6G_C (VM5a) model. (c) A fit of the ICE-6G (VM5a) model to the coral based record of

relative sea level history from the island of Barbados. The two major melt water pulses (MWP) are marked. (d) IC-6G_C (VM5a) model fit to the records of Holocene relative sea level history at sites near the centers of rebound of the Laurentide (d1) and Fennoscandian (d2) ice-sheets

Barbados relative to the present level was nearest to 26,000 years before present rather than to 21,000 years before present as has most often been assumed. The Barbados data provide a strong constraint on the total mass of ice that must have been removed from all glaciated regions of the land in order to cause the rise of sea level observed.

As demonstrated in Fig. 3 (also from Peltier et al. 2015), however, which shows in part (a) the locations on Barbados

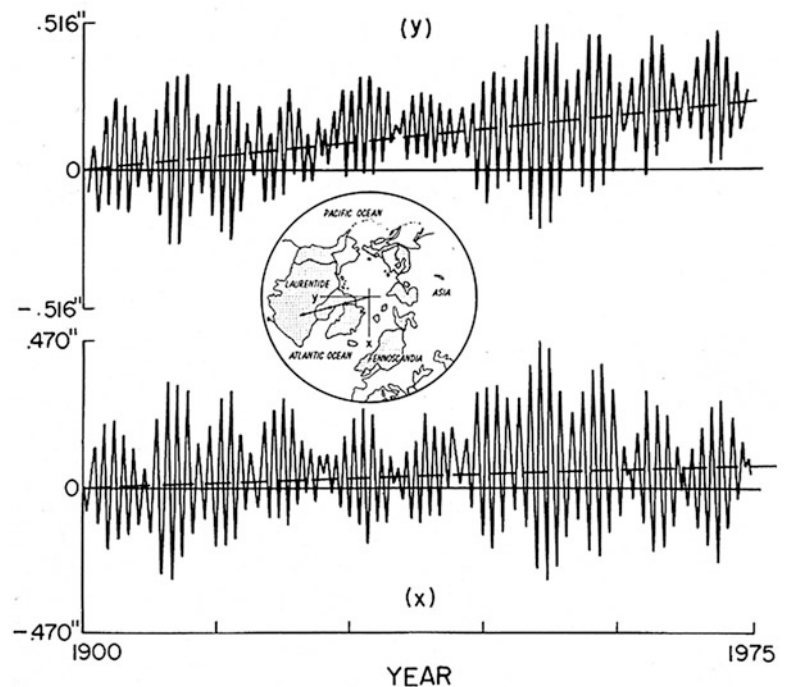
from which coral samples are available which enable a direct measurement of the long timescale tectonic uplift of the island that is occurring above the subducting Caribbean plate; these data demonstrate that this uplift has been occurring at a rate of 0.34 mm/year over the past several hundred thousand years of Earth history as shown in (b). This implies that the Relative Sea Level (RSL) record from Barbados must be corrected as shown on part (c) of the figure by adding to the depth from



Mantle Viscosity, Fig. 3 (a) Shows the three locations on the island of Barbados at which sub-aerial samples of the *Acropora palmata* species of coral are available that may be employed to determine the rate of tectonic uplift of the island that has been occurring over the past several glacial cycles, (b) rates of uplift inferred at each of these locations, (c) depths of

the radio-carbon dated samples adjusted for the inferred rate of uplift of the island, (d) fit of the prediction of the ICE-6G_C (VM5a) model of the glacial isostatic adjustment process to the inferred sea level history at Barbados

Mantle Viscosity, Fig. 4 Time series for x- and y- components of the location of the north pole of rotation of the planet relative to the coordinate system shown in the inset



which dated samples have been raised, the additional depth by which they have been uplifted due to the influence of local tectonics. This correction amounts to 7 m for samples of LGM age. Part (d) of the figure shows the fit of the ICE-6G_C (VM5a) model of the GIA process of Peltier et al. (2015) to the coral observations after the raw data have been corrected for this influence of local tectonics.

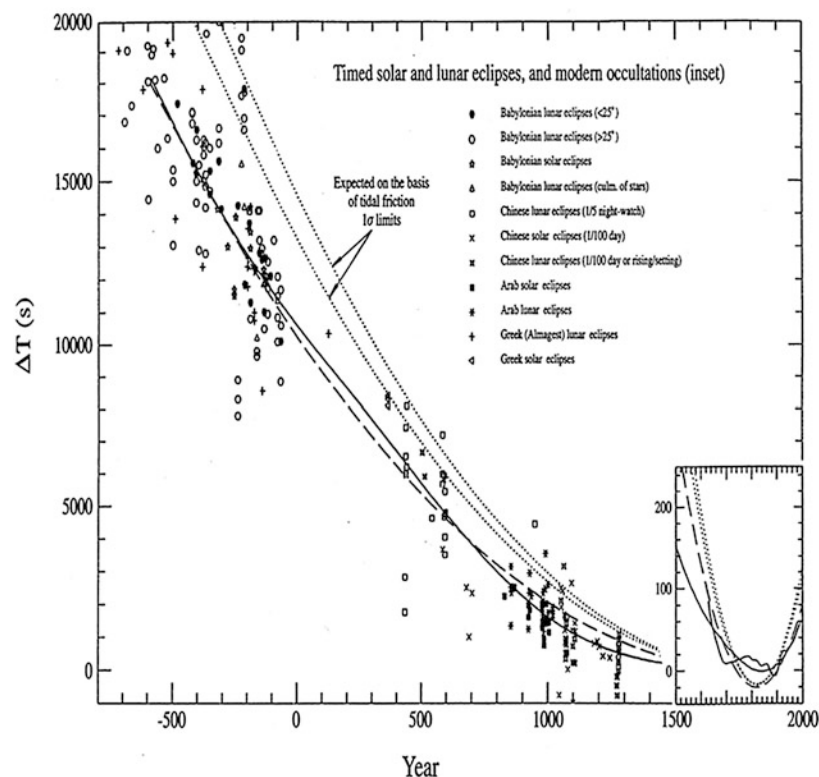
In Fig. 2 (d1,d2) are shown observations of RSL history from sites that were located near the center of the regions in North America and Eurasia, respectively, that were covered by glacial ice at LGM (from Peltier 1998). Whereas the Barbados data record the extent to which sea level was depressed at LGM, these data from the previously ice-covered regions demonstrate that subsequent to the removal of the ice, the land in these regions has been rising out of the sea. This is the process referred to as postglacial rebound of the crust or Glacial Isostatic Adjustment (GIA). The data shown in Fig. 2 (d1, d2) may be quantitatively analyzed to extract information concerning the amplitude and the relaxation time of the characteristically exponential individual sea level curves from such previously ice-covered locations. The relaxation time data inferred on the basis of such data depend most strongly on the mantle viscosity when such records are based upon data taken from the period of time when variations in the initial surface ice load are no longer occurring. Data such as these from hundreds of locations globally constitute one of the primary data sets that may be invoked to constrain the viscosity of Earth's mantle and its depth dependence.

The remaining data related to the GIA process that have come to play an increasingly important role in the inference of

mantle viscosity include those related to the Earth's rotational state. Figure 4 illustrates the first of these additional constraints. It shows the time dependent position of Earth's north pole of rotation in a coordinate system with x- and y-axes centered upon the Conventional International Origin (Vincente and Yumi 1969, 1970). Inspection of the time series in the period between 1900 and 1979, at which time the International Latitude Service which was responsible for maintaining these measurements was discontinued, shows that the measurements are dominated by "beating" at a period very close to 7 years. This is associated with the interference between the 12-month periodic annual wobble of the spin-axis forced by the seasonal variations of the atmosphere and oceans and the 14-month periodic Chandler wobble. The component of these observations that is of interest from the perspective of mantle viscosity measurement, however, consists of the slow secular drift of the pole upon which the oscillatory component is superimposed. These data reveal the secular drift of the pole at a rate near $0.95^\circ/\text{m.y.}$ approximately along the 79°W meridian, i.e., in the direction of present-day Hudson Bay. The rate of this "true polar wander" provides a further constraint upon the viscosity of Earth's mantle.

An additional observational datum that may contribute to the estimation of mantle viscosity consists of the "non-tidal acceleration" of planetary rotation. This datum is illustrated in Fig. 5 where specific times of occurrence of total eclipses of the Sun and Moon are plotted as a function of time spanning the past 2000 years of Earth's history, covering the periods in both Babylon and China during which naked eye astronomers

Mantle Viscosity, Fig. 5 Plot showing variation of time difference between the observed time of a total eclipse of the Sun or the Moon and the predicted time of the event based on the assumption of constant tidal torque over the past 2000 years. The inset (on a scale 25 times expanded) includes the continuous curve derived from the lunar occultations in the period 1620–1995.5 AD. The dashed curve is the best fitting parabola and the solid curve is fitted using cubic splines. (From Stephenson and Morrison 1995)

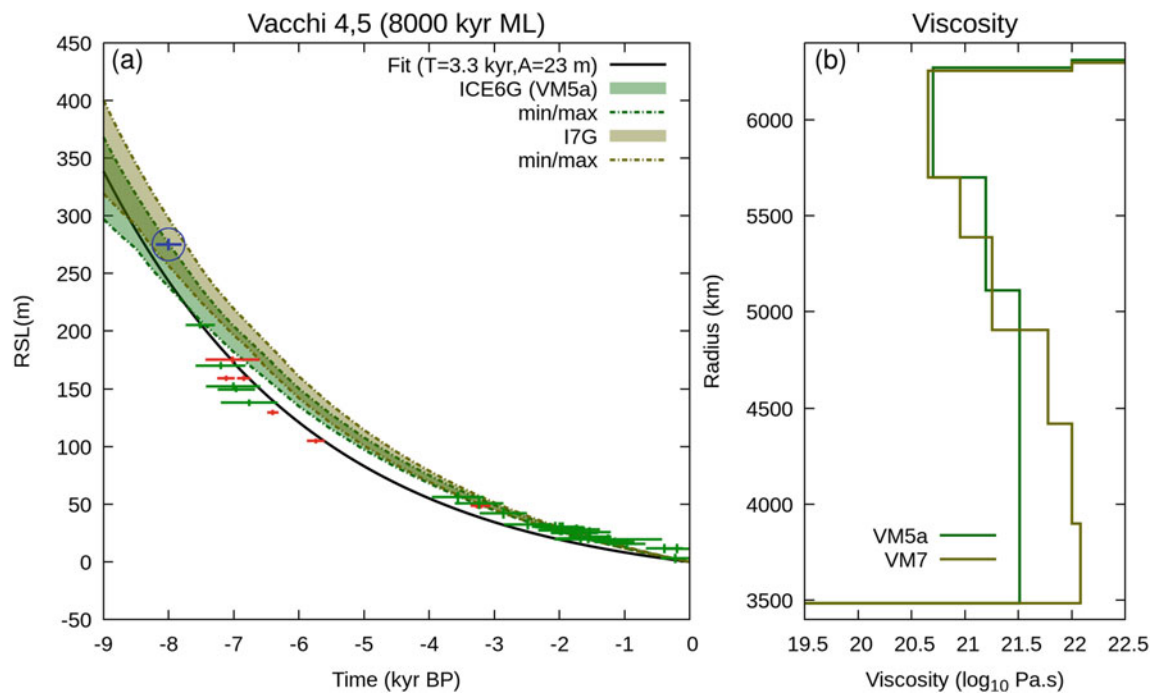


both recorded and attempted to make predictions of the timing of these “epochal” events (from Stephenson and Morrison 1995). These observations are compared with expectations described by two essentially parabolic curves. The curve that does not go through the data is based on the assumption that only tidal friction has contributed to variations in Earth’s rate of rotation over this period and that this has remained equal to its present-day value, a reasonable assumption. Because the predicted times of occurrence of total eclipses of the Sun and Moon do not agree with expectations based on this assumption, some physical process is missing. The nature of the misfit of this model to the data implies that the action of the missing process is such as to somewhat counteract the influence of tidal friction by decreasing the rate at which Earth’s angular velocity would otherwise decrease. This is called the nontidal acceleration of rotation and it too is a consequence of ice-age loading and unloading of the continents. The curve in Fig. 5 that does go through the data provides a precise measurement of this ice age influence and this may also be employed to constrain the viscosity of Earth’s mantle (e.g., see Peltier 2015).

Results of Inversion of the Totality of the Data to Determine Mantle Viscosity Depth Dependence

In the recent past (e.g., Vacchi et al. 2018), a very careful recompilation of the relative sea level data from the center of rebound in North America, which is located in South East Hudson Bay, has been produced. This has enabled a much

more detailed analysis of the relative merits of competing models of the depth dependence of mantle viscosity. An aspect of the data from this region that merits special attention because of its value, is the constraint provided by the so-called Marine Limit, the maximum height above present sea level ever reached by the sea. Figure 6a shows the complete data set from this region which consists of a very large number of Sea Level Index Points (SLIPs) through which the predictions of any viable model of the GIA process must pass, as well as the Marine Limit data point. Superimposed upon these observations are predictions of the ICE-6G_C (VM5a) model of Peltier et al. (2015) and the ICE-7G_NA (VM7) model of Roy and Peltier (2017). The important Marine Limit data point for SE Hudson Bay is constrained by geomorphological evidence to an elevation of 265 ± 15 m with an age of 8.1 ± 0.3 ka. Part (a) of Fig. 6 also shows the best fitting two parameter exponential model to the same data. A significant additional success of these models has been their ability to explain all of the available Global Positioning System (GPS) measurements of the ongoing rates of vertical motion of the crust across the entire North American continent as well as the Gravity Recovery And Climate Experiment (GRACE) observations of the time dependence of the gravitational field over both North America and Fennoscandia (Peltier et al. 2015). These models also fit the previously described earth rotation observations which are especially important in providing resolving power for viscosity in the deepest part of the lower mantle (Peltier 1998). The



Mantle Viscosity, Fig. 6 (a) Relative sea level observations at the south east Hudson Bay location which is the center of postglacial rebound of the crust that occurred following the complete deglaciation of the North American continent. The data consists not only of the rigorously quality controlled Sea Level Index Points (SLIPS) of Vacchi et al. (2018) but also the Marine Limit data point that records the highest

level on the modern landscape that was once at sea level. Comparisons of the fit to these data are shown for both the ICE-6G_C (VM5a) model of Peltier et al. (2015) and the ICE-7G_NA (VM7) model of Roy and Peltier (2015, 2017). (b) illustrates the VM5a and VM7 viscosity models that represent the current state of the art in spherically symmetric models of this physical property of the planet

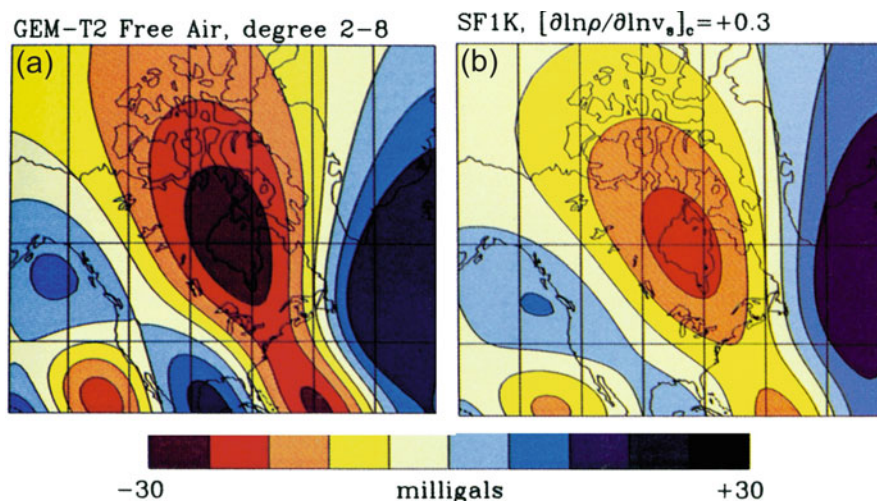
depth dependence of viscosity in the VM5a and VM7 models are compared in Fig. 6b. An important property of the VM7 viscosity model is that it is characterized by a continuous increase with depth in the lower mantle, a result that is in accord with expectations based upon the divergence with increasing depth of the mantle adiabat (to which vigorous convection adjusts the mantle geotherm) and melting curve.

Compatibility of the GIA-Based Viscosity Model with that Required by Models of the Mantle Convection Process

An early effort to determine the viscosity of Earth's deep mantle based on GIA data (e.g., Walcott 1970) appealed to the observed magnitude of the free air gravity anomaly over the Hudson Bay region to argue that the magnitude of this anomaly provided strong evidence that the viscosity of the deep Earth was much higher than implied by the analyses described in the previous sections. This anomaly, based on the GEM-T2 model of the Goddard Space Flight Center, is shown in Fig. 7a. It has a maximum amplitude of approximately -30 milligals when the field is restricted to include only spherical harmonic degrees 2 through 8. When the VM2 viscosity model (the precursor to VM5a which is simply a multilayer

fit to VM2) was used in conjunction with a model of ice thickness history in which the maximum thickness was assumed to be coincident with Hudson Bay as Walcott had assumed, the predicted free air gravity anomaly was found to have a peak amplitude of only -3 milligals. This issue concerning the origin of the free air gravity anomaly, was further analyzed in detail by Pari and Peltier (1996). Their work focused on the viability of the alternative hypothesis that the anomaly was supported by the mantle convection process rather than glacial isostasy. Models of mantle convection incorporating the results of seismic tomography (Forte et al. 1994) and based on internal loading theory were developed and applied in analyses of the nonhydrostatic geoid and free air gravity anomalies that would be supported by this mantle dynamic process. In the context of internal loading theory for an Earth model with spherically symmetric viscosity distribution there are only two distinct parameters that need to be specified. The first of these is the depth dependent mapping factor through which the lateral heterogeneity in seismic shear wave velocity is converted into lateral heterogeneity of density. The second is the depth dependence of mantle viscosity. Results were obtained for three different values of the S-wave velocity to density conversion factor in the subcontinental "tectosphere": -0.3 (positive buoyancy, upwelling flow), 0.0 (neutral buoyancy), and $+0.3$ (negative

Mantle Viscosity, Fig. 7 (a) Observed GEM-T2 degree 2 to 8 non-hydrostatic free-air gravity anomaly over the Hudson Bay region of Canada, and (b) the calculated anomaly based on the predictions of the loading theory of the mantle convection process, for the shear wave speed to density conversion factor for the cratonic root of +0.3. Because the pattern of the predicted free air gravity anomaly is extremely well correlated with the satellite observed pattern it is clear that a slight further increase in the local value of the conversion factor would only further improve the fit



buoyancy, downwelling flow). Of the three scenarios, the last model was shown to provide an excellent fit to the free air gravity anomaly (Fig. 7b). A full 24 milligals of the 30 milligal free air anomalies over Hudson Bay could thereby be explained by the mantle convection hypothesis. It needs to be understood that there has been no attempt to tune the model to achieve a best fit locally. Rather the model has been optimized to achieve a best fit to the entire global gravity field. It seems clear on the basis of these results that the free air gravity anomaly over the Hudson Bay region is essentially entirely due to the mantle convection process. Therefore, the original argument of Walcott (1970) that this anomaly must be due to remaining glacial isostatic disequilibrium and to imply that the viscosity of the mantle of the deep Earth below the seismic discontinuity at 660 km depth must be extremely high, is not supported. It seems rather to be explicable as a consequence of thermal convection in a mantle through which the variation of viscosity is rather modest.

The viscosity profile that delivers the high quality fit to the free air anomaly of the North American continent is shown previously in Fig. 5b where it is referred to as VM5a which is very close to the more recent profile referred to as VM7 in that figure. The quality of this model has been further verified by demonstrating it to accurately explain a further observable related to the GIA process, namely the time dependence of the gravitational field measured over the North American continent by the GRACE space geodetic system. As demonstrated in Peltier et al. (2015) the observed field is characterized by the existence of two extrema in gravity field time dependence, one in the south east Hudson Bay region previously discussed above, and another to the west of Hudson Bay coincident with the so-called Keewatin Dome of the Laurentide Ice Sheet. This is further confirmation of the interpretation above that the free air gravity anomaly over Hudson Bay must have an explanation in terms of the mantle convection process. Both

the time dependence of the planets gravitational field and the continent-wide pattern of vertical crustal motion are beautifully explained by the modern theory of the glacial isostatic adjustment process in which the variation with depth of the effective viscosity of the Earth is rather modest.

Summary

This article has reviewed the main sources of information on the viscosity of the deep Earth. These sources consist of data pertaining to the phenomena of postglacial rebound and mantle convection. In spite of modest differences between the models of mantle viscosity depth dependence derived in distinct ways, the overwhelming evidence is that the depth dependence of viscosity derived on the basis of GIA observations is in close accord with that required by the mantle convection hypothesis explanation of surface plate tectonics and continental drift.

Cross-References

- [Magnetic Anisotropy](#)
- [Mantle Convection](#)
- [Mantle D'' Layer](#)
- [Seismic Tomography](#)

Bibliography

- Forte AM, Woodward RL, Dziewonski AM (1994) Joint inversions of seismic and geodynamic data for models of three-dimensional mantle heterogeneity. *J Geophys Res* 99:21857–21877
- Goldsby DL, Kohlstedt DL (1997) Grain boundary sliding in fine grained ice I. *Scr Mater* 37:1399–1406

- Pari G, Peltier WR (1996) The free-air gravity constraint on sub-continental mantle dynamics. *J Geophys Res* 101:28105–28132
- Peltier WR (1996) Mantle viscosity and ice-age ice-sheet topography. *Science* 273:1359–1364
- Peltier WR (1998) Postglacial variations in the level of the sea: implications for climate dynamics and solid earth geophysics. *Rev Geophys* 36:603–689
- Peltier WR, Fairbanks RG (2006) Global glacial ice volume and last glacial maximum duration from an extended Barbados Sea level record. *Quat Sci Rev* 25:3322–3337
- Peltier WR, Argus DF, Drummond R (2015) Space geodesy constrains ICE-age terminal deglaciation: the global ICE-6G_C (VM5a) model. *J Geophys Res Solid Earth* 120(1):450–487
- Peltier WR (2015) The history of Earth's rotation: impacts of deep earth physics and surface climate variability. In: *Treatise on geophysics*, vol 9, 2nd edn. Elsevier, Oxford, pp 221–279
- Roy K, Peltier WR (2015) Glacial isostatic adjustment, relative sea level history and mantle viscosity: reconciling relative sea level predictions for the US east coast with geological constraints. *Geophys J Int* 201(2):1156–1181
- Roy K, Peltier WR (2017) Space geodetic and water level gauge constraints on continental uplift and tilting over North America: Regional convergence of the ICE-6G_C (VM5a) model. *Geophys J Int* 210:1115–1142
- Stephenson ER, Morrison LV (1995) Long term fluctuations in the Earth's rotation: 700 BC to AD 1990. *Philos Trans R Soc Lond Ser A* 351:165–202
- Vacchi M, Engelhart SE, Nikitina D, Ashe EL, Peltier WR, Roy K, Kopp RE, Horton BP (2018) Post-glacial relative sea level histories along the Eastern Canadian coastline. *Quat Sci Rev* 201:124–146
- Vincente RO, Yumi S (1969) Co-ordinates of the pole (1899–1968) returned to the conventional international origin. *Publ Int Latit Obs Mizusawa* 7:41–50
- Vincente RO, Yumi S (1970) Revised values (1941–1961) of the co-ordinates of the pole referred to the CIO. *Publ Int Latit Obs Mizusawa* 7:109–112
- Walcott RI (1970) Isostatic response to loading of the crust in Canada. *Can J Earth Sci* 7:716–727

Mapping of Magnetic Variations from Space

Nandini Nagarajan

National Geophysical Research Institute, Hyderabad, India

Definition

Magnetic measurements from space, the differences between ground and space measurements, and reduction of satellite-borne measurements are described. Early magnetic measurements in space, the POGO satellite series, with scalar magnetometers provided uniform coverage over the Earth, demonstrated the utility of total intensity measurements that provided the first model of the geomagnetic field, the International Geomagnetic Reference Field (IGRF). First tri-axial

vector measurements were made by MAGSAT Mission that produced IGRF, eliminating ambiguity in the potential field solutions. The Decade of Geopotential Research (1999–2009) wherein concurrent measurements with three satellites with technological improvements were made, expanded the utility and scope of these measurements rapidly. Following this, the present constellation of satellites with gradient measurements further refine studies of geomagnetism. A brief overview of the modelling approaches is presented. In the future, satellite geomagnetism would seek to provide Realtime monitoring, prediction of space weather events, as well as more precise mapping of polar wander, necessary for navigation. Improving measurements, advances in gradient measurements, operability, and cost-effectiveness of missions are some goals.

Measurements from Space

Direct measurements of the magnetic field have been made for a few hundred years, and currently, the field is monitored at approximately 150 geomagnetic observatories on ground. Ground measurements provide reliable temporal signatures of magnetic variation, and reduced to an epoch, produce spatial variations as well. Their spatial distribution is, however, uneven, with large gaps in the oceans, inhospitable terrain, and increasing industrialization. Global surveys required more even coverage and satellite-borne measurements commenced as the technology developed (Olsen et al. 2010; Friis-Christensen et al. 2009). Data obtained from satellites first need to be reduced: by accurately determining the altitude, ellipticity, and position of the satellite, to separate spatial and temporal variation. Reduction of satellite data and comparison with ground measurements presented challenges, since these satellites move at 8 km/s. As a consequence, it is not possible to decide whether an observed magnetic field variation is due to a temporal or spatial change of the field and the signals obtained need to be validated by both time and spatial controls. Another approach is to model the contributions from the Earth's core and crust and well-constrained ionospheric and magnetospheric contributions (Sabaka et al. 2004, 2018).

Efforts to make measurements of the potential field from space, using Low-Earth Orbiting (LEO) satellites started with the launch of OGO (Orbiting Geophysical Observatory) satellite series in the 1960s with an objective to obtain a better understanding of the Earth as a planet and to obtain more data on the Earth-Sun relationship. The LEO satellites attain an initial altitude less than 2000 km and are launched into different orbits based on its purpose. Inclination is an important parameter describing the shape and orientation of a celestial orbit, defined as the angular distance the orbital plane makes with a plane of reference, which is usually the equatorial or ecliptic plane. The LEO satellites having an inclination $>85^\circ$

are called polar LEO satellites. For polar LEO satellites having a perfectly circular orbit, the altitude of the satellite nearly remains constant, whereas for noncircular orbits, its altitude varies depending on the perigee and apogee of the satellite. The polar LEO satellites have an orbital period of 90–100 min, making about 15 orbits in a day, of which half would be in daylight and the other half at night. Except at the poles, the local time of the satellite remains nearly constant in its transit as the satellite moves over nearly constant longitude. With Earth rotating below, the longitude of polar LEO shifts by $\sim 23^\circ$ W after every complete orbit. With this constant shift in longitude every 100 min, on a given day, the polar LEO satellite witnesses nearly the same LT for all 15 day and night passes. The satellite orbit drifts in local time by approximately 1 min each day and hence the satellite visits all LT sectors in nearly 4 months. These attributes of polar LEO satellites make them suitable for the study of geomagnetic processes, since the satellite tracks provide data, covering the whole globe, but each pass remains in constant LT and therefore provides information along longitudes.

POGO Satellite Series

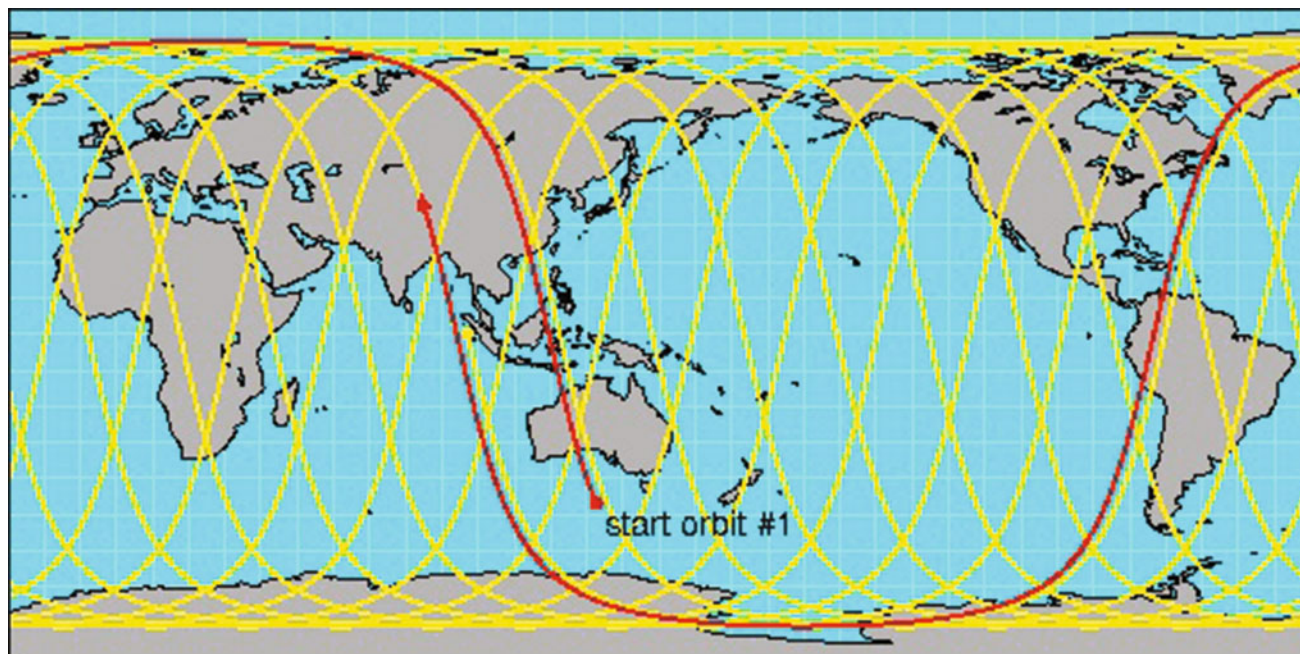
The POGO satellites (Polar Orbiting Geophysical Observatory) were launched in succession in 1965–1971, with near-polar orbits (inclination $>80^\circ$) and apogee heights from 400 to 1500 km. They were equipped with optically pumped rubidium vapor absolute magnetometers and measured only

magnetic field intensity. Intrinsic measurement error of all three satellites is believed to be below 1 nT, but contribution due to position uncertainty results in an effective magnetic error of about 4 nT (Cain and Sweeney 1973; Sabaka et al. 2004; Stockmann et al. 2015).

Though the POGO series of satellites, with scalar magnetometers, in 1960s provided limited coverage, it was enough to prepare the first International Geomagnetic Reference Field (IGRF) model (Cain et al. 1967). The spatial coverage provided by spacecraft was more uniform than any network of observatories, though accuracy was limited. Total magnetic field measurements obtained were used to develop early models of processes affecting the Earth's magnetic field. It was soon realized that such intensity measurements did not provide enough information to recover robust global models of the geomagnetic field. It was established in 1970 as the Backus effect. Backus (1970) proved that for a primarily dipole magnetic field, inversion from intensity data alone is formally nonunique, as two nontrivially different potential fields yield the same intensity everywhere on a spherical surface. It was also demonstrated that the errors of estimation of fields from POGO were linked to this discrepancy.

MAGSAT Mission

MAGSAT, launched after a decade, carried the first vector magnetometers (1979–80) and flew in a near-polar, dawn-



Mapping of Magnetic Variations from Space, Fig. 1 The ground paths of the multiple orbital revolutions during 1 day for a near-polar orbiting satellite. <https://ral.ucar.edu/~djohnson/satellite/coverage.html>

dusk orbit. Scalar and vector data were obtained (Fig. 1). It was equipped with a scalar (cesium) magnetometer, and a tri-axial fluxgate magnetometer which sampled the field at 16 Hz with a resolution of ± 0.5 nT. Attitude was measured using two star-trackers on the spacecraft; transformation of attitude determined by these star trackers to the vector magnetometer located at the tip of the boom was done using a sophisticated optical system. Attitude errors limited the vector data accuracy to about 4 nT rms (Langel et al. 1980; <https://earth.esa.int/web/eoportal/satellite-missions/m/magsat>). Even with a short life of less than a year, MAGSAT provided a data set that resulted in great improvement in the IGRF model of Earth's magnetic field eliminating the Backus effect in magnetic field modelling for the first time. Other achievements using this vector magnetic data set were: detailed lithospheric anomaly maps, delineating more features of the Equatorial electrojet, absolute strength of the magnetospheric ring current, and estimation of F-region ionospheric currents. More significantly, the properties of these vector and scalar data set enabled comprehensive modelling of the different external currents as well as those of internal origin.

Decade of Geopotential Field Research

The Decade of Geopotential Field Research commenced in 1999 with the launch of the Danish satellite Ørsted, nearly 20 years later. CHALLENGING Minisatellite Payload (CHAMP) and the Satélite de Aplicaciones Científicas-C (SAC-C) followed in this international effort to coordinate continuous measurements of the Earth's magnetic field. These satellites had near-polar orbits, lasting 60–90 min, at varying altitudes from 400 to 1500 km. Both Ørsted and CHAMP actively recorded/provided vector data for 10 years (2000–2010), completing 58,277 orbits, far exceeding their designed performance. Extensive tracking by all-sky cameras, tied in with orbital sensing parameters, made it possible to improve time and spatial stamping of data.

Ørsted was fitted with a proton precession Overhauser magnetometer measuring the magnetic field intensity with a sampling rate of 1 Hz and an accuracy < 0.5 nT. A CSC (Compact Spherical Coil) fluxgate vector magnetometer was mounted closely together with a Star Imager. The CSC samples the magnetic field at 100 Hz (burst mode, at polar latitudes) or 25 Hz (normal mode) with a resolution better than 0.1 nT. The measurements are calibrated against the field intensity measured by the scalar magnetometer. After calibration, the agreement between the two magnetometers is better than 0.33 nT rms. Due to attitude errors, the accuracy of the vector components (B_r , B_θ , B_ϕ) is limited to ~ 2 –8 nT (4 nT rms), depending on the component. Initial estimate of mission duration was 14 months, which extended to over 19 years. In

space, the satellite is still healthy and downlinking total intensity data in 2020 (Neubert et al. 2001; <https://directory.eoportal.org/web/eoportal/satellite-missions/o/oersted>)

The Argentinean satellite SAC-C was launched in November 2000, with a copy of the Ørsted boom and payload (but with a US Scalar Helium Magnetometer). SAC-C provided total intensity data, as the vector magnetometer malfunctioned (<https://directory.eoportal.org/web/eoportal/satellite-missions/s/sac-c>).

CHAMP had very similar instrumentation as Ørsted. Differences were: better position and attitude tracking. This enabled reduction of errors and when this system operated over mid and low-latitudes, resolution of ~ 0.2 nT was achieved. Lower orbits of CHAMP produced improved lithospheric field models, new observations of F-region, signatures of night-time currents. The most significant achievement was the unambiguous signatures of induced magnetic fields produced in the oceans (Tyler et al. 2003). CHAMP was also a pilot mission for use of space-borne GPS observations for atmospheric and ionospheric research and applications in weather prediction and space weather monitoring (<https://directory.eoportal.org/web/eoportal/satellite-missions/c-missions/champ>)

Another innovative aspect of Ørsted and CHAMP is that they were both designed to operate over a much more significant period of time. In the case of Ørsted, this is mainly because Ørsted was launched at a higher altitude than Magsat, with the disadvantage of making Ørsted less sensitive to the shortest wavelengths of the field of internal origin. The very different design of CHAMP, reducing weight and cross-section, made it possible to combine the advantages of a long-lasting mission with that of a low altitude orbit.

After calibrating vector magnetometers against the scalar magnetometer, the actual data is used to compute residuals: comparing the two measurements. Despite the improved resolution of magnetometers, errors in attitude and position were factors that resulted in precision of ~ 5 nT for vector measurements. Residuals were minimized in CHAMP measurements, using two all sky cameras and a precision of 2–5 nT was obtained. These satellites improved coverage significantly, overcoming the main limitation of the *Magsat* concept: the sun-synchronous characteristic of its dawn-dusk orbit. Thus, data was obtained in all local times, though each pass was in the same longitude (LT). Availability of two independent systems of magnetometers on board Ørsted and CHAMP provided a check on data quality. The sun-synchronous components of ionospheric origin could be estimated and differentiated. Ørsted and CHAMP provided vector data, concurrently, for several years, and continued to provide scalar data for 10 and 19 years, respectively. SAC-C could only provide scalar data, in 2001–2004. This data was used to validate measurements from the other two satellites and enhanced the accuracy of the combined data set that was

Mapping of Magnetic Variations from Space, Table 1 Details of orbit, duration, and observational capability of the LEO satellites of the past 50 years. (After Olsen and Stolle 2012)

Satellite	Operation	Inclination $^{\circ}$	Altitude	Data	Magnetometers/accuracy
OGO-2 $^{\circ}$	Oct. 1965–Sept. 1967	87	410–1,510 km	Scalar only	Rb vapor, 5 nT
OGO-4	July 1967–Jan. 1969	86	410–910 km	Scalar only	=do=
OGO-6	June 1969–June 1971	82	400–1,100 km	Scalar only	=do=
Magsat $^{\circ}$	Nov. 1979–May 1980	97	325–550 km	Scalar and vector	Cs Vapour, 3-axial fluxgate, 0.5 nT
Oersted $^{\circ}$	Feb. 1999– Till date	97	650–850 km	Scalar and vector	Overhauser, 0.5 nT CSC fluxgate 0.1 nT
CHAMP	July 2000–Sept. 2010	87 $^{\circ}$	260–450 km	Scalar and vector	Overhauser 0.5 nT CSC fluxgate 0.1 nT
SAC-C	Jan. 2001–Dec. 2004	97 $^{\circ}$	698–705 km	Scalar only	Helium 0.5 nT
Swarm A, B, C	2012– till date	88 $^{\circ}$ /87 $^{\circ}$	530/<450 km	Scalar and vector and gradient	Overhauser 0.5 nT CSC fluxgate 0.1 nT

Abbreviations: CHAMP, Challenging Mini-Satellite Payload; OGO, Orbiting Geophysical Observatories; SAC-C, Satellite de Aplicaciones Científico-C

acceptable to model Earth's magnetic field and its processes. After 2004, both Oersted and CHAMP provided only scalar data. A table of the operational details and instrumentation onboard LEO satellites of the past 50 years is provided here (Table 1).

During the Decade of Geopotential Research, the advantages of concurrent measurements by satellites launched in a constellation were demonstrated and this was improved upon with the SWARM Mission that became operational in 2014. The mission comprises of a constellation of three satellites, with two spacecraft to be launched side-by-side at low altitude to measure the East-West gradient of the magnetic field, and a third spacecraft to be launched (with the same launcher) at a slightly higher altitude. The orbital plane of this third spacecraft progressively shifts with respect to the orbital plane of the lower spacecraft pair, to provide data at different local times. Instrumentation that was largely successful in Oersted and CHAMP was deployed.

The constellation of the polar-orbiting trio of satellites was designed to provide north–south gradient information from each spacecraft and unique east–west gradient information from its low-altitude pair of fliers. The orbital planes of the high-altitude flier, known as “Swarm Bravo,” and the low pair, known as “Swarm Alpha” and “Swarm Charlie,” simultaneously sweep out different local times for improved determination of time-varying external fields. (Sabaka et al. 2018: <https://earth.esa.int/web/guest/missions/esa-co-missions/swarm/mission-fact-sheet>)

Modeling of Satellite Measurements

Satellite based measurements of the Earth's magnetic field, both scalar and vector, provided greater spatial coverage than earlier observatory records and ship tracks. It was possible to produce models of the potential field with far better accuracy than before. The generation of large

amounts of data in continuous sequence also posed challenges in estimating time and spatial stamping to preserve acceptable accuracy of measurement. Further, the volume of data required faster and multiple approaches to data reduction and modelling. Modelling methods have kept pace with each improvement in satellite data acquisition from the early satellite missions.

POGO satellite series resulted in some significant advances in estimation of the Earth's magnetic field: the first International Geomagnetic Reference Field (IGRF) (Cain et al. 1967), first global magnetic anomaly map, and observations of external field variations and offered some crucial constraints for reconstructing the core field evolution over the 1965–1971 time period (Regan et al. 1975; Sabaka et al. 2004). *POGO* data have contributed to the definition of lithospheric fields when combined with more recent missions.

Data collected by MAGSAT, in a short duration, contributed to computation of the first higher degree and order models of the field of internal origin, up to order 13, and demonstrated the separation of higher order fields of crustal origin. There were further improvements to lithospheric anomaly maps. In situ observations by MAGSAT also contributed significantly to studies of ionospheric and magnetospheric currents: by observation of meridional currents in the ionosphere, modelling of the Equatorial Electrojet, and estimation of absolute strength of magnetospheric currents. These steps laid the basis for jointly modeling the fields of internal and external origin,

Extensive developments in computational methods, cross-checking by independent teams, and upgrades in data processing provided a greatly improved data set, from the concerted collaborative efforts. The combined data sets of this Decade of Geopotential Research amounted to about 20,000 passes of satellite over any point on the Earth between latitudes $\pm 60^{\circ}$. The volume

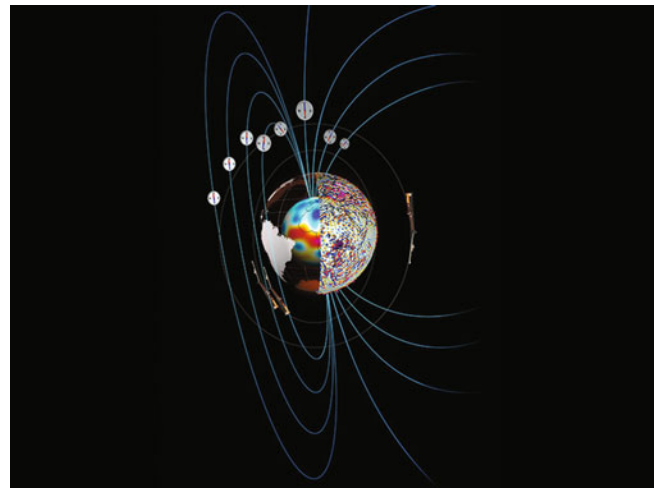
of data was huge, including both scalar and vector data. Concurrent data from CHAMP and Oersted provided sufficient quality measurements to model the main magnetic field using quiet night time values, quiet daytime variations due to ionospheric currents, and disturbance currents. First models of induction in oceans were produced from this data set.

Analyzing magnetic data from LEO satellites requires some care, in particular, because they do not orbit in the neutral atmosphere, but in the ionosphere. Some processes that are easily distinguished as being of external origin (ionosphere) and sun-synchronous systems have been characterized from observatory data. Model parameters of these sources were used in preliminary analysis of satellite data. But the same sources are seen as signatures of internal fields by satellites. Separation of these effects, from accurate, vector satellite data with repeated coverage, was achieved with Comprehensive Modeling. Different approaches are: provide a mathematical description of the additional fields, of varying complexity, whereby potential fields and the toroidal fields produced in the upper atmosphere are estimated. This approach, known as the “Comprehensive Modeling” (CM) approach (Sabaka et al. 2004), also makes it possible to take the temporal variations of the various fields into account, despite the fact that the satellites keep on moving.

The further improved approach, using parametrized models of magnetic variations (from ionospheric and magnetospheric sources), together with models of main field and secular variation evolved. This approach – Comprehensive Modeling – was successfully used to extract diurnal variations and effects of well-defined current systems, equatorial and polar electrojets, partial ring currents, etc. Further advances in modelling have been developed and applied to SWARM data, called Comprehensive Inversion (CI) (Sabaka et al. 2018).

Alternately, satellite data at specific times, eliminating atmospheric processes, is extracted to recover the fields of internal origin (main field and secular variation) This method has been successfully applied to data from the SWARM mission, using estimated gradients from two satellites orbiting close together.

Further improvements in modelling made possible with the combination of data obtained at different local times by the SWARM constellation, also allow for an improved determination of the spatiotemporal structure of large-scale external fields, especially of magnetospheric origins. Further, the accurate recovery of time-series of external variations, also makes it possible to separate high-frequency internal signals that are crucial for probing the mantle conductivity structure. Results of such attempts are in publication and further inferences of the internal structure of the Earth will emerge (Fig. 2).



Mapping of Magnetic Variations from Space, Fig. 2 Configuration of SWARM constellation and its sampling of the Earth's magnetic field. (Credit: ESA-DTU.medialabs)

The Future

Continuing evolution of space-borne geophysical measurements have resulted in more accurate and larger volumes of data, over longer durations. Methods of data analysis, comprehensive modelling, and inversion have differentiated between signals from different electromagnetic processes and sources, both external and internal. Mapping of the Earth's magnetic field is now a continuing process with coordinated global initiatives, providing more precise location of the magnetic poles, monitoring space weather, and discerning signatures of both internal and external processes.

Future programs would look to maintain desired accuracy and consistency of data. A significant parameter that has been evaluated by SWARM is measurement of both east-west and north-south gradients. Looking at the analogy between magnetic and gravity fields, the mission concept of GOCE already measures the full gravity gradient tensor of all second space derivatives of the gravity potential (Drinkwater et al. 2003). Aiming at measuring the analogous magnetic gradient tensor (nine component tensor of all three spatial derivatives of all three components of the magnetic field) would thus be a natural next step. In addition to an enhanced high-resolution mapping of the lithospheric magnetic field, such gradient measurements would in fact also allow for an improved in-situ determination of the electric current density in the near-Earth environment. Such an experiment has been conducted with magnetic data taken by ESA's four satellite Cluster constellation mission (Balogh et al. 1997). Constellation flying is, however, very demanding in terms of satellite control and therefore mission energy consumption. An alternative would be with free-flying satellites at varying heights and separations by which it would still be possible to obtain

gradient information at some points in space and time (as in the case of Oersted, SAC-C and CHAMP). A combination of a few very cheap polar orbiting scalar-only satellites augmented by a low-inclination satellite measuring the field vector could be a good and relatively cheap alternative.

Summary

The need for more uniform spatial coverage of coverage magnetic measurements gave rise to the development of satellite-based magnetometry in the 1960s. Several challenges to reducing space-borne measurements in a spatiotemporal frame, with sufficient precision, were overcome, with improving instrumentation and computational methods. The first International Geomagnetic reference field (IGRF) was prepared with total intensity measurements in space combined with ground measurements. Non-uniqueness of this model was later overcome by using both total field and vector magnetic measurements. Improvements in performance of satellites, and magnetometry and tracking systems, have provided a wealth of data by efforts of concerted global initiatives. This resulted in the production of more detailed magnetic anomaly models, distinguishing between external solar-terrestrial effects and internal contribution from the Earth's core, mantle, and crust. The present programs are set to closely monitor changes in the Earth's magnetic field with constellations of satellites, aiming to obtain gradient measurements that would further refine model accuracies.

Cross-References

- [Geomagnetic Field, IGRF](#)
- [Gravity Field, Temporal Variations from Space Techniques](#)
- [Lithospheric Magnetic Anomalies from Satellite Data](#)
- [Magnetic Anomaly Map, Global](#)
- [Magnetic Gradiometry](#)
- [Magnetic Methods, Principles](#)
- [Magnetic Modeling, Theory, and Computation](#)

Bibliography

- Backus GE (1970) Non-uniqueness of the external geomagnetic field determined by surface intensity measurements. *J Geophys Res* 75(31):6339–6341
- Balogh A, Dunlop MW, Cowley SWH et al (1997) The cluster magnetic field investigation. *Space Sci Rev* 79:65–91. <https://doi.org/10.1023/A:1004970907748>
- Cain JC, Sweeney RE (1973) The POGO data. *J Atmos Terr Phys* 35:1231–1247
- Cain JC, Hendricks SJ, Langel RA, Hudson WV (1967) A proposed model for the international geomagnetic reference field-1965. *J Geomagn Geoelectr* 19:335–355

- Drinkwater R, Floberghagen R, Haagmans R, Muzi D, Popescu A (2003) GOCE: ESA's first earth explorer core mission. *Space Sci Rev* 108(1):419–432
- Friis-Christensen E, Lühr H, Hulot G, Haagmans R, Purucker M (2009) Geomagnetic research from space. *Eos Trans AGU* 90(25):213–215 <https://directory.eoportal.org/web/eoportal/satellite-missions/c-missions/champ>
- <https://directory.eoportal.org/web/eoportal/satellite-missions/o/oersted>
- <https://directory.eoportal.org/web/eoportal/satellite-missions/s/sac-c>
- <https://earth.esa.int/web/eoportal/satellite-missions/m/magsat>
- <https://earth.esa.int/web/guest/missions/esa-eo-missions/swarm/mission-fact-sheet>
- Langel RA, Mead GD, Lancaster ER, Estes RH, Fabiano EB (1980) Initial geomagnetic field model from Magsat vector data. *Geophys Res Lett* 7:793–796
- Neubert TT, Mandea M, Hulot G, von Frese R, Primdahl F, Joergensen JL, Friis-Christensen E, Stauning P, Olsen N, Risbo T (2001) Ørsted satellite captures high-precision geomagnetic field data. *Eos Trans AGU* 82(7):81–88
- Olsen N, Stolle C (2012) Satellite geomagnetism. *Annu Rev Earth Planet Sci* 40:441–465. <https://doi.org/10.1146/annurev-earth-042711-105540>
- Olsen N, Hulot G, Sabaka TJ (2010) Measuring the Earth's magnetic field from space: concepts of past, present and future missions. *Space Sci Rev* 155:65–93. <https://doi.org/10.1007/s11214-010-9676-5>
- Regan RD, Cain JC, Davis WM (1975) A global magnetic anomaly map. *J Geophys Res* 80:794–802
- Sabaka TJ, Olsen N, Purucker ME (2004) Extending comprehensive models of the Earth's magnetic field with Ørsted and CHAMP data. *Geophys J Int* 159:521–547. <https://doi.org/10.1111/j.1365-246X.2004>
- Sabaka TJ, Clausen LT, Olsen N, Finlay CC (2018) A comprehensive model of Earth's magnetic field determined from 4 years of Swarm satellite observations. *Earth Planets Space* 70:130. <https://doi.org/10.1186/s40623-018-0896-3>
- Stockmann R, Christiansen F, Olsen N et al (2015) POGO satellite orbit corrections: an opportunity to improve the quality of the geomagnetic field measurements? *Earth Planets Space* 67:102. <https://doi.org/10.1186/s40623-015-0254-7>
- Tyler RH, Maus S, Luhr H (2003) Satellite observations of magnetic field due to ocean tidal flow. *Science* 299:239–241

Microcontinents

Carmen Gaina¹ and Joanne Whittaker²

¹Department of Geosciences, CEED – Centre for Earth Evolution and Dynamics, University of Oslo, Oslo, Norway

²Institute for Marine and Antarctic Studies (IMAS), University of Tasmania, Hobart, TAS, Australia

Definition

Microcontinents are tectonic blocks of continental origin that were detached from continental margins and isolated by oceanic lithosphere. Relocation of plate boundaries, the proximity of mantle plumes, wrench tectonics, and inherited lithospheric heterogeneities play crucial roles in the formation of microcontinents. In the geological record, carving of microcontinents from continental margins, and their transport

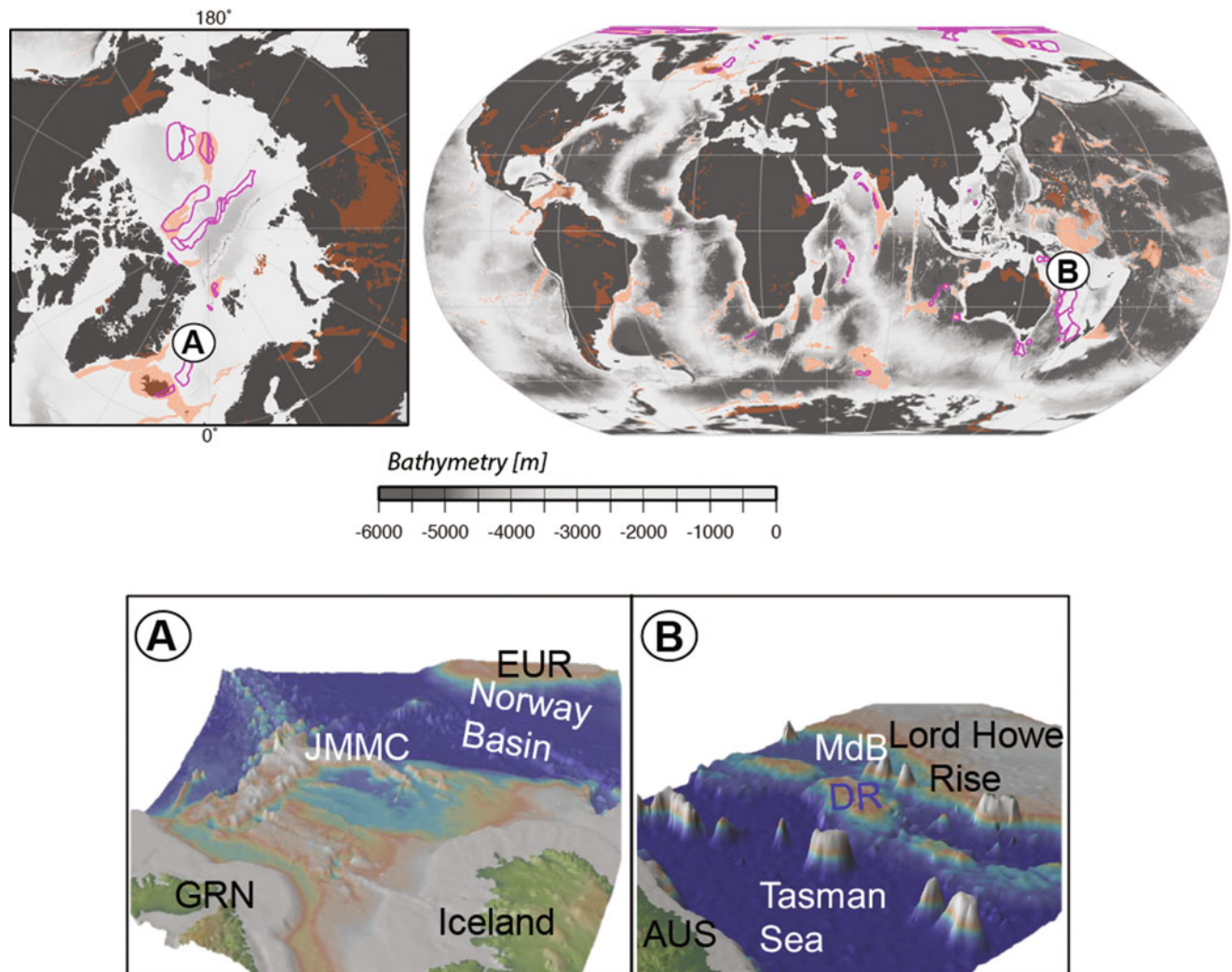
and accretion through seafloor spreading and subduction, has constituted an important mechanism for continental lithosphere redistribution. At the same time, microcontinents have likely played a significant role in paleogeography, paleo-ocean circulation, and biogeography.

Introduction

Continents are large masses of rocks with structures and composition that reflects their formation origin (island arcs, volcanism, etc.) and evolution during more than 4 billion years of Earth history. With the onset of plate tectonics 3 billion years ago, ancient continental masses started to be

broken-up or amalgamated with new continental material, which mainly formed in island arc settings. In 1929, Wegener pointed out that the granitic Seychelles Bank in the Indian Ocean (Fig. 1) is a chip of the larger, adjacent continental blocks. He considered the bank to be composed of the more plastic material of the deeper “sial” layers extracted during the rift process (Wegener 1924).

The name “microcontinent” was brought into wide use by Heezen and Tharp (1965) who applied it to various up-standing bathymetric features in the Indian and Atlantic Oceans. They considered any bathymetrically-defined oceanic plateaus and ridges as “microcontinents” regardless of their petrological composition. In the landmark paper “Microcontinents and their significance,” Scrutton (1976) proposed



Microcontinents, Fig. 1 Present-day location of microcontinents (magenta outline) in the Arctic and North Atlantic (upper left) and in all oceans (upper right). Lower panels show 3D bathymetric images of selected microcontinents: (a) Jan Mayen microcontinent (JMMC) in the

NE Atlantic and (b) SW Pacific Ocean (Tasman Sea). The global bathymetry grid is GEBCO 2014. Large igneous provinces-LIPs are shown in light orange. AUS-stands for Australia, DR for Dampier Ridge, EUR for Eurasia, GRN for Greenland, and MdB for Middleton Basin

several criteria for identifying microcontinents: (1) petrological (rocks of continental nature surrounded of oceanic lithosphere); (2) tectonic (the block should have been displaced horizontally relative to the main continent); and (3) morphological (the size of the block has to be significantly smaller than the originating continent).

Microcontinents received increased attention in the last two decades with many studies tackling their structure, formation, and evolution (e.g., Gaina et al. 2009; Whittaker et al. 2016). Today, 29.1% of Earth's lithosphere is continental in nature, but this does not consider additional continental crust that is scattered through the world's oceans as microcontinents (Fig. 1).

Examples of Present-Day Microcontinents

The Jan Mayen Microcontinent in the North Atlantic Ocean

This submerged microcontinent that occupies the central NE Atlantic (Figs. 1 and 2) was already described in the 1970s and 1980s by early scholars who studied the structure and evolution of North Atlantic. Early descriptions focused on its northern part, the Jan Mayen Ridge, a steep-flanked bathymetric horst structure that extends south from Jan Mayen Island. More recent studies based on geophysical data have mapped in more detail the extent of the microcontinent (Gaina et al. 2009), and it is now accepted that it encompasses a number of distinct, structurally controlled tectonic features that were formed by a succession of tectonic and volcanic events. In total, the Jan Mayen microcontinent is 400–450 km long, and varies in width from 100 km in the north to 310 km in the south. However, it has been postulated that the microcontinent may extend further south, into the Iceland plateau and eastern Iceland, as geochemical composition of igneous rocks in SE Iceland point to continental contamination from a source that was preserved under the thick volcanic plateau (Torsvik et al. 2015).

Until very recent, its continental crustal nature was inferred mainly from geophysical data, but now Polteau et al. (2019) describes rock samples recovered from the southern Jan Mayen microcontinent escarpment with ages from Permian/Triassic to Eocene. These rocks also include igneous samples related to the volcanism that accompanied the Early Eocene breakup from the western Eurasian margin. The subsequent isolation of the microcontinent from the western Greenland margin was gradual, with numerous plate boundary relocations to the south of Jan Mayen that led to the formation of its southern highly extended continental crust (Fig. 2). The block was affected by volcanism several times, and its eastern margin was deformed while the microcontinent was rotated to accommodate fan-shaped seafloor spreading in the adjacent Norway Basin.

The Batavia and Gulden Draak Microcontinents in the Eastern Indian Ocean

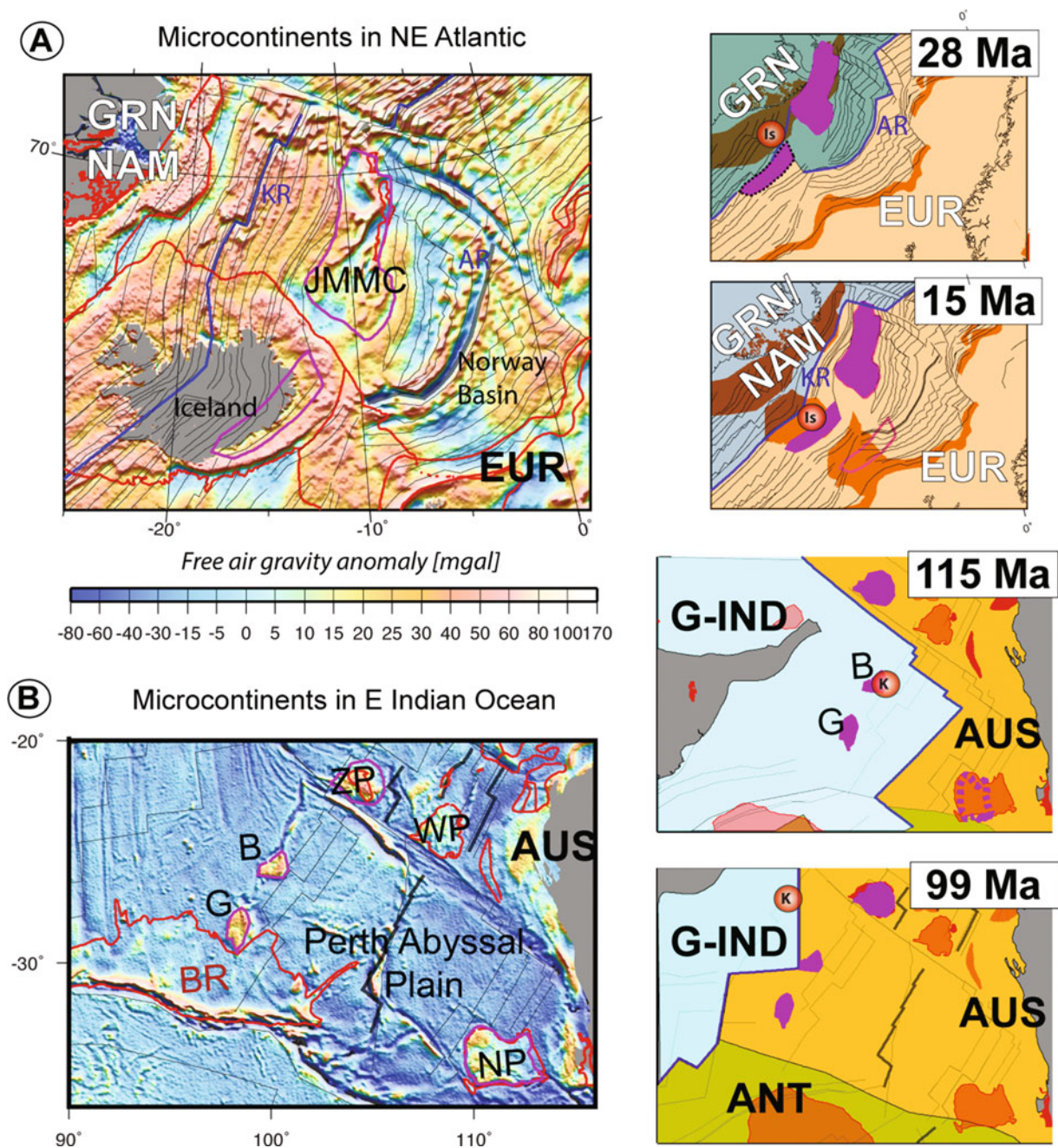
These microcontinents were discovered in 2011 in the historically sparsely sampled eastern Indian Ocean at the western boundary of the Perth Abyssal Plain (Fig. 1b; Gardner et al. 2015). The Batavia and Gulden Draak microcontinents are 100–200 km in width and length and rise steeply from the abyssal seafloor (~5000 mbsl) to 1500–2000 mbsl. Based on relatively sparse sampling, these two microcontinents are predominantly continental material, with only limited evidence for volcanism. They are composed of granulite facies basement, including pelitic paragneiss and mafic orthogneiss, with inferred Cambrian granite intrusions (Gardner et al. 2015). These continental basement rocks reveal that the microcontinents shared a tectonic history with India, Australia, and Antarctica before the disintegration of East Gondwana. The Batavia and Gulden Draak microcontinents are valuable as the only surviving remnants of the now-destroyed, or highly deformed, Greater India margin.

The microcontinents, as part of Greater India, initially separated from Australia/Antarctica during the Early Cretaceous (~137–136 Ma), followed by formation of ~1600 km of seafloor in the Perth Abyssal Plain over ~36 million years. Isolation of the microcontinents was relatively rapid, occurring through a large-scale relocation of the Perth Abyssal Plain spreading ridge, calving the microcontinents from Greater India 101–104 Myr ago (Whittaker et al. 2016). The calving was primarily driven by a regional/global plate tectonic reorganization event that initiated 105 Myr ago. The reorganization caused a significant change in the relative motion between India and Australia, which initiated transpression along a ~1500 km transform section of the plate boundary, and ultimately triggered the relocation of the plate boundary. The ridge jumped into a region of the Greater India passive margin that had been strongly modified and weakened by ~136–95 Ma volcanism from the Kerguelen mantle plume (Whittaker et al. 2016). This volcanism is widespread across the once conjugate Indian, Australian, Antarctic, and the intervening Indian oceanic lithosphere, including 116.6 Ma volcanism at the northern tip of the Gulden Draak microcontinent.

Formation Mechanisms of Microcontinents

Tectonic forces, inherited lithospheric heterogeneities, and the proximity of mantle plumes may play crucial roles in the formation of microcontinents. A large part of the identified in-situ microcontinents were affected by magmatic activity during continental break-up and sometimes post break-up. Proposed microcontinent formation mechanisms include:

1. The isolation of a small continental piece from a relatively young (less than 25 Myrs old) continental margin requires the proximity to a mantle plume which will



Microcontinents, Fig. 2 Microcontinents (magenta outline or fill) in central NE Atlantic (a) and (b) Perth Abyssal Plain (E Indian Ocean). Present-day maps show free air gravity anomaly in the background. Several submerged plateaus west of Australia (AUS) may be still attached to the main continent by extended continental crust, like the Naturaliste Plateau (NP) and the Wallaby Plateau (WP), whereas the Broken Ridge (BR) is part of a Cretaceous Large Igneous Province. Panels on the right show tectonic reconstructions indicating plate boundary relocations and microcontinent formation. Thin gray and brown lines

are isochrons based on magnetic anomalies showing the extent of oceanic lithosphere. Blue lines are active mid-ocean ridges. Thick grey lines are extinct spreading ridges. The locations of Iceland (Is) and Kerguelen hotspots (K) are shown by the pale red circles. Abbreviations: ANT – Antarctica, AR – Aegir Ridge, B – Batavia, G – Gulden Draak, G-IND – Greater India, KR – Kolbeinsey Ridge, NAM – North America and ZP – Zenith Plateau. Other abbreviations and LIPS polygons are the same as in Fig. 1

trigger one or multiple mid-ocean ridge relocations (Müller et al. 2001). This mechanism explains the volcanism and oceanic lithospheric asymmetries associated with most microcontinents.

2. Changing external plate boundary forces triggering tectonic plate relocations leading to calving of continental lithosphere, and therefore, microcontinent formation (e.g., Whittaker et al. 2016; Gaina et al. 2009).

3. Oblique rifting and strike-slip motion (or wrench tectonics) have been recognized as an efficient way to weaken continental margins and facilitate the isolation of microcontinents (Nemcok et al. 2016). Lithospheric underplating and inherited lithospheric heterogeneities facilitate continental break-up that leads to microcontinent formation when plate boundaries relocate.

Several microcontinents are now trapped in the vicinity of complex subduction zones, like in the Scotia Sea, SE Asia, and Mediterranean Sea, but the role of slab pull or slab roll-back in continental margin break-up and possible microcontinent formation awaits future research.

Summary and Conclusions

At present, microcontinents secluded in the world's oceans may represent 3–4% of the entire continental crust. Extensional and strike-slip motions facilitate continental break-up, and subsequent plate boundary relocations, sometimes due to mantle plume proximity, lead to the isolation of small continental slivers within oceanic lithosphere. Magmatic processes may later affect the microcontinents, and sometimes, they are completely hidden within Large Igneous Provinces. When microcontinents travel on subducting plates, they are subjected to subduction or accretion beneath or onto the upper plate, whose structure and dynamics may be affected both at the collision time and later, depending on the size and shape of the microcontinent. In geological time, microcontinents play important roles for ocean circulation, related climatic events, and paleobiodiversity.

Cross-References

- Continental Crustal Structure
- Mantle Plumes

- Plate-Driving Forces
- Plates and Paleoreconstructions
- Seafloor Spreading

Bibliography

- Gaina C, Gernigon L, Ball P (2009) Palaeocene-recent plate boundaries in the NE Atlantic and the formation of the Jan Mayen microcontinent. *J Geol Soc* 166(4):601–616
- Gardner RL, Daczko NR, Halpin JA, Whittaker JM (2015) Discovery of a microcontinent (Gulden Draak Knoll) offshore Western Australia: implications for East Gondwana reconstructions. *Gondwana Res* 28(3):1019–1031
- Heezen BC, Tharp M (1965) Horizontal displacements in earth's crust. 8. Tectonic fabric of Atlantic and Indian Oceans and continental drift. *Philos Trans R Soc S-A* 258(1088):90. <https://doi.org/10.1098/rsta.1965.0024>
- Müller RD, Gaina C, Roest WR, Hansen DL (2001) A recipe for microcontinent formation. *Geology* 29(3):203–206
- Nemcok M, Sinha ST, Dore AG, Lundin ER, Mascle JM, Ryba S (2016) Mechanisms of microcontinent release associated with wrenching-involved continental break-up; a review. In: *Transform margins: development, controls and petroleum systems*, London, pp 323–359. <https://doi.org/10.1144/SP431.14>
- Polteau S, Mazzini A, Hansen G, Planke S, Jerram DA, Millett J, Abdelmalak MM, Blischke A, Myklebust R (2019) The pre-breakup stratigraphy and petroleum system of the southern Jan Mayen ridge revealed by seafloor sampling. *Tectonophysics* 760:152–164. <https://doi.org/10.1016/j.tecto.2018.04.016>
- Scrutton RA (1976) Microcontinents and their significance. In: Drake CL (ed) *Geodynamics progress and prospects*. American Geophysical Union, Washington, DC, pp 177–190
- Torsvik TH, Amundsen HE, Trønnes RG, Doubrovine PV, Gaina C, Kuszniir NJ, Steinberger B, Corfu F, Ashwal LD, Griffin WL, Werner SC (2015) Continental crust beneath southeast Iceland. *Proc Nat Acad Sci* 112(15):E1818–E1827
- Wegener A (1924) *The origin of continents and oceans*. Methuen, London. 220 p
- Whittaker JM, Williams SE, Halpin JA, Wildd TJ, Stilwell JD, Jourdan F, Daczko NR (2016) Eastern Indian Ocean microcontinent formation driven by plate motion changes. *Earth Planet Sci Lett* 454:203–212. <https://doi.org/10.1016/j.epsl.2016.09.019>

Numerical Methods for Flow in Fractured Porous Media

Luca Formaggia¹, Alessio Fumagalli² and Anna Scotti²

¹MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy

²Department of Mathematics, Politecnico di Milano, Milan, Italy

Definition

Porous medium: A material containing voids (pores), whose size is small compared to the size of the sample.

Fracture: a break in a material with a small thickness compared to its global extension.

Fracture network: a network composed of several intersecting fractures.

Introduction

Fractures are ubiquitous in porous media. Here, with the term fracture we denote a void in the porous material that has the following characteristics:

- (i) One of its dimensions, the aperture, is orders of magnitude smaller than the other dimensions and the size of the domain of interest, but still large compared to pore size. We will indicate with extension the size of a fracture in the directions orthogonal to the aperture. The extension of fractures in a network has a distribution that is usually assumed to be governed by a power law, which implies the presence of a large variation of space scales. See Fig. 1.
- (ii) Fractures may be either open or infilled by a material, whose physical characteristics may be strongly different from those of the surrounding material.

- (iii) Fractures usually form networks, called fracture networks, often highly connected.

With the previous definitions one can consider as fractures, depending on the scale of interest, different objects such as: tectonic faults at the scale of sedimentary basins, cracks in glaciers, as well as fractures in concrete or in rocks.

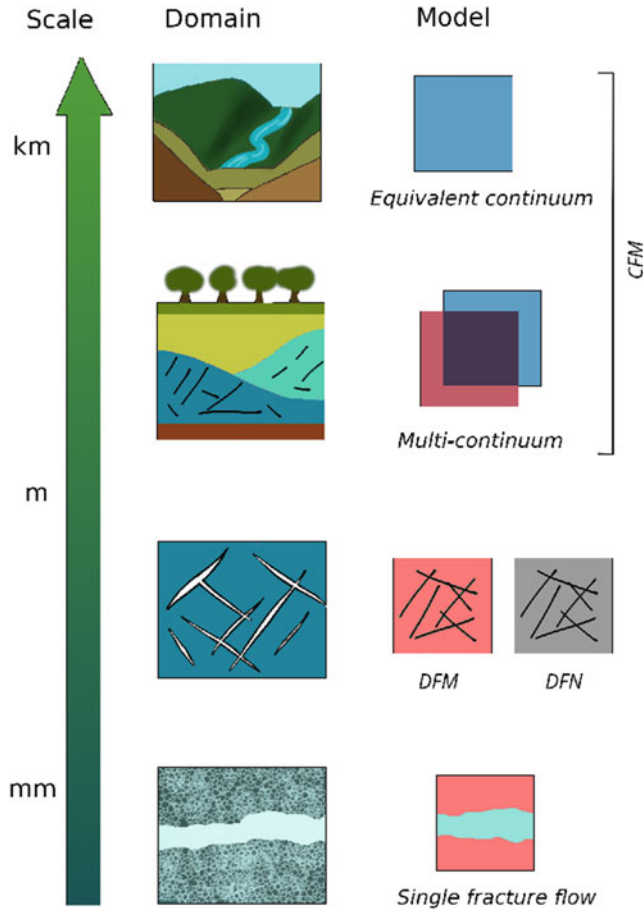
When they are empty or filled with highly permeable materials, fractures may provide a preferential path to fluid flow, but in some cases the deposits inside fractures can become nearly impermeable. We refer to the latter situation as blocking fracture.

The presence of fractures greatly alters the macroscopic properties of the material in a complex way, in particular its mechanical and flow characteristics. We will be here concerned with the second aspect, and specifically on the mathematical modeling and computing techniques that may be adopted in the presence of fractured porous media. For a more general treatment of fractures in porous media the reader may refer to the book “Fractured Porous Media” (Adler et al. 2012).

The irregular spatial distribution of fractures and the presence of multiple scales make it difficult and often impossible to account for fractures by deriving effective upscaled parameters, permeability for instance, by volume averaging or homogenization techniques. Indeed, such procedures assume a strong separation of scales. Therefore, methods have been developed that model fractures explicitly, at least those crucial for the flow. These models are based on the assumption that in the porous medium flow is governed by Darcy’s law, and often a similar model is adopted for the flow taking place in the fractures as well.

We recall the main characteristic of the Darcy’s model, considering, for simplicity, just the case of single-phase flow. In this mathematical framework the two main variables are the pressure p and the macroscopic velocity field u , also called Darcy’s velocity. The two quantities are related by Darcy’s law (Bear 1972; Helmig 1997),

$$\mu \mathbf{K}^{-1} u + \nabla p = \mathbf{0} \quad \text{in } \Omega, \quad (1a)$$



Numerical Methods for Flow in Fractured Porous Media, Fig. 1 Choice of the numerical model for domains at different space scales

where $\Omega \subset \mathbb{R}^d$ represents the domain occupied by the porous material and ∇ indicates the gradient. In the case where gravity effect are relevant, Eq. (1a) may be modified by replacing the pressure term with $p - \rho g z$, where ρ is the fluid density, g the magnitude of the gravity acceleration, and z is the vertical coordinate pointing upwards from the Earth surface. The main hypotheses behind the model are that fluid velocity is small, so we can neglect inertial effects, and the main model parameters are: μ , the fluid viscosity, and $\mathbf{K}$, the permeability tensor of the porous medium, which is a symmetric and positive definite tensor. Permeability may be heterogeneous in space and often with high variations.

The second equation expresses continuity of mass by the following differential equation (Bear 1972; Helmig 1997),

$$c\phi\partial_t p + \nabla \cdot \mathbf{u} = q \quad \text{in } \Omega, \quad (1b)$$

where $\nabla \cdot$ is the divergence operator, q a source/sink term, c accounts for the medium and fluid compressibility, and ϕ is the porosity. Sometimes one is interested in the steady-state

solution or the compressibility can be neglected, in which case $c\phi\partial_t p = 0$.

Equations (1a) and (1b) form a system of partial differential equations which, complemented by appropriate boundary and initial conditions, allows to describe the evolution of (u, p) in the porous medium.

Numerical Models for Fractured Porous Media

Numerical models for fractured porous media may be subdivided into two main categories: continuum fracture models (CFMs) and discrete fracture models (DFMs).

Continuum Fracture Models

CFMs are early models introduced in the 1960s (Warren and Root 1963), later justified mathematically in (Arbogast et al. 1990), and currently implemented in many industrial software. They assume a highly permeable and interconnected fracture network so that it can be modeled as a continuum superimposed to that of the porous medium.

A commonly adopted method is the dual-porosity/dual-permeability scheme. It assumes that at each point of the domain Ω we may use the Darcy's equations for flow in the fractures and in the porous medium, respectively, with a term representing the interchange of mass between them. A basic model of this type may be written as:

$$\begin{aligned} c_m \phi_m \partial_t p_m + \nabla \cdot \mathbf{u}_m + \alpha(p_m - p_f) &= q_m & \text{in } \Omega, \\ c_f \phi_f \partial_t p_f + \nabla \cdot \mathbf{u}_f - \alpha(p_m - p_f) &= q_f \end{aligned} \quad (2a)$$

where suffixes m and f refer to quantities related to the porous medium and fractures, respectively. The term $\alpha(p_m - p_f)$ represents the mass exchange between the two components, with α a rate parameter. The Darcy's velocities $\mathbf{u}_m$ and $\mathbf{u}_f$ are given by

$$\begin{aligned} \mu \mathbf{K}_m^{-1} \mathbf{u}_m + \nabla p_m &= 0 \\ \mu \mathbf{K}_f^{-1} \mathbf{u}_f + \nabla p_f &= 0 \end{aligned} \quad \text{in } \Omega. \quad (2b)$$

A simpler model, called dual-porosity/single-permeability, may be obtained formally by setting $\mathbf{K}_m = \mathbf{0}$. It assumes that the porous medium acts as storage volume for flow occurring only along fractures. A further simplification, appropriate only for highly connected networks of fracture of small extension relative to the domain size, consists in the use of a single equivalent continuum with upscaled properties that account for the combined effect of porous medium and fractures.

Discrete Fracture Models

DFMs represent fractures explicitly, modeled as a network of (typically planar) surfaces Γ , immersed in the porous medium. In the fractures we typically use a Darcy-type model where

some special source terms are added to account for the fluid exchange with the porous medium. The Darcy equations in the latter are also modified, with terms that act as interface conditions.

DFMs are computationally more demanding than CFMs, but also more accurate, particularly when fractures of relative large extension are present. For this reason the choice between CFM and DFM may depend on the spatial scale of interest, as illustrated in Fig. 1.

On each fracture we may identify a unit normal vector n_f , and thus a positive and a negative side of Γ , see Fig. 2. We indicate with $\llbracket f \rrbracket = f^+ - f^-$ the jump of a quantity f across the fracture.

A commonly used model is derived in Martin et al. (2005), in which the fracture permeability is split into a normal, $K_{f,n}$, and tangential, $K_{f,t}$, components to account for the fact that those quantities may be different and scale differently with the fracture aperture ϵ . In the porous medium $\Omega \setminus \Gamma$ we consider Eq. (1), while in the fractures we have, neglecting the compressibility term,

$$\begin{aligned} \mu K_{f,t}^{-1} u_f + \epsilon \nabla_\tau p_f &= 0 \\ \nabla_\tau \cdot u_f - [u_m \cdot n_f] &= q_f \end{aligned} \quad \text{in } \Gamma, \quad (3a)$$

where $\nabla_\tau \cdot$ and ∇_τ are the divergence and gradient operating on the tangent plane of the fractures, respectively. The jump of normal velocity across the fracture acts as a source term in the mass conservation equation and represent the net flux entering (or leaving) the fracture. This model must be complemented by appropriate boundary conditions. In the portion of the boundary of Γ that touches the boundary of Ω , the conditions are determined by the specific problem at hand: either pressure or mass flux is prescribed. In the case of a fracture tip that ends inside Ω , a zero mass flux condition is generally adopted.

A network normally exhibits intersection between fractures. A common approach is to assume that the pressure is continuous at the intersection and the net sum of fluxes is zero. More sophisticated models are available, to account for fractures with different hydraulic properties and flow along fracture intersections.

To close the problem we need also to specify interface conditions to couple Γ to the porous medium. In Martin et al. (2005) a family of models is presented, including the following, often adopted in practice

$$\begin{aligned} \mu \epsilon K_{f,n}^{-1} u_m^+ \cdot n_f + 2(p_f - p_m^+) &= 0 \\ \mu \epsilon K_{f,n}^{-1} u_m^- \cdot n_f + 2(p_m^- - p_f) &= 0 \end{aligned} \quad \text{on } \Gamma. \quad (3b)$$

These conditions can be interpreted as the application of a discrete Darcy law across the fracture, and indeed they link the flux across the fracture to pressure differences.

Remark 1 If the fractures are highly permeable in the normal direction, continuity of pressure across Γ is often assumed, that is, $p_m^+ = p_m^-$. This induces a certain simplification in the model, since it does not account for a net mass flow across the fracture, but only for porous medium-fracture exchanges.

Remark 2 For nearly impermeable porous media, a different simplification of these models, called discrete fracture networks (DFNs), consists in neglecting the effect of the porous material and simulates flow just in the fracture network. They can be used in the presence of highly connected and permeable fractures.

Discretization Schemes for DFM

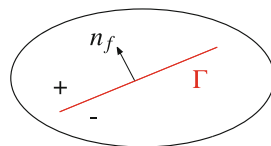
Numerical schemes for the approximation of the equations presented in the previous section are based on partitioning the domain Ω into a mesh of polyhedral elements $\mathcal{T}_h$. The unknowns are then discretized by assuming a given variation inside each element, for instance constant or linear. The continuous solution is then replaced by the discrete values at the mesh nodes.

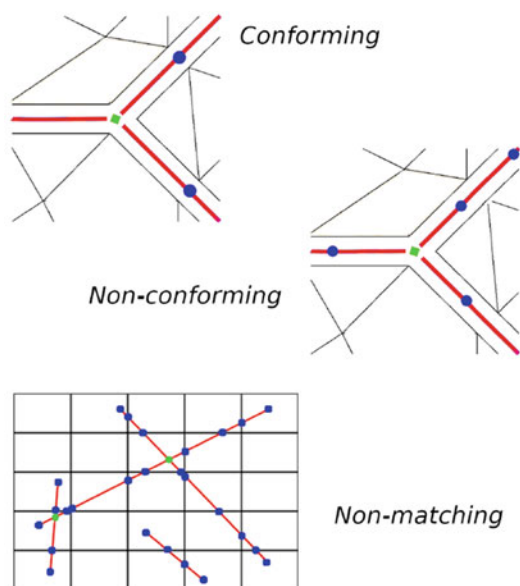
In the case of DFM we need to mesh both the porous medium and fracture domains and construct suitable ways to couple the two via (3b). Many discretization techniques available in the literature may be roughly subdivided into three categories, depending on the relation between porous medium and fracture grids, see Fig. 3.

- (i) Conforming methods. The mesh for the porous medium conforms to that adopted for the fracture network. It means that a fracture mesh element coincides (geometrically) with the faces of two porous medium mesh elements, on the positive and negative side of the fracture. Consequently, no porous medium mesh elements are cut by the fracture. This requirement poses strong constraints on the mesh generation process, which indeed can be the most time-consuming part of the simulation, particularly in the presence of complex networks. On the other hand, the implementation of (3b) is rather straightforward.
- (ii) Non-conforming methods. Still no elements in the porous medium grid are cut by the fractures: however, the grid at the two sides of the fracture and the fracture

Numerical Methods for Flow in Fractured Porous Media,

Fig. 2 Positive and negative side of Γ





Numerical Methods for Flow in Fractured Porous Media,
Fig. 3 Porous medium and fracture grids

grid are independent. The implementation of (3b) involves the setup of suitable operators to transfer the discrete solution in the fractures to the porous medium grid and vice versa. The so-called mortar technique, which is based on the set up of additional variables at the interface, is sometimes used to simplify the construction of the transfer operators. The process of grid generation is eased and it is simpler to avoid the generation of highly distorted elements: however, it is still rather demanding in complex situations.

- (iii) Non-matching methods. The mesh for the porous medium and the fracture network are completely independent and the fracture grids can cut the porous medium mesh elements in an arbitrary way. This simplifies mesh generation greatly, since porous medium and fracture grids can be generated independently. It is still necessary to find the intersections, but this is simpler than generating a conforming mesh and can be done with standard geometric search tools. However, implementation of conditions (3b) is more complex.

For every category, many different numerical schemes are at disposal. For the cases (i) and (ii), it is beneficial to use techniques able to operate on arbitrary polyhedral grids, like finite volumes or mimetic finite differences, to mention the more established one. The research in this field is, however, very active, and we mention also gradient schemes, the hybrid high order (HHO) method, and the virtual elements Method (VEM). A reference containing examples of numerical schemes applied in a DFM context is Fumagalli et al. (2019).

In the case (iii), as already stated, the main difficulty is how to impose the interface conditions. In that respect we have two

classes of procedures. The first is to represent the possible jumps in the solution across the elements cut by a fracture explicitly. This is what is done in eXtended Finite Elements (XFEM), where the finite element basis functions are locally enriched to allow discontinuous solutions that can satisfy the coupling conditions (3b). Contrarily, if one accepts a less accurate representation of the solution, some manipulations of the interface conditions are possible to transform them into source terms acting both on the fracture elements and on the elements in the porous medium that are cut by the fracture. The resulting source terms have some similarity with those present in CFM techniques. The embedded discrete fracture model (EDFM), usually coupled with a simple finite volumes approximation, falls into this second category of methods. A reference on EDFM techniques is in Lee et al. (2000).

The performances of many classes of methods are presented and discussed in Flemisch et al. (2018) for bi-dimensional problems and Berre et al. (2020) for three-dimensional problems.

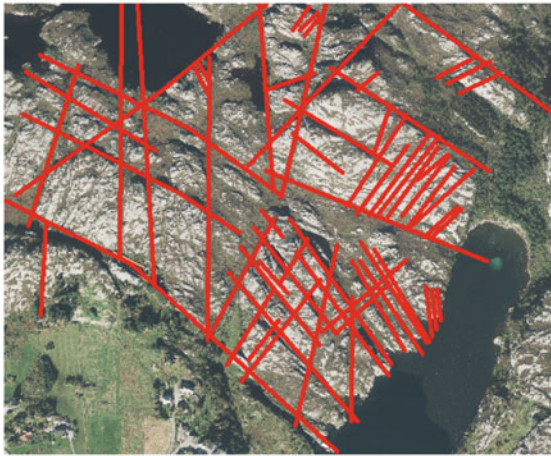
An Example of Computational Workflow

Numerical simulation of flow in fractured porous media is challenging due to the intrinsic geometrical complexity of the fractures, as well as the measurement of real fractures and their properties buried deep in the underground. These data are difficult to obtain and usually affected by large uncertainty which compromises the reliability of the numerical solutions.

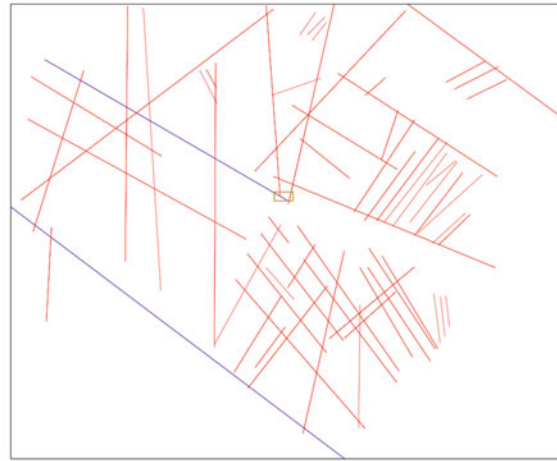
Several approaches can be considered to detect fractures in the underground, from seismic inversion to outcrop interpretation, the former effective to detect big fractures few kilometers below the surface while the latter normally used as an analogue of the underground. Once the fractures are collected and digitalized by means of one of these methods, a suitable mathematical model can be adopted to perform the simulations.

We focus our attention on the case of outcrop interpretation. From highly detailed photographs of the interested region, all the fractures are collected and interpreted up to a minimum sampling scale determined by the quality of the data or by computational constraints. Smaller fractures can be accounted for by a suitable change of the porous medium properties and seen as upscaled or homogenized. What quantity defines a fracture “small” and thus not explicitly represented is still an open question, many authors consider the fracture extension a good proxy for its importance. After this operation, the digitalized version of the outcrop is thus available, see Fig. 4 on the top as an example where natural and human factors limit its exposure. The outcrop is a portion of Sotra Island, near Bergen in Norway.

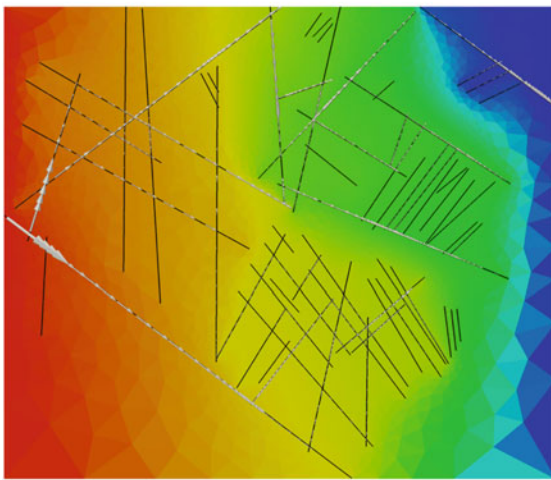
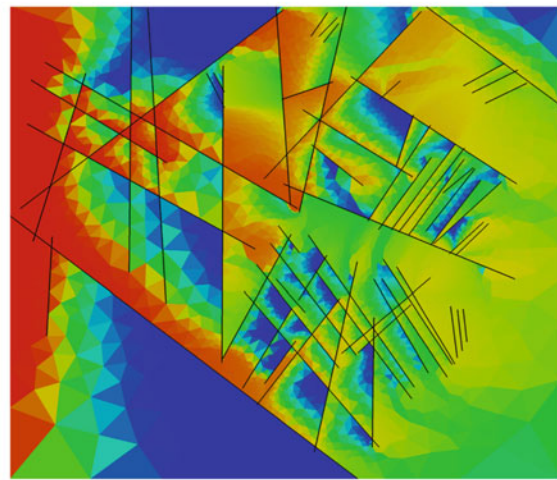
Another challenging aspect is the collection of the physical data, again affected by uncertainty, that are needed to run the



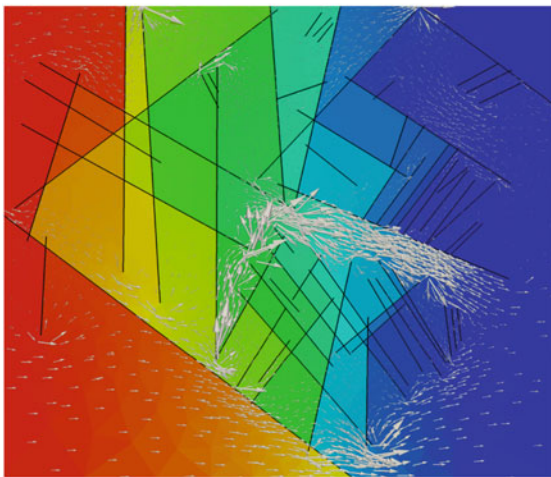
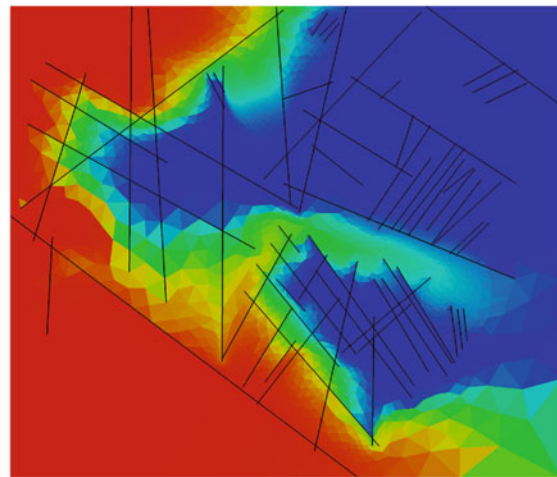
(a) Interpreted outcrop.



(b) Digitalized fractures.

(c) p and u with high permeable fractures.

(d) Scalar with high permeable fractures.

(e) p and u with low permeable fractures.

(f) Scalar with low permeable fractures.

Numerical Methods for Flow in Fractured Porous Media, Fig. 4 On the top the interpreted outcrop and digitalized fractures. The blue fractures are geometrically simplified due to software constraint. Images are taken from Flemisch et al. (2018). The others

represent pressure, velocity field, and a scalar tracer at a specific time. The color scheme spans from the lowest in blue to the highest in red. (Images are taken from Fumagalli and Keilegavlen (2019). Under CC BY 4.0 licence)

simulation. If some data are not available, it is possible to fill the gap considering models from the literature to, for example, relate the permeability with the aperture and the latter with the fracture extension. The model can, at this point, be discretized numerically by means of one of the methods mentioned in the previous sections.

Figure 4 reports numerical results of a simulation, in particular the pressure and concentration of a tracer (for instance a contaminant) in the interpreted outcrop. Two extreme cases are considered, if the fractures are more or less permeable than the surrounding porous medium. Data are homogeneous and a left to right pressure gradient is imposed. The solutions are obtained with the library PorePy, see <https://github.com/pmgbergen/porepy>.

Conclusion

Modeling flow in porous media in presence of fractures and fracture networks is a challenging task. Several approaches are available in literature, and a few of them implemented in specialized software. The choice depends on the scales at which the phenomenon has to be considered, on the connectivity of the fracture network and on the level of accuracy desired. Simpler continuum fracture models are suitable for highly connected and dense networks of fracture, while in presence of fracture with a larger extension and more sparse, discrete fracture models provide more accurate results. Clearly, it is also possible to use a combination of the two approaches. In this entry we gave a rapid review of the different strategies and, for the sake of simplicity, we focused the attention on single-phase flow. The general conclusions can however be extended to the more complex situation of multiphase flows.

Cross-References

- [Lithosphere, Mechanical Properties](#)
- [Numerical Methods, Finite Element](#)
- [Sedimentary Basins](#)

Acknowledgments The authors acknowledge Davide Losa-pio for his help in the research underlying this manuscript.

Bibliography

- Adler PM, Thovert JF, Mourzenko VV (2012) Fractured porous media. Oxford University Press, Oxford
- Arbogast T, Douglas J Jr, Hornung U (1990) Derivation of the double porosity model of single phase flow via homogenization theory. SIAM J Math Anal 21(4):823–836. <https://doi.org/10.1137/0521046>
- Bear J (1972) Dynamics of fluids in porous media. American Elsevier, New York
- Berre I, Boon WM, Flemisch B, Fumagalli A, Gläser D, Keilegavlen E, Scotti A, Stefansson I, Tatomir A, Brenner K, Burbulla S, Devloo P, Duran O, Favino M, Hennicker J, Lee IH, Lipnikov K, Masson R, Mosthaf K, Nestola MGC, Ni CF, Nikitin K, Schädle P, Svyatskiy D, Yanbarisov R, Zu-lan P (2020) Verification benchmarks for single-phase flow in three-dimensional fractured porous media. Tech Rep. arXiv:2002.07005 [math.NA]. <https://arxiv.org/abs/2002.07005>
- Flemisch B, Berre I, Boon W, Fumagalli A, Schwenck N, Scotti A, Stefansson I, Tatomir A (2018) Benchmarks for single-phase flow in fractured porous media. Adv Water Resour 111:239–258
- Fumagalli A, Keilegavlen E (2019) Dual virtual element methods for discrete fracture matrix models. Oil Gas Sci Technol – Revue d'IFP Energies nouvelles 74(41):1–17
- Fumagalli A, Berre I, Formaggia L, Keilegavlen E, Scotti A (eds) (2019) Numerical methods for processes in fractured porous media. Lecture notes in geosystems mathematics and computing. Birkhäuser, Basel
- Helmig R (1997) Multiphase flow and transport processes in the subsurface: a contribution to the modeling of hydrosystems. Springer, Berlin
- Lee S, Jensen C, Lough M (2000) Efficient finite-difference model for flow in a reservoir through discrete fracture networks and homogenized media. SPE Reserv Eval Eng 3:268–275
- Martin V, Jaffré J, Roberts JE (2005) Modeling fractures and barriers as interfaces for flow in porous media. SIAM J Sci Comput 26(5):1667–1691
- Warren EJ, Root PJ (1963) The behaviour of naturally fractured reservoir. Soc Petr Eng 3:245

Numerical Methods, Boundary Element

Michele Cooke

Geosciences Department, University of Massachusetts at Amherst, Amherst, MA, USA

Definition

The boundary element method is a numerical technique for investigating deformation within the Earth. The method is particularly adept at investigations of discontinuities in the crust, such as cracks, faults, and dikes. The numerical method is able to analyze complex problems by discretizing the surfaces of the discontinuities into small elements and solving the problem piecewise.

Introduction

The boundary element method (BEM) is one of several numerical methods used to investigate deformation within the Earth. The principal difference between boundary element

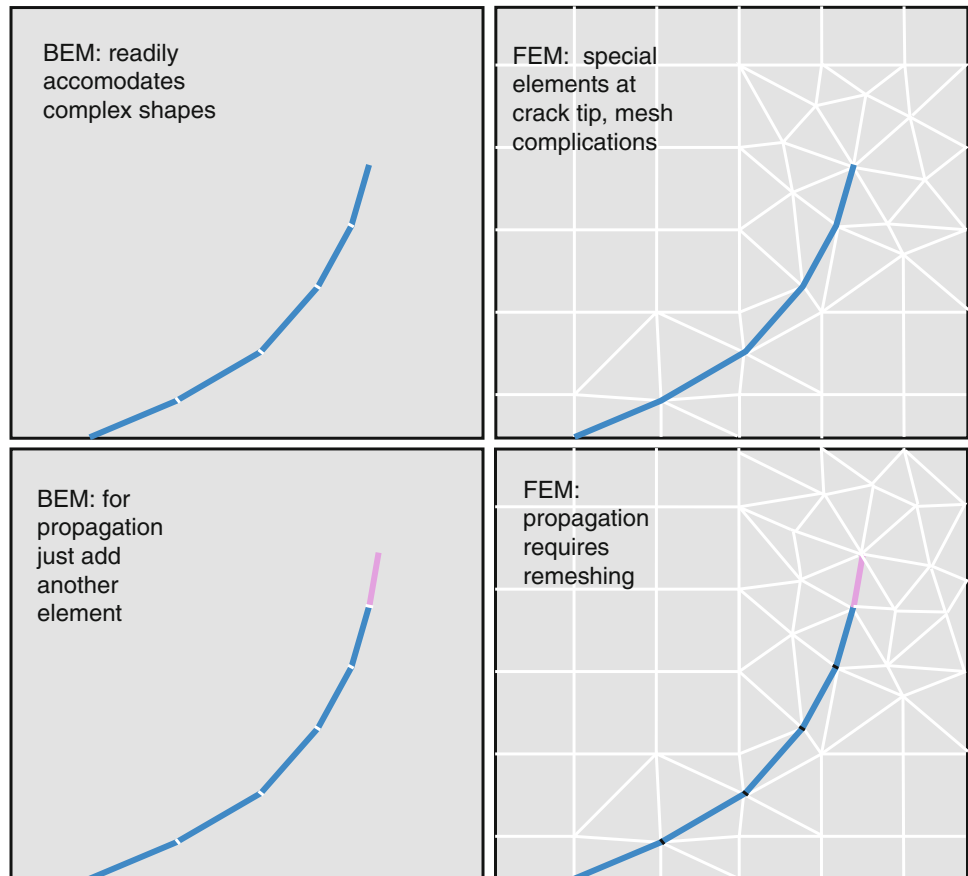
method and most other numerical methods of continuum mechanics, such as the finite element method (FEM), is that only the boundaries of the deforming body need to be meshed (Fig. 1). While other methods require meshing the entire volume of the deforming body into two-dimensional (2D) or three-dimensional (3D) elements, the BEM only requires discretization of the boundaries (e.g., fractures, bedding planes, or external boundaries) into one-dimensional (1D) or 2D elements. Minimizing discretization decreases model building and run time as well as errors due to discretization (Crouch and Starfield 1990). The greatest advantages of BEM modeling in geophysics are that (1) very complicated surface geometries, such as undulating fault surfaces, can be incorporated and (2) the method can readily accommodate fracture propagation by addition of new elements (Fig. 1). The disadvantages of BEM are (1) that the stiffness matrix relating stresses to strains on the elements is non-sparse so that inverting the matrix to find stresses can be very CPU intensive and (2) heterogeneous material properties are more complex to incorporate in BEM than methods that utilize volume discretization, such as FEM (Becker 1992).

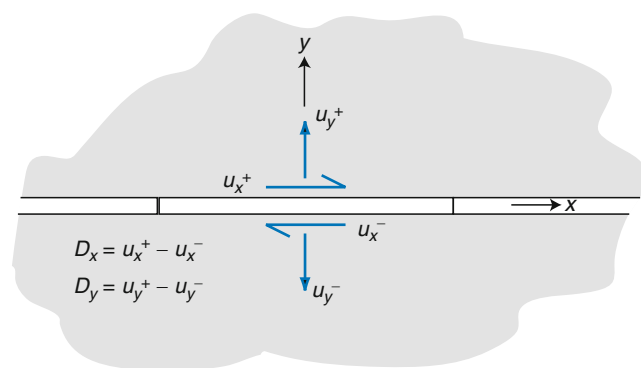
How Does BEM Work?

BEM models are based on the solution to Kelvin's problem for the deformation due to a point load within a body. The solution provides the stresses and displacements everywhere due to the point load. We can calculate the effects of distributed loads by integrating the point force. When the loads are applied over discrete elements, the relationships between the displacements at one element due to loads at another are used to build an influence coefficient matrix. When either the displacements or tractions are prescribed to all elements that comprise the BEM model, the method will solve for the unknown element tractions or displacements via the influence coefficient matrix. Once all the displacements and tractions along the boundary elements are known, the stresses and strains at any point within the body can be calculated using the solution of Kelvin's problem. So unlike FEM, where solutions are calculated at each node within the element mesh, within BEM, problems displacements and stresses can be calculated at any point within the body, once all the tractions and displacements along the elements are known.

Numerical Methods, Boundary Element, Fig. 1

Contrasting discretization approaches of the boundary element method and finite element method. BEM only discretizes the boundaries and/or internal surfaces (*blue line*) rather than the entire volume. Fracture propagation is facilitated by addition of elements (*red line*)





Numerical Methods, Boundary Element, Fig. 2 Formulation of the displacement discontinuity along each element within a boundary element method model. The normal and shear displacement discontinuity (D_x and D_y , respectively) are the difference between the displacements along the top and bottom of the element

Two variations of BEMs are commonly used, the *fictitious stress* and *displacement discontinuity* methods. The fictitious stress method is based on Kelvin's problem and applies uniform tractions along each boundary element. In contrast, the displacement discontinuity method applies displacements along each element such that the difference in displacement from one side of the element to the other is constant along the boundary element (Fig. 2). While fictitious stress method works well for homogeneous bodies under external loads, the displacement discontinuity method is better suited for problems of solid bodies containing cracks (Crouch and Starfield 1990). For this reason, the displacement discontinuity method is more common for geophysical analyses, which investigate deformation associated with dikes, sills, microcracks, joints, and faults within the Earth.

What Can BEM Do?

Boundary element method codes using the displacement discontinuity formulation have been developed to solve both two-dimensional (e.g., Crouch and Starfield 1990) and three-dimensional problems (Okada 1992; Thomas 1993; Meade 2007). Two-dimensional formulations use a linear dislocation for each element while three-dimensional formulations utilize either rectangular dislocations (Okada 1992) or angular dislocations (Cominou and Dundurs 1975) that can be assembled to form triangular elements of constant displacement discontinuity (Thomas 1993; Meade 2007). The advantage of triangular elements over rectangular elements is that complex nonplanar surfaces can be meshed without creating locations of overlaps and gaps.

Important refinements to the basic displacement discontinuity method for investigation of problems in geophysics include incorporation of frictional slip along crack elements

and prevention of interpenetration of crack walls during compressive loading of cracks. These refinements facilitate investigation of deformation associated with subsurface faults and joints. Algorithms for preventing interpenetration include the penalty method (Crouch and Starfield 1990; Cooke and Pollard 1997) and the complementarity method (Maerten et al. 2010). The penalty method employs normal and shear stiffness to each element to prevent interpenetration. The complementarity method uses an iterative solver to solve the inequality that opening should be equal to or greater than zero along each element (Maerten et al. 2010).

Heterogeneous properties can be incorporated within BEM formulations by way of a contact boundary. Stresses and displacements are prescribed to be uniform across the boundary between otherwise homogeneous portions of the model. With this constraint along the boundary, the solution for each homogeneous section is found separately. This approach has not been widely applied because it is difficult to implement; finite element methods are generally believed to be better suited for heterogeneous properties.

Applications of the Boundary Element Method

In most BEM implementations, gravitational body forces are superposed onto the solution to the applied loading. This precludes direct investigation of problems related to topographic loading. However, Martel and Muller (2000) developed a formulation for analyzing topographic stresses using a long stress-free crack to simulate the topography.

Investigations of fluid flow through fracture networks has benefited from the surface discretization approach of the boundary element method. Models of subsurface flow through fracture networks utilize a boundary element method model that incorporates fluid flow along each element (e.g., Dershowitz and Fidelibus 1999). Such approaches can permit the full coupling of mechanical deformation and fluid flow.

The boundary element method has found widespread use in the subfield of structural geology for solving problems of deformation and propagation of opening mode and sliding mode fractures (e.g., joints and faults). Both forward and inverse modeling techniques have utilized the boundary element method. Forward models apply loading to the crack system to determine the resulting deformation of the system. Such investigations may explore the development of secondary cracks around a fault or the aperture of a hydraulic fracture. In contrast, inverse models start with the observed deformation and invert to find the slip on the fault or opening on the crack that must have produced the observations. For example, deformation observed along the surfaces of the earth via GPS stations velocities or InSAR imagery can be inverted to find the slip distribution on the underlying fault (e.g., Maerten et al. 2005).

Cross-References

- Numerical Methods, Finite Difference
- Numerical Methods, Finite Element

Bibliography

- Becker AA (1992) The boundary element method in engineering: a complete course. McGraw-Hill, New York
- Cominou MA, Dundurs J (1975) The angular dislocation in a half space. *J Elast* 5:203–216
- Cooke ML, Pollard DD (1997) Bedding plane slip in initial stages of fault-related folding. *J Struct Geol* 19:567–581. Special Issue on Fault-Related Folding
- Crouch SL, Starfield AM (1990) Boundary element method in solid mechanics with applications in rock mechanics and geological engineering. Unwin Hyman, Boston, 322 p
- Dershowitz W, Fidelibus C (1999) Derivation of equivalent pipe network analogues for three-dimensional discrete fracture networks by the boundary element method. *Water Resour Res* 35(9):2685–2691
- Maerten F, Resor P, Pollard D, Maerten L (2005) Inverting for slip on three-dimensional fault surfaces using angular dislocations. *Bull Seismol Soc Am* 95(5):1654–1665
- Maerten F, Maerten L, Cooke ML (2010) Solving 3D boundary element problems using constrained iterative approach. *Comput Geosci*. <https://doi.org/10.1007/s10596-009-9170-x> [in press]
- Martel S, Muller J (2000) A two-dimensional boundary element method for calculating elastic gravitational stresses in slopes. *Pure Appl Geophys* 157:989–1007. <https://doi.org/10.1007/s000240050014>
- Meade BJ (2007) Algorithms for the calculation of exact displacements, strains, and stresses for triangular dislocation elements in a uniform elastic half space. *Comput Geosci* 33:1064–1075. <https://doi.org/10.1016/j.cageo.2006.12.003>
- Okada Y (1992) Internal deformation due to shear and tensile faults in a half-space. *Bull Seismol Soc Am* 82:1018–1040
- Thomas AL (1993) Poly3D: a three-dimensional polygonal-element displacement discontinuity boundary element computer program with applications to fractures, faults and cavities in the earth's crust. MS thesis, Stanford University, California

Neumann boundary condition

A condition imposed on the fluxes associated to the differential problem at hand, that is, stresses, heat flux, etc. Also called natural boundary condition.

Parallel computer

A computer with more than one processing unit capable of computing operations concurrently.

Parallel computation

A numerical procedure executed on a parallel computer.

Preconditioner

In the context of the solution of a linear system by an iterative method, a preconditioner is an easily invertible matrix spectrally similar to the matrix governing the problem.

Introduction

Domain decomposition (DD) method is a technique for the solution of partial differential equations which can be instrumental to the development of *parallel computations*. It can be used in the framework of discretization methods, for example, finite elements, finite volumes, finite differences, or spectral element methods.

It is based on the reformulation of the given boundary-value problem on a partition of the computational domain Ω into M subdomains Ω_i , with $i = 1, \dots, M$. Typically M is also the number of processors at disposal, even if it is possible to have more subdomains per processor.

The DD method also provides a convenient framework for the solution of heterogeneous or multi-physics problems, that is, those that are governed by differential equations of different kinds in different subregions of the computational domain. In this case, the subdomains conform with the subregions. In this work, however, we will address only homogeneous DD methods, yet some of the concepts presented here may be readily adapted for the nonhomogeneous case.

We will refer to a partial differential equation of the form

$$-\operatorname{div}(\mathbf{T}(\mathbf{u})) = \mathbf{f} \text{ in } \Omega \quad (1)$$

with suitable boundary conditions on $\partial\Omega$.

For instance, in an elastostatic problem, $\mathbf{u}$ is the displacement and $\mathbf{T}(\mathbf{u})$ the stress tensor; in a heat conduction problem, $\mathbf{u}$ is a scalar variable, the temperature, and $\mathbf{T}$ the Laplace operator. With a rather similar formalism we may also account for the Stokes problem. The method is in fact applicable also to time-dependent problems, since in most cases the time advancing scheme required by the numerical simulation leads eventually to a sequence of problems of type (1). For instance, if we consider the evolution equation

Numerical Methods, Domain Decomposition

Alfio Quarteroni^{1,2} and Luca Formaggia¹

¹MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy

²CMCS-MATHICSE, EPFL, Lausanne, Switzerland

Definition

Computational domain	A bounded spatial region where a partial differential equation is solved.
Dirichlet boundary condition	A condition imposed directly on the unknown of the differential problem, also called essential boundary condition.

$$\partial_t \mathbf{u} - \operatorname{div}(\mathbf{T}(\mathbf{u})) = \mathbf{f},$$

a time discretization by the implicit Euler method leads to solve for the unknown $\mathbf{u}^n$ at each time step t^n a differential problem of the type

$$\mathbf{u}^n - \delta t \operatorname{div}(\mathbf{T}(\mathbf{u}^n)) = \delta t \mathbf{f}^n + \mathbf{u}^{n-1}.$$

It is still of the form (1) with $\operatorname{div}(\mathbf{T})$ replaced by $\delta t \operatorname{div}(\mathbf{T}) - \mathbf{I}$, δt being the time step, and $\mathbf{I}$ the identity operator. A similar form is obtained also for the elastodynamic equations, discretized in time, for instance, by a Newmark method (Quarteroni and Valli 1994).

Typically, and for evident practical reasons, the partition into subdomains is made after having triangulated Ω into a finite element mesh τ_h , each subdomain being in fact formed by a set of elements of the original grid (see Fig. 1). The partition is often made automatically, using libraries like METIS (Karypis and Kumar 1998b) or PARMETIS (Karypis and Kumar 1998a), the latter able to exploit a parallel architecture also for this preprocessing stage.

More in particular, there are two ways of subdividing the computational domain into subdomains. The first uses disjoint subdomains, where $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$. Here, the interface between subdomains reduces to surfaces (in 3D) and lines (in 2D). The second strategy adopts *overlapping subdomains*, usually built by letting the subdomains of an initial nonoverlapping partition grow of a certain factor δ . In most of the cases δ is dictated by the number of “layers” of grid elements that are added to the original partition. The minimal overlap is of one element, like that shown in Fig. 1. For a more complete review of DD methods one may refer to Smith et al. (1996), Quarteroni and Valli (1999), Wohlmuth (2001),

Toselli and Widlund (2005), and Mathew (2008); several examples with emphasis on parallel computations are reported, for instance, in Bruaset and Tveito (2006).

Algorithms Without Overlap

We indicate with Γ_{ij} the interface between subdomains Ω_i and Ω_j and we set $\Gamma = \cup_{ij} \Gamma_{ij}$. We exploit the fact that problem (1) is equivalent to solving the following M coupled problems for $i = 1, \dots, M$,

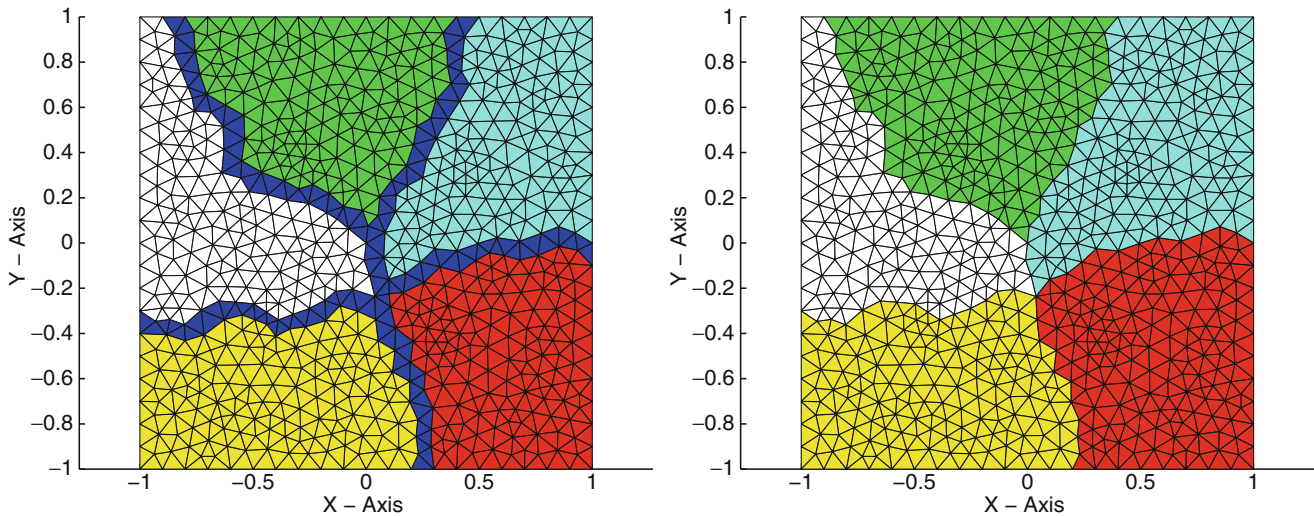
$$-\operatorname{div}(\mathbf{T}(\mathbf{u}_i)) = \mathbf{f}_i \text{ in } \Omega_i \quad (2)$$

with the same boundary conditions of the original problem applied to $\partial\Omega_i \cap \partial\Omega$, while on each Γ_{ij} we set the continuity of the local solutions $\mathbf{u}_i$ and of the fluxes, that is,

$$\mathbf{u}_i = \mathbf{u}_j \quad \mathbf{T}(\mathbf{u}_i) \cdot \mathbf{n}_{ij} + \mathbf{T}(\mathbf{u}_j) \cdot \mathbf{n}_{ji} = 0, \quad (3)$$

where $\mathbf{n}_{ij}$ is the normal to Γ_{ij} outward oriented w.r.t. Ω_i . It may be proved that $\mathbf{u}_i = \mathbf{u}|_{\Omega_i}$ (Quarteroni and Valli 1999).

Problems (2) are coupled because of relations (3). A parallel algorithm may be obtained by an iterative procedure where conditions (3) are enforced in such a way to generate at each iteration decoupled problems that can be run on different processes, with only a small amount of communication needed at the beginning of each iteration. For the sake of space we give an example of just one of these procedures, called *Dirichlet-Neumann* for the case of two subdomains Ω_1 and Ω_2 . Starting from a guess for $u_1^{(0)}$ and $u_2^{(0)}$, the algorithm solves for $k = 1, 2, \dots$ the following sequence of independent problems,



Numerical Methods, Domain Decomposition, Fig. 1 Partition of a computational domain starting from a given triangulation. *Left*: A partition into subdomains with an overlap equal to one layer of elements. *Right*: Partition of a domain with disjoint subdomains

$$\begin{cases} -\operatorname{div}\left(\mathbf{T}\left(\mathbf{u}_1^{(k+1)}\right)\right)=\mathbf{f}_1, & \text{in } \Omega_1, \\ \mathbf{u}_1^{(k+1)}=\theta \mathbf{u}_2^{(k)}+(1-\theta) \mathbf{u}_1^{(k)}, & \text{on } \Gamma_{12} \end{cases} \text{ and } \begin{cases} -\operatorname{div}\left(\mathbf{T}\left(\mathbf{u}_2^{(k+1)}\right)\right)=\mathbf{f}_2, & \text{in } \Omega_2 \\ \mathbf{T}\left(\mathbf{u}_2^{(k+1)}\right) \cdot \mathbf{n}_{21}=\mathbf{T}\left(\mathbf{u}_1^{(k)}\right) \cdot \mathbf{n}_{21} & \text{on } \Gamma_{12} \end{cases}.$$

until a measure of the difference $\mathbf{u}_1^{(k+1)} - \mathbf{u}_2^{(k+1)}$ on Γ_{12} is below a given tolerance; θ here is a convenient relaxation factor. The Dirichlet-Neumann technique is not easily extendable to an arbitrary number of subdomains (unless suitable *coloring techniques* are used) and its convergence characteristics strongly depend on the geometry of the subdomains as well as on the possible jump of characteristic coefficients (for instance, the viscosity of different rocks forming a sedimentary basin). Other more favorable techniques like the Neumann-Neumann, Robin-Robin, and FETI methods are described in Quarteroni and Valli (1999), Wohlmuth (2001) and Toselli and Widlund (2005).

The DD method may be also set from an algebraic viewpoint. Indeed when discretized, for instance, by a finite element method, (2) and (3) reduce to a system of linear equations, which may be written in a block form as

$$\begin{pmatrix} A_{II} & A_{I\Gamma} \\ A_{\Gamma I} & A_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{u}_I \\ \mathbf{u}_\Gamma \end{pmatrix} = \begin{pmatrix} \mathbf{f}_I \\ \mathbf{f}_\Gamma \end{pmatrix},$$

where A_{II} is a *block diagonal* matrix with M blocks of dimension equal to the number of unknowns internal to each subdomain, the latter being collected in $\mathbf{u}_I$, while $\mathbf{u}_\Gamma$ is the vector of unknowns on the interface Γ . If A_{II} is invertible (and normally this is the case), we may obtain a problem for the $\mathbf{u}_\Gamma$ only (*Schur complement system*), $\Sigma_\Gamma \mathbf{u}_\Gamma = \chi_\Gamma$, where $\Sigma_\Gamma = A_{\Gamma\Gamma} - A_{\Gamma I} A_{II}^{-1} A_{I\Gamma}$ is the so-called *Schur complement matrix* w.r.t. A_{II} . Having solved for $\mathbf{u}_\Gamma$ the computation of $\mathbf{u}_I$ can be done in a perfect parallel fashion by solving the block diagonal problem $A_{II} \mathbf{u}_I = \mathbf{f}_I - A_{I\Gamma} \mathbf{u}_\Gamma$.

A DD scheme with no overlap may be interpreted as *preconditioned iterative scheme* (Quarteroni and Valli 1994; Quarteroni 2009) for the Schur complement system, where the preconditioner can be efficiently applied in a parallel setting. A crucial issue for parallel computing is that of *scalability*. In the DD setting, an algorithm is said to be scalable if its convergence properties do not depend on the number of subdomains M , and in particular does not degrade if we keep the ratio M/N between number of subdomains and the total number N of unknowns of our problem fixed. Indeed in this case (if we neglect communication overheads), we may solve in the same time a problem twice as large by doubling the number of processors.

A scalable parallel preconditioner cannot be built using only local (i.e., at the subdomain level) approximations of the Schur matrix; we need to add also a *coarse operator* that has the role of transferring information among far away subdomains. The typical form of the preconditioner is (we write directly the inverse operator since it is the one actually required by the iterative procedure (Quarteroni et al. 2007))

$$P_\Sigma^{-1} = \sum_{i=1}^M R_i^T \sum_i^* R_i + R_0^T \sum_0^* R_0, \quad (4)$$

where R_i is a *restriction operator* that selects from all the $\mathbf{u}_\Gamma$ those local to the i -th subdomain, and $\sum_i^*$ is a local approximation of the inverse Schur matrix, which typically can be built using just data related to the i -th subdomain (and thus in parallel). Finally, $\sum_0^*$ is the coarse operator, of small dimension (typically of the order of M), whose role is to guarantee the coupling among all subdomains and it is necessary for scalability. The application of the preconditioner (i.e., the computation of $P_\Sigma^{-1} \mathbf{x}$, where $\mathbf{x}$ is any vector of the right length) can be done in parallel a part from the coarse operator, which however, being small in size, is irrelevant in the computational cost balance. Several scalable preconditioners are available, see Toselli and Widlund (2005) and Canuto et al. (2007).

Methods with Overlap

A typical DD scheme with overlapping subdomains is the Schwarz method. In its basic form, it is an iterative algorithm for the solution of (1) that for $k = 1, 2, \dots$ solves a series of local problems on each subdomain Ω_i , where on the non-empty interfaces $\Gamma_{ij} = \partial\Omega_i \cap \Omega_j$, we apply Dirichlet boundary conditions using the latest values available from Ω_j , that is,

$$\begin{cases} -\operatorname{div}\left(\mathbf{T}\left(\mathbf{u}_i^{(k+1)}\right)\right)=\mathbf{f}_i, & \text{in } \Omega_i, \\ \mathbf{u}_i^{(k+1)}=\mathbf{u}_j^{(k)}, & \text{on } \Gamma_{ij}. \end{cases}$$

The iteration continues until the difference between two successive iterations is sufficiently small. The convergence analysis of the Schwarz method may be found in Smith et al. (1996); Quarteroni and Valli (1999); Toselli and Widlund (2005), and Canuto et al. (2007) in the context of spectral element discretizations.

The method is seldom used in this form, however. Yet, it is probably the most adopted method to build parallel preconditioners for an iterative solver of the global problem, giving rise the so-called *Krylov-Schwarz* (for linear problems) and *Newton-Krylov-Schwarz* methods.

A Schwarz parallel preconditioner P_S may be written similarly to (4), that is,

$$P_S^{-1} = \sum_{i=1}^M R_i^T A_i^{-1} R_i + R_0^T A_0^{-1} R_0,$$

where here the restriction operator R_i extracts from a vector of length N the elements corresponding to the unknowns internal to the extended subdomain Ω_i , $A_i = R_i A R_i^T$ is the local matrix extracted from the matrix A of our problem, and finally A_0 is again a global coarse operator of size of the order M needed for scalability, and R_0 being the corresponding restriction matrix.

Again, apart from the coarse operator, the computation of $P_S^{-1} \mathbf{x}$ can be done in parallel since the matrices A_i are local and can be handled by each processor.

Application to Geophysical and Geological Problems

DD methods have been applied successfully in the context of geophysical problems, like acoustic and elastodynamic wave propagation (Faccioli et al. 1997), or full-waveform tomography (Soubier et al. 2009).

It has been also applied to speed up the simulation of the evolution of sedimentary basins. Here, different types of sediments, such as gravel, sand, rocks, and biological remains that have been transported by the wind, the rivers, and, sometimes, by the sea, accumulate, are buried deeper and deeper, and are transformed eventually into rocks by a complex process of compaction and diagenesis.

On geological scales, the evolution of a sedimentary basin may be tackled by a fluid approach where each rock layer is modeled as a viscous fluid. This is particularly convenient in the presence of salt tectonics (Massimi et al. 2007).

The computational domain Ω is split in several subdomains Ω_i , which usually correspond to the rock layers, as shown in Fig. 2. At each time step, we have to solve in the domain Ω a Stokes problem, possibly with non-Newtonian rheology, to compute the instantaneous velocity field,

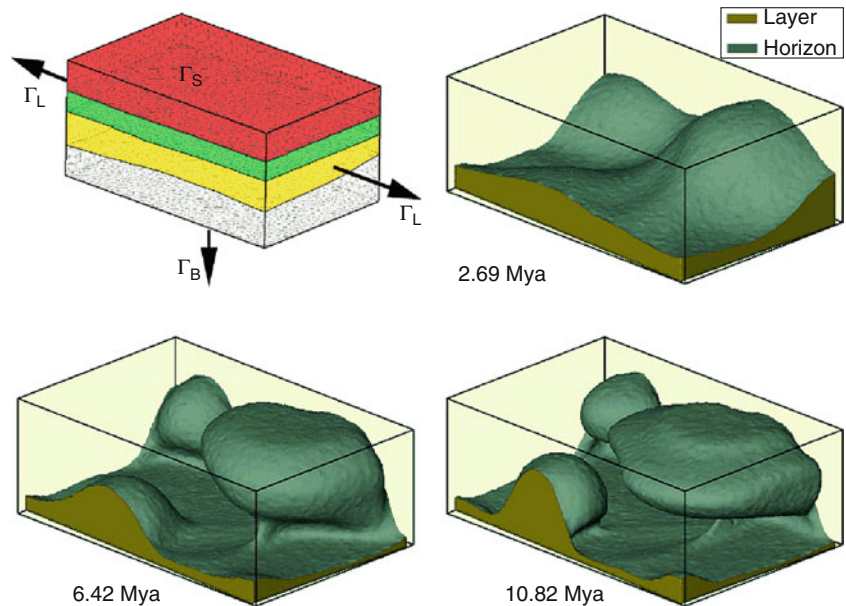
$$\begin{cases} -\text{div}(\mathbf{T}(\mathbf{v})) + \nabla p = \mathbf{f}, \\ \text{div}(\mathbf{v}) = \phi, \\ \frac{\partial \rho}{\partial t} + (\mathbf{v} \cdot \nabla) \rho = 0. \end{cases} \quad (5)$$

Here, $\mathbf{v}$ is the velocity, p the pressure, $\mathbf{f}$ the external gravitational field, and ρ the density. The stress tensor $\mathbf{T}$ depends on characteristics of the rocks, such as viscosity and density, which may be discontinuous across layers. The function ϕ may account for compaction processes or may be simply set to zero when the hypothesis of isochoric flow is acceptable. The last equation describes the evolution of the density, which is simply advected by the velocity field. The presence of faults is accounted for by appropriately reducing the rock viscosity in the vicinity of the faults.

The movement of the layers has been tracked using a level set technique. A finite element scheme has been adopted for the discretization of the Stokes problem, while a conservative finite volume scheme has been used for the tracking of the layer interfaces. The parallel implementation has been carried out with the help of the TRILINOS library (Heroux et al.

Numerical Methods, Domain

Decomposition, Fig. 2 The simulated evolution of a salt diapir. In the first graphic, the computational domain showing the subdivision of the layers. The other pictures show snapshots of the evolution of the salt layer with the formation of a diapir



2005), using a Schwarz algorithm with a coarse operator built by aggregation (Sala 2004).

Figure 2 shows the evolution of a salt dome. Salt is less compressible than the surrounding rock, so during the sedimentation process it ends having a smaller density than the overburden. We are then facing a Rayleigh-Taylor instability and any small perturbation will cause the salt to rise, with large movements that cause the so-called salt diapirism.

Cross-References

- Numerical Methods, Finite Element
- Numerical Methods, Multigrid
- Sedimentary Basins

Bibliography

- Bruaset A, Tveito A (eds) (2006) Numerical solution of partial differential equations on parallel computers. Lecture notes in computational science and Engineering, vol 51. Springer, New York
- Canuto C, Hussaini MY, Quarteroni A, Zang TA (2007) Spectral methods. Evolution to complex geometries and application to fluid dynamics. Springer, Berlin/Heidelberg
- Faccioli E, Maggio F, Paolucci R, Quarteroni A (1997) 2D and 3D elastic wave propagation by a pseudo-spectral domain decomposition method. *J Seismol* 1(3):237–251
- Heroux M et al (2005) An overview of the TRILINOS project. *ACM Trans Math Softw* 31(3):397–423. (TOMS)
- Karypis G, Kumar V (1998a) A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J Sci Comp* 20:359–392
- Karypis G, Kumar V (1998b) METIS: unstructured graph partitioning and sparse matrix ordering system. Technical report 98-036. University of Minnesota, Department of Computer Science
- Massimi P, Quarteroni A, Saleri F, Scrofani G (2007) Modeling of salt tectonics. *Comput Methods Appl Mech Eng* 197(1–4):281–293
- Mathew T (2008) Domain decomposition methods for the numerical solution of partial differential equations. Lecture notes in computational science and engineering. Springer, New York
- Quarteroni A (2009) Numerical models for differential problems. *MS&A*, vol 2. Springer, Berlin
- Quarteroni A, Valli A (1994) Numerical approximation of partial differential equations. Springer, Berlin/Heidelberg
- Quarteroni A, Valli A (1999) Domain decomposition methods for partial differential equations. Oxford Science, Oxford
- Quarteroni A, Sacco R, Saleri F (2007) Numerical mathematics, 2nd edn. Springer, Berlin/Heidelberg
- Sala M (2004) Analysis of two-level domain decomposition preconditioners based on aggregation. *Math Model Numer Anal* 38(5):765–780
- Smith B, Björstad P, Gropp W (1996) Domain decomposition, parallel multilevel methods for elliptic partial differential equations. Cambridge University Press, New York
- Soubier F, Operto S, Virieux J, Amestoy P, L'Excellent JY (2009) FWT2D: a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data—part 2: numerical examples and scalability analysis. *Comput Geosci* 35(3):496–514
- Toselli A, Widlund O (2005) Domain decomposition methods – algorithms and theory. Springer series in computational mathematics, vol 34. Springer, New York
- Wohlmuth B (2001) Discretization methods and iterative solvers based on domain decomposition. Lecture notes in computational science and engineering. Springer, Berlin

Numerical Methods, Finite Difference

Johan O. A. Robertsson¹ and Joakim O. Blanch²

¹Institute of Geophysics, Department of Earth Sciences, ETH-Zürich, Zürich, Switzerland

²BHP, Houston, TX, USA

Definition

Finite-difference method. A method to approximate derivatives between neighboring points in a grid. The method can be applied to solve partial-differential equations, such as the wave equation.

Introduction

The finite-difference (FD) method is among the most commonly used methods for simulating wave propagation in a heterogeneous Earth. In this article, we describe the FD method for modeling wave propagation on Cartesian grids in acoustic, elastic isotropic, elastic anisotropic, as well as viscoacoustic/elastic media. The basic equations for wave propagation can be formulated in various formally equivalent ways. We will restrict our description to systems of first-order partial-differential equations with pressure/stress and particle velocities as wavefield variables. Due to its versatility, attractive stability, and dispersive properties, particularly for modeling wave propagation in elastic media, this formulation has been by far the most popular formulation over the last several decades. A more extensive review of the FD method is given by Moczo et al. (2007).

We start with a brief review of wave propagation theory and introduce some fundamental concepts including the basic structure (staggering) of the FD grid. In the following two sections, we discuss the choice of FD approximation and introduce two basic numerical properties of the FD method: numerical dispersion and stability. The last two sections are devoted to different boundary conditions and source implementations.

Theory of Wave Propagation and Fundamental Concepts

We focus on wave propagation in 3D Cartesian coordinates with coordinate axes (x, y, z) . The special cases of 1D or 2D can easily be generalized from the expressions presented here, for example, in 2D, omitting terms or expressions containing components in the y -direction. The particle velocity wavefield is denoted by $\vec{v} = (v_x, v_y, v_z)$. In acoustic media, the second wavefield variable is the scalar quantity pressure, p , whereas

in elastic media it is the (symmetric) stress tensor, S . We will often be representing stresses and particle velocities using stress and strain vectors defined as (the so-called Voigt notation):

$$\vec{\sigma} = (\sigma_{xx} \ \sigma_{yy} \ \sigma_{zz} \ \sigma_{yz} \ \sigma_{xz} \ \sigma_{xy})^T,$$

and

$$\vec{\epsilon} = (\epsilon_{xx} \ \epsilon_{yy} \ \epsilon_{zz} \ 2\epsilon_{yz} \ 2\epsilon_{xz} \ 2\epsilon_{xy})^T,$$

where the strain vector is related to the particle velocity vector components through:

$$\frac{\partial \vec{\epsilon}}{\partial t} = \left(\frac{\partial v_x}{\partial x} \ \frac{\partial v_y}{\partial y} \ \frac{\partial v_z}{\partial z} \ \frac{\partial v_y}{\partial z} + \frac{\partial v_z}{\partial y} \ \frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x} \ \frac{\partial v_y}{\partial x} + \frac{\partial v_x}{\partial y} \right)^T.$$

The system of equations for wave propagation consists of two (dependent) sets of equations. First, the equation of motion (or Newton's second law):

$$\rho \frac{\partial \vec{v}}{\partial t} = \nabla \cdot \vec{S} + \vec{f}, \quad (1)$$

where ρ is the density of the medium, $\nabla \cdot \vec{S}$ denotes divergence of the stress tensor and $\vec{f}$ is a point-force source (source formulations driving the FD simulations will be described in detail later). In an acoustic medium, only the diagonal elements of the stress tensor are non-zero and equal to the negative pressure $-p$.

The second set of equations for wave propagation is the constitutive stress-strain relation:

$$\frac{\partial \vec{\sigma}}{\partial t} = C \frac{\partial \vec{\epsilon}}{\partial t}, \quad (2)$$

where C is the 6-by-6 stiffness matrix. For an elastic isotropic medium, the stiffness matrix takes the form:

$$C = \begin{pmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}, \quad (3)$$

whereas in a general anisotropic medium, all components of the stiffness matrix may be non-zero. As we shall see, this has profound implications on the choice of grid geometry to discretize the constitutive relation. Finally, although redundant, the acoustic constitutive relation can also be written using the notation of the stiffness matrix and the stress-strain

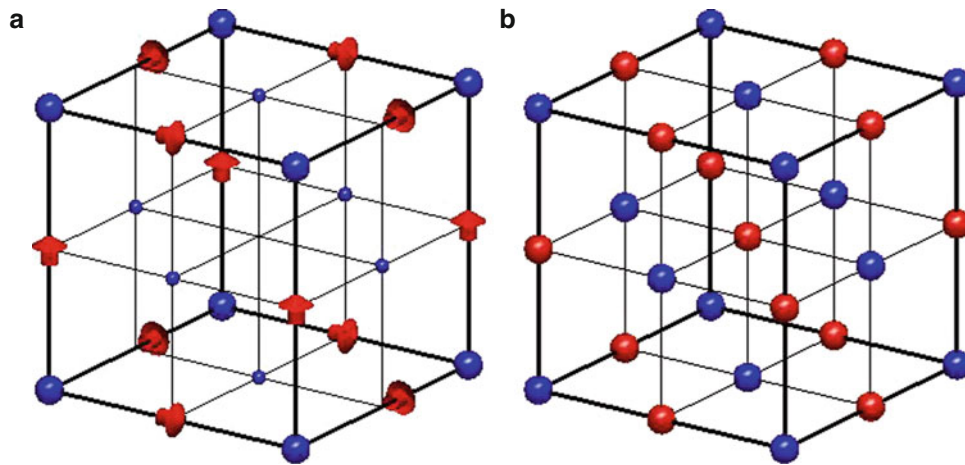
constitutive relation. As noted above, instead of a stress vector with three independent components, we obtain:

$$\vec{\sigma} = -(p \ p \ p \ 0 \ 0 \ 0)^T. \quad (4)$$

In an acoustic medium, only the (3-by-3) upper left quadrant of the stiffness matrix is non-zero, with all elements equal and identical to the bulk modulus of the acoustic medium κ . Clearly this results in three identical equations – the familiar Hooke's law: $\frac{\partial}{\partial t} p = -\kappa \nabla \cdot \vec{v}$.

The FD method comprises solving Eqs. 1 and 2 by discretizing them in time and space and stepping forward in small incremental time steps. In particular, the exact choice of discretization in space turns out to be of fundamental importance. The so-called staggered grid was introduced for modeling isotropic elastic wave propagation by Virieux (1986) and Levander (1988, 1989). In the following, we will be referring to this as the Virieux grid. The choice of this grid results in schemes with attractive stability and dispersion properties, which other non-staggered grids do not necessarily possess. FD approximations to first-order derivatives naturally result in outputting the derivative of the wavefield at a location in between the discretized wavefield. By shifting the exact location of individual wavefield components by half a grid step in certain directions within each grid cell, it is possible to ensure that spatial derivatives can be computed exactly at the locations where they are needed to advance canonical wavefield quantities in time using Eqs. 1 and 2. The structure of the Virieux staggered grid is illustrated in Fig. 1a. The Virieux grid is also a good choice for modeling wave propagation in acoustic media. The resulting sparsely populated staggered-grid cell is equivalent to Fig. 1a but without the shear stresses populating the grid cell (small blue balls in Fig. 1a). In many situations, it is desirable to have a computational model that may contain both acoustic and elastic regions (e.g., modeling a marine seismic experiment or the seismic response of magma chambers). This is straightforward using a Virieux grid, by simply setting $\lambda = \kappa$ and $\mu = 0$ in Eq. 3 in the acoustic regions. Updating the equations will guarantee that the “acoustic stress vector” in Eq. 4 is satisfied to within-machine precision.

The Virieux grid is the natural choice for modeling wave propagation in elastic media as long as the stiffness matrix, C , in Eq. 2 belongs to a class of anisotropic materials referred to as orthorhombic media and as long as the symmetry axes of the medium are aligned with the Cartesian grid (isotropic media is a special case that belongs in this family of media). For this class of anisotropic materials, all elements that are zero in Eq. 3 remain zero. However, as soon as we introduce more complex anisotropic symmetry classes or rotate the symmetry axes with respect to the grid, the elements that are zero in Eq. 3 will become non-zero. As a consequence, Igel et al. (1995) noted that spatial derivatives are no longer



Numerical Methods, Finite Difference, Fig. 1 Staggered-grid geometries. (a) (left): Virieux staggered grid (1986). Large blue ball: normal stress components; small blue ball: shear stress components; red arrows: particle velocity components in the directions indicated. (b) (Right): Lebedev staggered grid (Lisitsa and Vishnevskiy 2010; Bernth

and Chapman 2010). Large blue ball: all stress components; large red ball: all particle velocity components. The authors acknowledge Chris Chapman and Henrik Bernth (Schlumberger Cambridge Research) for providing the figure

available at all locations where they are needed. Igel et al. (1995) solved this by interpolating strains between their natural locations. However, the scheme is both expensive and results in fairly complex implementations. Saenger (2000) solved the problem in 2D by observing that a rotation of the staggered grid by 45° results in a natural choice for staggering the wavefield quantities so that no interpolation is necessary. The so-called rotated staggered scheme can also be generalized to 3D, although it can no longer be seen as a simple rotation of the Virieux grid. Another staggered grid that avoids the necessity to interpolate wavefield quantities in anisotropic media is the Lebedev grid (Lisitsa and Vishnevskiy 2010), which is illustrated in Fig. 1b. Bernth and Chapman (2010) analyzed and compared the different staggered-grid geometries and concluded that whereas the rotated staggered grid and the Lebedev grid are equivalent in 2D, the Lebedev grid is different and a better choice compared to the rotated staggered grid in 3D. The Lebedev grid is both computationally more efficient and also lends itself to a simpler implementation as it can be regarded as a combination of four independent Virieux grids that decouple in isotropic media.

So far, the discussion in this section has concerned lossless media. In order to account for attenuation, a viscoacoustic or viscoelastic model is introduced by using a different constitutive stress–strain relation compared to Eq. 2. Equation 2 can be thought of as a model of springs connecting particles in a lattice. When compressing the springs, energy is stored for an infinite time or until the compressed springs are released after which the particles return to their original positions. In a viscoelastic medium, dashpots are introduced in parallel or in series with the springs causing energy to dissipate after compression. The stiffness matrix becomes time-dependent and the multiplication with the time derivative of the strains is

replaced by a convolution in time. The new viscoelastic constitutive relation becomes:

$$\frac{\partial}{\partial t} \vec{\sigma} = G(t) * \frac{\partial}{\partial t} \vec{\epsilon}, \quad (5)$$

where $G(t)$ is the new relaxation stiffness matrix, which contains elements that are functions of time. Such a particularly useful function corresponds to a spring/dashpot configuration called a standard linear solid (Robertsson et al. 1994). First, due to the exponential kernel (in time) in the relaxation function of a standard linear solid, the convolution in Eq. 5 can be eliminated at the expense of introducing a new set of differential equations to be solved at each time step and spatial location. No additional complications due to various grid geometries arise after introducing the new differential equations solving for the so-called viscoelastic memory variables as these equations are fairly simple ordinary differential equations in time (as opposed to partial-differential equations with additional spatial derivatives). Second, Robertsson et al. (1994) and Blanch et al. (1995) showed how arrays of standard linear solid elements can advantageously be used to model constant quality factor Q versus frequency and also how to model attenuation of P and S waves in elastic isotropic media separately. However, exactly what attenuation to model for different wave types in an anisotropic medium is unclear and remains a topic for research (there are no pure P and S waves in an anisotropic medium).

Finite-Difference Approximations

The FD method approximates derivatives by combining neighboring function values on a grid, where the particular

combination is commonly derived using the Taylor expansion of the function at the different sample points. The simplest example of a FD approximation is the first derivative of a function p at x_0 using only two samples:

$$\frac{\partial p(x_0)}{\partial x} = \frac{1}{\Delta x} \left(p\left(x_0 + \frac{\Delta x}{2}\right) - p\left(x_0 - \frac{\Delta x}{2}\right) \right) + C_1 \frac{\partial^3 p(x_0)}{\partial x^3} (\Delta x)^2 + O((\Delta x)^4), \quad (6)$$

where Δx is the sampling interval and C_1 is a constant. The lowest-order error term in this expansion is proportional to the square of the sampling interval. Hence, as we decrease Δx , the FD approximation will become increasingly accurate (the increase in accuracy proportional to the square of Δx) and in the limit be equivalent to the first derivative of p . The approximation is therefore considered second-order accurate since the lowest-order error term depends on the square of the sampling interval. Higher-order approximations of the first derivative can be expressed as a weighted sum of additional adjacent sample points:

$$\frac{\partial p(x_0)}{\partial x} = \frac{1}{\Delta x} \sum_{i=1}^M \alpha_i \left(p\left(x_0 + \frac{(2i-1)\Delta x}{2}\right) - p\left(x_0 - \frac{(2i-1)\Delta x}{2}\right) \right) + O((\Delta x)^{2M}). \quad (7)$$

The coefficients α_i describe what is often called the FD stencil and are chosen such that terms containing lower-order error factors are canceled (Table 1).

The accuracy of the FD approximations for the temporal derivatives and the spatial derivatives are often different and schemes are accordingly named with two accuracy orders. An $O(2,8)$ accurate scheme would be second-order accurate in time and eighth-order accurate in space.

From Eq. 6, we note that the derivative estimate of a spatial derivative is computed in between sample points of the wavefield quantity in question. As described in the previous section, this is the reason why the staggered-grid formulation is particularly well suited for FD modeling of the system of first-order differential equations describing wave propagation. The grid staggering is also applied in time, typically utilizing the simple second-order accurate approximation in Eq. 6, resulting in a time-stepping method referred to as a “leap-frog” scheme. The leap-frog scheme is in turn referred

to as an explicit FD scheme since the differential equations only contain one unknown after discretization in space and time, and which, therefore, can be computed directly without solving further systems of equations. For example, if we are in the process of updating particle velocities at a new time step $t_0 + \Delta t$ using Eq. 1, we can compute the particle velocities directly from the known particle velocities at time t_0 and through spatial FD approximations of the known stresses at time $t_0 + \Delta t/2$. We then proceed to update stresses at time $t_0 = 3\Delta t/2$ using Eq. 2 explicitly from the stresses at $t_0 + \Delta t/2$ and particle velocities at $t_0 + \Delta t$. In this fashion, we march through the FD simulation until we reach the desired maximum time of the simulation. Clearly, we need to initiate the simulation with so-called initial conditions for particle velocities at time 0 and stresses at time $\Delta t/2$.

Accuracy and Numerical Dispersion

The most visible error introduced through the FD approximation is called numerical dispersion and appears when either the sampling interval is too large or the approximation is not of sufficiently high order. Numerical dispersion is defined as the effect when different frequencies of the wavefields propagate slower or faster compared to the correct speed of wave propagation in the medium. Figures 2 and 3 show wave propagation simulations with two different order schemes using the same spatial sampling and clearly demonstrate severe effects of numerical dispersion of the lower-order scheme.

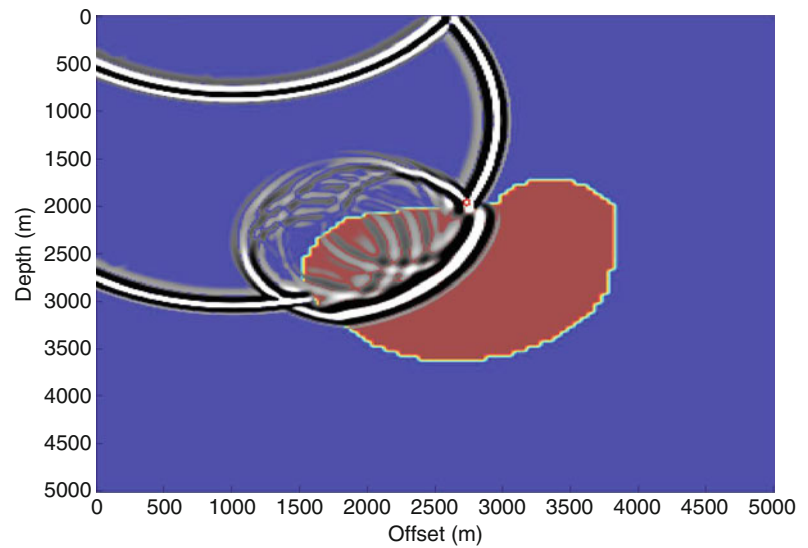
The numerical dispersion can be predicted by Fourier transforming the FD expressions in time and space. Using the Fourier-transformed expressions, we obtain a relationship that shows how the propagation velocity in the medium is affected by numerical dispersion. It is a function of several parameters including the spatial grid step and the so-called Courant number (see stability section below). Figure 4 shows several such dispersion curves (phase velocity error as a function of the number of wavelengths per gridpoint) for different choices of time steps (resulting in different Courant numbers, see below). The numerical dispersion is dominated by the slowest wave speed in the medium, where the wavelength is the shortest. The number of sampling/gridpoints per wavelength necessary for a certain level of accuracy is thus determined by the dispersion curve and the slowest wave speed in a medium.

In practice, it is still difficult to define exactly how fine the wavefields must be sampled to avoid numerical dispersion as the severity of the numerical dispersion depends on the accuracy of the approximation and how far the wave has to propagate. However, as general rules of thumb, in order to roughly limit effects of numerical dispersion within a 2.5% error in propagation velocity, an $O(2,4)$ scheme requires at

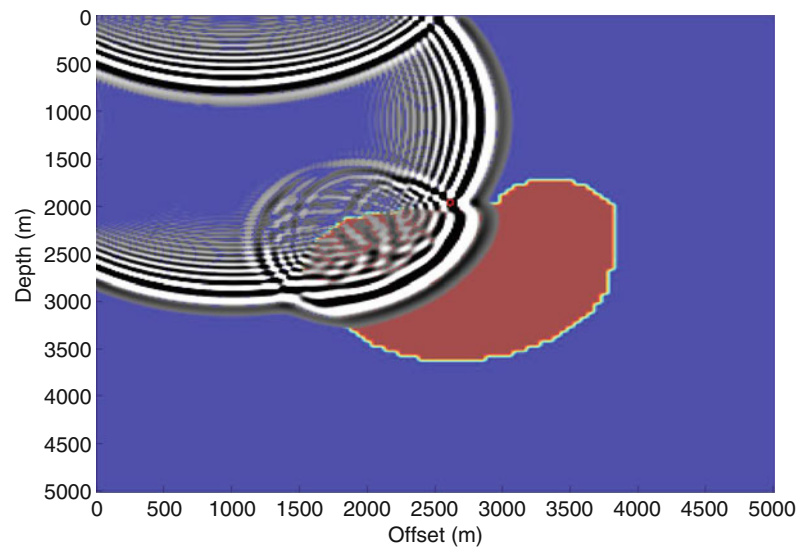
**Numerical Methods,
Finite Difference,
Table 1** Examples of α_i for
different accuracies

Order	α_i		
	$i = 1$	$i = 2$	$i = 3$
2	1	N/A	N/A
4	9/8	-1/24	N/A
6	75/64	-25/384	3/640

Numerical Methods, Finite Difference, Fig. 2 Snapshot in time of a simulation using the acoustic wave equation. The black-and-white wave front indicates the location of a travelling wave. The wave has been reflected in the upper free surface and changed polarity. The left side shows the radiating boundary condition, which in this case is a PML. The blue area has slower propagation velocity than the red area. The solution has been computed using an $O(2,16)$ scheme



Numerical Methods, Finite Difference, Fig. 3 Similar snapshot in time to that shown in Fig. 2. The spatial sampling and medium is the same as in Fig. 2, but in this case the simulation has been performed with an $O(2,2)$ scheme. The trailing ripples of the wave front are typical for numerical dispersion



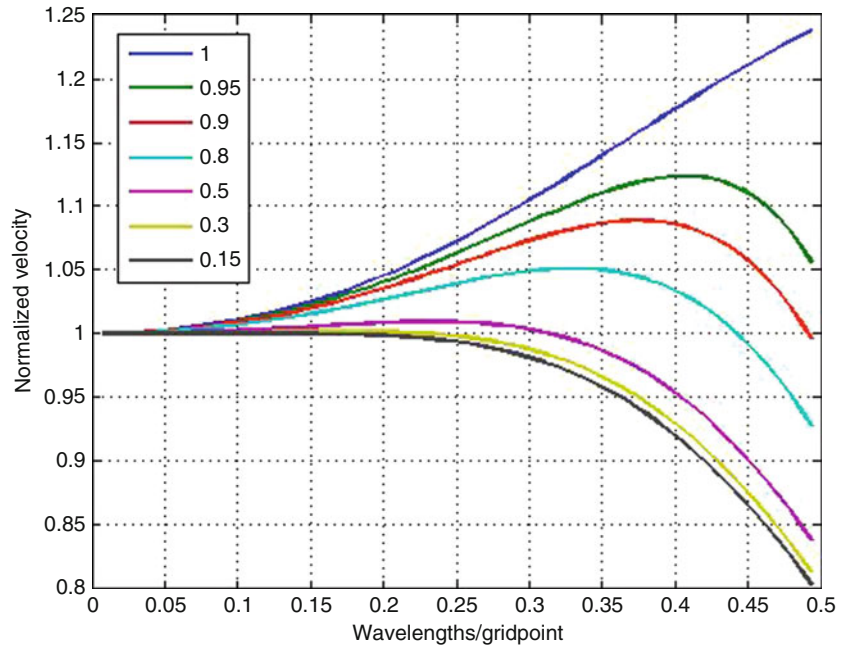
least eight grid points per minimum wavelength, whereas an $O(2,6)$ scheme requires six grid points per minimum wavelength.

It is obvious from Eq. 8 that a higher-order derivative approximation requires more computations than a lower-order approximation. However, in general higher-order approximations (up to some limit) are preferred as the computational requirements increase linearly with accuracy order, whereas a finer grid increases the computational requirement to the fourth power of the refinement in 3D. The fourth power dependence originates from the fact that the temporal step must be proportional to the spatial step (see [Time Stepping and Stability](#) below). Thus, if the spatial step is reduced by a factor n , then a factor n^3 more gridpoints are required in 3D in addition to another factor n more time steps, therefore resulting in an increase of a factor n^4 in the number of computations required.

Lax–Wendroff Corrections, Optimally Accurate FD Schemes and Time Dispersion Correction

The higher-order FD approximations are usually applied to spatial derivatives. To construct a functional scheme that has a higher order of accuracy in time, it is often necessary to use implicit time stepping, which results in a substantial increase in computational requirements. A solution to increase the formal order of accuracy (in a Taylor sense) without resorting to implicit schemes or saving several time levels is to apply a so-called Lax–Wendroff correction (Dablain 1986). The Lax–Wendroff correction works by using the system of equations to express higher-order derivatives in time as spatial derivatives. The approximations of the higher-order time derivatives can be used to construct correction terms to cancel errors of successively higher order. Using the system

Numerical Methods, Finite Difference, Fig. 4 Dispersion curves for a $O(2,6)$ FD Scheme. A perfect solution should be equal to one. The dispersion curves vary depending on the ratio of the velocity of the medium and the maximum velocity used to determine the time step, that is, the (normalized) Courant number displayed in the legend. To yield a sufficiently accurate solution, this scheme should probably not be run with less than five to seven gridpoints per minimum wavelength (0.14–0.2 wavelengths/gridpoint)



$$\begin{cases} \frac{\partial p}{\partial t} = -\kappa \frac{\partial v}{\partial x} \\ \frac{\partial v}{\partial t} = -\frac{1}{\rho} \frac{\partial p}{\partial x} \end{cases} \quad (8)$$

the corresponding Lax–Wendroff correction for a third-order time derivative is:

$$\frac{\partial^3 p}{\partial t^3} = \kappa \frac{\partial^3 v}{\partial x \partial t^2} = \kappa \frac{\partial}{\partial x} \frac{1}{\rho} \frac{\partial^2 p}{\partial x \partial t} = \kappa \frac{\partial}{\partial x} \frac{1}{\rho} \frac{\partial}{\partial x} \kappa \frac{\partial v}{\partial x}. \quad (9)$$

However, the method is somewhat cumbersome in multi-dimensions as it requires mixed derivatives and also derivatives of the material parameters, that is, ρ and κ in Eqs. 8 and 9 (see for instance Blanch and Robertsson (1997) for applications of the Lax–Wendroff correction).

The FD stencil can also be adjusted such that it is optimal for a certain frequency/wavelength range and thus achieves a higher accuracy than a standard scheme (Holberg 1987; Geller and Takeuchi 1995; Zingg 2000). Note that in a strict (and in this case misleading) Taylor sense, these schemes do not necessarily appear to be high-order schemes.

Fornberg (1987) noted that spatial and temporal dispersion can be considered as separate issues that can be minimized separately. Dispersion associated with the approximation of the time derivatives is not a function of the propagation path or type (e.g., P or S) of the synthesized data but only of the number of time-steps in the simulation since the source was introduced and can therefore be completely removed from FD computed data with a pre- and post-processing filtering procedure (Stork 2013; Koene et al. 2018). The computational

cost of obtaining this spectral accuracy in time is negligible. It can also be used in combination with Lax–Wendroff related schemes (e.g., Amundsen and Pedersen 2017) which generally also allow for larger time-steps without violating the stability condition of the FD scheme. Thus, stability may be increased without increasing time dispersion.

Time Stepping and Stability

As discussed above, the time stepping is most commonly explicit using a second-order approximation resulting in a leap-frog scheme. The explicit time stepping introduces a so-called CFL (Courant–Friedrich–Levy) stability condition prescribing the maximum size of the time step Δt . The CFL condition can be derived from the same Fourier transformation as described above (in the context of numerical dispersion) and requiring the resulting difference equation to have roots of absolute value less than or equal to one. As a result, the maximum time step is determined by the fastest wave speed in a medium, the spatial step, and a constant depending on the particular FD approximations used:

$$\Delta t \leq \frac{K}{\sqrt{D}} \frac{\Delta x}{c_{\max}}, \quad (10)$$

where the constant K depends on the spatial accuracy and is slowly decreasing with increasing accuracy of the scheme from a value of 1 for an $O(2,2)$ scheme, D is the dimension of the simulation (e.g., $D = 2$ for 2D) and $c_{\max}$ is the maximum wave propagation speed in the medium. Rapid changes

in material parameters may require an even smaller time step, but do not, in general, cause problems (Haney 2007). Using a temporal step larger than determined by the CFL condition yields an unstable simulation, that is, parasitic solutions will grow at an exponential rate and swamp the physical solution.

The Courant number is often used in the context of discussing stability and numerical dispersion. It is defined as

$$\gamma = c_0 \frac{\Delta t}{\Delta x}, \quad (11)$$

where c_0 is the local velocity of the medium. The Courant number can be interpreted physically as the fraction of the spatial increment that a wave can advance in a time step. From Eq. 11, we see that the Courant number controls the stability of the numerical scheme. In a heterogeneous medium, there exists a large range of velocities and it is thus important to study a FD scheme's behavior (i.e., numerical dispersion) for a range of Courant numbers (e.g., see Fig. 4). In the literature, the normalized Courant number is often used:

$$\gamma_{norm} = c_0 \frac{\Delta t}{\Delta x} \frac{\sqrt{D}}{K}, \quad (12)$$

which needs to be less than or equal to 1 for a stable simulation.

Amundsen and Pedersen (2017) showed that through a Lax-Wendroff related procedure it is possible to choose time steps that can be doubled, tripled, or generally, n -tupled beyond the classical CFL limit. However, like the Lax-Wendroff procedure, they require more computational effort per time step.

Boundary Conditions

The most important boundary conditions for wave simulations are radiating and free-surface boundary conditions. Radiating boundary conditions are applied at the edge of the finite computational domain to annihilate undesired reflections from the edge of the domain. The radiating boundary conditions can be implemented as “one-way” wave equation propagators using special operators at the boundaries. A first-order simple version is to implement the following equation at the boundary:

$$\frac{\partial p}{\partial t} = \pm c_0 \frac{\partial p}{\partial x}, \quad (13)$$

where the sign depends on which boundary (maximum or minimum x coordinate) and c_0 is the (local) propagation velocity. For a single propagation velocity, Eq. 13 is fairly straightforward to implement. However, significant complications arise for elastic or anisotropic media and for higher dimensions than 1D where it is the apparent velocity

perpendicular to the boundary that is the relevant parameter (Clayton and Engquist 1977; Higdon 1986, 1990).

The radiating boundary conditions are more commonly implemented as absorbing (or attenuating) boundary conditions applied in a finite region just inside the computational domain (Cerjan et al. 1985). Absorbing boundaries do, however, need to have a transition zone, where the attenuation of the wavefield is gradually increased, since a rapid change in attenuation will cause reflections as well. A particularly efficient absorbing boundary condition is the “Perfectly Matched Layer” (PML) (Bérenger 1994; Collino and Tsogka 2001). The PML allows the attenuation to be increased extremely rapidly without causing reflections. The PML boundary conditions can be thought of as introducing artificial anisotropy to the boundary and matching the reflection coefficient in different propagation directions. The most straightforward realization of a PML boundary is through the split-field technique, where the wavefield is split depending on which spatial derivatives are used to update it. For the system in Eq. 8 in two dimensions, an implementation for the p variable would be

$$\begin{cases} \frac{\partial p_x}{\partial t} + \beta_x p = -K \frac{\partial v_x}{\partial x} \\ \frac{\partial p_z}{\partial t} + \beta_z p = -K \frac{\partial v_z}{\partial z} \\ p = p_x + p_z \end{cases} \quad (14)$$

The coefficients β_i control the attenuation in either the x or z directions independently.

The straightforward split-field implementation is equivalent to a fully anisotropic implementation (Teixeira et al. 2002) but suffers from a mild instability for waves impinging on the boundary at an incidence angle higher than 45° . A slightly more complex implementation avoids this instability (Bécache et al. 2002).

The free-surface boundary condition is used to approximate the interface between a water/solid material and air. Since the acoustic impedance of air is significantly different from that of water or the Earth, explicit modeling results in a computationally much more expensive simulation. In fact, it is possible to use a (pressure release) free-surface condition instead (water or Earth in contact with vacuum). For a water/air contact it is simply implemented through mirroring,

$$\begin{cases} p(-z) = -p(z) \\ v_z(-z) = v_z(z) \end{cases}, \quad (15)$$

assuming the boundary is located at $z = 0$. The upper boundary in Fig. 2 is implemented using a free-surface condition. The free-surface condition for a solid elastic material is more complicated and cannot be achieved solely through mirroring. Care must be taken to ensure that all wavefield components satisfy the wave equation and the free-surface condition in

order to accurately model wave phenomena such as Rayleigh waves at the free surface (see Robertsson 1996, for details).

Topography and Conformal Mapping of Grids

Topography can either be implemented explicitly into a FD scheme (Robertsson 1996) or by stretching the grid through conformal mapping of the regular computational grid onto a grid with continuously varying grid spacings where the top of the grid follows the topography (Fornberg 1988; Hestholm and Ruud 1994). The explicit method will lose some of its flexibility to implement rapidly varying topography if a higher-order scheme is used for the solution. If a conformal mapping is used, all derivatives will depend on the mapping function through the chain-rule. This may lead to stricter stability conditions and stricter requirements on spatial sampling.

Source Implementations

Seismic sources can be introduced in terms of body forces as indicated in Eq. 1 or as moment-tensor sources using an equivalent source term in the constitutive equation (e.g., Eqs. 2 or 5). A straightforward approach is to simply drive a body force at one point by feeding Eq. 1 with the value of a desired source wavelet in a single spatial grid point at each time step. Although this approach tends to work fairly well, researchers have sometimes reported on high-frequency noise being introduced. The origin of this noise has been attributed to the intrinsic difficulty in representing delta functions on a discretely sampled model with continuous wavefields.

Practitioners in exploration seismology often mimic explosive sources by equal excitation of the diagonal elements of the stress tensor at the source location or alternatively model compressional or shear vibroseis source by either exciting vertical or horizontal body forces. Earthquake seismologists who wish to model a moment-tensor source can do this either through a stress or body-force representation as described by Moczo et al. (2007).

An alternative way to introduce a source wavefield on a FD grid is by means of a boundary condition on an artificial surface interior to the grid. The method was first described by Alterman and Karal (1968) and followed by many authors clarifying the concept (Kelly et al. 1976; Levander 1989; Robertsson and Chapman 2000; Moczo et al. 2007). The technique relies on two prerequisites. First, that the spatial FD operators have a limited extent in space and second, that the (isolated) source wavefield can be computed in a region around the source location without reverberations from the rest of the model at the (simulated) time of source insertion.

The method of introducing a source wavefield as a boundary condition along an internal surface can also be used in seismic data processing and imaging applications (e.g.,

Amundsen and Robertsson 2014) and in so-called hybrid methods where two different numerical solutions are coupled together (e.g., Robertsson et al. 1996; Zahradník and Moczo 1996; Robertsson and Chapman 2000; van Manen et al. 2007). However, the wavefield on opposite sides of the artificial surface in the technique above is reversed so that the source wavefield is present inside the artificial surface only.

Finally, Robertsson et al. (2015) showed that by introducing a source wavefield as a boundary condition, it is possible to output up- and down-going as well as P- and S-wave constituents to machine precision accuracy. This enables the use of FD methods in research where such outputs are required instead of relying on so-called propagator matrix methods that are limited to 1D media.

Conclusions

We have outlined the basic principles of FD approximations to the systems of first-order partial-differential equations describing acoustic, isotropic elastic, anisotropic elastic, or viscoacoustic/elastic wave propagation. The method is based on explicit leap-frog time-stepping and staggered-grid representations. Most commonly Taylor-series-derived stencils are used to approximate the spatial derivatives, although other means to optimize FD stencils have also been described.

Two types of boundary conditions exist. First, radiating or absorbing boundary conditions (e.g., PML) are used to truncate the model while avoiding boundary reflections. Second, a free-surface boundary condition can be used to simulate the surface of the Earth with or without topography.

Sources are introduced either as point-force or stress sources, or along closed (artificial) surfaces inside the FD grid.

The maximum possible spatial grid size in order to limit numerical error is computed from the slowest propagation velocity in the model (e.g., a shear-wave velocity in the near-surface) using the so-called numerical dispersion relation for the FD scheme. Once the grid spacing has been chosen, the maximum possible time step is computed from the CFL-stability condition. The maximum possible time step is proportional to the spatial grid size and inversely proportional to the maximum propagation velocity in the grid. Since the computational cost of a 3D FD simulation increases to the power of four with a linear reduction of grid size, high-order accurate FD stencils are usually preferred.

Cross-References

- [Earthquake Source Theory](#)
- [Magnetic Anisotropy](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Imaging, Overview](#)
- [Seismic Viscoelastic Attenuation](#)

- Seismic, Migration
- Seismic, Waveform Modeling and Tomography

Bibliography

- Alterman Z, Karal FC (1968) Propagation of elastic waves in layered media by finite-difference methods. *Bull Seismol Soc Am* 58:367–398
- Amundsen L, Pedersen Ø (2017) Time step n-tupling for wave equations. *Geophysics* 82:T249–T254
- Amundsen L, Robertsson JOA (2014) Wave equation processing using finite-difference propagators, part 1: wavefield dissection and imaging of marine multicomponent seismic data. *Geophysics* 79:T287–T300
- Bécache E, Petropoulos P, Gedney S (2002) On the long-time behavior of unsplit perfectly matched layers. INRIA report de Recherche, No 4538. [ftp://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-4538.pdf](http://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-4538.pdf)
- Béranger JP (1994) A perfectly matched layer for the absorption of electromagnetic waves. *J Comput Phys* 114:185–200
- Berth H, Chapman CH (2010) A comparison of Lebedev and rotated staggered grids for anisotropic finite difference modeling. In: Extended abstracts from the 72nd EAGE conference and exhibition
- Blanch JO, Robertsson JOA (1997) A modified Lax–Wendroff correction for wave propagation in media described by Zener elements. *Geophys J Int* 131:381–386
- Blanch JO, Robertsson JOA, Symes WW (1995) Modeling of constant Q: methodology and algorithm for an efficient and optimally inexpensive viscoelastic technique. *Geophysics* 60:176–184
- Cerjan C, Kosloff D, Kosloff R, Reshef M (1985) A non-reflecting boundary condition for discrete acoustic and elastic wave equations. *Geophysics* 50:705–708
- Clayton RW, Engquist B (1977) Absorbing boundary conditions for acoustic and elastic wave equations. *Bull Seismol Soc Am* 6:1529–1540
- Collino F, Tsogka C (2001) Application of the PML absorbing layer model to the linear elastodynamic problem in anisotropic heterogeneous media. *Geophysics* 66:294–307
- Dablain MA (1986) The application of high-order differencing to the scalar wave equation. *Geophysics* 51:54–66
- Fornberg B (1987) The pseudospectral method: comparison with finite differences for the elastic wave equation. *Geophysics* 52:483–501
- Fornberg B (1988) Generation of finite difference formulas on arbitrary spaced grids. *Math Comput* 51:699–706
- Geller RJ, Takeuchi N (1995) A new method for computing highly accurate DSM synthetic seismograms. *Geophys J Int* 123:449–470
- Haney M (2007) Generalization of von Neumann analysis for a model of two discrete half-spaces: the acoustic case. *Geophysics* 72:SM35–SM46
- Hestholm S, Ruud BO (1994) 2-D finite difference elastic wave modeling including surface topography. *Geophys Prospect* 42:371–390
- Higdon RL (1986) Absorbing boundary conditions for difference approximations to the multi-dimensional wave equation. *Math Comput* 47:437–459
- Higdon RL (1990) Radiating boundary conditions for elastic wave propagation. *SIAM J Numer Anal* 27:831–870
- Holberg O (1987) Computational aspects of the choice of operator and sampling interval for numerical differentiation in large-scale simulation of wave phenomena. *Geophys Prospect* 35:629–655
- Igel H, Mora P, Rioulet B (1995) Anisotropic wave propagation through finite-difference grids. *Geophysics* 60:1203–1216
- Kelly KR, Ward RW, Treitel S, Alford RM (1976) Synthetic seismograms: a finite-difference approach. *Geophysics* 41:2–27
- Koene EFM, Robertsson JOA, Broggini F, Andersson F (2018) Eliminating time dispersion from seismic wave modeling. *Geophys J Int* 213:169–180
- Levander AR (1988) Fourth-order finite-difference P-SV seismograms. *Geophysics* 53:1425–1435
- Levander AR (1989) Finite-difference forward modeling in seismology. In: James DE (ed) *The encyclopedia of solid earth geophysics*. Van Nostrand Reinhold, New York, pp 410–431
- Lisitsa V, Vishnevskiy D (2010) Lebedev scheme for the numerical simulation of wave propagation in 3D anisotropic elasticity. *Geophys Prospect* 58:619–645
- Moczo P, Robertsson JOA, Eisner L (2007) The finite-difference time-domain method for modelling of seismic wave propagation. In: Wu RS, Maupin V (eds) *Advances in wave propagation in heterogeneous Earth*, vol 48, *Advances in geophysics* (ed Dmowska R). Elsevier-Pergamon, Oxford, pp 421–516
- Robertsson JOA (1996) A numerical free-surface condition for elastic/viscoelastic finite-difference modeling in the presence of topography. *Geophysics* 61:1921–1934
- Robertsson JOA, Chapman CH (2000) An efficient method for calculating finite-difference seismograms after model alterations. *Geophysics* 65:907–918
- Robertsson JOA, Blanch JO, Symes WW (1994) Viscoelastic finite-difference modeling. *Geophysics* 59:1444–1456
- Robertsson JOA, Levander A, Holliger K (1996) Modeling of the Acoustic Reverberation Special Research Program deep ocean seafloor scattering experiments using a hybrid wave propagation simulation technique. *J Geophys Res* 101:3085–3101
- Robertsson JOA, van Manen DJ, Schmeltzbach C, van Renterghem C, Amundsen L (2015) Finite-difference modelling of wavefield constituents. *Geophys J Int* 203:1334–1342
- Saenger EH (2000) Wave propagation in fracture media: theory and applications of the rotated staggered finite-difference grid. PhD thesis, Karlsruhe University
- Stork C (2013) Eliminating nearly all dispersion error from FD modeling and RTM with minimal cost increase. In: Extended abstracts from the 75th EAGE conference and exhibition
- Teixeira FL, Moss CD, Chew WC, Kong JA (2002) Split-field and anisotropic medium PML-FDTD implementations for inhomogeneous media. *IEEE Trans Microwave Theory* 50:30–35
- van Manen DJ, Robertsson JOA, Curtis A (2007) Exact wave field simulation for finite-volume scattering problems. *J Acoust Soc Am* 122:EL115–EL121
- Virieux J (1986) P-SV wave propagation in heterogeneous media: velocity-stress finite-difference method. *Geophysics* 51:889–901
- Zahradnik J, Moczo P (1996) Hybrid seismic modeling based on discrete-wavenumber and finite-difference methods. *Pure Appl Geophys* 148:21–38
- Zingg DW (2000) Comparison of high-accuracy finite-difference methods for linear wave propagation. *SIAM J Sci Comput* 22:476–502

Numerical Methods, Finite Element

J. N. Reddy
Department of Mechanical Engineering, Texas A&M
University, College Station, TX, USA

Definition

The finite element method is a numerical method, like the finite difference method, for solving differential equations arising in the study of physical phenomena. In the finite

element method, a given domain is viewed as a set of non-intersecting subdomains, called *finite elements*, and over each element the governing equation is approximated by any of the traditional variational methods (e.g., Ritz, Galerkin, and least-squares methods). The main reason behind seeking approximate solution on a set of elements is the fact that it is easier to represent a complicated function as a sum of simple polynomials. Of course, each individual segment of the solution should fit with its neighbors in the sense that the function and possibly its derivatives up to a chosen order are continuous along the interface between elements. This entry is a brief introduction to the finite element modeling of diffusion processes and the Navier–Stokes equations governing viscous incompressible fluids. Both phenomena play crucial role in modeling a variety of geological or geomechanics processes.

General Introduction

Scientists and engineers model (i.e., develop mathematical models and numerically simulate) natural phenomena with the objective of understanding it. Virtually every phenomenon in nature, whether aerospace, biological, chemical, geological, or mechanical, can be described in terms of algebraic, differential, and/or integral equations relating various quantities of interest. Determining the stress distribution in a mechanical structure with oddly shaped members and numerous stiffeners and subjected to mechanical, thermal, and/or aerodynamic loads, finding the concentration of pollutants in lakes and estuaries or in the atmosphere, predicting geophysical and geological events, and simulating weather in an attempt to predict the formation of tornadoes and thunderstorms are a few examples of many important practical problems that scientists and engineers are occupied with.

In the last four decades, the finite element method has emerged as a powerful computational tool for solving non-linear partial differential equations over irregular domains with complex domain properties (Reddy 2015, 2019; Reddy and Gartling 2010; Bathe 1996; Belytschko et al. 2000). Applications of the method to geology, geophysics, and geomechanics problems are numerous (see, e.g., Reddy et al. 1982; Wickham et al. 1982; Bird 1989; Parsons 2002; Dyksterhuis et al. 2005 and references therein). This brief introduction is meant to provide some insight into the workings of the finite element method as applied to the Poisson equation as well as the Navier–Stokes equations that are used to model certain geomechanics processes in two dimensions.

Finite Element Model of 2-D Poisson Equation

Consider the problem of finding $u(x, y)$ such that the following partial differential equation (a generalized Poisson's equation) is satisfied

$$-\left[\frac{\partial}{\partial x}\left(a_{xx}\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a_{yy}\frac{\partial u}{\partial y}\right)\right] = f(x, y) \text{ in } \Omega \quad (1)$$

where Ω is a two-dimensional domain with boundary Γ . Here a_{xx} and a_{yy} are material coefficients in the x and y directions, respectively, and $f(x, y)$ is the known source. For example, in a ground water flow problem u denotes the water head (i.e., velocity potential), a_{xx} and a_{yy} are the permeabilities in the x and y directions, respectively, and $f(x, y)$ is distributed water source. Equation 1 also arises in many areas of science and engineering.

In the finite element method, the domain $\bar{\Omega} = \Omega \cup \Gamma$ is divided into a set of subdomains $\bar{\Omega}^e = \Omega^e \cup \Gamma^e$, called finite elements. Any geometric shape qualifies as an element, provided that the approximation functions ψ_i^e can be derived uniquely for the shape.

Suppose that the dependent unknown u is approximated over a typical finite element $\bar{\Omega}^e$ by the expression

$$u(\mathbf{x}) \approx u_h^e(\mathbf{x}) = \sum_{j=1}^n u_j^e \psi_j^e(\mathbf{x}), \mathbf{x} = (x, y) \in \bar{\Omega}^e \quad (2)$$

where $u_h^e(\mathbf{x})$ represents an approximation of $u(\mathbf{x})$ over the element $\bar{\Omega}^e$, parameters u_j^e denote the values of the function $u_h^e(\mathbf{x})$ at a selected number of points (i.e., element nodes) in the element $\bar{\Omega}^e$, and ψ_j^e are the Lagrange interpolation functions associated with the element.

We seek to satisfy the governing differential equation in a weak-form sense, with the weight functions being the same as the approximation functions. The resulting finite element model, i.e., set of algebraic equations) is called the *weak-form Galerkin finite element model* (Reddy 2019). The weak form is

$$0 = \int_{\Omega^e} \left(a_{xx} \frac{\partial w_i^e}{\partial x} \frac{\partial u_h^e}{\partial x} + a_{yy} \frac{\partial w_i^e}{\partial y} \frac{\partial u_h^e}{\partial y} - w_i^e f \right) dx dy - \oint_{\Gamma^e} w_i^e q_n ds \quad (3)$$

Suppose that u_h^e is represented over a typical finite element Ω^e by expression of the form 2. Substituting the finite element approximation 2 into the weak form 3, we obtain

$$\mathbf{K}^e \mathbf{u}^e = \mathbf{f}^e + \mathbf{q}^e \equiv \mathbf{F}^e \quad (4)$$

where the coefficients K_{ij}^e , f_i^e and q_i^e are defined by

$$\begin{aligned}
K_{ij}^e &= \int_{\Omega^e} \left(a_{xx} \frac{\partial \psi_i^e}{\partial x} \frac{\partial \psi_j^e}{\partial x} + a_{yy} \frac{\partial \psi_i^e}{\partial y} \frac{\partial \psi_j^e}{\partial y} \right) dA \\
&+ \oint_{\Gamma^e} h_c \psi_i^e \psi_j^e ds \\
f_i^e &= \int_{\Omega^e} f \psi_i^e dA, q_i^e = \int_{\Gamma^e} (\hat{q}_n + h_c u_c) \psi_i^e ds
\end{aligned} \quad (5)$$

We note that $K_{ij}^e = K_{ji}^e$ (i.e., $\mathbf{K}^e$ is symmetric).

Finite Element Models of the 2-D Navier–Stokes Equations

In this section, we develop the finite element models of steady flows of viscous, incompressible fluids in two-dimensional domains. The governing equations are the conservation of linear momentum and conservation of mass, expressed in terms of the Cartesian components

$$\rho \left(u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} \right) - \frac{\partial \sigma_{xx}}{\partial x} - \frac{\partial \sigma_{xy}}{\partial y} - f_x = 0 \quad (6)$$

$$\rho \left(u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} \right) - \frac{\partial \sigma_{xy}}{\partial x} - \frac{\partial \sigma_{yy}}{\partial y} - f_y = 0 \quad (7)$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \quad (8)$$

$$\begin{aligned}
\sigma_{xx} &= 2\mu \frac{\partial u}{\partial x} - P, \sigma_{yy} = 2\mu \frac{\partial v}{\partial y} - P, \\
\sigma_{xy} &= \mu \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right)
\end{aligned} \quad (9)$$

Here we present two different finite element models associated with Eqs. 6, 7, 8, and 9. The first one is a direct formulation in which the three equations in (u, v, P) are used in their original form. This formulation is known as the *velocity–pressure formulation*. The other formulation is based on the interpretation that the continuity equation (Eq. 8) is a constraint on u and v , and the constraint is satisfied in a least-squares (i.e., approximate) sense. This particular method of including the constraint in the formulation is known as the *penalty function method*, and the model is termed as the *penalty-finite element model*. It is informative to note that the velocity–pressure formulation is the same as the Lagrange multiplier formulation, wherein the constraint is included by means of the Lagrange multiplier. The Lagrange multiplier turns out to be the negative of the pressure.

Velocity–Pressure (Mixed) Model

The weak forms of Eqs. 6, 7, and 8 over an element Ω^e can be constructed following the standard procedure (Reddy 2015). The weight functions have the following physical interpretation:

$$w_1 \sim u, w_2 \sim v \text{ and } w_3 \sim -P \quad (10)$$

Assuming approximations of the form

$$u = \sum_{j=1}^m v_j \psi_j^e, \quad v = \sum_{j=1}^m v_j \psi_j^e, \quad P = \sum_{j=1}^n P_j \phi_j^e \quad (11)$$

where ψ_j^e and ϕ_j^e denote the Lagrange type interpolation functions ($n < m; n + 1 = m$) and substituting into the weak forms of Eqs. 6, 7, and 8, we obtain the following finite element model:

$$\begin{bmatrix} \mathbf{K}^{11} & \mathbf{K}^{12} & \mathbf{K}^{13} \\ \mathbf{K}^{21} & \mathbf{K}^{22} & \mathbf{K}^{23} \\ \mathbf{K}^{31} & \mathbf{K}^{32} & \mathbf{K}^{33} \end{bmatrix} \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{P} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F}^1 \\ \mathbf{F}^2 \\ \mathbf{F}^3 \end{Bmatrix} \quad (12)$$

where $F_i^3 = 0$

$$\begin{aligned}
K_{ij}^{11} &= \int_{\Omega^e} \left[\mu \left(2 \frac{\partial \psi_i}{\partial x} \frac{\partial \psi_j}{\partial x} + \frac{\partial \psi_i}{\partial y} \frac{\partial \psi_j}{\partial y} \right) \right. \\
&\quad \left. + \rho \psi_i \left(u \frac{\partial \psi_j}{\partial x} + v \frac{\partial \psi_j}{\partial y} \right) \right] dx dy \\
K_{ij}^{12} &= \int_{\Omega^e} \mu \frac{\partial \psi_i}{\partial y} \frac{\partial \psi_j}{\partial x} dx dy, \\
K_{ij}^{13} &= - \int_{\Omega^e} \frac{\partial \psi_i}{\partial x} \psi_j dx dy, \quad K_{ij}^{33} = 0 \\
K_{ij}^{22} &= \int_{\Omega^e} \left[\mu \left(\frac{\partial \psi_i}{\partial x} \frac{\partial \psi_j}{\partial x} + 2 \frac{\partial \psi_i}{\partial y} \frac{\partial \psi_j}{\partial y} \right) \right. \\
&\quad \left. + \rho \psi_i \left(u \frac{\partial \psi_j}{\partial x} + v \frac{\partial \psi_j}{\partial y} \right) \right] dx dy \\
K_{ij}^{23} &= - \int_{\Omega^e} \frac{\partial \psi_i}{\partial y} \psi_j dx dy, \quad K_{ij}^{21} = K_{ji}^{21}, \\
K_{ij}^{31} &= K_{ji}^{13}, \quad K_{ij}^{32} = K_{ji}^{23} \\
F_i^1 &= \int_{\Omega^e} f_x \psi_i dx dy + \oint_{\Gamma^e} t_x \psi_i ds, \\
F_i^2 &= \int_{\Omega^e} f_y \psi_i dx dy + \oint_{\Gamma^e} t_y \psi_i ds
\end{aligned} \quad (13)$$

Here (t_x, t_y) denote the components of the stress vector $\mathbf{t}$ on the boundary

$$t_x = \sigma_{xx} n_x + \sigma_{xy} n_y, \quad t_y = \sigma_{xy} n_x + \sigma_{yy} n_y \quad (14)$$

and (n_x, n_y) are the components of the unit normal vector $\hat{\mathbf{n}}$.

Penalty-Finite Element Model

Use of the penalty function method amounts to replacing the pressure with

$$P = -\gamma \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) \quad (15)$$

where γ is known as the penalty parameter. For complete details, the reader may consult (Reddy 2015, 2019; Reddy and Gartling 2010). Note that Eq. 15 is used to post-compute P once the velocity field is available.

Assuming interpolation of the form

$$u = \sum_{j=1}^n u_j^e \psi_j^e, \quad v = \sum_{j=1}^n v_j^e \psi_j^e \quad (16)$$

where ψ_i^e are Lagrange interpolation functions. Substituting Eq. 16 into the weak forms of Eqs. 6 and 7, we obtain the finite element model

$$\begin{bmatrix} \mathbf{K}^{11} & \mathbf{K}^{12} \\ \mathbf{K}^{21} & \mathbf{K}^{22} \end{bmatrix} \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F}^1 \\ \mathbf{F}^2 \end{Bmatrix} \quad (17)$$

where

$$\begin{aligned} K_{ij}^{11} &= \int_{\Omega^e} \left[\mu \left(2 \frac{\partial \psi_i^e}{\partial x} \frac{\partial \psi_j^e}{\partial x} + \frac{\partial \psi_i^e}{\partial y} \frac{\partial \psi_j^e}{\partial y} \right) \right. \\ &\quad \left. + \rho \psi_i^e \left(u \frac{\partial \psi_j^e}{\partial x} + v \frac{\partial \psi_j^e}{\partial y} \right) \right] dx dy + \int_{\Omega^e} \gamma \frac{\partial \psi_i^e}{\partial x} \frac{\partial \psi_j^e}{\partial x} dx dy \\ K_{ij}^{12} &= \int_{\Omega^e} \mu \frac{\partial \psi_i^e}{\partial y} \frac{\partial \psi_j^e}{\partial x} dx dy + \int_{\Omega^e} \gamma \frac{\partial \psi_i^e}{\partial x} \frac{\partial \psi_j^e}{\partial y} dx dy \\ K_{ij}^{22} &= \int_{\Omega^e} \left[\mu \left(\frac{\partial \psi_i^e}{\partial x} \frac{\partial \psi_j^e}{\partial x} + 2 \frac{\partial \psi_i^e}{\partial y} \frac{\partial \psi_j^e}{\partial y} \right) \right. \\ &\quad \left. + \rho \psi_i^e \left(u \frac{\partial \psi_j^e}{\partial x} + v \frac{\partial \psi_j^e}{\partial y} \right) \right] dx dy + \int_{\Omega^e} \gamma \frac{\partial \psi_i^e}{\partial y} \frac{\partial \psi_j^e}{\partial y} dx dy \\ F_i^1 &= \int_{\Omega^e} f_x \psi_i^e dx dy + \oint_{\Gamma^e} t_x \psi_i^e ds, \\ F_i^2 &= \int_{\Omega^e} f_y \psi_i^e dx dy + \oint_{\Gamma^e} t_y \psi_i^e ds \end{aligned} \quad (18)$$

The numerical evaluation of the coefficient matrices appearing in Eq. 17 requires special consideration. Eq. 17 is of the general form

$$(\mathbf{K}^\mu + \mathbf{K}^\rho + \mathbf{K}^\gamma) \Delta = \mathbf{F} \quad (19)$$

where $\mathbf{K}^\mu$, $\mathbf{K}^\rho$, and $\mathbf{K}^\gamma$ denote the contributions from the viscous, inertia, and penalty terms, respectively. In theory, as we increase the value of the penalty parameter γ , the conservation of mass is satisfied more exactly. However, in

practice, for some large value of γ , the contribution from the viscous and inertia terms would be negligibly small compared to the penalty terms in the computer. Thus, if $\mathbf{K}^3$ is a nonsingular (i.e., invertible) matrix, the solution of the final equations associated with Eq. 19 for a large value of γ is trivial

$$\lim_{\gamma \rightarrow 0} (\mathbf{K}^\mu + \mathbf{K}^\rho + \mathbf{K}^\gamma) \Delta = \mathbf{F} \rightarrow \mathbf{K}^\gamma \Delta = \frac{1}{\gamma} \mathbf{F} \quad (20)$$

which will yield $\Delta = 0$. While the trivial solution satisfies the continuity equation, it does not satisfy the momentum equations for nontrivial boundary data or body forces. In this case, the discrete problem (Eq. 17) is said to be over-constrained or “locked.” If $\mathbf{K}^\gamma$ is singular, then the sum $\mathbf{K}^\mu + \mathbf{K}^\rho + \mathbf{K}^\gamma$ is nonsingular (because $\mathbf{K}^\mu$ is nonsingular after the imposition of proper boundary conditions), and a nontrivial solution to the problem may be obtained.

The numerical problem described above is eliminated by proper evaluation of the integrals in $\mathbf{K}^\gamma$. It is found that if the coefficients of $\mathbf{K}^\gamma$ (i.e., penalty terms) are evaluated using a numerical integration rule of one order less than that required to integrate them exactly, the finite element equations (Eq. 17) will give acceptable solutions for the velocity field. This technique of under-integrating the penalty terms is known in the literature as *reduced (order) integration*. For example, if a linear rectangular element is used to approximate the velocity field in a two-dimensional problem, the matrix coefficients $\mathbf{K}^\mu$ as well as $\mathbf{K}^\rho$ are evaluated using the 2×2 Gauss quadrature, and $\mathbf{K}^\gamma$ are evaluated using the one-point 1×1 Gauss quadrature. The one-point quadrature yields a singular $\mathbf{K}^\gamma$. Therefore, Eq. 17 cannot be inverted, whereas $\mathbf{K}^\mu + \mathbf{K}^\rho + \mathbf{K}^\gamma$ is nonsingular and can be inverted (after assembly and imposition of boundary conditions) to obtain a good finite element solution of the original problem. When a quadratic rectangular element is used, the 3×3 Gauss quadrature is used to evaluate $\mathbf{K}^\mu$ and $\mathbf{K}^\rho$, and the 2×2 Gauss quadrature is used to evaluate $\mathbf{K}^\gamma$.

The choice of the penalty parameter is largely dictated by the ratio of the magnitude of penalty terms to the viscous and convective terms (or compared to the Reynolds number, Re), the mesh, and the word length in the computer. The following range of γ is suggested in computations

$$\gamma = 10^4 Re \text{ to } \gamma = 10^{12} Re.$$

Summary

Numerical simulation of geomechanical processes requires a good understanding of computational fluid mechanics, heat

transfer, and solid mechanics and their couplings. The increase in computing power in both single processor and parallel environments has allowed realistic geomechanics problems of significant complexity and fidelity to be routinely solved and utilized in technological advances. Commercial software has made rapid progress in providing a broad spectrum of analysis capabilities to a variety of industries. Though software is increasingly robust, accurate simulations still require a knowledgeable user, with a background in both mechanics and numerical methods. This entry only provides an introduction to an individual who is interested in the use of the finite element method as a numerical simulation tool for the study and understanding of geomechanical phenomena. The Poisson equation and the Navier–Stokes equations visited here provide the necessary background for the study of diffusion processes and viscous flow problems. Interested readers may consult the references listed.

Cross-References

- [Numerical Methods, Boundary Element](#)
- [Numerical Methods, Finite Difference](#)
- [Numerical Methods, Multigrid](#)

Acknowledgments The author gratefully acknowledges the support provided by the Oscar S. Wyatt Endowed Chair.

Bibliography

- Bathe KJ (1996) Finite element procedures. Prentice Hall, Englewood Cliffs
- Belytschko T, Liu WK, Moran B (2000) Nonlinear finite elements for continua and structures. Wiley, Chichester
- Bird P (1989) New finite element techniques for modeling deformation histories of continents with stratified temperature-dependent rheology. *J Geophys Res* 94(B4):3967–3990
- Dyksterhuis S et al (2005) Finite-element modelling of contemporary and palaeo-intraplate stress using ABAQUS. *Comput Geosci* 31(3):297–307
- Parsons T (2002) Post-1906 stress recovery of the San Andreas fault system calculated from three-dimensional finite element analysis. *J Geophys Res* (B8). <https://doi.org/10.1029/2001JB001051>
- Reddy JN (2015) An introduction to nonlinear finite element analysis, 2nd ed. Oxford University Press, Oxford
- Reddy JN (2019) An introduction to the finite element method, 4th edn. McGraw-Hill, New York
- Reddy JN, Gartling DK (2010) The finite element method in heat transfer and fluid dynamics, 3rd edn. CRC Press, Boca Raton
- Reddy JN, Stein RJ, Wickham JS (1982) Finite-element modeling of folding and faulting. *Int J Numer Anal Methods Geomech* 6:425–440
- Wickham JS, Tapp GS, Reddy JN (1982) Finite-element modeling of fracture density in single layer folds. *Int J Numer Anal Methods Geomech* 6:441–459

Numerical Methods, Multigrid

Wim A. Mulder

Faculty of Civil Engineering and Geosciences, Department of Geoscience and Engineering, Delft University of Technology, Delft, The Netherlands

Shell Global Solutions International BV, Amsterdam, The Netherlands

Definition

A multigrid method is an algorithm for the iterative solution of partial differential equations using a sequence of discretizations on multiple scales.

Introduction

The numerical solution of a partial differential equation (PDE) requires its discretization and a method to solve the resulting large system of algebraic equations. For linear equations, the resulting system is often a sparse matrix, and a direct solution method suffices if the size of the problem is modest. In three space dimensions, the computational cost of a direct solver may be too high. An iterative method that improves the accuracy of an approximate solution step by step can be a good alternative. Multigrid is an example.

The multigrid method has optimal complexity: the amount of work required to solve a problem with N unknowns is $O(N)$, meaning that it scales with the problem size N . It achieves its efficiency by employing several discretization grids for the same problem.

History

Fedorenko (1964) introduced the multigrid method as an iterative scheme for solving Poisson's equation on a square and showed that the number of computations required to determine a solution with a prescribed accuracy is proportional to the number of unknowns, N . Therefore, the method has an optimal computational complexity. Brandt (1973) found that the actual computational cost for a sufficiently accurate result was about 10 work units, where a work unit is the cost of evaluating the discretized equations. He connected the method to local adaptive grid refinement and introduced nonlinear multigrid. Hackbush (1976, 1985) discovered the method independently and provided a mathematical foundation. Since then, the method was developed further to handle PDEs other than Poisson's, which is elliptic, and to go beyond PDEs.

Textbooks include those by Hackbush (1976), who includes convergence proofs; by Wesseling (1992), with chapters on computational fluid dynamics; by Briggs et al. (2000), an easy to read introduction; and by Trottenberg et al. (2001), among others.

Two-Grid Scheme

A discrete representation of a PDE provides an algebraic system of equations with relations between solution values at neighbouring grid points. It is fairly easy to make local corrections that reduce the solution error on a short range but much harder to correct the long-range or long-wavelength components of the solution that have a more global character. By projecting the solution onto a coarser grid, the long wavelengths become shorter and can effectively be solved for. Combining the corrections to the solution from coarser and finer grids yields an efficient solver.

A simple one-dimensional example can help to understand the fundamentals of the multigrid method. The PDE is $Lu = f$, with $u(x)$ the unknown solution as a function of position x on a finite interval $[x_{\min}, x_{\max}]$ on the real axis, $f(x)$ a source term or forcing function, and L a linear differential operator, for instance, minus the Laplacian, which in 1D is $-\frac{d^2}{dx^2}$. Dirichlet boundary conditions let the solution be zero at the endpoints of the interval. To obtain a discrete representation of the problem, an equidistant one-dimensional grid is defined with grid points $x_k = x_{\min} + kh$, index $k = 0, \dots, N+1$, where $N+1 = 2^M$ and M is a positive integer. The grid spacing is $h = (x_{\max} - x_{\min})/(N+1)$. A standard second-order finite-difference scheme leads to

$$\frac{-u_{k-1} + 2u_k - u_{k+1}}{h^2} = f_k, \quad k = 1, \dots, N,$$

where u_k approximates $u(x_k)$ and $f_k = f(x_k)$. At the boundaries, $u_0 = u_{N+1} = 0$. The discrete equations represent the problem $L^h u^h = f^h$, where L^h is a $N \times N$ sparse matrix and u^h and f^h are vectors of length N . The residual is defined as $r^h = f^h - L^h u^h$ and should vanish once the numerical solution has been found.

A simple iterative method is Jacobi relaxation, in which the matrix L^h is replaced by its diagonal $D^h = 2/h^2$. One step of Jacobi relaxation amounts to

$$u^h := u^h + \omega (D^h)^{-1} r^h.$$

The symbol “:=” indicates that the solution values are replaced by the expression on the right-hand side. The factor ω controls the amount of damping. Convergence requires $0 < \omega \leq 1$. Fourier analysis (Hackbush 1985) shows that the convergence rate, the factor by which the norm of the difference between the current and the exact numerical

solution is reduced, is $1 - O(h^2)$ for Jacobi relaxation. Convergence slows down on increasingly finer grids.

The Fourier analysis reveals that the slow convergence is caused by the long-wave components of the solution. Because long waves can be represented on coarser grids, it makes sense to have look at a two-grid scheme, outlined in Fig. 1 and its caption. The coarse grid consists of every other grid point of the fine grid: $x_K = x_0 + KH$, with $H = 2h$ and $K = 0, \dots, \frac{1}{2}(N+1)$. A restriction operator $\tilde{I}_h^H$ maps the current fine-grid residual r^h to the coarser grid: $r^H = \tilde{I}_h^H r^h$. The simplest restriction operator is injection: $r_K^H = r_{2K}^h$, $K = 1, \dots, \frac{1}{2}(N-1)$. Full weighting, which lets $r_K^H = \frac{1}{4}r_{2K-1}^h + \frac{1}{2}r_{2K}^h + \frac{1}{4}r_{2K+1}^h$, applies some smoothing to the residual and is more common. The exact solution of the coarse-grid problem yields the coarse-grid correction $v^H = (L^H)^{-1} r^H$, which should be interpolated back to the fine grid and added as a correction to the latest fine-grid solution. The interpolation or prolongation operator is denoted by I_H^h and lets $u^h := u^h + I_H^h v^H$.

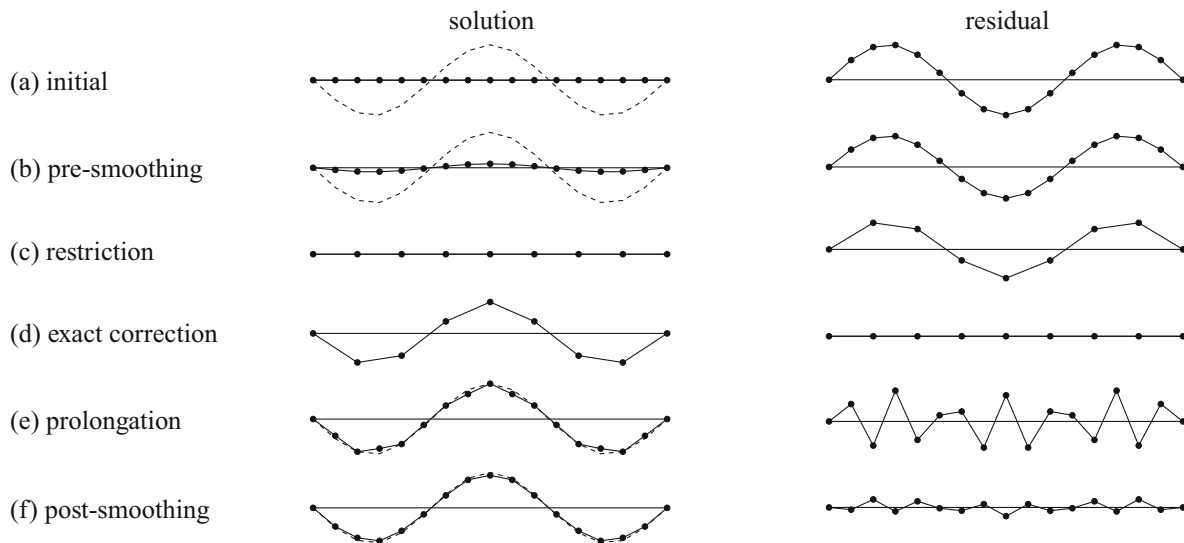
If linear interpolation is used for prolongation, this becomes

$$u_{2K}^h := u_{2K}^h + v_K^H, \quad K = 1, \dots, \frac{1}{2}(N-1),$$

$$u_{2K-1}^h := u_{2K-1}^h + \frac{1}{2}(v_{K-1}^H + v_K^H), \quad K = 1, \dots, \frac{1}{2}(N+1).$$

Here, it is assumed that $v_0^H = v_{(N+1)/2}^H = 0$. After prolongation, an additional relaxation step with damped Jacobi further removes oscillatory error components of the solution. Jacobi relaxation inside this two-grid scheme has a different purpose than when used as an iterative solver by itself. Instead of removing both the short- and long-wave components of the numerical solution error, it only has to deal with those components that cannot be represented on the coarser grid without aliasing. Therefore, a damped version with $\omega < 1$ can be more effective as it can be optimized to remove the short-wave or oscillatory components. For Fig. 1, $\omega = 2/3$ was used. The optimal choice requires a more detailed analysis (Hackbush 1985).

A relaxation scheme geared towards removing oscillatory components is called a smoother. Another popular choice is Gauss-Seidel relaxation. The operator is then approximated by a lower or upper triangular matrix, which is easy to invert. The implementation of the scheme is similar to Jacobi, but the residual is evaluated with the most recent solution available. The result will depend on the order in which the grid is traversed. With lexicographic Gauss-Seidel, one follows the natural index k in increasing order or in the opposite direction. Symmetric Gauss-Seidel performs both these smoothing steps in sequence. An alternative is red-black Gauss-Seidel, where first the points with an odd and then those with an even



Numerical Methods, Multigrid, Fig. 1 Steps in a two-grid cycle. (a) The initial solution on the fine grid is set to zero and the residual equals the forcing function. The dash line represents the exact numerical solution. (b) After one step of pre-smoothing with damped Jacobi relaxation, the solution error is still large, and the residual has hardly changed in this example. Its restriction to the coarser grid (c) is solved exactly, using a coarse-grid version of the discrete differential operator.

The resulting correction to the fine-grid solution (d) is interpolated or prolonged back to the fine grid and added to the fine-grid solution (e). The error in the solution is now dominated by the short wavelengths and appears as an oscillatory function, which is reflected in the corresponding residual. A post-smoothing step (f) removes most of the solution error. Repeating the whole cycle will further reduce the error

index are updated, always using the latest solution values for the residual. In 2D, this would follow a checkerboard pattern.

The grid transfer operator, restriction and prolongation, should obey $m_p + m_r > 2m$ (Hackbush 1985) where $2m$ is the order of the differential equation and $m_p - 1$ is the highest degree of the polynomial that is interpolated exactly. The scaled adjoint of the restriction operator is an interpolation operator for which m_r can be defined in the same way as m_p . In the example above, $2m = 2$, full-weighting has $m_r = 2$, and prolongation based on linear interpolation has $m_p = 2$. More advanced grid transfer operators are based on the differential operator L^h , leading to operator-weighted restriction and prolongation.

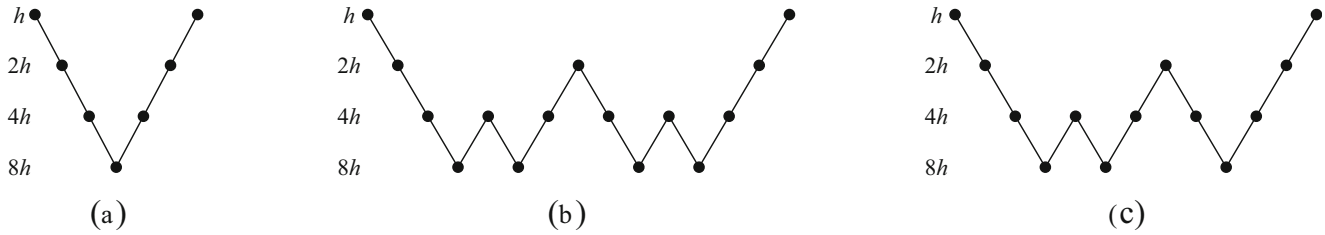
The coarse-grid operator L^H can be based on the same discretization as the fine-grid operator L^h . An alternative is the Galerkin coarse-grid approximation $\tilde{I}_h^H L^h I_H^h$. In the current example with full weighting, these happen to be the same. Operator-weighted grid transfer operators can accelerate the convergence by using the coefficients of the differential operator in the construction of restriction and prolongation operators, which in turn will affect the Galerkin coarse-grid approximation of the differential operator.

Multigrid

Instead of using the exact numerical solution on the coarser of the two grids, a two-grid scheme can be applied to the coarse-grid problem. Extending this recursively to the coarsest grid

with three points and one unknown, a multigrid solver is obtained. All these grids can be visited in different orders. The standard approach is to start at the finest grid, perform a number of pre-smoothing steps, then move to the coarser grid, again perform pre-smoothing, and continue this until the coarsest grid is reached. There, the exact numerical solution is computed, and the resulting coarse-grid correction is prolonged back to the finer level, followed by a number of post-smoothing steps. This is repeated up to the finest grid. Such a sequence is called a V-cycle. A single V-cycle may be insufficient to obtain a convergence rate on the coarser grid that is similar to the smoothing rate on the finer grid. Convergence on coarser grids can be improved by performing more than one cycle. In the W-cycle, the number of cycles doubles on each coarser grid. Figure 2 illustrates the order in which grids are visited for a V-, W-, and F-cycle. The last one is less expensive than a W-cycle, as the number of cycles increases by only one for increasingly coarser grids. Another option, adaptive cycling, decides on the sequence of grids by monitoring the decrease of the residuals.

With a proper choice of relaxation scheme or smoother grid transfer operators, the multigrid method can reach a grid-independent or h -independent convergence rate. This does not, however, mean that the amount of work required to reach a sufficiently accurate solution is proportional to the number of grid points N . The exact numerical solution differs from the true solution of the PDE because of the numerical error due to the discretization, typically with some power p of h . In the earlier example, the discretization is second-order



Numerical Methods, Multigrid, Fig. 2 (a) V-cycle, (b) W-cycle, and (c) F-cycle

accurate, so $p = 2$. Given this numerical error, convergence to machine precision is more than needed. An iteration error somewhat smaller than the discretization error should suffice. Therefore, more iterations should be carried out if a more accurate solution on a finer grid has to be computed. This leads to an overall complexity of $O(N \log N)$ rather than $O(N)$. Successive grid refinement, in which one first computes a solution on a coarse grid and uses that as an initial guess for the next finer grid and so on, enables removal of the factor $\log N$, assuming that a fixed number of multigrid cycles are used at each level. The combination of successive grid refinement with a multigrid solver is called full multigrid (FMG).

Nonlinear Multigrid

One way to apply the multigrid method to nonlinear PDEs is the use of Newton's method. A multigrid solver is then applied to a linearization of the discretized nonlinear problem. Another approach is the use of the discretized nonlinear PDE inside the multigrid algorithm. This requires the full solution to be available on the coarser grids. In the example above, only corrections to the solution were represented on coarser grids. The full approximation scheme (FAS) is a reformulation of the multigrid method that includes the full solution (Brandt 1982). It requires a restriction of the full solution, $u^H = I_h^H u^h$, and of the coarse-grid forcing function, $f^H = L^H(I_h^H u^h) + \tilde{I}_h^H r^h$. The prolongation needs to be changed to $u^h := u^H + I_h^H(u^H - I_h^H u^h)$, where u^H is the sufficiently converged solution on the coarser grid and $I_h^H u^h$ the restriction of the latest fine-grid solution. The restriction operators for the residuals, $\tilde{I}_h^H$, and for the solution, I_h^H , do not have to be the same. Note that the coarse-grid problem, $L^H u^H = f^H$, can be interpreted in a different way: $L^H u^H = f^H = \tilde{I}_h^H f^h + \tau_{hH}^H$. The fine-to-coarse defect correction, $\tau_{hH}^H = L^H(I_h^H u^h) - \tilde{I}_h^H(L^h u^h)$, ensures that the coarse-grid equations maintain the accuracy of the fine-grid equations.

Nonlinear equations of the form $L^h(u^h) = f^h$ fit into this scheme. For $L^H(u^H)$, the same discretization scheme as on the fine grid can be adopted. Smoothing operators can be based on a local linearization of the residuals.

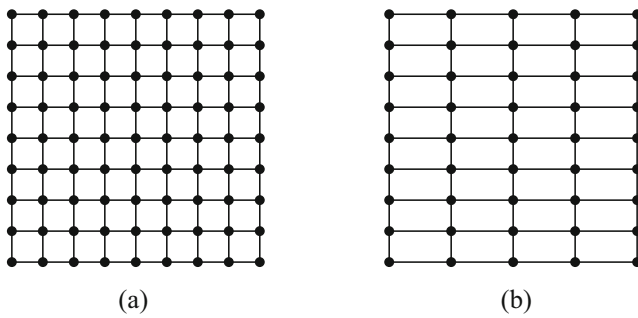
Generalizations

Multigrid is highly efficient for the solution of Poisson's equation on the square, discretized by the finite-difference method on a Cartesian grid. For other PDEs, describing convection, diffusion, waves, and flow, the method may be less easy to apply. The same is true for unstructured grids, for instance, based on triangles or tetrahedra. With suitable modifications and at the expense of additional code complexity, the multigrid method can still be the optimal approach. With unstructured grids, the coarser and finer grid are generally not nested, leading to grid transfer operations that are less easy to code than on a Cartesian grid.

Anisotropy in the PDE can degrade performance. Consider the equation:

$$-\frac{\partial^2 u}{\partial x^2} - a \frac{\partial^2 u}{\partial y^2} = f(x, y).$$

For $0 < a \ll 1$, smoothing operators may have difficulty with the y -direction where there is only weak coupling between neighbouring unknowns. Effective anisotropy in the discretized equation can also happen with grids that have widely different spacings in the various coordinate directions. A more powerful relaxation scheme may repair the poor performance of a simple smoother. For 2D problems, line relaxation will smooth in one direction and solve in the other. An example is line Jacobi relaxation, which approximates the matrix L^h by dropping the off-diagonals in one coordinate direction, while keeping them in the other. The dropped off-diagonals can be subtracted from the main diagonal with a suitable weight factor to obtain better smoothing properties. For 3D problems, plane relaxation is required, which may be too expensive. An alternative is semi-coarsening, where the coarse grid has the same number of points in the direction of weak coupling and is coarsened only in the direction in which the smoothing effectively removes the shorter wavelengths. Figure 3 shows an example of a grid before and after semi-coarsening in the horizontal direction. In the more general case where semi-coarsening is required in all coordinates, it can be applied in alternating directions.



Numerical Methods, Multigrid, Fig. 3 Semi-coarsening of grid (a) in the horizontal direction provides grid (b)

A more costly but powerful approach is simultaneous semi-coarsening in all coordinates (Mulder 1989), which still retains optimal complexity.

Beyond Partial Differential Equations

The application of the multigrid method is not restricted to the solution of PDEs. There are generalizations to eigenvalue problems, integral equations, optimization and optimal control, statistical physics, image processing, image segmentation, and edge detection. In all cases, the simultaneous use of different scales overcomes the problem of slowly converging solution components.

Algebraic multigrid (AMG), or, better, the algebraic multi-level method, refers to a class of solvers that construct sets of coarse-level equations without referring to grids or geometrical properties (Trottenberg et al. 2001; Shapira 2008). Coarsening is applied to subsets of unknowns for which smoothing is effective. If the problem is specified by a large matrix, these subsets can be determined by examining the coefficients of the matrix. Strongly coupled unknowns, say u_i and u_j , are related by a matrix coefficient a_{ij} that is relatively large. In that case, smoothing is effective, and one of the unknowns can be removed from the equations on the coarser level. The coarser level will then be mainly populated by weakly coupled variables. Operator-weighted grid transfer operators are natural in this context, as well as the Galerkin approach for the construction of the coarse-level operator. Algebraic multilevel methods are a popular choice for finite-element discretizations on unstructured grids, as they can be used as a black-box solver and do not require the tedious coding of grid transfer operations between non-nested elements.

The multigrid method bears some resemblance to other techniques such as wavelets, hierarchical-basis finite elements, multi-scale techniques, cyclic reduction, fast multipole methods, and other divide-and-conquer methods.

Geophysical Applications

Multigrid has been applied successfully to a wide range of partial differential equations, describing static and dynamic problems, diffusion, convection, and flow problems. Examples in geophysics include elliptic problems in potential theory, such as gravity (Kusche 2002), magnetostatics (De Giersem and Hameyer 2001), electrostatics (Ascher and Haber 2003), and Darcy's law in porous media flow (Mulder and Gmelig Meyling 1993). Two-dimensional phase unwrapping for satellite interferometric synthetic aperture radar (Pritt 1996) can also be stated as an elliptic problem. Controlled-source electromagnetics (Werthmüller et al. 2019), magnetotellurics (Li et al. 2015), mantle convection, and geodynamics (Gerya 2019) are examples of parabolic problems.

The wave equation for seismic applications is a hyperbolic problem. Iterative solution of its frequency-domain formulation, the Helmholtz equation, was a notoriously difficult numerical problem but can nowadays be accomplished by using a damped version as preconditioner in a conjugate-gradient-type iterative scheme. Multigrid efficiently deals with the approximate inversion of the preconditioner (Riyanti et al. 2006). In the generalization of the method to the elastic wave equation, simple Point-Jacobi smoothing turns out to be insufficient for the secondary or shear waves, and a more powerful smoother is required (Rizzuti and Mulder 2016).

Summary

The multigrid method provides an optimal iterative solution method for a wide class of problems governed by partial differential equations. Its application to a finite-difference discretization of Poisson's equation on a square is the easiest. For other problems, a bit more effort may be required.

Cross-References

- Numerical Methods, Finite Difference
- Numerical Methods, Finite Element

Bibliography

- Ascher UM, Haber E (2003) A multigrid method for distributed parameter estimation problems. *Electron Trans Numer Anal* 15:1–17
- Brandt A (1973) Multi-level adaptive technique (MLAT) for fast numerical solution to boundary value problems. In: Cabannes H, Temam R (eds) *Proceedings of the third international conference on numerical methods in fluid mechanics*. Springer Berlin Heidelberg, Berlin/Heidelberg, pp 82–89. <https://doi.org/10.1007/BFb0118663>

- Brandt A (1982) Guide to multigrid development. In: Hackbusch W, Trottenberg U (eds) *Multigrid methods*. Springer Berlin Heidelberg, Berlin/Heidelberg, pp 220–312. <https://doi.org/10.1007/BFb0069930>
- Briggs WL, Henson VE, McCormick SF (2000) *A multigrid tutorial*, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia. <https://doi.org/10.1137/1.9780898719505>
- De Gersem H, Hameyer K (2001) Full multigrid for magnetostatics using unstructured and non-nested meshes. *IEEE Trans Magn* 37(5):3460–3464. <https://doi.org/10.1109/20.952637>
- Fedorenko RP (1964) The speed of convergence of one iterative process. *USSR Comput Math Math Phys* 4(3):227–235. [https://doi.org/10.1016/0041-5553\(64\)90253-8](https://doi.org/10.1016/0041-5553(64)90253-8)
- Gerya TV (2019) *Introduction to numerical geodynamic modelling*, 2nd edn. Cambridge University Press, Cambridge. <https://doi.org/10.1017/9781316534243>
- Hackbusch W (1976) Ein iteratives Verfahren zur schnellen Auflösung elliptischer Randwertprobleme. Report 76-12. Mathematisches Institut, Universität zu Köln, Köln
- Hackbusch W (1985) *Multi-grid methods and applications*. Springer series in computational mathematics, vol 4. Springer, Berlin/Heidelberg. <https://doi.org/10.1007/978-3-662-02427-0>
- Kusche J (2002) On fast multigrid iteration techniques for the solution of normal equations in satellite gravity recovery. *J Geodyn* 33(1): 173–186. [https://doi.org/10.1016/S0264-3707\(01\)00062-X](https://doi.org/10.1016/S0264-3707(01)00062-X)
- Li G, Hao T, Zhang L (2015) Performance of preconditioned iterative and multigrid solvers in solving the three-dimensional magnetotelluric modeling problem using the staggered finite-difference method: a comparative study. *J Geophys Eng* 13(1):1–10, 12. <https://doi.org/10.1088/1742-2132/13/1/1>
- Mulder WA (1989) A new multigrid approach to convection problems. *J Comput Phys* 83(2):303–323. [https://doi.org/10.1016/0021-9991\(89\)90121-6](https://doi.org/10.1016/0021-9991(89)90121-6)
- Mulder WA, Gmelig Meyling RHJ (1993) Numerical simulation of two-phase flow using locally refined grids in three space dimensions. *SPE Adv Technol Ser* 1:36. <https://doi.org/10.2118/21230-PA>
- Pritt MD (1996) Phase unwrapping by means of multigrid techniques for interferometric SAR. *IEEE Trans Geosci Remote Sens* 34(3): 728–738. <https://doi.org/10.1109/36.499752>
- Riyanti CD, Erlangga YA, Plessix R-É, Mulder WA, Vuik C, Oosterlee C (2006) A new iterative solver for the time-harmonic wave equation. *Geophysics* 71(5):E57–E63. <https://doi.org/10.1190/1.2231109>
- Rizzuti G, Mulder WA (2016) Multigrid-based ‘shifted-Laplacian’ preconditioning for the time-harmonic elastic wave equation. *J Comput Phys* 317:47–65. <https://doi.org/10.1016/j.jcp.2016.04.049>
- Shapira Y (2008) *Matrix-based multigrid: theory and applications*. Numerical methods and algorithms, vol 2, 2nd edn. Springer, Boston. <https://doi.org/10.1007/978-0-387-49765-5>
- Trottenberg U, Oosterlee CW, Schüller A (2001) *Multigrid*. Academic, London. ISBN 9780127010700
- Werthmüller D, Mulder WA, Slob EC (2019) emg3d: a multigrid solver for 3D electromagnetic diffusion. *J Open Source Softw* 4(37):1463. <https://doi.org/10.21105/joss.01463>
- Wesseling P (1992) *An introduction to multigrid methods*. Wiley, Chichester

Ocean Bottom Seismics

Ingo A. Pecher^{1,2}, Jörg Bialas³ and Ernst R. Flueh³

¹GNS Science, Lower Hutt, New Zealand

²Institute of Petroleum Engineering, Heriot-Watt University, Edinburgh, UK

³Leibniz-Institut für Meereswissenschaften, University of Kiel, Kiel, Germany

Synonyms

Ocean bottom seismology

Introduction

Ocean bottom seismics, recording of seismic waves on the ocean floor by seismometers placed on the ocean floor, is becoming increasingly common both in academic research surveys and for hydrocarbon exploration. The main advantages of ocean bottom seismics over reflection seismics with hydrophone streamers (see ► [“Seismic Data Acquisition and Processing”](#)) are that the use of receivers on the seafloor yields sufficient source-receiver offsets to analyze the deeper crustal velocity structure, deploying multicomponent receivers allows recording of shear waves, and that instruments can be deployed for an extended time to record natural seismicity. In most cases, the seismic sources are either airguns towed on the sea surface or natural seismicity. Seafloor sources are rare.

Instrumentation

Seismic seafloor receivers fall broadly into three categories, (1) ocean bottom seismometers (OBSs), individual receivers that can be deployed and recovered without the need of any

specialized vessels, (2) ocean bottom cables (OBCs), streamer-like arrays of receivers that are deployed from a specialized vessel, and (3) ocean bottom seismic nodes, individual receivers that require underwater vehicles for deployment and retrieval.

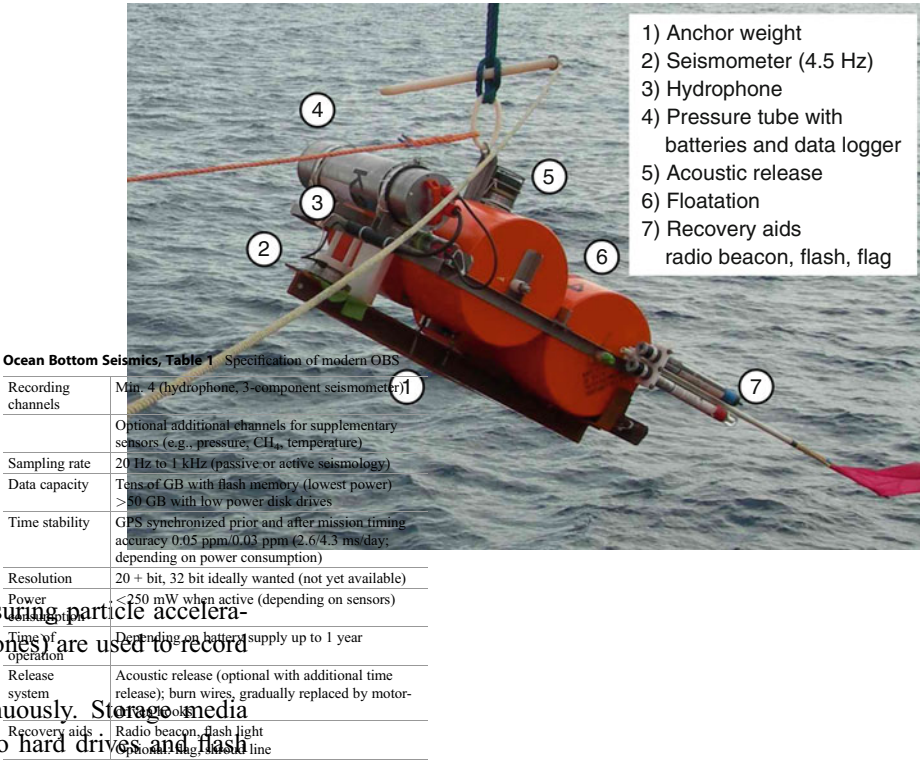
Ocean Bottom Seismometers

OBSs are individual stations that contain sensors and electronics for data storage (Fig. 1). These instruments are deployed from a vessel, and descend to the seafloor weighted by an anchor. A release mechanism is activated toward the end of the survey, detaching the instrument from the anchor. The instrument then rises to the surface buoyed by flotation devices. The instrument is recovered onto the vessel where the recorded data are downloaded. The term ocean bottom seismometer is sometimes used for any instrument that records seismic waves at the seafloor, including OBCs and ocean bottom seismic nodes.

Communication with OBSs while on the seafloor is usually performed with acoustic transmission. This includes releasing the instrument from the anchor; although sometimes, a time-based release is used as backup in case communication fails. Acoustic communication is also used for determining the position of the instrument at the seafloor by triangulation and for checking the status of the instrument.

OBSs that contain only a hydrophone as receiver are often called ocean bottom hydrophones. Hydrophones record pressure waves in the water. Some instruments contain vertically oriented geophones, mostly in combination with hydrophones. But, more commonly used OBSs have 3-component geophones (one vertical, two horizontal components) or 4-component (4-C) OBSs that contain 3-component geophones and a hydrophone. The type of geophone varies, depending on application – usually 4.5 Hz geophones (geophones with a resonance frequency at 4.5 Hz) for active-source experiments. Broadband seismometers with varying frequency bands (with periods as low as several tens of seconds) are used for recording of natural seismicity.

Ocean Bottom Seismics,
Fig. 1 Ocean bottom
seismometer (instruments used by
IfM-Geomar, Kiel, Germany)



Ocean Bottom Seismics, Table 1 Specification of modern OBS

Recording channels	Min. 4 (hydrophone, 3-component seismometer) Optional additional channels for supplementary sensors (e.g., pressure, CH ₄ , temperature)
Sampling rate	20 Hz to 1 kHz (passive or active seismology)
Data capacity	Tens of GB with flash memory (lowest power) > 50 GB with low power disk drives
Time stability	GPS synchronized prior and after mission timing accuracy 0.05 ppm/0.03 ppm (2.6/4.3 ms/day; depending on power consumption)
Resolution	20 + bit, 32 bit ideally wanted (not yet available)
Power consumption	< 250 mW when active (depending on sensors)
Time of operation	Depending on battery supply up to 1 year
Release system	Acoustic release (optional with additional time release); burn wires, gradually replaced by motor-
Recovery aids	Radio beacon, flash light (optional flag, surface line)

In a few cases, accelerometers (measuring particle acceleration rather than velocity as in geophones) are used to record seismic waves.

Seismic data are recorded continuously. Storage media have evolved from analogue tapes to hard drives and flash drives. Exact time stamps from an internal clock are critical for OBS records. GPS time synchronization is completed before and after deployment. Modern high precision clocks keep the deviation as low as 0.05 ppm (4.3 ms/day). For active-source experiments, the records are split into receiver gathers using shot times, usually in SEG-Y format. Data processing varies widely depending on the application and resembles that of land data. Some peculiarities of OBS records include unknown receiver locations, unknown orientations of the horizontal components, and a lack of continuous GPS time stamps. Typical specifications of a modern OBS are listed in Table 1.

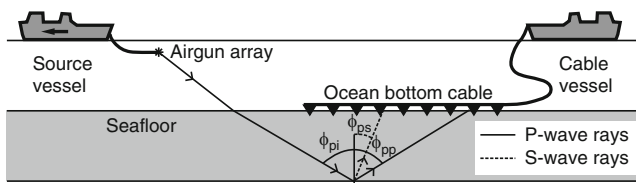
Until the 1990s active-source OBS surveys typically involved 10–50 OBSs deployed along a seismic transect. Spacing between instruments for a survey targeting the deeper crustal structure typically was between 5 and 20 km, allowing travelt ime analysis of wide-angle reflected and refracted arrivals. In the past decade, significant improvements in hardware and electronics have made instruments less expensive and shortened turnaround times. This has led to a number of experiments with considerably more OBSs allowing imaging of deep crustal structures and 3D travelt ime analysis (e.g., Kodaira et al. 2000; Morgan et al. 2002).

Advances in batteries and storage devices also make it possible to continuously record seismicity with OBSs for up to ~1 year. The use of additional batteries may extend this time window but even for frames and housing made of titanium, corrosion is a problem. Therefore a service recovery after ~1 year is good practice. Such passive- source experiments are becoming more common to study the earth’s deep

structure, subduction zones on regional scales, as well as seismically active local features such as mud volcanoes (Brueckmann et al. 2009). A combination of onshore and offshore seismological networks has shown to significantly improve the determination of near-shore hypocenters (Husen et al. 1999; Arroyo et al. 2009).

Ocean Bottom Cables

OBCs are densely spaced receivers attached to a cable. The cable is deployed from a vessel, usually a specialized ship (Fig. 2). Seismic data are transmitted along the cable to be recorded on the vessel. Typical receiver spacing is 25 m. Most modern cables are equipped with 4-C receivers. Cables are often dragged along the seafloor to achieve deployment in a



Ocean Bottom Seismics, Fig. 2 OBC acquisition. Part of the downgoing P-wave energy is converted upon reflection to upgoing S-wave energy. ϕ_{pi} : angle of incidence, downgoing P-wave. ϕ_{pp} : exit angle, reflected P-wave. ϕ_{ps} : exit angle, P-to-S converted wave. Ray angles according to Snell's Law, that is, $\phi_{pi} = \phi_{pp}$ and $\sin(\phi_{ps})/V_s = \sin(\phi_{pi})/V_p$

straight line. Some OBCs use geophones that are being pushed into the seafloor by remotely operated vehicles (ROVs) to improve coupling to the seafloor. OBCs may also be deployed in trenches to improve coupling and reduce noise, such as for the Life of Field experiment for longterm monitoring of the Valhall oil field, North Sea (Barkved and Kristiansen 2005).

OBCs allow imaging of the subsurface with similar coverage as surface-towed streamers. Furthermore, coupling between receivers and seafloor is generally far better than for OBSs partly because the mass of the recording equipment is much smaller, and partly because the deployment techniques for OBCs facilitate slight “burrowing” of cables into the seafloor. OBCs are widely used in the oil and gas exploration industry. The need for a specialized cable deployment vessel, however, normally makes OBCs prohibitively expensive for academic research to date.

Ocean Bottom Seismic Nodes

Ocean bottom seismic nodes are individual receivers with independent recording units – essentially, OBSs without anchors and flotation devices. They are typically deployed with ROVs making their deployment relatively expensive compared to OBSs. Nodes are often deployed to fill coverage gaps in OBC surveys, for example, in the immediate vicinity of platforms where it is unsafe to deploy OBCs (or OBSs). Nodes are often pushed into the seafloor, which potentially leads to improved receiver coupling (Caldwell 1999).

Ocean Bottom Seismic Sources

By far the majority of active-source OBS surveys use pressure waves generated by airguns towed behind a vessel, although some ocean bottom sources have been pioneered, mainly to generate shear waves.

Explosives have been used as sources at the seafloor for an experiment to record shear waves in boreholes during Ocean Drilling Program Leg 156 but because of the strong increase of shear (S-) wave velocity with depth beneath the seafloor, rays were bent away from the borehole and did not reach borehole receivers (Peacock et al. 1997).

Furthermore, environmental and health and safety concerns preclude large-scale use of seafloor explosives.

Imploding glass spheres are the source of the SEEBOSEIS system, for which glass spheres containing air at atmospheric pressure are punctured on the seafloor (Nguyen et al. 2009).

Shear-wave “sleds” are occasionally used to study the high-resolution shear-wave structure of the first few meters beneath the seafloor. One such system uses an electromagnetic coil to generate seismic shear energy at the seafloor. Receivers are dragged along the seafloor behind the source. This system has been used successfully, for example, to measure near-seafloor S-wave velocity offshore California (Huws et al. 2000).

Applications

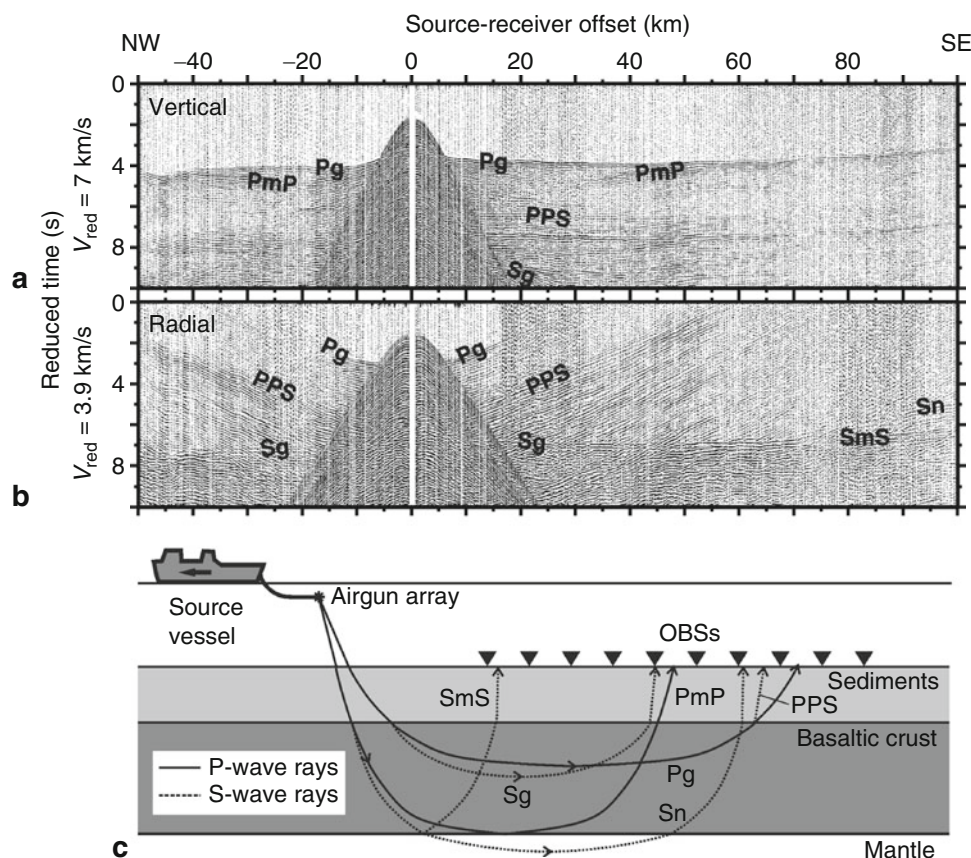
P-Waves

OBSs have been used extensively in the last 4 decades to study the crustal structure beneath the seafloor (see ► “Deep Seismic Reflection and Refraction Profiling”) from P-wave traveltimes analysis. A typical survey, for example, for a transect across a subduction zone consists of 20+ instruments spaced between 5 and 20 km. Seismic shot lines with maximum offsets of often >100 km are acquired with a large ($\sim 4,000$ – $8,000$ cu-in = 65 – 130 l) airgun array. Often, a coincident seismic reflection profile is acquired with the shot vessel although, in order to avoid noise from previous shots to interfere with arrivals from deep targets, shot intervals for crustal OBS surveys should be much larger (~ 60 s, equivalent to 100 – 150 m shot spacing) than would be desired for reflection surveying (~ 15 s) (Christeson et al. 1996) (see ► “Single and Multichannel Seismics”).

Such a configuration typically allows recording of reflected and refracted arrivals down to the base of the crust as well as refracted waves through the upper mantle (Fig. 3). Records are similar to those from land surveys, albeit often at considerably higher signal-to-noise ratio, and receivers and shots are reversed; land crustal wide-angle surveys typically have sparse shot and dense receiver spacing, whereas marine OBS surveys have dense shot but sparse receiver spacing.

Data processing often only involves bandpass filtering and amplitude scaling. Typical sources of noise or data contamination include reverberations particularly on horizontal components, previous-shot noise, water-column multiples, and noise from underwater currents that may shake the instruments.

As for land surveys, data are often displayed using a reduced traveltimes ($t_{red} = t - x/v_{red}$, where t_{red} : reduced time, x : source-receiver offset, v_{red} : reduction velocity). The origin of arrivals is identified and their traveltimes are picked, followed by traveltimes analysis to determine seismic velocities (see ► “Traveltimes Tomography Using Controlled-Source Seismic Data”). The latter is performed based on ray-trace modeling (e.g., Zelt and Smith 1992). In the past



Ocean Bottom Seismics, Fig. 3 Example of an OBS record from offshore the Faeroe Islands with interpreted arrivals (after Eccles et al. 2009). (a) vertical component, (b) rotated radial component, (c) schematic ray paths. See text in chapter on converted waves for details on radial component. Reduction velocities 7 km/s for vertical component, 3.9 km/s for radial component. Arrivals P-waves: Pg: diving wave in the crust, PmP: reflection from the top of the mantle (Mohorovičić discontinuity). Converted S-waves: PPS: S-wave, converted upon transmission

from upgoing P-wave at basalt-sediment interface. Sg: S-wave, converted upon transmission from downgoing P-wave at basalt-sediment interface and diving as S-wave in the basaltic crust. SmS: S-wave, converted upon transmission from downgoing P-wave at basalt-sediment interface and reflected as S-wave at the top of the mantle. Sn: S-wave, converted upon transmission from downgoing P-wave at basalt-sediment interface and diving as S-wave in the mantle. Courtesy Jennifer Eccles (University of Auckland, New Zealand)

decade, traveltimes inversion has become common allowing better determination of the uncertainties and uniqueness of resulting velocity models (e.g., Korenaga et al. 2000). However, the uncertainty from potentially misinterpreting the origin of arrivals is difficult to quantify.

Coincident OBS and seismic reflection surveys are being used for high-resolution studies of the first few hundreds of meters beneath the seafloor, in particular to obtain velocity information in the gas-hydrate zone. Data analysis usually comprises joint analysis of reflected and refracted P-waves often with additional structural information from seismic reflection data (Hobro et al. 1998). Many of these experiments are designed to achieve 3D ray coverage, sometimes coincident with 3D seismic reflection data (Plaza-Faverola et al. 2010). OBS experiments designed for 3D traveltimes analysis have also become more common for surveys targeting the deeper crust (e.g., Morgan et al. 2002).

The increased availability of OBSs has also led to a number of surveys using OBSs with sufficiently dense spacing to allow imaging of deep crustal structures similar to seismic

reflection data (e.g., Kodaira et al. 2000). Advancements in wide-angle migration techniques are key toward successful imaging with OBSs (e.g., Talwani and Zelt 1998). Imaging using OBSs is also improved by increased reflection coverage by utilizing rays from seafloor multiples, that is, rays that are reflected at the seafloor and sea surface, often referred to as mirror imaging (e.g., Grion et al. 2007).

High-resolution (sub-seismic-wavelength) P-wave velocities have been extracted from OBS data using 1D full-waveform inversion techniques similar to those applied to multichannel reflection data (Korenaga et al. 1997). Increases in computing power now also allow 2D visco-acoustic full-waveform tomography (Sourbier et al. 2009a, 2009b) (see ▶ “Seismic, Waveform Modeling and Tomography”).

P-to-S Converted Waves

Identification of P-to-S converted (PS-) waves in horizontal-component records has led to a paradigm shift in ocean bottom seismics in the late 1990s and early 2000s by the use of OBCs for multicomponent seismic exploration (Fig. 2) (see

► “Energy Partitioning of Seismic Waves” and ► “Seismic Data Acquisition and Processing”). P-waves are generated on a shooting vessel. They are partly converted at layer interfaces to S-waves. Upgoing S-waves are then recorded with OBCs or sometimes, densely spaced OBSs. The most prominent PS-arrivals are usually arrivals that are converted from downgoing P-waves upon reflection to upgoing S-waves.

Processing sequences for PS-waves require additional steps compared to conventional P-wave seismic reflection data acquired with surface-towed streamers (Stewart et al. 2002). In particular, processing of PS-waves requires separation of P- and S-waves, a rotation of the horizontal components such that a maximum amount of energy is projected into a single component, the radial component, and various modifications to adjust for the asymmetry of the PS-raypaths. A key step of data analysis is the correlation of P- and PS-events. Once P- and PS-events from the same horizons have been identified, determination of the ratio of P-wave velocity (V_p) over S-wave velocity (V_s) is straight-forward (e.g., Granli et al. 1999).

Multicomponent PS-wave surveys are being conducted routinely by the exploration industry, chiefly with the following objectives:

Imaging through gas clouds	P-waves are attenuated significantly in gas-bearing layers, whereas S-waves are almost unaffected by gas. Converted-wave surveys often yield superior images of structures in and beneath gas-bearing sediments even though in many cases the downgoing P-wave still has to penetrate gas (Granli et al. 1999).
Fluid discrimination	Reflection strength of PS-waves is only marginally affected by the pore fill (through its effect on density). Therefore, flat spots, oil-water or gas-water contacts at the base of hydrocarbon reservoirs, which generate P-wave reflections, usually do not cause any noticeable PS-reflections (e.g., MacLeod et al. 1999). For the same reason, PS-waves may help establish whether high negative-polarity P-wave reflections (bright spots) are likely to be caused by gas in sediment pores.
Mapping of sand–shale interfaces	Sand and shale layers often have similar V_p and density, which leads to low P-wave reflectivity. However, because V_s in sands is usually considerably higher than in shales, sand–shale interfaces often cause strong P-to-S conversion and have been imaged successfully with PS-waves (e.g., MacLeod et al. 1999).
Lithology from V_p/V_s	$V_p - V_p/V_s$ crossplots may be employed to delineate between sands and shales and to

Fracture detection from shear-wave splitting

constrain the shale content of reservoir sands (Margrave et al. 1998). V_p/V_s may also allow to distinguish between different types of carbonates (Rafavich et al. 1984). Nearvertical fractures common, for example, in carbonate reservoirs, causes an azimuthal dependence of velocities, azimuthal anisotropy (see ► “Magnetic Anisotropy”). The propagation of S-waves that are polarized perpendicularly to fractures is slower than that of S-waves polarized parallelly to fractures. This leads to PS-waves being split into fast and slow waves, depending on the azimuth of the incoming P-wave, and allows determination of azimuth and magnitude of anisotropy, from which fracture orientation can be inferred (Gaiser et al. 2002).

In recent years, P-to-S converted waves have been utilized increasingly to study V_s in the Earth’s crust (e.g., Eccles et al. 2009). Unlike for hydrocarbon exploration, with the aim of imaging the sedimentary section with S-waves converted from P-waves upon reflection, studies of the deeper crust often analyze traveltimes of S-waves converted from downgoing P-waves upon transmission. The sediment-basalt interface on the North Atlantic margin between the UK and Iceland, for example, provided a conversion horizon to allow determination of V_s in the deeper crust (Fig. 3). $V_p - V_p/V_s$ crossplots allowed lithological conclusions indicating mixing between continental crust and mafic intrusions across the ocean-continent transition and the presence of sub-basalt sediments on the Faeroe Margin (Eccles et al. 2009).

Wide-Aperture (Wide-Azimuth) Studies

Wide-aperture studies are increasingly conducted in the hydrocarbon exploration for “undershooting” in order to illuminate the subsurface beneath features that are difficult to penetrate with seismic energy and/or exhibit prominent localized velocity anomalies, such as gas clouds or salt domes: Rays in wide-aperture surveys may bypass these features and reach deeper targets of interest. The length of seismic streamers towed behind the same vessel that tows the airgun sources may not be sufficient for wide-aperture studies. Seismic arrivals at wide apertures may be recorded with surface-towed streamers by using separate vessels for sources and receivers. Alternatively, receivers deployed at the seafloor may be used. An acquisition campaign partly aimed at obtaining wide-aperture data across the Atlantis Field in the deep-water Gulf of Mexico involved 900 ocean bottom seismic nodes (Barley and Summers 2007). 3D surveys with ocean bottom receivers also provide illumination of seismic targets from different azimuths (multi-azimuth surveys),

which allows analysis for azimuthal anisotropy and may improve image quality. Wide-aperture seismic in the hydrocarbon exploration industry is often acquired as part of “seismic-on-demand” 4D surveys over permanently installed seismic surveillance systems such as at the Valhall Field (Barley and Summers 2007 and references therein).

Passive-Source Seismology

The use of modern storage media has led to a decrease in power consumption and an increase of storage capacity allowing deployments of OBSs for several months to over a year to record natural seismicity (Fig. 4). Because of the lower frequencies of natural seismic sources usually, broadband geophones are used with a lower resonance frequency than the 4.5 Hz geophones typically used for active-source surveys. Both local seismicity and teleseismic events are being evaluated similarly to data from land-seismic records. Unlike for land data, however, receiver location and geophone orientation need to be established. Local noise conditions in the deep oceans vary significantly throughout the seasons, but in general, noise is higher than on well-placed land stations. In addition, the seismic waveforms are contaminated by water-column multiples.

A very different application of passive-source ocean bottom seismology is to measure seafloor compliance. Pressure variations from ocean waves (gravity or infragravity waves) are being transmitted into the seafloor. Seafloor compliance is the transfer function between pressure loading and seafloor deformation. This transfer function is determined as a function of frequency and depends largely on the subsurface shear modulus (Crawford et al. 1991). Pressure variations are typically recorded with differential pressure gauges, seafloor

deformation derived from acceleration measured with self-leveling gravimeters. The frequency range of the analyzed pressure variations depends largely on water depth – because of the dispersion of ocean-surface waves, short-period waves are not transmitted into deeper waters. The resolution of subsurface shear-modulus analyses from compliance measurements decreases with decreasing frequencies. As an example, the frequency range for studying gas-hydratebearing sediments at 1.3 km water depth and 2 km beneath the seafloor was 0.003–0.049 Hz (Willoughby et al. 2008).

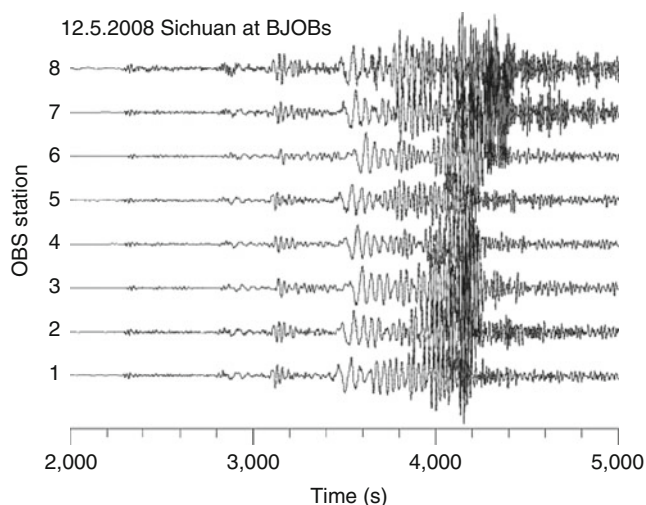
Future Developments

One of the most significant challenges for autonomous OBSs is still the power supply. Both data loggers and storage media are becoming increasingly energy efficient. Together with new battery technology and further improvement of fuel cells, this is leading to reduction of weight and size of instrumentation. Some study areas may provide additional power sources, such as thermal energy. Weight and size reduction in modern instruments is achieved in most cases by installing all the components in a single glass sphere. On the other hand, glass spheres require increased service efforts and precautions compared to more robust, but larger and heavier titanium pressure casing.

The use of ocean bottom sensors in permanent ocean-floor observatories is likely to become more common. OBSs have already been installed for long-term monitoring at seafloor observatories (e.g., GEOSTAR, Beranzoli et al. 2003) and have been integrated in deployments for tsunami warning systems. These observatories generally have the capability of data transmission to shore. Regional studies usually require a deployment of an array of sensors distributed over an area of several square kilometers. Several seismic seafloor observatories have been established to study the seismogenic zone offshore of Japan and a network of 20 instruments is being developed as part of the Japanese Dense Oceanfloor Network System for Earthquakes and Tsunamis (Kaneda et al. 2009).

Combined seismic and electromagnetic field experiments are becoming more common as the combination of both methods is proving powerful for studying subsurface lithology (Constable and Srnka 2007 and references therein). Despite significant technical challenges, especially with contamination of electromagnetic signals from the seismic instrumentation, it is conceivable that future ocean bottom instruments will include sensors for both seismic and electromagnetic waves.

The use of OBCs and nodes for hydrocarbon exploration and field appraisal appears to have become a “mainstream” technique. While costs for deployments of either type of instruments are still high compared to conventional 3D streamer seismic, they need to be weighed against the risk of mispositioning considerably more expensive appraisal



Ocean Bottom Seismics, Fig. 4 Teleseismic event on array of broadband OBSs at ~74°N 17°E (between Norway and Spitsbergen), the Mw 7.9 Sichuan Earthquake (China), May 12, 2008. Courtesy Frank Krüger (University of Potsdam, Germany)

wells (Barley and Summers 2007). Furthermore, several hydrocarbon fields are now being monitored with permanent installation of ocean bottom sensors for repeat-seismic surveys in order to guide field development (Barley and Summers 2007 and references therein).

Conclusions

The last decade has seen a substantial increase of ocean bottom seismic studies. While seafloor seismic sources are rarely deployed, the use of seafloor seismic sensors has established itself primarily for recording offshore wide-angle data, P-to-S converted waves, and natural seismicity. Academic research surveys, for studying the deeper crustal structure, are usually conducted with autonomous OBSs, which are relatively inexpensive and versatile. Surveys by the exploration industry for evaluating and developing hydrocarbon reservoirs mostly involve deployment of ocean bottom cables or ocean bottom seismic nodes, which give superior data quality compared to OBSs but at much higher cost. In recent years, permanent deployment of instruments has become more common in ocean bottom observatories and for monitoring the development of hydrocarbon fields. The rapid development of ocean bottom seismic instruments is leading to improvements in reliability, data quality, and efficiency, which makes it likely that ocean bottom seismics will continue to expand significantly in the future.

Cross-References

- Deep Seismic Reflection and Refraction Profiling
- Energy Partitioning of Seismic Waves
- Magnetic Anisotropy
- Seismic Data Acquisition and Processing
- Seismic, Waveform Modeling and Tomography
- Single and Multichannel Seismics
- Traveltime Tomography Using Controlled-Source Seismic Data

Acknowledgments We would like to thank Stuart Henrys (GNS Science) and an anonymous external reviewer for constructive comments. We also thank Jennifer Eccles (University of Auckland, New Zealand) and Frank Krüger (University of Potsdam, Germany) for providing the images in Figs. 3 and 4, respectively.

Bibliography

Arroyo IG, Husen S, Flueh ER, Gossler J, Kissling E, Alvarado GE (2009) Three-dimensional P-wave velocity structure on the shallow part of the Central Costa Rican Pacific margin from local earthquake tomography using off- and onshore networks. *Geophys J Int* 179:827–849

- Barkved OI, Kristiansen T (2005) Seismic time-lapse effects and stress changes: examples from a compacting reservoir. *Lead Edge* 24:1244–1248
- Barley B, Summers T (2007) Multi-azimuth and wide-azimuth seismic: shallow to deep water, exploration to production. *Lead Edge* 26:450–457
- Beranzoli L, Braun T, Calcara M, Casale P, DeSantis A, D'Anna G, DiMauro DGE, Favali P, Fuda J-L, Frugoni F, Gamberi F, Marani M, Millot C, Montuori C, Smriglio G (2003) Mission results from the first GEOSTAR observatory (AdriaticSea, 1998). *Earth Planets Space* 55:361–373
- Brueckmann W, Tyron MD, Bialas J, Feseker T, Lefeldt MR (2009) Monitoring the dynamic properties of an active Mud Volcano in the West Nile Delta. *Eos Trans AGU*, 90, Fall meeting supplement, Abstract OS21A-1156
- Caldwell J (1999) Marine multicomponent seismology. *Lead Edge* 18:1274–1282
- Christeson GL, Nakamura Y, McIntosh KD, Stoffa PL (1996) Effect of shot interval on ocean bottom seismograph and hydrophone data. *Geophys Res Lett* 23:3783–3786
- Constable S, Srnka LJ (2007) Special section – marine controlled-source electromagnetic methods: an introduction to marine controlled-source electromagnetic methods for hydrocarbon exploration. *Geophysics* 72:WA3–WA12
- Crawford WC, Webb SC, Hildebrand JA (1991) Seafloor compliance observed by long-period pressure and displacement measurements. *J Geophys Res* 96(B):16151–16160
- Eccles JD, White RS, Christie PAF (2009) Identification and inversion of converted shear waves: case studies from the European North Atlantic continental margins. *Geophys J Int* 179:381–400
- Gaiser J, Loinger E, Lynn H, Vetri L (2002) Birefringence analysis at Emilio field for fracture characterization. *First Break* 20:505–514
- Granli JR, Arntsen B, Sollid A, Hilde E (1999) Imaging through gas-filled sediments using marine shear-wave data. *Geophysics* 64:668–677
- Grion S, Exley R, Manin M, Miao X-G, Pica A, Wang Y, Granger P-Y, Ronen S (2007) Mirror imaging of OBS data. *First Break* 25:37–42
- Hobro JWD, Minshull TA, Singh SC (1998) Tomographic seismic studies of the methane hydrate stability zone in the Cascadia margin. In: Henriot JP, Mienert J (eds) *Gas hydrates; relevance to world margin stability and climatic change*. Geological society of London special publication, vol 137. Geological Society, London, pp 133–140
- Husen S, Kissling E, Flueh E, Asch G (1999) Accurate hypocentre determination in the seismogenic zone of the subducting Nazca Plate in northern Chile using a combined on-/offshore network. *Geophys J Int* 138:687–701
- Huws DG, Davis AM, Pyrah JR (2000) A nondestructive technique for predicting the in situ void ratio for marine sediments. *Mar Georesour Geotechnol* 18:333–346
- Kaneda Y, Kawaguchi K, Araki E, Sakuma A, Matsumoto H, Nakamura T, Kamiya S, Ariyoshi K, Baba T, Ohori M, Hori T (2009) Dense Ocean floor network for earthquakes and tsunamis (DONET) – development and data application for the mega thrust earthquakes around the Nankai trough. *Eos Trans AGU*, 90, Fall meeting supplement, Abstract S53A-1453
- Kodaira S, Takahashi N, Kato A, Park J-O, Iwasaki T, Kaneda Y (2000) High pore fluid pressure may cause silent slip in the Nankai Trough. *Science* 308:1295–1298
- Korenaga J, Holbrook WS, Singh SC, Minshull TA (1997) Natural gas hydrates on the southeast US margin: constraints from full waveform inversion and traveltime inversion of wide-angle seismic data. *J Geophys Res* 102(B):15345–15365
- Korenaga J, Holbrook WS, Kent GM, Kelemen PB, Detrick RS, Larsen HC, Hopper JR, Dahl-Jensen T (2000) Crustal structure of the southeast Greenland margin from joint refraction and reflection seismic tomography. *J Geophys Res* 105(B):21,591–21,614

- MacLeod MK, Hanson RA, Bell CR (1999) The Alba field ocean bottom cable seismic survey: impact on development. *Lead Edge* 18:1306–1312
- Margrave GF, Lawton DC, Stewart RR (1998) Interpreting channel sands with 3C-3D seismic data. *Lead Edge* 17:509–513
- Morgan JV, Christeson GL, Zelt CA (2002) Testing the resolution of a 3D velocity tomogram across the Chicxulub crater. *Tectonophysics* 355:215–226
- Nguyen XN, Dahm T, Grevemeyer I (2009) Inversion of Scholte wave dispersion and waveform modeling for shallow structure of the Ninetyeast ridge. *J Seismol* 13:543–559
- Peacock S, Westbrook GK, Graham DP (1997) Seismic velocities in the northern Barbados ridge accretionary complex, site 949. In: Shipley TH, Ogawa Y, Blum P, Bahr JM (eds) *Proceedings ODP, scientific results*, vol 156. Ocean Drilling Program, College Station, pp 263–275
- Plaza-Faverola A, Bünnz S, Mienert J (2010) Fluid distributions inferred from P-wave velocity and reflection seismic amplitude anomalies beneath the Nyegga pockmark field of the mid-Norwegian margin. *Mar Pet Geol* 27:46–60
- Rafavich F, Kendall CHSC, Todd TP (1984) The relationship between acoustic properties and the petrographic character of carbonate rocks. *Geophysics* 49:1622–1636
- Sourbier F, Operto S, Virieux J, Amestoy P, L'Excellent J-Y (2009a) FWT2D: a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data – part 1 algorithm. *Comput Geosci* 35:487–495
- Sourbier F, Operto S, Virieux J, Amestoy P, L'Excellent J-Y (2009b) FWT2D: a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data – part 2 numerical examples and scalability analysis. *Comput Geosci* 35:496–514
- Stewart RS, Gaiser JE, Brown RJ, Lawton DC (2002) Converted-wave seismic exploration: methods. *Geophysics* 67:1348–1363
- Talwani M, Zelt B (1998) Some recent developments in the acquisition and processing of seismic data. *Tectonophysics* 286:123–142
- Willoughby EC, Latychev K, Edwards RN, Schwalenberg K, Hyndman RD (2008) Seafloor compliance imaging of marine gas hydrate deposits and cold vent structures. *J Geophys Res* 113:B07107
- Zelt CA, Smith RB (1992) Seismic traveltimes inversion for 2-D crustal velocity structure. *Geophys J Int* 108:16–34

Ocean, Spreading Center

K. S. Krishna

Centre for Earth, Ocean and Atmospheric Sciences,
University of Hyderabad, Hyderabad, India

Definition

The *spreading center*, also known as a segment of the mid-ocean ridge, is an underwater mountainous feature with a rift valley lying between the two lithospheric plates, through which new magma material being continuously spread out and led to construction of the lithospheric plates on either side.

The *ridge internal structure*, the partially melted mantle up welled through the mid-ocean ridge segments/spreading centers, forms a basaltic crust of 3–6 km thick, but at a later stage, the same crust moves away from the ridge crest and allows growth of mantle component.

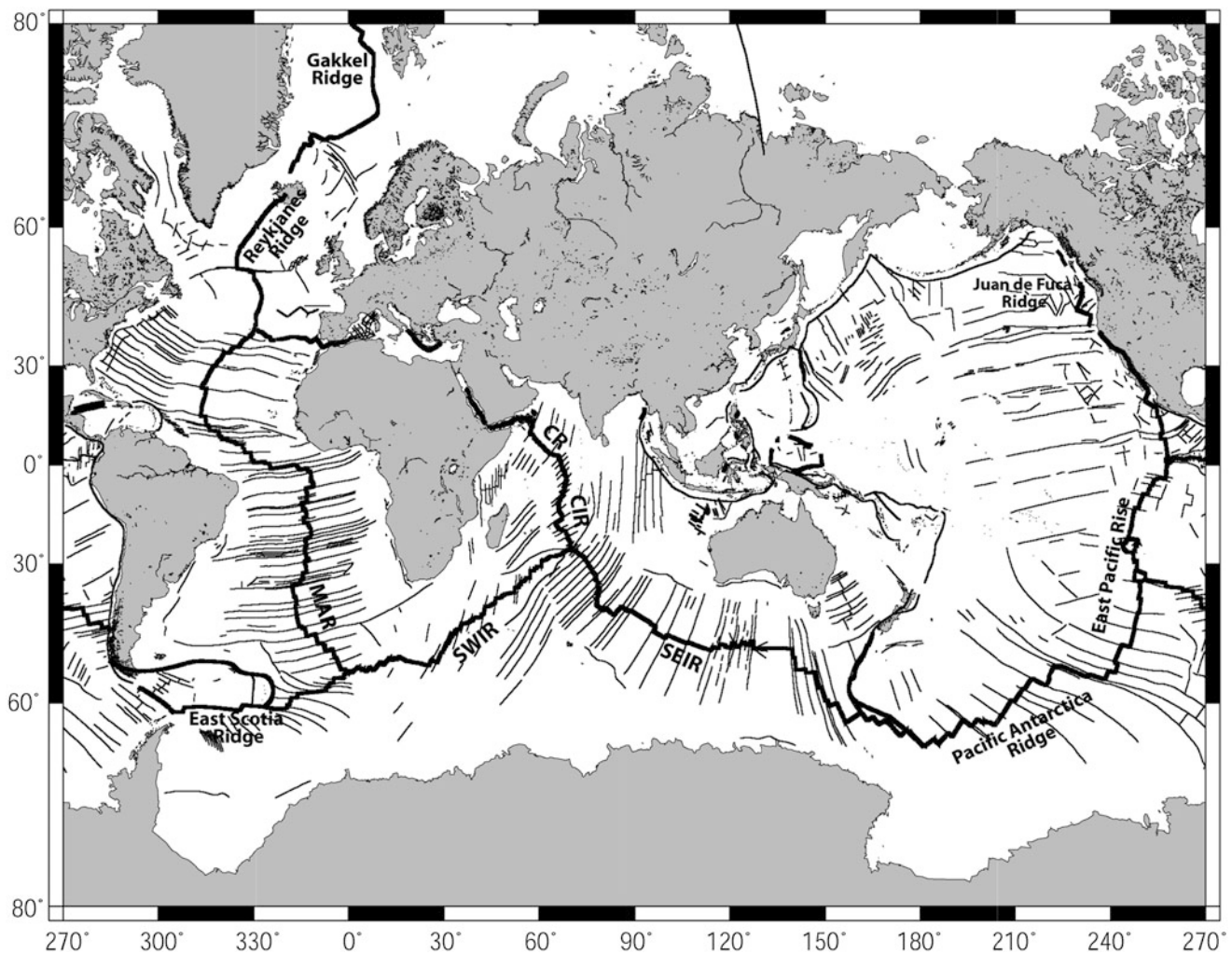
Introduction

The rigid outer layer of the Earth comprising the crust and uppermost mantle (generally about 70–100 km thick below the oceans), *termed as oceanic lithosphere*, is divided into a number of tectonic plates. Presently there are about eight major tectonic plates, viz., North American, South American, African, Indian, Australian, Antarctica, Eurasian, and Pacific; seven minor and several more smaller plates are covering the Earth's surface (for details, see Wikipedia website: http://en.wikipedia.org/wiki/List_of_tectonic_plates). The lithospheric plates are bounded by one of the three main types of geologic features: (1) mid-oceanic ridges, (2) subduction zones, and (3) transform faults. The boundaries are narrow deforming zones and are associated with intense seismicity, whereas the plate's interiors are relatively stable. The plates, on which continents and ocean floor commonly lie, are in continuous motion at a speed of few centimeters per year. Each plate is in relative motion with respect to the other on the surface of the Earth. The relative motion produces new crust at mid-oceanic ridges and consumes old lithosphere at subduction zones. Apart from these tectonic processes, on large geologic framework, plates do undergo breakups and unifications.

Morphology and Spreading Rates of Mid-Oceanic Ridge

A mid-ocean ridge is an underwater linear physiographic feature and also serves as a divergent tectonic boundary between two lithospheric plates. The mid-ocean ridges are physically connected within the ocean and also with the ones lying in other oceans (Fig. 1). Thus, the ridge systems encircle the entire globe from one ocean to the other and extend for a total length of about 65,000 km as a single global entity (Fig. 1). Many of the major oceanic ridges in the world oceans have been reconfigured in the geologic past that led to unification and split of the lithospheric plates. One of the well-mapped ancient oceanic ridges in the Indian Ocean is the Wharton Ridge, once separated the Indian and Australian plates, and the ridge cessation at about 42 million years ago led to unification of both aforesaid plates into a single Indo-Australian plate (Liu et al. 1983; Krishna et al. 1995, 2012).

In deep-sea regions, long, linear, and elevated features with variable undulations are observed (Fig. 2). They are the locations where hot magma is continuously raising and being added to the previously formed oceanic lithosphere on either side of the spreading ridge system. The upwelling magma rates along the global mid-oceanic ridges are not fixed; they greatly vary with a range from 6 to 200 mm per year. Following the seafloor expansion rates, the mid-oceanic ridges are classified into categories of slow-, intermediate-, and fast-spreading ridges. The Gakkel mid-oceanic ridge located in the Arctic



Ocean, Spreading Center, Fig. 1 General map of the global oceans showing mid-oceanic ridge system, subduction zones, and major fracture zones (after Royer et al. 1989 and Müller et al. 1997). MAR indicates

Mid-Atlantic Ridge in the Atlantic Ocean. CR, CIR, SWIR, and SEIR represent Carlsberg Ridge, Central Indian Ridge, Southwest Indian Ridge, and Southeast Indian Ridge, respectively, in the Indian Ocean

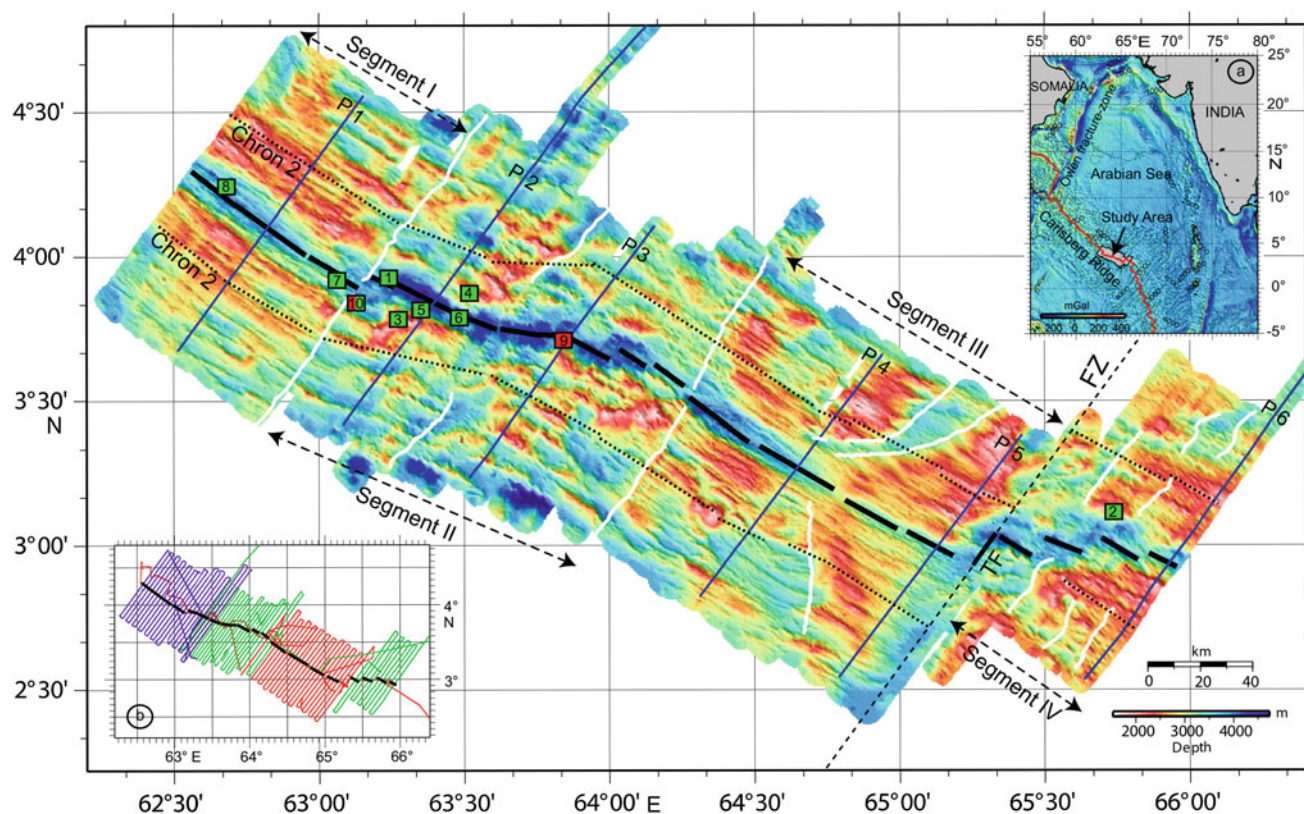
Ocean between Greenland and Siberia (Fig. 1) is the slowest-spreading ridge in world oceans with an upwelling magma rate of about 6 mm per year. On the other hand, the East Pacific Rise, particularly in equatorial region, generates magma at the fastest rates up to 200 mm per year. The seafloor morphology on flanks of the mid-oceanic ridge is largely controlled by the rate of seafloor accretions. Fast-spreading rates contribute to generally smooth and flat seafloor surfaces in the vicinity of the ridge, while the slow-spreading rates add exceptionally rough surfaces. The slow-spreading ridges are generally marked in their crest parts by an axial valley, called rift valley, which is clearly distinguishable along the Carlsberg Ridge in northern Indian Ocean (Fig. 2). The ridge crest is strongly characterized by rugged topography, steep valley walls, and wide rift valley floor (Kamesh Raju et al. 2008). Further the long ridge segment is bisected by less known non-transform discontinuities, while the fast-spreading ridge systems such as East Pacific Rise in Southeast Pacific Ocean and East Scotia Ridge in South

Atlantic Ocean do have crest-shaped features over the ridge axis without manifestation of rift valleys.

Internal Structure of the Mid-Oceanic Ridge

Along the mid-ocean ridge system, the lithospheric plates do expand on either side of it, which eventually lead to a partial melting of upwelling mantle and to form a layer of basaltic crust 3–6 km thick. Close to the ridge crest, the lithosphere may consist of oceanic crust only, but at a later stage, the same crust moves away from the ridge crest and allows growth of mantle component due to thermal cooling. Both lithosphere and elastic plate thicken continuously with the growing age of the ocean floor until about 60 Ma; thereafter the lithosphere and elastic plate thickness remains roughly stable.

A seismic reflection section across the East Pacific Rise is shown in Fig. 3 (Detrick et al. 1987). It shows the presence of



Ocean, Spreading Center, Fig. 2 Seafloor image of the Carlsberg Ridge segments, northern Indian Ocean. (Inset) Maps a and b show the study area (shaded box) and cruise tracks over the segment of the Carlsberg Ridge. (The image is reproduced from Kamesh Raju et al. 2008)

strong cohesive reflections on both sides of the section with a kind of discontinuity in the middle. These reflections possibly mark the Moho boundary separating the crust from the mantle, whereas the reflections lying in a shallow depth show a top of crustal low-seismic-velocity zone (LVZ). Detrick et al. (1987) interpreted the top of the LVZ as the roof of an axial magma chamber (AMC) within the crust beneath the ridge axis. On either side of the chamber, the reflections are weaker and perhaps represent the frozen top of the magma chamber body (shown with arrows in Fig. 3). Also, the reflections from Moho boundary are absent below the magma chamber. A sketch of the interpreted seismic structure of the East Pacific Rise is presented in Fig. 4.

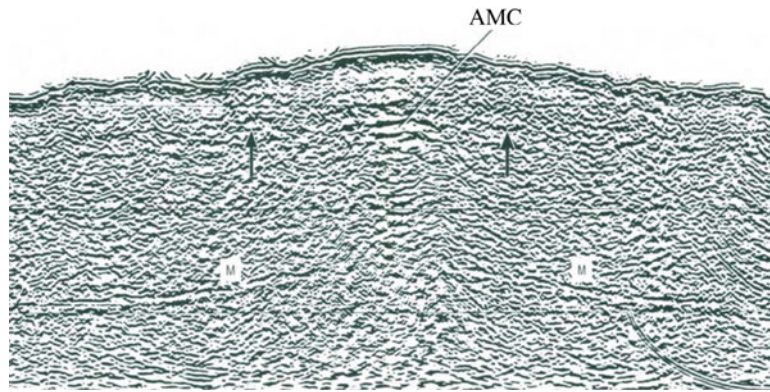
Below a few of mid-oceanic ridge segments, it is found that the size of the axial magma chamber is either small or absent altogether. For example, multichannel reflection experiment carried out on the Mid-Atlantic Ridge (Detrick et al. 1990) did not find evidences for the presence of intracrustal reflections associated with an axial magma chamber. Seismic experiments, in general, demonstrated that along slow-spreading ridges magma chambers are either absent or extremely short-lived, if at all they exist. The Southwest Indian Ridge, an ultraslow-spreading ridge is a classic example for demonstrating the migration of melt supply. Geophysical studies in close proximity of the Prince Edward FZ

revealed that the melt supply center has migrated and reduced magmatic activity after some time (Sato et al. 2013).

Seismic refraction experiments carried out on a ridge segment between 8° and 9°S of the Mid-Atlantic Ridge revealed largely a different seismic structure from that of the East Pacific Rise. Interestingly, an axial magma chamber is not traceable within the Mid-Atlantic Ridge crust. Velocity models of the Mid-Atlantic Ridge segment (Minshull et al. 2003) suggest a total thickness of the crust ranging from 6.5 to 10 km with individual layers 2 and 3 having thickness of approximately 3 and 5 km, respectively (Fig. 5). The velocity heterogeneities are much greater in layers 2A and 2B, ranging from 2.65 to 3.0 km/s and from 4.6 to 6.8 km/s, respectively, while the velocities in layer 3 are remarkably uniform.

Relationships Between Spreading Rate, Seismic Structure, and Ridge-Axis Morphology

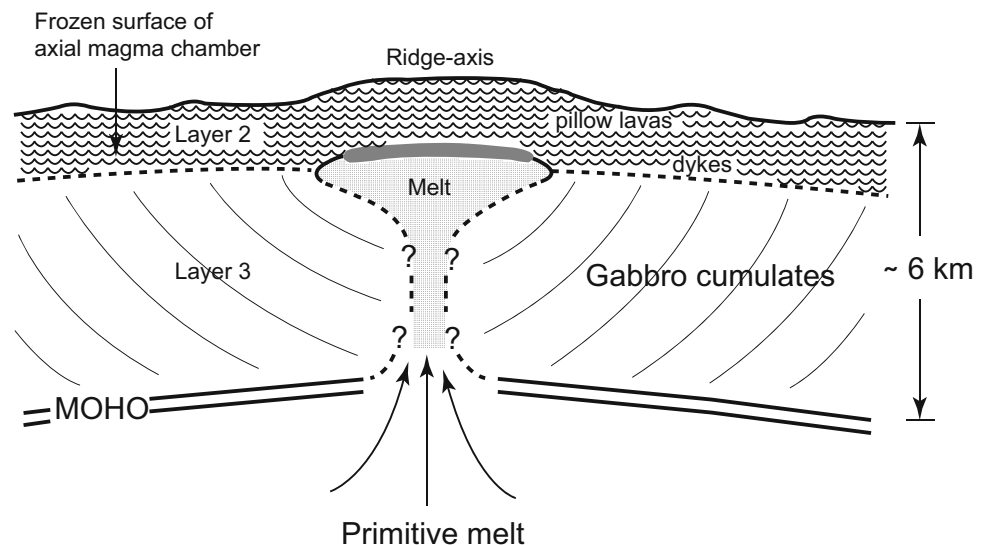
Geophysical experiments over the global mid-oceanic ridges have led to the discovery of some explicit relationships between spreading rate, seismic structure, and ridge-axis morphology. Purdy et al. (1992) found that as full spreading rate decreases from 150 to 20 mm yr.⁻¹ the depth to the top of the axial low-seismic-velocity zone (LVZ) increases from ~1



Ocean, Spreading Center, Fig. 3 Seismic reflection cross section across the East Pacific Rise shows the internal structure of the oceanic crust beneath the ridge axis. AMC represents significant seismic signals reflected from the roof of an axial magma chamber. Solid arrows on

either side mark relatively weaker reflections, perhaps representing the frozen top of the magma chamber, away from the ridge axis. M indicates the Moho boundary. (The image is reproduced from Detrick et al. 1987)

Ocean, Spreading Center, Fig. 4 A sketch showing the crustal structure beneath the East Pacific Rise (Reproduced from Detrick et al. 1987)



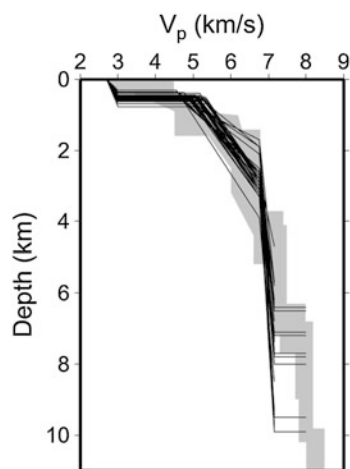
to ~4 km. In another study, Minshull et al. (2003) established a relationship between crustal thickness and ridge-axis morphology at a constant spreading rate. The ridge segments having large crustal thicknesses are associated with rifted axial highs, whereas the segments with reduced crustal thickness have deep axial valleys.

Large-scale detachment faults are found to occur on slow- and ultraslow-spreading ridges. At the slow-spreading ridges, detachment faults tend to occur where local magma supply is relatively low, while at the ultraslow-spreading rates, they tend to occur where local magma supply is relatively high (Zhiteng et al. 2013). Resource explorations over the global mid-oceanic ridges have also proven that they are important locations for metallic mineral depositions (Rona 1980). In subsequent studies, Zhiteng et al. (2013) have found the useful relationship that the detachment faults are accompanied by hydrothermal activities. Recently, Suo et al. (2017),

by analyzing the data of spreading rate, bathymetry, gravity, and geochemical data along the Southwest Indian Ridge, established a weak positive trend between spreading rates and occurrences of hydrothermal sites, thereby inferring that the long-term heat supply is the first-order control on global distribution of hydrothermal activity.

Summary

A mid-ocean ridge, an underwater mountainous feature, extends for a length of about 65,000 km as a single continuous body in the global oceans. This serves as a window into the Earth's interior, through which magma is continuously upwelling and drives the lithospheric plates on either side in order to accommodate newly accreted crust. Many of the oceanic ridges in the world oceans have been abandoned in the geologic past



Ocean, Spreading Center, Fig. 5 Velocity-depth profiles (solid lines) beneath the Mid-Atlantic Ridge segment between 8 and 9°S. Shaded region marks the range of velocities for 0–7 Ma Atlantic oceanic crust (White et al. 1992). (The figure is reproduced from Minshull et al. 2003)

and led to resume the activity elsewhere either in the intra-oceanic or intracontinental plate regions. Partially melted mantle upwelled through the mid-ocean ridge forms a layer of basaltic crust 3–6 km thick, but at a later stage, the same crust moves away from the ridge crest and allows growth of mantle component. The upwelling magma rates along the mid-oceanic ridges, in general, control the internal structure. Fast-spreading ridge such as East Pacific Rise has an axial magma chamber (AMC) at a shallow depth within the crust, whereas beneath an intermediate-spreading ridge, the Mid-Atlantic Ridge, an axial magma chamber is not traceable. Geophysical experiments over the global mid-oceanic ridges have found some explicit relationships between spreading rate, seismic structure, and ridge-axis morphology.

Cross-References

- Earth's Structure, Global
- Earth's Structure, Upper Mantle
- Earthquakes and Crustal Deformation
- Lithosphere, Oceanic
- Lithosphere, Oceanic: Thermal Structure
- Plate-Driving Forces
- Plate Motions in Time: Inferences on Driving and Resisting Forces
- Seafloor Spreading
- Seismic Structure at Mid-ocean Ridges

Bibliography

Detrick RS, Buhl P, Vera E, Mutter J, Orcutt J, Madsen J, Brocher T (1987) Multi-channel seismic imaging of a crustal magma chamber along the East Pacific rise. *Nature* 326:35–41

- Detrick RS, Mutter JC, Buhl P, Kim II (1990) No evidence from multi-channel reflection data for a crustal magma chamber in the MARK area on the Mid-Atlantic Ridge. *Nature* 347:61–64
- Kamesh Raju KA, Chaubey AK, Amarnath D, Mudholkar AV (2008) Morphotectonics of the Carlsberg Ridge between 62° 20' and 66° 20'E, northwest Indian Ocean. *Mar Geol* 252:120–128
- Krishna KS, Gopala Rao D, Ramana MV, Subrahmanyam V, Sarma KVLNS, Pilipenko AI, Shcherbakov VS, Radhakrishna Murthy IV (1995) Tectonic model for the evolution of oceanic crust in the northeastern Indian Ocean from the Late Cretaceous to the Early Tertiary. *J Geophys Res* 100:20011–20024
- Krishna KS, Abraham H, Sager WW, Pringle M, Frey FA, Gopala Rao D, Levchenko OV (2012) Tectonics of the Ninetyeast Ridge derived from spreading records in adjacent oceanic basins and age constraints of the ridge. *J Geophys Res* 117:B04101. <https://doi.org/10.1029/2011JB008805>
- Liu CS, Curray JR, McDonald JM (1983) New constraints on the tectonic evolution of the Eastern Indian Ocean. *Earth Planet Sci Lett* 65:331–342
- Minshull TA, Bruguier NJ, Brozena JM (2003) Seismic structure of the Mid-Atlantic Ridge, 8–9°S. *J Geophys Res* 108:2513. <https://doi.org/10.1029/2002JB002360>
- Müller RD, Roest WR, Royer J-Y, Gahagan LM, Sclater JG (1997) Digital isochrons of the world's ocean floor. *J Geophys Res* 102:3211–3214
- Purdy GM, Kong LS, Christenson GL, Solomon SC (1992) Relationship between spreading rate and the seismic structure of Mid-Ocean Ridges. *Nature* 355:815–817
- Rona PA (1980) TAG hydrothermal field: Mid-Atlantic Ridge crest at latitude 26°N. *J Geol Soc Lond* 137:385–402
- Royer J-Y, Sclater JG, Sandwell DT (1989) A preliminary tectonic fabric chart of the Indian Ocean. *Proc Indian Acad Sci (Earth Planet Sci)* 98:7–24
- Sato T, Okino K, Sato H, Mizuno M, Hanyu T, Seama N (2013) Magmatic activities on the Southwest Indian Ridge between 35E and 40E, the closest segment to the Marion hotspot. *Geochim Geophys Geosyst* 14:5286–5307. <https://doi.org/10.1002/2013GC004814>
- Suo Y, Li S, Li X, Zhang Z, Ding D (2017) The potential hydrothermal systems unexplored in the Southwest Indian Ocean. *Mar Geophys Res* 38:61–70. <https://doi.org/10.1007/s11001-016-9300-5>
- White RS, McKenzie D, O'Nions RK (1992) Oceanic crustal thickness from seismic measurements and rare earth element inversions. *J Geophys Res* 97:19683–19715
- Zhiteng Yu, Li J, Liang Y, Han X, Zhang J, Zhu L (2013) Distribution of large-scale detachment faults on Mid-Ocean Ridges in relation to spreading rates. *Acta Oceanol Sin* 32(12):109–117. <https://doi.org/10.1007/s13131-013-0397-y>

Orogenic Belts

A. M. Celâl Şengör

Faculty of Mines, Department of Geology and Eurasia
Institute of Earth Sciences, Istanbul Technical University,
Istanbul, Turkey

Synonyms

Fold belt; Folded belt; Folded mountains; Orogen

Partial Synonyms

Mountain belt; Height Systems, Vertical Datums and Their Unification

Definition

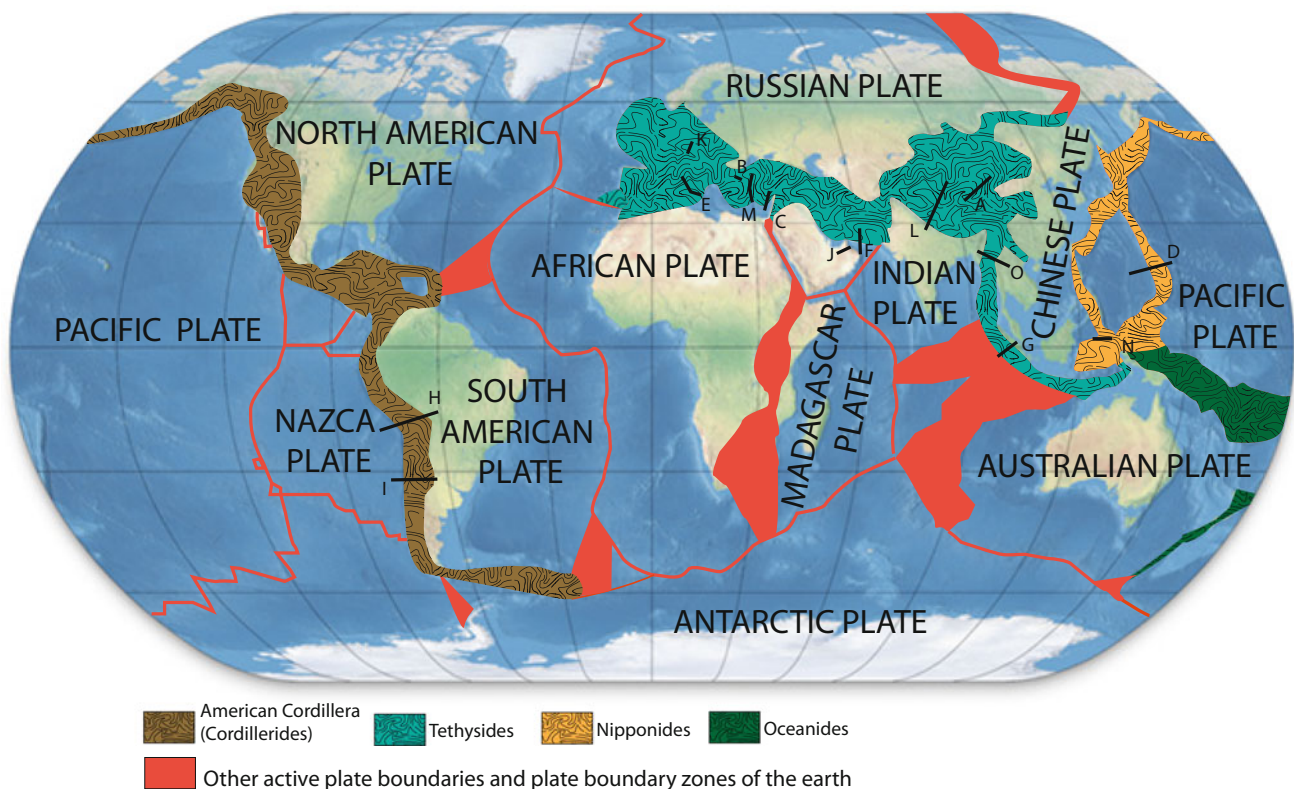
An orogenic belt is a lithospheric structure that is formed by the collective work of any group of convergent plate margin processes (see ► “[Lithosphere, Continental](#)”, and ► “[Lithosphere, Mechanical Properties](#)”). Kober (1921, p. 21) suggested the term “orogen” to designate the orogenic belts and his term is now generally used. Although orogeny is derived from the ancient Greek *oros* (mountain) and *genesis* (birth), just as old rifts and taphrogens may no longer have any topographic expression because of tectonic and/or erosional events, some, especially old, orogenic belts may no longer appear as linear/arcuate mountain ranges for the same reasons. Their original height with respect to their surroundings can be gauged by their stratigraphically older contents, or metamorphically higher grade of their uplifted corpus plus trains of intermediate and felsic igneous intrusions they usually contain.

Introduction

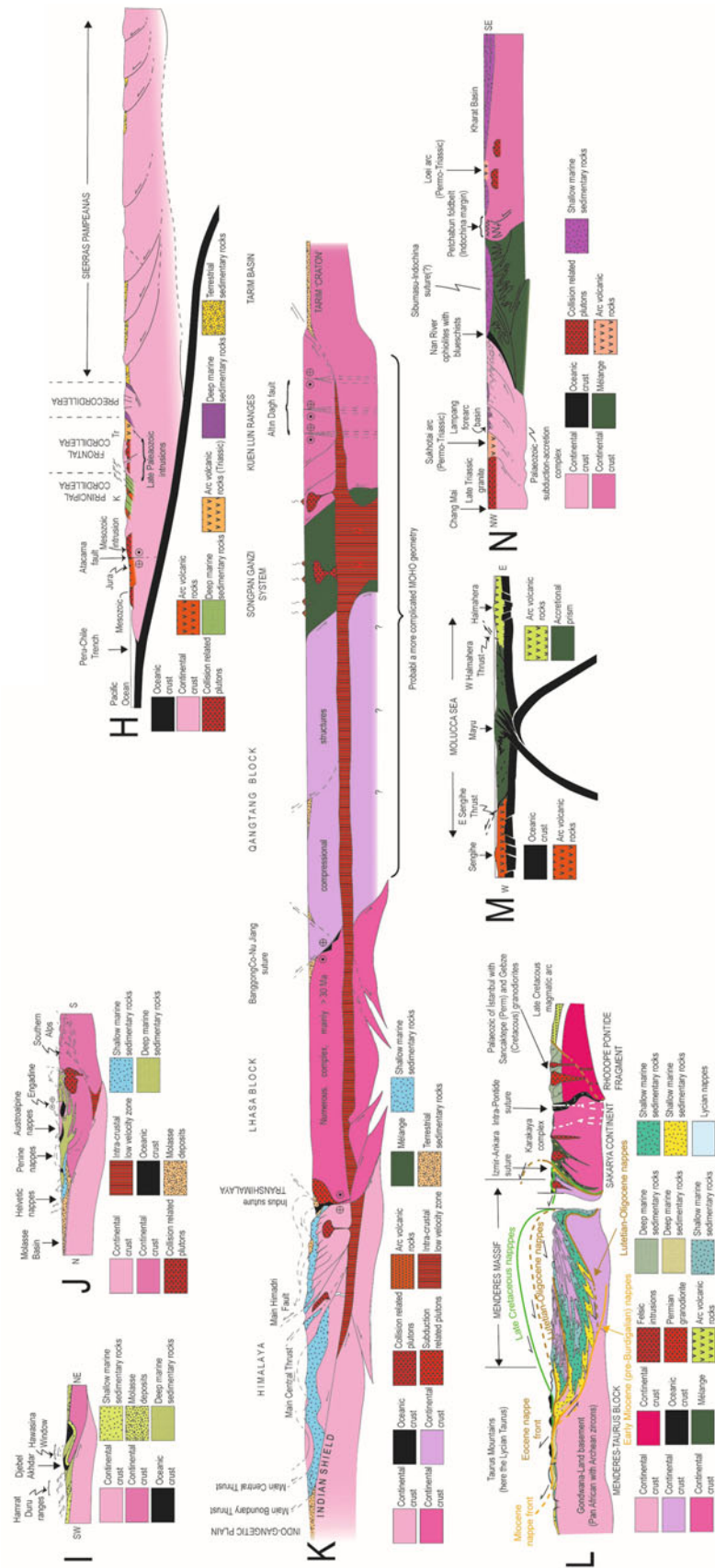
Orogenic belts or orogens constitute one of the three main classes of lithospheric-scale structures resulting from differential motion of parts of the lithosphere (Figs. 1 and 2):

lithospheric convergence creates the orogens, just as lithospheric divergence creates taphrogens and lithospheric horizontal simple shear along subvertical planes creates keirogens (Figs. 1 and 2). However, at present, only 20.5% of the active plate boundaries show normal convergence and 21% normal divergence. Some 8% show pure dextral strike-slip and some 6% pure sinistral strike-slip. The remaining 58.5% display some deviation from the three end-members, with relative motion vector angles to boundary strikes varying between 0° and 67° (Fig. 3: Woodcock 1986). Plate boundaries have probably shown a similar behavior in the past, so only about half of all the orogens a geologist may encounter is likely to show normal shortening, although tectonic escape may add significant strike-slip into an orogenic architecture even when plate convergence is normal to the trench and/or collision front (see ► “[Plate Motions in Time: Inferences on Driving and Resisting Forces](#)” and ► “[Plates and Paleoreconstructions](#)”). It should also be borne in mind that in the past continental collisions at times eliminated significant lengths of subduction zones during the construction of supercontinents such as the late Palaeozoic Pangaea resulting in the diminution of orogenic activity leading to a global slowing down of plate motions and lowering sea-level.

Orogenic belts are major features of the lithosphere on which 98% of the world’s largest earthquakes occur (<https://>



Orogenic Belts, Fig. 1 Active orogens of the world in the framework of world’s active plate boundaries and plate boundary zones. The lettered black lines refer to the cross-sections displayed in Fig. 2



Orogenic Belts, Fig. 2 Selected true-scale cross-sections of various types of active orogenic belts in the world. All of the cross-sections have the same scale. (Modified from Şengör (1990a)). The locations of the cross-sections are shown in Fig. 1 Key to abbreviations: *A* Aar massif, *AA* Austroalpine nappes, *B* Bergell pluton, *D4* Djebel Akhdar, *E* Eocene (in A) and Engadine (in K), *EST* East Senghe Thrust, *G* Gotthard massif, *H* Helvetic nappes, *HD* Hamrat Duru ranges, *HW* Hawasina Window, *M* Molasse basin, *MBT* Main Boundary Thrust, *MCT* Main Central Thrust, *MHF* Main Himaladri Fault, *MR* Mediterranean Ridge, *N* Neogene, *NAF* North Anatolian Fault, *P* Pennine nappes, *Pz* Palaeozoic, *S4* Southern Alps, *T* Tavetsch massif, *Tz* Triassic

www.usgs.gov/faqs/where-do-earthquakes-occur; last visited on 12th December 2019). Out of these 81% alone occur along the Pacific rim, the so-called “Ring of Fire,” and 17% along the Alpine-Himalayan mountain ranges. About 1/3 of the world’s active volcanoes occur along the Ring of Fire (total of 452 volcanoes) and 1/6 of them along the Alpine-Himalayan mountains including Indonesia (total 216 volcanoes). Since the diameter of the Earth does not change to any appreciable degree, ocean-floor spreading and subduction must balance each other. Spreading occurs along an 80,000 km-long mid-oceanic ridge system, whereas subduction and postcollisional shortening occur along a 55,000 km long front along the active orogens of the world. Thus, the rate of total deformation along orogenic belts must be almost 1.5 times higher than along the spreading centers, which is corroborated by the concentration of the largest earthquakes along them making them the most active parts of the lithosphere.

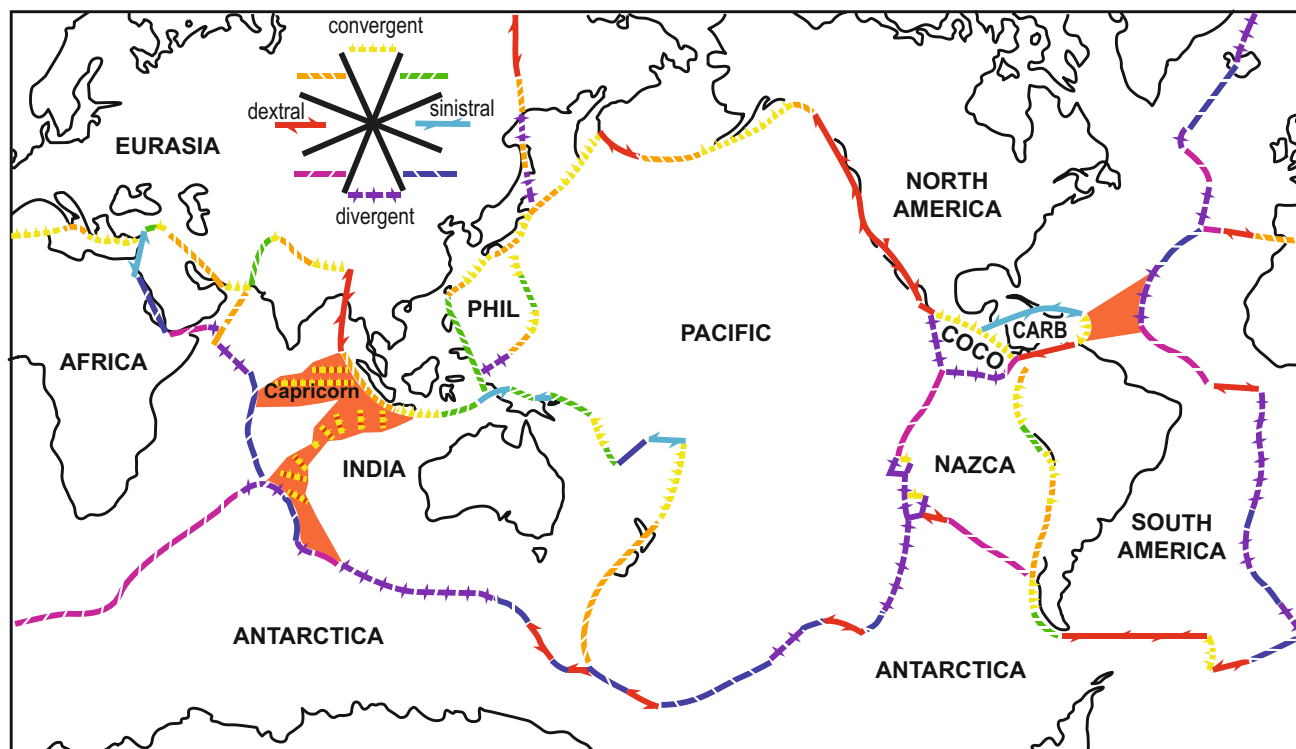
Orogenic belts are the main factories of the continental crust, at least since after the Hadean. Because the Earth is the only planet in the Solar System with a continental crust, orogeny, that is, the making of orogenic belts, must be an essential part of any theory of the Earth. Orogens house the world’s main mineral wealth and an understanding of their nature is critical for successful mineral exploration. The basins flanking them and genetically associated with them

contain a significant portion of the world’s hydrocarbon resources. Knowing how they evolve is an indispensable part of oil and gas exploration on this planet. Finally, orogenic belts create mountains which control a good portion of the world’s climate and climate in turn partly controls the growth of mountain belts.

For good reviews of orogenic belts and orogeny, see Smith (1976), Dewey (1977, 1980, 2002, 2005), Johnson and Harley (2012), Miyashiro et al. (1982), Molnar (1986), Rodgers (1987, 1990, 1991, 1995, 1997), Şengör (1990a, 2006). For the timing question in orogenic belts, see Şengör (1991a and the references therein; Dewey 2005). This very inadequate list of references, imposed by the available space in an encyclopedia article that was synthesized from the publications of more than three hundred geologists, would lead the reader at least to some other sources discussing the detailed aspects of orogens and orogeny. The rest the reader can search in the internet using the concepts discussed herein as keywords.

The Orogenic Architecture

Figure 4 illustrates the main elements of orogens. The defining feature of an orogen is its subduction zone (Fig. 5) which is commonly lost after continental collision. Eduard Suess first emphasized in 1875 the fundamental asymmetry of

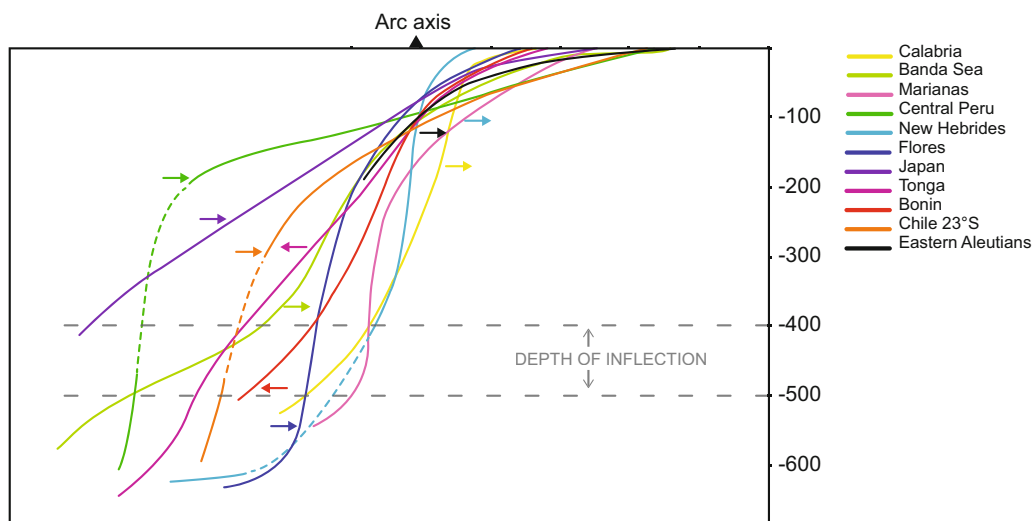
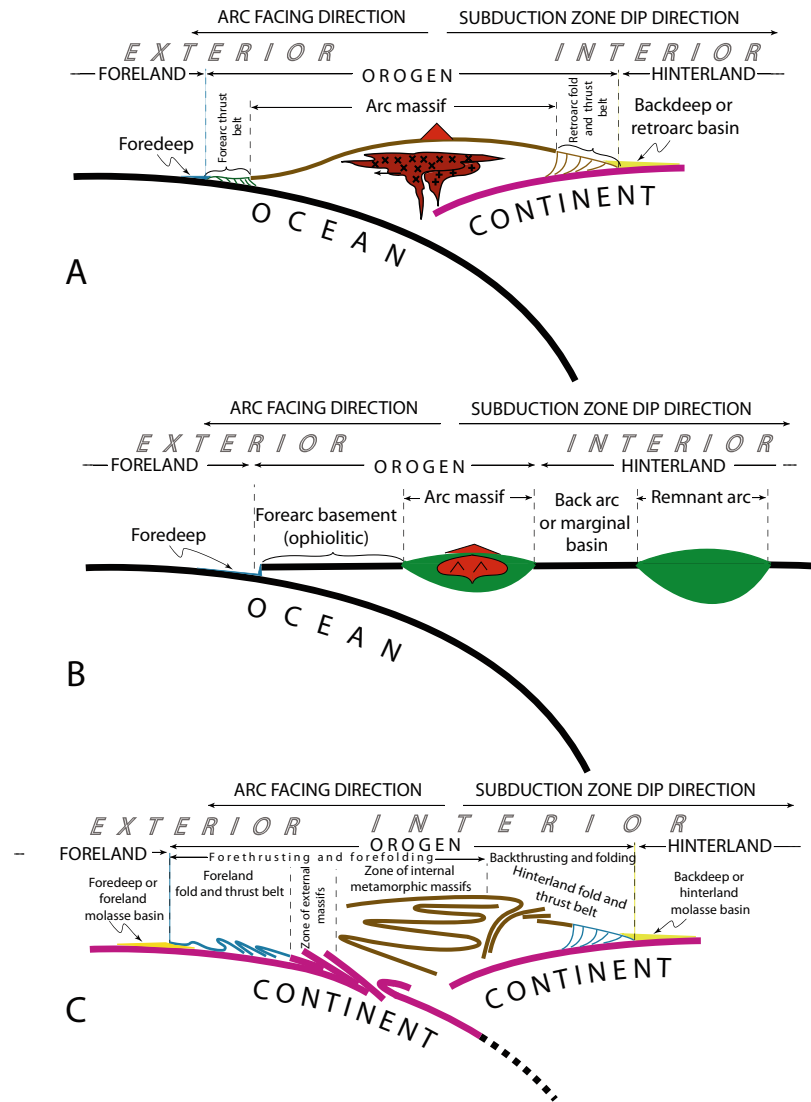


Orogenic Belts, Fig. 3 World plate boundary map, modified from Woodcock (1986) showing plate boundaries smoothed with respect to oceanic ridge-transform systems and are ornamented according to the

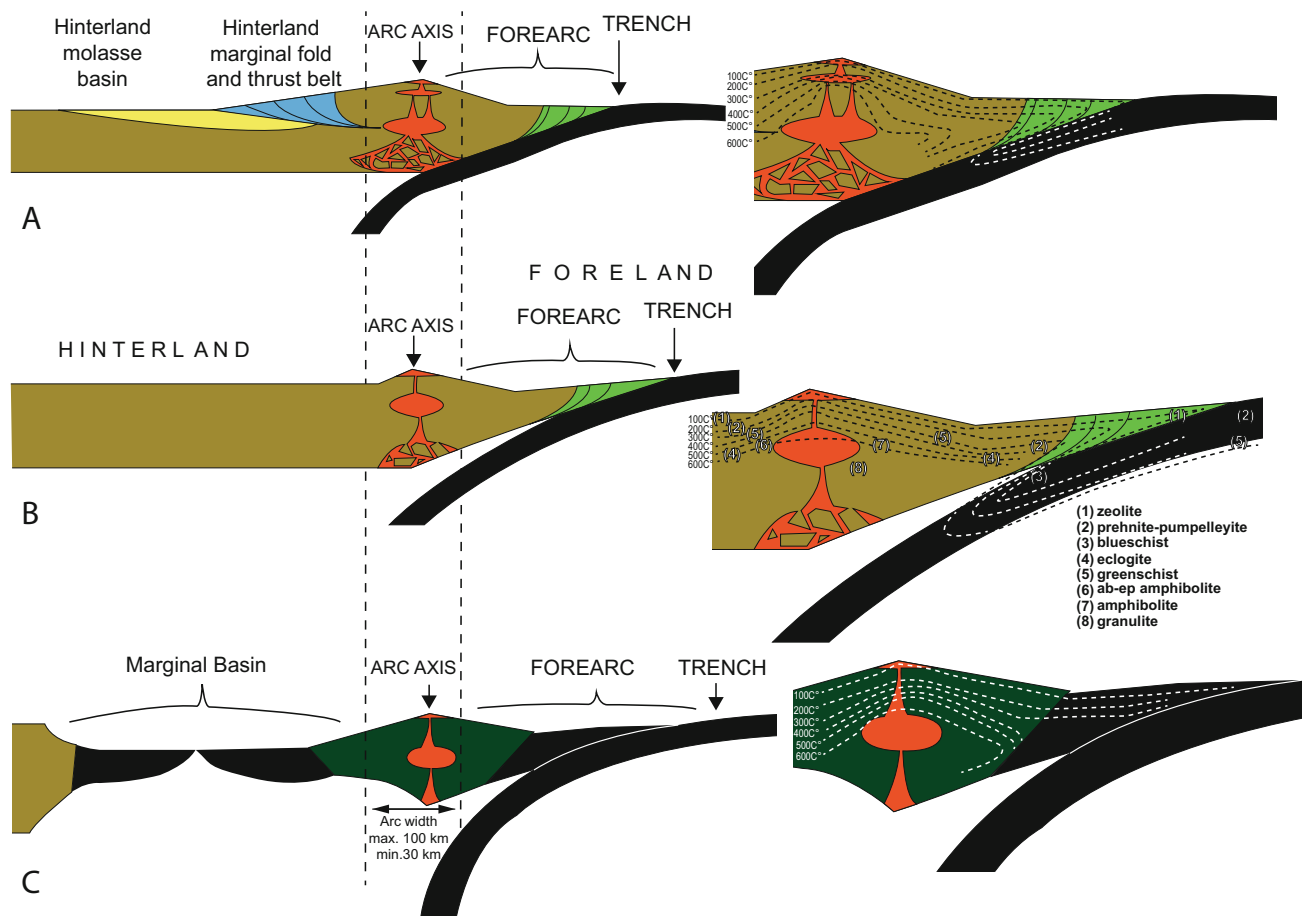
angle between the boundary and the local relative velocity vector (shown in the inset)

Orogenic Belts,

Fig. 4 Orogenic terminology illustrating the basic orogenic architecture. (a) A subduction-controlled shortening continental margin orogen such as the Andes. (b) An extensional subduction-controlled orogen such as the Marianas/Kyushu-Palau System. (c) A collisional orogen



Orogenic Belts, Fig. 5 Various subducting slabs from the world's subduction zones compiled from different sources. Note the great variability of the subduction angles, but the fixed position of the arc axis above a certain depth of the slab. Arrows indicate advance or retreat of the slab



Orogenic Belts, Fig. 6 Simplified cross-section across three types of subduction-related, magmatic arc orogens. (a) Shortening, (b) neutral, (c) extending; detail to the right of a, b and c show the schematic metamorphic isotherms in each

orogenic belts and distinguished forelands from hinterlands and external parts of orogens from their internal parts. The fundamental architectural elements in a convergent plate boundary is the *foreland*, made up of the subducting lithosphere, the *orogen*, the deformed edges of the converging plates with magmatism being confined mainly to the upper plate, and the little to undeformed parts of the overlying plate called the *hinterland*. Parts of an orogen closer to the foreland are called its *externides* and are commonly little metamorphosed and deformed, whereas those parts closer to the hinterland constitute the *internides* of an orogen characterized by intense metamorphism and deformation, which, however, dies out towards the more internal parts as the orogen merges with the hinterland. An orogen is said to *face* towards its foreland, that is, in the direction opposite to the dip direction of its formative subduction zone. Dominant vergence of orogenic structures is usually towards the foreland, towards which the main tectonic transport occurs, but this is not universally so.

Figure 4 shows the main architectural elements of subduction- and collision-controlled orogens. The subduction controlled-orogens (Fig. 6) may be shortening (Fig. 6a),

neutral (Fig. 6b) or extensional (Fig. 6c) type depending on the age of the subducting slab. Opinions about the effect of the age of the slab on creating extension arcs have been divided. While some argue that young slabs induce shortening, others have argued the opposite because of deep-mantle buckling of young slabs. Although collisional orogens (Fig. 4c) may have topography-induced secondary extension, wholesale extensional collapse of orogens commonly requires a different process, such as transtension or a switch from plate convergence to plate divergence, to interfere.

Subduction-Controlled Orogens

The fore-arc: In most subduction-controlled orogens, the formative subduction zone meets the overlying plate usually at a deep-sea trench (e.g., Fig. 2c; world's deepest point, the Challenger Deep, located at 11°22.4 N and 142°35.5 E, with a depth range of –10,902 to 10,929 m is in the Mariana Trench; also Fig. 2g, h); where such a trench is not present (e.g., Fig. 2d, E), it is a consequence of its overflowing with sediment as in the Makran case. The forward lip of the overlying plate in subduction-controlled orogens called the arc-trench gap or the fore-arc (Dickinson and Seeley 1979) may be

characterized by a subduction-accretion complex of variable size (Fig. 2b, d, e, f, m). A trapped ophiolite (Fig. 2c, m), or continental basement, may form directly the inner wall of the trench (as in the case of the Arequipa Massif, Peru-Chile trench: Fig. 2g). This latter case may come about because of tectonic erosion or strike-slip removal of a preexisting inner wall consisting of an accretionary complex or an ophiolitic fore-arc basement. The accretionary complexes are made up of material scraped off the subducting slab including both sedimentary and igneous rocks of the ocean floor (the former usually dominating by far) and/or the turbidites filling the trench fed both across, from the arc and from the fore-trench bulge, and along, its axis. Accretion occurs where sediment thickness in trench exceeds 500 m to 1 km, whereas in places where that thickness is below 500 m subduction erosion takes place. The scraped off material is stacked into thrust slices commonly verging towards the trench, but in some (e.g., Washington State and Gulf of Alaska trench inner slopes) a landward vergence can also develop already in the trench inner wall, possibly resulting from the presence of low-strength, probably overpressured, zones in trench and slope sediments and the presence of oceanward-dipping surfaces of weakness in the trench fill. In some accretionary complexes, especially the large ones such as the Alaskan wedge, major vergences both synthetic and antithetic to subduction may develop. Deformation in accretionary complexes is penetrative leading to much dewatering, in places lubricating the decollement under the accretionary wedges (e.g., in the Seram Trench in Indonesia, in the Franciscan and Kodiak complexes in the North American Cordillera, and in the Ligurian units in the Appenines), and foliation development. In most large accretionary wedges, the wedge overthrusts the backstop, that is, the arc massif, along later-developed “secondary” thrusts. Accretionary wedges may be less than 2 km thick (e.g., in the Peru Trench: Fig. 2h) to some 36 km thick (e.g., in the Cascadia Accretionary Wedge). Where they are caught up in collision zones, they may further thicken to more than 70 km as in the Songpan Ganzi System in eastern Tibet (Fig. 2k). Metamorphism in them ranges from none at their surface to granulite grade along the bottoms of thick wedges where, along with much underplating, spreading centers have been subducted, as in the Chugach Mountains in Alaska. Where trench sedimentation is extremely limited as in intraoceanic arc systems such as the Marianas, the arc-trench gap may form from trapped oceanic lithosphere, essentially an ophiolite, because it is now incorporated into an orogenic edifice (Fig. 2c). In the Marianas, serpentinite mud volcanoes bring fragments of early Eocene blueschists formed at around 16 kb and 590 °C to the fore-arc surface indicating that such subduction systems do preserve products of their earliest history. Fore-arc magmatism occurs sporadically when spreading centers are subducted, as, for example, along the Aleutian and the Nankai trenches or where the fore-arc is

cracked by normal faulting perpendicular or subperpendicular to the arc axis, as between Mt. Hood and Portland, Oregon, in the Cascade arc in the northwestern United States. In extensional intraoceanic arcs, basalts to andesitic basalts to dacites, in places even rhyolites, although very sparse, are also observed in the Mariana, Izu-Bonin, and Tonga fore-arcs and these seem to be fed by the closest arc volcanoes along dykes and sills.

Above the fore-arc are usually located one or more fore-arc basins, such as the Great Valley in California, the Paita-Yunguilla Basin, and the older Celica-Lancones Basin in Ecuador, Pampa del Tamarugal in the Central Andes (Fig. 2g) or the Cretan Basin in the southern Aegean, the Andaman/Nicobar/Sunda fore-arc basins (Fig. 2f). Of the above examples the first group contains shortening, the second strike-slip and the third extensional basins. Such basins may in turn be of various subtypes, that is, shelved *vs.* sloped, terraced: ponded (filled *vs.* unfilled), imbricated; ridged: narrow, broad, and terrestrial upland types (Dickinson and Seeley 1979) depending on the predominant strain creating them, the topography and nature of the fore-arc region (continental *vs.* oceanic, or accretionary), the availability of sediment influx and climate. In the case of the intraoceanic ensimatic arcs, the fore-arc basement may consist entirely of oceanic lithosphere, commonly devoid of thick fore-arc basin fills (e.g., in the case of the Mariana fore-arc: Fig. 2c).

In many fore-arc systems, an upper slope discontinuity separates the fore-arc from the arc massif (Dickinson and Seeley 1979), forming the core of a subduction-controlled orogen. That discontinuity manifests itself as a break in slope determined by a fault that may be of any kind depending on the strain regime governing the evolution of the arc.

The arc massif and shortening retro-arc areas with flexural hinterland basins: In subduction-controlled orogens, the arc massif is the core of the orogen (Fig. 4a, b). It shows great variability in terms of rock composition, structure, and topography, ranging from (1) the dominantly mafic, extensional, low-lying island arcs, such as the Marianas (Fig. 2c), (2) through the deformationally neutral continental arcs of moderate topography characterized by dominantly andesitic volcanism, such as the Sumatran arc (Fig. 2f) (3) to the majestic mountains showing great across-orogen shortening, flanked by thrust belts and flexural basins on both sides, crowned by andesitic to rhyolitic volcanoes with subordinate basaltic volcanism and immense batholithic intrusions ranging from tonalitic through granodioritic to granitic compositions, such as the Andes (Fig. 2g, h). Within the arc cores, deformation is characterized by ductile deformation in levels where temperature exceeds about 580 to 600 °C and pressures reach and exceed 4 kb. Large, close to isoclinal folds (nappes of the first genre) dominate the structure at such levels (Fig. 4c), such as the Islay Anticline and the Tay Nappe formed from the Ardisaig Anticline in the Scottish Caledonides or the refolded isoclinal to

close fold nappes of the Thor-Odin Dome in the Canadian Cordillera. At higher structural levels, one sees typical fold-and-thrust belt geometries with little to no metamorphism, into which some intermediate to felsic plutons rise (e.g., in the Cordillera Huayhuash in Peru; in the Cariboo Mountains in the southern Canadian Cordillera). In the highest parts, in Altiplano-type high plateaux, brittle, nonpenetrative, germanotype structures take up orogen-parallel extension and orogen-normal shortening. Usually, the upper, less metamorphosed carapace is detached along thrusts and/or low-angle normal faults from the highly metamorphic core. In such orogens, large fold and thrust belts in the hinterland area are in turn flanked by wide flexural hinterland molasse basins of the type of the Llanos and Pampas of South America (Fig. 4a), or the vast Western Interior Basin behind the North American Rocky Mountains. The tectonic difference among the various types is dependent on whether the arc is extensional, neutral, or shortening. In all three types, arc-subparallel strike-slip faulting, both of arc-slicing and arc-shaving types, may be present depending on the angle of relative convergence of the subducting and the overlying plates. Especially the arc-shaving strike-slip faults give rise to significant coastwise transport of fore-arc material. Thus, not all fore-arcs have developed in front of the arc segments they now neighbor. Arc cores of especially Andean-type shortening orogens house not only high-grade metamorphic rocks, but also large, composite intrusions, such as the tonalitic/granodioritic/granitic Coast Mountains Batholith in Canada, Idaho, Sierra Nevada, and Peninsular Ranges batholiths in the US Cordillera, the immense Coastal Batholith of southernmost Ecuador and Peru, and the northern and southern Kuen-Lun plutons (e.g., Shahidulla) of felsic composition in northern Tibet, softening the orogenic edifice. This leads to underthrusting of the hinterland under the arc massif for considerable distances leading to the formation of the thin-skinned hinterland fold-and-thrust belts (Figs. 2g and 6a). This brings up the problem of the strain history in such orogenic cores: in the USA and Canadian Cordillera, it is widely believed that the core eventually collapsed under its own weight and created the extensional metamorphic core complexes (e.g., Armstrong 1982; Coney and Harms 1984). Yet, in the Andes, nowhere do we see any significant orogen perpendicular extension, despite the prodigious height of the orogen and a crustal thickness exceeding 75 km, especially in its central part. Moreover, Glombick et al. (2006) disputed the earlier interpretations of the Shuswap metamorphic complex: they wrote, “The shear zones, which yield upper plate to the west shear-sense indicators, are interpreted as the northern extension of the Okanagan Valley fault. Farther east, near the north-south axis of the complex, superstructure and infrastructure are separated by an attenuated metamorphic section, but evidence of noncoaxial strain is lacking. Discrete detachments were not found. Steeply dipping normal faults cut low-angle shear zones and do not merge with

them at depth. Middle Eocene volcanic and sedimentary rocks rest unconformably on metamorphic basement. The continuity of superstructure indicates that infrastructure was not exhumed by crustal-scale detachments.” (their p. 245). They propose that the late Cretaceous to early Tertiary partial melting of the middle crust resulted in the development of a zone of channel flow. As the channel was underthrust by a crustal-scale ramp in the underlying, more competent Palaeoproterozoic basement, it was exhumed from depths of 20–30 km and thinned vertically. Shear zones between infrastructure and superstructure are interpreted as being a transient rheological interface at the upper boundary of the channel, which in places may very well show a normal separation. Such an interpretation, also involving upper crustal stretching but without “gravitational collapse,” which seems very difficult without massive magmatic weakening of the crust, seems more compatible with world-wide observations. Where arc magmatism is feeble to nonexistent because of a very low angle of the subducting slab, shortening may disrupt the cratonic edge rimming the orogen in the hinterland area giving rise to large, thrust-bound, thick-skinned basement uplifts such as those in the US Rockies or the Sierras Pampeanas in northwest Argentina (Fig. 2h).

Back-arc: One or more extensional basins, which may or may not be floored by oceanic lithosphere, are located behind extensional arcs (Fig. 6c), such as those of the Sea of Japan, or the Parece Vela Basin and the Mariana Trough (Fig. 2c), or the Tyrrhenian Sea (Fig. 2d). Where oceanic, such back-arc, or marginal basins commonly have crusts thicker than normal oceanic crusts ranging from about 9 to 30 km. In continental back-arc basins, the crust may be thinned down to 6–8 km. Within the back-arc basins, volcanism is commonly, but not exclusively, tholeiitic basaltic and in places may be alkalic. Not all back arc basins are extensional. If a subduction zone develops far from a continent within an ocean, it may “trap” oceanic lithosphere behind the arc massif giving rise to a “marginal” basin. The Philippine and the Aleutian Basins are examples of such trapped marginal basins. In the case of such basins, the nature of the arc is immaterial for the formation of the basin.

Marginal basins do not often develop into major oceans, although examples of such development are well-known from the history of the Tethysides. Most seem to have life-spans of about 60 Ma and when starting to close back, fold and thrust belts may develop between the arc massif and the marginal basin behind it (e.g., as in the Sado and Okushiri Ridges along the western margin of Honshu and Hokkaido in the Japan Sea).

In case of neutral arcs, the epi-arc and hinterland areas have small basins commonly controlled by strike-slip tectonics parallel with the arc massif, such as the early Cainozoic coal and hydrocarbon basins behind the Sumatra arc axis, partly controlled by strike-slip-related secondary extension, especially in the North (Lho Sukon, Jawa, and Tamang deeps

bordering the Barisan mountain front) and Central (Balam, Kiri, Amon, Kampar Kanan, and Bengkalis grabens) basins.

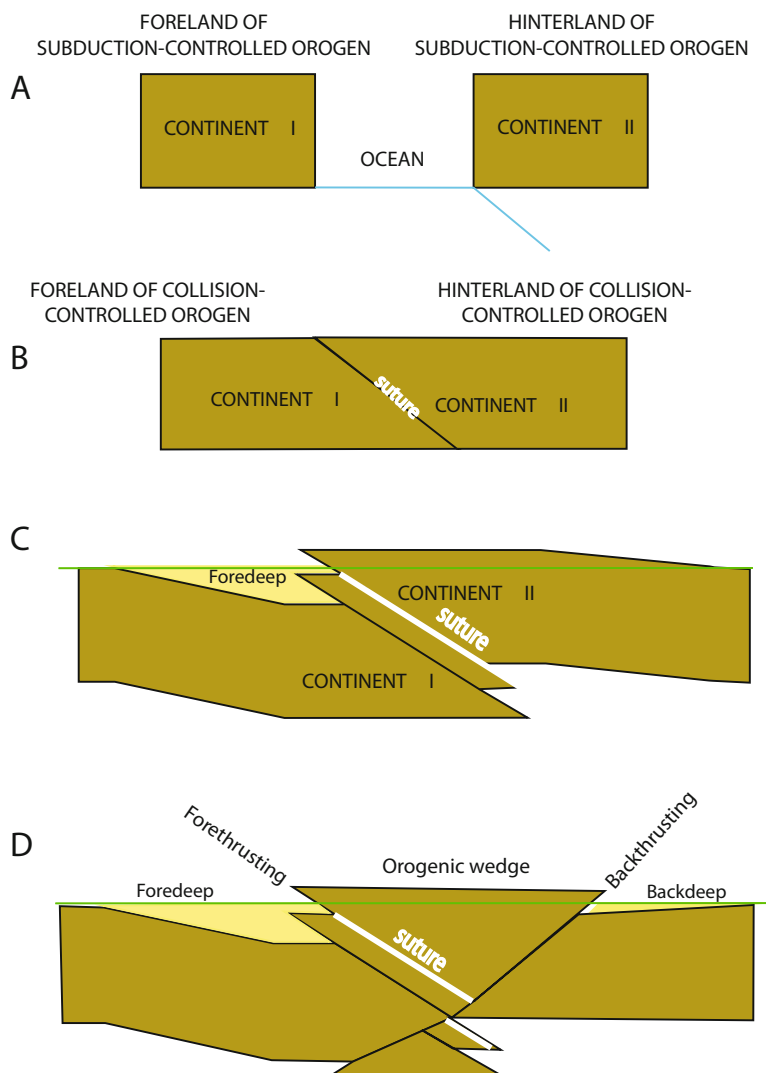
Collision-Controlled Orogens

In collisional systems (Fig. 3c), the asymmetry of structure is dictated by the foregoing subduction. Collision begins when a nonsubductable continent, carried by the subducting plate, is fed into the subduction zone. In such a situation, the fore-arc is usually obliterated (e.g., along the Indus Suture south of Lhasa; partial preservation farther west near Xigatze). In other cases, major parts of the fore-arc may be preserved in the form of overriding ophiolitic and *mélange* nappes (e.g., along the Intra-Pontide suture in Turkey or along the Himalayan margin). More extreme preservation of a fore-arc may occur in case of an extended Atlantic- or Californian-type continental margin, on which a magmatic arc had not developed to any extent, overrides the incoming continent (as the Austroalpine units did in the Alps). The incoming continent may be a major cratonic piece with its peripheral continental margins such as

India, or it may be just an ensialic island arc, such as the Rhodope-Pontide Fragment in northern Turkey, or even a well-developed ensimatic arc, as in Oman. *In collisional orogens, the incoming, that is, the underthrusting, continent is always the foreland* (e.g., India in the case of the Himalaya; Central Europe in the case of the Alps; Arabia in the case of the Zagros; Eastern Europe in the case of the Urals; Siberian Craton in the case of the Verkhoyansk Mountains).

Collision-controlled orogens are at least as multifarious as the subduction-controlled ones, but their basic architecture is the same in all (Fig. 7). In Fig. 7a, two continents are approaching each other by eliminating the intervening ocean by subduction under continent II. Notice that no details are given of the structure to keep the figure as general as possible. Figure 7b shows the situation once the collision is accomplished. Here it is implied that the foregoing subduction vergence also determines the orogenic vergence initially. As orogeny evolves, a foredeep develops because of the isostatic loading of the incoming Continent I by the developing orogen

Orogenic Belts, Fig. 7 A highly simplified and schematic series of cross-sections (a–d) illustrating the basic stages in a continent-continent collision



as seen in Fig. 7c. Finally as the orogenic edifice rises, gravitational potential makes it more efficient to develop backthrusts, in front of which this time a backdeep develops (Fig. 7d). When well-developed Cordilleran orogens collide with other continents, their already well-developed backthrust/backdeep areas become inherited by the collisional orogen, as, for instance, in the Transhimalaya Range. Here a terminological muddle needs clarification: in most geological publications dealing with the American Cordillera, the vast backdeeps are almost universally referred to as “foredeeps” and the cratons bordering the Pacific-facing orogen as “foreland.” This led to the development of such absurd and inconsistent terminology as “retro foreland basin,” meaning “back front basin.” The confusion has deep roots going back to James Dwight Dana in the middle of the nineteenth century, who considered all orogens to face continents away from bordering oceans. This was rectified by Suess in 1885 who correctly recognized the Pacific Ocean floor as the foreland of the entire Cordillera, but later, in 1909, he too became confused and mentioned both Laurentia and Brasilia (Guiana+Amazonian+São Francisco+Rio Apa+Rio de la Plata cratons) as forelands as a result of the discovery of the extensive thrusting in the Subandean chains that led him to consider the Andes as a bilaterally symmetric chain. However, plate tectonic theory restored the asymmetric structure of the Cordillera by recognizing the Andean subduction as the main motor of the orogeny there. Clarifying this confusion is important, because in many publications with titles involving the word “foreland” hinterlands and hinterland basins are also (in places, exclusively!) described.

There are two main types of collision-controlled orogens: override and nonoverride types (Fig. 8). The “override” refers to overriding of one of the colliding continents above the other one for considerable distances at shallow structural levels (~ 15 km; Fig. 8a). This happens when the maximum width of the closed ocean was not sufficient to generate a well-developed continental margin magmatic arc upon subduction (about 400–600 km), as, for instance, in the case of the Alps. In that case, two thinned continental crustal pieces collide and the overlying plate easily thrusts above the underthrusting one for very considerable distances (Şengör 1991b). In the nonoverride types, none of the colliding continents override the other one at shallow structural levels, because the intervening ocean was wide and therefore its subductive removal generated a well-developed continental margin arc with a thick crust prohibiting it from mounting the incoming continent owing to excessive gravitational energy needed (Fig. 8b). Himalaya is the best-known example of such an orogen. When the subduction zone of the arc has accumulated an excessively large subduction-accretion complex, as in Makran or Alaska or Sikhote Alin today, its thickening upon collision with the incoming continent may stop continental convergence and the two continents may never come into

direct contact (Fig. 8c). The Kuen-Lun and the Altaids are examples of such “Turkic-type” orogens.

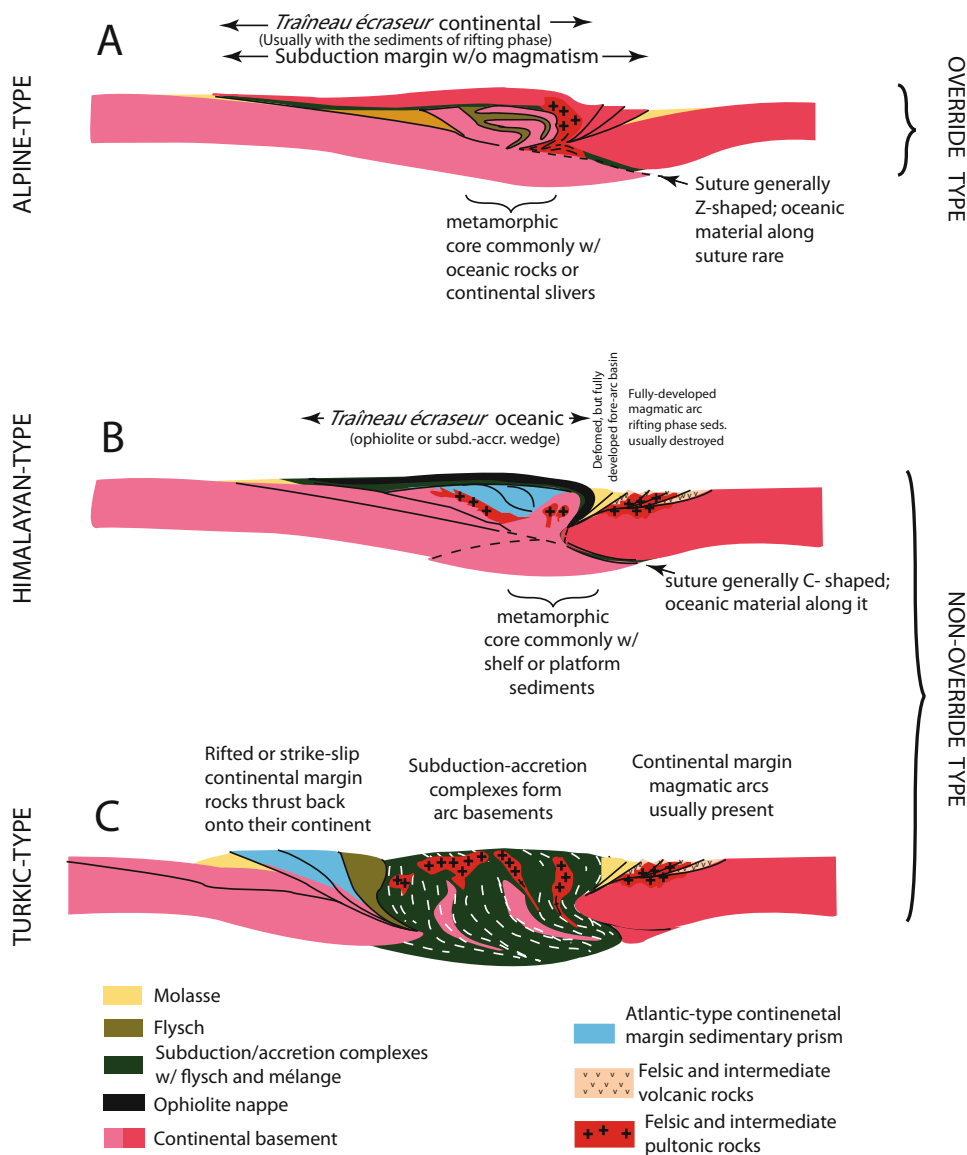
Despite the fairly simple outlines of orogenic architecture in both subduction- and collision-controlled orogens, the detailed structure of orogens on our globe is extremely varied depending on the nature of the interacting lithospheres and crusts (including their history), composition of the mantle underlying them, nature of the magmas that intrude them and extrude out of them, availability and nature of sediment, prevailing climate in where they develop, duration of their development, and the rate of plate interaction creating them. All these factors are in turn interdependent generating extremely complicated structures and evolutionary histories. In the next section, the various species of orogens are presented, the term *species* here taken in its original Latin meaning: “appearance, quality” with no genetic implications. However, the reader must bear in mind that species of orogens do not maintain their identity throughout their history, but, almost as a rule, evolve into each other. Therefore, the “species” presented below should only be taken as entities often existing only during a specific interval of the life of an orogen. Such metamorphoses might be compared with the metamorphoses of certain insects, but such a comparison would be misleading, because insects follow a predetermined, unchanging course (barring mutations) guided by genetics, whereas the evolution of orogens is entirely contingent on the factors listed above. Even within one species, orogens may display a great variety that are not uniform enough to be designated as “races.” They are varieties, but the number of varieties are as many as there are orogens in any one species. This may be compared to the variety seen in human beings. No two humans are identical, but they all have the same *Bauplan*. It is that *Bauplan* the species concept in tectonics unveils to us and enables us to employ a kind of “comparative anatomy of orogenic belts.”

Species of Orogens

Figure 9 illustrates a hierarchical classification of orogens. All orogens are products of convergent plate boundaries. Four types of convergence may create orogenic structures:

Transpressional orogens (Fig. 9i): For a transpressional structure to count as an orogen, it must involve subduction. Subductionless transpressional structures may be very major edifices, such as the Transverse Ranges in California or the Qinghai Nan Shan in northeastern Tibet (Fig. 2a), yet they are parts of the strike-slip regime, often parts of major keirogens, and not orogens.

Transpressional orogens may form along continental margins (Fig. 9A) or entirely within the ocean (Fig. 9B). Example for the former is the entire Sumatran subduction zone controlling a transpressional orogen above it that has given rise to strain partitioning by creating the Sumatran Fault. The Hikurangi Trench in the North Island of New Zealand carries

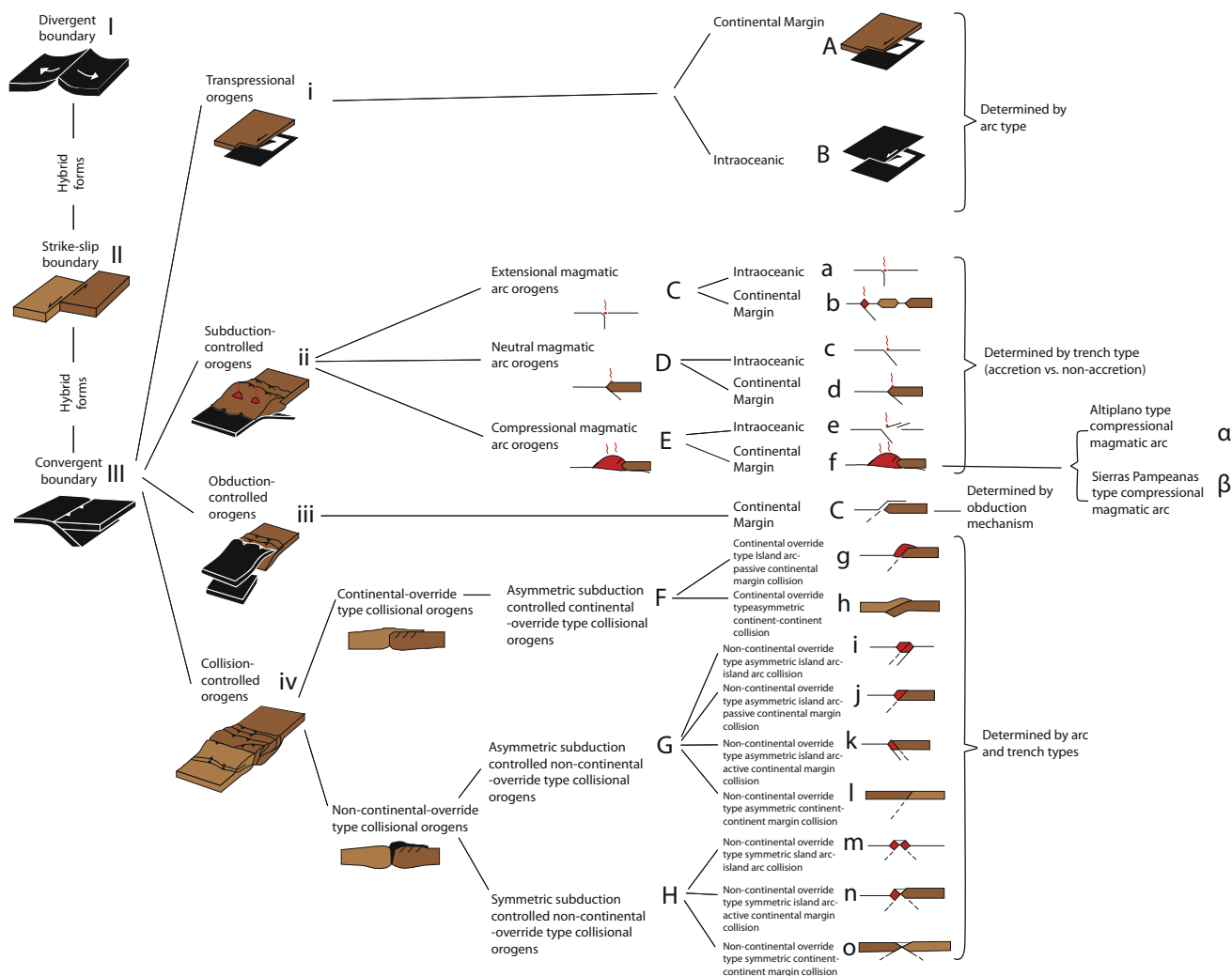
Orogenic Belts, Fig. 8 Styles of collisional orogens

TYPES OF COLLISIONAL OROGENIC BELTS

above it also a transpressional orogen with strain partitioning expressed by the Kekerengu Strike-Slip Fault. The Tonga-Kermadec subduction system, the northerly prolongation of the Hikurangi Trench, is also transpressional with strain partitioning creating a right-lateral strike-slip system within the arc and head-on subduction in the trench. The Aleutian subduction zone/magmatic arc system, connecting the Queen Charlotte Transform Fault and the Bering-Kresta Transform Fault, is partly a transpressional orogen, at least west of the Atka Island ($52^{\circ}08'17''$ N $174^{\circ}26'43''$ W). Similarly, the Florence Rise and the entire magmatic system behind it in Central Anatolia is a perfectly good transpressional orogen with a part

of the Eastern Mediterranean Ridge in front of it. The Macquarie transpressional system with increasing amounts of convergence to the south along the Hjort Trench is a purely intraoceanic example. Thus, transpressional orogens may be of various sorts depending on the type of crust involved, the degree of obliquity of the subduction along them, and, if present, the influence of the bounding strike-slip boundaries.

Subduction-controlled orogens (Fig. 9ii): subduction controlled orogens may be of various types depending on the nature of the interacting lithospheres. The simplest form is that of intraoceanic arcs formed by intraoceanic subduction zones (Fig. 9Ca). If the arc is extensional, a system of low-



Orogenic Belts, Fig. 9 A hierarchical classification of orogenic belts. I, II, and III are three end-member plate boundaries. Orogens originate only from convergent boundaries, that is, from III. (Modified from Şengör (1990a))

lying, dominantly mafic intraoceanic island arc (e.g., the Mariana arc) and various generations of marginal basins behind it (e.g., the Mariana Trough and the Parece Vela Basin; Fig. 2c) separated from one another by remnant arcs (e.g., West Mariana and Palau-Kyushu Ridges) constitute the orogen. The arc and the expanding youngest marginal basin are often the only active parts of the orogenic system.

The continental margin arcs are the most prominent subduction-controlled orogen species (Fig. 9Cb, Dd, Ef). The extensional varieties (Fig. 9Cb) create ensialic arcs torn from continents or from other ensialic arcs opening back-arc basins behind them. The Japanese islands with the Japan Sea marginal basin behind them are the best-studied examples of ensialic arcs torn from a continent. Arcs torn from continental arcs are best represented by the Inner Apennine massifs torn from Sardinia/Corsica, which in turn had been torn from continental Europe in Iberia opening behind them the Tyrrhenian and the Algero-Provençal Basins respectively (Fig. 2d).

If the arc is neutral (Fig. 9D), no marginal basin develops (unless one is trapped). An intraoceanic example (Fig. 9Dc) is the Central Aleutian island arc system between the islands of Unalaska and Atka (a length of about 500 km; the Aleutian Basin oceanic lithosphere behind it is a trapped piece of the Pacific Ocean floor). For neutral continental margin arcs (Fig. 9Dd), the Makran in Pakistan and Iran (Fig. 2e) is possibly the best continental example with the majestic arc volcanoes of Kuh-e Bazman, Kuh-e Taftan, and Kuh-e Sultan.

Shortening arcs (Fig. 9E) in intraoceanic settings (Fig. 9Ee) almost always results in subduction polarity flipping, as, for example, happened in the Vitiaz/New Hebrides Trench systems in the Southwest Pacific Ocean. Shortening intraoceanic arcs almost always come into being when a buoyant oceanic object, such as a seamount chain or an oceanic plateau, reaches the trench associated with the arc and makes subduction difficult. In the case of the Vitiaz

Trench, it was the arrival of the Samoan island chain of oceanic volcanoes that choked the trench and led to arc-perpendicular shortening. Shortening of continental margin arcs (Fig. 9Ef) commonly create majestic cordilleran chains with doubly vergent deformation on both sides of the arc massif housing large plutons such as those of North, Central, and South America (Fig. 2g, h) and the Sikhote-Alin Orogen in Northeast China and the Russian Far East.

Collision-controlled orogens (iii, iv): There are two fundamental types of collision-controlled orogens as mentioned above. The continental override types result from the collision of arcs or two major continents across the closure of small oceans of the Alpine type (Figs. 8a and 9ivF). The closure of the Antalya Ocean in southern Turkey is an example of the first (e.g., Okay and Özgül 1984) and the closure of the Alpine Tethys (e.g., Pfiffner 2014) is the example of the second one. Inevitably the arcs associated with such orogens are either very poorly-developed, as in the Alps, or not developed at all, reducing the orogeny to the collision of a small continental sliver with another arc. In fact, all Alpidic orogens west of the Carpathians belong to one or the other of the override type collisional orogens (Wildi 1983; Pfiffner 2014). All major ophiolite obductions actually belong to this category (Figs. 2i and 9iiiC), because almost all represent collision with a continental margin of an ensimatic arc or a piece of oceanic lithosphere generated above a subduction zone by prearc spreading (Şengör 1990a).

The noncontinental override type orogenic belts are basically of three kinds (Fig. 8b, c): arc-arc collisions (Fig. 9Gi) of the kind seen in Japan where the Izu-Bonin arc is ramming the Honsu arc and creating the syntaxis of the Fossa Magna, arc-continent collisions (Fig. 9Gj), and continent-continent collisions (Fig. 9Gl). However, notice that all island arc-island arc collisions must be in reality the collision of two subduction-controlled orogens and should not be considered under fundamental orogenic types. Similarly, the collision of an island arc with a continental margin arc, facing each other (Figs. 2m, n and 9Hn) or facing in the same direction (Fig. 9Gk), similar to the collision of the Panama Arc with the northern Andes also represents the collision of two orogens and should similarly fall out of the list of the fundamental types of orogens. By contrast the collision of an island arc with an Atlantic- or California-type continental margin creates a single, continent-facing orogen (Fig. 9Gj). An island-arc facing away from a nearby continent would not be able to generate a collisional orogen, because the arc would not be able to approach the continent. Ensialic arc-continent collisions with an arc facing the continent are well-known: in northern New Guinea, in Kamchatka, in western Turkey, west of Ankara (Fig. 2l), in Southeast Asia, in Taiwan and the Luzon Arc in the northern Philippines. As pointed out above, all supra-subduction zone ophiolite obduction events and the ensuing orogenic belts result

from ensimatic arc-continent collisions; more developed ensimatic arcs that collided with Atlantic-type continental margins are also known, for example, the Kohistan Arc in the western Himalaya. In arc-continent collisions, there are cases where the geological evolution makes the slotting of an orogen into a category impossible: in the southern Andes, the northern end of the Rocas Verdes Basin closed without any subduction, but by squashing and folding the Sarmiento ophiolite without much of a dominant vergence. In this case would one consider the structure as an orogen resulting from an arc-continent collision, or simply a continental margin arc orogen? Towards the south, the Rocas Verdes Basin became wider and here an arc-continent collision is more obvious with a clear northerly vergence of the Tortuga ophiolite and that is why the closure of the Rocas Verdes Basin is considered an orogenic event by itself. However, the problem becomes much more acute in case of even narrower intra-arc basins floored by oceanic lithosphere, such as those behind the New Hebrides arc. Subduction in such a case is essentially impossible and, consequently, the closure of such basins must be considered as an intra-arc event and not an independent orogeny.

Orogenic Collages

The major mountain belts of our globe, even the youngest and still active ones, rarely consist of a single orogen. For example, the Caledonian/Appalachian System has many orogens of diverse types within it and the end result is what James Helwig in 1974 called an “orogenic collage.” The reason that the correct recognition of the constituents of orogenic collages is important is to be able to dissect the orogenic complex formed from them correctly and thus reconstruct its evolution accurately. Failure to do this would result in subdividing an orogenic edifice on the basis of inappropriate criteria resulting in the proliferation of nameless “tectonic units” that serve no purpose in understanding geological processes and the history of orogenic complexes. The rise of “terranoology” in the North American Cordillera is one example of such a sterile methodology (see Şengör 1990b). Fig. 2m, n illustrates further examples of “symmetric orogenic complexes” formed from more than one orogen, but often, erroneously, called “an orogen” (e.g., the Caledonian “Orogen” or the Appalachian “Orogen”).

Magmatism in Orogens

Magmatism in orogens occurs in two main settings: subduction magmatism and collision magmatism. Although fore- and hinterland magmatism may be caused by orogeny (e.g., the rifting-related magmatism of the Upper Rhine Rift being caused by the Alpine collision to the south), it is not considered a part of the orogenic phenomena.

Subduction Magmatism

Subduction magmatism shows a very great variety on the basis mainly of the nature of the arc (extensional, neutral, or shortening, with an ensimatic or ensialic arc massif), subduction dip angle, and the nature and amount of the sediment being dragged into the subduction zone which are functions both of the nature, and the amount of sedimentary fill, of the trench and the prevailing climate.

Subduction magmatism is generated by the release of volatiles (mainly water and CO₂) from subducting slabs and their percolation upwards in the mantle wedge above. Subduction generates a pressure minimum at the apex of the wedge where much of the generated high-Mg mafic magma resulting from 2 to 8% melting is gathered. From here the magma escapes directly upward, most likely by creating its own fractures and not along the so-called “megathrust” of the subduction zone because of the low temperature there. This magma creates the magmatic arc axis and the magmatic front. “Hydrous phases in the subducted slab decompose at levels shallower than 100 km, that is, beneath the fore-arc region. The slab-derived H₂O, enriched in incompatible elements with larger ionic radii, reacts with the overlying mantle wedge materials to form polluted amphibole peridotite. The polluted peridotite is transported downward on the slab. . . . Amphibole in the dragged peridotite decomposes at a depth of about 110 km, just beneath the volcanic front. H₂O released migrates upward to reform amphibole peridotite in the higher-temperature and lower-pressure region. When the front of amphibolization reaches a level at which the solidus temperature of amphibole peridotite is distributed, initial magmas are produced and rise as a form of mantle diapir.” (Tatsumi 1986, p. 717). The interesting thing is that the base of the oceanic lithosphere, commonly at about 100 km, coincides with the peridotite solidus. Thus, there seems to exist a double control of the position of the magmatic front in subduction-controlled orogens. Rise of magmas from initial partial melting to the surface takes about 30 Ma (Turner et al. 2004). The magmatic front has been used as a *file-d’Ariane* to reconstruct now disrupted subduction-controlled orogens (Şengör et al. 1993). The 30 Ma gap mentioned above shows that when magmatic fronts are used for reconstruction purposes, time intervals >30 Ma must be considered (ideally 50–60 Ma).

Sediments coming into the subduction zone are in part dragged down into the magma generation zone, but not much beyond. In sediment starved subduction zones, the generated magmas are mainly basaltic, being island arc tholeiites with subordinate dacites and even rhyolites resulting from local differentiation. As arc magmatism matures, more and more calc-alkalic magmas begin replacing tholeiitic magmatism. In continental margin arcs andesites dominate where the arc is neutral and andesites and rhyolites dominate where the arc is actively shortening. In Sumatra, for example, where the arc is neutral, magmatism begins with andesites and

basalts with subordinate felsic rocks, goes through a phase of explosive volcanism with andesites and rhyolites, and then returns to an andesitic basaltic phase. During the Plio-Pleistocene, basaltic effusions cease almost entirely and the arc magmatism is dominated by andesites and rhyolites with rare basalts.

Where the arc is actively shortening felsic magmatism dominates together with andesites, as in the Andes, Central and North American Cordilleras; in the Transhimalayan Arc before the Himalayan collision; in the Lachlan Orogenic Belt in eastern Australia. In the hinterland parts of the magmatic arcs in shortening orogens, continental crustal melting becomes significant and tin mineralization is widespread. Some of the Na-rich trondjemite/tonalite/ignimbrite magmatism in the Andes of Peru may result from the partial melting of underplated basaltic lower crust and probably did not come either from the subducted slab or from the mantle wedge. The generation of trondjemite/tonalite complexes during the Mio-Pliocene has important implications for the origin of their Archaean analogues. Low-angle subduction may lead to slab melting and adakitic arc magmatism, as for example, in the Yanshan Belt of north China, where it occurs alongside basement-involved shortening.

The morphology of volcanoes is controlled by the composition of their products and represents an important factor in governing the topography of the magmatic arc and the resulting efficiency of subaerial erosion and concomitant sedimentation. The table (e.g., Columbia River Flood Basalt Province) and shield (Marianas) volcanoes have very subdued relief, with few of the volcanoes rising above sea-level as in the case of the Marianas or the Tonga-Kermadec arc. In such arcs, sediment production is slow and consequently the trenches remain empty and accretionary complexes grow very slowly or not at all. By contrast, where volcanoes are andesitic, steep-sided stratovolcanoes dominate the arc creating a sharp, but regular relief with much erosion and sedimentation. The associated trenches are usually full and give rise to substantial subduction-accretion complexes as seen in Japan, Sumatra, in eastern Alaska, the Cascade subduction zone in the northwestern USA, Central America or in Makran. In shortening arcs, crustal shortening is mainly responsible for the generation of high ground, but there also, the stratovolcanoes and rhyolite domes and ignimbritic blankets contribute to the creation of the rough and jagged topography that facilitates erosion of the high-standing edifices and the transportation of the produced sediment. It is to be remembered that the highest point above the Earth’s center is not Mt. Everest, but the summit of Mount Chimborazo in Ecuador, an Andean stratovolcano located one degree south of the Equator, where the Earth’s rotational bulge is greatest.

Magmatism in extended back-arc basins is basaltic, generally with significant subduction contribution, but may be even alkalic as in the case of the Utsuryoto Island in

southwestern Japan Sea. The thickness of the oceanic crust in them is more variable than that in the major oceans, reaching in places thicknesses of 30 km, not dissimilar to oceanic plateaux (Bransfield Strait: 30 km; Algero-Provençal Basin: >20 km; Yamato Basin of the Japan Sea: 14–16 km; Japan Basin of the Japan Sea; Banda Sea: 9–10 km) but in others showing normal oceanic thicknesses (e.g., Shikoku Basin: 5–7 km; Kuril Basin: 5–7 km; Tyrrhenian Basin: 6 to 8 km); yet others displaying significant variation between 5.5 and 9 km within the same structure as in the Lau Basin and from 20 to 28 km in the Black Sea basins. Not all back arc basin floors are homogeneously oceanic: the Japan Sea, for example, is underlain mostly by continental crustal patches of various thickness and only a triangular area between northernmost Honshu, Hokkaido, and southernmost Sikhote Alin is truly oceanic. The same thing is true for the Black Sea where the Arkhangelsky and the Andrusov ridges, forming the Mid-Black Sea Ridge, separate the two oceanic halves of the basin. This has important implications in interpreting the geology of the now-closed back-arc basins encountered in orogenic belts. For example, an extreme case is illustrated by the evolution of the Tasman Sea, which is the best analogue for the Mesozoic/Cainozoic evolution of the Mediterranean Apides in Turkey and surrounding regions.

Collision Magmatism

Magmatism does not cease in an orogenic belt after continental collision, but lasts until all orogenic activity comes to a close. Along collision belts two types of collisional magmatism dominate: (1) Slab break-off magmatism and (2) crustal thickening magmatism. In addition to these, the formation of important strike-slip related extensional basins within collisional orogens may localize basaltic volcanism in places accompanied by shallow crustal melts (as, for example, in the Erzincan pull-apart basin along the North Anatolian Fault in Turkey) and so can crustal fissuring at high angles to the orogenic convergence direction (e.g., the Nemrud Volcano in eastern Turkey). The most spectacular present-day example of a slab break-off magmatism is seen in the actively shortening Turkish-Iranian High Plateau. Over half of the Turkish-Iranian High Plateau is covered with collision-related young volcanic rocks (~43,000 km²), ranging in age from 11.4 Ma to the seventeenth century. Although it is difficult to calculate the total volume of the volcanic material produced because of variations in the thickness of the volcanic succession and erosion in the region, the estimated total volume is a minimum of 15,000 km³. The erupted volume may, however, represent only a small fraction of the melt generated beneath the region, because a greater proportion presumably was emplaced deeper in the crust in the form of plutons. Although fissure eruptions dominated the volcanic activity (with some N-S fissures as in the case of Nemrud), there are over 20 major volcanic centers and numerous smaller ones on the plateau,

corresponding to central eruption sites. The volcanic rocks of the plateau span, remarkably, the whole compositional range from basalts to rhyolites. There is, however, a significant north-south variation in lava chemistry between the north and the south parts of the plateau. The volcanic units of the northern areas are calc-alkalic, whereas those of the south are mildly alkalic to alkalic in character (even phonolites are known). Slab break-off-related magmatic massifs, housing large porphyry copper/gold deposits, with attendant tectonism has been studied in detail in New Guinea. The magmatic rocks are mostly intermediate in composition resembling arc rocks in their major element content, but also contain those giving “intraplate” signatures. There are also basalts and lamproites. Where this magmatism is west of the Tasman Line, it involves signatures of the Precambrian crust, to the west, it only has younger influences as expected from the composition of the crust through which they erupted. The mainly felsic Eocene magmatic rocks in northwestern Turkey were similarly results of a slab break-off following the closure of the İzmir-Ankara suture south of the Sakarya continent.

In Himalaya and Tibet, the collisional magmatism appears to be related to the thickening of the crust, much in the way originally suggested by Dewey and Burke (1973). In the Himalaya, the “suture-crossing” collisional magmatic rocks are represented by true leucogranites, the younger members of which (27 to 16 Ma “hair granites”) have facilitated the extensional collapse of the mountain front (Burchfiel et al. 1992). Whether the loss of the mantle lithosphere under Tibet actually happened is still controversial from the viewpoint of the available seismic data and thus difficult to consider in its rôle in generating magmas under the high plateau. However, the Oligocene to Recent vulcanicity on the Plateau north of the Indus-Yarlung suture began with highly undersaturated lavas in Central Tibet in the Oligocene. Typical calc-alkalic volcanism was active in the southern part of the plateau during the Miocene and highly incompatible element enriched ultrapotassic vulcanicity has characterized the Pliocene to recent activity in northwestern Tibet. This is not thought to be compatible with a wholesale underthrusting of the Indian craton under Tibet. The “suture-crossing” post-collisional tin-bearing granites of Cornwall, which extend westward into the Isles of Scilly and the submarine outcrop of Haig Fras, are considered similar in origin to the Himalayan granites, in that they too occur as multiple intrusive sheets into nappes on the foreland side of the Hercynian suture. High-K calc-alkalic postcollisional granites are also known from the Neoproterozoic Pan-African orogenic system, where in places they are associated with extensive strike-slip faulting, much like in northern Tibet today.

Rift-related alkalic and carbonatitic magmatism that in places accompanies the opening of oceans the closure of which gives rise to orogeny is usually not considered a part of orogenic magmatism, except in one case. Burke et al.

(2003) noticed that 90% of nepheline syenite gneisses and deformed carbonatites occur along Proterozoic suture zones in ancient collisional orogens. They infer from this observation that deformed alkalic rocks and carbonatites (DARCs) mark the places where vanished oceans have opened and then closed. They further infer that DARCs subducted down to ~100 km depths during collisions, as indicated by diamond-grade ultra-high pressure metamorphism in some collisional orogens, could provide source material for later alkalic magmatism and that this could account for the observation of recurrent alkalic magmatic activity over hundreds of million years in old orogenic provinces.

Metamorphism in Orogens

Metamorphism in orogens is related to magmatic events in arc cores and to tectonic burial in both magmatic arcs and collisional orogenic belts (Fig. 6). Holland and Lambert (1969) presented a synthesis of structural regimes corresponding to a given metamorphic grade. Their first regime is that seen below the level of diagenesis to the depth where metamorphic reactions begin. Here creep has a logarithmic relation to time, but other factors such as fluid pressure greatly influence the actual strains generated. The second regime is that between the onset of metamorphism and the first appearance of the epidote-amphibolite facies, where similar folds accompanied by schistosity and recumbent folds on all scales form. The third regime is that accompanying the epidote-amphibolite and amphibolite facies, in which deformation occurs by grain boundary or intralattice diffusion following the β creep law ($\epsilon = \beta t^n$; $n < 1$). Regime four is where a higher pressure amphibolite facies dominates and is characterized by partial melting, formation of gneissose structure, extensive flow folding, and possible onset of diapirism. The final regime is that of the granulites where laminar flow seems to dominate. The Holland and Lambert classes naturally apply to both subduction and collision metamorphism. Folding and thrusting under metamorphic conditions also fold and repeat the isotherms resulting in very complex metamorphic rock/structure relationships such as “inverted metamorphism.”

Below orogenic metamorphism is considered under the two headings of subduction- and collision-related metamorphism.

Subduction-Related Metamorphism

Before the advent of the theory of plate tectonics, Miyashiro (1961) pointed out that in active orogens around the Pacific Ocean a metamorphic belt of the andalusite/sillimanite and/or low-pressure intermediate type lies always next to the one characterized by jadeite-glaucophane and/or high-pressure intermediate type and the latter is always found closer to the ocean. In 1973, Miyashiro pointed out that the high-pressure

member of the pair is probably formed below a trench and the low-pressure member formed beneath a volcanic chain in the adjacent island arc or continental magmatic arc. He further observed that the low and intermediate pressure belt is associated with granitic/andesitic/rhyolitic magmatism, whereas the high-pressure belt is usually accompanied by tectonic slices of mafic and ultramafic rocks.

Our understanding of subduction-related metamorphism has not changed in principle since Miyashiro's perspicacious papers. Figure 6 shows the highly schematized isotherms in arcs and the corresponding mineral facies. The corresponding pressures may be obtained by looking up the metamorphic facies expressed by the index minerals and rocks in the figure. Figure 6a illustrates a shortening magmatic arc orogen. The normal geothermal gradients there (Fig. 6a) are disturbed both by the adiabatically rising plutons and the insertion of the cold oceanic lithosphere into the trench on the ocean, that is, foreland side and the insertion of the cold continental lithosphere in the hinterland. The insertion of the continental lithosphere on the hinterland side greatly augments the crustal thickness in the orogenic core and introduces into it radioactive minerals that generate heat.

Figure 6b is the situation of the isotherms in a neutral magmatic arc orogen and the corresponding rock-types. The difference between the neutral and the shortening arc orogens is that the former has no contribution of extra heat from the hinterland and, consequently, its magmatic activity is less vigorous. Figure 6c shows the situation in an intraoceanic arc where there is no preexisting continental crust and the only source of substantial heat comes from the magmatic activity of the arc and a slightly higher geothermal gradient commensurate with that of the oceanic crust. This higher geothermal gradient is offset by the introduction of the cold subducted slab into the subduction zone beneath the arc. Consequently, the intraoceanic arcs have the most subdued metamorphic zonation of all the arc types. The arc metamorphism is episodic in many arc terrains, which is a consequence of the episodic melt segregation and intrusion of plutons and the temperature sensitive rheology of the upper crust controlling the strength of the lithosphere.

Because subduction introduces cold material to very considerable depths (in places in excess of 100 km), a high pressure-low temperature metamorphism, also called “blueschist metamorphism” since the appellation used by E. H. Bailey in 1964 for the glaucophane schists in California, develops. Beyond about 39 km-depth coesite forms and near 100 km and beyond diamond crystallizes in deeply subducted rocks. The ultra-high pressure rocks are no longer low temperature-high pressure rocks, because their formation occurs near 17–20 kb and between 600 and 850 °C; they are simply “very high grade.” A particularly thorny problem has been to understand how such very deeply buried rocks return to the surface.

Collision-Related Metamorphism

Collision-related metamorphism is basically burial metamorphism, essentially Barrovian. This burial occurs by crustal thickening, most commonly by thrusting at the upper crustal levels and bulk shortening by folding and foliation development and even rock flowage in deeper levels as explained by Holland and Lambert (1969). Figure 10 shows its basic principles. In Fig. 10a two continents collide, whereby Continent I forms the foreland. The arrows in Fig. 10b indicate how the rocks in the descending continent will heat up as they are depressed below the normal continental isotherms, whereby heating is assumed to be instantaneous for graphic simplicity. In Fig. 10b, the orogenic welt rises isostatically bringing with it the heated, that is, metamorphosed rocks. The situation depicted in Fig. 10b is an idealized one, before any backthrusting begins. In reality, backthrusting would begin during the isostatic rise, or even earlier, as in the Alps, for example. Figure 10c depicts the situation after the advent of backthrusting and the development of an orogenic root. This would further raise the orogenic welt and the bottom of the root may begin partially melting creating the postcollisional granites such as those of Bergell and Novate in the Alps. As the orogenic welt rises, it gets denuded both tectonically and by attrition through the exogenic agencies and the updomed metamorphic isograds successively expose the mostly low pressure high temperature metamorphic facies within the core of the orogen, as, for example, is well-known in the Alps, where metamorphic rocks from prehnite-pumpellyite facies in the marginal thrust belts (Helvetic/Dauphine units in the external zones and the Southern Alps in the internalmost zones) are present as well as the granulite facies rocks within the orogenic core (e.g., the Ivrea Zone). The scenario shown in Fig. 10 is naturally overly simplified as it assumes a two-dimensional evolution where transition from one state to the other is instantaneous and the deformation is entirely brittle. In reality, convergence across a collisional orogen is rarely purely orthogonal to the collision front, the rate at which both the shortening proceeds and the rocks deform are very variable ranging from essentially zero in the western Alps to some 2 mm/a in the eastern Alps to 6 cm/a in the Himalaya and the structures range from highly ductile flow folds (where temperatures are about 600 °C or higher) to brittle structures as outlined by Holland and Lambert (1969). For example, in many collisional orogens, large-scale thrusting inverts the metamorphic isograds, as for example in the Himalaya or diapirism very much steepens, even overturns them and the loss of the lithospheric mantle beneath an orogen or the rise of a pluton may in places create even Buchan-type metamorphism (e.g., during the Grampian collision in the Scottish Caledonides or in the south Portuguese Hercynides).

In some orogenic belts, the foreland continent is dragged down to the subduction zone to depths of about 100 km to

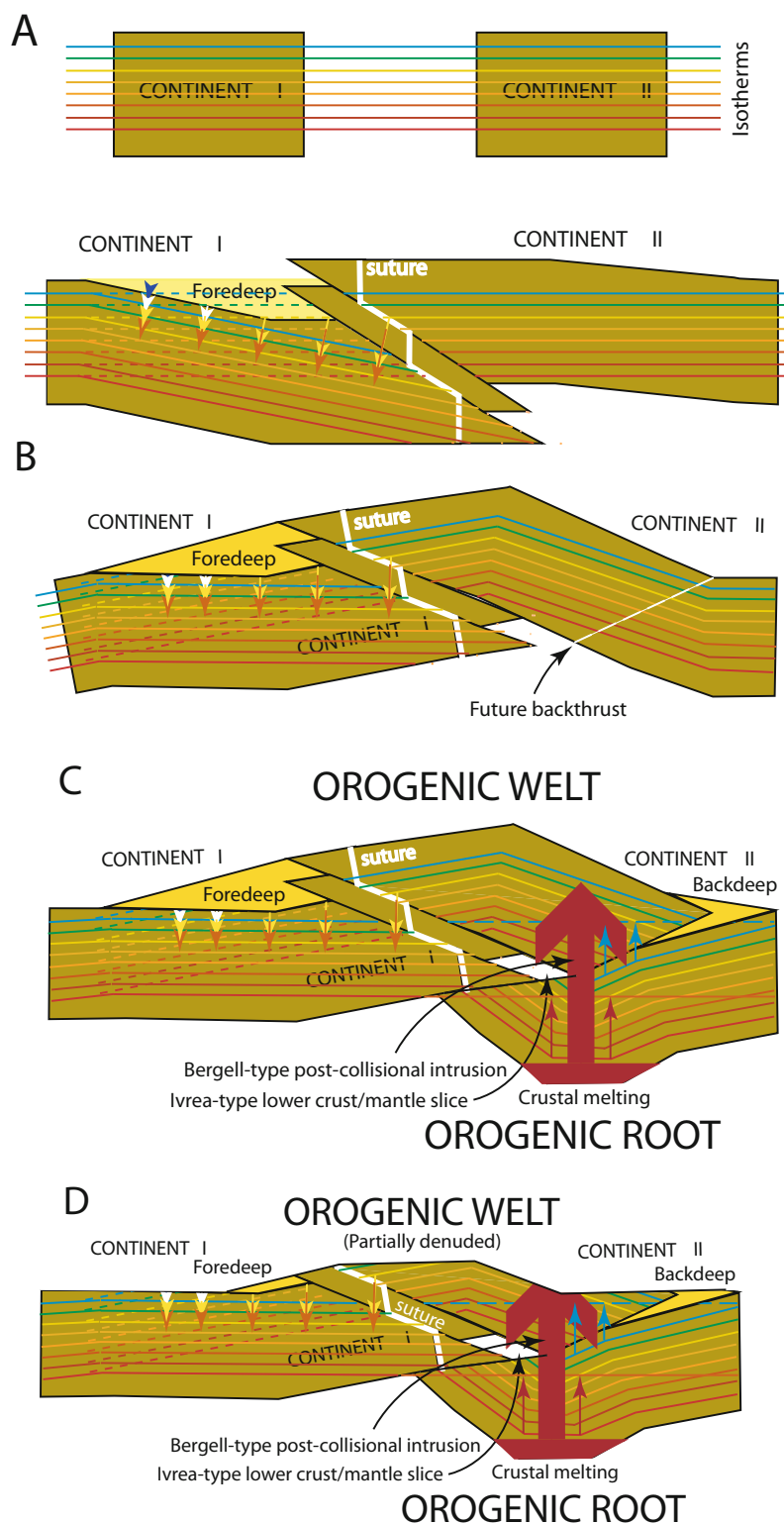
create ultra-high pressure metamorphism, but it is important to realize that this metamorphism is *not* a product of the collision, but of the subduction that happens to outlive for a certain time the collision.

Geomorphology of Orogenic Belts

The sculpting of mountain belts is a complex process governed by internal processes that generally tend to increase topographic differentiation (except normal-fault-governed tectonic denudation decreases it) and external processes that diminish relief. For many years, it was believed that mountain erosion was very rapid and that mountains would disappear in a few million years, if not repeatedly elevated by ongoing internal processes. In the latter half of the twentieth century, it was realized that the denudation of a mountain range by attrition resulting from atmospheric and hydrological processes was critically dependent on isostasy and sea-level. Pitman and Golovchenko (1991) studied quantitatively the morphological evolution of mountain belts in the light of large (>200 m), long-period (>50 ma) sea-level changes. Based on the available data, they assumed that the rate of denudation of such old landscapes is proportional to the average regional elevation by a constant ($K \approx 10^{-4}/1000$ years). Thus, if the sea-level is approximately stable, it will take more than 300 m.y. and more likely more than 450 m.y. to degrade a Himalayan size mountain belt to a peneplain (in its original Davisian sense). For an old mountain belt that has been eroded to an average elevation of 500 m, it will take an additional 220 m.y. and more likely 390 m.y. to reduce to a peneplain. In a system in which the river valleys are graded with respect to base level, the rate of downcutting is enhanced by the rate of uplift caused by the isostatic response to regional denudation. With rising base level, the rate of valley incision decreases by the rate of rise. However, transgression of the river valleys cannot take place until the rate of river valley uplift caused by isostatic response to the denudation becomes equal to the rate of sea level rise. Under usual, that is, nonglacial, conditions of sea level rise, this will not occur until the landscape has been degraded to an average elevation of 60 m or less. However, under these conditions the decrease in the rate of downcutting in the river valleys greatly enhances the rate of degradation. As a consequence, older mountain belts already eroded to a regional elevation of 500 m or less may be further reduced to a peneplain during an episode of persistent sea level rise at rates of several tenths of a centimeter per 1000 years for several tens of millions of years. With sea level falling, the landscapes evolve toward an equilibrium surface at which point the rate of degradation is equal to the rate of sea level fall. Under such conditions, peneplains cannot form. The

Orogenic Belts,

Fig. 10 Evolution of metamorphism and collisional magmatism in an asymmetric collisional orogen



conditions for a steady-state of an orogen are difficult to realize because of the mobility of the interacting agencies such as the changes in the base level. Faulting is probably the dominant internal process that determines the geomorphology of the mountain ranges.

Economic Deposits in Orogenic Belts

Hydrocarbons

Hydrocarbons in orogenic belts occur in some fore-arc (e.g., in the Great Valley Forearc Basin in California) and hinterland

basins (e.g., in the Alberta Basin in the Canadian Rockies) in subduction-controlled orogens and in the fore- and hinterland-basins of the collisional orogens. In fact, the earliest uses of petroleum are known from three foreland basins (Zagros foreland: third millennium; BCE Sichuan Basin in China, which is also a hinterland basin with respect to the Tibetan Alpides, fourth century CE; 1859: Appalachian foreland basin in Pennsylvania). Some pull-apart basins within the orogen in places produce hydrocarbons (e.g., the Vienna Basin). Although some intramontane ramp-valley basins, such as those in eastern Turkey, have reported petroleum potential, no economic deposits have yet been discovered in them. The fore- and hinterland deformation areas of both subduction-controlled and collision-controlled orogens generate germanotype basins that contain important petroleum reserves (e.g., the Denver Basin in the US Rockies; Karamay Field in the Junggar Basin in NW China).

Coal

Coal mainly occurs in the fore- and hinterland basins of orogenic belts; but, if such deposits are later involved in a younger orogeny, some anthracite deposits may be encountered within the orogen itself (e.g., the ancient Heraclea coal basin in Zonguldak, NW Turkey).

Metals

It has long been known that almost all metals occur in orogenic belts, because they sweep together so many different environments and expose so many levels within the crust and upper mantle.

Building Materials

The world's most beautiful and famous, decorative building stones are obtained from orogenic belts, commonly from their metamorphic cores. All rocks from orogens in one way or another contribute to the building stone industry in the whole world, even the molasse basins (in fact, the word "molasse" is a Swiss stonemason's term).

Summary

An orogenic belt is a lithospheric structure that is formed by the collective work of any group of convergent plate margin processes.

Orogenic belts are major features of the lithosphere on which 98% of the world's largest earthquakes occur. Since the diameter of the Earth does not change to any appreciable degree, ocean-floor spreading and subduction must balance each other. Thus, the rate of deformation along orogenic belts must be almost 1.5 times higher than along the spreading centers, which is corroborated by the concentration of the largest earthquakes along them making them the most active

parts of the lithosphere. Orogenic belts are the main factories of the continental crust, at least since after the Hadean. Orogens house the world's main mineral wealth and an understanding of their nature is critical for successful mineral exploration. The basins flanking them and genetically associated with them contain a significant portion of the world's hydrocarbon resources. Knowing how they evolve is an indispensable part of oil and gas exploration on this planet. Finally, orogenic belts create mountains which control a good portion of the world's climate and climate in turn influences the growth of mountain belts.

Bibliography

- Armstrong R I (1982) Cordilleran metamorphic core complexes – from Arizona to Southern Canada. *Annu Rev Earth Planet Sci* 10:129–154
- Burchfiel BC, Chen Z, Hodges KV, Liu Y-P, Royden LH, Deng C-R, Xu J (1992) The South Tibet detachment system, Himalayan orogen: extension contemporaneous with and parallel to shortening in a collisional mountain belt. *Geol Soc Am Spec Paper* 269:1–41
- Burke K, Ashwal LD, Webb SJ (2003) New way to map old sutures using deformed alkaline rocks and carbonatites. *Geology* 31:391–394
- Coney PJ, Harms TA (1984) Cordilleran metamorphic core complexes: Cenozoic extensional relics of Mesozoic compression. *Geology* 12:550–554
- Dewey JF (1977) Suture zone complexities: a review. *Tectonophysics* 40:54–67
- Dewey JF (1980) Episodicity, sequence and style at convergent plate boundaries. *Geol Assoc Can Spec Pap* 20:553–573
- Dewey JF (1988) Extensional collapse of orogens. *Tectonics* 7:1123–1139
- Dewey JF (2002) Transtension in arcs and orogens. *Int Geol Rev* 44:402–439
- Dewey JF (2005) Orogeny can be very short. *Proc Natl Acad Sci* 102(43):15286–15293
- Dewey JF, Burke K (1973) Tibetan, Variscan and Precambrian basement reactivation: products of continental collision. *J Geol* 81:683–692
- Dickinson WR, Seeley DR (1979) Structure and stratigraphy of forearc regions. *Am Assoc Pet Geol Bull* 63:1–31
- Glombick P, Thompson RI, Erdmer P, Daughtry KL (2006) A reappraisal of the tectonic significance of early Tertiary low-angle shear zones exposed in the Vernon map area (82 L), Shuswap metamorphic complex, southeastern Canadian Cordillera. *Can J Earth Sci* 43:245–268
- Holland JG, Lambert RSJ (1969) Structural regimes and metamorphic facies. *Tectonophysics* 7:197–217
- Johnson MRW, Harley SL (eds) (2012) *Orogenesis – the making of mountains*. Cambridge University Press, Cambridge, x+388 pp
- Kober L (1921) *Der Bau der Erde*. Gebrüder Borntraeger, Berlin, [II]+324 pp.+1 foldout map
- Miyashiro A (1961) Evolution of metamorphic belts. *J Petrol* 2:277–311
- Miyashiro A, Aki K, Şengör AMC (1982) *Orogeny*. Wiley, Chichester, 242 pp
- Molnar P (1986) The structure of mountain ranges. *Sci Am* 254(7):70–80
- Okay AI, Özgül N (1984) HP/LT metamorphism and the structure of the Alanya Massif, Southern Turkey: an allochthonous composite tectonic sheet: in Dixon, J. E. and Robertson, A. H F., editors, *The Geological Evolution of the Eastern Mediterranean*, Geological Society Special Publication no. 17 London, pp. 429–439

- Pfiffner OA (2014) *Geology of the Alps*: Blackwell, Wiley, Oxford, Chichester and Hoboken, X+376 pp
- Pitman WC III, Golovchenko X (1991) The effect of sea level changes on the morphology of mountain belts. *J Geophys Res* 96:6879–6891
- Rodgers J (1987) Chains of basement uplifts within cratons marginal to orogenic belts. *Am J Sci* 287:661–692
- Rodgers J (1990) Fold and thrust belt in sedimentary rocks. Part 1: typical examples. *Am J Sci* 290:321–359
- Rodgers J (1991) Fold and thrust belt in sedimentary rocks. Part 2: other examples, especially variants. *Am J Sci* 291:825–886
- Rodgers J (1995) Lines of basement uplifts within the external parts of orogenic belts. *Am J Sci* 295:455–487
- Rodgers J (1997) Exotic nappes in external parts of orogenic belts. *Am J Sci* 297:174–219
- Şengör AMC (1990a) Plate tectonics and orogenic research after 25 years: a Tethyan perspective. *Earth-Sci Rev* 27:1–201
- Şengör AMC (1990b) Lithotectonic terranes and the plate tectonic theory of orogeny: a critique of the principles of terrane analysis. In: Wiley TJ, Howell DG, Wong FL (eds) *Terrane analysis of China and the Pacific Rim*: Houston, Circum-Pacific Council for Energy and Mineral Resources, Earth science series, vol 13. pp 9–44
- Şengör AMC (1991a) Timing of orogenic events: a persistent geological controversy. In: Müller DW, McKenzie JA, Weissert H (eds) *Modern controversies in geology*. Proceedings of the Hsü symposium. Academic, London, pp 405–473
- Şengör AMC (1991b) Orogenic architecture as a guide to size of ocean lost in collisional mountain belts. *Bull Tech Univ Istanbul (Ketin Festschrift)* 44:43–74
- Şengör AMC (2006) Orogenic style, sea-level, Sr isotopes and a brief history of the Earth for the last 600 million years. *Russ Geol Geophys* 47:26–34. (N.L. Dobretsov volume)
- Şengör AMC, Natal'in BA, Burtman VS (1993) Evolution of the Altaid tectonic collage and Palaeozoic crustal growth in Eurasia. *Nature* 364:299–307
- Smith AG (1976) Plate tectonics and orogeny: a review. *Tectonophysics* 33:215–285
- Tatsumi Y (1986) Formation of the volcanic front in subduction zones. *Geophys Res Lett* 13:717–720
- Turner S, Black S, Berlo K (2004) ^{210}Pb – ^{226}Ra and ^{228}Ra – ^{232}Th systematics in young arc lavas: implications for magma degassing and ascent rates. *Earth Planet Sci Lett* 227:1–16
- Wildi W (1983) La chaîne tello-rifaine (Algérie, Maroc, Tunisie): structure, stratigraphie et évolution du Trias au Miocène, *Revue de Géologie Dynamique et de Géographie Physique*, Vol. 24, p. 201–297, Paris
- Woodcock NH (1986) The role of strike-slip fault systems at plate boundaries. *Philos Trans R Soc Lond A* 317:13–29

Further Reading

<https://eies.itu.edu.tr/arastirma/%C3%B6nemli-makaleler>
<https://eies.itu.edu.tr/en/research-groups/important-papers>

Paleomagnetic Field Intensity

Andrew Biggin¹, Greig A. Paterson¹, Neil Suttie² and John Shaw¹

¹Geomagnetism Lab, Department of Earth, Ocean and Ecological Sciences, University of Liverpool, Liverpool, UK

²School of Archaeological and Forensic Sciences, University of Bradford, Bradford, UK

Synonyms

Archaeointensity; Palaeointensity

Definition

The strength of the ancient geomagnetic field and its variability through time can be deduced from various different types of measurements. *Absolute palaeointensities* are calibrated using a known laboratory magnetic field and can be recovered from any rocks retaining a thermoremanent magnetization yielding a palaeomagnetic measurement of the geomagnetic field intensity recorded at the time and place that the rock cooled down from above the Curie temperature of the magnetic grains. Similarly, *archaeointensity* measurements are made from heated archaeological artifacts (e.g., fired pottery). By contrast, *relative palaeointensities* provide uncalibrated records of geomagnetic intensity variations measured from the natural remanent magnetization of sediments, and *cosmogenic isotope palaeointensities* achieve the same using concentrations of cosmogenic isotopes as a proxy.

Introduction

Direct measurements of the intensity of the geomagnetic field only became possible in 1832 when a method of obtaining them

was developed by Carl Frederich Gauss (Stern 2002). If we wish to know how the field intensity has varied before this time, then we are required to use the geological and archaeological records to make indirect measurements. Palaeomagnetic approaches are based on the principle that, for weak fields such as Earth's, the intensity (M_{anc}) of the natural remanent magnetization (NRM) recorded by a material is linearly proportional to the strength (H_{anc}) of the palaeofield. For absolute palaeointensity measurements, the material in question must (except in very rare cases; e.g., Games 1980) have acquired a thermoremanent magnetization (TRM) whereas for relative palaeointensity, it is required to be a depositional remanent magnetization (DRM). Unfortunately, in both cases, the constant of proportionality linking the NRM to the palaeofield intensity is a sensitive and complicated function of the magnetic properties of the recording material and extremely difficult to measure directly. Absolute palaeointensity experiments obviate this difficulty by imparting the specimen with a new TRM (M_{lab}) in the lab by heating the specimen up and cooling it in a controlled field of known intensity (H_{lab}). The linearity relationships of the two types of M and H can then be combined to eliminate the constant and solve for the ancient field intensity:

$$H_{anc} = \frac{M_{anc}}{M_{lab}} H_{lab} \quad (1)$$

Depositional-type remanences cannot be reproduced reliably in the laboratory and therefore palaeointensity experiments performed on sediments produce records only of relative variations in H_{anc} .

In order for any type of palaeointensity measurement to be useful, an independent estimate for the age of the magnetization (often assumed or demonstrated to be the same age as the material) is usually required.

The first attempts to measure the ancient field intensity were made by Folgheraiter (1899) using archaeological vases, but with no reported success. Later, Koenigsberger (1936, 1938) and then Thellier (1941) made more serious attempts to recover past geomagnetic field strengths using igneous

rocks, baked archaeological materials and a sedimentary contact rock, respectively. It was somewhat later that the idea of Koenigsberger (1936) to obtain palaeointensity measurements from much older igneous rocks by comparing their natural thermoremanence to a laboratory-imparted field was properly tested (Wilson 1961). Absolute palaeointensity measurements from igneous rocks of all ages began in earnest in the mid- to late-1960s (e.g., Briden 1966; Coe 1967), and the discipline has been subject to numerous attempts to improve the methodology since (Table 1). Palaeointensity

measurements of all types are typically curiosity-driven, although the desire to understand geomagnetic polarity reversal transitions and geodynamo behavior more generally has provided a clear focus for some studies (Valet 2003).

Absolute Palaeointensity Measurements

Absolute palaeointensity experiments may only be performed using materials that have cooled from elevated temperatures

Paleomagnetic Field Intensity, Table 1 Selection of absolute palaeointensity methods currently in use

Family	Variants	Benefits	Criticisms
Thellier (Thellier and Thellier 1959)	Coe (1967); Aitken-Walton (Aitken et al. 1988; Walton et al. 1979); Perpendicular (Kono and Ueno 1977); IZZI (Tauxe and Staudigel 2004)	Sound theoretical basis for single domain grains; potential to recover estimate from a sample that contains both secondary magnetizations at low temperatures and that alters at high temperatures; additional checks for alteration (Prévot et al. 1981) and multidomain behavior (Krasa et al. 2003; Riisager et al. 2001) can easily be added	Multiple heatings are time-consuming and increase likelihood of sample alteration; bias due to multidomain effects may be prevalent in lava samples (Biggin 2010)
Microwave (Hill and Shaw 2000; Shaw et al. 1996)	All Thellier-type variants may be used	In principle, all benefits of Thellier, but with reduced sample heating in terms of both temperature and time; faster experiments that can be individually tailored to samples	Requires highly specialist equipment (microwave palaeointensity system); theoretical basis not yet firmly established; TRMs may not thermally unblock as sharply and as completely in microwave as in thermal experiments
Shaw (1974)	ARM-alteration corrections (Kono 1978; Rolph and Shaw 1985); Double heating (Tsunakawa and Shaw 1994); low temperature demagnetization (Yamamoto et al. 2003)	Original method is fast; arguably domain-state independent as uses full TRMs	The use of ARMs to check and correct for alteration in terms of the TRM properties is not always valid (Tanaka and Komuro 2009)
Continuous high temperature measurement	Wilson (1961); Le Goff and Gallet (2004)	Very fast methods; Wilson method is arguably domain-state independent as it uses full TRMs; Le Goff–Gallet method incorporates cooling rate and anisotropy corrections	Requires specialist equipment (high temperature magnetometer); samples must survive heating to Curie temperature without altering in Wilson method; unknown effect of multidomain grains in Le Goff–Gallet method
Multispecimen (Hoffman et al. 1989)	Hoffman and Biggin (2005); Dekkers and Böhnelt (2006); Fabian and Leonhardt (2010)	Fast; <i>inter</i> -specimen consistency is implicitly considered; domain state effects can be minimized in Dekkers–Böhnelt experiments; fewer heating steps implies less alteration	Fewer checks for <i>intra</i> -specimen consistency; no internal checks for alteration; domain-state benefits of Dekkers–Böhnelt method may be lost if secondary magnetizations require removing (Michalk et al. 2010); original Dekkers–Böhnelt method may not be domain state independent, but can be corrected (Fabian and Leonhardt 2010)
Calibrated pseudo-Thellier	Stephenson and Collinson (1974); de Groot et al. (2013); Paterson et al. (2016)	No heating required; fast, especially on automated systems	ARM/TRM ratio is grain size dependent, which manifests as a ~ 25% uncertainty on the calibration factor (Paterson et al. 2016)
IRM Normalization (Cisowski and Fuller 1986)	REM (Kletetschka et al. 2000); REM (Gattacceca and Rochette 2004)	No heating required	Precision limited to orders of magnitude
First-order reversal curve (FORC)	Muxworthy and Heslop (2011); Muxworthy et al. (2011); Di Chiara et al. (2017)	No heating required; moderately fast	Limited testing; considered accurate to a factor of 2–3

in nature and which therefore possess a TRM imparted by the ambient (geomagnetic) field. Materials that may be used include any fired archaeological materials (ceramics, kiln lining, hearths, burnt bricks, etc.) and most types of igneous rocks.

Producing unequivocally accurate measurements of the absolute palaeointensity is extremely demanding and requires at least the following:

1. The remanence carriers must retain a pure TRM and have undergone no chemical or physical changes since cooling beneath their Curie temperature (see e.g., Fabian 2009).
2. Any secondary components of magnetization (e.g., viscous or isothermal overprints, see ► “[Remanent Magnetism](#)”) must be removed so that the primary TRM is isolated.
3. The remanence carriers must undergo no chemical or physical alteration as a consequence of the laboratory heating that it is required to obtain the palaeointensity (or this must be corrected for; see e.g., Valet et al. 1996).
4. It must be shown that multidomain, vortex state, or magnetically interacting grains are not biasing the measurement by violation of Thellier’s laws of thermoremanence (see ► “[Remanent Magnetism](#)”).
5. Any anisotropy of TRM must be avoided or corrected for (see e.g., Veitch et al. 1984).
6. Any effect that the difference in the cooling rate between the natural and laboratory cooling cycles has on the magnitude on the calculated palaeointensity must be corrected for (see e.g., Perrin 1998).
7. Any deviation from Thellier’s law of linearity must be detected and corrected for (Selkin et al. 2007).
8. The number (N) of individual sample estimates from a single rock unit or archaeological horizon must be *sufficient* to obtain a reasonable mean (μ) and to demonstrate *adequate* internal consistency (see e.g., Paterson et al. 2010).

The high failure rates and time-consuming nature of many absolute palaeointensity studies have stimulated a wealth of innovation in terms of both the types of materials that are used and the manner in which the experiments are performed. In the former respect, two noteworthy materials are submarine basaltic glass (SBG) generally taken from ocean drilling cores and ophiolites (Tauxe 2006), and single silicate crystals isolated from both extrusive and intrusive rocks (Tarduno et al. 2006). Both of these have been argued to be higher fidelity recorders (less prone to problems numbered 1, 4, and/or 6 in the list above) than more conventional palaeointensity recorders (e.g., whole-rock subaerial lava samples). As an alternative approach to favoring any one specific recorder, the Q_{PI} criteria (Biggin and Paterson 2014) were developed as a universal qualitative measure of palaeointensity reliability. These attempt to emulate the success of a similar tool (Van der

Voo 1990) long-applied to palaeomagnetic poles in deriving a set of qualitative “check boxes” for minimizing the risk of bias associated with any of the problems listed above.

In terms of the techniques developed to measure the absolute palaeointensity, Table 1 summarizes some of those which are more widely used and provides the relevant references. The Thellier method and its derivatives remain the most widely used and trusted technique; for more information on this and other techniques listed in Table 1, the reader is directed to comprehensive reviews by Tauxe and Yamazaki (2015) and Valet (2003). Accompanying these techniques is a broad range of statistics that are used as selection criteria to screen less reliable results. For a comprehensive description of these statistics and comparisons of the effectiveness of various sets of selection criteria, see Paterson et al. (2014, 2015).

Relative Palaeointensity Measurements

Palaeomagnetic analyses of sedimentary rock and soft sediment samples can produce continuous time series of geomagnetic intensity variations which may be calibrated using absolute palaeointensity estimates or used alone as records of geomagnetic behavior. Relative palaeointensity studies are generally undertaken on the assumption that the samples retain a DRM acquired as the magnetic grains in the sediment physically align themselves during or shortly after deposition from the water column. The DRM acquisition process is sensitively related to the size, shape, and composition of the remanence carriers and is further complicated by flocculation effects in saline environments. Much later, the intensity of the magnetization can be significantly affected by consolidation and compaction processes (see review of Tauxe and Yamazaki 2015).

Some normalization of the measured magnetization intensity is required to account for changes in the concentration and nature of the magnetic recorders (which should ideally be small) and produce a record of geomagnetic intensity variations. Candidates for normalization include ARM, IRM, and low-field susceptibility. These should ideally display a strong coherence with the NRM signal (after removal of any secondary components of magnetization), but no coherence with the normalized record (Valet 2003). The reliability of a relative palaeointensity record can be further supported by the presence of antipodal directions (indicating no inclination flattening) and by the coherence of multiple curves from the same region covering over the same time interval (as calibrated using independently obtained timescales).

The sedimentation rate and the interval over which the magnetization is “locked-in” obviously set a limit on the time resolution of any relative palaeointensity record. In deep-marine sediments, a standard 2.5-cm palaeomagnetic

sample can represent a time period of 10 kyr or more, which will produce significant smoothing of secular variation (see ► “[Geomagnetic Field, Secular Variation](#)”). Roberts et al. (2013) provide a detailed discussion of DRM processes, relative paleointensity determinations, and normalization techniques.

Cosmogenic Isotope Palaeointensity Measurements

This method of palaeointensity measurement, applicable to the last few hundreds of kyr, is based on the fact that production rate of certain radioisotopes (^{14}C , ^{36}Cl , and ^{10}Be) is strongly influenced by the degree of shielding the geomagnetic field provides the Earth from incoming cosmic rays. Records of the concentration of these isotopes in sediment and ice cores (and tree rings and corals in the case of ^{14}C) can therefore provide a proxy for geomagnetic dipole moment once the necessary corrections for latitude and relevant physical, chemical, and biological processes have been made. See review by Valet (2003) for more information.

Geophysical Applications

Palaeointensity measurements provide information about the geodynamo across a variety of time scales. Since direct observations of field strength were first made in the 1830s, the main axial dipole field has fallen by around 10%. The rate and duration of this dipole decay is a focus of many archaeointensity studies, but is still debated. Using VADM values derived directly from archaeointensity estimates, Poletti et al. (2018) suggest that the dipole intensity has been decreasing since ~750 CE – over a millennium of weakening. Recent studies based on global field models (e.g., Arneitz et al. 2019; Hellio and Gillet 2018), however, argue that the axial dipole has been weakening only for the last few centuries with variable trends prior to this.

Archaeointensity and absolute palaeointensity data are particularly important for constructing global field models, which now extend to 100 ka (Panovska et al. 2018). Each of the observable palaeomagnetic field components (declination, inclination, and intensity) can be mapped, via their respective Green’s functions, to the radial field at the core mantle boundary (Johnson and Constable 1997). Measurements of field intensity yield different information than directional data about the core mantle boundary (particularly at higher latitudes; Johnson and McFadden 2015).

Relative palaeointensity records from varved sediments are capable of providing good temporal resolution over the Holocene and agree well with absolute intensity records as well as records of cosmogenic nuclide production (Snowball and Sandgren 2002). Correspondingly, absolute archaeointensities

have been combined with the ^{14}C record from tree rings and the ^{10}Be record from ice-cores to infer variation in solar activity, with implications for understanding solar forcing of climate (Solanki et al. 2004). Marine sediments can provide relative palaeointensity records over longer timescales. Here it is usually assumed that a typical sample will represent a sufficiently long period for non-dipole features to be averaged out (Valet 2003) allowing composite stacks of relative palaeointensity from different areas of the world to be constructed. A global stack for the past 2 Myr, Sint 2000 (Valet et al. 2005), agrees well with absolute palaeointensity from lava flows and confirms a low field strength during excursions. This has been incorporated into a continuous model of axial dipole moment variation for the past 2 Ma (Ziegler et al. 2011). Importantly, this record is long enough to include reversals and finds a correlation between polarity interval length and field strength. This same correlation is seen, albeit weakly, in the 11 Myr Oligocene record from site Deep Sea Drilling Project, site 522 (Constable et al. 1998) and in SBG absolute intensity records (Tauxe 2006).

It becomes increasingly difficult to obtain relative palaeointensities from sediments more than a few million years old so longer term variations in field strength are studied using absolute palaeointensity derived from igneous rocks. The emerging picture characterizes the time-averaged dipole moment over the last few hundreds of millions of years as substantially lower than its current instantaneous value, but subject to dramatic variations on timescales ranging from centennial or shorter to at least hundreds of millions of years (Biggin et al. 2012; Tauxe and Yamazaki 2015).

Prévot et al. (1990) found evidence of a long period of low field intensity which they called the *Mesozoic dipole low*. However, subsequent work (Kulakov et al. 2019; Tarduno et al. 2006; Tauxe and Staudigel 2004) has indicated that this did not extend into the Cretaceous normal superchron (CNS; 84–121 Ma) and its age of onset remains unclear (Anwar et al. 2016). Other notable lows have recently been identified in the mid-Palaeozoic (Hawkins et al. 2019) and Ediacaran (Bono et al. 2019). The timing of these lows, on the order of 200 million years apart, together with a similar temporal variability seen in Phanerozoic polarity reversal frequency (the dipole lows being coincident with inferred intervals of reversal hyperactivity) gives rise to the hypothesis that dipole moment is quasi-periodic on these long timescales. The likely candidate for the source of such variations is changes in the total heat flowing across the core-mantle boundary and/or its spatial pattern, which might themselves be linked to substantial fluctuations on similar timescales in the global flux of slab material subducted into the mantle (Hounslow et al. 2018).

Claims of even longer, billion-year variations in dipole moment linked to the secular evolution of the core (including inner core nucleation) and mantle remain controversial (Biggin et al. 2015; Bono et al. 2019; Smirnov et al. 2016; Tarduno et al. 2015). Similarly, the magnitude and

timing of the expected variations under different core thermal evolution scenarios is not yet clear (Driscoll 2016; Landeau et al. 2017). Vigorous activity in this potentially fruitful research area is anticipated for years to come.

Palaeointensity experiments are not confined to terrestrial rocks, with numerous studies on lunar, Martian, and meteorite materials. Lunar regolith from the Apollo missions is thought to hold a thermoremanence and a variety of palaeointensity techniques suggest a strong dynamo (20–100 μT) operating between ~ 4.2 and ~ 3.6 Gyr (Cournède et al. 2012; Suavet et al. 2013), prior to a possible prolonged period of weak field (~ 5 μT) that extended to at least ~ 2.5 Ga (Tikoo et al. 2017). Martian meteorites have been used to estimate the ancient Martian field (Shaw et al. 2001; Weiss et al. 2008b), and paleointensity studies on pallasite meteorites indicate that small (~ 200 km) planetesimals are also capable of sustaining dynamo-generated magnetic fields with field strengths of ~ 30 – 130 μT (Bryson et al. 2015; Tarduno et al. 2012). Some of the most ancient meteorites, formed at the onset of our solar system, are also capable of remembering magnetic fields in the solar nebula, yielding insight into solar system evolution (Wang et al. 2017; Weiss et al. 2008a).

Summary

From 450-year-old pots to 4.5 billion-year-old meteorites, studies on the strength of Earth's ancient magnetic field continue to flourish, bringing with them new innovations in methods, materials, and analyses. And as paleointensity studies grow, we are gaining new understanding on the evolution of the magnetosphere in near-Earth space, the planet beneath our feet, and the solar system beyond.

Cross-References

- [Archaeomagnetism](#)
- [Curie Temperature](#)
- [Geomagnetic Excursions](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Geomagnetic Field, Secular Variation](#)
- [Paleomagnetism, Measurement Techniques and Instrumentation](#)
- [Remanent Magnetism](#)

Bibliography

- Aitken MJ, Allsop AL, Bussell GD, Winter MB (1988) Determination of the intensity of the Earth's magnetic field during archeological times: reliability of the Thellier technique. *Rev Geophys* 26:3–12
- Anwar T, Hawkins L, Kravchinsky VA, Biggin AJ, Pavlov VE (2016) Microwave paleointensities indicate a low paleomagnetic dipole moment at the Permo-Triassic boundary. *Phys Earth Planet Inter* 260:62–73
- Ameitz P, Egli R, Leonhardt R, Fabian K (2019) A Bayesian iterative geomagnetic model with universal data input: self-consistent spherical harmonic evolution for the geomagnetic field over the last 4000 years. *Phys Earth Planet Inter* 290:57–75
- Biggin AJ (2010) Are systematic differences between thermal and microwave Thellier-type palaeointensity estimates a consequence of multidomain bias in the thermal results? *Phys Earth Planet Inter* 180:16–40
- Biggin AJ, Paterson GA (2014) A new set of qualitative reliability criteria to aid inferences on palaeomagnetic dipole moment variations through geological time. *Front Earth Sci* 2(24):21–29
- Biggin AJ, Steinberger B, Aubert J, Suttie N, Holme R, Torsvik TH, van der Meer DG, van Hinsbergen DJJ (2012) Possible links between long-term geomagnetic variations and whole-mantle convection processes. *Nat Geosci* 5:526–533
- Biggin AJ, Piispa EJ, Pesonen LJ, Holme R, Paterson GA, Veikkolainen T, Tauxe L (2015) Palaeomagnetic field intensity variations suggest Mesoproterozoic inner-core nucleation. *Nature* 526:245–248
- Bono RK, Tarduno JA, Nimmo F, Cottrell RD (2019) Young inner core inferred from Ediacaran ultra-low geomagnetic field intensity. *Nat Geosci* 12:143–147
- Briden JC (1966) Variation of intensity of the palaeomagnetic field through geological time. *Nature* 212:246–247
- Bryson JFJ, Nichols CIO, Herrero-Albillos J, Kronast F, Kasama T, Alimadadi H, van der Laan G, Nimmo F, Harrison RJ (2015) Long-lived magnetism from solidification-driven convection on the pallasite parent body. *Nature* 517:472–475
- Cisowski S, Fuller M (1986) Lunar paleointensities via the IRM (s) normalization method and the early magnetic history of the moon. In: Hartmann W, Phillips R, Taylor G (eds) *The origin of the moon. Lunar and Planetary Science Institute, Houston*, pp 411–421
- Coe RS (1967) Palaeointensities of the Earth's magnetic field determined from tertiary and quaternary rocks. *J Geophys Res-Solid Earth* 72:3247–3262
- Constable CG, Tauxe L, Parker RL (1998) Analysis of 11 Myr of geomagnetic intensity variation. *J Geophys Res B: Solid Earth* 103:17735–17748
- Cournède C, Gattacceca J, Rochette P (2012) Magnetic study of large Apollo samples: possible evidence for an ancient centered dipolar field on the Moon. *Earth Planet Sci Lett* 331–332:31–42
- de Groot LV, Biggin AJ, Dekkers MJ, Langereis CG, Herrero-Bervera E (2013) Rapid regional perturbations to the recent global geomagnetic decay revealed by a new Hawaiian record. *Nat Commun* 4:Art 2727
- Dekkers MJ, Böhm H (2006) Reliable absolute palaeointensities independent of magnetic domain state. *Earth Planet Sci Lett* 248:508–517
- Di Chiara A, Muxworthy A, Trindade RIF, Bispo-Santos F (2017) Paleoproterozoic geomagnetic field strength from the Avanavero mafic sills, Amazonian Craton, Brazil. *Geochem Geophys Geosyst* 18:3891–3903
- Driscoll P (2016) Simulating 2 Ga of geodynamo history. *Geophys Res Lett* 43:5680–5687
- Fabian K (2009) Thermochemical remanence acquisition in single-domain particle ensembles: a case for possible overestimation of the geomagnetic paleointensity. *Geochem Geophys Geosyst* 10. <https://doi.org/10.1029/2009GC002420>
- Fabian K, Leonhardt R (2010) Multiple-specimen absolute paleointensity determination: an optimal protocol including pTRM normalization, domain-state correction, and alteration test. *Earth Planet Sci Lett* 297:84–94
- Folgeraiter G (1899) Sur les variations séculaires de l'inclinaison magnétique dans l'antiquité. *J Phys Theor Appl* 8:660–667. <https://doi.org/10.1051/jphysap:018990080066001>

- Games KP (1980) The magnitude of the archaeomagnetic field in Egypt between 3000 and 0 BC. *Geophys J Roy Astronom Soc* 63:45–56
- Gattacceca J, Rochette P (2004) Toward a robust normalized magnetic paleointensity method applied to meteorites. *Earth Planet Sci Lett* 227:377–393
- Hawkins L, Anwar T, Shcherbakova VV, Biggin AJ, Kravchinsky VA, Shatsillo A, Pavlov V (2019) An exceptionally weak Devonian geomagnetic field recorded by the Viluy Traps, Siberia. *Earth Planet Sci Lett* 506:134–145
- Hellio G, Gillet N (2018) Time-correlation-based regression of the geomagnetic field from archeological and sediment records. *Geophys J Int* 214:1585–1607
- Hill MJ, Shaw J (2000) Magnetic field intensity study of the 1960 Kilauea lava flow, Hawaii, using the microwave palaeointensity technique. *Geophys J Int* 142:487–504
- Hoffman KA, Biggin AJ (2005) A rapid multiple-sample approach to the determination of absolute paleointensity. *J Geophys Res-Solid Earth* 110. <https://doi.org/10.1029/2005JB003646>
- Hoffman KA, Constantine VL, Morse DL (1989) Determination of absolute palaeointensity using a multi-specimen procedure. *Nature* 339:295–297
- Hounslow M, Domeier M, Biggin AJ (2018) Subduction flux modulates the geomagnetic polarity reversal rate. *Tectonophysics* 742–743:34–49
- Johnson C, Constable CG (1997) The time-averaged geomagnetic field: global and regional biases for 0–5 Ma. *Geophys J Int* 131:643–666
- Johnson C, Mcfadden PL (2015) Time-averaged field and paleosecular variation, chapter 11. In: Kono M (ed) *Treatise on geophysics*. Elsevier, New York
- Kletetschka G, Wasilewski PJ, Taylor PT (2000) Mineralogy of the sources for magnetic anomalies on Mars. *Meteorit Planet Sci* 35:895–899
- Koenigsberger J (1936) Die abhaengigkeit der natuerlichen remanenten magnetisierung bei eruptivgesteinen von deren alter und zusammensetzung. *Beitr Angew Geophys* 5:193–246
- Koenigsberger JG (1938) Natural residual magnetism of eruptive rocks (part 2). *Terres Magn Atmos Electr* 43:119–130
- Kono M (1978) Reliability of paleointensity methods using alternating-field demagnetization and anhysteretic remanence. *Geophys J Roy Astronom Soc* 54:241–261
- Kono M, Ueno N (1977) Palaeointensity determination by a modified Thellier method. *Phys Earth Planet Inter* 13:305–314
- Krasa D, Heunemann C, Leonhardt R, Petersen N (2003) Experimental procedure to detect multidomain remanence during Thellier-Thellier experiments. *Phys Chem Earth* 28:681–687
- Kulakov EV, Sprain CJ, Dubrovine PV, Smirnov AV, Paterson GA, Hawkins L, Fairchild L, Biggin AJ (2019) Analysis of an updated paleointensity database (QPI-PINT) for 65–200 Ma: implications for the long-term history of dipole moment through the Mesozoic. *J Geophys Res*. <https://doi.org/10.1029/2018JB017287>
- Landeau M, Aubert J, Olson P (2017) The signature of inner-core nucleation on the geodynamo. *Earth Planet Sci Lett* 465:193–204
- Le Goff M, Gallet Y (2004) A new three-axis vibrating sample magnetometer for continuous high-temperature magnetization measurements: applications to paleo- and archeo-intensity determinations. *Earth Planet Sci Lett* 229:31–43
- Michalk DM, Biggin A, Knudsen MF, Böhnell HN, Nowaczyk N, Ownby S, Lopez-Martinez M (2010) Application of the multi-specimen palaeointensity method to Pleistocene lava flows from the trans-Mexican volcanic belt. *Phys Earth Planet Inter* 179:139–156
- Muxworthy AR, Heslop D (2011) A Preisach method for estimating absolute paleofield intensity under the constraint of using only isothermal measurements: 1. Theoretical framework. *J Geophys Res-Solid Earth* 116:B04102
- Muxworthy AR, Heslop D, Paterson GA, Michalk D (2011) A Preisach method for estimating absolute paleofield intensity under the constraint of using only isothermal measurements: 2. Experimental testing. *J Geophys Res-Solid Earth* 116:B04103
- Panovska S, Constable C, Korte M (2018) Extending global continuous geomagnetic field reconstructions on timescales beyond human civilization. *Geochem Geophys Geosyst* 19:4757–4772
- Paterson GA, Heslop D, Muxworthy AR (2010) Deriving confidence in paleointensity estimates. *Geochem Geophys Geosyst* 11:Q07Z18. <https://doi.org/10.1029/2010gc003071>
- Paterson GA, Tauxe L, Biggin AJ (2014) On improving the selection of paleointensity data. *Geochem Geophys Geosyst* 15:1180–1192
- Paterson GA, Biggin AJ, Hodgson E, Hill MJ (2015) Thellier-type paleointensity data from multidomain specimens. *Phys Earth Planet Inter* 245:117–133
- Paterson GA, Heslop D, Pan Y (2016) The pseudo-Thellier palaeointensity method: new calibration and uncertainty estimates. *Geophys J Int* 207:1596–1608
- Perrin M (1998) Paleointensity determination, magnetic domain structure, and selection criteria. *J Geophys Res-Solid Earth* 103:30591–30600
- Poletti W, Biggin AJ, Trindade RIF, Hartmann GA, Terra-Nova F (2018) Continuous millennial decrease of the Earth's magnetic axial dipole. *Phys Earth Planet Inter* 274:72–86
- Prévot M, Lecaillon A, Mankinen EA (1981) Magnetic effects of maghematization of oceanic crust. *J Geophys Res* 86:4009–4020
- Prévot M, Derder ME, McWilliams M, Thompson J (1990) Intensity of the Earth's magnetic-field – evidence for a Mesozoic dipole low. *Earth Planet Sci Lett* 97:129–139
- Riisager J, Perrin M, Riisager P, Vandamme D (2001) Palaeomagnetic results and palaeointensity of late cretaceous Madagascan basalt. *J Afr Earth Sci* 32:503–518
- Roberts AP, Tauxe L, Heslop D (2013) Magnetic paleointensity stratigraphy and high-resolution Quaternary geochronology: successes and future challenges. *Quat Sci Rev* 61:1–16
- Rolph TC, Shaw J (1985) A new method of paleofield magnitude correction for thermally altered samples and its application to Lower Carboniferous lavas. *Geophys J Roy Astronom Soc* 80:773–781
- Selkin PA, Gee JS, Tauxe L (2007) Nonlinear thermoremanence acquisition and implications for paleointensity data. *Earth Planet Sci Lett* 256:81–89
- Shaw J (1974) A new method of determining the magnitude of the palaeomagnetic field, application to five historic lavas and five archaeological samples. *Geophys J Roy Astronom Soc* 39:133–141
- Shaw J, Walton D, Yang S, Rolph TC, Share JA (1996) Microwave archaeointensities from Peruvian ceramics. *Geophys J Int* 124:241–244
- Shaw J, Hill MJ, Openshaw SJ (2001) Investigating the ancient Martian magnetic field using microwaves. *Earth Planet Sci Lett* 190:103–109
- Smirnov AV, Tarduno JA, Kulakov EV, McEnroe SA, Bono RK (2016) Palaeointensity, core thermal conductivity and the unknown age of the inner core. *Geophys J Int* 205:1190–1195
- Snowball I, Sandgren P (2002) Geomagnetic field variations in northern Sweden during the Holocene quantified from varved lake sediments and their implications for cosmogenic nuclide production rates. *Holocene* 12:517–530
- Solanki SK, Usoskin IG, Kromer B, Schussler M, Beer J (2004) Unusual activity of the Sun during recent decades compared to the previous 11,000 years. *Nature* 431:1084–1087
- Stephenson A, Collinson DW (1974) Lunar magnetic field palaeointensities determined by an anhysteretic remanent magnetization method. *Earth Planet Sci Lett* 23:220–228
- Stern DP (2002) A millennium of geomagnetism. *Rev Geophys* 40:1–30
- Suavet BP, Weiss WS, Cassata DL, Shuster J, Gattacceca J, Chan L, Garrick-Bethell I, Head JW, Grove TL, Fuller M (2013) Persistence and origin of the lunar core dynamo. *Proc Natl Acad Sci USA* 110:8453–8458

- Tanaka H, Komuro N (2009) The Shaw paleointensity method: can the ARM simulate the TRM alteration? *Phys Earth Planet Inter* 173:269–278
- Tarduno JA, Cottrell RD, Smirnov AV (2006) The paleomagnetism of single silicate crystals: recording geomagnetic field strength during mixed polarity intervals, superchrons, and inner core growth. *Rev Geophys* 44:RG1002. <https://doi.org/10.1029/2005RG000189>
- Tarduno JA, Cottrell RD, Nimmo F, Hopkins J, Voronov J, Erickson A, Blackman E, Scott ERD, McKinley R (2012) Evidence for a dynamo in the main group pallasite parent body. *Science* 338:939–942
- Tarduno JA, Cottrell RD, Davis WJ, Nimmo F, Bono RK (2015) A hadean to paleoarchean geodynamo recorded by single zircon crystals. *Science* 349:521–524
- Tauxe L (2006) Long-term trends in paleointensity: the contribution of DSDP/ODP submarine basaltic glass collections. *Phys Earth Planet Inter* 156:223–241
- Tauxe L, Staudigel H (2004) Strength of the geomagnetic field in the Cretaceous Normal Superchron: new data from submarine basaltic glass of the Troodos Ophiolite. *Geochem Geophys Geosyst* 5:Art. No. Q02H06
- Tauxe L, Yamazaki T (2015) Paleointensities. In: Kono M (ed) *Geomagnetism*. Elsevier, Amsterdam, pp 461–509
- Thellier E (1941) Sur la vérification d'une méthode permettant de déterminer l'intensité du champ terrestre dans le Passé. *C R Acad Sci Paris* 212–283
- Thellier E, Thellier O (1959) Sur l'intensité du champ magnétique terrestre dans la passé historique et géologique. *Ann Géophys* 15:285–376
- Tikoo SM, Weiss BP, Shuster DL, Suavet C, Wang HF, Grove TL (2017) A two-billion-year history for the lunar dynamo. *Sci Adv* 3: e1700207
- Tsunakawa H, Shaw J (1994) The Shaw method of paleointensity determinations and its application to recent volcanic rocks. *Geophys J Int* 118:781–787
- Valet JP (2003) Time variations in geomagnetic intensity. *Rev Geophys* 41:1004. <https://doi.org/10.1029/2001RG000104>
- Valet JP, Brassart J, LeMeur I, Soler V, Quidelleur X, Tric E, Gillot PY (1996) Absolute paleointensity and magnetomineralogical changes. *J Geophys Res-Solid Earth* 101:25029–25044
- Valet JP, Meynadier L, Guyodo Y (2005) Geomagnetic dipole strength and reversal rate over the past two million years. *Nature* 435:802–805
- Van der Voo R (1990) The reliability of paleomagnetic data. *Tectonophysics* 184:1–9
- Veitch RJ, Hedley IG, Wagner JJ (1984) An investigation of the intensity of the geomagnetic-field during Roman times using magnetically anisotropic bricks and tiles. *Arch Sci* 37:359–373
- Walton D (1979) Geomagnetic intensity in Athens between 2000 BC and AD 400. *Nature* 277:643–644
- Wang H, Weiss BP, Bai X-N, Downey BG, Wang J, Wang J, Suavet C, Fu RR, Zucolotto ME (2017) Lifetime of the solar nebula constrained by meteorite paleomagnetism. *Science* 355:623–627
- Weiss BP, Berdahl JS, Elkins-Tanton L, Stanley S, Lima EA, Carporzen L (2008a) Magnetism on the angrite parent body and the early differentiation of planetesimals. *Science* 322:713–716
- Weiss BP, Fong LE, Vali H, Lima EA, Baudenbacher FJ (2008b) Paleointensity of the ancient Martian magnetic field. *Geophys Res Lett* 35(23):L23207. <https://doi.org/10.21029/22008GL035585>
- Wilson RL (1961) Palaeomagnetism in Northern Ireland. 1. The thermal demagnetization of natural magnetic moments in rocks. *Geophys J Roy Astronom Soc* 5:45–58
- Yamamoto Y, Tsunakawa H, Shibuya H (2003) Palaeointensity study of the Hawaiian 1960 lava: implications for possible causes of erroneously high intensities. *Geophys J Int* 153:263–276
- Ziegler LB, Constable CG, Johnson CL, Tauxe L (2011) PADM2M: a penalized maximum likelihood model of the 0–2 Ma palaeomagnetic axial dipole moment. *Geophys J Int* 184:1069–1089

Paleomagnetism, Magnetostratigraphy

Donald R. Prothero

Department of Geology, Occidental College, Los Angeles, CA, USA

Synonyms

Magnetic polarity stratigraphy; Magnetic stratigraphy

Definition

- | | |
|-----------------------|---|
| Magnetic stratigraphy | Application of magnetic methods to obtain the magnetic polarity pattern in stratified rocks, allowing them to be correlated to the global magnetic timescale. |
| Remanent magnetism | Permanent magnetism in rocks, caused by the orientation of the Earth's magnetic field at the time the rock was formed. |

Magnetic Stratigraphy

Introduction

Since the 1960s, one of the most important tools for dating and correlating fossiliferous sedimentary rocks has been magnetic stratigraphy. First developed to aid in the correlation of deep-sea sediment cores from the Deep Sea Drilling Project, by the 1970s it had also been applied to dating terrestrial sedimentary rocks as well. The early efforts were hampered by the slowness and insensitivity of old spinner magnetometers, but with the widespread availability and the increased reliability, sensitivity, speed, and precision of cryogenic magnetometers in the 1970s, the field has grown tremendously. It has now been applied to many of the important marine and terrestrial sequences that were the type sections for biostratigraphic zones and timescales, and is routinely used whenever there are suitable outcrops and age control to allow global correlation.

Magnetic stratigraphy works by obtaining a dense sampling of oriented samples from stratigraphic sections of layered sedimentary rocks or volcanic rocks. These samples record the direction of the magnetic field (*normal*, as in today's field, or *reversed*, about 180° from today's field) in a random pattern of polarity zones. If there is a sufficiently long pattern of polarity zones, it can be matched up to the global magnetic polarity timescale by calibrating it with radiometric ages, or by means of biostratigraphy, or both. Once sections have been correlated, the magnetic polarity

reversals represent time planes, and their ages can be resolved to less than 10,000 years, much better dating and resolution than obtainable by any other means.

Detrital Remanent Magnetization

Although it is possible to correlate lava flows by magnetic methods, by far the most common application of magnetic stratigraphy is to layered sedimentary rocks. These rocks become magnetized by *detrital remanent magnetization* (DRM). When the sediment is loose and fluid, it will contain a small number of tiny (less than 60 μm in diameter) grains of magnetic minerals, such as magnetite or hematite. While these sediments settle out and become liquefied, the magnetic mineral grains are influenced by the ambient magnetic field at the time the sediment was deposited, and thus line up to be parallel to it. Thus, the magnetic grains act like tiny bar magnets, or like iron filings near a bar magnet. As the sediment becomes compacted or cemented and turned into sedimentary rock, the orientation of these magnetic grains is preserved. Although there are instances of sediment compaction causing the grains to rotate and thus flattening the inclination angle, most sedimentary rocks seem to record field directions that closely match those produced from the cooling of igneous rocks, which are not subject to compaction (Butler 1992, pp. 183–187).

Field and Laboratory Techniques

In order to undertake a magnetostratigraphic study, a large number of oriented samples densely spanning the total stratigraphic interval of exposed rock are required. The density of sampling is dictated by how many exposures are suitable, and the logistics of how many people and man-hours of field work are available to conduct the sampling, since it is a tedious and exhausting task. Typically, each stratigraphic level or “site” is sampled at least in three places, and more if possible, so that Fisher statistics can be calculated. In deep-sea cores, sampling is conducted on the split core, with small subsamples taken from the core so that an excessive amount of material is not removed.

The lithology of rock is also critical. Most remanence is held in tiny (less than a few tens of microns) *singledomain grains* of magnetic mineral, or in slightly larger *pseudo-single domain* grains. Above a certain size, it is possible to have two or more magnetic domains in a grain, and then they will tend to form antipodally oriented domains within a single grain and cancel each other out, giving no net remanence. Thus, a fine-grained component is essential for a sedimentary rock to retain useful remanence. For this reason, sampling usually concentrates on shales, mudstones, and fine-grained

sandstones. Coarser sandstones may work if they have a fine-grained matrix. A well-sorted, very clean coarse sandstone with no matrix should be avoided for this reason.

If the rocks are very hard and well cemented, it is possible to sample using a coring drill, as is commonly practiced in igneous and metamorphic rocks. Most sedimentary rocks, however, are too friable and crumbly to stand up to the stress of drilling and the flow of water through the drill bit (required to cool the bit and flush out the drilled rock chips). In this case, samples of terrestrial rocks are taken by scraping oriented surfaces on hand samples, and then subsampling them into cubes or cores in the laboratory using a coring drill or band saw, or hammer and sandpaper for very soft rocks. If the samples are very friable (such as loose soils or unconsolidated sediments), it is still possible to obtain oriented samples. Typically, this is done by pressing open tubes of a high-temperature glass (such as quartz glass or sometimes Pyrex) into the exposure, measuring the orientation, then removing it and pouring a high-temperature-tolerant hardener (such as sodium silicate) into the sediment. The tube is then taped shut, and in the lab it can be sealed at the open end by a high-temperature ceramic, such as Zircar aluminum cement. The sealed glass tube is then measured as if it were a normal drill core. In addition, samples may crumble during preparation until they are too small for a core or cube. They can be salvaged by placing them oriented face down on a piece of wax paper, surrounding them with a plastic ring the diameter of a core (such as a cut piece of plastic test tube), and pouring in Zircar aluminum ceramic to make them into a molded ring that will survive the high temperatures of the thermal demagnetization.

Once samples are fully prepared, they can be measured in a magnetometer to determine the direction and intensity of their magnetic remanence. Before the late 1970s, all the studies were done with older-generation *spinner magnetometers*, which required over 30 min to measure a single sample, and were not sensitive enough to measure most sedimentary rocks. For this reason, only a few measurements could be made on each sample, and only a limited number of samples could be taken if the project were to remain reasonable in size.

In the 1970s, the invention and widespread adoption of *cryogenic magnetometers* revolutionized the study of paleomagnetism. These devices have a sealed jacket of liquid helium at 4 K surrounding a superconducting region in the bottom of the magnetometer. The electronics in the cryogenic chamber are thus sensitive to extremely small magnetic moments, and can make measurements in a few seconds. They are four to five orders of magnitude more sensitive than spinner magnetometers, so samples can be measured that are extremely weakly magnetized. These include not only every kind of rock but even organic tissues like the magnetite in bacteria or birds or bees or butterflies. Early cryogenic magnetometers were very temperamental and

prone to problems. They bled off the evaporating helium at such a fast rate that a single shipment costing hundreds of dollars would only last a few weeks. Only a few measurements could be taken on each sample, and the lab had to work 24/7 to use every precious second of magnetometer time. Now most cryogenic magnetometers have a cryocooler pump that circulates liquid helium or nitrogen around the cryogenic container, and keeps it cold for many months at a time, greatly reducing the cost of operation, wear and tear on the magnetometer, and improving reliability. In many labs (especially those equipped with a Caltech-style automatic sample changer), it is possible to make dozens of measurements in an hour, allowing paleomagnetists to make dozens of different measurements on each sample.

Each sample contains a mixture of remanences acquired during the long history of the rock unit. There will be a primary, or characteristic, component from when the rock originally formed, which may be overprinted by one or more magnetizations that were acquired at later times. For a reliable paleomagnetic study, the researcher must attempt to remove these overprinted components by gradually demagnetizing the sample, and gradually stripping off the younger, less stable magnetizations until only the characteristic remanence is left. Consequently, it is typical for a modern paleomagnetic study to measure the samples first at NRM (*natural remanent magnetization*, or the magnetization of the sample before treatment). The samples can then be demagnetized using *alternating fields* (AF), which helps remove remanence due to low-coercivity minerals such as magnetite, and eliminates magnetization from transient fields (such as lightning strikes) or from VRM (*viscous remanent magnetization*). Examination of the magnetic behavior and the drop in intensity at each AF step allows the paleomagnetist to determine how much of the remanence is held in a low-coercivity mineral, or whether significant high-coercivity minerals (which do not respond to AF demagnetization) are present. AF treatment is usually sufficient for samples from deep-sea cores, which have not undergone any significant diagenesis or alteration. Another method, known as *thermal demagnetization*, heats the samples in field-free space to a preset temperature, then cools them down to room temperature without acquiring any new magnetization from the earth's field. In some cases, the oven is saturated by a jet of nitrogen gas to reduce the possibility of minerals oxidizing as they cool down. Thermal demagnetization in many steps of increasing temperature allows the paleomagnetist to see how the remanence changes as higher and higher temperatures are reached and the *blocking temperature* of the magnetic minerals (578 °C for pure magnetite; 630° for hematite) is approached. At high temperatures, the behavior of the sample above the blocking temperature of magnetite tells the researcher about the intensity and direction of the magnetic components that are held in hematite. Thermal

demagnetization has an additional advantage. Above 200 °C, iron oxides and hydroxides such as goethite are converted to hematite, and their remanence is removed. These “rusty” minerals are often formed by weathering of sedimentary rocks at the surface, and usual form as a *chemical remanence* deposited in the rock long after the primary or characteristic remanence was formed. Thus, they are a form of overprinting that must be removed, but cannot be affected by AF demagnetization. For this reason, nearly all terrestrial sedimentary rocks (even if they are not visibly reddish) should be thermally demagnetized to insure that there are no false overprinted magnetic components. Many studies use both AF and thermal demagnetization on every sample to strip off overprints due to iron oxides, to determine the relative contribution of high-coercivity and low-coercivity minerals, and to obtain a clear sense of the dominant magnetic mineralogy of each sample. The samples can also be studied in polished thin sections, where the light reflects differently off magnetite, goethite, and hematite, and the depositional fabric of the magnetic mineral and size of the grains can be determined.

These magnetic directions can be plotted and analyzed in a number of different ways. For simple visual plotting of the three-dimensional changes in directions, a conventional *stereonet* can be used. But a stereonet only shows direction, where as an *orthogonal demagnetization plot* (“Zijderveld plot”) shows both the direction and intensity of each demagnetization step (Zijderveld 1967). Most paleomagnetics labs have software that rapidly creates these plots, so the behavior of the sample can be studied as the demagnetization proceeds. Once a stable remanence has been obtained and overprints isolated and removed, the inferred characteristic remanence components are usually summarized by the *least squares method* of Kirschvink (1980). These directions for each sample can then be averaged using Fisher (1953) statistics, which give the mean and ellipse of 95% confidence for each sample. Using Watson's (1956) criteria, the statistical significance of each site can be assessed. Frequently they are ranked using the criteria of Opdyke et al. (1977), where Class I sites are statistically significant at the 95% confidence level, that is, there is a 95% chance that the sample distribution differs from a random distribution. These sites were labeled “Class A” by Johnson et al. (1985). Class II sites have fewer than three samples (so statistics cannot be calculated), and Class III sites (“Class B” sites of Johnson et al. 1985) have at least two sample directions that are in agreement, but one which is divergent due to incomplete removal of overprints. These directions can then be plotted against a stratigraphic column to show the overall polarity pattern for the section, and also indicate the degree of reliability of individual samples at each site.

As in any other paleomagnetic study, there are a number of techniques for assessing the stability and reliability of the sample directions (see Butler 1992, pp. 123–128). The most commonly used in magnetostratigraphy is the *reversal test*,

which compares the statistical mean of all normal directions and reversed directions to see if they are antipodal (as they should be if they represent the original magnetic field of the earth). Most long stratigraphic sections yield both normal and reversed polarity samples, providing the data for a stability test like this. In some cases, there are conglomerates in the section, so a *conglomerate test* can be performed. In this test, samples are taken from a large sample of clasts in a conglomerate. If their directions are randomly scattered, then they have not been remagnetized since the conglomerate formed. If, on the other hand, they yield directions that are clustered, then there has been overprinting not only of that conglomerate, but also of the rocks in the rest of the section. Another common test in rocks that have been deformed is the *fold test*. The researcher can compare the directions of the samples with and without dip correction. If these directions are clustered with dip correction, and more scattered without dip correction, then the directions were acquired before folding. If the directions are more clustered in their uncorrected, folded orientation, then the directions were acquired after folding (and thus are not primary or characteristic directions). Not all these stability tests are available in every study, but at least one (usually the reversal test) should be attempted to obtain credible results. If other tests are also available, they should also be analyzed to increase the rigor of the study.

Calibration and Correlation

Once characteristic directions that have passed stability tests have been obtained from each site and they are plotted, then the overall polarity pattern for each individual section can be assessed. Polarity zones are not diagnostic or unambiguous by themselves, but must be correlated based on other criteria, such as lithostratigraphy or biostratigraphy. In the early days of magnetic stratigraphy, deep-sea cores (Fig. 1) could be sampled down from the modern sediments at the top of the core, and each polarity zone could be calibrated by the Brunhes normal polarity magnetozones at the top. Most deep-sea cores, however, do not have a continuous sedimentary record from the present backward, but represent an older geologic span of time separated by unconformities. In such cases, the biostratigraphy of the planktonic microfossils enables the scientist to calibrate a core directly against the magnetic polarity timescale.

In another example, Fig. 2 shows the correlation of magnetozones of many different stratigraphic sections across the Eocene-Oligocene White River Group from northeastern Colorado to Nebraska and Wyoming to the Big Badlands of South Dakota to North Dakota. The correlations are based on the lithostratigraphic units and the biostratigraphic zonation, with the common datum based on a widespread ash horizon found throughout most of the sections. Note that the polarity

magnetozones are roughly consistent in thickness from one section to another, although each of the magnetozones is proportionally thicker where the total sediment accumulation rate is higher (e.g., Slim Buttes, South Dakota).

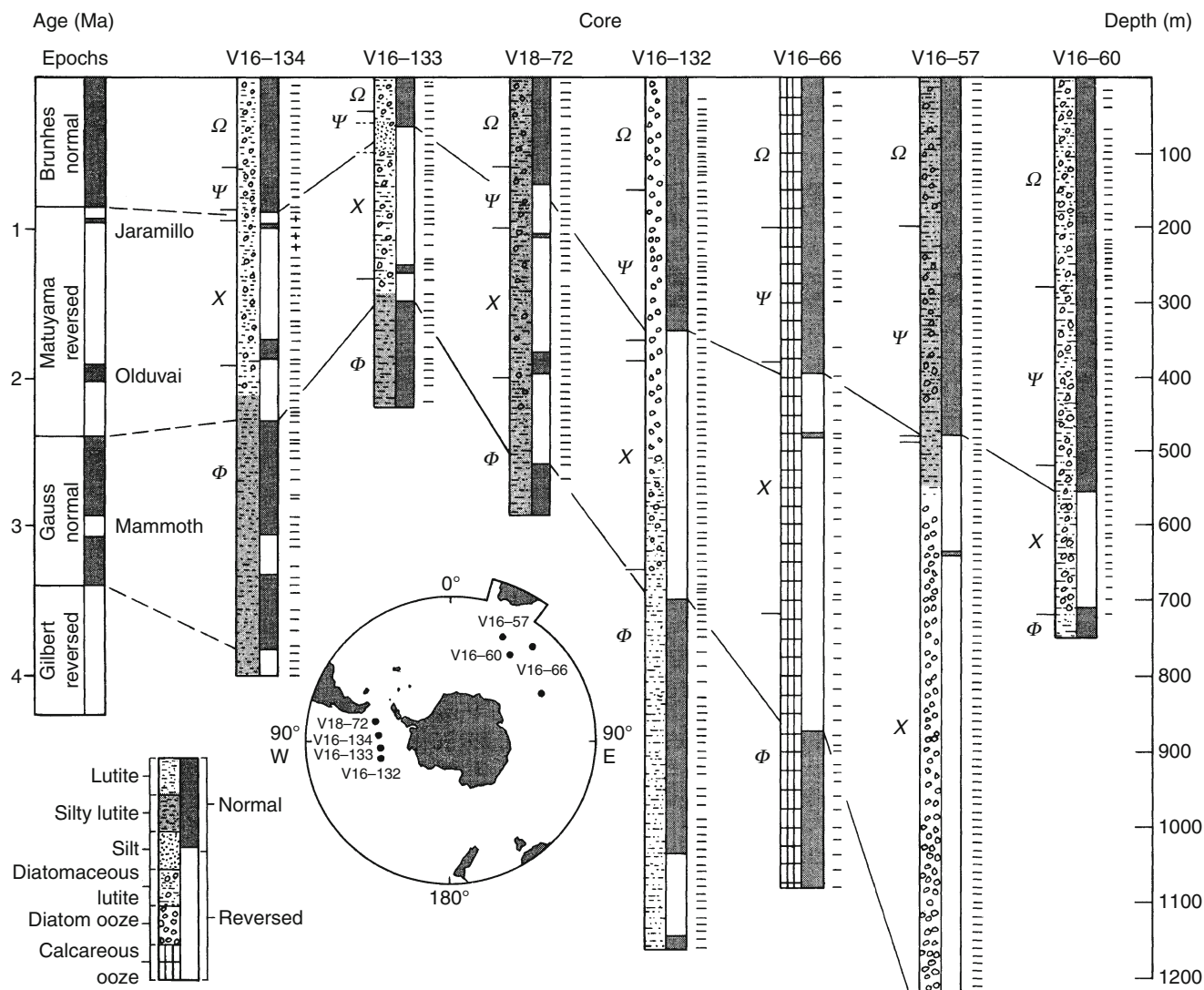
But these polarity matches do not establish correlation with the global magnetic polarity timescale. Magnetic polarity can only be “normal” or “reversed,” and that pattern has occurred thousands of times in the geologic past, so it is not sufficient just to get a raw polarity pattern and attempt to correlate it to the global timescale (unless you are working down from modern sediments at the top of a section, in which case the uppermost normal magnetozones is the Brunhes Chron, or Chron 1n). For reliable correlation, the section needs to be calibrated by some sort of independent means of time control. Ideally, this is provided by at least one and preferably more radiometric dates, which will place the local section on the global magnetic polarity timescale (Fig. 3). Where radiometric dates are not available, it is often possible to use a finely resolved well-calibrated biostratigraphy (whose precise correlation to the global timescale is already known) to correlate another section containing those key fossils.

Strengths and Weaknesses of Magnetic Stratigraphy

Like any other stratigraphic technique (see Prothero and Schwab 2003), magnetic stratigraphy has great strengths – and also some weaknesses. The strengths include:

1. *Worldwide*: magnetic polarity events happen around the globe simultaneously. By contrast, most other stratigraphic markers (volcanic ashes, biostratigraphic events, lithologic boundaries, even isotopic events) occur only locally.
2. *Independent of facies or lithology*: any fine-grained sedimentary rock or volcanic rock will work, allowing us to correlate between deep-sea muds to shallow-marine silts to terrestrial floodplain mudstones. By contrast, rock units are typically time-transgressive as facies migrate, and are not found on both land and sea. Stable isotope curves are usually restricted to marine rocks.
3. *Geological instantaneous*: polarity reversals take place over 4000–5000 years, which is a geological instant for most studies of the deep past. By contrast, most lithostratigraphic units are time-transgressive over millions of years.
4. *Constant resolution regardless of age*: the duration of magnetic polarity transitions is constant whether the rocks are Pleistocene or Permian, so the resolution is always in the order of a few thousand years.

By contrast, radiometric dates have an error bar that is a proportion of the mean age and therefore gets larger as the rock gets older. For example, if the error estimates are $\pm 5\%$



Paleomagnetism, Magnetostratigraphy, Fig. 1 Magnetostratigraphic correlation of Antarctic cores, based on the modern Brunhes normal polarity sediments at the top of each section, and the biostratigraphic zones of radiolaria (shown by the capital Greek letters) in each

core. To the right of each column, minus signs indicate normally magnetized specimens, while plus signs are reversed specimens. (After Opdyke et al. 1966, Fig. 1; by permission of the American Association of the Advancement of Science)

for a rock of 1 Ma, then the true age is 1 million $\pm$ 50,000 years. If the age is 100 Ma, however, then the true age is 100 million $\pm$ 5 million years. In other words, there is a 95% probability that the true age lies between 95 and 105 Ma. These error bars get larger with older and older rocks.

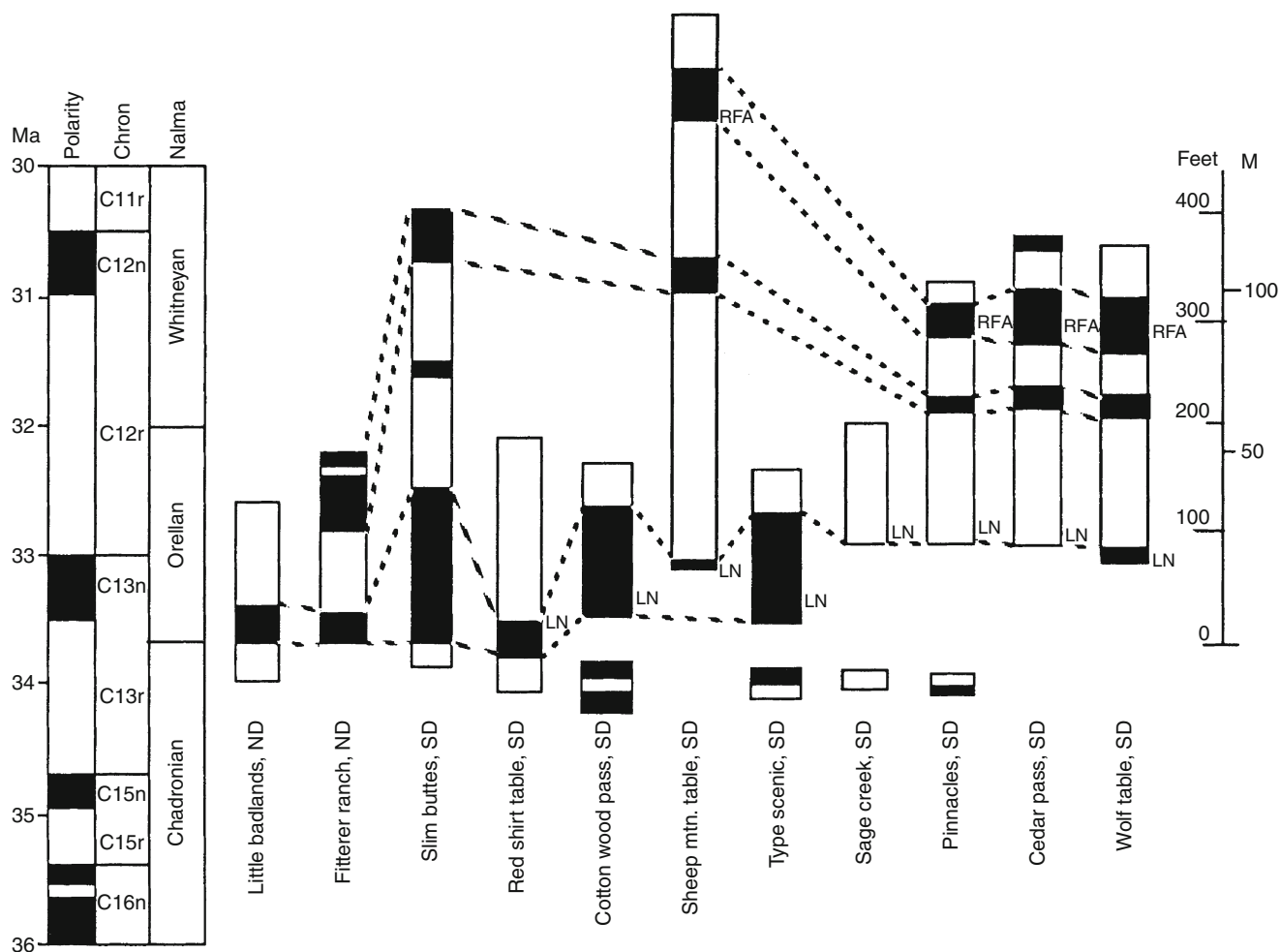
There are some limitations to magnetic stratigraphy as well:

1. Only *finer-grained lithologies*, or sandstones with clay matrix, are suitable (as discussed above).
2. If there are *unconformities* in the section, then magnetostratigraphic zones can be shortened or deleted altogether. This problem can best be addressed by biostratigraphic methods, the usual way in which we test for evidence of slow or no deposition.

3. *Demagnetization problems* can hamper the quality of the data, particularly in yielding spurious normal overprinted zones that have not yet been removed by thermal demagnetization of iron oxides and hydroxides (e.g., Butler and Lindsay 1985).
4. *Independent form of age control* is required to calibrate the section.

Some Applications and Implications

Since the late 1960s, magnetic stratigraphy has become one of the most powerful tools available for precise dating of sedimentary sequences and geologic events, and for correlation of events between the terrestrial, shallow-marine, and deep-sea



Paleomagnetism, Magnetostratigraphy, Fig. 2 Correlation of magnetostratigraphic sections in the White River Group from North and South Dakota. All of the sections are correlated based on the datum of the base of the Orellan. Abbreviations as follows:

NALMA = North American land mammal ages; LN = Lower Nodular Zone of the Big Badlands; RFA = Rockyford Ash. (From Prothero (1996, Fig. 9))

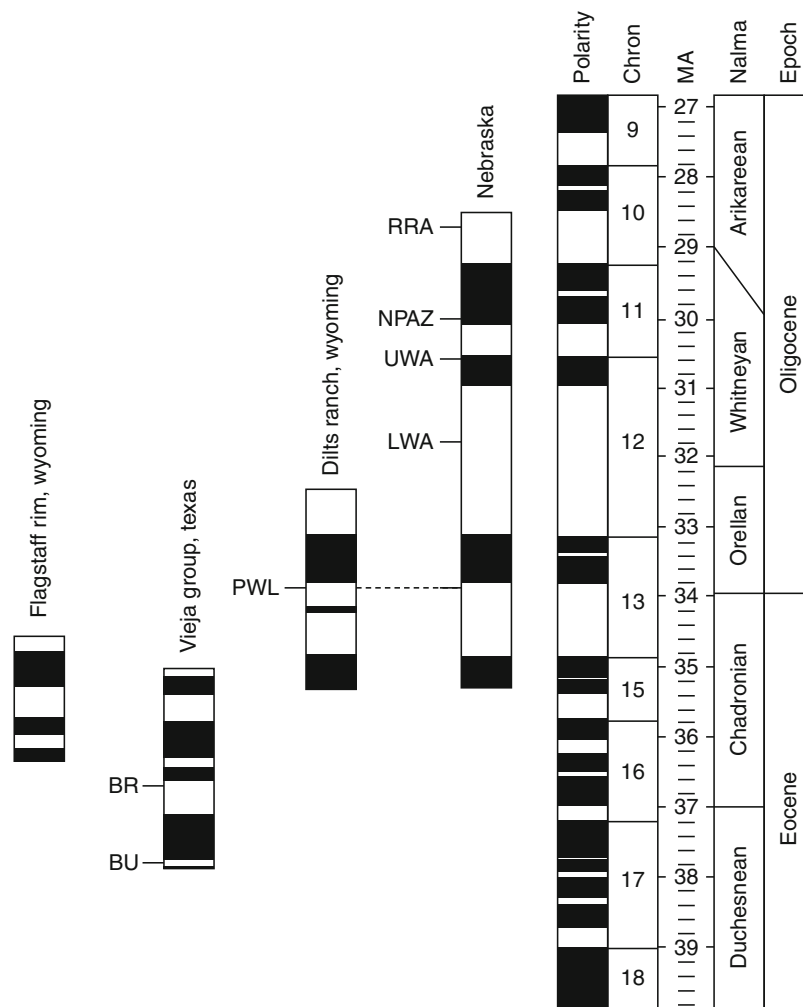
records. In most detailed stratigraphic studies, it is routine to do magnetostratigraphic analysis along with biostratigraphy to form a network of correlation now called *magnetobiostratigraphy* (Fig. 4). Combined with stable isotope stratigraphy (especially in deep marine cores), it is possible to obtain very high-resolution, high-precision dating of many stratigraphic sections and correlate their events with high precision on a global basis. This, in turn, has allowed reliable and detailed comparison between climatic and biotic events on the land and the sea, and has revolutionized our understanding of climatic change and global correlations of climate and evolution.

The application of magnetic stratigraphy has also changed our conceptions of many geologic problems. For example, the dating of strata before the advent of magnetic stratigraphy has been radically rethought as these correlations improve. In North America, for example, the Uintan and Duchesnean land mammal ages of Wood et al. (1941) were once thought to be late Eocene, the Chadronian land mammal age early

Oligocene, and the Orellan and Whitneyan land mammal ages middle and late Oligocene, respectively. However, with the advent of magnetobiostratigraphy and high-resolution argon/argon dating, all these correlations have changed radically (Prothero and Swisher 1992). The Uintan and Duchesnean are now middle Eocene, the Chadronian late Eocene, and the Orellan and Whitneyan early Oligocene. The Arikarean, once thought to be early Miocene, is now largely late Oligocene. Every one of the original mid-Cenozoic stratigraphic interpretations of the Wood Committee (1941) was mis-correlated to the standard European timescale by at least one stage of the Eocene or Oligocene (and the same was true of their North American Miocene-Pliocene stages). Prior to magnetic stratigraphy and radiometric dating, the Desadean land mammal age of South America was thought to be early Oligocene (27–37 Ma). Today it is considered late Oligocene to early Miocene (22–27 Ma) in age (MacFadden et al. 1985), or about 10 million years younger than originally thought.

Paleomagnetism, Magnetostratigraphy,

Fig. 3 Calibration of the White River-Arikaree Group magnetic stratigraphy of Fig. 2, based on $^{40}\text{Ar}/^{39}\text{Ar}$ dating of key volcanic ashes. Abbreviations: BR = Bracks Rhyolite; BU = Buckshot Ignimbrite; LWA = Lower Whitney Ash; NPAZ = Nonpareil ash zone; RRA = Roundhouse Rock Ash; UWA = Upper Whitney Ash. (Modified from Prothero and Swisher 1992, Fig. 2.3)



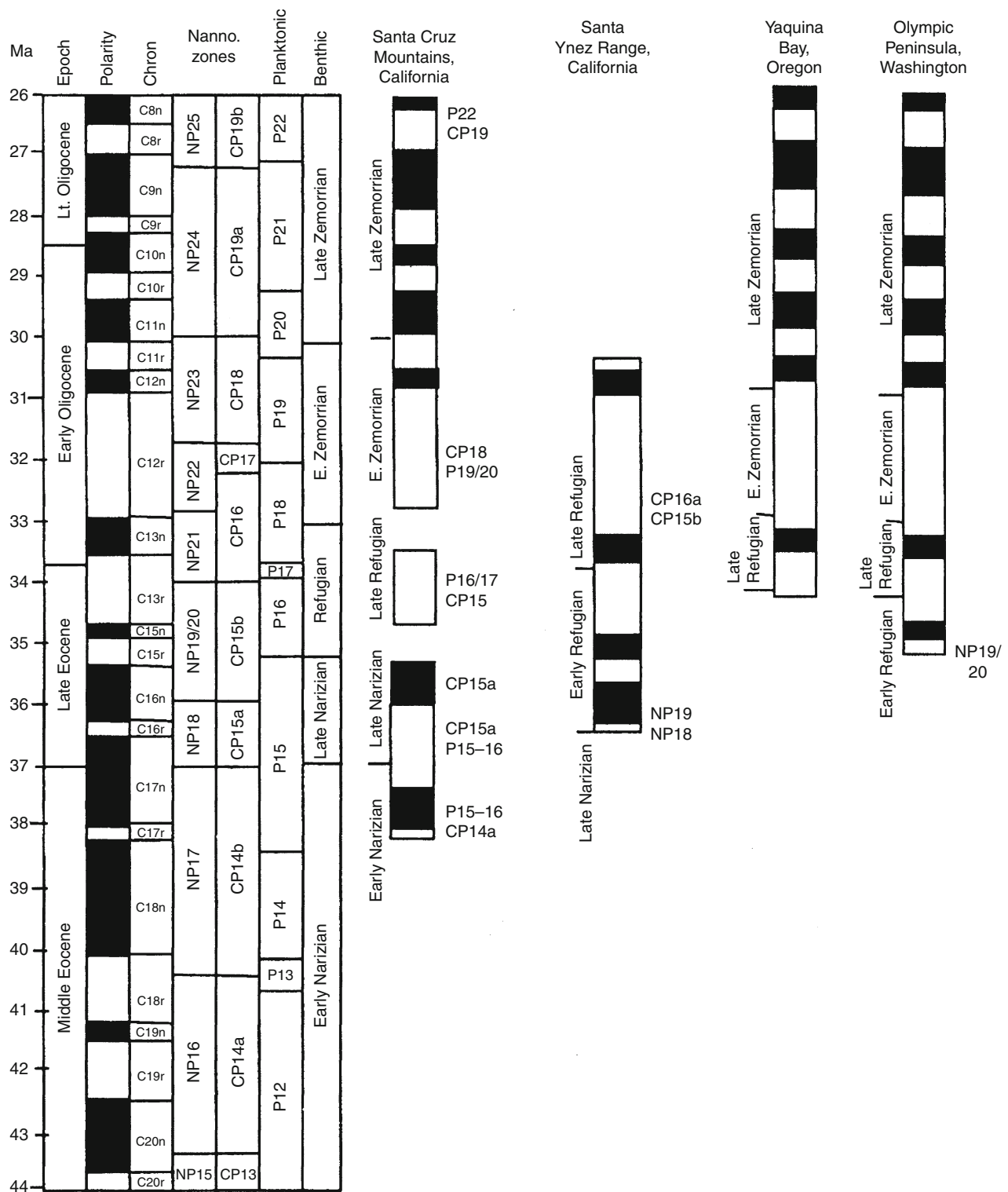
Magnetic stratigraphy has also been used as a device for testing other geologic ideas. Because it produces geologically instantaneous time planes, it is valuable in testing whether biostratigraphic datums are synchronous or not. By and large, this has proven true, confirming the importance of biostratigraphy as a time indicator. Prothero (1982) showed this in the Eocene-Oligocene strata and fossils of the White River Group of the American High Plains. Lindsay et al. (1984) did the same for the biostratigraphy of the late Cenozoic of southwestern North America and Europe, as did Flynn et al. (1984) for the Eocene of the Rocky Mountains. A few biostratigraphic events, like *Lepus* event in North America (Lindsay et al. 1990), or the *Hipparion* horse datum in Europe, are demonstrably time-transgressive (Woodburne et al. 1981; Woodburne 1989), but these are the exceptions to the rule.

In some cases, magnetic stratigraphy produces powerful inferences about the rock record. For example, Johnson et al. (1975) showed that magnetozones could be mapped along the Miocene-Pliocene exposures in the San Pedro Valley of Arizona. These “invisible magnetic time planes” (Fig. 5) exist in the outcrops and can be mapped and interpreted using

standard magnetostratigraphic methods. Magnetic stratigraphy allows precise determination of sedimentation rates in marine and terrestrial sections, and relative completeness of sections, which are important to many different kinds of stratigraphic studies (Sadler 1981; May et al. 1985; Johnson et al. 1988).

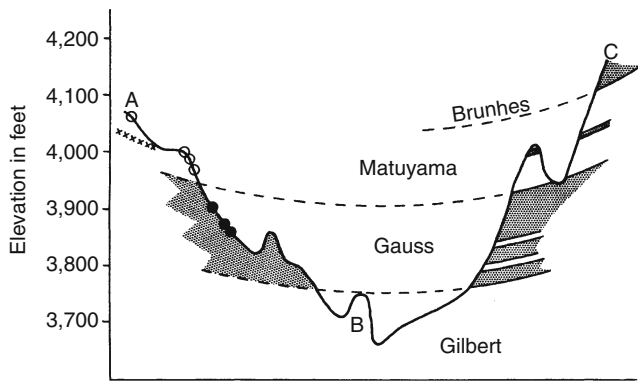
Summary

Magnetic stratigraphy is now one of the principal tools that allow the precise, high-resolution correlation of geologic events in the stratigraphic record, especially climatic events. It is unique among stratigraphic methods in that global high-resolution correlations are possible between terrestrial and marine records, and does not suffer from issues of local facies changes or limited lateral extent of most other stratigraphic markers. Of course, the researcher must be aware of the limitations: relatively complete well-exposed sections of fine-grained rocks are required, as is an independent form of time control to allow calibration. Every sample must be



Paleomagnetism, Magnetostratigraphy, Fig. 4 Magnetostratigraphic correlation of marine sections from the Pacific Coast, based on the magnetic polarity pattern and the biostratigraphy of

microplankton. On the left is the standard geological time scale and polarity chrons with the planktonic zones of nannofossils and foraminifera. (After Prothero 2001)



Paleomagnetism, Magnetostratigraphy, Fig. 5 Topographic profile across the San Pedro Valley of Arizona, showing the interpretation of “invisible magnetic zones” lying within the outcrops based on the sampling of Johnson et al. (1975). (Modified from Johnson et al. 1975, Fig. 8)

carefully analyzed by multiple techniques to insure that the primary or characteristic remanence has been recovered and overprinting removed. Nonetheless, magnetic stratigraphy has improved our correlations of global events by several orders of magnitude, and corrected many erroneous correlations that were once widely accepted. It has allowed for many other interesting studies, such as analysis of rates of sedimentation, and tests of the isochroneity of biostratigraphic events. In short, the horizon is limitless for the application of paleomagnetism to stratified rocks.

Cross-References

- [Magnetic Domains](#)
- [Magnetometers](#)
- [Paleomagnetism, Principles](#)
- [Remanent Magnetism](#)

Bibliography

- Aïssaoui DM, McNeill DF, Hurley NF (1993) Applications of paleomagnetism to sedimentary geology. *SEPM Spec Publ* 49:1–215
- Butler RF (1992) *Paleomagnetism*. Blackwell, Boston, p 320
- Butler RF, Lindsay EH (1985) Mineralogy of magnetic minerals and revised magnetic polarity stratigraphy of continental sediments, San Juan Basin, New Mexico. *J Geol* 93:535–554
- Fisher RA (1953) Dispersion on a sphere. *Proc R Soc Lond Ser A* 217:295–305
- Flynn JJ, MacFadden BJ, McKenna MC (1984) Land mammal ages, faunal heterochrony, and temporal resolution in Cenozoic terrestrial sequences. *J Geol* 92:687–705
- Hailwood EA (1989) *Magnetostratigraphy*. Blackwell, Oxford, pp 1–84
- Johnson NM, Opdyke ND, Lindsay EH (1975) Magnetic polarity stratigraphy of Pliocene-Pleistocene terrestrial deposits and vertebrate faunas, San Pedro Valley, Arizona. *Geol Soc Am Bull* 86:5–12
- Johnson NM, Stix J, Tauxe L, Cerveny PF, Tahirkheli RAK (1985) Paleomagnetic chronology, fluvial processes, and tectonic implications of the Siwalik deposits near Chinji Village, Pakistan. *J Geol* 93:27–40
- Johnson NM, Shiekh KA, Dawson-Sanders E, McRae L (1988) The use of magnetic reversal time lines in stratigraphic analysis: a case study in measuring variability in sedimentation rates. In: Kleinspehn K, Paola C (eds) *New perspectives in basin analysis*. Springer, New York, pp 307–330
- Kennett JP (ed) (1980) *Magnetic stratigraphy of sediments*. Dowden, Hutchinson & Ross, Stroudsburg, p 438
- Kirschvink JL (1980) The least-squares line and plane and the analysis of paleomagnetic data: examples from Siberia and Morocco. *Geophys J R Astron Soc* 62:699–718
- Lindsay EH, Opdyke NM, Johnson NM (1984) Blancan-Hemphillian land mammal ages and late Cenozoic dispersal events. *Annu Rev Earth Planet Sci* 12:445–488
- Lindsay EH, Smith GA, Haynes GV, Opdyke ND (1990) Sediments, geomorphology, magnetostratigraphy, and vertebrate paleontology in the San Pedro Valley, Arizona. *J Geol* 98:605–619
- MacFadden BJ, Campbell KE, Cifelli RL, Siles O, Johnson NM, Naeser CW, Zeitler PK (1985) Magnetic polarity stratigraphy and mammalian fauna of the Deseadan (late Oligocene-early Miocene) Salla beds in northern Bolivia. *J Geol* 93:223–250
- May SR, Butler RF, Roth FA (1985) Magnetic polarity stratigraphy and stratigraphic completeness. *Geophys Res Lett* 12:341–344
- Opdyke ND, Channell JET (1996) *Magnetic stratigraphy*. Academic, San Diego, p 341
- Opdyke ND, Glass B, Hays JD, Foster J (1966) Paleomagnetic study of Antarctic deep-sea cores. *Science* 154:349–357
- Opdyke ND, Lindsay EH, Johnson NM, Downs T (1977) The paleomagnetism and magnetic polarity stratigraphy of the mammal-bearing section of Anza-Borrego State Park, California. *Quat Res* 7:316–329
- Prothero DR (1982) How isochronous are mammalian biostratigraphic events? *Proc 3rd N Am Paleontol Convention* 2:405–409
- Prothero DR (1988) Mammals and magnetostratigraphy. *J Geol Educ* 34(4):227–236
- Prothero DR (1996) Magnetostratigraphy of the White River group in the high plains. In: Prothero DR, Emry RJ (eds) *The terrestrial Eocene-Oligocene transition in North America*. Cambridge University Press, Cambridge, pp 247–262
- Prothero DR (ed) (2001) *Magnetic stratigraphy of the Pacific Coast Cenozoic*. Pacific Section SEPM Special Publication, vol 91. 394 p
- Prothero DR, Schwab F (2003) *Sedimentary geology*, 2nd edn. W.H. Freeman, New York, p 557
- Prothero DR, Swisher CC III (1992) Magnetostratigraphy and geochronology of the terrestrial Eocene-Oligocene transition in North America. In: Prothero DR, Berggren WA (eds) *Eocene-Oligocene climatic and biotic evolution*. Princeton University Press, Princeton, pp 46–74
- Sadler PM (1981) Sediment accumulation rates and completeness of the stratigraphic record. *J Geol* 89:569–584
- Watson GS (1956) A test for randomness of directions. *Mon Not R Astron Soc Geophys Suppl* 7:160–161
- Wood HE II, Chaney RW, Clark J, Colbert EH, Jepsen GL, Reeside JB Jr, Stock C (1941) Nomenclature and correlation of the North American continental Tertiary. *Bull Geol Soc Am* 52:1–48
- Woodburne MO (1989) Hipparchon horses: a pattern of worldwide dispersal and endemic evolution. In: Prothero DR, Schoch RM (eds) *The evolution of perissodactyls*. Oxford University Press, Oxford, pp 197–233
- Woodburne MO, MacFadden BJ, Skinner MF (1981) The North American Hipparchon Datum and the implications for the Neogene of the Old World. *Geobios* 14(4):493–524
- Zijderveld JDA (1967) AC demagnetization of rocks: analysis of results. In: Collinson DW, Creer KM, Runcorn SK (eds) *Methods in paleomagnetism*. Elsevier, Amsterdam, pp 254–286

Paleomagnetism, Measurement Techniques and Instrumentation

T. Radhakrishna^{1,2}, J. D. A. Piper³ and Asanulla R. Mohamed¹

¹National Centre for Earth Science Studies, Trivandrum, India

²GITAM University, Nagadenehalli, Bengaluru, India

³Geomagnetism Laboratory, Department of Earth and Ocean Sciences, University of Liverpool, Liverpool, UK

Synonyms

Ancient magnetism; Fossil magnetism

Definition

Palaeomagnetism is the study of the Earth's magnetic field in past times. A permanent record of the ancient magnetic field is recorded by certain ferromagnetic minerals contained in rocks either precipitated as primary crystals, or as secondary alteration products formed during later geological events. The ancient direction of the geomagnetic field can be readily determined from this record, and the time of acquisition can usually be dated when carefully related to the geological history. The information is essential for quantifying motions of the tectonic plates comprising the outer shell of the Earth and resolving the nature of global tectonics in past times. The ancient intensity of the field is a more challenging parameter to resolve, but is key to resolving the long-term history of the geodynamo, while techniques for recovering it are improving all the time. Although this contribution is concerned primarily with the record of fossil magnetism in the rock record and the means of determining it, the acquisition of magnetism is an ongoing process within materials forming at the present day. In addition to continuing igneous activity and sedimentation, it is acquired, for example, by diagenetic processes in soils and by the secretion of magnetotactic bacteria within living things.

Introduction

The permanent magnetism residing in a natural material, lodestone containing the mineral magnetite (Fe_3O_4), has been known for many centuries for its ability to align with the polar direction and has long been used for this purpose by navigators. It was originally believed that the pole star (Polaris) or a large magnetic island at the North Pole attracted lodestone compasses. However, William Gilbert (1544–1603)

in his book *De Magnete, Magneticisque Corporibus, et de Magno Magnete Tellure* published in 1600, concluded that the Earth behaved like a giant magnet. He argued that the center of the Earth was iron. His other important achievement was documenting the special property of magnets by showing that they could be cut so that each new magnet possessed north and south poles. By the nineteenth century, it had been discovered that certain rocks were capable of becoming permanently magnetized in the direction of the ambient geomagnetic field. Bernard Brunhes (1867–1910) found young volcanic lavas in central France magnetized in a direction opposite to the prevailing field. This is the phenomenon of geomagnetic reversal, and in 1929 Japanese scientist Motonori Matuyama (1884–1958) was able to show that reversals of the field appeared to have occurred at specific times during the recent geological past.

The geometry of the magnetic field direction at the surface of the Earth is described in terms of declination and inclination. *Declination* refers to the angle in the horizontal plane between geographic north and the geomagnetic direction along which the lines of magnetic field are directed. Since the magnetic field is also directed towards, or away from, the source within the Earth (with the exception of the magnetic equator), it is typically inclined to the horizontal; the angle above (negative) or below (positive) the horizontal is the *inclination* of the field. In the characteristic surface expression of an inner dipole source, it will be latitude-dependent and range smoothly from vertically up or down at the magnetic poles through zero at the equator.

A geocentric dipole source with an axis aligned to geographic north-south rotation axis is the model used for relating a palaeomagnetic direction to a pole position relevant to unraveling ancient continental movements. For this assumption to be true, it is important to demonstrate that the magnetic poles are always constrained to the geographic poles of rotation in the long term. Historic records of the magnetic field recorded in London from 1580 show that the declination of the magnetic field changed from 11°E to 24°W by 1820 with the calculated position of the magnetic pole executing a path around, and close to, the geographic pole of rotation. This movement is *secular variation* and when time-averaged the migration of the magnetic pole proves to be very close to the geographic pole. Subsequent analyses found that Late Tertiary palaeomagnetic poles, a time interval over which continental motions had been relatively small, yielded magnetic pole estimates very close to the pole of rotation when averaged over intervals of thousands of years. In 1926 this led P-L Mercanton (1876–1963) to suggest that fossil magnetism could be applied during more ancient times to monitor movements of the continents.

An outcome of magnetic research conducted during World War II was the development by Patrick Blackett (1897–1974)

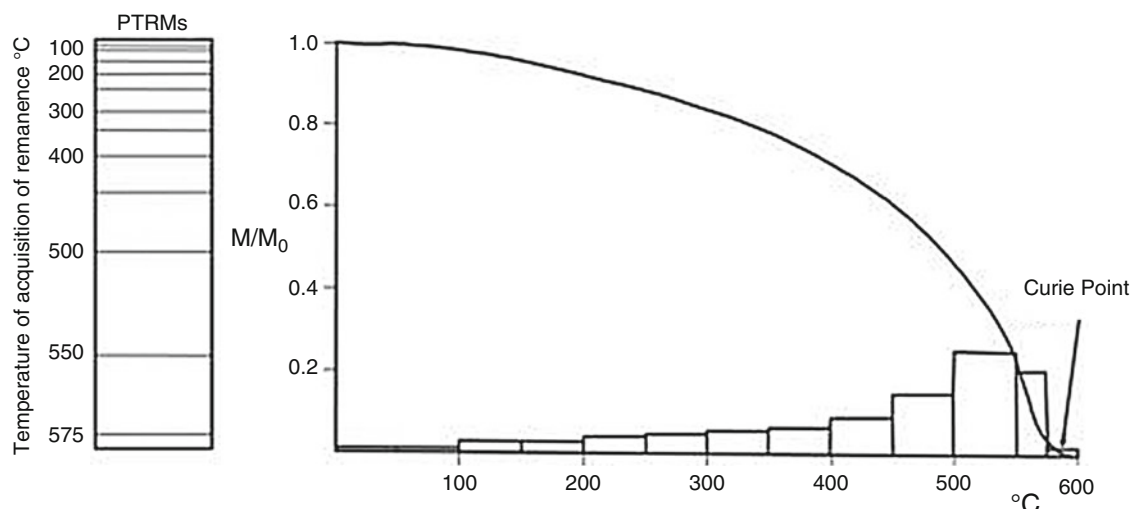
of a sensitive astatic magnetometer able to accurately measure magnetism in relatively weakly magnetized materials such as rock samples. It was first employed from the early 1950s by Patrick Blackett and Keith Runcorn (1922–1995). Analysis of the evidence, particularly by the seminal work of Edward Irving (1927–2014), showed that the continents had moved over the surface of the Earth in ways that matched movement through latitude conforming to ancient climatic evidence from the source rocks. This evidence placed the theory of continental drift, which had languished since proposed by Alfred Wegener (1880–1930) in 1915, onto a firm quantitative basis. When it was realized that the ocean crust could also preserve a palaeomagnetic record from the time of its formation, continental drift was rapidly integrated with the temporal evidence from the magnetic lineaments defining crustal growth of the ocean floors (sea floor spreading). This provided the key quantitative evidence for the Theory of Plate Tectonics in the 1960s, a theory formally articulated in 1967. The Plate Tectonic paradigm accommodated extensive additional evidence from earthquake and volcanic activity and has prevailed with much refinement, as the explanation for the growth and destruction of Earth's surface plates up to the present day. Palaeomagnetism has thus provided the primary data for quantifying the rate of plate movements and remains one of the most important techniques in the Earth Sciences. It is also versatile and has applications to many other fields including environmental sciences and archaeology.

In addition to its primary long-term value for tracing the movements of continents and evaluating palaeogeography, palaeomagnetism is also used to unravel the history of reversals of the geomagnetic field polarity. Reversals have not occurred at regular intervals although there is a long-term trend for intervals of frequent reversal to become less frequent with time and move into long periods when there were no reversals (superchrons). Such intervals may last for tens of millions of years and when reversals recommence their frequency tends to increase again; this is a geomagnetic definition of a long term behavioral transition within the Earth's interior. The Cretaceous Normal "Superchron" (121–83 Ma) was the last long interval when reversals were absent. Subsequently the frequency of reversals has increased to a maximum about 10 Ma ago; since then the reversal rate has been generally declining. This history of magnetic reversals comprises the *geomagnetic polarity time scale* (GPTS), a chronology of periods of normal and reversed polarity. It forms the basis of magnetostratigraphy whereby the patterns of magnetic reversal are matched with the stratigraphic record of palaeomagnetism to date rocks and correlate them from place to place. The parallel investigation of the intensity of the geomagnetic field from rocks such as lava flows, which are rapidly cooled and magnetized, provides the essential data for monitoring the influence of secular variation and

investigating the behavior of the geodynamo. A further application of palaeomagnetism is the study of the anisotropy (fabric) of both the remanent magnetism and the magnetic susceptibility. This has applications to processes of sedimentation, the resolution of ancient tectonic stress fields, to fluid-migration and to economic mineral precipitation; a comprehensive guide to methods and applications can be found in Martin-Hernandez et al. (2004).

In common with all-natural materials, rocks possess a weak *diamagnetism* in applied magnetic fields as the electron spins align with, and oppose, an applied field. Minerals containing transition elements, the most important being iron and manganese, also possess a stronger *paramagnetism* as unbalanced electron orbits align with, and reinforce, the field. However, these effects are transient and only survive as long as the applied field is present. Although important to the interpretation of the ambient magnetic fields and of significance to ground surveys, they are not relevant to palaeomagnetism. Instead this relies on a memory of the magnetic field aligned with the ambient magnetic field at the time of rock formation, or at some later time period with an acquisition age determined from the geological history of the rock. In igneous and metamorphic rocks, this magnetic memory resides in grains with ferromagnetic properties which are usually iron-titanium oxides (titanomagnetites) comprising a few percent of the total rock volume. Although the intensity of the Earth's magnetic field is relatively weak (~ 0.5 Tesla), as these rocks cool through the Curie point they are capable of acquiring a strong and stable magnetism from this field. At high temperatures, the ferromagnetic minerals only have a transient weak diamagnetism and paramagnetism in the ambient field, but they acquire a permanent magnetization, *ferromagnetism*, as they cool through a critical temperature called the *Curie point* (generally $< 680^\circ\text{C}$). This is much lower than the temperatures at which many igneous rocks start to crystallize ($\sim 1200^\circ\text{C}$) and is lower than the higher grades of metamorphism (typically up to $\sim 800^\circ\text{C}$). If an igneous rock such as a dyke or lava flow cools rapidly at, or near to, the surface it will cool rapidly to this temperature over period of only days or weeks. It will then record a field direction which is instantaneous in geological terms; this becomes a *geomagnetic field direction* and part of the secular variation record rather than a long-term average of the field representative of the Earth's rotation axis. In contrast igneous rocks which cool slowly deep within the Earth's crust will integrate a record of the magnetic field over long periods of time; they may even record reversals of the geomagnetic field and yield a true *palaeomagnetic pole*.

Since rocks contain magnetic grains with a wide diversity of sizes and magnetic domain properties, their magnetizations are acquired continuously over a broad range of temperature below the Curie Point. This trapping of remanence from the Curie



Paleomagnetism, Measurement Techniques and Instrumentation, Fig. 1 A blocking temperature spectrum of Thermal Remanent Magnetization (TRM) showing the progressive fraction reduction of an initial magnetization M_0 . Each segment of the TRM is referred to as a

Partial Thermal Remanence (PTRM) and these segments obey a Law of Additivity: $TRM = PTRM(550-500^\circ) + PTRM(500-450^\circ) + PTRM(450-400^\circ) + \dots$

point down to ambient temperatures defines a blocking temperature spectrum (Fig. 1). In igneous rocks, the total spectrum is a *Thermal Remanent Magnetization (TRM)*. This can be considered to be the cumulative sum of *Partial Thermal Remanences (PTRMs)* obeying a Law of Additivity: the remanence fraction acquired over any particular interval is discrete and independent of the fraction acquired over any other interval (Fig. 1).

Many sedimentary rocks include grains of ferromagnetic minerals derived from the erosion of older rocks. When these grains are deposited they can be geomagnetically aligned if the influence of the Earth's magnetic field is stronger than gravitational and hydraulic forces. This is a *Depositional Remanent Magnetism (DRM)*. Usually sediments undergo a protracted history of compaction and cementing (lithification) and sometimes chemical alteration (diagenesis) and may well also be disturbed by the burrowing of organisms (bioturbation). As these processes operate, the ferromagnetic grains usually have the opportunity to become more precisely aligned into the geomagnetic field to give the rock a *Post-Depositional Remanent Magnetism (PDRM)* and define a true palaeomagnetic pole. Chemical changes resulting in the precipitation of new ferromagnetic minerals produce a *Chemical Remanent Magnetism (CRM)* which may also yield a magnetization integrated over a long period of time. In sedimentary rocks such as limestone ferromagnetic grains may be formed directly by diagenetic processes or they may incorporate magnetite grains precipitated by magnetotactic bacteria. Although these secondary processes are often still poorly understood, they operate over long time periods. They therefore always tend to integrate the geomagnetic field to provide information which is of palaeomagnetic, rather than of short-term geomagnetic, significance.

Palaeomagnetism is a time relaxation phenomenon so that all-natural magnetizations decay with time and relax into the field prevailing at the present day. Nevertheless, this relaxation time is very long, and longer than the age of the Earth for many small ferromagnetic grains. It is the very great length of time required for these natural magnetizations to decay which means that a record of permanent magnetism is still preserved, even within the oldest rocks on Earth. Nevertheless, it also means that if a rock is metamorphosed at an elevated temperature below the Curie point and this temperature is sustained for many millions of years, it is capable of being totally remagnetized. It will then acquire a new magnetism dating to the time of metamorphism.

In the late 1950s, it was also recognized that the rocks commonly acquire secondary magnetizations by a range of processes that include not only the influence of later temperature elevation, but also chemical change. The total magnetism in a rock, referred to as the *Natural Remanent Magnetization (NRM)*, is then the vector resultant of the primary magnetization acquired when the rock was formed and the secondary magnetization (which may include several components) acquired during subsequent geological times, and even imparted during coring and sampling. The chief goal of palaeomagnetic investigations is therefore to isolate the various components of remanence and ascribe an origin, age, and reliability to them. The resolved directions from each component may then be used to compute geomagnetic (short-term) or palaeomagnetic (long-term) pole positions. The following sections summarize the laboratory procedures and techniques of analysis employed in palaeomagnetism to achieve these goals.

Techniques of Sampling, Measurement and Analysis

Field Sampling

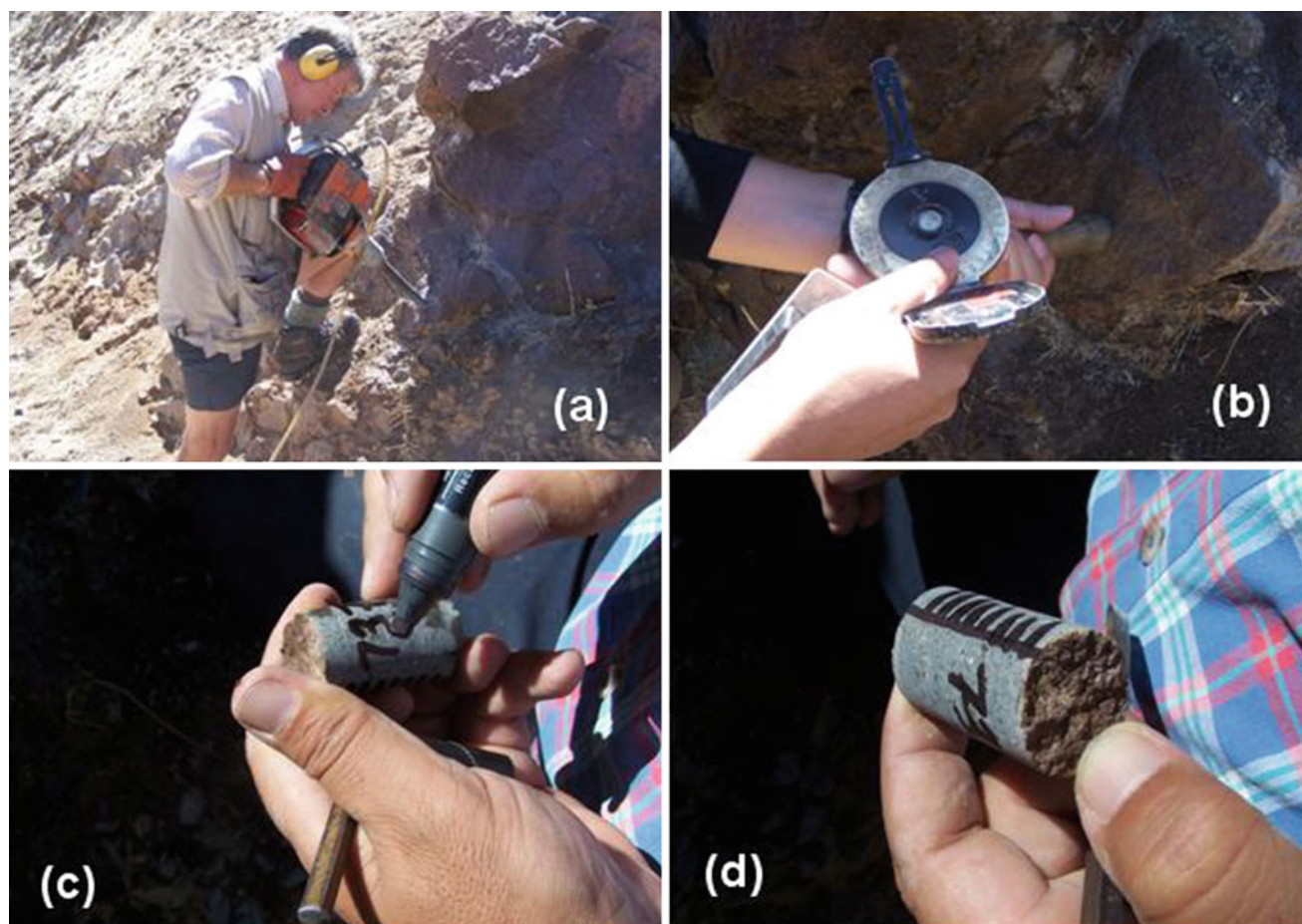
Palaeomagnetic sampling procedures follow a hierarchical approach in order to average the influence of secular variation and compensate for other inaccuracies such as the limitations of sample orientation. Sampling strategies are also adapted to the time scale of the scientific objective. Typically, oriented samples are collected from a number of different localities termed *sites*, each of which can be considered to represent a single time-instant record of the geomagnetic field. A number of such sites will be sampled within a single rock formation. The rock formation may comprise, for example, a suite of lava flows recording a volcanic episode, an igneous intrusion, or a succession of sedimentary beds. The sites clearly need to be *in situ*, cover a significant regional area, have tectonic histories which can be resolved, and collectively comprise a representative distribution through the time interval represented by the rock formation. This will ensure that although units such as lavas may record an instantaneous record of the field, the time interval covered by the survey is adequate ($\sim 10^5$ – 10^6 years) to average the effects of secular variation and any longer-term excursions of the geomagnetic field. It then derives a time-averaged assessment of the palaeomagnetic field direction. Each site will constitute a number of separately oriented samples (N) to compensate for orientation and instrumental errors and for possible failures to completely separate magnetic components during laboratory treatment. Sample orientation procedures typically have an uncertainty of $\sim 3^\circ$ but precision is gained in proportion to $N^{1/2}$ where N is the number of samples; thus, to improve orientation precision from 3° to 1° , nine or ten individually oriented samples would be required. In the case of magnetostratigraphy, where determination of the polarity of a rock level is the major objective, a minimum of three samples will determine the polarity unambiguously.

The samples are normally taken as oriented cores drilled in situ using portable hand-held drills which have been commercially available since the 1970s (Fig. 2), or may be extracted in the laboratory by drilling oriented block samples; the former method is preferred both because initial accuracy is higher and because accuracy is lost when block samples are reoriented for laboratory drilling. There are several orientation schemes currently in use, but each one employs an orientation device designed to provide an *in situ* geographic orientation using magnetic and Sun compasses (Fig. 2). Since the magnetism of the rock outcrop can influence the reading of a magnetic compass, Sun compass observations are preferred and use the direction of the shadow, together with a record of the time and sampling location; the azimuth is determined with the aid of the Astronomical Almanac or a simple algorithm. A new orientation device developed by

C. Constable and F. Vernon at the Scripps Institution of Oceanography (Lawrence et al. 2009) uses the Global Positioning System (GPS) and laser technology to produce similar accuracies to the Sun compass, although its application requires at least an additional half hour. Nonetheless, achieving sun-compass accuracy is a major breakthrough for palaeomagnetic field procedures in regions where siting of the Sun may not be possible. The palaeomagnetic specimens recovered for laboratory analysis are typically cylinders ~ 1 -inch-long and 1 inch in diameter (Fig. 2).

A palaeomagnetic field sample should be integrated with geological evidence so that it incorporates widely separated localities linked to defined rock formations for which the tectonic setting is properly understood; this will enable any effects of tectonic deformation on the recovered directions of magnetization to be accounted for. The survey will also need to accommodate effects of regional or local heating by metamorphism or igneous activity, and also the possible influence of chemical alteration, for example, by hydrothermal activity. It is also best to avoid elevated exposed locations because these tend to focus lightning. Lightning is a strong line current and the associated magnetic fields can induce a large *Isothermal Remanent Magnetization* (IRM) which is capable of completely replacing an ancient remanent magnetism in the rock.

There are a number of palaeomagnetic field tests which can be applied to help constrain the age of the magnetization in a rock formation. The *fold test* compares the directions of magnetization derived from limbs of fold with contrasting attitudes before and after adjusting for the tilt of the beds (Fig. 3); an improved grouping after adjustment indicates that the magnetization was acquired before the folding episode occurred. Sometimes the optimum grouping of directions is attained at an intermediate stage of untilting; this indicates that the magnetization was acquired during the tectonic event responsible for the folding. The *contact test* evaluates the remagnetization influence on the country rocks bordering an igneous intrusion. If the thermal influence of the intrusion diminishes with distance, it is likely that the intrusion has a primary remanence dating from initial cooling; if however, the magnetization across a widespread bordering region in the country rock is the same as the intrusion it will be apparent that the whole region, including the intrusion, has been subject to blanket overprinting. The *conglomerate test* compares the directions of magnetization in a number of pebbles in a boulder bed, preferably when these pebbles can be linked to a source in an underlying bed (Fig. 3). If the directions are random it demonstrates that pebble bed has not been remagnetized since formation; the magnetization in the bed from which the pebbles were derived is then also likely to be primary, although this would need to be independently established. If the directions of magnetization in the pebbles are uniformly directed, it is apparent that a later remagnetization event has influenced the pebble bed as well

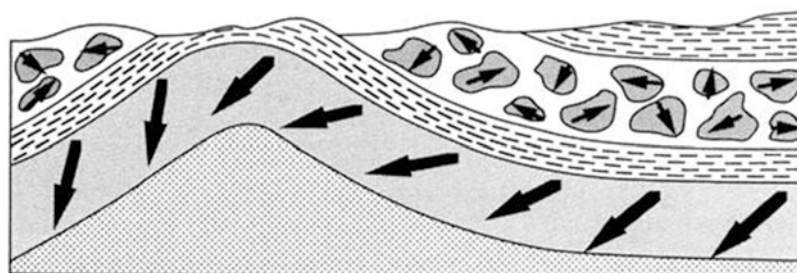


Paleomagnetism, Measurement Techniques and Instrumentation, Fig. 2 Collecting and orienting rock samples for palaeomagnetic study. The cores are drilled (a) with a portable motorized coring machine. (b) An orienting tool is inserted around the core to determine

the inclination and the azimuth by Sun and magnetic compasses. (c) The core is extracted and coded. (d) A line on the core is used to fix the extracted core with respect to the orientation readings using a directed line on the top or the bottom

Paleomagnetism, Measurement Techniques and Instrumentation,

Fig. 3 Schematic cross section of ground showing a fold and conglomerate layer with magnetic components (black arrows) illustrating the concepts behind the fold and conglomerate tests used in palaeomagnetism



as the source formation beneath. The *reversal test* compares directions of opposite polarity to evaluate whether they are statistically antiparallel and therefore likely to be the record of a reversing dipolar field. Each of these tests can be placed on a sound statistical basis using methods described in textbooks such as Tauxe (1998).

Measurement and Treatment

The Natural Remanent Magnetization of a rock sample is the vector sum of the remanent magnetic moments of all the

magnetized particles contained within it. These particles are likely to have a wide range of sizes and magnetic domain structures, so the record of primary to secondary magnetization (or even no preserved remanent magnetism at all) will vary greatly. Magnetometers measure the magnetization of rock samples in three mutually perpendicular directions (X, Y, and Z) and the values are compounded to yield the magnitude and direction of the resultant vector. Using the field orientation parameters for the sample (usually as noted above, in the form of short cylindrical cores), the data from the specimen

coordinate system can be transformed into geographic coordinates and displayed graphically on a stereonet. The geographic vector can be considered in terms of declination (D), inclination (I), and magnitude (M); M does not change during the transformation of coordinate systems.

The observed NRM of rocks may consist of any combination of primary and secondary components and it is essential to separate these to unravel the magnetic history of the rock. Since any sample will comprise multiple grain domain properties, this separation can be achieved by partial demagnetization referred to as “cleaning.” The methods rely on the observation that magnetic components of different ages and modes of acquisition usually have contrasting thermal and magnetic blocking temperature spectra. Alternating field and thermal demagnetization techniques are a routine part of palaeomagnetic investigations and depend on the relationship of relaxation time to coercivity and temperature to subtract (demagnetize) lower coercivity or blocking temperature components. Alternating field (AF) demagnetization selectively removes components with the short relaxation times that typically have lower coercivities. Thermal demagnetization is also applied progressively, this time employing the observation that grains with the lower unblocking temperatures also tend to have the lower relaxation times, τ , and are therefore more likely to carry a later secondary magnetization.

All ferromagnetic grains are subject to a number of competing internal molecular-level forces (see, for example, Butler 1992; Tauxe 1998) and larger grains tend to subdivide into zones of homogeneous magnetization in order to minimize the total magnetic energy. These *domains* are separated by narrow domain walls, typically up to 100 atoms thick, where the magnetic dipoles are canted around from one orientation to another. In magnetite (Fe_3O_4) bearing rocks, the secondary NRM tends to be carried mostly by multi-domain (MD) grains with lower relaxation times, while the primary characteristic remanence (ChRM) tends to be retained by single domain (SD) or pseudo-single domain (PSD) grains; although these latter grains have sizes larger than SD, SD properties are fixed by shape or defect properties which enables them to retain a comparable remanence to SD. MD grains have coercivities predominantly <20 mT, whereas SD and PSD grains typically have a broad range of higher coercivities. AF demagnetization can therefore selectively remove secondary NRM and leave the primary magnetization largely unaffected. Rocks containing the less common ferromagnetic minerals pyrrhotite (Fe_{1-x}S) and maghemite ($\gamma\text{Fe}_2\text{O}_3$) have properties which are closely comparable to magnetite, whereas hematite ($\alpha\text{Fe}_2\text{O}_3$), which tends to be the commonest ferromagnetic constituent in sedimentary rocks deposited in an oxygenated environment, has much higher coercivity and cannot be effectively subtracted by the fields achieved by most AF demagnetization systems. However, since hematite has a higher Curie point ($\sim 670^\circ\text{C}$) than

the other common ferromagnetic minerals, it can usually be isolated by demagnetization to temperatures above $\sim 580^\circ\text{C}$. Because hematite is also found as an oxidation product of magnetite in igneous, metamorphic, and sometimes in sedimentary rocks, primary magnetization residing in magnetite may first be unblocked below a secondary magnetization residing in hematite during progressive thermal treatment. The nature of the magnetizations resolved by palaeomagnetism is determined by rock magnetic studies. These usually involve the application of a controlled laboratory magnetism or temperature. The methods used in rock magnetic investigations are described in all the standard textbook on palaeomagnetism.

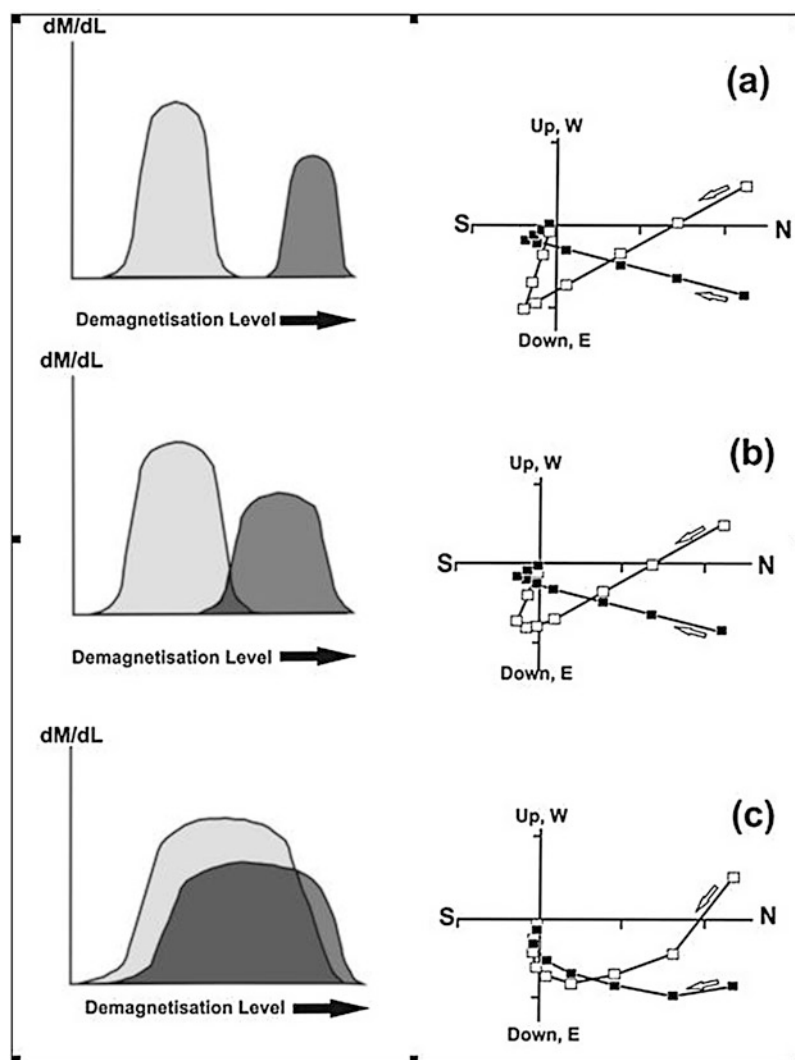
Since AF demagnetization is unable to completely subtract the magnetism in a rock because the applied alternating fields required to do this are simply unattainable in the laboratory, and would in any case present their own problems, only thermal demagnetization is capable of fully subtracting magnetization if continued beyond the Curie point. Nevertheless, while bulk AF treatment of rock collections is now amenable to automation and can permit rapid data acquisition, thermal demagnetization requires stepwise heating and cooling in a zero magnetic field, so that all magnetization blocked below the assigned temperature is then unblocked and removed. Inevitably this is a longer procedure and only amenable to limited automation. In both cases, the successive measurements of magnetization should be conducted in a field-free space so that magnetization cannot readily be imparted from the ambient field before the treated samples have been measured.

Analysis

A stereographic plot of magnetization directions at each stage of demagnetization treatment, although generally used in older studies (see Collinson 1983 for details of plotting), is not a satisfactory way of resolving the components of remanence because these directions will likely record the resultant of two or more components. Instead the results of progressive treatment are compiled as vectors in two-dimensional space by projecting the magnetization vector (i.e., with both magnitude, M, and direction resolved into D and I) onto two planes at right angles. These projections of the horizontal and vertical projection are combined together onto a single projection describing an orthogonal plot. They were first applied to a palaeomagnetic study by R.L. Wilson (1961); the display technique was formalized by Zijdeveld (1967) and has been in routine application ever since. The D, I, and M data of the progressive demagnetization experiments are most commonly resolved into north, east, and downward (or upward) components to yield a diagram like the one illustrated in Fig. 4 where it is common practice to denote projections onto the horizontal plane and vertical planes by solid and open symbols, respectively. The origin of the Cartesian coordinate

Paleomagnetism, Measurement Techniques and Instrumentation, Fig. 4

The orthogonal projection of demagnetization data during progressive demagnetization. When (a) there is no overlap between the blocking temperature/coercivity spectra of the sample, the lower and higher components are defined by discrete straight-line segments as they are unblocked by thermal demagnetization or subtracted by AF demagnetization. As the overlap of the spectra increases ((b) and (c)) the orthogonal projection becomes progressively more curved and linear segments are no longer observed when both spectra are being subtracted at the same time (c). Note that declinations can be read directly from the horizontal projection of the vector but a direct reading of inclination from the plot is not possible. (After Collinson 1983)



system is the point at which all magnetization has been subtracted; it is routinely reached with comprehensive thermal treatment, but will rarely be attained by AF demagnetization unless only very low coercivity MD grains are present in the sample.

A critical observation is that a linear trend of projected points defines the progressive removal of one component only (Fig. 4a). If this line migrates towards the origin of the projection, it demonstrates that only the high unblocking temperature or coercivity magnetization is left to be removed. The successive removal of the lower unblocking temperature or coercivity magnetizations, and the isolation of the highest unblocking temperature/coercivity component, can be achieved by the recognition of successive linear trajectories. The last component to be resolved is designated as the *Characteristic Remanent Magnetization* (ChRM) because it is usually the most interesting component for interpretive purposes and likely to be the earliest magnetization acquired by the rock, and often of primary origin; the secondary

components removed below this may be of interest for resolving the later history of the rock, especially if they can be linked to an event such as a thermal pulse or a diagenetic episode. The direction of any component represented by a straight-line segment on an orthogonal plot is resolved by extracting the component from the data above and below, and treating it as if it were a unique record.

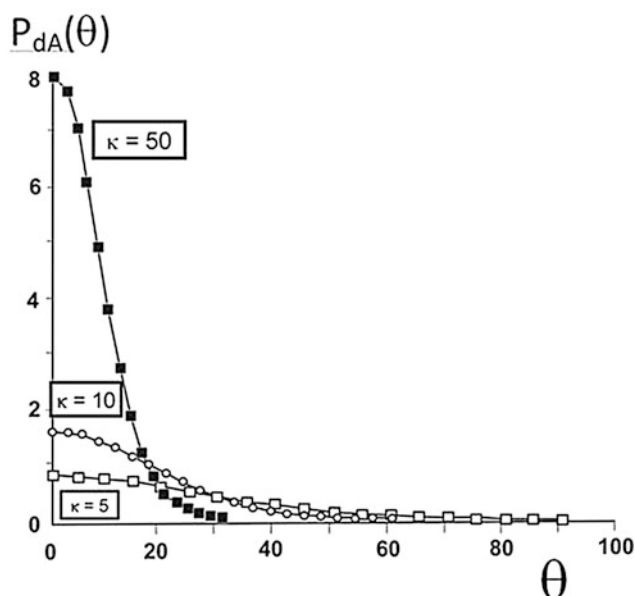
The trajectories of vector projections depend upon the degrees of overlap of the coercivity or blocking temperature spectra (Fig. 4). When the spectra do not overlap, the vector diagram displays successive linear trajectories, each one recording the subtraction of a component. If overlap is only partial, straight line trajectories are still identified but intermediate points corresponding to the overlap display a curved trajectory (Fig. 4b). If there is complete overlap between the spectra, the vector component diagram does not possess linear segments (Fig. 4c), although the trajectories may sometimes be used to isolate the components if they define a great circle, a "circle of remagnetization." A single great circle on its own

provides insufficient information to estimate the ChRM direction, but if more than one circle is available from different specimens and these circles converge, then the convergence point, or its antiparallel equivalent, can yield an estimate of the ChRM. Halls (1979) reports a procedure for calculating the best fitting convergence point of several great circles. The majority of convincing palaeomagnetic results come from one and two component structures, although palaeomagnetic workers have now ventured into the study of rocks with varied histories comprising multicomponent NRMs, with combinations of AF and thermal demagnetization used to isolate complex remanence histories.

The procedure for separating and resolving the directions of individual components is known as *Principal Component Analysis* (PCA) and comprises a method for determining the direction of the lines or planes which best-fit the data described by the orthogonal projections (Kirschvink 1980); essentially a component segment is extracted from the orthogonal projection and isolated as though it is the only component present in the rock. An advantage of this method is that it compensates for errors due to signal to noise ratio, and provides a quantitative measure of the precision of the direction. The precision of the direction determined from the best-fit line is described in terms of a *Maximum Angular Deviation* (MAD), while MAD values $<12^\circ$ are usually considered acceptable, much higher degrees of precision should be aimed for, especially in youthful rock material.

Statistics

As with all experimental data, the directions resolved by PCA will be associated with a number of random errors arising from factors such as instrument noise, inaccuracies in sample orientation, imperfections in demagnetization, and secular variation of the recorded geomagnetic field. The net result of all these factors is a dispersion from which the researcher requires the best estimate of the mean direction. Determining a mean and confidence limit from any set of values by a statistical method is based on a probability density function. A function applicable to palaeomagnetic directions was developed by Fisher (1953) and provides a simple procedure for treating palaeomagnetic data. Fisher's analysis treats each accepted direction in the population as having a unit weight and represents it as a point on a sphere of unit radius. The dispersion of these points on the sphere is assumed to have a circular Gaussian distribution (Fig. 5) and the mean of the set of directions is calculated by vector addition: each unit vector is resolved into three orthogonal components in Cartesian coordinates. The components are summed to obtain the length of the resultant vector (R) and the mean direction (D and I). D and I define the direction of the best estimate of the mean, while R is used to assess confidence in the result.



Paleomagnetism, Measurement Techniques and Instrumentation, Fig. 5 The Fisher circular distribution derived from the Gaussian probability function. $P_{dA}(\theta)$ is the probability per unit angular area of finding a direction within an angular area, dA , centered at an angle θ from the true mean. Distributions are shown for values of the precision parameter κ , of 50, 10, and 5

Two simple parameters, k and α_{95} , are used to illustrate the quality of the calculated mean in palaeomagnetic studies. The probability of finding a direction at an angle (θ) from the calculated mean is expressed in terms of a probability function involving a precision parameter κ . Fisherian statistics use an approximation for κ defined by a parameter denoted by $k = (N - 1)/(N - R)$ where N is the number of directions and R is the length of the resultant vector. The α_{95} value in degrees defines the angular radius of the cone of confidence projected from the center of the unit sphere within which we have 95% confidence that the true direction actually lies. It is simply calculated from $\alpha_{95} = 140/(k \cdot N)^{1/2}$ and is comparable to the standard estimate error of the mean in Gaussian statistics; 5% of the observed directions are expected to fall outside of this circle. High reliability in palaeomagnetic analysis is therefore expressed by large k and small α_{95} . As with all statistical analysis, precision estimates are improved by increasing the sample population although in terms of experimental effort the increase becomes progressively less worthwhile for large values of N . Sample populations from a single unit such as a lava flow are generally considered satisfactory if N is 7 or more, whilst k values <10 are unsatisfactory.

Determination of Palaeomagnetic Poles

The resolved mean direction of magnetization is used directly in much analysis and interpretation – in regional small to

medium scale tectonic problems, for example. However, for larger scale analysis such as the determination of past global geometries and analysis of the geomagnetic field, it is necessary to convert it into an equivalent pole position. This also provides a meaningful way of comparing ChRM directions from rock formations of same age, but widely separated on the same continent or tectonic plate; for this purpose, a common reference system is necessary and is provided by the palaeomagnetic pole. The pole position is the position where the geomagnetic field axis responsible for the observed palaeomagnetic field direction meets the surface of the Earth. A model is required to perform this calculation. Since more than 90% of the present geomagnetic field can be explained in terms of an axial geocentric dipole source, this is the model normally used to calculate palaeomagnetic poles and the position is expressed in the present geographic longitude–latitude grid. For a palaeomagnetic field of inclination I , the ancient latitude (λ) of the site is given by the formula $\tan I = 2 \tan \lambda$. The equations required to calculate the position of the palaeomagnetic pole require the declination D and the coordinates of the sample locality in addition to the inclination; they are more complex but are provided in all palaeomagnetic text books (see for example, Tauxe 1998).

The geomagnetic pole calculated from a single geomagnetic field observation (from a single lava flow for example) yields a *Virtual Geomagnetic Pole* (VGP) and will include components of the nondipole field incorporating the secular variation. A palaeomagnetic pole is calculated by averaging a number of VGPs derived, for example, from several lava flows or levels through a sedimentary section; it is expected to have averaged nondipole field components and represent the Earth's rotation axis at the time that the ChRM was acquired by the rocks. The angular dispersion of the VGPs constituting the collection can be used to test whether this is effectively achieved by comparison with the dispersions recorded by secular variation during historic and recent geological times (McFadden et al. 1991). The expected angular dispersion is a function of latitude, although the comparison is presently of limited value because the comparative database is sparse in the southern hemisphere and temporal coverage is poor. Confidence limits for a palaeomagnetic pole are derived from α_{95} , the radius of the cone of 95% confidence about the mean direction. These translate into an oval of confidence within which we have 95% confidence that the pole actually lies. The maximum and minimum radii of this oval are given by $dp = 2 \alpha_{95} (1/(1 + 3\cos^2 I))$ and $dm = \alpha_{95} (\sin p/\cos I)$ where p is the magnetic colatitude or the great-circle distance from the sampling location to the pole.

Instrumentation

The intensity of magnetization of rocks can vary by several orders of magnitude. Young basic lavas are often so strong

that their magnetizations can often be approximately determined with a hand compass. At the other extreme rocks such as pure limestones are so weak that it is only the presence of sparse magnetotactic bacteria that may give them a very weak remanence; this is only detectable with sophisticated cryogenic magnetometers developed in recent decades. The early studies in the 1950s and 1960s were inevitably concentrated on the more strongly magnetized igneous rocks and red sediments. The pioneering instrument for this was the parastatic magnetometer developed by Blackett (1952) which employed the deflection of a suspended unstable astatic magnet system. Samples were brought into proximity to disturb this system; they would be individually measured in three orthogonal directions, each in antiparallel configurations yielding six measurements in all, and used to calculate a resultant direction. Each measurement required a time integration for the deflection to stabilize and was inevitably relatively slow with the instrument often subject to local vibrations and fluctuations in the ambient field. Vibration and magnetic noise issues were overcome and measurement speed enhanced with the development of spinner magnetometers. These were amenable to successive improvements; with the speed of rotation of the sample near the detecting system enhanced, sensitivity could be improved and this expanded the range of rock types available for investigation, notably into a broader range of weakly magnetized sediments. High speed spinner magnetometers can achieve sensitivities approaching superconducting magnetometers but they are not amenable to the treatment of more friable rocks. Superconducting quantum interference magnetometers (SQUIDs) are now most widely used to evaluate remanence in more weakly magnetized rock types.

Spinner Magnetometers

Spinner magnetometers operate in accordance with Faraday's law whereby the magnetic field averaged along an axis of a spinning sample generates an electric current in a detection coil, or sensor, to produce an electromotive force (EMF). The sample will typically be spun within, or adjacent to, a fluxgate detecting system. The amplitude of the output EMF is proportional to the component of magnetic moment perpendicular to the rotation axis ($P_{\perp}$), and the phase of the voltage is utilized to relate the direction of the measured component to a reference direction in the sample. The sample is rotated successively about three axes to obtain average values of the NRM components and reduce the effect of inhomogeneity to $\sim 2^\circ$ in direction and $\sim \pm 5\%$ in intensity. Measurement of weak magnetizations has been achieved by increasing the rotation speed and careful design of the pick-up coil systems. Since electric motors operating at a frequency of ~ 100 Hz are required for rotating the samples, a challenge to the design has been the requirement to shield this unavoidable source of

electromagnetic noise from the measuring assembly. The problem can be reduced by using an air turbine and in modern magnetometers, the frequencies used are in the range 5–80 Hz with signal discrimination circuits employed to reduce the effects of noise.

The usual procedure for NRM measurement is to spin each sample in six mutually orthogonal orientations so that any two of X, Y, and Z are in turn perpendicular to the spin axis. Since two orthogonal components of magnetization are obtained from each spin, the orientations are chosen to provide four determinations each of X, Y, and Z, usually two of each sign for each component. This procedure tends to average out any random noise on the signals, the effects of inhomogeneity of NRM, and any remanent magnetism in the part of the sample holder permanently attached to the shaft. The six sample orientations used should ideally satisfy the following: (a) provide two values each of +X, +Y, and +Z, (b) distribute the displaced dipole caused by inhomogeneity of the NRM as uniformly as possible, (c) average the possible error arising from imperfect calibration of the NRM direction of the magnetometer, and (d) measure each NRM component in each “channel” of the magnetometer (i.e., the sample will be placed in the holder with X, Y, and Z in turn parallel and perpendicular to the reference direction).

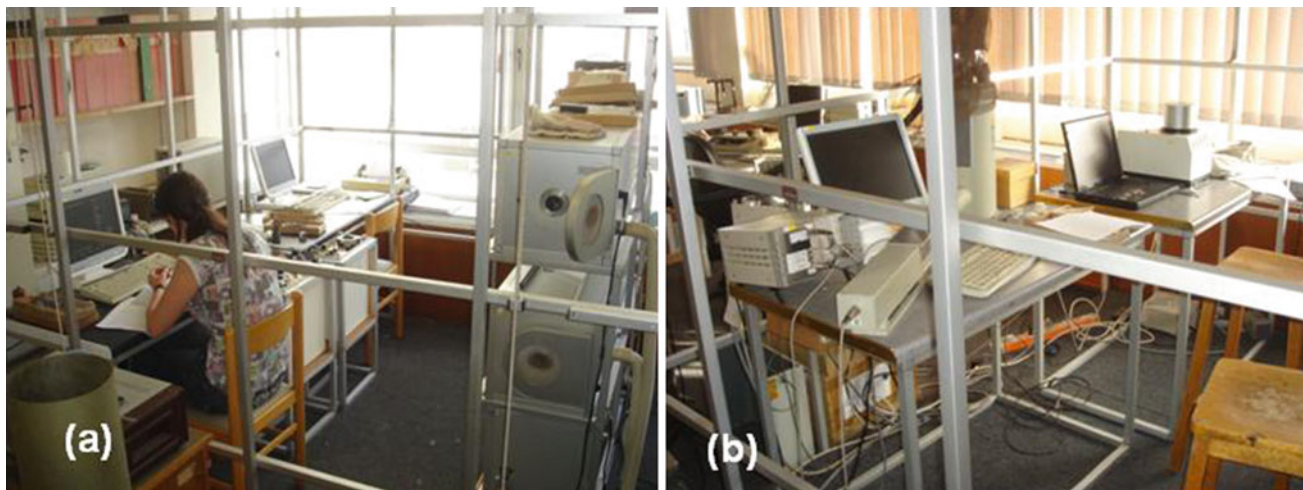
Some commercial spinner magnetometers (Fig. 6b) can now achieve sensitivities close to those attainable with SQUID magnetometers and supplied with software for online analysis of the data, although due to their need to run at high speed, they are unsuitable for friable or poorly shaped samples. The dual speed JR-6 (AGICO Ltd; Czech) series spinner magnetometers have sensitivities of $\sim 2.4 \times 10^{-6}$ A/m and can measure up to 12,500 A/m. A much cheaper compact spinner

magnetometer, the “Minispin” (MOLSPIN Ltd; UK), is no longer available but its modified version “Portable Spinner Magnetometer” suitable for the routine measurement of more strongly magnetized samples came onto the market recently.

Cryogenic Magnetometers

Developments in cryogenic technology have allowed two important advances in palaeomagnetic research. Firstly, they have permitted an increase in sensitivity by one to three orders of magnitude which, combined with rapidity of measurement, allows an unlimited range of rock types to be studied. Secondly, they have allowed magnetism in a sample to be measured by a single insertion into the sensing coils because only one axis of the sample needs to be presented to the magnetometer. Magnetic and thermal cleaning can then potentially be automated with the measuring procedure. Cryogenic magnetometers use the properties of a superconducting ring made of a metal with perfect diamagnetism which causes the expulsion of all magnetic flux from the ring below a certain temperature. When an external field is applied in the superconducting state, all flux is expelled from the interior of the metal but not from the hole within the ring; as a result, induced currents flow in one direction on the outer face of the ring and in the opposite direction on the inner face. If the applied field is removed, the currents on the outside of the ring disappear, while the current on the inside face persists to maintain a field within the ring.

The ability of the ring to trap and maintain magnetic fields to better than one part in 10^9 can be used in magnetometry to trap a zero field within the loop and use this volume for



Paleomagnetism, Measurement Techniques and Instrumentation, Fig. 6 Instrumentation in a typical palaeomagnetic laboratory. Measurements and demagnetization procedures take place within large sets of Rubens coils designed to continuously nullify the ambient geomagnetic field. In (a) two thermal demagnetizers are stacked on the right; these are designed for the treatment of trays of multiple cores to specified

temperatures for programmed time intervals. On the left the operator is using one of two interfaced spinner magnetometers for measurement of relatively strongly magnetized rock samples. (b) Two magnetometers used for the measurement of weakly magnetized rock samples comprise a FIT hydrogen SQUID magnetometer on the left and a JR6 high speed spinner magnetometer on the right

measurement. These magnetometers are operated at the superconducting-resistive boundary; the ring current is quantized so any flux entering the ring at this transition state is also quantized and measured in terms of a feed-back current. The magnetometer is surrounded by an insulating evacuated space cooled by liquid helium or nitrogen, while pick-up coils connected to SQUID detectors are employed to measure the NRM along one axis, or simultaneously along two or three mutually perpendicular axes. Mumetal shields are employed to blanket external field variations and some automated systems allow complete demagnetization within this field-free environment. Cryogenic magnetometers are faster and typically more sensitive than spinner magnetometers but cost much more to purchase and operate; the liquid nitrogen represents an ongoing cost, and while helium can achieve lower temperatures, it is more expensive and less commonly used these days. For further details, the reader is referred to Collinson (1983) and Tauxe (1998).

Alternating Field Demagnetizers

The AF demagnetization procedure applies an oscillating field to the specimen in a nullified magnetic field environment. The equipment comprises a coil, a capacitor bank for tuning the coil, a power supply, and current reduction mechanism; a coil system or magnetic shield aims to minimize the ambient field over the sample and sample tumbler. The rock specimen is rotated within the center of the coil in a tumbling system incorporating simultaneous rotation about horizontal and vertical axes. The ratio of the two tumbling rates is designed to expose as many configurations of the sample as possible to the applied field acting along the axis. The applied current is then smoothly reduced to zero as the specimen continues to rotate. It is essential to ensure that the geomagnetic field within the demagnetizer is completely cancelled or the specimen may acquire an *Anhyseretic Remanent*

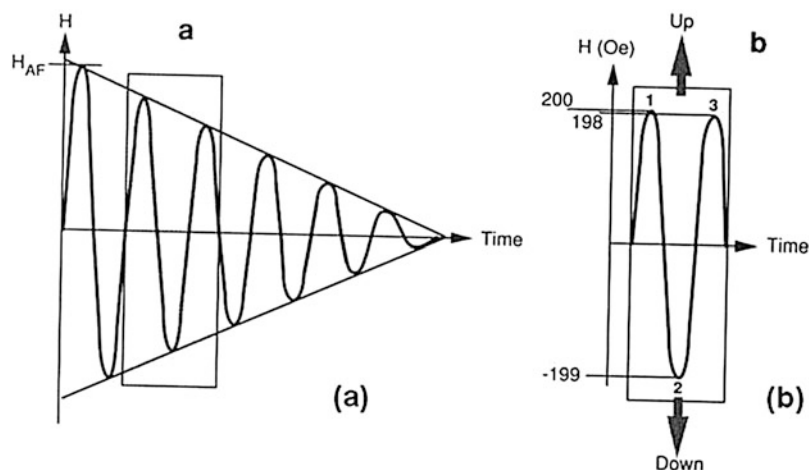
Magnetization (ARM), and an alternating field with a very pure waveform is essential for the best results to avoid rotational and gyro magnetic effects. Most modern AF demagnetizers are designed to achieve initial fields in excess of 100 mT.

The theory of AF demagnetization is explained by Butler (1992). The waveform of the applied alternate magnetic field is a sinusoid with linear decrease in magnitude with time (Fig. 7). The expanded figure (b) emphasizes the smooth decrease required. All the grain moments with coercivities below the peak AF will track the field. These entrained moments will become stranded however, as the peak field drops below the coercivities of individual magnetic domains. The directions of stranding will be effectively random if the tumbler has been appropriately designed to present as many configurations as possible of the sample to the declining applied field. By progressively increasing the applied field, the domains of successively higher coercive force are randomized or subtracted from the total measured remanence. The coercive spectra of the sample can then be resolved up to the maximum field of demagnetization and contributions of successive components to the NRM resolved by application of orthogonal projections and PCA. As noted above, AF demagnetization is unable to fully subtract the NRM and is ineffectual with hematite-bearing rocks due to their high coercivity.

Commercially available alternating field demagnetizers include the D-2000-DTECH AF demagnetizer and the AGICO-LDA5 AF demagnetizer systems. Both are in wide use and can demagnetize samples up to a peak field of 200 mT with multiple field decay rates using both manual and tumbling facilities. The D-2000 can demagnetize five to six samples simultaneously at any chosen peak field, and both units can be used for rock magnetic experiments by imparting complete or partial anhysteretic magnetizations (ARM and PARM) using a controlled direct current applied continuously across the AF demagnetizer coil.

Paleomagnetism, Measurement Techniques and Instrumentation, Fig. 7

(a) The generalized wave form of the magnetic field used during AF demagnetization and (b) expanded version of a small portion of AF demagnetization between two successive peaks. (After Butler 1992)



Thermal Demagnetization

Thermal demagnetization is achieved by demagnetizers that serve the purpose of both demagnetizing the NRM in palaeomagnetic specimens and applying a partial thermal remanent magnetization (pTRM); the latter procedure is used for investigating the record of the ancient intensity (palaeointensity) preserved in suitable rocks. If the temperature of the sample is raised until the relaxation time of some fraction of the particles carrying NRM has been reduced to a few minutes or less, the NRM of these particles will be unblocked. Their contribution to the NRM of the rock will be lost if they are cooled to the ambient temperature in a zero field. In this way, the contribution of lower stability grains to the NRM can be randomized. Alternatively, if a DC field is applied during cooling, the grains whose unblocking temperatures have been exceeded will be realigned into the new applied field direction; they will then have acquired a pTRM. If the rock sample is subjected to successively higher temperature treatments, the lower blocking temperature components will be progressively removed. The components contributing to the NRM are then progressively subtracted and can be resolved by projecting the successive magnetizations onto an orthogonal projection and applying PCA.

Thermal demagnetizers comprise a magnetic shielding system to exclude the ambient field, a furnace, a thermocouple, and temperature control unit. A prime concern is the neutralization of the external magnetic field. To balance the times for measurement against the times taken for heating and cooling, the oven needs to be of sufficient size to hold a reasonable number of samples (10–50) and apply a uniform temperature through them ($<\pm 10^\circ\text{C}$ at 600°C). It is required to be capable of heating samples to a temperature in excess of the highest Curie points of the ferromagnetic minerals ($\sim 700^\circ\text{C}$) over a modest time (~ 1 h). The heating coil is wound onto the curved surface of a silica tube forming the inside wall of the furnace in a noninductive way to reduce stray magnetic fields. The maximum available voltage will normally be that of the electrical mains. The maximum current is dictated by the temperature controller using a variable transformer determining the resistance of the furnace winding. Between 20°C and 700°C the resistance of the winding increases by 5–10%. If the resistance determined from the available voltage and current is the cold resistance, the current (and heating rate) will decrease at high temperatures; if it is the hot resistance then the cold current will be slightly above that calculated. The wire diameter is controlled by the size of the furnace, while the number of turns and their minimum separation is set by the requirement for adequate insulation between them; as separation is increased, the uniformity of temperature deteriorates.

Although large furnaces are useful for dealing with many samples at once, their thermal inertia is considerable and may cause problems with temperature control. A suitable

thermocouple material is Pt/13%Pt-Rh, with an output of ~ 6.7 mV at 700°C . The thermocouple tip may be covered with a standard rock sample with an axial hole drilled in the center. The controller then senses the furnace temperature via a thermocouple placed close to the furnace winding. By ensuring that the thermal inertia of rock and thermocouple are similar, the rock samples should be uniformly heated throughout the indicated temperature. Several commercial temperature controllers are available for automatically raising and holding the furnace temperature to a preselected value.

Some furnaces are designed so that it is possible to conduct heating/cooling cycles in an inert atmosphere or in a vacuum to minimize oxidation. Although thermal demagnetizers are in principle, capable of completely removing the NRM in a rock and thereby recovering the entire component structure in a way that is not possible with AF demagnetization, this is seldom fully achievable in practice. The limitations are the difficulty of completely cancelling the ambient magnetic field without excessive shielding and the possibility of chemical alteration during the successive heating and cooling cycles. To surmount the latter problem, microwave systems have been developed to provide an alternative source of heating. These have been successfully applied to palaeointensity studies in a small number of laboratories (Walton et al. 1992), but have not yet been produced with a capability of routine demagnetization of NRMs.

The MMTD Thermal demagnetizer (Magnetic Measurements; UK) and the TSD1 Thermal Specimen Demagnetizer (Schonstedt; USA) are the two popular Commercial units in common use today. Both achieve a very low magnetic field environment (< 10 nT) and can attain furnace temperatures up to 800°C . The new generation MMTD demagnetizer is capable of handling a large number (~ 80) of samples. In the TSD1, both heating and cooling chambers are separately arranged in a coaxial manner. Although only handling about ten samples in a batch, cooling of one set can accompany heating of the next batch and achieve relatively rapid processing. In the MMTD successive heating and cooling of a single but large batch takes place in the same chamber.

A Perspective on Seventy Years of Palaeomagnetism

Seven decades have now passed since rock samples were first introduced to sensitive suspended magnet systems to evaluate their magnetism. Following pioneer work in Japan and the UK, the 1960s saw a rapid extension of studies to the USA, Australia, Africa (former Rhodesia), Holland, and the former USSR (extensive results but only incompletely tested and reported). The intervening time has seen palaeomagnetic laboratories established in most countries of the world, specifically to address the palaeomagnetic record in their geological

successions. Great technological improvements in the equipment used for these studies have accompanied developments in the field of electronics. The prime emphasis has been on expanding the numbers of samples that can be measured and speeding up the treatment by improved AF and Thermal Demagnetizers. While the original equipment was usually built in-house, an impressive range of commercially available equipment is now able to cater for the expanded world market. It is now possible to conveniently evaluate component structures of magnetizations in more detail than was possible in times past. Although the conclusions may not be substantially altered, this provides a strong motivation for revisiting earlier studies where component separation may well have been incomplete.

We illustrate modern developments with two examples. Wack and Gilder (2012) describe an automated system for palaeomagnetic investigations. Based on a migrating SushiBar system, up to 99 samples can be rapidly loaded and unloaded, and AF demagnetized. The progressive treatment system also enables the researcher to examine magnetic properties using an applied anhysteretic remanent magnetization (ARM) through any coercivity window up to the peak field. This paper also compares the properties of the SQUID systems developed since 1984. As the emphasis on palaeomagnetic studies has diminished somewhat in recent years, the studies of palaeointensity have expanded as we seek to understand more about the long-term behavior of the geodynamo. A recent development is reported by Fukuma and Kono (2016) who describe an automated system which couples a magnetometer with an electric furnace controlled by software that ensures an efficient and identical double heating of samples. This is required by the Thellier method (1959) which likely remains the most reliable method for determining ancient intensities of the magnetic field.

Summary

A record of the past magnetic field, known as magnetic remanence, is preserved by ferromagnetic mineral grains in rocks and other natural materials. These minerals are mostly oxides of iron (magnetite, hematite, maghemite) and the sulfide mineral pyrrhotite. Although they usually comprise no more than a few percent of the total rock volume, they provide key information in the disciplines of geology and geophysics and have further applications in the environmental sciences and archaeology. The material for study is collected in the form of small oriented samples in sufficient numbers for robust statistical analysis and in distributed amounts to be representative of a sample of a fossil magnetic field. The record can apply to short periods of time in the study of secular variation, short term geomagnetic field fluctuations, and the record in archaeological samples such as thrown

bricks, heated pottery, and kiln material. Alternatively, it can apply to long periods of time when applied to determining pole positions for rock formations spanning millions of years to evaluate past continental motions, or to establish polarity and link it to the Geomagnetic Polarity Time Scale. The magnetizations are measured using magnetometers designed to determine natural magnetizations ranging from strong to very weak. The instruments used for determining these magnetizations nowadays are either spinner magnetometers which measure the current induced in a detector system as the sample is spun near to a coil, or SQUID magnetometers based on cryogenic technology. The latter systems are more rapid and versatile but are expensive to operate and beyond the scope of many laboratories to afford.

The magnetizations in most materials are composite effects comprising the vector sum of a primary magnetization often imparted when the rock was formed, and secondary magnetizations acquired during later events that may have influenced the rock. These component magnetizations are separated and resolved by demagnetization procedures that progressively subtract the magnetization resolved either in terms of the coercivity spectrum of the magnetization (AF demagnetization) or the blocking temperature spectrum (thermal demagnetization). The magnetizations are resolved by projecting the successive values of the magnetization vector during this progressive treatment onto a two-dimensional orthogonal plot and identifying the constituent components from their contributions to the coercivity/blocking temperature spectra. The populations of magnetic directions resolved from a palaeomagnetic study are statistically analyzed using a system of (Fisherian) statistics assuming that each direction is of equal importance, and that the population can be represented in terms of a circular Gaussian distribution. Two simple parameters comprising a precision parameter, k , and a cone of 95% confidence α_{95} are used to express the quality of the result. The mean direction is normally converted into an equivalent geomagnetic or palaeomagnetic pole by assuming that the source of the field direction resolved at the surface of the Earth originates in a geocentric dipole source within the Earth's core. In the long-term application spanning at least thousands of years, this dipole may also be assumed to be axial and constrained to the rotation axis of the Earth.

Cross-References

- [Archaeomagnetism](#)
- [Curie Temperature](#)
- [Geodynamics](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Paleomagnetism, Magnetostratigraphy](#)
- [Plate Tectonics, Precambrian](#)

Bibliography

- Blackett PMS (1952) A negative experiment relating to magnetism and the Earth's rotation. *Phil Trans R Soc Lond A* 245:309–370
- Butler RF (1992) *Paleomagnetism: magnetic domains to geologic terranes*. Blackwell Scientific Publications, Boston, 319p
- Collinson DW (1983) *Methods in rock magnetism and paleomagnetism*. Chapman and Hall, London, 500p
- Fisher RA (1953) Dispersion on a sphere. *Proc R Soc London Ser A* 217:295–305
- Fukuma K, Kono M (2016) A LabVIEW software for Thellier paleo intensity measurements with an automated three-component spinner magnetometer TSpin. *Earth Planets Space* 68:43. <https://doi.org/10.1186/s40623-016-0424-2>
- Halls HC (1979) Separation of multicomponent NRM: combined use of difference and resultant magnetization vectors. *Earth Planet Sci Lett* 43:303–308
- Kirschvink JL (1980) The least-squares line and plane and the analysis of paleomagnetic data. *Geophys J R Astron Soc* 62:699–718
- Lawrence KP, Tauxe L, Staudigel H, Constable C, Koppers A, McIntosh WC, Johnson CL (2009) Paleomagnetic field properties at high southern latitude. *Geochem Geophys Geosyst* 10. <https://doi.org/10.1029/2008GC002072>
- Martin-Hernandez F, Luneburg CM, Aubourg C, Jackson M (2004) *Magnetic fabric: methods and applications*. Geological Society of London special publication no 238. Geological Society, London, 551pp
- McFadden PL, Merrill RT, McElhinny MW, Lee S (1991) Reversals of the Earth's magnetic field and temporal variations of the dynamo families. *J Geophys Res* 96:3023–3933
- Tauxe L (1998) *Paleomagnetic principles and practise*. Kluwer Academic Publishers, Dordrecht, 299pp
- Thellier E, Thellier O (1959) Sur l'intensité du champ magnétique terrestre dans le passé historique et géologique. *Ann Geophys* 15:285–376
- Wack MR, Gilder SA (2012) The SushiBar: an automated system for palaeomagnetic investigations. *Geochem Geophys Geosyst* 13(12). <https://doi.org/10.1029/2011GC003985>
- Walton D, Shaw J, Share JA, Hakes J (1992) Microwave demagnetization. *J Appl Phys* 71:1549–1551
- Wilson RL (1961) Palaeomagnetism in Northern Ireland. Part 1. The thermal demagnetization of natural magnetic moments in rocks. *Geophys J R Astr Soc* 5:45–58
- Zijderveld JDA (1967) A.C. demagnetization of rocks: analysis of results. In: Collinson D, Creer K, Runcorn S (eds) *Methods in paleomagnetism*. Elsevier, Amsterdam, pp 254–286t

Paleomagnetism, Polar Wander

Jean Besse¹, Vincent Courtillot² and Marianne Greff²

¹Laboratoire de Paléomagnétisme, Institut de Physique du Globe de Paris, Paris, France

²Institut de Physique du Globe de Paris, Paris, France

Definition and Introduction

For more than two centuries, geoscientists have suggested, based on geological field observations (reconstruction of paleoclimate belts based on fossils or certain rock types),

that the past Earth's equator (and Equatorial conditions) must have at some time been located far from its present position. In the early 1950s, paleomagnetists such as Runcorn (1956) provided quantitative evidence that the instantaneous geographic or rotation pole had moved with respect to certain continents. The paths followed by the poles in the geological past were termed “apparent” polar wander paths (APWPs), because it was not clear whether it was the pole or the continent that had moved. Since then, we have learned that oceanic and continental plates have moved with respect to each other and that a significant part of APW was actually due to these relative motions. A remaining fraction in polar wander, which would be a characteristic of “Earth as a whole” and which would not be accounted for by plate tectonics, is called “true” polar wander (TPW). True polar wander refers thus to the large-scale motions of the Earth's rotation axis through geological time. External forces driving shorter-term changes such as precession are not taken into account here.

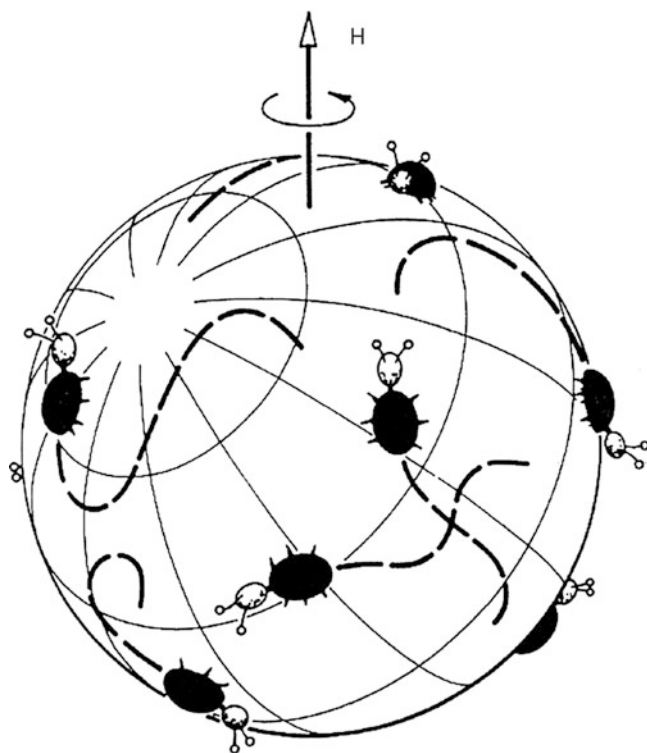
True polar wander could result from centrifugal forces acting on mass anomalies distributed on or inside the Earth: excess mass, for instance, will lead to slow deformation of the Earth and change its rotation axis, causing the mass to move towards the equator without being displaced with respect to the solid Earth. The problem is to define precisely in a deformable Earth which part wanders with respect to which reference frame. What is defined (e.g., Steinberger and O'Connell 2002) is the axis of rotation of a reference frame (RF) relative to an inertial frame of reference (i.e., determined by the angular momentum vector of Earth). The reference frame can be selected as the one that has zero net rotation when motions are integrated over the entire mantle, called the “mean mantle” reference frame. Or one can also define a “mean lithosphere” frame, in which the lithospheric plates have no net rotation. These frames in general do not coincide (Ricard et al. 1991; O'Connell et al. 1991). The “mean mantle” frame can be computed in numerical simulations in which the density distribution is given over the entire mantle. However, if one attempts to constrain it with actual paleomagnetic observations, one has to assume that the “mean mantle” frame is identical to some frame based on observations: the “hotspot reference frame” will most often be the one involved. Steinberger and O'Connell (2000) show that the approximation (discussed later in this paper) is quite good even if hotspots move in the convecting mantle.

The crust and mantle actually deform actively at the typical velocities of plate tectonics (~10–100 mm/year). Thus, polar wander involves (relative) motions of the rotation axis with respect to the Earth's mantle.

Goldreich and Toomre (1969) hypothesized that large-scale polar wandering and continental drift shared a common explanation through mantle convection and the resulting redistribution of masses within the mantle. Goldreich and Toomre (1969) started their analysis of the motions of the

rotation axis of a quasi-rigid Earth with the classical figure, where a colony of beetles (the continents or other mass anomalies) moves slowly at the surface of a rigid rotating sphere (Fig. 1). They simulated very large polar wander, with particularly intense and rapid swings.

The physics of TPW is explained among others by Ricard et al. (1993a): TPW is seen as the result of conservation of angular momentum of a rotating, deformable body, whose inertia tensor slowly changes with time. The reasons for the changes are mass redistributions linked to mantle convection. In simple terms, the inertia tensor component that affects polar wander is related to subducting slabs, upwelling plumes, and more generally thermal and chemical mass heterogeneities resulting from a long history of mantle convection. With estimates of degree 2 components of the geopotential derived from satellite data, it can be shown that the non-hydrostatic part of the Earth's inertia ellipsoid is triaxial, the present rotation axis coinciding with the largest of the non-hydrostatic moments of inertia. The rotation axis tracks the maximum principal inertia axis (PIA) of the non-hydrostatic inertia tensor as fast as adjustment of the hydrostatic bulge allows.



Paleomagnetism, Polar Wander, Fig. 1 Diagram illustrating polar wander on a quasi-rigid planet (After Goldreich and Toomre 1969). The beetles represent mass anomalies that change with time (drifting plates or mantle heterogeneities), inducing changes in the principal non-hydrostatic moments of inertia. The coordinate system attached to the (slowly deforming) solid planet moves with respect to the rotation (or spin) axis

Measure of True Polar Wander

The position of the rotation axis depends on the time scale. According to astronomers and geodesists, polar wandering is occurring at present. For instance, a determination by Gross and Vondrak (1999) uses the star catalog established from satellites during the past century. The result is a mean drift of the Earth's rotation pole over the period 1900–1992 of 107 mm/year along the direction of the 79°W meridian.

It has also been proposed that this long-term polar wander is largely due to post-glacial rebound, i.e., lithospheric and mantle response to melting of ice (e.g., Sabadini and Peltier 1981; Mitrovica and Milne 1998). The pole is moving towards Hudson Bay, where a large ice mass melted and the lithosphere is currently rising.

On time scales longer than 10^6 years, TPW most likely reflects convection of the whole mantle expressed at the surface by horizontal plate motions Steinberger and O'Connell (2002), and paleomagnetism has become the primary source of information. The apparent polar wander paths from all the major plates are being defined with increasing accuracy, and their differences reflect continental drift and plate tectonics. This demonstration of the reality of "apparent" polar wander seen from any plate has led some to propose that the concept of "true" polar wandering might be irrelevant or unnecessary. Are remaining fractions in polar wander, which would be a characteristic of "Earth as a whole," and which would not be accounted for by plate tectonics, actually resolvable? The concept is worthwhile if a reference frame external to the plates themselves can be defined, in which there was significant motion of the pole. Most studies performed in the last two decades followed an approach, based on the possibility, originally suggested by Morgan (1972, 1981), that hotspots could provide a valid reference frame for the mantle. Its meaning and existence has been the subject of continued debate, up to this day.

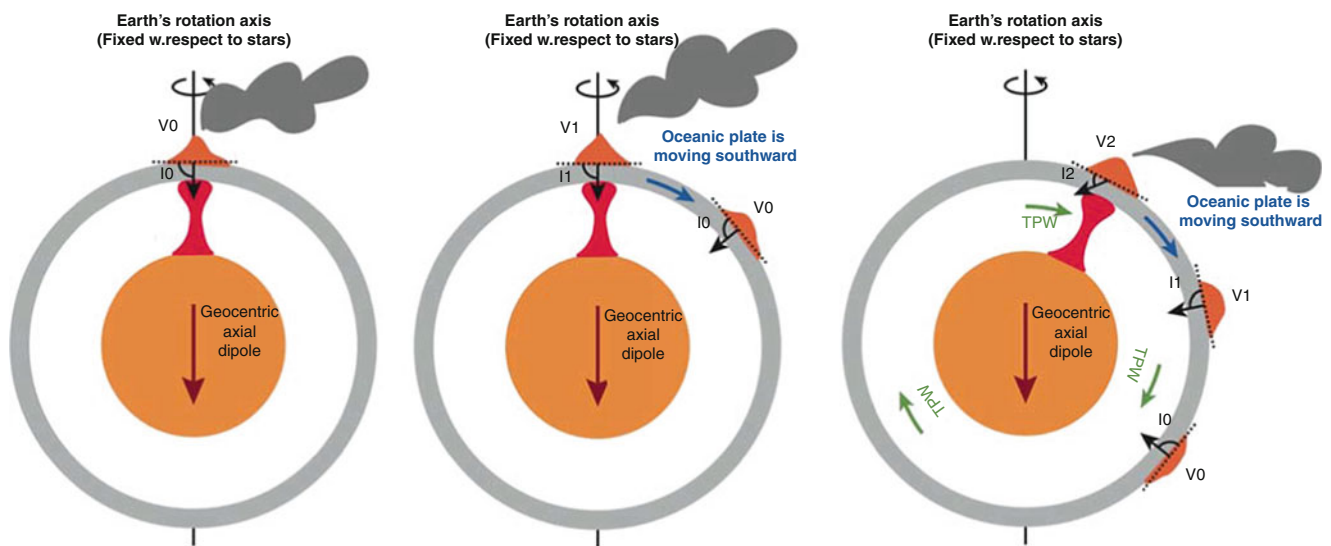
Hotspots are active volcanoes apparently unrelated to plate boundaries, which have been active for tens of millions of years. Hotspots are supposed to be the surface manifestation of plumes, which have left traces in the form of nearly linear chains of extinct volcanoes on the plate(s), which passed over them. Morgan (1972, 1981) proposed that these plumes have a "deep" mantle origin: these convective instabilities may be anchored at least down to the transition zone and possibly all the way down to the core-mantle boundary. As a plate moves over such a plume, the relative motion can be described by a rotation about a pole. If hotspots are indeed fixed with respect to each other and are fixed within the mantle, then hotspot apparent polar wander describes the wander of the Earth's rotation axis with respect to the mantle. It is this motion of the pole with respect to the hotspots, assumed to act as a reference frame for the mantle, which has been taken as an estimate of true polar wander. Determining TPW requires the best

possible knowledge of APWPs of all plates (i.e., the best possible data base of paleomagnetic poles to determine the paleolatitudes and orientations of plates with respect to the rotation axis), of their relative motions (determined by marine magnetic anomalies and azimuths of transform faults), and finally the positions and ages of individual volcanoes along hotspot tracks on plates. Paleolatitudes of these seamounts are used for comparison to test model predictions.

APWPs and kinematic models were developed from the 1960s to the 1980s under the key assumption that when averaged over a sufficient amount of time, in excess of a few thousand years, the Earth's magnetic field could be described accurately by a geocentric axial dipole (so-called GAD hypothesis). Besse and Courtillot (1991, 2002), among others, proposed to blend two data sets: the paleomagnetic poles used to derive the APWPs and the oceanic data used to derive plate kinematic models. They used then available paleomagnetic and kinematic models and produced a “synthetic” APWP, which was defined with respect to South Africa and then transferred to all other plates. Motion of hotspots with respect to the plates (Müller et al. 1993) was then integrated to derive an estimate of TPW (Fig. 2).

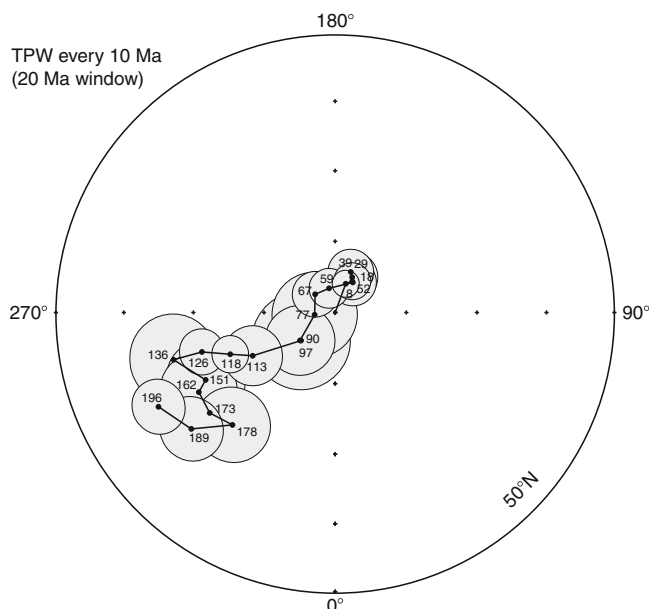
The resulting TPW curve is shown in Fig. 3 with a 20 Myr time resolution (each point corresponds to the successive positions with respect to time of a hotspot presently situated at the North geographic pole). Points are shown with their 95% confidence circles every 10 Myr and every other point is statistically independent (estimates at 5 Myr intervals are also

available in that study). The first points, corresponding to the period 8–59 Ma, are all in the same quadrant, between 3.3° and 5.9° away from the present rotation pole, with 95% uncertainties ranging from 2° to 3.8° . They are not statistically distinct from each other and therefore could correspond to a standstill. A mean position can be calculated from all data points in that time window: It is found to lie significantly displaced from the pole. The youngest mean pole at ~ 5 Ma is also significantly displaced from the pole but not statistically different from the 55 Ma overall mean. It seems that TPW may have been negligible for an extended period of about 50 Myr, but accelerated a few million years ago, with a velocity on the order of 100 km/Myr. Prior to that, the path displays a succession of a standstill at 160–130 Ma, a quasi-circular track from 130 to 70 Ma, then the standstill at 50–10 Ma and the faster motion up to the present. The true polar wander rate between 130 and 70 Ma averages 30 km/Myr. True polar wander appears episodic in nature, with periods of (quasi) standstill alternating with periods of faster TPW. The typical duration of these standstill periods is on the order of a few tens of millions of years (50 Myr). Because of all the uncertainties in models of hotspot kinematics prior to ~ 130 Ma, it is not safe to place too much weight on behavior prior to that time. The major event since then is therefore the end of the 130–60 Ma period of relatively fast polar wander, with a standstill (i.e., no or little TPW) from 50 Ma (actually because of larger 95% confidence circles, possibly 80 Ma) to 10 Ma.



Paleomagnetism, Polar Wander, Fig. 2 How mantle TPW is determined: At Time T0, a deep plume (located arbitrarily under the North Geographic Pole) produces a volcano V0 on an oceanic plate. In the hypothesis of a Geocentric Axial Dipole field, a magnetization of inclination $I_0 = 90^\circ$ is fossilized in the lavas (the inclination (I) depends on the colatitude (θ) such as $\text{TgITg}\theta = 2$). Time T1: the oceanic plate has moved southward and a hotspot trace is formed, but the Plume remains at the same position (i.e., no mantle TPW). A new volcano (V1) is fed and

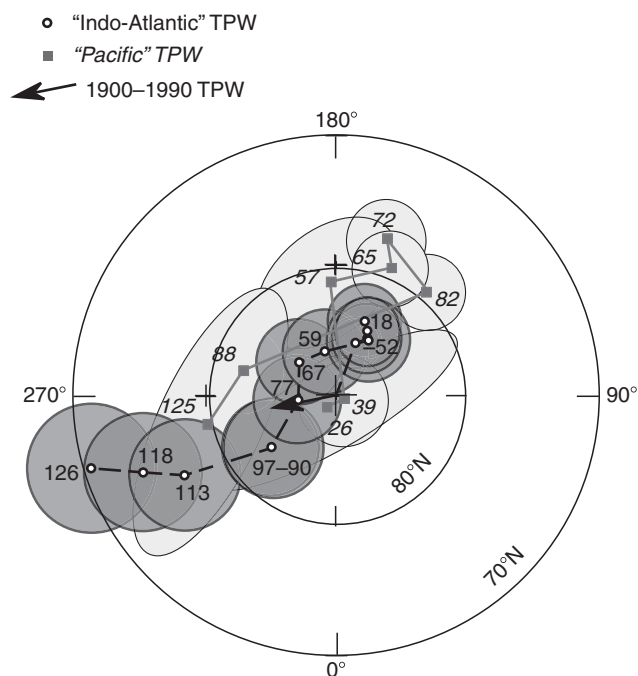
records the same 90° inclination (I_1) as in the volcano V0. Time T2: an episode of TPW occurs: the plume is tilted southward (with the rest of the mantle). The third volcano V2 will fossilize a lower inclination than V0 and V1. Later on, a paleomagnetist will measure past inclinations in lavas: as lithosphere only moves on the mantle, volcanoes will record the same inclination; the plume remains stationary. In case of TPW, the difference of paleolatitudes deduced from measured inclination will indicate the amount of motion



Paleomagnetism, Polar Wander, Fig. 3 True polar wander paths deduced from the hotspot model of Müller et al. (1993) going from the Present back to 130 Ma, and that of Morgan (1981) going from 130 back to 200 Ma, with associated 95% confidence ellipses (shaded light gray); 20 Ma sliding window. Ages are in Ma and correspond to the actual mean age of the data in the corresponding window, which is why they are not regularly distributed when the original time windows were. Note that the external circle is at 50° latitude. (After Besse and Courtillot 2002)

However uncertain, evidence for the fact that Earth emerged from a long standstill to enter a new period of faster polar wander in a different direction 10 Ma ago (~5 Ma at the higher resolution) is particularly interesting. That period would then still be going on. The youngest pole at ~5 Ma, taken at face value, would imply a TPW rate of 130 km/Myr in the direction ~15°W (with a large azimuthal uncertainty from 10°E to 40°W, due to proximity of the 5 Ma pole to the present rotation axis). The reasonable agreement with recent astronomical determinations (107 km/Myr in the direction 79°W, with a possible range from 65°W to 85°W) has reinforced the idea that this might be a valid estimate and that TPW direction and rate might have been rather stable over the last five million years corresponding to significant, large-scale changes in ice cover. Ice is not considered to have played a significant role before. It was absent prior to ~30 Ma.

A legitimate concern regarding the above conclusions is due to the fact that the analysis is not truly global, in that it fails to encompass the Pacific plate. Petronotis and Gordon (1999) have compiled an APW path (nine poles from 125 to 26 Ma) for the Pacific plate, based on the analysis of skewness of ocean crust magnetic anomalies and on seamount magnetic anomaly modeling. Using the Pacific plate versus hotspot kinematic model of Engebretson et al. (1985), based on dating of volcanoes that are part of hotspot tracks such as Hawaii, Louisville, or MacDonald, Besse and Courtillot (2002) have



Paleomagnetism, Polar Wander, Fig. 4 Comparison between the “Indo-Atlantic” (open dots and dashed line) and “Pacific” (squares and solid line) TPW Paths. The arrow shows the mean pole motion between 1900 and 1990. (After Besse and Courtillot 2002)

determined a corresponding 125–26 Ma “Pacific hotspot only” TPW curve. The “Pacific” and “Indo-Atlantic” TPW curves are compared in Fig. 4. This comparison is interesting because the data sets they are based on are entirely different and independent.

Despite some significant differences, it is worth emphasizing that the two curves are altogether similar in shape (tracks, amplitudes, azimuths), particularly the 300° longitude trending track from 130 to 70 Ma, though the two are offset (in the same general direction) by about 7°. More precisely, the confidence intervals intersect (though means are not in the intersection) near 125, 90, 60, 40, and 30 Ma. The main differences occur between 80 and 70 Ma: the 82 and 65–72 Ma poles derived from Petronotis and Gordon (1999) and the 77 and 67 Ma poles derived by Besse and Courtillot (2002) are clearly distinct. Tarduno and Cottrell (1997) and Tarduno et al. (2003), however, expressed doubts on the reliability of APWPs for an oceanic plate, such as the Pacific, when they are based solely from inversions of magnetic surveys over seamounts and/or analysis of skewness of marine magnetic profiles. They have determined the paleolatitude (based on inclination-only data from cores) of the 81 Ma old Detroit Seamount, which is part of the Emperor chain, and noticed that the paleolatitude is distinct from that based on the 81 Ma pole of the Pacific APWP (Sager and Pringle 1988) based on seamount anomaly poles. Tarduno and Cottrell (1997) exclude the possibilities of inadequate

sampling of secular variation, bias due to unremoved overprints or off-vertical drilling. Di Venere and Kent (1999) and Cottrell and Tarduno (2000) argue that the reliability of the Pacific paleopoles based on either modeling of seamount magnetic anomalies or determination of skewness of marine magnetic anomalies should be considered suspect. Both are prone to numerous biases and could yield errors in excess of 10° in the position of the mean poles derived from them. This debate is not closed.

Is a Reference Frame Based on Hotspots Valid?

Another question discussed in the last decade is Pacific hotspot fixity with respect to one another and to the “Indo-Atlantic” hotspots. Norton (1995) has suggested that the famous 43 Ma Hawaiian bend in the Hawaiian–Emperor hotspot track was actually a “nonevent,” i.e., indicated a change in motion of the Hawaiian hotspot with respect to the mantle rather than a change in Pacific plate motion. Koppers et al. (2001) find that the 0–43 Ma Hawaiian and Foundation seamount trail pair is the only one compatible with the fixed hotspot hypothesis. The 43–80 Ma Emperor/Line pair shows particularly large discrepancies, requiring motions of at least 30 km/Myr. Koppers et al. (2004) studied $^{40}\text{Ar}/^{39}\text{Ar}$ age progressions of the Hawaiian and Louisville hotspot tracks. They showed that the progression is nonlinear and can be better fitted with a moving hotspot model rather than a fixed hotspot frame of reference. The primary Hawaiian and Louisville hotspots did not move in concert, showing significant inter-hotspot motion between 80 and 47 Ma. While the Hawaiian hotspot experienced a substantial 15° southward motion prior to 47 Ma (Tarduno et al. 2003), the Louisville hotspot is predicted to have experienced an eastward motion of 5° during the same geological period, followed by a minor 2° latitudinal shift over the last 30 Myr. Clouard and Bonneville (2001) showed that Pacific seamounts are actually created by different processes, many being short-lived. Only the Hawaii and Louisville chains qualify as long-lived hotspots that can robustly be tested for fixity. Their preferred explanation for this is southward drift of Atlantic hotspots prior to ~95 Ma. Courtillot et al. (2003) reanalyzed the characteristics of the world catalog of some 50 hotspots. They suggest that surface hotspots on Earth may have three distinct origins with only seven “primary” hotspots (Hawaii, Easter, Louisville, Iceland, Afar, Reunion, and Tristan da Cunha) originating from the deepest part of the lower mantle, possibly anchored on chemical heterogeneities deposited in the D'' layer. Courtillot et al. (2003) found no evidence for inter-hotspot motion significantly larger than 5 km/Myr either within the Pacific hemisphere or the Indo-Atlantic hemisphere hotspots. Such rms velocities of 5 km/Myr or less, i.e., an order of magnitude less than rms

plate velocities, are to first order “small.” So, some primary hotspots indeed seem to provide a quasifixed frame in each hemisphere over the past 80–100 Myr.

Was there then any motion between the two hotspot ensembles? This raises the well-known difficulty of establishing a reliable kinematic connection between the two hemispheres through Antarctica. This has been addressed by Raymond et al. (2000). Based on updated kinematics (assuming that there is no missing plate boundary or unaccounted for motion between East and West Antarctica), these authors predict the location of the Hawaiian hotspot back in time, under the hypothesis that Reunion and Hawaii have remained fixed with respect to each other; for this, they use the dated tracks left on the African and Indian plates by the Reunion hotspot since it started as the Deccan traps 65 Ma ago. The plot of misfit between the predicted and observed positions for Hawaii as a function of time indicates that the two hotspots have actually drifted slowly, at ~10 km/Myr, for the last 45 Myr, but at a much faster rate (~50 km/Myr) prior to that, vindicating conclusions of Pacific hotspot motion reached by Tarduno and Cottrell (1997) or Tarduno (2007). Courtillot et al. (2003) conclude that the primary hotspots may form two distinct subsets in each one of the two geodynamically distinct hemispheres. Each subset would deform an order of magnitude slower than typical plate velocities. The two subsets would have been in slow motion for the last 45 Myr, but in faster motion prior to that.

Using their paleomagnetic results from Marie Byrd Land in West Antarctica, Di Venere and Kent (1999) demonstrated that some motion must have taken place since 100 Ma between West (Marie Byrd Land) and East Antarctica; they concluded that this motion cannot account for more than 20%, and possibly as little as 4% of the 14.5° offset between the observed and predicted positions of the 65 Ma Suiko seamount on the Emperor continuation of the Hawaiian hotspot track. They also discussed the integrity of the Pacific plate and the role of missing plate boundaries and errors in kinematic plate circuits, and find that they play a small role. Di Venere and Kent (1999) concluded that most of the apparent motion between the two main groups of hotspots is real, with an average drift of about 25 km/Myr since 65 Ma. Steinberger et al. (2004) have combined hotspot motion in the large-scale flow field of the mantle and intraplate deformation to account for most of the post-65 Ma misfits between computed and observed hotspot tracks. Only a combination of the two effects (hotspot motion and intraplate deformation) successfully predicts the geometry of the four selected seamount chains in the Pacific and Indo-Atlantic hemispheres. If correct, this analysis at the same time invalidates the notion of a reference frame or of separate hemispheres and replaces it with limited, predictable, inter-hotspot motion. But TPW can still be calculated. The Hawaiian hotspot would have moved faster between 80 and 50 Ma, and at the same time TPW

would have been faster prior to 50 Ma. TPW alone would account for paleolatitudinal changes of the Reunion hotspot, whereas the two effects would be required for the Hawaiian hotspot.

The estimate of true polar wander (TPW) over the last 200 Myr is thus still a matter of strong debates, as recalled above, being based only on hotspots from the Indo-Atlantic hemisphere. The TPW curve estimated using Pacific-only data is similar to the Indo-Atlantic TPW, seemingly validating to first order the concept that TPW is a global phenomenon. On the other hand, Besse and Courtillot (2002) suggest that TPW pole positions for the two hemispheres are significantly displaced between ~50 and ~90 Ma. Hotspot tracks become fewer and more uncertain as one goes back in the past and pre-130 Ma TPW estimates should be regarded with caution. However, there are strong indications (Besse and Courtillot 2002) that the phase of true polar wander, which ended around 50 Ma, may have started ~130 Ma ago.

Modeling of True Polar Wander

In an early attempt, Ricard et al. (1993b) and Richards et al. (1997) derived a model of mantle density heterogeneities based on plate motions recorded by the ocean floor, including slabs crossing the upper/lower mantle transition zone slowing down by a factor between 2 and 5. Based on this, they computed a synthetic geoid, which agreed rather well with observations, provided viscosity increased by a factor 40 going from the upper to the lower mantle. Changes in the degree 2 geoid should reflect the history of the inertia tensor and therefore TPW as revealed by paleomagnetism. They predicted TPW velocities on the order of 300 km/Myr in the direction 130°W, much larger than observed rates. Their three-layer model could not account simultaneously for the present-day geoid and observed TPW.

In other studies, seismic tomography has been used to infer 3D maps of density heterogeneities that drive flow in the viscous mantle. Steinberger and O'Connell (1997, 2002) have developed an algorithm to calculate changes in the Earth's rotation axis for rather general viscoelastic Earth models. Their model combines hotspot motion, polar wander, and plate motions in a single dynamically consistent model, allowing all effects to be simultaneously determined and their relative effects sorted out. Changes in rotation are due to advection of "realistic" density anomaly distributions inferred from a number of seismic tomography models in a "realistic" flow field constrained by observed plate kinematics at the surface. Steinberger and O'Connell (2002) compare their model results for a number of tomographic models and discuss the effects of compressibility, deep mantle viscosity, thickness of a high viscosity layer in the deep mantle, and chemical boundaries on rates of TPW. Their TPW estimates

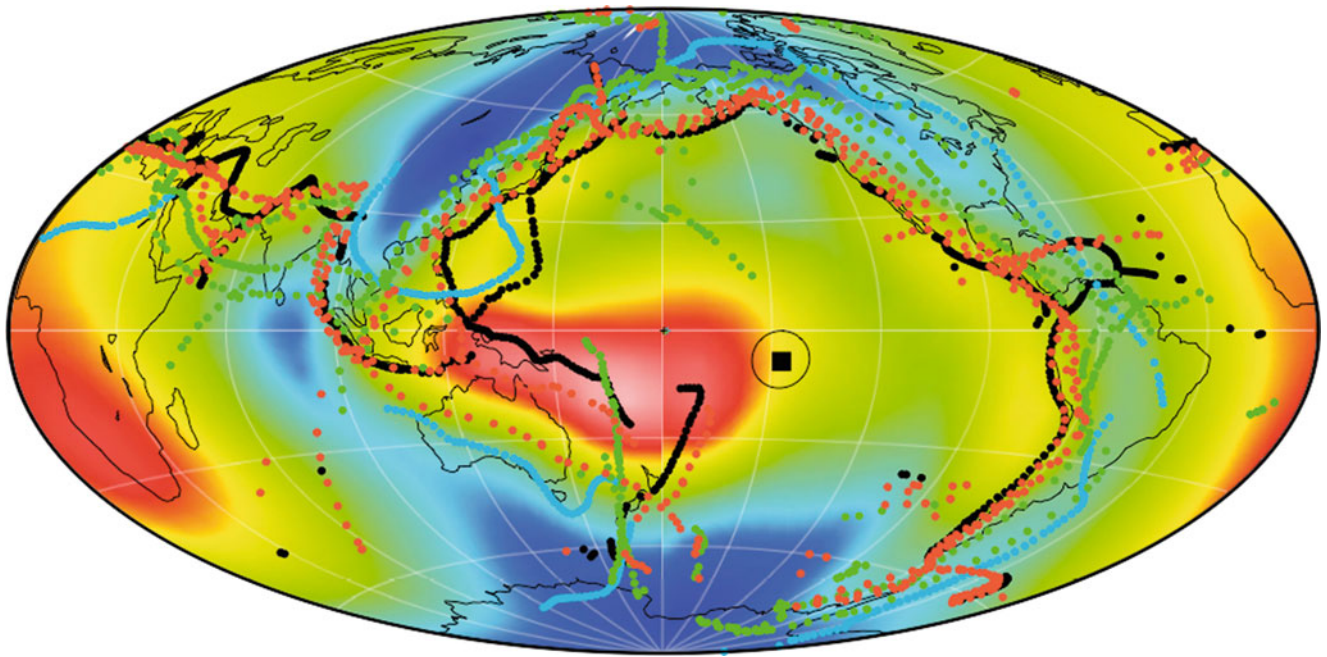
are in qualitative agreement with observations for the last 100 Ma. Steinberger and Torsvik (2010) modeled the present-day non-hydrostatic inertia tensor by combining contributions due to large low-shear velocity provinces in the lowermost mantle and due to subduction and found a spin axis at 67°N, 96°E (north Siberia). The distribution of recent subduction, with largest amounts in the northwest Pacific (beneath East Asia) and the southeast Pacific (beneath South America), adds a secondary contribution that moves the spin axis toward the observed poles.

Rouby et al. (2010) also investigated the effect of internal mass redistributions on the position of the Earth's rotational pole for the last 120 Myr. They also used a geodynamic model based on plate reconstructions and estimated the location and rate of subducted slabs under the assumption that they sink vertically into the mantle (Fig. 5). The new model takes into account the effect of large-scale mantle upwellings derived from tomography (one is situated under Africa, the other, nearly antipodal under the Pacific Ocean). The temporal variation of the mantle density heterogeneities was essentially found to reflect changes in subduction history. Because the tomographic images seem to show that subducted plates do not systematically reach the CMB, they investigated the influence of the depth down to which these plates retain significant density contrast with respect to the surrounding mantle on the present-day geoid and on the temporal evolution of the Earth's PIA (we recall the rotational axis is aligned to the maximum PIA). They proposed a model, which yields to a present-day geoid in good agreement with the observed one. The associated minimum principal inertia axis seems to be relatively stable since 120 Ma and close to Africa, within the upwelling (Fig. 6). Their model displays that some of the features of the computed TPW, which are directly linked to the temporal evolution of the subductions, are reminiscent of observed phases of slow and fast true polar wander, with similar peak velocities. Although their computed TPW is almost in the same meridian as the observed one, the poles are still drifting in the opposite direction, as was already the case for Ricard et al. (1997).

An Integrated Explanation for Polar Wander?

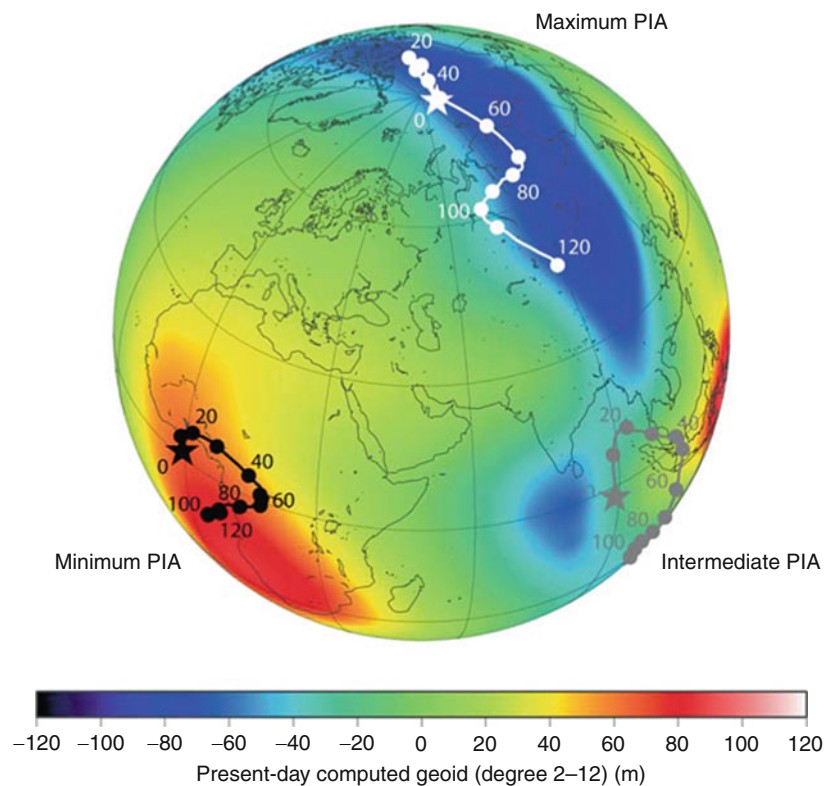
Let us now put this study in the broader context of the history of the Earth's rotation.

Paleomagnetic indications that would support rather fast true polar wander in pre-Mesozoic time have been put forward by Van der Voo (1994; with rates between 70 and 110 km/Myr) and for very fast true polar wander by Kirschvink et al. (1997; with rates in excess of 600 km/Myr in the early Cambrian). Recall that Steinberger and O'Connell's (2002) numerical models place an upper limit of about 100 km/Myr on TPW velocity. For pre-Jurassic time,



Paleomagnetism, Polar Wander, Fig. 5 Subduction zones at the Earth surface, every 40 Myr, since 200 Myr, superimposed to the present-day final computed geoid (degree 2–12). The *square* represents the mean center of small circles fitting the subductions zones. (After Rouby et al. 2010)

Paleomagnetism, Polar Wander, Fig. 6 Principal Inertia Axes since 120 Myr (dots every 10 Myr, stars denote present-day values) for our mantle density heterogeneities model. The corresponding present-day geoid (degree 2–12) is also shown. (After Rouby et al. 2010)



ocean floor has been destroyed (together with evidence of plate kinematics and most hotspot traces) and TPW cannot be estimated in the way indicated above. Evans (2003) proposes to obtain a crude estimate of TPW in the case when TPW velocity would have been much faster than between-plate

motions. Kirschvink et al. (1997) and Evans (2003) discussed the fact that the Earth may have in the past catastrophically exchanged two of its principle axes of inertia (inertial interchange true polar wander or IITPW), i.e., the whole lithosphere (and mantle) may have rotated by some 90° in only a

few million years, bringing for instance parts of the Earth that were close to the poles at the Equator. Evans (2003) interpreted the large rotations as TPW about a common, long-lived, minimum inertial axis ($I_{\min}$) near eastern Australia, which may also be regarded as the prolate axis of Earth's figure, inherited from Rodinia (Fig. 7).

Satolli et al. (2007) on the basis of a high definition magnetostratigraphy in the Apennines argued that a major trend of the “mean” plate motion since at least 200 Ma can be described as large successive clockwise or counter-clockwise rotations linked to TPW, and roughly centered on Africa, this last plate being a “central plate” from which most continents have diverged since the break-up of Pangea (Fig. 8). An animated sequence derived from Besse and Courtillot (1991) APWP displays visually those features available at <http://www.ipgp.fr/~besse/pw>. Steinberger and Torsvik (2008) have also found that the global axis of rotation for plates since 250 Ma also roughly centered in Africa, and argued that they may also correspond to maxima of the degree 2 non-hydrostatic geoid, corresponding to the mantle upwellings.

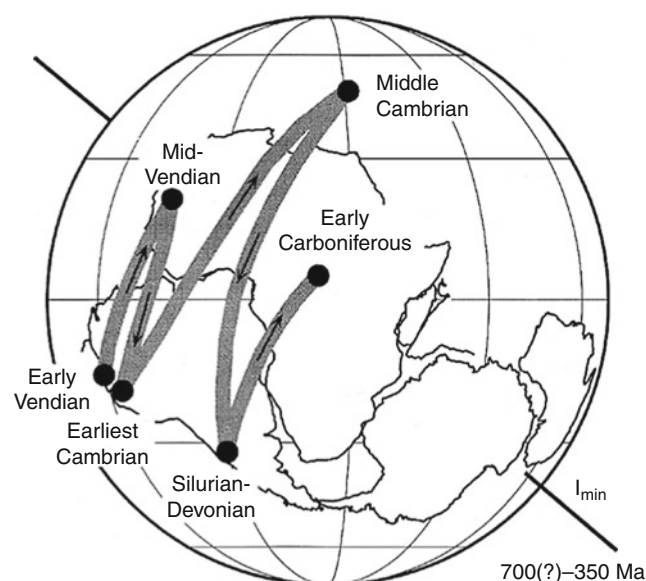
Le Pichon and Huchon (1984) already remarked that during Late Triassic about 200 Ma ago, the outlines of Pangea lay along a great circle passing through the paleopoles of rotation and also noticed that the present geoid also displays hemispheric symmetry with an axis located in the equatorial plane. The stability of subduction zones back to the middle of the Paleozoic in a nearly N-S plane was also recognized by Collins (2003). All these studies pointed out a strong link

between pole positions and the geometry of the convection: this coincidence might be explained by a deep-seated, quadrupolar convection as was first advocated by Busse (1983). Rouby et al. (2010) addressed this question by computing a present geoid day (degree 2–12) using the plate model described above (Figs. 5 and 6).

The persistent location of subductions in the western Tethyan and Pacific regions for at least 250 Ma (and possibly earlier according to Collins) have generated a ring of downwelling material inducing hot return flows (the upwellings mentioned above) confined to two main regions (presently Indo-Atlantic and Pacific). This geometry defines the location of PIAs and pins the minimum PIA in the region of upwelling. This large zone is roughly centered on the equator also close to the maxima of the degree 2 non-hydrostatic geoid. The minimum PIA has been relatively stable since 120 Ma and close to Africa. On the contrary, both the maximum and intermediate PIAs have moved in a plane perpendicular to Africa along a meridian corresponding to the large-scale pattern low of the geoid. This meridian seems to be stable with time, similar to the large-scale pattern of subduction around the Pacific and western Tethyan zone of subductions.

The temporal evolution of the rotation axis is that of the maximum PIA. Oscillatory polar wander is attributed to TPW about a common, long-lived minimum inertial axis, which would be the prolate axis of Earth figure, inherited from the past plates/continent configurations and following the quadrupolar convection model of Busse (1983).

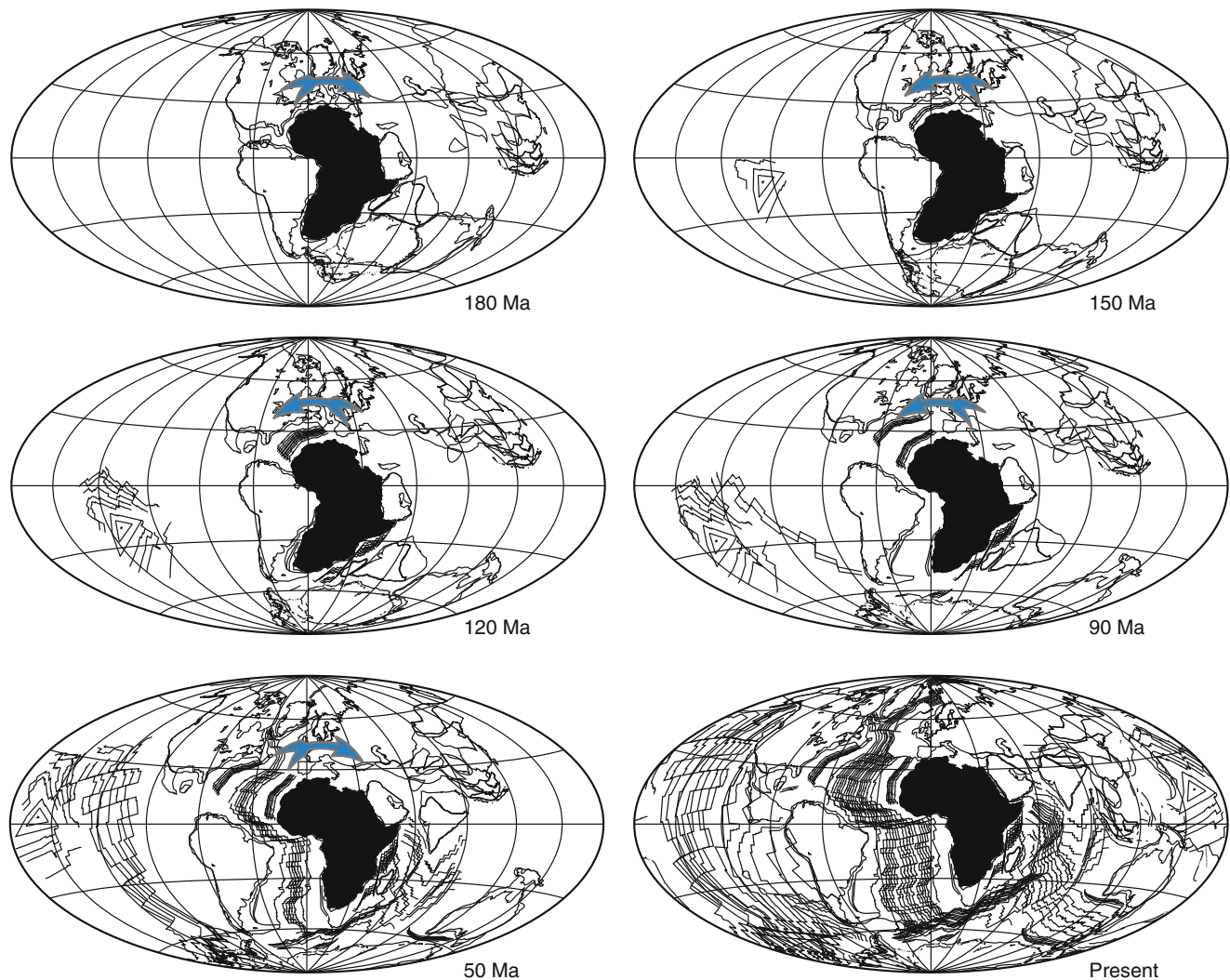
This peculiar pattern may constitute a long-lived feature of our planet, at least since 200–250 Ma, which forces continents to be roughly in an EW position, on a single hemisphere bounded by a plane containing the rotation axis, with the Earth's rotational pole not far from the main ring of subduction zones. TPW (and possibly APW) may occur around a nearly equatorial euler pole (axis) of rotation, driven by slight changes of magnitude between the maximum and intermediate PIAs, principally caused by subduction locations and rate changes with time. This view is coherent with the successive CW and CCW rotations of all continents (including the Pacific Ocean) observed by Evans (2003), Satolli et al. (2007), and Steinberger and Torsvik (2008) around a point located somewhere in Africa and close to the equator for these last two authors.



Paleomagnetism, Polar Wander, Fig. 7 Terminal Proterozoic to late Paleozoic apparent polar wander path for Gondwanaland. Oscillatory APW rotations are interpreted as TPW about a common, long-lived, minimum inertial axis (near eastern Australia). (After Evans 2003)

Conclusions

For the time being, the hypothesis that TPW is a go-stop-go phenomenon, with long standstills in the tens of millions of years, and periods of faster, rather uniform polar wander at



Paleomagnetism, Polar Wander, Fig. 8 Global continent and ocean reconstructions, from 180 Myr to present-day, based on techniques and APWPs described in Besse and Courtillot (2002). Magnetic anomalies are from Royer et al. (1992), the 150 and 130 Myr frames from Satolli et al. (2007). A major trend of the “mean” plate motion since at least

200 Myr can be described as successive clockwise or counter-clockwise rotations (linked to TPW), and roughly centered in Africa (close to the minimum PIA, see Fig. 8). The blue arrow indicates the direction of rotation until the next reconstruction. (After Rouby et al. 2010)

velocities often not exceeding 30 km/Myr, cannot be discounted. Evidences for superfast events at any time in the past remain to be demonstrated.

The similarity between TPW estimates for the Pacific plate and the rest of the world, which are based on completely different and independent data sets, may be taken as support to the idea that small yet significant TPW, on the order of 10° , occurred since the Cretaceous, though it still requires careful verification.

Most current modeling still fails to some extent to account for the slow values of typical TPW velocity (0–30, and rarely up to 100 km/Myr), and even more so to account for the prolonged (~50 Myr) periods with almost no TPW (standstills). Steinberger and O’Connell (2002) give reasonable velocities but predict smooth evolutions, rather than the

alternating episodes, revealed by paleomagnetic data. Rouby et al.’s (2010) model restores the present geoid, displays some of the features of computed TPW that are directly linked to the temporal evolution of subductions, reminiscent of observed phases of slow and fast true polar wander, with similar peak velocities, but the poles are still drifting in the wrong direction, as was also the case for Ricard et al. (1997).

TPW seems to be linked to a pattern of quadrupolar convection permanent (at least on periods of time in excess of 200 Ma). If the two upwellings remain more or less in the same locations, the pole may move on the great circle perpendicular to the axis they form, the system being stabilized by subductions not far from the pole. These concepts remain to be tested for earlier periods of time.

The importance of the Pacific plate and severe limitations on presently available data from that plate, the fact that the global paleomagnetic database still is quite scant for certain continents and time windows, the fact that many hotspot traces are not yet sufficiently constrained in terms of age and paleolatitude, all point to the need for many more direct (paleomagnetic) measurements as opposed to indirect/remote sensing determinations of magnetization direction (i.e., “skewness” or “seamount” data).

The reason for the discrepancies between models and data could be due to the quality of reconstruction of the Pacific Ocean (specially before 130 Ma) and its relationship with bordering continents, and hence, the poor knowledge of the exact amount of subduction at a given time (to which these kind models are so sensitive). Some of this information is unfortunately lost forever, since large parts of the Pacific Ocean have been subducted.

Future research should aim to understand the discrepancies between observation and computations in order to estimate possible uncertainties on the associated mantle mass anomalies by testing other reconstructions. It will be particularly useful to build a single reference frame for both Indo-Atlantic and Pacific hemispheres: this implies that one must be able to use Antarctica to link the two hemispheres. This will not be easy, given the poor knowledge of deformations suffered by the continent and the possibility of existence of active margins on its Pacific border.

Cross-References

- [Geomagnetic Excursions](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Geomagnetic Field, Secular Variation](#)
- [Geomagnetic Field, Theory](#)
- [Paleomagnetism, Magnetostratigraphy](#)
- [Paleomagnetism, Principles](#)

Bibliography

- Besse J, Courtillot V (1991) Revised and synthetic polar wander paths of the African, Eurasian, North American and Indian plates, and true polar wander since 200 Ma. *J Geophys Res* 96:4029–4050
- Besse J, Courtillot V (2002) Apparent and true polar wander and the geometry of the geomagnetic field in the last 200 million years. *J Geophys Res* 107(2300):31. <https://doi.org/10.1029/2000JB000050>
- Busse FH (1983) Quadrupole convection in the lower mantle? *Geophys Res Lett* 10(4):285–288
- Clouard V, Bonneville A (2001) How many Pacific hotspots are fed by deep-mantle plumes? *Geology* 21:695–698
- Collins WJ (2003) Slab pull, mantle convection, and Pangaean assembly and dispersal. *Earth Planet Sci Lett* 205:225–237
- Cottrell RD, Tarduno JA (2000) Late cretaceous true polar wander: not so fast. *Science* 288:2283a
- Courtillot V, Davaille A, Besse J, Stock J (2003) Three distinct types of hotspots in the Earth’s mantle. *Earth Planet Sci Lett* 205:295–308
- Di Venere V, Kent DV (1999) Are the Pacific and Indo-Atlantic hotspots fixed? Testing the plate circuit through Antarctica. *Earth Planet Sci Lett* 170:105–117
- Engelbreton DC, Cox A, Gordon, RG (1985) Relative motions between oceanic and continental plates in the Pacific Basin. *Geological Society of America Special Paper* 206, 59 pp
- Evans DAD (2003) True polar wander and supercontinents. *Tectonophysics* 362:303–320
- Goldreich P, Toomre A (1969) Some remarks on polar wandering. *J Geophys Res* 74:2555–2567
- Gordon RG (1987) Polar wandering and paleomagnetism. *Annu Rev Earth Planet Sci* 15:567–593
- Gordon RG, Horner-Johnson BC, Petronotis K, Acton GD (2004) Apparent polar wander of the Pacific plate and Pacific hotspots: implications for true polar wander and hotspot fixity. *EOS Transactions, AGU, Spring meeting, Abstract #U52A-04*
- Gross RS, Vondrak J (1999) Astrometric and space-geodetic observations of polar wander. *Geophys Res Lett* 26:2085–2088
- Kirschvink JL, Ripperdan RL, Evans DA (1997) Evidence for a large-scale early Cambrian reorganization of continental masses by inertial interchange true polar wander. *Science* 277:541–545
- Koppers AAP, Morgan JP, Morgan JW, Staudigel H (2001) Testing the fixed hotspot hypothesis using Ar-40/Ar-39 age progressions along seamount trails. *Earth Planet Sci Lett* 185(3–4):237–252
- Koppers A, Duncan R, Steinberger B (2004) Implications of a nonlinear 40Ar/39Ar age progression along the Louisville seamount trail for models of fixed and moving hot spots. *Geochem Geophys Geosyst* 5(6):Q06L02. <https://doi.org/10.1029/2003GC000671>
- Le Pichon X, Huchon P (1984) Geoid, Pangea and convection. *Earth Planet Sci Lett* 67:123–135
- Mitrovica JX, Milne GA (1998) Glaciation-induced perturbations in the Earth’s rotation: a new appraisal. *J Geophys Res* 103:985–1005
- Morgan WJ (1972) Plate motions and deep mantle convection. *Geol Soc Am Mem* 132:7–22
- Morgan JW (1981) Hotspot tracks and the opening of the Atlantic and Indian oceans. In: Emiliani C (ed) *The sea*. Wiley, New York, pp 443–487
- Müller DM, Royer JY, Lawver LA (1993) Revised plate motions relative to the hotspots from combined Atlantic and Indian Ocean hotspot tracks. *Geology* 21(3):275–278
- Norton IO (1995) Plate motion in the North Pacific: the 43 Ma nonevent. *Tectonics* 14:1080–1094
- O’Connell RJ, Gable CW, Hager BH (1991) Toroidal- poloidal partitioning of lithospheric plate motions. In: Sabadini R, Lambeck K (eds) *Glacial isostasy, sea level and mantle rheology*. Kluwer, Dordrecht, pp 535–551
- Petronotis KE, Gordon RG (1999) A Maastrichtian paleo- magnetic pole for the Pacific plate from a skewness analysis of marine magnetic anomaly 32. *Geophys J Int* 139:227–247
- Raymond CA, Stock JM, Cande SC (2000) Fast Paleogene motion of the Pacific hotspots from revised global plate circuit constraints. In: Richards M, Gordon R, van der Hilst R (eds) *The history and dynamics of global plate motions*. Geophysical monograph, vol 121. American Geophysical Union, Washington, DC, pp 359–375
- Ricard Y, Doglioni C, Sabadini R (1991) Differential rotation between lithosphere and mantle: a consequence of lateral mantle viscosity variations. *J Geophys Res* 96:8407–8416

- Ricard Y, Spada G, Sabadini R (1993a) Polar wandering of a dynamic Earth. *Geophys J Int* 113:284–298
- Ricard Y, Richards M, Lithgow-Bertelloni C, Le Stunff Y (1993b) A geodynamic model of mantle density heterogeneity. *J Geophys Res* 98:21895–21909
- Richards MA, Ricard Y, Lithgow-Bertelloni C, Spada G, Sabadini R (1997) An explanation of the Earth's long-term rotational stability. *Science* 275:372–375
- Rouby H, Greff M, Besse J (2010) Mantle dynamics, geoid, inertia and TPW since 120 Myr. *Earth Planet Sci Lett*. <https://doi.org/10.1016/j.epsl.2010.01.033>
- Royer J-Y, Müller RD, Gahagan LM, Lawver LA, Mayes CL, Nürnberg D, Sclater JG (1992) A global isochron chart. University of Texas Institute for Geophysics, Technical report, 117
- Runcorn SK (1956) Paleomagnetic comparisons between Europe and North America. *Proc Geol Assoc Can* 8:77–85
- Sabadini R, Peltier WR (1981) Pleistocene deglaciation and the Earth's rotation: implications for mantle viscosity. *Geophys J R Astron Soc* 66:553–578
- Sager WW, Pringle S (1988) Mid Cretaceous to early tertiary apparent polar wander path of the Pacific plate. *J Geophys Res* 93:11753–11771
- Satolli S, Besse J, Speranza F, Calamita F (2007) The 125–150 Ma high-resolution Apparent Polar Wander Path for Adria from magnetostratigraphic sections in Umbria–Marche (Northern Apennines, Italy): timing and duration of the global Jurassic–Cretaceous hairpin turn. *Earth Planet Sci Lett* 257:329–342
- Spada G, Ricard Y, Sabadini R (1992) True polar wander for a dynamic Earth. *Nature* 360:452–454
- Steinberger B (2000) Plumes in a convecting mantle: models and observations for individual hotspots. *J Geophys Res* 105:11127–11152
- Steinberger BM, O'Connell RJ (1997) Changes of the Earth's rotation axis inferred from advection of mantle density heterogeneities. *Nature* 387:169–173
- Steinberger B, O'Connell RJ (2000) Effects of mantle flow on hotspot motions. In: Richards M, Gordon R, van der Hilst R (eds) *The history and dynamics of global plate motions*. Geophysical monograph, vol 121. American Geophysical Union, Washington, DC, pp 377–398
- Steinberger B, O'Connell RJ (2002) The convective mantle flow signal in rates of true polar wander. In: Mitrovica J, Vermeersen L (eds) *Ice sheets, sea-level and the dynamic Earth*. Geodynamics series, vol 29. American Geophysical Union, Washington, DC, pp 233–256
- Steinberger B, Torsvik TH (2008) Absolute plate motions and true polar wander in the absence of hotspot tracks. *Nature* 452(7187):620
- Steinberger B, Torsvik TH (2010) Toward an explanation for the present and past locations of the poles. *Geochem Geophys Geosyst* 11:Q06W06. <https://doi.org/10.1029/2009GC002889>
- Steinberger B, Sutherland S, O'Connell R (2004) Prediction of Emperor–Hawaii seamount locations from a revised model of plate motion and mantle flow. *Nature* 430:167–173
- Tarduno JA (2007) On the motion of Hawaii and other mantle plumes. *Chem Geol* 241:234–247
- Tarduno JA, Cottrell RD (1997) Paleomagnetic evidence for motion of the Hawaiian hotspot during formation of the Emperor seamounts. *Earth Planet Sci Lett* 153:171–180
- Tarduno JA, Gee J (1995) Large-scale motion between Pacific and Atlantic hotspots. *Nature* 378:477–480
- Tarduno JA, Duncan RA, Scholl DW, Cottrell R, Steinberger B et al (2003) The Emperor seamounts: southward motion of the Hawaiian hotspot plume in Earth's mantle. *Science* 301:1064–1069
- Van der Voo R (1994) True polar wander during the mid-Paleozoic? *Earth Planet Sci Lett* 122:239–243

Paleomagnetism, Principles

William Lowrie

Institute of Geophysics, Swiss Federal Institute of Technology, Zürich, Switzerland

Definition

Paleomagnetism is the study of the Earth's ancient magnetic field, based on the measurement and analysis of the *remanent* (i.e., permanent) magnetization acquired by rocks during their formation.

Introduction

Paleomagnetism provides data that are important in the interpretation of geological processes. For paleomagnetic results to be reliable for this purpose, several stringent conditions must be fulfilled. First, a rock under investigation must be able to record accurately the direction and intensity of the weak geomagnetic field during its formation or later alteration. This requires an understanding of the magnetic properties of rocks and minerals. Second, the age of acquisition of the rock's magnetization, or its subsequent alteration, must be well known. Field tests based on the local and regional geology of a sampling site are valuable in this respect. Third, the remanent magnetization of the rock must possess outstanding stability in order to be able to persist unchanged for geological eons.

The Geocentric Axial Dipole

The Earth's magnetic field has been measured and analyzed systematically since Carl Friedrich Gauss and Wilhelm Weber established a global network of geomagnetic observatories in the late 1830s. In the past 40 years, Earth-orbiting satellites carrying magnetometers have greatly increased the amount of high-quality magnetic data and also improved the global coverage. The measurements are analyzed using spherical harmonic analysis to define the potential W_i of the field of internal origin by the series

$$W_i = R \sum_{n=1}^N \sum_{m=0}^n \left(\frac{R}{r}\right)^{n+1} (g_n^m \cos m\phi + h_n^m \sin m\phi) P_n^m(\cos \theta)$$

The Gauss coefficients g_n^m and h_n^m in this series can be computed accurately to a very high degree n . For example, a

combination of satellite data with marine magnetic and aeromagnetic data produced a high-resolution model of the field (NGDC-720) with Gauss coefficients g_n^m and h_n^m up to degree $n = 720$. The high-degree terms with $n \geq 15$ are attributed to magnetic sources in the lithosphere. The Gauss coefficients with $n \leq 14$ define the International Geomagnetic Reference Field (see ► “Geomagnetic Field, IGRF”), which is the global magnetic field that originates in the Earth’s core.

The most important part of the magnetic field is the dipole component, which corresponds to the terms with $n = 1$. All terms with $n \geq 2$ in the spherical harmonic analysis are referred to collectively as the *non-dipole field*. The term g_1^0 is the strongest component of the field; it represents a magnetic dipole located at the center of the Earth and aligned with the rotation axis. The coefficients g_1^1 and h_1^1 correspond to orthogonal equatorial dipole components. Together, the three terms with $n = 1$ describe a geocentric dipole whose axis is inclined to the rotation axis. At present the axial tilt is about 9° , and the geomagnetic pole does not coincide with the geographic pole. A compass needle aligns with the horizontal component of the magnetic field – the magnetic meridian – and not with the geographic meridian; the angular deviation is the *declination* (D), which varies from place to place on the Earth’s surface.

The *inclination* (I) of the magnetic field is the angle that a magnetic field line makes with the horizontal at the Earth’s surface. For a dipole field the inclination is related to the latitude λ of the observation site by the equation

$$\tan I = 2 \tan \lambda = 2 \cot p \quad (1)$$

In this equation $p = (90^\circ - \lambda)$ is the co-latitude of the observation site, i.e. the angular distance of the magnetic pole from the site.

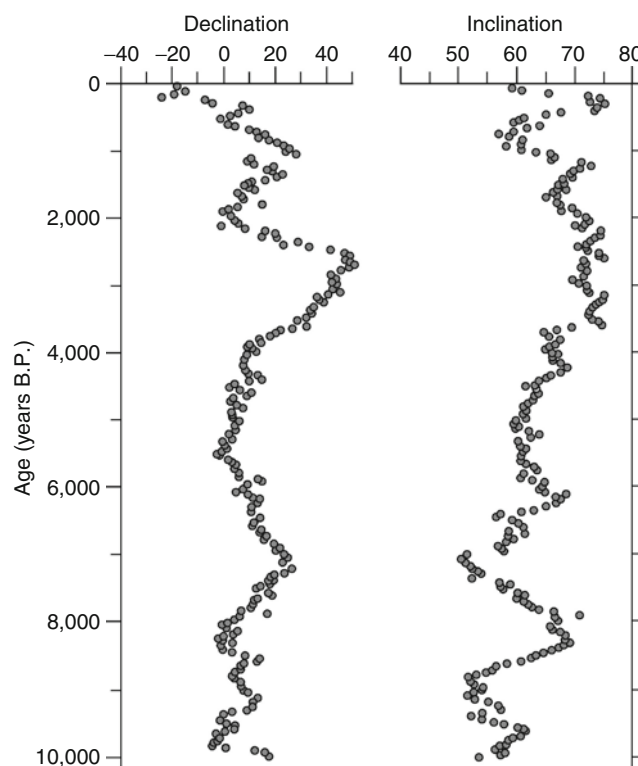
The basic premise of paleomagnetism is the *geocentric axial dipole* (GAD) hypothesis. This postulates that the equatorial dipole components, together with the non-dipole field, average to zero when averaged over sufficiently long intervals of time, leaving only the GAD component as the geologically important record of the ancient field. Consequently, the magnetic pole position calculated in a paleomagnetic investigation defines the location of the paleogeographic axis. Discrepancies from the present axis are interpreted in terms of geological processes.

The dipole and non-dipole fields change slowly with time, a characteristic known as *secular variation*. Time-dependent changes of the Gauss coefficients are observed in repeated analyses of the recent field. Secular variations affect both the direction and the intensity of the field. It has been possible to reconstruct geomagnetic secular variation since the late sixteenth century by using the magnetic observations of mariners (Jackson et al. 2000). Moreover, the intensity of the geomagnetic field has decayed by about 5% per century since direct

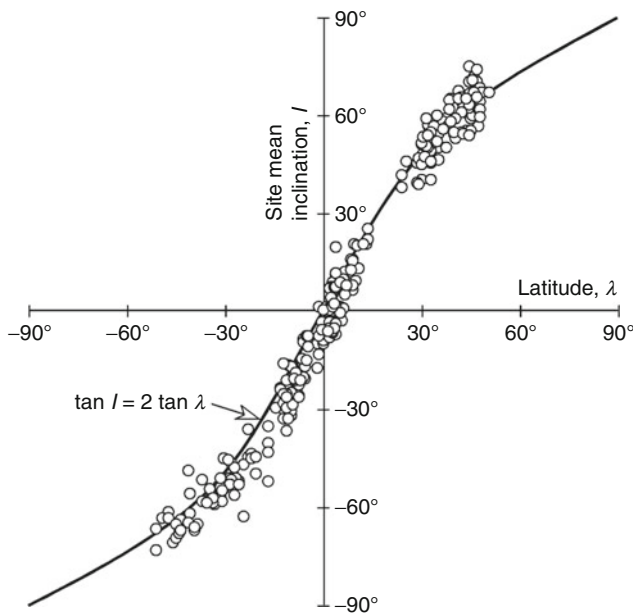
measurements began about 180 years ago. The rate of decrease prior to 1840 may have been slower (Gubbins et al. 2006).

On a longer time-scale, fluctuations in magnetization direction and intensity that are not due to mineralogical variations have been documented in paleomagnetic analyses of sedimentary records. The changes take place over centuries and millennia and are called *paleosecular variations* to distinguish them from changes in the present and historic field. Figure 1 shows oscillations in both inclination and declination, compiled by averaging the magnetization directions measured in cores of sediments from British lakes. The sediments were dated using radiocarbon and palynological age determinations, which enabled correlation of their records from lake to lake and subsequent stacking. The figure represents paleosecular variation in direction of the geomagnetic field in Britain during the past 10,000 years.

For the GAD hypothesis to be valid, the paleosecular variations must average to zero over a few thousands of years, so that the inclination and declination of the magnetization recorded in a rock or sediment represent the GAD field at the time of their formation. A test of the hypothesis, carried out in cores of marine sediments deposited in the last 2.5 Million years, is illustrated in Fig. 2. The mean inclinations



Paleomagnetism, Principles, Fig. 1 Master curves of paleosecular variation of inclination and declination for the past 10,000 years, compiled from British lake sediments. (After Fig. 5 in Turner and Thompson 1981. Redrawn by permission of Oxford University Press)



Paleomagnetism, Principles, Fig. 2 The mean inclinations of magnetization in marine sediment cores are plotted against the latitude of each sampling site. (After Fig. 3 in Schneider and Kent 1990. Redrawn by permission of John Wiley and Sons. Copyright 1990 by the American Geophysical Union). The continuous curve is the theoretical relationship given by Eq. (1) for a GAD field. The small deviations are partly due to paleosecular variation that has not been averaged out, but systematic bias indicates persistent long-term quadrupole and octupole components

of magnetization measured in the sediments agree well with the values expected for a GAD field predicted by Eq. (1). However, the agreement is not perfect, and the optimum fit of a theoretical curve to the data is obtained when small geocentric axial quadrupole (g_2^0) and octupole (g_3^0) components are included, each equivalent to a few percent of the dipole. The magnetohydrodynamic theory of the origin of the geomagnetic field gives no rigorous theoretical justification for the GAD hypothesis or explanation of possible persistent zonal quadrupole and octupole components. However, experimental tests of the hypothesis indicate that it is a good first-order model of the time-averaged geomagnetic field. The discrepancies may be due to a geological factor, such as sedimentary inclination flattening (Tauxe 2005).

Rock Magnetism

The ability of rocks and sediments to acquire a stable magnetization that preserves the direction of the GAD is due to the presence of specific magnetic minerals that behave in a similar way to ferromagnetic metals. Ferromagnetism is the type of magnetic behavior found in iron and similar metals. It is characterized by a Curie temperature, above which the ferromagnetism disappears. Ferromagnetic metals are strongly

magnetic. They have properties such as magnetic hysteresis, saturation magnetization, and remanent magnetization. The important rock-forming minerals exhibit *diamagnetism* or *paramagnetism*, which are very weak forms of magnetism that do not show magnetic hysteresis. Because they cannot acquire a remanent magnetization, they play no part in paleomagnetism. A few minerals – iron oxides and sulfides – exhibit *ferrimagnetism*, which is a form of ferromagnetic behavior associated with the mineral's crystal structure. Ferrimagnetic minerals are able to acquire a *remanent magnetization*. This means that when the magnetic field changes or is removed, they retain a magnetization imparted by it. Although they occur in tiny concentrations in rocks, the paleomagnetic properties of a rock or sediment are due to their presence.

The most common ferrimagnetic minerals are members of the titanomagnetite and titanohematite solid solution series, whose end-members are magnetite and hematite, respectively. Magnetite is usually a primary magnetic mineral coeval with the origin of igneous rocks or pre-dating the deposition of sediments. Hematite, pyrrhotite, and greigite are other important magnetic minerals, but they often originate in the rock after its formation, so the age of their magnetization can be dubious.

The magnetization of a rock is due to the partial alignment of the magnetic moments of its constituent grains. Within any particular grain, the magnetic moment has a direction, conventionally called the “easy” direction of magnetization, which is determined by the magnetic anisotropy of the grain. This, in turn, depends on the intrinsic magnetic properties of the magnetic mineral. For example, magnetite has a cubic structure and a comparatively strong spontaneous magnetization ($M_s = 4.8 \times 10^5 \text{ A m}^{-1}$). The shape of a non-spherical magnetite grain dominates its anisotropy so that the magnetic moment favors the direction of the longest axis of the grain. In contrast, hematite has a rhombohedral structure and a weak spontaneous magnetization ($M_s \sim 2.5 \times 10^3 \text{ A m}^{-1}$). It has a strong magnetocrystalline anisotropy related to the crystal structure that causes the magnetic moment to lie in the basal plane of the crystal at room temperature.

In very fine grains of ferrimagnetic minerals, the anisotropic magnetic energy of the grain constrains the spontaneous magnetization direction to lie along an easy direction, and the entire grain is uniformly magnetized as a *single domain*. The high spontaneous magnetization of magnetite creates anisotropic magnetostatic energy related to the particle shape. The theoretical range of single domain sizes in magnetite is narrow, from about 0.03 to 0.1 μm in equant grains and up to about 1 μm in elongate grains. In contrast, hematite has a large magnetocrystalline anisotropy energy, and single domain grain sizes may extend up to 15 μm . The magnetization of a single domain particle is very stable, because to change it requires rotating the entire uniform

magnetization of the grain against the grain anisotropy. The magnetic field required to reverse the direction of magnetization of a single domain grain is its *coercivity* B_c . In magnetite, this field is commonly in the range 10–100 mT, but may be up to 0.3 T in acicular grains. The magnetic properties of hematite are variable and its maximum coercivity can exceed 2 T. Because of their stable remanent magnetizations, single domain particles play a very important role in paleomagnetism.

When the grain size of a magnetic mineral exceeds a critical value, the magnetostatic energy of the grain becomes too large for the grain to exist as a single domain and the magnetization subdivides into oppositely directed domains. *Multidomain* magnetizations can be changed by merely moving the wall between adjacent domains, so are less stable against change than single domains. As a result, coarse-grained rocks are often unsuitable for paleomagnetism. Laboratory methods have been developed to identify the magnetic minerals in a rock, and to evaluate whether the grain size is single domain, multidomain, or intermediate (the latter category being termed pseudo-single domain).

The magnetic moment of each grain in a rock aligns with a magnetizing field as closely as permitted by the magnetic anisotropy of the grain. Thermal energy acts to destroy the partial alignment of magnetic moments. As a result, a magnetization that originally has the value M_0 decays exponentially with time, so that after time t the remaining magnetization M_r is given by

$$M_r = M_0 \exp(-t/\tau)$$

The rate of decay is governed by the *relaxation time* τ , which is a function of the magnetic properties of the grain and the temperature T . For single domain grains of magnetite

$$\tau = \frac{1}{f} \exp\left(\frac{VB_c M_s}{2kT}\right) \quad (2)$$

Here f is a frequency factor ($\sim 10^9 \text{ s}^{-1}$) related to the vibrational frequency of the crystal lattice; V is the volume of the grain, B_c its microscopic coercivity, and M_s its spontaneous magnetization; k is Boltzmann's constant. A single domain grain of magnetite, 30 nm in diameter, has a relaxation time of about 10^9 years at 20 °C. Other ferrimagnetic minerals can also carry magnetizations that are very stable against thermal decay. The high stability enables a rock to retain the remanent magnetization it acquired during rock formation for very long intervals of geological time. The relaxation time of the magnetization decreases with increasing temperature and with decreasing grain size. Thus, heating during tectonic deformation, or a change of grain size or mineral chemistry as a result of weathering, can modify or destroy a primary remanent magnetization.

Remanent Magnetizations

The development of paleomagnetism as a scientific discipline was made possible by the invention in the early 1950s of the astatic magnetometer, a device that was able to measure accurately the weak magnetizations of rocks. The instruments commonly used in modern paleomagnetic laboratories are spinner magnetometers and cryogenic magnetometers, which are very sensitive and can measure the weak magnetization of a rock very quickly. The concentrations of ferrimagnetic minerals are low, so rocks have very weak remanent magnetizations, measuring typically $\sim 10^{-1} \text{ A m}^{-1}$ for a basaltic lava and $\sim 10^{-3} \text{ A m}^{-1}$ for a pelagic limestone. The mechanism by which a rock becomes magnetized in the Earth's magnetic field has a strong effect on the intensity of the remanent magnetization it acquires.

When an igneous rock forms, its temperature is initially well above the Curie temperature of ferrimagnetic minerals (e.g., 580 °C in magnetite, 675 °C in hematite). The magnetic behavior is paramagnetic, incapable of retaining a magnetization. As it cools below the Curie point, the magnetite or hematite becomes ferrimagnetic. At high temperature, the relaxation time is short and the magnetic behavior is classified as *superparamagnetic*. With further cooling the relaxation time increases rapidly. A few tens of degrees below the Curie point a *blocking temperature* is reached, at which the relaxation time is so large that the magnetization becomes blocked by the anisotropy of the grain. If this occurs in a magnetic field, the ferrimagnetic grains acquire a *thermoremanent magnetization* (TRM) in the direction of the field. Equation (2) controls this process and ensures that the TRM has a high stability. On the other hand, if subsequent chemical alteration of the grain occurs, the grain volume V or spontaneous magnetization M_s may alter, which – again according to Eq. (2) – can give rise to a new *chemical remanent magnetization* (CRM). This can be an undesirable stable secondary component.

Sedimentary rocks that form from slowly deposited sediments can often be excellent archives of the ancient magnetic field. Magnetite grains form a tiny fraction of a sediment (e.g., less than 0.01% of a limestone). They acquire a partial alignment with the magnetic field in which they settle, forming a *depositional remanent magnetization* (DRM). In some very fine-grained sediments, further magnetite alignment takes place in the water-logged pore spaces between the non-magnetic matrix minerals. The sedimentation rates of pelagic marine sediments are very low and paleosecular variations are in principle averaged out within the size of a standard 1-in paleomagnetic sample. However, in lake sediments, the sedimentation rate is much higher and it is possible to study both the paleosecular variation (Fig. 1) and paleoclimatic effects in these sediments. Shallow-water marine limestones also have a relatively high deposition rate, but the detrital and biogenic

fractions that supply the magnetite are diluted in these limestones and they tend to have weak or unstable remanent magnetizations. However, deep-water marine sediments with high sedimentation rates from the north Atlantic have been used to simultaneously match relative paleointensity variations and oxygen isotope records that reflect paleoclimate (Channell et al. 2009).

In a paleomagnetic study, special laboratory techniques are used to isolate the components of magnetization present in the rock and to identify the primary component (see ► “Paleomagnetism, Measurement Techniques and Instrumentation”). This is the remanent magnetization acquired during formation of the rock. The remanent magnetization of a rock in situ is called its *natural remanent magnetization* (NRM). It may contain a primary component of geological significance, but this can be accompanied by secondary components of later origin. These can arise from geological processes, but some may also be induced during sampling. The secondary components can partially – in extreme cases, completely – overprint the primary component. The components of a rock magnetization must be understood for a paleomagnetic analysis to have geological significance.

By progressively demagnetizing the samples, the structure of the NRM can be analyzed and stable components isolated. There are two main methods of achieving this. When a rock sample is placed in an alternating field (AF), the magnetizations of all domains with coercivity less than the peak field are randomized. If the peak field is slowly reduced to zero and the equipment is shielded from the Earth’s field, this part of the rock magnetization is demagnetized. The part that remains has been “magnetically cleaned.” If the procedure is carried out in progressively increasing peak fields, the NRM decay shows the components present. Similarly, when a rock sample is heated to a given temperature T , components that have lower blocking temperatures than T are thermally randomized. If the sample is then cooled in field-free space, this part of the NRM remains demagnetized. The progressive destruction of the NRM by AF or thermal demagnetization reveals the components present in the NRM. The stable primary component (also called the *characteristic magnetization*) is identified computationally by principal component analysis of the directions during magnetic cleaning (Kirschvink 1980) and graphically using vector diagrams (Zijderveld 1967).

In addition to the remanent magnetizations acquired in a geological process, synthetic remanent magnetizations may be imparted to rock samples in the laboratory for the purpose of identifying and describing the carrier of a paleomagnetic signal. If a sample is placed in a magnetic field, the magnetic grains with coercivity less than the field are realigned; when the field is removed, an *isothermal remanent magnetization* (IRM) is left in the sample. If the field is strong enough to magnetically saturate the sample, a *saturation IRM* ensues.

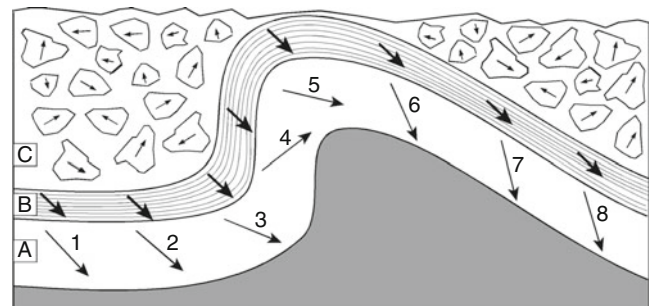
Another important laboratory magnetization is produced by placing a sample in an alternating magnetic field in the presence of a constant bias field. When the alternating field is reduced to zero, an *anhysteretic remanent magnetization* (ARM) is produced in the sample. Both IRM and ARM can be used in various ways to gain information about the magnetic mineral and its grain size.

For paleomagnetic results to be a valid record of the ancient magnetic field, it is important to know that the remanent magnetization has the same age as the rock. Identification of the magnetic mineral is important, but this is not a sufficient criterion to establish that the magnetization is likely to be unaltered. Field tests play a decisive role in establishing whether a magnetization component is primary or secondary. The most important are the structural fold test and the magnetic reversals test; less common, but also useful, are the conglomerate test and baked contact test.

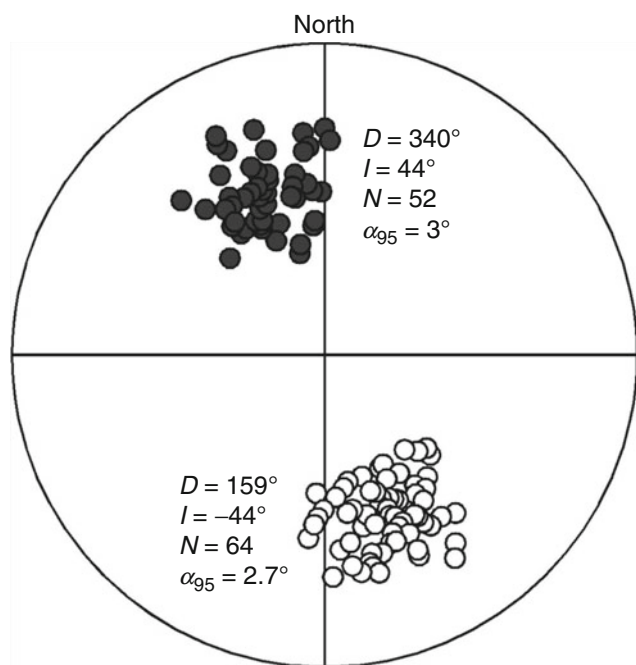
The principles underlying the fold and conglomerate tests are illustrated in Fig. 3. Layers A and B are folded rocks, and layer C is a conglomerate layer of large cobbles. The *fold test* examines the directions of magnetization around a fold. If the magnetization of a formation is stable, as in layer A, its direction will change from site to site around the fold; if it is unstable, as in layer B, all sites in the layer will have a common magnetization direction.

If a sedimentary formation is deposited over a long time interval, or if a collection of similarly aged igneous rocks covers enough time so that both normal and reversed intervals of geomagnetic polarity are represented, the *reversals test* may be applied. If paleosecular variation is adequately averaged out, the reversed directions should be antipodal to the normal directions, as in Fig. 4.

Less commonly used are the conglomerate and baked contact tests. The *conglomerate test* examines whether the cobble magnetizations have random directions (Fig. 3), which would indicate that the magnetization of the parent formation



Paleomagnetism, Principles, Fig. 3 The fold and conglomerate tests of the paleomagnetic stability of magnetization directions. The cobbles in the conglomerate (layer C) have random directions of magnetization, which indicates these are stable. The directions are also stable in layer A, changing around the fold, but in layer B their uniformity indicates they have been remagnetized after folding



Paleomagnetism, Principles, Fig. 4 Positive reversals test in a section of the *Scaglia cinerea* formation in Umbria, Italy. The mean of 52 normal directions (solid points) has declination 340° , inclination 44° ($\alpha_{95} = 3.0^\circ$) and differs by less than 1° from the antipode of the mean of 64 reversed directions (open points), which has declination 159° , inclination -44° ($\alpha_{95} = 2.7^\circ$)

from which the cobbles derive is stable. The *baked contact test* is applied to igneous flows and intrusions. The heat of a lava dike can remagnetize the contact zone of a rock it intrudes, or on which it flows. If the paleomagnetic directions in the baked contact zone agree with those of the lava and differ from the rest of the host rock, the lava has a primary magnetic direction.

In each of these field tests, rigorous statistical methods are used to decide the significance of the magnetization directions.

Statistical Testing of Paleomagnetic Data

Paleomagnetic studies are based on a hierarchical sampling system. A comparatively small number of samples, n , is taken at each of N sampling sites in a geological formation of interest; N and n are each commonly around 10–20. Statistical tests are carried out on the measured data at each hierarchical level. Testing the significance of the results is based on a method developed by Fisher (1953) that represents the paleomagnetic directions as unit vectors. These can be visualized as a distribution of points on a sphere, represented by a stereographic plot. The vector mean of the unit vectors is the best estimate of the population mean direction. Fisher suggested that the probability density function P of the distribution of

deviations between individual sample directions and their mean direction is given by

$$P(\theta, \kappa) = \frac{\kappa}{4\pi \sinh \kappa} \exp(\kappa \cos \theta)$$

The “precision parameter” κ describes the dispersion of the points. If $\kappa = 0$, the directions are uniformly or randomly distributed; if κ is large, the points cluster tightly about their mean direction. The best estimate of κ is k , defined as

$$k = \frac{N - 1}{N - R}$$

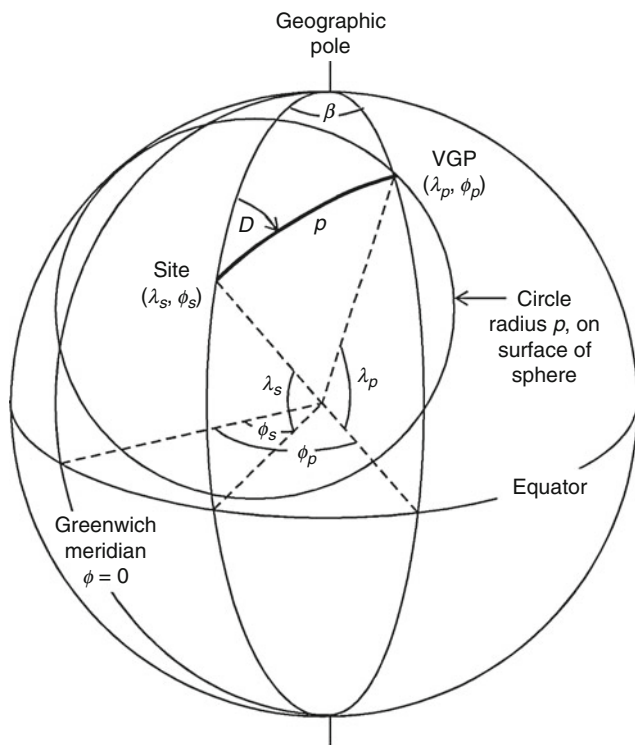
in which R is the vector sum of the N unit vectors. In paleomagnetism, the level of 95% confidence is used to define a circle with radius α_{95} around the estimated mean direction. This means that there is a 95% probability that the true mean of the distribution lies within this circle. An approximate formula (McElhinny and McFadden 2000) for the radius of the circle of confidence, especially appropriate for the small numbers of samples or sites in a paleomagnetic study, is

$$\alpha_{95} = \frac{140}{\sqrt{kR}}$$

The statistical definition of a confidence circle around the mean direction allows comparison of a paleomagnetic result with other data of the same age and from the same region. It allows critical evaluation of the directions before and after tectonic corrections (the fold test), as well as the equivalence of antipodal directions in a reversals test. Statistical analysis lends authority to the results and is an important aspect of all paleomagnetic investigations. The statistical methods can be applied to paleomagnetic directions and also to distributions of paleomagnetic pole positions.

Virtual Geomagnetic Poles and Apparent Polar Wander Paths

Assuming the GAD hypothesis, the location of the magnetic pole at the time of formation of a given rock formation is coincident with the geographic pole. As a result of subsequent plate tectonic motions, the location of the paleomagnetic pole becomes displaced. When the magnetizations of the rocks are measured, they appear to have been magnetized by a pole at the displaced location, which is accordingly designated the *virtual geomagnetic pole* (VGP). The principle of locating a VGP from the measured declination and inclination of the stable component of remanent magnetization in a rock is illustrated in Fig. 5. First, the inclination is used to calculate the angular distance p to the magnetic pole, employing Eq. (1). This defines a circle with radius p centered on the sampling site; it is the locus of possible pole locations. The



Paleomagnetism, Principles, Fig. 5 Method for locating the virtual geomagnetic pole in a paleomagnetic study. The distance p to the pole is computed from the paleomagnetic inclination I ; the VGP is located at the intersection of the circle with radius p and the meridian with declination D

VGP is located at the intersection of this circle with the meridian defined by the declination; a magnetic pole at this location gives the measured inclination and declination at the sampling site.

The VGP positions of samples from geologically young rocks cluster near to the geographic pole, corresponding to the GAD hypothesis. Older rocks from the same continent give VGP positions that group well away from the present rotation axis. They plot systematically along a curved path, called an apparent polar wander (APW) path, which appears to trace the motion of the paleomagnetic pole relative to that continent. However, paleomagnetic data define different APW paths for each continent. Assuming the validity of the GAD hypothesis, the mean paleomagnetic pole at any time must lie on the rotation axis. The Earth has only a single rotation axis, so it follows that each APW path is a record of the motion of its continent relative to the rotation axis. The APW paths for different continents diverge because the continents have been displaced relative to each other. This motion of the continents was originally called “continental drift” and is the result of plate tectonics (see ► [“Plates and Paleoreconstructions”](#)).

The use of paleomagnetic data to trace continental movements is illustrated by comparing the APW paths for Europe and North America prior to the Jurassic. They have similar

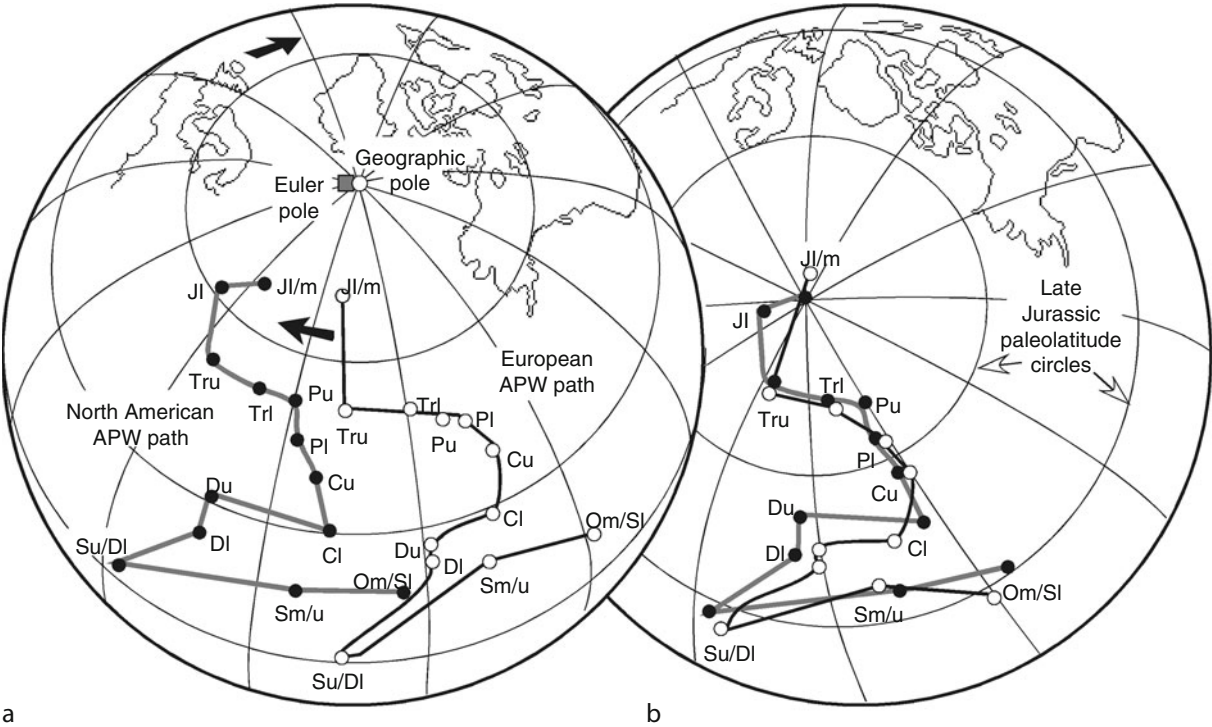
shapes (Fig. 6a), particularly for the Upper Carboniferous (Cu) to the Upper Triassic (Tru). The relative motions of tectonic plates on the surface of the spherical Earth are described by rotating pairs of plates about a common pole, known as the *Euler pole* of rotation. For example, the Euler pole of the European and North American plates in Fig. 6a is close to the present-day geographical pole. The APW path of a continent is constrained to move with the continent. Thus, the Cu-Tru segments of the two APW paths can be made to overlap by a clockwise rotation of Europe through 38° about this Euler pole (Fig. 6b). This indicates that from the Upper Carboniferous to the Upper Triassic Europe and North America belonged to the same supercontinent, *Laurasia*, itself the northern part of *Pangea* (see ► “[Plates and Paleoreconstructions](#)”). Later segments of the APW paths diverge as a result of the relative motion between Europe and North America since the Early Jurassic.

Continental reconstruction based on paleomagnetic data requires that the APW paths are securely defined. In order to ensure this, stringent selection criteria have been developed to ensure that the paleomagnetic data accepted for incorporation into the reference database are of satisfactory quality (Table 1).

Magnetic Polarity Reversals

The geomagnetic field has reversed polarity numerous times, apparently spontaneously and randomly. The record of polarity reversals has been studied in detail in igneous and sedimentary rocks and in unconsolidated sediments. Beginning in the 1950s, systematic studies of paleomagnetic polarity were made in radiometrically dated lavas for the most recent 5 Million years of Earth history. The history of magnetic reversals was confirmed by investigations in slowly deposited marine sediments covering the same time interval. The agreement between the igneous and sedimentary polarity sequences established that the reversals are a geomagnetic phenomenon and not merely a rock magnetic artifact.

Intervals of constant normal or reverse polarity are called polarity *chrons*; they last typically for 5×10^4 years to 5×10^6 years. The chrons may be interrupted at irregular intervals by shorter polarity *subchrons* that last on the order of 10^4 years. The *transition* from one polarity to the opposite one is estimated to take about 3–7 thousand years, which is much shorter than a polarity chron or subchron. The polarity transition is accompanied by a decrease in intensity of the paleofield, which may last up to twice as long as the directional change. The polarity record also shows occasional large *excursions*, in which the magnetic pole appears to spend $\sim 10^3$ years in equatorial latitudes, but instead of completing a reversal it returns to its initial location on the rotation axis.



Paleomagnetism, Principles, Fig. 6 Reconstruction of earlier positions of Europe and North America based on APW paths. (After Fig. 5 in Van der Voo 1990a. Redrawn by permission of John Wiley and Sons. Copyright 1990 by the American Geophysical Union). In (a) the APW

paths are shown in present coordinates; in (b) the European continent has been rotated by 38° clockwise about the Euler pole shown in (a) to bring the Cu-Tru segments of the APW paths into coincidence

Paleomagnetism, Principles, Table 1 Reliability criteria for paleomagnetic data. Van der Voo (1990b) suggested that acceptable paleomagnetic data should meet at least 3 of the following criteria

No.	Brief description of reliability criterion
1	Well-determined rock age and a presumption that the magnetization has the same age
2	Sufficient number of samples ($N \geq 24$), adequate precision of directional grouping ($k \geq 10$), and confidence angle $\alpha_{95} \leq 16^\circ$
3	Progressive demagnetization that demonstrably includes vector subtraction information
4	Field tests that constrain the age of magnetization
5	Structural control and tectonic coherence with the craton or block involved
6	The presence of reversals
7	No resemblance to paleomagnetic poles of younger age (by more than a period)

The irregular pattern of polarity intervals forms a geological “fingerprint,” which can be used to date and correlate other magnetic polarity sequences in both igneous and sedimentary rocks, and in sediments. The most important application has been the interpretation and dating of marine magnetic anomalies. New oceanic lithosphere forms at oceanic ridge systems and is then transported away from the ridges by the sea-floor spreading process (see ► “Sea-floor Spreading”). As the basaltic lava of the ocean crust cools, it acquires a TRM parallel to the Earth’s magnetic field. Sea-floor spreading persists for millions of years at a ridge, during which the magnetic

field changes polarity many times. This leaves some blocks of oceanic crust normally magnetized parallel to the field while their neighbors are reversely magnetized. A sequence of positive and negative magnetic anomalies, corresponding to normal and reverse polarities of the oceanic crust, is observed over the ridges; from this pattern, the sequence of polarity of the Earth’s magnetic field is obtained. The oceanic polarity pattern correlates with the polarity sequence in lavas that have been dated radiometrically, thus the ages of magnetized blocks of the ocean floor can be determined. Knowing the distances of the polarity changes from the oceanic ridge, the rate of sea-floor spreading can be calculated.

Older reversal patterns have been dated by magnetostratigraphy. This entails stratigraphic sampling of a sedimentary section that has a uniform lithology and was deposited at a uniform rate. Standard paleomagnetic techniques are used to evaluate the characteristic direction of each sample and establish the existence of *magnetozones* of constant polarity. As is the case with paleomagnetic data, strict reliability criteria regulate the acceptance of a magnetic stratigraphy into a global database (Table 2).

An example of coordinated magnetostratigraphy and biostratigraphy that correlates well with marine magnetic reversal history is shown in Fig. 7. The *Scaglia cinerea*, a marly limestone in Umbria, Italy, was deposited in the Late Paleogene and acquired a DRM that preserves the direction and polarity reversals of the geomagnetic field during deposition.

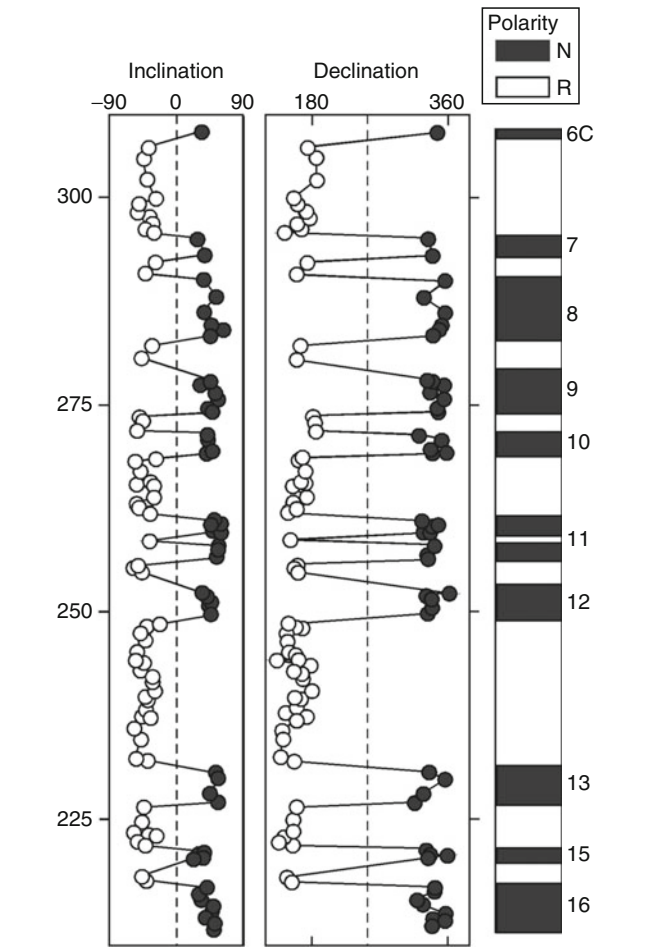
Paleomagnetism, Principles, Table 2 Reliability criteria for magnetostratigraphy. Opdyke and Channell (1996) proposed that an acceptable magnetostratigraphic section should pass at least 5 of the following 10 tests

No.	Brief description of reliability criterion
1	Stratigraphic age known to the level of the stage, and associated paleontology presented adequately
2	Sampling localities placed in a measured stratigraphic section
3	Complete thermal or alternating field demagnetization performed, and analysis of magnetization components carried out using orthogonal vector projections
4	Stable paleomagnetic directions determined from principal component analysis
5	Data published completely as plots of VGP latitudes and/or of declination and inclination against stratigraphic position. The statistical parameters should be fully documented
6	Magnetic mineralogy determined
7	Field tests to constrain the age of magnetization undertaken where possible
8	A positive reversals test should be carried out
9	Radiometric ages, especially ⁴⁰ A/ ³⁹ Ar or U-Pb ages from volcanic ashes or bentonites, available in the stratigraphic section
10	Multiple overlapping sections confirm the polarity sequence

Paleontological dating of the limestones enabled correlation of major stage boundaries with the polarity sequence. The absolute ages of many stage boundaries are known independently, so the reversal pattern could be dated. Correlation with marine magnetic anomalies allowed the corresponding regions of the oceanic crust to be dated. The polarity sequence correlates well with the history of geomagnetic polarity derived from marine magnetic anomalies for chrons numbered 6C to 16 (Cande and Kent 1992).

Magnetozones in sediments can now be dated more accurately by means of cyclostratigraphy. This method is based on cyclical changes in the orientation of Earth’s rotation axis and the shape and orientation of its orbit. These are caused by the gravitational attractions of the Sun, Moon, and other planets (especially Jupiter). The changes (known as Milankovitch cycles) modify the insolation of the Earth and have long-term paleoclimatic effects on the deposition of sediments. They show as cyclical changes in color, bedding thickness, magnetic concentration, and isotopic ratios. The durations of the cycles are well known from astronomy, so by identifying and counting them they can be used to date a sedimentary section with greater accuracy. Cyclostratigraphy is an important dating method for determining the ages and durations of polarity zones in a magnetic polarity stratigraphy. Originally applied to Neogene sediments, the method has been extended to older Mesozoic rocks.

The correlation of dated polarity sequences from sections on land and in marine cores with the polarity record from oceanic magnetic anomalies is used to produce a geomagnetic polarity time-scale (GPTS). The record of geomagnetic polarity is well established from the present back to the Late Jurassic



Paleomagnetism, Principles, Fig. 7 Declination and inclination profiles in the Late Paleogene *Scaglia cinerea* formation, whose directions are shown in Fig. 4, correlated to the history of geomagnetic polarity derived from marine magnetic anomalies for chrons numbered 6C to 16

(Oxfordian), equivalent to the last 155 Million years of geological time. Three distinct episodes of reversal behavior are identified in this time interval. The oldest episode, comprising frequent reversals in the Late Jurassic and Early Cretaceous, is referred to as the M-sequence. The youngest episode, comprising the Late Cretaceous and Cenozoic, consists of frequent reversals known as the C-sequence. The M-sequence and C-sequence are separated by an interval, about 40 Ma in length, called the Cretaceous Normal Polarity Superchron (CNPS), in which correlated magnetic anomalies are absent.

The episodes of reversing behavior are long enough to make tentative interpretations about the reversal process. It is evident from the patterns of reversals that they are not periodic. Making allowance for the time spent in transition between opposite polarity states, statistical analysis indicates that the reversals in the C- and M-sequences occur randomly. The CNPS is interpreted as a time in which the Earth’s magnetic field did not reverse polarity. There are competing interpretations of why this should be. Whereas some investigators consider that the CNPS represents a special behavior of

the geodynamo, others regard it as an exceptionally long feature in a continuous reversal process.

The break-up of Pangea and the onset of sea-floor spreading date from the Early Jurassic, about 180 Million years ago. Between then and the earliest part of the marine polarity record, the corresponding regions of the ocean floor have low-amplitude magnetic anomalies but their confirmation in magnetostratigraphic records is fragmentary. For earlier eras, there is no marine polarity record to which a magnetic stratigraphy can be correlated. Despite this absence, a geomagnetic polarity time scale, dated by cyclostratigraphy, has been established for the Late Triassic and Early Jurassic (Kent et al. 2017). This was achieved by correlating magnetic stratigraphies in overlapping cores of continental sediments. The durations of magnetozones were dated by cyclostratigraphy, and the combined sequence was anchored by a radiometric age. The successful study indicates how a GPTS for earlier times may be established.

Environmental Magnetism

Paleomagnetism depends on the abilities of magnetic minerals to retain a stable magnetization for geologically long periods of time. The same properties make the minerals useful carriers of information about past environments and paleoclimates. They also enable investigations of present-day regional environmental pollution by heavy industry and

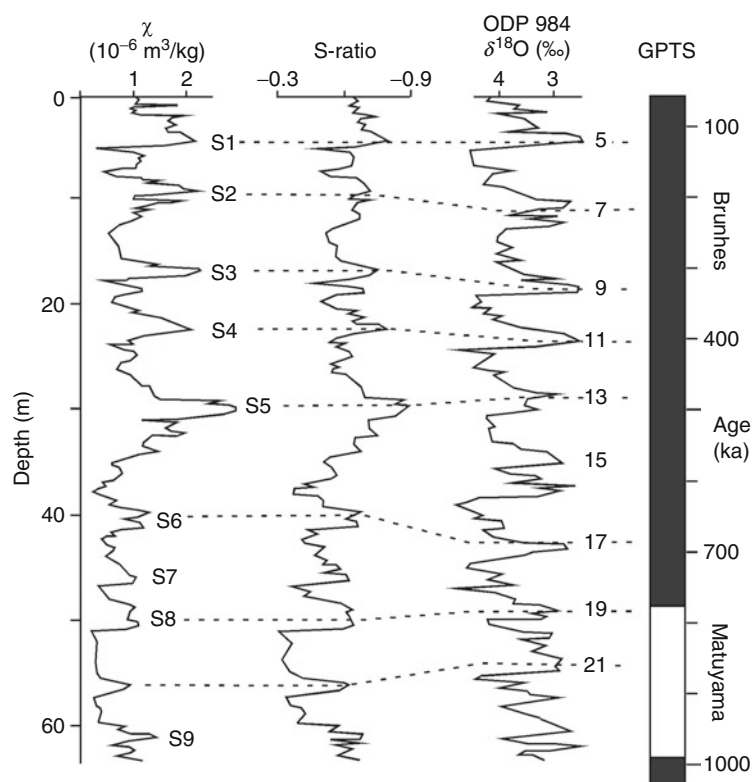
traffic, by using the concentrations of ferromagnetic minerals to trace heavy metal contamination.

Most environmental magnetic studies have been made in sedimentary archives. The magnetic susceptibility of a sediment reflects both the concentration and type of the magnetic minerals, whereas variations in grain size may be revealed in the stratigraphic variations of magnetic coercivity and remanent magnetizations produced in the laboratory. The proportion of an IRM carried by grains with coercivity above a certain value (the so-called S-ratio) can be an indicator of the relative size of the fine-grain fraction of the ferromagnetic grains. Lake sediments and wind-blown loess deposits have comparatively high deposition rates. Detailed evaluation of the magnetic signal in these sedimentary archives can reveal variations in environmental conditions during deposition. The magnetic properties of Chinese loess deposits illustrate the way in which magnetic parameters can act as proxy for paleoclimatic indicators.

The boundary between the Brunhes and Matuyama polarity chrons (age 780 ka), and the upper and lower boundaries of the Jaramillo normal polarity subchron (990 ka and 1070 ka, respectively), tie the loess section at Sanmenxia, China, to deep-sea sediments cored at Ocean Drilling Program Site 694 (Fig. 8). The wind-blown loess sediments in central China were deposited in cold climatic conditions. Paleosols formed under warmer, moister conditions that were conducive to enhancement of the magnetic susceptibility. Alternations of loess deposits and paleosols are

Paleomagnetism, Principles,

Fig. 8 Profiles of magnetic susceptibility (χ), hardness of remanent magnetization (S-ratio), and paleomagnetic polarity in a loess section at Sanmenxia, China, correlated to the marine oxygen-isotope ratio ($\delta^{18}\text{O}$) at ODP site 984. (Compiled from data in Wang et al. 2006). The peaks marked S1, S2, etc., correspond to paleosols; the lows between them are loess beds



accompanied by fluctuations in magnetic susceptibility, with low values marking the loess and higher values in the paleosols. Oxygen isotope results at the ODP site reveal cycles of climatic variation, which correlate well with the magnetic susceptibility and S-ratio data from the Sanmenxia section. The results demonstrate the usefulness of magnetic mineralogy parameters as proxies for isotopic data that document paleoclimatic variation.

Summary

Apart from isolated but important early pioneering studies, in which the characteristics of the geomagnetic field and the rudiments of rock magnetism were explored, paleomagnetism began to expand as a scientific discipline in the 1950s. As a result of the continuous development and refinement of instrumentation and analytic techniques, it has made many notable contributions to geological knowledge. Initially, it focused on the tectonic problem of “continental drift” and, by describing apparent polar wander paths for the different continents, it documented incontrovertibly the mobility of the continents. Matching of APW paths led to the reconstruction of supercontinents such as Rodinia, Pangea, Gondwana, etc. In parallel with this work, paleomagnetic analysis in conjunction with geological dating worked out the history of geomagnetic reversals. On the basis of the established reversal history and the rock magnetic properties of oceanic basalts, the hypothesis of sea-floor spreading was developed. This provided the key to the mechanism of plate tectonics. It allowed the documentation and dating of plate motions, and explained how “continental drift” took place. In turn, the pattern of magnetic reversals in the oceanic crust established a history of geomagnetic polarity for the past 155 Million years that has been confirmed by magnetostratigraphy and dated by radiometric, paleontological, and astrochronological methods. Paleomagnetism and rock magnetism have found wider application in addressing environmental problems and in contributing to an understanding of paleoclimatic history.

Cross-References

- [Curie Temperature](#)
- [Geomagnetic Field, IGRF](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Geomagnetic Field, Secular Variation](#)
- [Paleomagnetism, Magnetostratigraphy](#)
- [Paleomagnetism, Measurement Techniques and Instrumentation](#)
- [Paleomagnetism, Polar Wander](#)
- [Plates and Paleoreconstructions](#)
- [Remanent Magnetism](#)
- [Seafloor Spreading](#)

Bibliography

- Cande SC, Kent DV (1992) A new geomagnetic polarity time scale for the late cretaceous and Cenozoic. *J Geophys Res* 97:13917–13951
- Channell JET, Xuan C, Hodell DA (2009) Stacking paleointensity and oxygen isotope data for the last 1.5 Myr (PISO-1500). *Earth Planet Sci Lett* 283:14–23
- Fisher RA (1953) Dispersion on a sphere. *Proc R Soc London, Ser A* 217:295–305
- Gubbins D, Jones AL, Finlay CC (2006) Fall in Earth’s magnetic field is erratic. *Science* 312:900–901
- Jackson A, Jonkers ART, Walker MR (2000) Four centuries of geomagnetic secular variation from historical records. *Philos Trans R Soc Lond A* 358:957–990
- Kent DV, Olsen PE, Muttoni G (2017) Astrochronostratigraphic polarity time scale (APTS) for the late Triassic and early Jurassic from continental sediments and correlations with standard marine stages. *Earth Sci Rev* 166:153–180
- Kirschvink JL (1980) The least-squares line and plane and the analysis of paleomagnetic data. *Geophys J R Astron Soc* 62:699–718
- McElhinny MW, McFadden PL (2000) *Paleomagnetism: continents and oceans*. Academic, San Diego
- Opdyke ND, Channell JET (1996) *Magnetic stratigraphy*. Academic, San Diego
- Schneider DA, Kent DV (1990) The time averaged paleomagnetic field. *Rev Geophys* 28:71–96
- Tauxe L (2005) Inclination flattening and the geocentric axial dipole hypothesis. *Earth Planet Sci Lett* 233:247–261
- Turner GM, Thompson R (1981) Lake sediment record of the geomagnetic secular variation in Britain during Holocene times. *Geophys J R Astron Soc* 65:703–725
- Van der Voo R (1990a) Phanerozoic paleomagnetic poles from Europe and North America and comparisons with continental reconstructions. *Rev Geophys* 28:167–206
- Van der Voo R (1990b) The reliability of paleomagnetic data. *Tectonophysics* 184:1–9
- Wang X, Yang Z, Løvlie R, Sun Z, Pei J (2006) A magnetostratigraphic reassessment of correlation between Chinese loess and marine oxygen isotope records over the last 1.1 Ma. *Phys Earth Planet Inter* 159:109–117
- Zijderveld JDA (1967) AC demagnetization of rocks. In: Collinson DW, Creer KM, Runcorn SK (eds) *Methods in paleomagnetism*. Elsevier, Amsterdam, pp 254–286

Paleoseismology

Shinji Toda
Disaster Prevention Research Institute, Kyoto University, Uji,
Kyoto, Japan

Synonyms

Earthquake archaeology; Earthquake geology

Definition

Paleoseismology An interdisciplinary field dealing with prehistoric or pre-instrumental earthquakes, in particular, their location, timing, and magnitude.

Introduction

Modern instrumental records of earthquakes cover only a little more than a century since 1890s (see ► [“Seismic Instrumentation”](#)). Historic accounts in several countries such as China, Japan, and Mediterranean regions allow us to extend the records of large earthquakes up to more than a couple of thousand years. However, for most parts of the globe, no earthquake data are available for the time periods of even the past hundreds to thousands of years that are comparable to repeat time of large earthquakes on a fault. Nevertheless, geologic sedimentary strata and geomorphic features preserve evidences for strong shaking, ground breaks, and landscape changes associated with large earthquakes. Although geologic and geomorphic records are incomplete and poorly constrained in time and space, they greatly help to understand faulting process and recurrence of earthquakes that are unavailable to seismologists due to the brevity of historic documents. We have defined this discipline as “paleoseismology” and such individual earthquakes are referred to as “paleoearthquakes.”

Paleoseismology is not a subbranch of seismology, but rather it is an interdisciplinary field of research, which borrows many concepts from structural geology and active tectonics. Methodology and techniques in paleoseismology are derived primarily from Quaternary geology and related disciplines such as geomorphology, soil mechanics, sedimentology, archaeology, paleoecology, photogrammetry, radioisotope dating or pedology (Grant 2007; McCalpin 2009).

The history of paleoseismology is as short as that of instrumental seismology. Although several pioneer geologists described surface faulting accompanying earthquakes in the late 1800s, geology and seismology at that time were separated (Yeats and Prentice 1996). But a few leading geologists, e.g., R. E. Wallace, C. R. Allen, and D. B. Slemmons, who studied San Andreas fault and active scarps in Nevada, and radiocarbon dating technique pervasive since about 1960s–1970s, played prominent roles in bridging the gulf between instrumental seismology and geology (Yeats and Prentice 1996). Since the 1970s, paleoseismic investigations have been yielding information about the paleo-movements on faults, dates of previous earthquakes, recurrence times, average slip rate, and earthquake effects over time intervals ranging from decades to thousands of years (Grant 2002). Consequently, numerous paleoseismic data (e.g., Sieh 1978) and fundamental concepts of recurrent large earthquakes (e.g., Schwartz and Coppersmith 1984) derived from field observations have largely contributed to progress in understanding the mechanics of earthquakes and faulting.

Unlike many subbranches of geology, the uniqueness of paleoseismology lies in its potential to decipher the future from the past and present. The traditional geological doctrine of “uniformitarianism,” which is used to reconstruct the past events from phenomena observable in the present, can be stretched to the future in paleoseismology to perform seismic hazard estimates and earthquake forecasting. The final goal and

major contribution of paleoseismology is to not only understand long-term faulting processes, but also to provide fundamental data and speculations to mitigate seismic hazard.

Paleoseismic Records

In seismology, estimation of earthquake magnitude depends on the performance of seismometers and on network coverage (see ► [“Seismological Networks”](#) and ► [“Statistical Seismology”](#)). Geomorphic markers, surface soil, and sediments are like poorly performing strong-motion seismograms. If an earthquake is large and shallow enough, the seismogenic faulting normally reaches the Earth’s surface. For inland intraplate and interplate earthquakes occurring shallower than ~20 km (“seismogenic layer,” see entries on ► [“Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat”](#) and ► [“Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity”](#)), $M \geq 6$ earthquakes may leave tectonic surface breaks and permanent ground offsets (Wells and Coppersmith 1994). The minimum magnitude of a paleoearthquake we can identify in strata and landscape is about $M \sim 6$ or higher (discussed later). For offshore earthquakes, including subduction associated mega-thrust events, $M \geq 7.5$ earthquakes leave records of significant coastal uplift and subsidence (e.g., Berryman et al. 1989). However, the paleoseismic record is spatially and temporally incomplete. Unlike instrumental seismology, recording processes by sedimentary layers are intermittent, and erosion often destroys the evidence for paleoearthquakes.

Fault Identification and Slip Rate

Identifying an active fault is the first step for paleoseismology. As described above, a large earthquake often leaves surface breaks and warping with significant offset when the faulting reaches the Earth’s surface. Detailed excavation studies often reveal evidence for each individual earthquake at an individual site. However, recognizing active traces relies on identification of geomorphic expression due to repetition of surface-faulting earthquakes as “fault landforms” for which deformation rate must be higher than local rates of erosion and deposition. Although the definition of “active fault” partly depends on the purposes (engineering or scientific) and on the usage of the terminology in different countries, it is usually identified by its association with tectonically deformed Quaternary-age materials or surfaces and by its potential to cause a large earthquake in near future.

To identify active faults and associated active landforms, aerial photographs and/or satellite images are commonly used (“aerial photographic interpretation”). High-resolution satellite images may be used to identify faulted landforms, even finding coseismic surface ruptures, around the globe, particularly in hardly accessible areas (e.g., Lin et al. 2006). The most frequently encountered feature is topographic lineaments. But lineaments are often developed by a combination of constructive geologic features and erosion. Therefore,

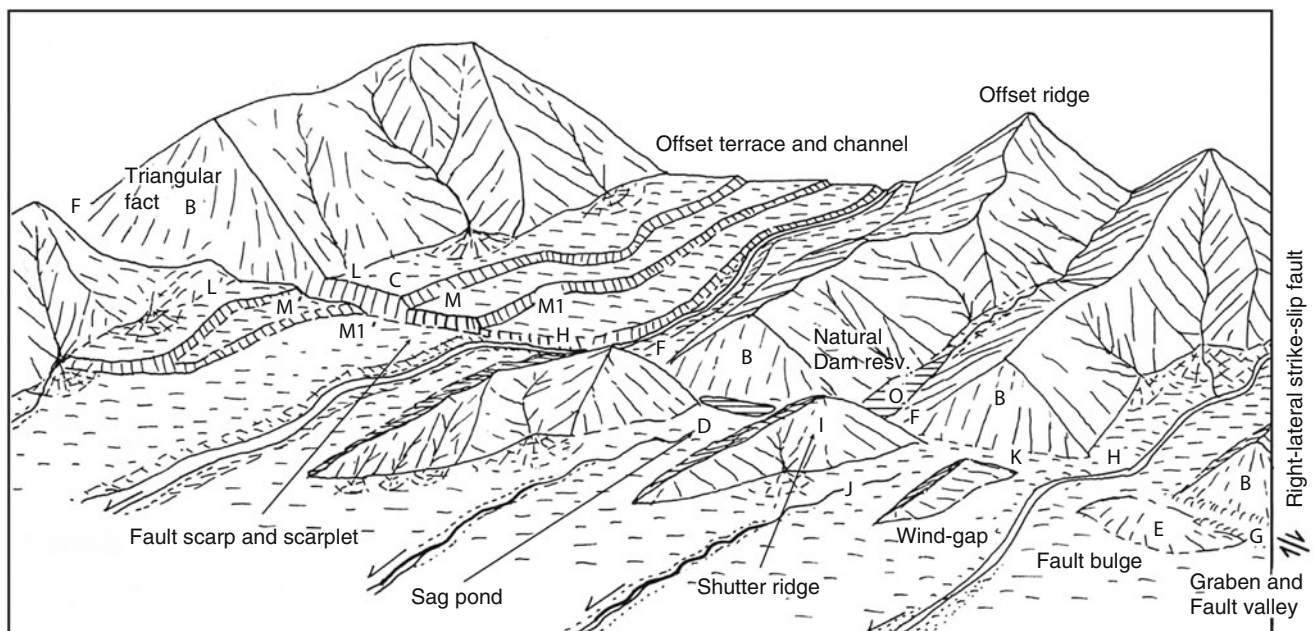
other characteristics associated with faulting such as triangular facets, fault scarps, sag ponds, and bulges should be found along a fault lineament (Fig. 1). Strike-slip faults accompany offset streams, shutter ridges, beheaded streams, wind gaps, and offset terrace risers. Multiple and systematic fault landforms along a single trace increase the reliability of its identification as an active fault trace. Most of such traces tend to be roughly along geologic faults juxtaposing different geologic units so that a traditional geologic reconnaissance is also useful to constrain an active fault zone. Fresh fault gouge (clay) and/or fault breccia are often exposed on outcrop along geomorphic traces. Geologic drilling and shallow coring are auxiliary approaches to constrain the location of a fault trace and sometimes help to estimate vertical offsets of stratigraphic units, particularly much deeper than trench excavation (e.g., Dolan et al. 2003). Deeper geologic information down to a few kilometers depth from industry oil wells provides a larger picture of fault structure and associated folds (e.g., Tsutsumi and Yeats 1999). Several other methods reveal subsurface geologic structures for the areas having shallow water table and urban areas with limited open space. For example, cone penetrometer testing (CPT) and large diameter borings can be applied for these conditions (Grant et al. 1997). The “Geoslicer” (Nakata and Shimazaki 1997), which consists of a sample tray and a shutter plate, can be inserted into the ground by a weighted vibrator to extract a sliced soil sample of larger than a few meters dimension (Takada and Atwater 2004). Geoslicer can directly sample a fault itself without losing information of strike, dip, and geologic context of the fault.

Spatial resolution and accuracy in aerial photographic images are limited. Recently, airborne Light Detection And Ranging (LiDAR) surveys, which measure bare topography irradiating a laser beam from an airplane or helicopter, have been widely used to image very detailed fault-zone topography beneath dense vegetation cover (e.g., Sherrod et al. 2004) and urbanized areas (e.g., Kondo et al. 2008). These data are at the appropriate scale (meters) and accuracy (decimeter) to provide useful measurements of the fine features of fault landforms (Arrowsmith and Zielke 2009). Airborne Laser Swath Mapping (ALSM) is also expected to find hidden active faults.

To locate an active fault trace and to confirm deformed subsurface structure various geophysical imaging techniques such as seismic reflection and refraction, ground-penetrating radar (GPR), aeromagnetic surveys, and gravity surveys are employed. The advantage of these techniques is that these are nondestructive methods for seeking deeper structures (Bond et al. 2007). Highresolution seismic reflection surveys are most pervasive to visualize shallow geologic structure associated with details of fault activity, even to reveal blind fault and fault-related fold (Fig. 2).

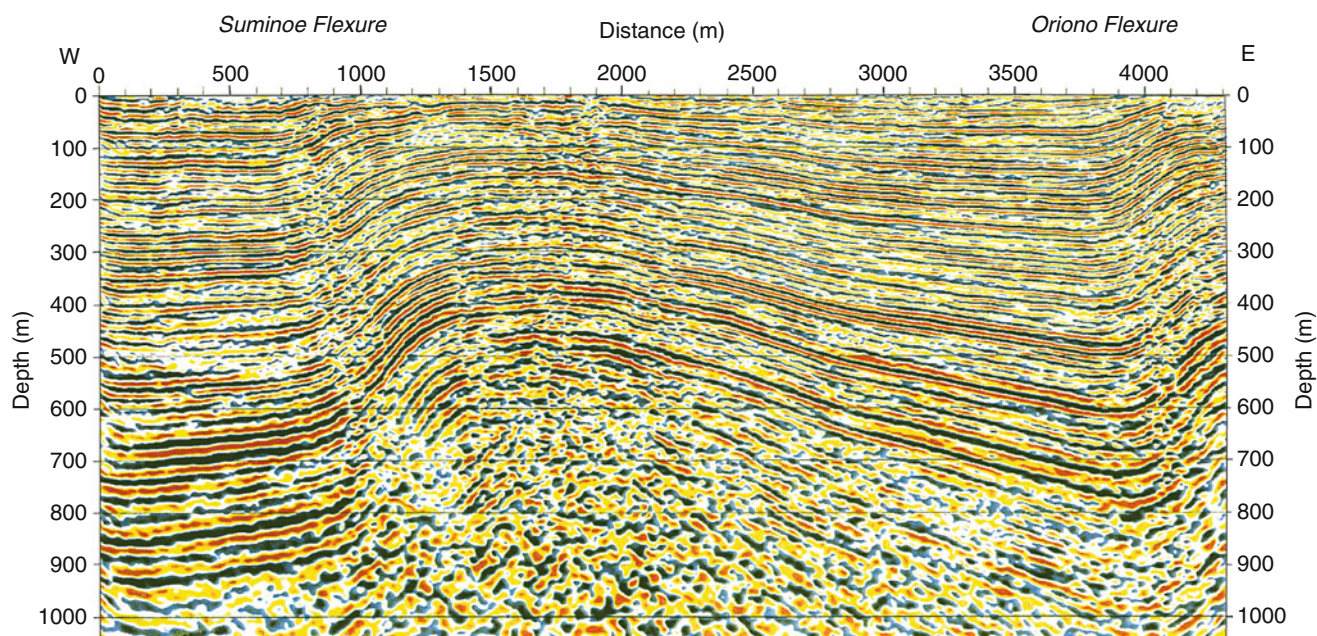
These geomorphic, geologic, and geophysical explorations for fault identification also provide information about long-term slip rates. The slip rate on a fault, S , is calculated by dividing cumulative displacement (D) across the fault by time interval T (Fig. 3),

$$S = D/T. \quad (1)$$



Paleoseismology, Fig. 1 Topographic features formed by right-lateral offset (modified from Okada 1991 and Research Group for Active Faults of Japan 1992). B, triangular facet; C, fault scarplet; D, sag pond; E,

tectonic bulge; F, fault saddle; G, graben; H, offset stream; I, shutter ridge; J, beheaded stream; K, wind gap; L-L', offset of piedmont line; M-M', offset of terrace edge; Q, fault pond



Paleoseismology, Fig. 2 An example of fault-related fold revealed by a seismic reflection survey. Depth-converted P-wave seismic reflection profile across two flexures in south Osaka, Japan (Sugiyama et al. 2003). The strata are composed of Plio-Pleistocene non-marine and marine sediments

To obtain S , we need to find “piercing lines” or “piercing points” which intersect a fault zone as a line or a point. The piercing lines and points should be formed before the subsequent faulting events occurred. Geomorphologically, offset ridges, streams, and gullies become piercing lines, but more often well-reconstructed and dateable terraces and terrace risers are employed as the best piercing lines to estimate slip rates (e.g., L-L', M-M', M1-M1' in Fig. 1). If offsets increase with terrace ages, the fault is considered to have been continuously active. Offset bedrock geology is also used as piercing lines and points to roughly estimate the very long-term slip rate. Cultural features such as roads, fences, walls, buried pipes, and rice paddy boundaries across a fault line are often used to measure the amount of coseismic slip for very recent earthquakes and historic events. In coastal and fluvial areas, where uplifted terraces are commonly seen, the vertical component of the slip rate S , is more easily obtained than the lateral component. Such slip rate estimates are used for classifying the fault activity (e.g., in Japan, class A: order of m/1,000 years; class B: order of 0.1 m/1,000 years; class C: order of 0.01 m/1,000 years), which is an important index for the estimation of the recurrence time of major earthquakes or for future seismic activity (Research Group for Active Faults of Japan 1992).

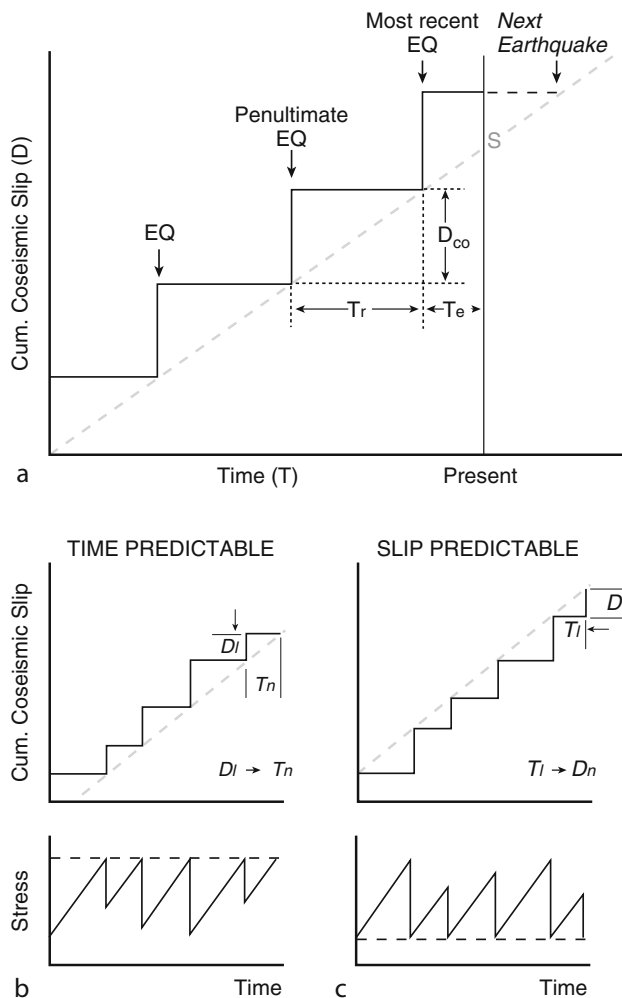
Evidence for Paleoearthquake

Paleoseismic evidence is either primary or secondary (McCalpin 2009). Primary evidence is formed by tectonic deformation resulting from coseismic slip along a source fault plane. Secondary evidence is produced by earthquake shaking. Paleoseismic features are also distinguished as being on or above fault trace (on-fault or near-field features), or

away from or far above a fault trace (off-fault or far-field features) (McCalpin 2009). Because off-fault evidence for a paleoearthquake is mostly associated with strong shaking, the problem with the off-fault evidence is its difficulty in finding the corresponding seismic source. Thus, the majority of paleoseismic studies focus only on “on-fault” evidence.

Encountering a natural exposure of faulted young sediment is extremely rare. We thus excavate trenches across a fault zone (sometimes parallel to the fault zone) to expose faults and young strata that record evidence of paleoearthquakes. The site for trenches is chosen based on the fault location, land usage, sediments preservation, paleoearthquake records, and abundant dating material. The best trench site is considered as the one where: (1) deformation zone is narrow, or fault strand is single and representative of the entire fault zone; (2) frequent but dilute sedimentation processes occur with numerous dateable materials. The trench walls are not long-lived exposures and mostly not reproducible so that detailed observations and objective records by as many observers as possible are required (Grant 2007). A trench log, which is a geologic sketch, is used for defining and mapping sedimentary units and recording evidence for surface-rupturing earthquakes, which we call an “event horizon.”

An event horizon is used as the ground surface at the time of a paleoearthquake (Pantosti et al. 1993). The event horizon may be preserved as scarp-derived colluvial wedge deposits, unconformities that terminate faults, and warping (Fig. 4), each of which is defined as an “event indicator” (Scharer et al. 2007). It is most likely to conclude that surface rupture and significant deformation such as folding and tilting along active traces are the direct products of seismogenic faulting.



Paleoseismology, Fig. 3 (a) Schematic diagram showing the cumulative displacement (D) at a site along a fault as a function of time (T). D_{co} coseismic displacement, T_r recurrence interval, T_e elapsed time since the most recent earthquake, S slip rate estimated from D/T . (b) Time-predictable model. (c) Slip predictable model. (After Shimazaki and Nakata 1980)

There are various kinds of evidences for paleoearthquakes which can be rated for their reliability.

Upward fault termination: This is the most common and easily identified event indicator in which faults and faulted strata are overlain by a non-deformed horizon (Fig. 4a). But one tends to misinterpret an event horizon in special situations where fault slip near the ground surface tapers and dies out upward (Bonilla and Lienkaemper 1990). Thus the quality and confidence level of this indicator depend on the amount of slip and stratigraphic conditions between different exposures at a site.

Fissure fills: Fissures and fissure fills are a sort of open crack associated with faulting and younger material filled during postseismic process (Fig. 4b). They are commonly seen as a wide zone of alternating tension gashes and compressional mole tracks immediately after a large surface-rupturing earthquake due to strike-slip faulting.

Colluvial wedge: The colluvial wedge is formed on the footwall side as a result of collapse of the fault scarp exposed at the time of a scarp-forming earthquake (Fig. 4c, d). A trench wall across a fault normally exposes a wedge-shaped section tapering toward or away from the fault scarp, often showing upward-coarsening structure, similar to talus deposition. When a bedrock fault plane is exposed at the earthquake, the collapsed sequence of gouge, breccia, and intact bedrock block are sequentially observed in the wedge. When it occurs within the loose sediments, scarp degradation is involved. Colluvial wedges are not only seen along normal faults but also along strike-slip and thrust faults.

Angular unconformity associated with tilted and folded strata: An angular unconformity is produced by broad deformation of subsurface sediments due to faulting and subsequent erosion and sedimentation (Fig. 4e, f). If a fault plane is simple and it is hard to find an upward termination, angular unconformity would be a substitutive criterion to identify an event horizon. But, except for significant inclination of strata, horizontal sedimentation must be guaranteed. Adhesive composition with clay and silt often shows a veneer deposit along a slope, which often mimics crustal deformation.

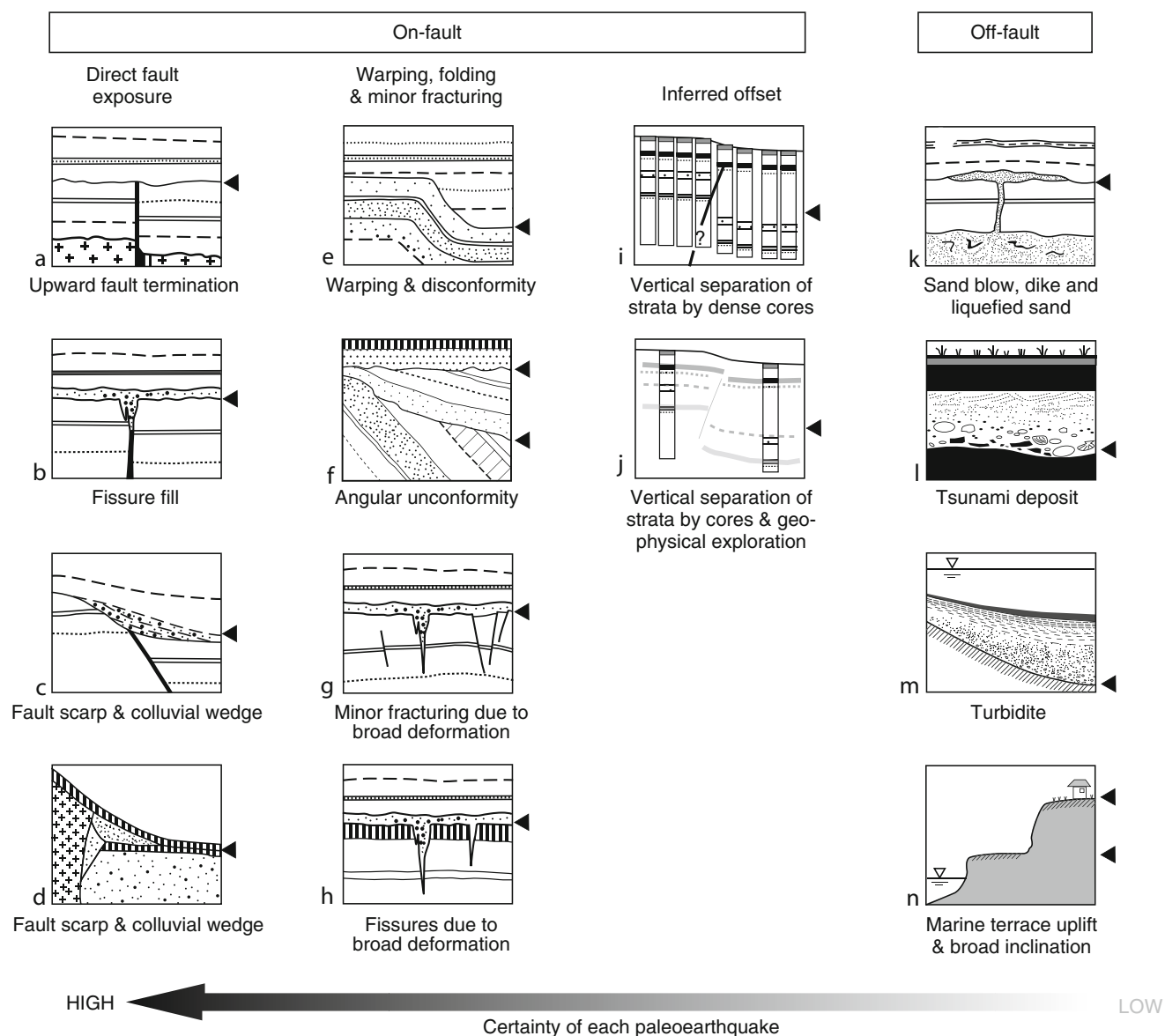
Minor cracks and faults associated with subtle deformation: Where the amount of displacement is small, and/or surface condition is dry, numerous minor cracks or tension fissures associated with deformation can be produced (Fig. 4g, h).

The deeper we expose stratigraphic units, the more events and associated structures are likely to be exposed. Since evidence associated with older events tends to be destroyed and obscured by the most recent event, we must restore the stratigraphy and structures prior to the most recent event and then reinterpret the older events to identify multiple paleoearthquakes at a site (“Retrodeformation analysis,” McCalpin 2009). It means that reliability of the paleoearthquake tends to be reduced when we go farther back in the past. To have higher confidence for a paleoearthquake, multiple exposures of a fault and/or significant deformation (event indicators) in the same stratigraphic horizon should be confirmed at a site (Scharer et al. 2007).

Time Constraints of Paleoearthquakes

Although there are several methods to directly determine ages of fault activities from fault gouges (e.g., Fukuchi 2001), timing of each paleoearthquake is commonly constrained by ages of deformed stratigraphic units and overlying undeformed strata, for which many dating approaches using different isotopes have been used (see details in Grant 2002 and references therein).

Among them, radiocarbon dating is the most powerful and pervasive technique in paleoseismic studies. In this technique, the time since the death of a plant or animal is estimated by measuring the sample’s residual ^{14}C content relative to modern standards and the half-life decay constant of ^{14}C .



Paleoseismology, Fig. 4 Schematic diagrams of typical stratigraphic event indicators of paleoearthquakes in both on-fault and off-fault regions. Solid triangle beside each panel indicates an event horizon. Several on-fault diagrams are based on Allen (1986) and Okada

(1991). Note that confidence level of paleoearthquakes increases from the diagrams in off-fault regions to the ones showing direct fault exposure

(5,568 years). The beta particle counting method conventionally requires 2–200 g of carbon and allows us to date back to ~40,000 years before present (yBP; actually years before 1950 AD). Meanwhile, newly developed Accelerator Mass Spectrometry (AMS) dating requires samples as small as a few milligrams and extends the measurable date back to 50,000–60,000 yBP. Since the ratio of ^{14}C to ^{12}C has varied with time due to variations in number of cosmic rays associated with the strength of the Earth's magnetic shield, measured radiocarbon age needs to be converted to calendar date. From dendrochronological approaches with tree rings, glacial varves, corals that annually record the amount of ^{14}C in the atmosphere, radiocarbon ages are calibrated to corresponding

calendar dates back to about 24,000 years before present (Stuiver et al. 1998).

Such dendrochronological approaches to constrain timing of a paleoearthquake have been commonly employed since Jacoby (1989). Time constraint of a paleoearthquake requires at least two radiocarbon dates that bracket an event horizon, in which probabilistic density is calculated from the calibrated dates of both samples. To have better constraints for paleoearthquakes, there are several calibration programs, one of which “OxCal” (Bronk Ramsey 2001; Lienkaemper and Bronk Ramsey 2009) not only calibrates the radiocarbon ages to the calendar years but also implements rigorous means of estimating these event ages and their uncertainties using Bayesian statistics.

As mentioned before, several countries have historical accounts longer than a thousand years or more and thus may cover more than a few seismic cycles of very active faults. However, the longer we go back in the past, the fewer and less reliable the historical accounts are. In addition, most of the historical earthquakes roughly control isoseismal areas from documents of earthquake damage, but not the specific source of each earthquake. Thus a combination of such historical documents and paleoseismic evidence with well-determined dates are used to confirm the source fault for such historical earthquakes (e.g., Armijo et al. 2005).

Archaeoseismology reaches further back in the past to reveal prehistoric earthquakes based on their effects on man-made objects, usually buildings or other monuments (Karcz and Kafri 1978, see entry on ► “Earthquake, Archaeoseismology”). The time range depends on history and archaeology in each country. Historic accounts are often available and tied with ruins due to devastating shaking, paleo-liquefactions, and paleo-landslides (rock-slides) at archaeological sites. The advantage of archaeoseismology is narrow time constraints of earthquake occurrence by archaeological and historical knowledge. Characteristic off-fault damage types include horizontal shifting of large building blocks, downward sliding of one or several blocks from masoned arches, collapse of heavy, stably-built walls, chipping of corners of building blocks, and aligned falling of walls and columns (Marco 2008). In some rare cases, faults intersect with archaeological sites, which enables us to measure sense and size of slip and constrain its occurrence time. Several recent archaeoseismological studies in the Eastern Mediterranean region in particular have been unveiling the detailed slip history and rupture behavior of the Dead Sea fault (e.g., Meghraoui et al. 2003), and large earthquakes associated with mythology along the Aegean Extensional province, Cyprus, and Hellenic Arcs (e.g., Stiros 2001).

Magnitude and Frequency of Paleoearthquakes

Like “uniformitarianism,” no paleoearthquakes can be inferred without detailed knowledge of recent observed earthquakes that left important clues between subsurface seismic processes and surface deformation. Since paleoseismology relies on paleo-surface information preserved in geology and landforms, seismic and geologic data compilations of recent large shocks associated with surface-ruptures, tilt, warping, and other earthquake-related deformation are necessary.

As seismologists found simple self-similar scaling laws for source faults, dimension and magnitude, surface ruptures compiled from worldwide data have similar scaling relations (Wells and Coppersmith 1994). For example, the empirical relation between surface rupture length (L in km) and magnitude (M) is expressed as $M = 5.16 + 1.12 \log L$ for strike-slip fault. For seismic hazard assessment, such an empirical relation is increasingly applied to estimate magnitude from the

length of an identified fault trace. Similar empirical scaling relations are found between coseismic surface displacement and magnitude. For maximum displacement $D_{\max}$ in meter, $M = 6.69 + 0.74 \log D_{\max}$. For average displacement, D_{ave} in meters, $M = 6.93 + 0.82 \log D_{\text{ave}}$. Besides fault length, coseismic slip, either $D_{\max}$ or D_{ave} , is also increasingly used to estimate the magnitude of a paleoearthquake. There are two ways to have such coseismic slip from paleoseismic data. One gives an indirect estimate from slip rate (S) and average recurrence interval (T_r) of paleoearthquakes along a fault as $D = S \times T_r$. The other provides direct measurement from vertical separation of faulted strata on trench walls in particular for reverse faults and three-dimensional reconstruction of lateral offset of piercing points for strike-slip faults.

In terms of magnitude and frequency estimates for an active fault, one of the greatest impacts of geology on seismology is the “characteristic earthquake model” (Schwartz and Coppersmith 1984, see entry on ► “Characteristic Earthquakes and Seismic Gaps”). This model illustrates the concept of the repetitions of similar size of large earthquakes with a similar amount of slip, rupture length, and magnitude along a fault. The model also emphasizes that paleoseismic records from a fault suggest more frequent occurrence of large earthquakes compared to the ones extrapolated from the Gutenberg–Richter relation of regional seismicity (Wesnousky 1994). If only scanty paleoseismic information on a fault is available, one simply uses the characteristic earthquake model to estimate the size of earthquakes potentially caused by the fault.

If the characteristic earthquakes on a fault occur periodically or quasi-periodically with characteristic slip, we would not have any difficulty to approximately forecast the time and magnitude for the next large earthquake (Fig. 3a). However, as summarized below, most of the recurrences of earthquakes are not periodic and may be influenced by recent rupture history. Regarding the slifetime relation at a site along a fault, under the assumption of constant long-term slip rate, two conceptual models are proposed (Shimazaki and Nakata 1980). The “time-predictable model” can predict the time of the next earthquake from a knowledge of the previous slip (Fig. 3b), whereas “slip predictable model” cannot predict the time but amount of slip, taking the slip deficit since the last one into account (Fig. 3c). In reality, however, at least 30 prehistoric earthquakes in a 6000-year-long record at Wrightwood along the San Andreas fault, southern California, follow neither time-predictable nor slip-predictable model (Weldon et al. 2004). The paleoseismic history might prove that slip occurs at a wide range of stress levels and does not return to the same postseismic stress level, keeping quasi-periodic recurrence behaviors (Scharer et al. 2010).

To furnish the magnitude estimate using the length of a fault, one must define the extent of a fault. In nature, we rarely encounter an isolated long single fault trace. Instead, we more often map a group of faults or a network of faults, which require

us to define each seismogenic fault. To subdivide faults into portions, the concept of “fault segmentation” is employed. “Fault segment” is a general term for the portions of a fault or fault system or fault zone. There are several types of fault segments ordered in terms of earthquake occurrence; *earthquake segment*, *behavioral segment*, *structural segment*, *geologic segment*, and *geometric segment* (McCalpin 2009). Identifying and defining earthquake segments that directly determine the size of earthquakes is the ultimate goal. However, numerous studies started to focus on changes in fault orientation and boundaries of fault traces (branches, intersections, separations, and step-overs) to identify each structural and geometric fault segment. There are also varieties of scale of segment and “discontinuous gaps.” But since we assess seismogenic fault, kilometer-scale traces are generally taken into account. To identify each seismogenic fault among many complex mapped fault traces, “5-km separation criteria” are now broadly employed for the hazard assessment (e.g., Earthquake Research Committee in Japan (ERC), 2005). Several studies done in the late 1980s and early 1990s compiled the surface ruptures, associated historic earthquakes, and mapped active traces around the globe and then concluded that rupture propagation beyond the 5-km step width is highly unlikely (Barka and Kadinsky-Cade 1988; Wesnousky 1988, 2006). Although these should be case-dependent with many factors, “5 km” is still a standard separation distance to identify each seismogenic fault segment.

A long continuous fault (system) cannot be simply divided by the 5-km separation criterion instead. It generates a variety of rupture patterns and sizes of earthquakes. Several conceptual models are proposed with slip rates and event age constraints at several sites along a single fault or a single fault system. The characteristic earthquake model is one of the several conceptual models. In reality, a space-time diagram of fault rupture patterns estimated from intensive paleoseismic surveys is not so simple to apply the conceptual model. Detailed examples from the southern San Andreas fault (Fig. 5, Weldon et al. 2004) and the North Anatolian fault, Turkey (Hartleb et al. 2003) show the complicated rupture patterns on a fault system.

Off-Fault Paleoseismology

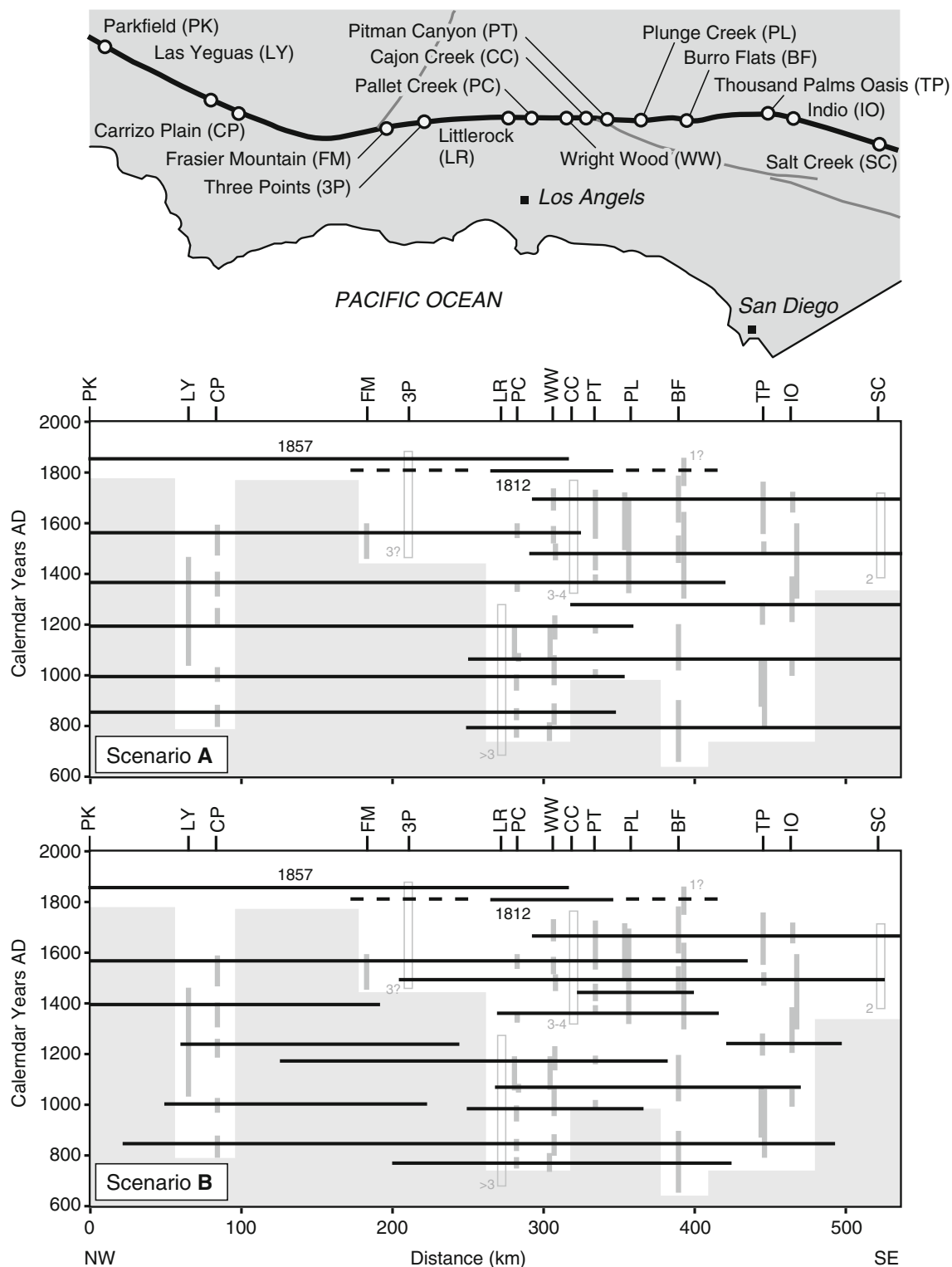
Although a limitation of off-fault evidence is its difficulty in identifying the seismic source, it becomes supportive evidence for paleoearthquakes, together with on-fault evidence. There are mainly two types of off-fault evidence for paleoearthquakes: one is due to strong shaking, and the other is related to features associated with broadly distributed coseismic and postseismic deformation (see ► “Earthquakes and Crustal Deformation”). The former triggers liquefaction, terrestrial and submarine landslides, and tsunamis. The latter leaves uplifted, tilted, and subsided ground surfaces recorded on shorelines and marine and river terraces, where paleo-altitude can be estimated and a flat surface is guaranteed.

Besides on-fault geology and geomorphology, preservation of such paleoseismic evidence is necessary. In addition, synchronicity and spatial extent of off-fault events are keys to distinguish paleoearthquakes from other causes of such secondary geologic features.

Liquefaction and associated features such as sand blows, sand boils, sand volcanoes, lateral spreading, clastic dikes, hydraulic fracturing, and rarely gravel blows and gravel dikes are the products of strong ground acceleration due to a large earthquake. Worldwide data on historical earthquakes show that the features having a liquefaction origin can be developed at earthquake magnitudes as low as about 5, but that a magnitude of about 5.5–6 is the lower limit at which liquefaction effects become relatively common (Ambraseys 1988; Obermeier 2009). Therefore, paleo-liquefaction is strong evidence for large earthquakes in the past. Although it is difficult to excavate trenches or pits targeted only for paleo-liquefaction, we often encounter evidence for paleo-liquefaction at archaeological sites, which can constrain the earthquake occurrence time much better than ^{14}C date due to cover-up by the non-damaged younger ruins and horizons (e.g., Fig. 6, Sangawa 2010). Since liquefaction can be commonly repeated at the same site where the recent large earthquake occurred, we can sometimes find the paleo-liquefaction on trench walls and confirm the recurrence of strong shaking at the site (e.g., Wesnousky and Leffler 1992).

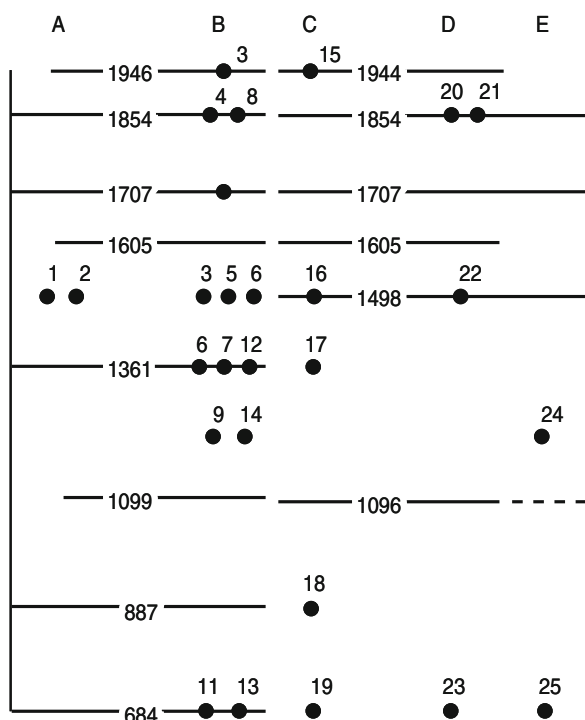
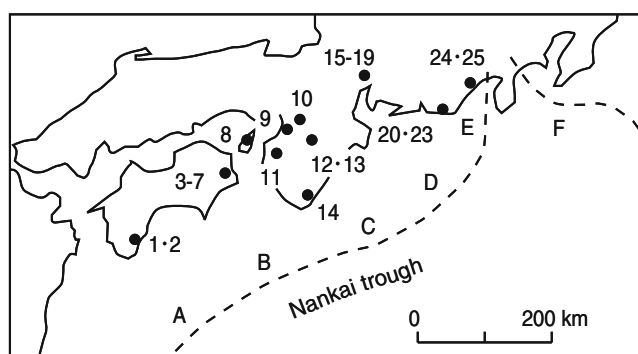
Strong ground motions in mountainous regions often trigger slope failures. In the regions characterized by steep unstable topography, we can identify massive landslides and scarps occurred in the past in aerial photographs. However, interpreting an earthquake-induced landslide or group of landslides excluding other factors like heavy rainfall is particularly hard, and levels of confidence in the resulting interpretation vary widely. Comprehensive geologic surveys with dating of landslides and several criteria (Jibson 2009) that discriminate landslides from other factors are required.

Widespread thin sediments in coastal regions reveal occurrence of paleo-tsunamis. Sand layer often containing pebble and shell deposit is extensively deposited by tsunami waves and its inundation on the coastal areas (see ► “Tsunami”). Although there is no single convenient feature to distinguish tsunami deposits from those generated by other events, sedimentary facies analysis with several key features is indispensable for identifying tsunami deposit (Shiki et al. 2008). To better identify paleo-tsunami deposits, there are limited conditions to preserve such sand layers; paleo-tsunami deposits are distinguished from black humus and mud deposits if the paleo-tsunami reaches the areas in swales, marsh and other calm environments where no sand sources can normally flow. Numerous studies, which employed shallow coring, pits, and trenches, exposed tsunami sand sheets intervened by finer and datable organic soils (e.g., Atwater 1987). The recent significant contribution of the tsunami deposit studies to the subduction process is to find prehistoric M9 class earthquakes



Paleoseismology, Fig. 5 Two possible rupture sequences on the southern San Andreas fault (Weldon et al. 2004). Vertical grey bars are ranges in age for earthquakes at the sites listed at the lower margin and

horizontal bars are rupture lengths. Open boxes represent multiple events age ranges; the individual event ages are unknown. Grey shading indicates regions and times with no data



Paleoseismology, Fig. 6 History of mega-thrust earthquakes along the Nankai trough, Japan (Sangawa 2010). The subduction zone is divided into five rupture segments A–E (F is the source of 1923 $M = 7.9$ Kanto earthquake along the Sagami trough). The horizontal lines with number in lower panel indicate the rupture extents and occurrence years of large earthquakes. Dots with number denote archaeological sites with exposed significant liquefactions at specific ruins and their calendar years. Recent archaeo-seismological data filled gaps in long recurrence intervals originally estimated by only historic accounts (e.g., possible rupture on A and B segments at the 1,498 earthquake), and making the Nankai subduction earthquake cycles more regular and frequent. Tsunami deposits recovered from coastal lakes, lagoons, swamps, and lowlands are also filling the gaps (e.g., Takada et al. 2002) and also provide much longer rupture history back to the entire Holocene period (Komatsubara and Fujiwara 2007)

and their recurrence intervals along the 2004 Sumatra-Andaman (Jankaew et al. 2008), and along the Kuril trench (Nanayama et al. 2003), and relation with the M8 class predecessors of the 1960 M9.5 Chile earthquake (Cisternas et al. 2005), which breaks through the conventional idea on regular

occurrences of subduction earthquakes (see ► “Great Earthquakes”).

Besides paleo-tsunami deposits, submarine turbidites are another geologic clue to reveal large submarine and coastal earthquakes. Goldfinger et al. (2007) demonstrated their validity and significance of sediments associated with earthquake-triggered turbidity currents by analyzing numerous piston cores collected from submarine channels and canyon systems draining the northern California continental margin. By studying synchronous triggering of the turbidity currents, they found 15 turbidites during the last ~2,800 years, equivalent to an average repeat time of ~200 years which is similar to the onshore value of ~230 years along the northern San Andreas fault. Along-strike correlation also suggests the ~320-km rupture extent of the most of the younger events.

Broad deformation of off-fault areas is another indicator for paleoearthquakes (see ► “Earthquakes and Crustal Deformation”). Vertical deformation can induce changes in local rates of deposition and erosion that provide evidence of a paleoearthquake, particularly in fluvial and coastal environment (McCalpin and Carver 2009). Subduction associated mega-thrust earthquakes and dip-slip intraplate earthquakes largely change horizontal markers such as shorelines prior to an earthquake and record off-fault coseismic uplift (e.g., Meghraoui et al. 2004) which are preserved as uplifted terraces and shorelines. Coastal areas facing subduction zones, such as southwest Japan, Taiwan, New Zealand, and Crete in Greece, have several steps of elevated terraces and shorelines associated with the mega-thrust earthquakes (see ► “Seismicity, Subduction Zone”). But significant coseismic uplifts are also inferred to have been formed by shallow near-coast offshore active faults that might be branched from subduction interface (e.g., Maemoku 2001). Although dating paleo-marine terraces and shorelines is not easy, coral reefs, coral microatolls (e.g., Zachariasen et al. 1999), fossils of sessile organisms inhabit intertidal level and become both paleo-shoreline markers and indicators of timing of earthquake occurrences. For older terraces over 50 ka, other dating techniques and regional tephra fallouts are used to estimate the emergent dates. In Japan, altitudes of uplifted marine terraces formed at marine oxygen isotope stage 5e (~120 ka) often provide us with the best estimates of long-term uplift rates associated with interplate events (Koide and Machida 2001). In contrast to coseismic uplift, studies for paleoseismic subsidence are much fewer. Subsidence associated with a subduction earthquake normally occurs inland from the rebounded uplift zone. Here, buried soil of subsided marsh and forest records the sudden vertical deformation (e.g., Atwater et al. 1995).

Permanent deformation of landscapes associated with seismogenic faulting is sometimes modified or amplified by postseismic crustal movement. Recent mega-thrust subduction events caused not only coseismic uplift, but also postseismic transient deformation that lasts more than a few decades (see

► “Earthquakes and Crustal Deformation”). Although contribution of postseismic deformation associated with paleoearthquakes is hardly measured in the uplifted terraces, Sawai et al. (2004) used fossil diatom assemblages to infer the gradual changes in sedimentary environments corresponding to the transient postseismic deep creep that occurred in the seventeenth century along the Kuril subduction zone.

Seismic Hazard Assessment Based on Paleoseismology

Paleoseismic data are directly used for making probabilistic seismic hazard maps (see ► “Seismic Hazard”) which incorporate information on fault location, dimension, geometry, largest earthquake from such dimension, slip rates, and recurrence intervals (e.g., Working Group on California Earthquake Probabilities 1995). Two types of statistical models of earthquake occurrence are commonly used to estimate earthquake probability: a stationary Poisson model and a conditional quasi-periodic model (see ► “Statistical Seismology”). The former is used for active faults or active seismic regions where only frequency of large earthquakes is available. The latter considers fluctuations of repeat times (coefficient of variation in recurrence time) and elapsed time since the most recent event (see Fig. 3), which is time-dependent and compatible with the strain accumulation process. Several probability density functions such as normal, lognormal, Weibull, or gamma distributions are used for the recurrence process. Among them, the Brownian Passage Time function (Matthews et al. 2002) is evaluated as so far the best function because of its capability to account for the stress perturbations and has been adopted for the probabilistic seismic hazard studies in California and Japan (Field et al. 2009; ERC 2005). Time-dependent seismic hazard considering conditional probabilities on most major active faults provides us with more realistic estimates (Fujiwara et al. 2009).

If further detailed paleoseismic information is available for a specific fault, a “scenario earthquake model (seismic hazard maps for specified seismic sources)” is often proposed to show the strong shaking for the areas being assessed when the specific earthquake occurs (ERC 2005). When providing paleoseismic data to seismic engineers, possible rupture patterns for the next earthquake, slip distribution, and plausible locations of asperities becomes additional information (Somerville et al. 1999).

In addition to the shaking damage, surface rupture and displacement associated with a large earthquake often destroy man-made structures. Structures for human occupancy (e.g., hospital and school) and critical buildings (e.g., nuclear power plant) can be so located to avoid the traces of active faults (e.g., the Alquist-Priolo Earthquake Fault Zoning Act in California, see entry on ► “Seismic Zonation”). To mitigate direct damage due to surface-faulting, detailed positions of active faults and their paleoseismological properties are progressively been opened to the public through the internet.

Issues in Paleoseismology and Perspectives

There are fundamental limitations of the geologic approach to infer paleoearthquakes in time and space. The following are major caveats in using paleoseismological information:

1. Surface ruptures are commonly associated with $M > 6$ earthquakes (Wells and Coppersmith 1994) that occur at shallow depths (mostly < 20 km), even though $M \sim 5$ earthquakes may also leave surface ruptures in the case of extremely shallow hypocenter (< 5 km) and volcanic environments (Payne et al. 2009, see entry on ► “Earthquakes, Volcanogenic”). However, we must know that geomorphic and stratigraphic evidence for all $M > 6$ paleoearthquakes is not always produced. It totally depends on regional tectonic setting, depth of hypocenter (thickness of seismogenic crust), fault slip sense, and surface geologic conditions. For example, revisiting the inland earthquakes in Japan since 1923, no more than half of the $M \sim 7$ earthquakes left the surface breaks comparable to the seismic faults, which obviously underestimates the probabilities of even $M \sim 7$ shocks (Toda and Awata 2008). Even along major strike-slip faults, we experienced enigmatic large earthquakes that did not leave clear surface rupture (e.g., 1989 Loma Prieta earthquake and 2010 Haiti earthquake).
2. Rates of erosion and sedimentation at regional and local scales control preservations of fault landforms and evidence for paleoearthquakes at excavation sites. Human destruction of the surface geology and topography due to urbanization or farming often destroys evidence of very recent paleoseismic events. Furthermore, obtaining better time constraints for paleoearthquakes is rare due to the lack of abundant dating material.
3. Geologic and topographic evidence normally includes interpretation with epistemic uncertainties. Complete agreement is rare in aerial photographic interpretation. “Trench parties” (Grant 2007) often create indecisive debate that is not visible in the final publication in which simple trench logs are only illustrated. Photos and detailed sketches are objective records, but they are infrequently opened to the public and are hardly accessible. Therefore one should be aware that publications often underestimate or overestimate numbers of paleoearthquakes and need extra effort to clarify quality measurements and other paleoseismic interpretations.
4. Paleoseismic study cannot precisely reveal rupture extent, thus magnitude of an earthquake without any help of historic accounts. Space-time diagrams of paleoseismic events on a fault or fault system may only conclude the cases adjacent segments did not simultaneously rupture through. Displacement data of a paleoearthquake would be useful information to estimate magnitude based on the empirical scaling law (Wells and Coppersmith 1994). But at least five to ten measurements are stochastically required

to characterize magnitude of a paleoearthquake (Hemphill-Haley and Weldon II 1999).

5. Although we simply apply the characteristic earthquake model or extended segmentation model under the elastic rebound framework (see ► “Earthquakes and Crustal Deformation”), there may be intrinsic uncertainties of repeated faulting processes, for example, nonlinear strain accumulation and complex rupture process. This might be a subject seismologists have to reveal. But paleoseismology instead can be suggestive with data that also conquer “uniformitarianism.” For example, the 2004 $M = 9.2$ Sumatra-Andaman earthquake was an extraordinary earthquake which can be hardly expected from historic accounts. However, intensive detailed studies about coral microatolls (Zachariasen et al. 1999) and tsunami deposits (Jankaew et al. 2008) suggested there were frequent huge events in the past, even if we do not see them in historic ages. This is an important role of paleoseismology.

Summary

Paleoseismology, which is the study of prehistoric or pre-instrumental earthquakes, is an interdisciplinary field of research mainly encompassing seismology, geology and geomorphology. Starting from identifying active faults, various field and laboratory techniques such as aerial photography interpretation, trench excavation, and radiocarbon dating are used to reveal the spatial extent and occurrence times of paleoearthquakes, and associated long-term movements of active faults. Evidence for paleoearthquakes is not only preserved in strata across a fault but also found far from the source as paleo-liquefaction, paleo-landslide, and paleo-tsunami deposits. Empirical relations between rupture length and slip versus earthquake magnitude from recent instrumentally recorded shocks enable us to estimate the size of future earthquakes on a fault or fault zone. Detailed rupture history of a fault also allows us to estimate time-dependent probabilistic seismic hazard. Although there are fundamental limitations and incompleteness in field data, contribution of paleoseismology in understanding the long-term faulting process is enormous and indispensable. Recent progress in paleoseismic data accumulations also allows seismologists and seismic engineers to make better seismic hazard maps to mitigate seismic damage in the future.

Cross-References

- Characteristic Earthquakes and Seismic Gaps
- Earthquake, Archaeoseismology

- Earthquake, Focal Mechanism
- Earthquakes and Crustal Deformation
- Earthquakes, Volcanogenic
- Great Earthquakes
- Seismic Hazard
- Seismic Instrumentation
- Seismic Zonation
- Seismicity, Subduction Zone
- Seismological Networks
- Statistical Seismology
- Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
- Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity
- Tsunami

Bibliography

- Allen CR (1986) Seismological and paleoseismological techniques of research in active tectonics. In: Wallace RE (ed) *Active tectonics: studies in geophysics*. National Academy Press, Washington, DC, pp 148–154
- Ambraseys N (1988) Engineering seismology: earthquake engineering and structural dynamics. *J Int Assoc Earthq Eng* 17:1–105
- Armijo R et al (2005) Submarine fault scarps in the sea of Marmara pull-apart (north Anatolian fault): implications for seismic hazard in Istanbul. *Geochem Geophys Geosyst* 6:Q06009. <https://doi.org/10.1029/2004GC000896>.
- Arrowsmith JR, Zielke O (2009) Tectonic geomorphology of the San Andreas fault zone from high resolution topography: an example from the Cholame segment. *Geomorphology* 113:70–81
- Atwater BF (1987) Evidence of great Holocene earthquakes along the outer coast of Washington state. *Science* 236:942–944
- Atwater BF, Nelson AR, Clague JJ, Carver GA, Bobrowsky PT, Bourgeois J, Darienzo ME, Grant WC, Hemphill-Haley E, Kelsey HM, Jacoby GC, Nishenko SP, Palmer SP, Peterson CD, Reinhart MA, Yamaguchi DK (1995) Summary of coastal geologic evidence for past great earthquakes at the Cascadia subduction zone. *Earthquake Spectra* 11:1–18
- Barka A, Kadinsky-Cade K (1988) Strike-slip fault geometry in Turkey and its influence on earthquake activity. *Tectonics* 7:663–684
- Berryman KR, Ota Y, Hull AG (1989) Holocene paleoseismology in the fold and thrust belt of the Hikurangi subduction zone, eastern North Island, New Zealand. *Tectonophysics* 163:185–195
- Bond CE, Gibbs AD, Shipton ZK, Jones S (2007) What do you think this is? “Conceptual uncertainty” in geoscience interpretation. *GSA Today* 17:4–10
- Bonilla MG, Lienkaemper JJ (1990) Visibility of fault strands in exploratory trenches and timing of rupture events. *Geology* 18:153–156
- Bronk Ramsey C (2001) Development of the radiocarbon calibration program OxCal. *Radiocarbon* 43:355–363
- Cisternas M, Atwater BF, Torrejon F, Sawai Y, Machuca G, Lagos M, Eipert A, Youlton C, Salgado I, Kamataki T, Shishikura M, Rajendran CP, Malik JK, Rizal Y, Husni M (2005) Predecessors of the giant 1960 Chile earthquake. *Nature* 437:404–407
- Dolan JF, Christofferson SA, Shaw JH (2003) Recognition of paleoearthquakes on the Puente Hills blind thrust fault, California. *Science* 300:115–118
- Earthquake Research Committee (2005) Report: ‘National Seismic Hazard Maps for Japan (2005)’, headquarters for earthquake research

- promotion, 162 pp. Available from world wide web: <http://www.jishin.go.jp/main/index-e.html>
- Field EH, Dawson TE, Felzer KR, Frankel AD, Gupta V, Jordan TH, Parsons T, Petersen MD, Stein RS, Weldon RJ II, Wills CJ (2009) Uniform California earthquake rupture forecast, version 2 (UCERF 2). *Bull Seismol Soc Am* 99:2053–2107
- Fujiwara H, Morikawa N, Ishikawa Y, Okumura T, Miyakoshi J, Nojima N, Fukushima Y (2009) Statistical comparison of national probabilistic seismic hazard maps and frequency of recorded JMA seismic intensities from the K-net strong-motion observation network in Japan during 1997–2006. *Seismol Res Lett* 80:458–464
- Fukuchi T (2001) Assessment of fault activity by ESR dating of fault gouge; an example of the 500 m core samples drilled into the Nojima earthquake fault in Japan. *Quat Sci Rev* 20:1005–1008
- Goldfinger C, Morey AE, Nelson CH, Gutierrez-Pastor J, Johnson JE, Karabanov E, Chaytor J, Eriksson A, Shipboard Scientific Party (2007) Rupture lengths and temporal history of significant earthquakes on the offshore and north coast segments of the northern San Andreas fault based on turbidite stratigraphy. *Earth Planet Sci Lett* 254:9–27
- Grant LB (2002) Paleoseismology. In: Lee WH, Kanamori H, Jennings PC, Kisslinger C (eds) *IASPEI International handbook of earthquake and engineering seismology*, vol 81A. International Association of Seismology and Physics of the Earth's Interior Press, New York, pp 475–489
- Grant LB (2007) Historical seismicity – paleoseismology. In: Schubert G (ed) *Treatise of geophysics*, vol 4. Elsevier, Amsterdam, pp 567–589
- Grant LB, Waggoner JT, von Stein C, Rockwell TK (1997) Paleoseismicity of the north branch of the Newport–Inglewood fault zone in Huntington Beach, California, from Cone penetrometer test data. *Bull Seismol Soc Am* 87:277–293
- Hartleb RD, Dolan JF, Akyuz HS, Yerli B (2003) A 2000-year-long paleoseismologic record of earthquakes along the central north Anatolian fault, from trenches at Alayurt, Turkey. *Bull Seismol Soc Am* 93:1935–1954
- Hemphill-Haley MA, Weldon RJ II (1999) Estimating prehistoric earthquake magnitude from point measurements of surface rupture. *Bull Seismol Soc Am* 89:1264–1279
- Jacoby GC (1989) Application of tree-ring analysis to paleoseismology. *Geology* 17:226
- Jankaew K, Atwater BF, Sawai Y, Choowong M, Charoentitrat T, Martin ME, Prendergast A (2008) Medieval forewarning of the 2004 Indian Ocean tsunami in Thailand. *Nature* 455:1228–1231
- Jibson RW (2009) Using landslides for paleoseismic analysis. In: McCalpin JP (ed) *Paleoseismology*. Academic Press, Burlington, pp 565–601
- Karcz I, Kafri U (1978) Evaluation of supposed archeoseismic damage in Israel. *J Archaeol Sci* 5:237–253
- Koide K, Machida H (eds) (2001) *Atlas of quaternary marine terraces in the Japanese Islands*. University of Tokyo Press, Tokyo
- Komatsubara and Fujiwara (2007) Overview of Holocene tsunami deposits along the Nankai, Suruga, and Sagami troughs, Southwest Japan. *Pure Appl Geophys* 164:493–507
- Kondo H, Toda S, Okumura K, Takada K, Chiba T (2008) A fault scarp in an urban area identified by LiDAR survey: a case study on the Itoigawa-Shizuoka tectonic line, Central Japan. *Geomorphology* 101:731–739
- Lienkaemper JJ, Bronk Ramsey C (2009) OxCal: versatile tool for developing paleoearthquake chronologies – a primer. *Seismol Res Lett* 80:431–434
- Lin A, Guo J, Kano K, Awata Y (2006) Average slip rate and recurrence interval of large-magnitude earthquakes on the western segment of the strike-slip Kunlun fault, northern Tibet. *Bull Seismol Soc Am* 96:1597–1611
- Maemoku H (2001) Reexamination of coseismic uplift of Cape Muroto, Southwest Japan, using AMS ^{14}C ages of raised sessile organisms. *J Geogr* 110:479–490. (in Japanese)
- Marco S (2008) Recognition of earthquake-related damage in archaeological sites: examples from the Dead Sea fault zone. *Tectonophysics* 453:148–156
- Matthews MV, Ellsworth WL, Reasenber PA (2002) A Brownian model for recurrent earthquakes. *Bull Seismol Soc Am* 92:2233–2250
- McCalpin JP (ed) (2009) *Paleoseismology*, International Geophysics Series, vol 95, 2nd edn. Academic Press, Burlington. 613 pp
- McCalpin JP, Carver GA (2009) Paleoseismology of compressional tectonic environments. In: McCalpin JP (ed) *Paleoseismology*. Academic Press, Burlington, pp 315–419
- Meghraoui M, Gomez F, Sbeinati R, Van der Woerd J, Mouty MN, Darkal A, Radwan Y, Layyous I, Al-Najjar H, Darawchah R, Hijazi F, Al-Ghazzi R, Barazangi M (2003) Evidence for 830 years of seismic quiescence from paleoseismology, archaeoseismology and historical seismicity along the Dead Sea fault in Syria. *Earth Planet Sci Lett* 210:35–52
- Meghraoui M, Maoche S, Chemaa B, Cakir Z, Aoudia A, Harbi A, Alasset PJ, Ayadi A, Bouhadad Y, Benhamouda F (2004) Coastal uplift and thrust faulting associated with the Mw = 6.8 Zemmouri (Algeria) earthquake of 21 May, 2003. *Geophys Res Lett* 31:L19605. <https://doi.org/10.1029/2004GL020466>
- Nakata T, Shimazaki K (1997) Geo-slicer, a newly invented soil sample, for high-resolution active fault studies. *J Geogr* 106:59–69. (in Japanese)
- Nanayama F, Satake K, Furukawa R, Shimokawa K, Atwater BF, Shigeno K, Yamaki S (2003) Unusually large earthquakes inferred from tsunami deposits along the Kuril trench. *Nature* 424:660–663
- Obermeier SF (2009) Using liquefaction-induced and other softsediment features for paleoseismic analysis. In: McCalpin JP (ed) *Paleoseismology*. Academic Press, New York, pp 497–564
- Okada A (1991) Studies of active faults by trenching method. In: Yonekura N, Okada A, Moriyama A (eds) *Tectonic landforms*. Kokon-Shoin, Tokyo, pp 18–44. (in Japanese)
- Pantosti D, Schwartz DP, Valensise G (1993) Paleoseismology along the 1980 surface rupture of the Irpinia fault; implications for earthquake recurrence in the southern Apennines, Italy. *J Geophys Res* 98:6561–6577
- Payne SJ, Hackett WR, Smith RP (2009) Paleoseismology of volcanic environments. In: McCalpin JP (ed) *Paleoseismology*. pp 271–314
- Research Group for Active Faults of Japan (1992) *Maps of active faults in Japan with an explanatory text*. University of Tokyo Press, Tokyo. 74 pp
- Sangawa A (2010) A study of paleoearthquakes at archeological sites – a new interdisciplinary area between paleoseismology and archeology. *Synthology* 2:84–94. National Institute of Advanced Industrial Science and Technology (AIST)
- Sawai Y, Satake K, Kamataki T, Nasu H, Shishikura M, Atwater BF, Horton BP, Kelsey HM, Nagumo T, Yamaguchi M (2004) Transient uplift after a 17th-century earthquake along the Kuril subduction zone. *Science* 306:1918–1920
- Scharer KM, Weldon RJ II, Fumal TE, Biasi GP (2007) Paleoeearthquakes on the southern San Andreas fault, Wrightwood, California, 3000 to 1500 B.C.: a new method for evaluating paleoseismic evidence and earthquake horizons. *Bull Seismol Soc Am* 97:1054–1093
- Scharer K, Biasi GP, Weldon RJ II, Fumal TE (2010) Quasi-periodic recurrence of large earthquakes on the southern San Andreas fault. *Geology* 38:555–558
- Schwartz DP, Coppersmith KJ (1984) Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas fault zones. *J Geophys Res* 89:5681–5698
- Sherrod BL, Brocher TM, Weaver CS, Bucknam RC, Blakely RJ, Kelsey HM, Nelson AR, Haugerud RA (2004) Holocene fault scarps near Tacoma, Washington, USA. *Geology* 32:9–12

- Shiki T, Tachibana T, Fujiwara O, Goto K, Nanayama F, Yamazaki T (2008) Characteristic features of tsunamites. In: Shiki T (ed) *Tsunamites – features and implications*. Elsevier, Amsterdam. <https://doi.org/10.1016/B978-0-444-51552-0.00018-7>
- Shimazaki K, Nakata T (1980) Time-predictable recurrence model for large earthquakes. *Geophys Res Lett* 7:279–282
- Sieh KE (1978) Pre-historic large earthquakes produced by slip on the San Andreas fault at Palmett Creek, California. *J Geophys Res* 83:3970–3939
- Somerville PG, Irikura K, Graves R, Sawada S, Wald D, Abrahamson N, Iwasaki Y, Kagawa T, Smith N, Kowada A (1999) Characterizing crustal earthquake slip models for the prediction of strong ground motion. *Seismol Res Lett* 70:59–80
- Stiros SC (2001) The AD 365 Crete earthquake and possible seismic clustering during the fourth to sixth centuries AD in the eastern Mediterranean: a review of historical and archaeological data. *J Struct Geol* 23:545–562
- Stuiver M, Reimer PJ, Bard E, Beck JW, Burr GS, Hughen KA, Kromer B, McCormac FG, van der Plicht J, Spurk J (1998) INTCAL98 radiocarbon age calibration, 24000–0 cal BP. *Radiocarbon* 40:1041–1083
- Sugiyama Y, Mizuno K, Nanayama F, Sugai T, Yokota H, Hosoya T, Miura K, Takemura K, Kitada N (2003) Study of blind thrust faults underlying Tokyo and Osaka urban areas using a combination of high-resolution seismic reflection profiling and continuous coring. *Ann Geophys* 46:1071–1085
- Takada K, Atwater BF (2004) Evidence for liquefaction identified in peeled slices of Holocene deposits along the lower Columbia River. Washington. *Bull Seismol Soc Am* 94:550–575
- Takada K, Satake K, Sangawa A, Shimokawa K, Kumagai H, Goto K, Haraguchi T (2002) Survey of tsunami deposits at an archaeological site along the eastern Nankai trough. Annual report on active fault and paleoearthquake researches, Geological Survey of Japan, AIST, vol 2. pp 235–243 (in Japanese)
- Toda S, Awata Y (2008) Does the 2007 Noto Hanto earthquake reveal a weakness in the Japanese national seismic hazard map that could be remedied with geologic data? *Earth Planets Space* 60:1047–1052
- Tsutsumi H, Yeats RS (1999) Tectonic setting of the 1971 Sylmar and 1994 Northridge earthquakes in the San Fernando Valley, California. *Bull Seismol Soc Am* 89:1232–1249
- Weldon R, Fumal T, Biasi G (2004) Wrightwood and the earthquake cycle: what a long recurrence record tells us about how faults work. *GSA Today* 14(9):4–10
- Wells DL, Coppersmith KJ (1994) New empirical relationships among magnitude, rupture length, rupture area, and surface displacement. *Bull Seismol Soc Am* 84:974–1002
- Wesnousky SG (1988) Seismological and structural evolution of strike-slip faults. *Nature* 335:340–343
- Wesnousky SG (1994) The Gutenberg-Richter or characteristic earthquake distribution, which is it? *Bull Seismol Soc Am* 84:1940–1959
- Wesnousky SG (2006) Predicting the endpoints of earthquake ruptures. *Nature* 444:358–360
- Wesnousky SG, Leffler LM (1992) The repeat time of the 1811 and 12 New Madrid earthquakes; a geological perspective. *Bull Seismol Soc Am* 82:1756–1784
- Working Group on California Earthquake Probabilities (WGCEP) (1995) Seismic hazards in southern California: probable earthquakes, 1994–2024. *Bull Seismol Soc Am* 85:379–439
- Yeats RS, Prentice CS (1996) Introduction to special section: paleoseismology. *J Geophys Res* 101:5847–5853
- Zachariasen J, Sieh K, Taylor FW, Edwards RL, Hantoro WS (1999) Submergence and uplift associated with the giant 1833 Sumatran subduction earthquake: evidence from coral microatolls. *J Geophys Res* 104:895–919

Plate Motions in Time: Inferences on Driving and Resisting Forces

Giampiero Iaffaldano¹ and Hans-Peter Bunge²

¹Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark

²Geophysics Section, Department of Earth and Environmental Sciences, Ludwig Maximilians University of Munich, Munich, Germany

Definition

Plate motions: horizontal piecewise movements of Earth's outer shell that occur over million years time scales and that can be measured or inferred from a range of geophysical techniques.

Introduction

Plate tectonics is a remarkable theory. Its central tenet, which is widely known, states that the surface of the Earth moves in a piecewise coherent fashion via a number of rigid plates (Wilson 1965; Morgan 1968), with most of the deformation focused along plate boundaries (Stein 1987). The underlying dynamic mechanism responsible for plate motions has been long identified in lateral buoyancy variations within the Earth mantle (Forte 2010). However, 40 years after the acceptance of plate tectonics, the details of how these forces may vary in time to modify plate movements remain limited, and we still fail to answer basic questions of fundamental importance: *Why do plates change their motions as shown by the geologic and geodetic records? – What are the spatial and temporal patterns of forces responsible for such changes?* The main difficulty stems from our poor knowledge of the *force balance* in plate tectonics. It is widely agreed that plate motions are driven by convection in the Earth's mantle (Hager and O'Connell 1981; Davies and Richards 1992). But the influence of other driving and resisting mechanisms, especially along plate margins, remains unclear. Forces concentrated along plate margins are known as plate boundary forces (Forsyth and Uyeda 1975) and it is likely that they play an important role in modulating plate motions (Humphreys and Coblenz 2007).

Portions of this entry, including all figures, were previously published in Iaffaldano, G., Bunge, H.-P., 2009 – Relating rapid plate-motion variations to plate-boundary forces in global coupled models of the mantle/lithosphere system: Effects of topography and friction, *Tectonophysics* v. 474, pp. 393–404 – and are reprinted here with permission of Elsevier.

A key component of plate boundary forces is the gravitational collapse associated with large mountain belts extending along plate margins. Climate-induced variations in topography, related to uplift and erosion in orogens, have been long recognized as possible controls in tectonics (Koeppen and Wegener 1924), and may therefore be capable of initiating variations in plate motion on regional and global scales. Conversely, geodynamicists can use inferences on topography and plate motion variations as powerful probes into the underlying force balance of plate tectonics.

Temporal variations in plate motions provide significant constraints on the budget of forces acting upon plates. This is because, by virtue of Newton's first law of motion, any change in plate motion is necessarily driven by variations in one or more driving or resisting forces. With the advent of space geodetic techniques (Dixon 1991), increasingly accurate estimates of global plate motions and their temporal variations are now available (Sella et al. 2002). While geodesy measures the current rates of plate motion, one can also glean long-term constraints on plate motion from paleomagnetic observations derived from the magnetic isochron record of the ocean floor. In fact, recent paleomagnetic plate motion models approach temporal resolutions on the order of 1–2 Myrs and reveal a number of rapid *velocity-changes* (few cm/year) occurring over time periods of a few Myrs or less (Müller et al. 2008). Some of these are particularly evident in the Southern Atlantic (Fig. 1a) and in the Indian Ocean (Fig. 1b). Their short duration makes it unlikely to attribute the changes to variations in the internal distribution of mantle buoyancies. In fact, these evolve globally on longer time

scales on the order of 50–100 Myrs (Bunge et al. 1998), although locally changes may occur over some 10 Myrs (Forte et al. 2009). It is reasonable to link these rapid changes to variations in plate boundary forces.

Plate Tectonics and the Rheology of the Lithosphere

The regime of plate tectonics requires high strain rates and low resistive stresses to coexist simultaneously along plate margins. But it is challenging to model this regime on a computer, because it is difficult to simulate shear failure along plate boundaries. To overcome this challenge, some models of the lithosphere incorporate highly non-Newtonian, viscous creep, strain-rate weakening rheologies together with viscoplastic yielding. Moresi and Solomatov (1998) explored the effects of temperature-dependent viscosity in combination with a plastic yield stress: The former causes the cold upper boundary layer (lithosphere) to be strong, while the latter allows the boundary layer to fail locally in regions of high stress. The success, measured through a so-called *plateness*, is evident when extreme strain-softening rheologies, known as pseudo-stick-slip (Bercovici 1995), are employed (Bercovici 2003). Unfortunately, the rheological parameters required for pseudo-stick-slip agree poorly with laboratory experiments of ductile deformation performed on olivine (Kirby 1983), particularly at the pressure and temperature conditions typical of the upper mantle (Karato and Jung 2003). Figure 2 shows normalized strength envelopes of a simplified two-phase

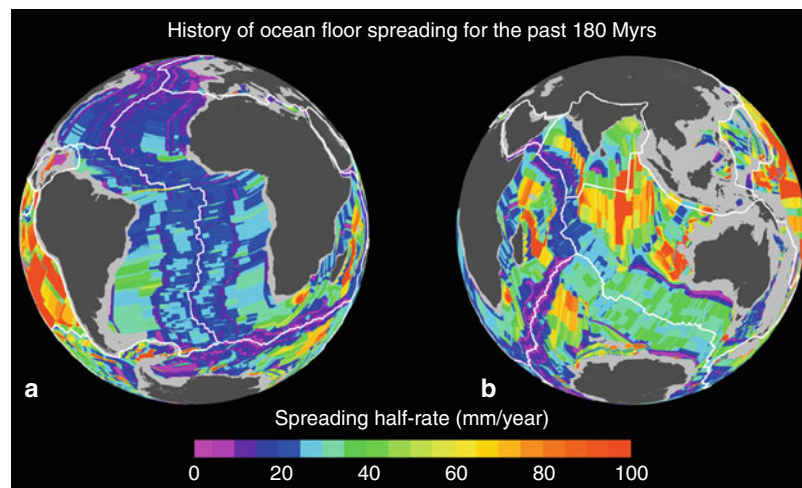


Plate Motions in Time: Inferences on Driving and Resisting Forces, Fig. 1 Observed oceanic spreading half-rates for the past 180 Myrs after a recent global compilation by Müller et al. (2008). Plate boundaries are in white, continents in dark gray. Abrupt changes in spreading rates reveal short-term variations in global plate motions, particularly visible in the South Atlantic (a) as well as in the Indian Ocean (b). Such rapid variations are unlikely to originate from global

changes in mantle driving forces, which occur on a longer time scale on the order of 50–100 Myrs as indicated by mantle circulation models. Instead, they are related to short-term variations in plate boundary forces caused, for example, by rapid growth of surface topography at convergent margins (see text). These observations point to the first-order importance of plate boundary forces in controlling global plate motions. (Courtesy of Elsevier © 2009)

lithosphere. Depth-dependent strength is parameterized via empirical laws established through laboratory experiments performed on quartzite, abundant in the upper 20 km of continental lithosphere, and olivine, which dominates at greater depths. The laboratory results indicate that strength increases linearly with overburden pressure in the upper, brittle part of the lithosphere; it then decreases exponentially with increasing temperature in the lower, ductile part (Kohlstedt et al. 1995). High strength in the upper lithosphere reflects the resistance of rocks to failure at low temperature, or to sliding past each other when already faulted. Experimental results indicate a simple linear relationship to parameterize this behavior (Byerlee 1978), with shear stress proportional to the normal pressure through a friction coefficient typically on the order of 0.6 (dashed envelope in Fig. 2). There is, however, mounting evidence for significantly lower values (in the range 0.01–0.15, see solid envelopes in Fig. 2) along faults and plate boundaries (Hickman 1991; Bird 1998; Suppe 2007).

Computer Models of the Faulted Lithosphere Coupled with Global Mantle Circulation Models

Lithospheric faults exhibit low friction, as we have noted. In other words, they are mechanically weak and experience low

stresses independently of the strain rate. It is therefore attractive to represent faults directly within the computational grid of numerical models of the lithosphere. This can be done, for instance, through the use of so-called contact elements, in an approach known as neo-tectonic modeling. Neo-tectonic models have reached considerable sophistication. They solve the equations of mass and momentum conservation, and compute the instantaneous force balance and associated plate velocities. They implement empirical, depth-dependent rheologies of the lithosphere, and account for ductile deformation in the lower crust as well as brittle deformation along faults. In some cases, they take advantage of the so-called thin-sheet approximation to reduce the computational complexity from 3-D to 2-D (Kong and Bird 1995) and to achieve greater computational efficiency.

Independent of advances in neo-tectonic models, there has been much progress in modeling the global circulation of the Earth's mantle (Tackley et al. 1994; Bunge et al. 1997; Zhong et al. 2000). Mantle circulation models (MCMs) account for the dynamic effects from a weak asthenosphere on the horizontal length-scales of the flow (Bunge et al. 1996), include internal heat generation from radioactivity, as well as a significant amount of heat from the core (Bunge 2005; Quere and Forte 2006; van der Hilst et al. 2007). Combined with constraints on the history of subduction, these models place first-

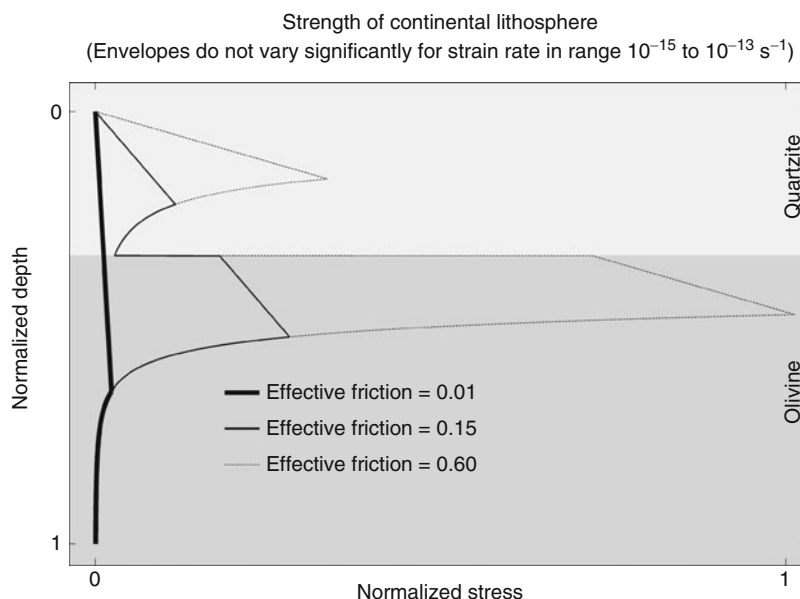


Plate Motions in Time: Inferences on Driving and Resisting Forces, Fig. 2 Normalized strength/depth profiles for a simplified two-layer continental lithosphere, plotted for three different fault friction coefficients and a strain rate in the range of 10^{-15} to 10^{-13} s $^{-1}$, typical of plate tectonics. Laboratory experiments performed mainly on quartzite and olivine (see text) indicate that lithosphere strength increases linearly with overburden pressure in the upper brittle part and decreases exponentially with increasing temperature in the lower ductile part. Note that the brittle part contributes a significant portion (70–80%)

of the total lithospheric strength, independently of the particular choice of friction coefficient and strain rate values. Experimental results suggest a friction coefficient around 0.6 (*dashed envelope*) for rocks under lithostatic pressure conditions, known as Byerlee's law. However, various independent evidences for weak faults suggest much lower values between 0.01 and 0.15 (*solid envelopes*), so that plate boundaries experience considerably lower stresses even for high strain rates. (Courtesy of Elsevier © 2009)

order estimates on the internal mantle buoyancies driving plate motions. Figure 3 shows the temperature distribution in the mantle from one recent model with more than 100 million grid points (Oeser et al. 2006), equivalent to a grid point spacing of 20 km and less throughout the model mantle. MCMs provide first-order estimates of mantle buoyancy forces but do not account for the complex processes in the lithosphere such the brittle failure. Similarly, neo-tectonic models include stresses originating within the lithosphere and realistic plate boundary forces. But they rely on assumptions of the mantle buoyancy field to complete the force balance. The logical step is merging the two model classes to simulate the coupled global mantle convection/plate tectonics system. This makes it possible to account simultaneously for plate boundary forces and mantle-related components of lithospheric force balance, and to predict variations in global plate velocities that one may test explicitly against the geologic record. Several authors have recently

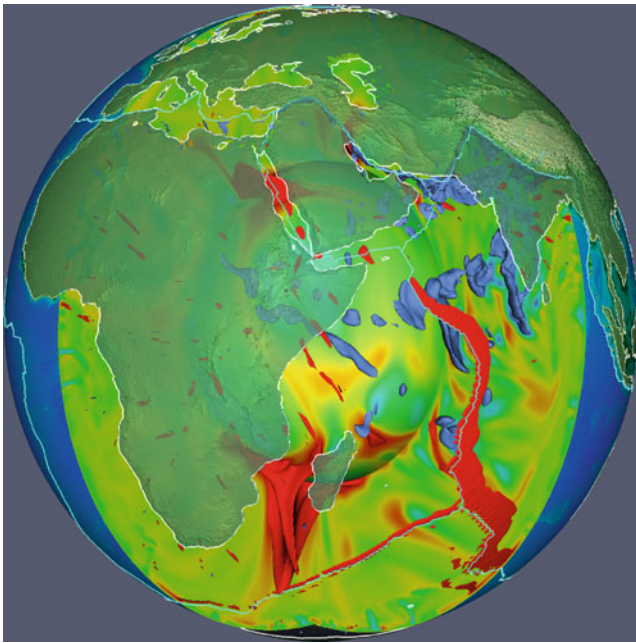


Plate Motions in Time: Inferences on Driving and Resisting Forces, Fig. 3 Temperature distribution in the Earth's mantle from a recent, high-resolution 3-D global circulation model. Blue represents cold, denser material whereas red is hot and buoyant mantle. View is on Africa, coastline is in white, with continental topography in transparent green color scale. Present-day plate boundaries are outlined in blue. More than 100 million grid points discretize the Earth's mantle, equivalent to an average grid spacing of 20 km or less. Circulation models include radial variations in mantle viscosity (factor 40 increase from the upper to the lower mantle), internal heat generation from radioactivity, bottom heating from the core, and a history of subduction spanning the past 120 Myrs. A cold downwelling is visible beneath Tibet where the ancient Tethys Ocean subducted under Eurasia; a hot and buoyant upwelling is visible as well beneath the spreading triple-junction of the Antarctica, Africa, and Australia plates. The circulation model provides a realistic, first-order estimate of internal buoyancy forces driving global plate motions. (Courtesy of Elsevier © 2009)

undertaken such approach (Iaffaldano et al. 2006; Ghosh et al. 2008). Results are encouraging, and in the following, we review recent predictions of climate-related topography variations and their influence on plate velocities in the Southern Pacific, Southern Atlantic, as well as in the Indian Ocean that compare well with observations.

Recent Plate Motion Changes

Kinematics of the Southern Pacific and Atlantic, and Its Relations to Climate Variations and Topographic Growth in the Andes

Paleomagnetic (Gordon and Jurdy 1986; DeMets et al. 1994) and geodetic (Norabuena et al. 1999) data indicate a significant reduction (as high as 30%) of the Nazca/South America plate convergence over the past 10 Myrs. The timing of the slowdown is significant in that it is coeval with major growth of the Andes (Allmendinger et al. 1997) inferred from a variety of independent data (Gregory-Wodzicki 2000). Iaffaldano et al. (2006) tested the effect of topography on plate convergence by computing plate velocities before and after Andean topography growth, assuming mantle shear tractions from a MCM, as well as a low fault friction coefficient of 0.03 for the tectonic model. They performed two separate simulations of global plate motions, in order to estimate the increase of horizontal deviatoric force in the frictional region of the Nazca/South America interface arising from the gravitational collapse of the Andean belt, and its effect on the convergence rate. One simulation was based on present-day topography from the ETOPO 5 data set, as high as 5 km in the central Andes, to compute global plate velocities. A second simulation accounted for a lower topographic relief of continental South America, based on a reconstruction of Andean paleoelevation 10 Myrs ago (Gregory-Wodzicki 2000). Comparison of the two computed velocity fields predicts a 30% convergence reduction between the Nazca and South America plates (Fig. 4), and agrees well with the record of present and past plate motions. More importantly, the contribution of the momentum from the gravitational collapse of the large plateau has been estimated to be of some 10^{13} N/m along the central Andes. It should be pointed out that other mechanisms, such as temporal variations of friction along plate boundaries (Iaffaldano and Bunge 2008) and lateral buoyancy variations (Forte et al. 2009), may contribute to modify the convergence regime along the Nazca/South America margin to a minor extent. From Fig. 4, it is evident that the total convergence reduction is unevenly partitioned between the Nazca and South America plates. This can be understood by recalling that mantle shear tractions exerted on the lithosphere-base scale to first order with the basal surface area of the plate times its velocity. Because the Nazca plate is smaller than South America, a

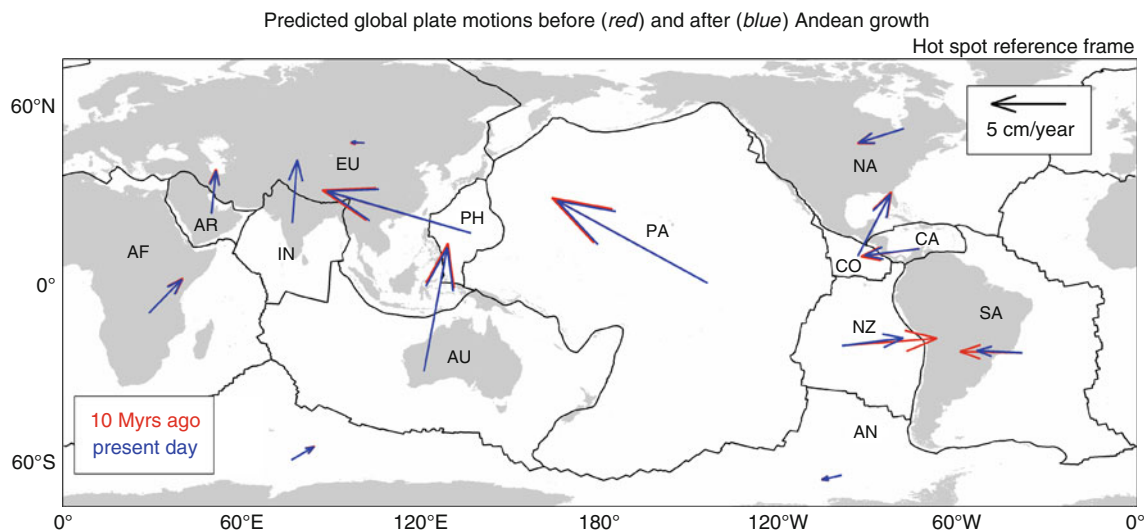


Plate Motions in Time: Inferences on Driving and Resisting Forces, Fig. 4 Predicted plate motions in the hot spot reference frame from coupled global mantle convection/lithosphere dynamics simulations, corresponding to assumed Andean paleotopography 10 Myrs ago (red) and present-day topography (blue). Plate boundaries are in black and continents in gray. AF Africa, AN Antarctica, AR Arabia, AU Australia, CA Caribbean, CO Cocos, EU Eurasia, IN India, NA North America, NZ Nazca, PA Pacific, PH Philippine, and SA South America. IN and AU are treated as separate in the computational grid, based on recent evidences of plate separation (see text and Figs. 6 and 7). Note that a lower paleotopography assumed for the Andes 10 Myrs ago results in a

predicted NZ/SA convergence of 10.1 cm/year at long 71.5°W, lat 25°S, whereas present-day topography results in a predicted convergence of 6.9 cm/year at the same position (see text). The rates compare remarkably well with observations inferred from a variety of data, which indicate a 30% reduction of NZ/SA plate convergence from 10.3 cm/year to 6.7 cm/year over the past 10 Myrs. The modeling results suggest that the reduction of NZ/SA convergence is caused by resisting plate-margin forces associated with the topographic load of Andes (see text). Similar numerical models moreover confirm that frictional variations along the boundary, arising from variations in trench sediment infill, are insufficient to explain the record of plate motion. (Courtesy of Elsevier © 2009)

higher velocity reduction is required in the momentum balance for mantle shear tractions to equilibrate topography-generated plate-boundary forces along the margin. Iaffaldano and Bunge (2009) took a step forward and used global models to predict the history of relative motion for plates adjacent to Nazca and South America (Pacific, Africa, Antarctica), which are shown in Fig. 5. Agreement between models and observations is remarkable, implying that the resisting forces along the Nazca/South America plate boundary are responsible for driving plate motion changes also in the Southern Atlantic and Southern Pacific regions.

Plate Motion Changes in the Indian Ocean

While instantaneous calculations of the plate tectonic momentum balance cannot be taken to model the temporal evolution of plate boundaries, they do allow us to test the effects of variations in plate geometry on global plate motions, and in particular the creation of new plate boundaries. From the principle of inertia, it follows that any such event would invariably trigger plate motion changes due to repartitions in the budget of basal drag and plate boundary forces. A recent such episode is thought to have occurred in the Indian Ocean, where a variety of evidence has been interpreted as the generation of a diffuse boundary between the India and Australia plates, dated between 8 and 20 Myrs ago (Wiens et al. 1985; Gordon et al. 1998). Ocean-floor

deformation at about 8 Myrs is documented from buckling of marine sediments (Weissel et al. 1980). Ongoing deformation in the Indian Ocean is also supported by pronounced ($M_s > 6.5$) and localized seismicity (Stein and Okal 1978) along the northern portion of the Ninety East Ridge (see inset in Fig. 6), suggestive of left-lateral strike-slip motion. Figure 6 shows the observed convergence history of India and Australia relative to Eurasia since early Miocene based on geodetic (Sella et al. 2002) as well as paleomagnetic (Cande and Stock 2004; DeMets et al. 1994; Gordon and Jurdy 1986; Merkouriev and DeMets 2006) data collected along the Carlsberg and South East Indian ridges (labeled respectively as CBr and SEIr in the inset). Within error-bars, convergence rates are almost indistinguishable between 20 and 11 Myrs ago, suggesting that India and Australia behaved as one single plate with presumably little deformation occurring in between. Over the past 11 Myrs, however, their convergence to Eurasia differs distinctly. While India slowed down by almost 2 cm/year, convergence of Australia to Eurasia remained almost steady, with only some 0.5 cm/year of reduction. Timing of the India/Eurasia plate-motion change coincides reasonably well with the occurrence of diffuse deformation in the Indian Ocean. More relevant is the fact that Tibet had attained most of its current elevation (Tapponnier et al. 2001) prior to the slowdown of the Indian plate and prior also to the presumed formation of the India/

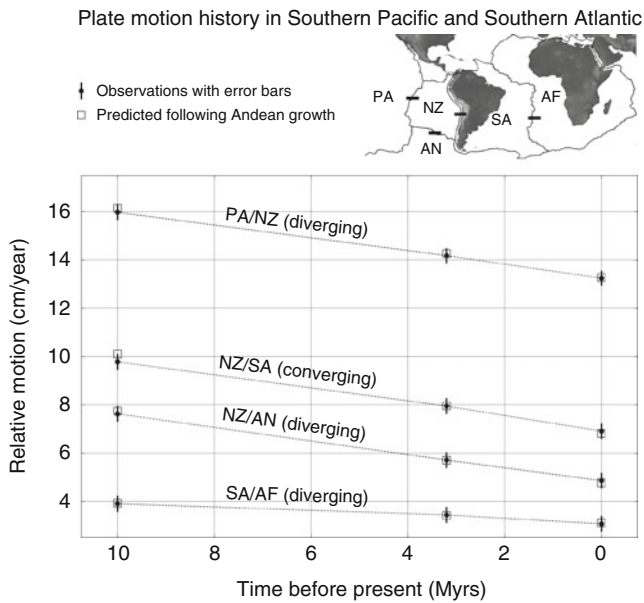


Plate Motions in Time: Inferences on Driving and Resisting Forces, Fig. 5 Predicted and observed relative plate motions in the South Atlantic and South Pacific over the past 10 Myrs for a set of adjacent plate pairs: PA/NZ, NZ/SA, NZ/AN, and SA/AF (abbreviations as in Fig. 6). *Black bold segments* in the small inset indicate positions along plate boundaries (*thin black*) at which relative motions have been computed. Observed plate motions (with *error bars*) inferred from paleomagnetic and geodetic data are represented by *black dots*, while *empty squares* indicate relative motions predicted from our simulations of the global coupled mantle/lithosphere system. The models explicitly account for the growth of the Andes over the past 10 Myrs, and demonstrate that the relative plate motion record can be entirely explained with the history of Andean orogeny. Our simulations thus point to the importance of far-field effects in plate tectonics and imply that resisting plate margin forces due to Andean growth account for about 18% of global plate motion changes over the past 10 Myrs (see Fig. 8). (Courtesy of Elsevier © 2009)

Australia plate boundary, implying that resistive plate boundary forces arising from the gravitational load of Tibet were already in place to act against convergence. Iaffaldano and Bunge (2009) tested explicitly whether plate-boundary forces from high Tibet are sufficient to explain the observed reduction of India/Eurasia plate convergence, once the former is separated from Australia by an additional plate boundary (Fig. 7). Specifically, they performed two distinct simulations of global plate motions, one with India and Australia cast as one single plate and the other with two plates built into the computational finite-element grid. A single India/Australia plate results in a predicted convergence of 5.2 cm/year at long 86°E, lat 27°N (Fig. 7a), whereas India being separated from Australia implies a convergence of 3.5 cm/year at the same position (Fig. 7b). The latter prediction compares remarkably well with the geodetic estimate (see Fig. 6). Finally, it is worth mentioning that simulations also predict an increased convergence between India and Australia,

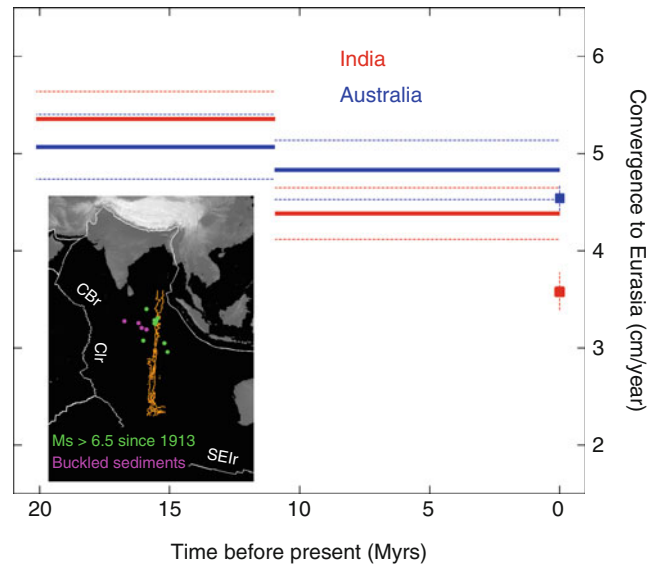


Plate Motions in Time: Inferences on Driving and Resisting Forces, Fig. 6 Observed convergence of India (*red*) and Australia (*blue*) relative to fixed Eurasia over the past 20 Myrs. Convergence rates are computed through rigid-rotation Euler poles at long 86°E, lat 27°N of the India/Eurasia margin. Present-day values (*squares*) are derived from geodetic techniques while the paleomagnetic record (*solid lines*) is computed by averaging finite rotations of magnetic anomalies identified along the Carlsberg and South-East Indian ridges (labeled respectively, as CBr and SEIr in inset). Note that convergence rates are very similar between 20 and 11 Myrs ago, when India and Australia appear to behave as one single plate with presumably little internal deformation. The convergence relative to Eurasia differs more distinctly over the last 11 Myrs, when India and Australia slowed down by almost 2 cm/year and 0.5 cm/year, respectively. Inset shows locations of identified unconformities of sediments (*magenta dots*) as well as great (*Ms > 6.5*) earthquakes (*green dots*) indicating left-lateral strike-slip motion in the northern portion of the Ninety East Ridge (*orange contours*). Those evidences suggest diffuse deformation in the Indian Ocean particularly pronounced during late Miocene and have been interpreted as separation between the India and Australia. Plate boundaries are in *white* and continental topography in *gray color* scale. (Courtesy of Elsevier © 2009)

concentrated in the Indian Ocean, compatible with the aforementioned geologic and geodetic record.

Conclusions

Recent results indicate that joint modeling of the mantle/lithosphere system begins to achieve a level of maturity that allows explicit testing of a range of hypotheses on the force balance in plate tectonics and identifying key controlling parameters. While buoyancy forces from MCMs contribute significantly to the dynamics of plate motion, it is clear that plate boundary forces are of sufficient magnitude relative to these driving forces to affect plate motions and plate deformation, and to initiate rapid plate motion changes. One key controlling parameter in regulating plate velocity is the elevation of large mountain belts, because their topographic load

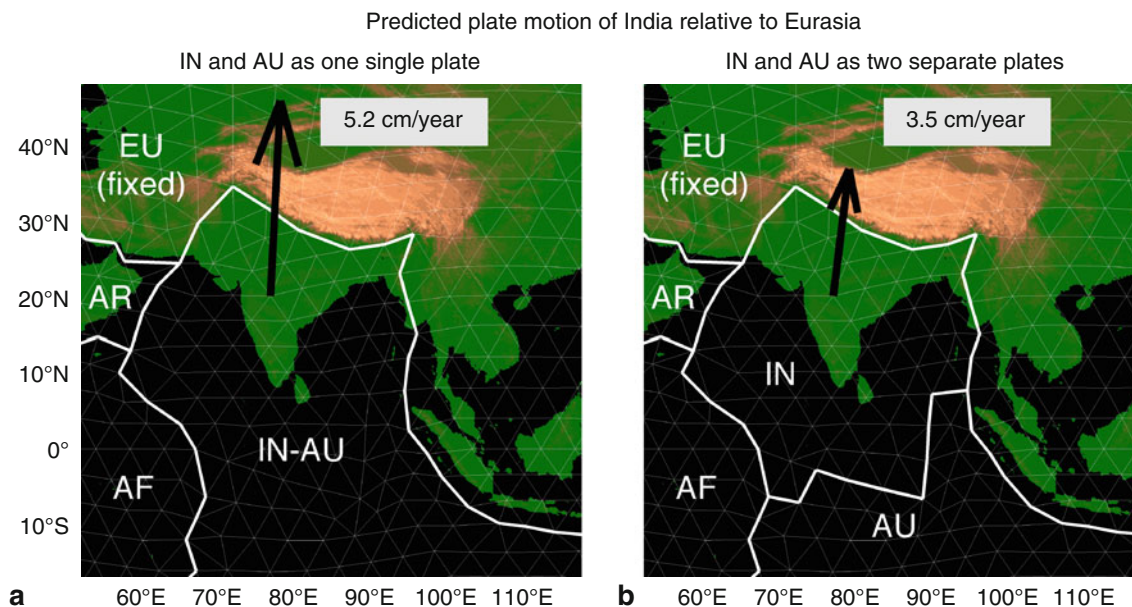


Plate Motions in Time: Inferences on Driving and Resisting Forces, Fig. 7 Predicted India (IN) plate motion relative to Eurasia (EU) from two distinct simulations with India and Australia (AU) acting, respectively, (a) as one single and (b) as two separate plates, where plate boundaries in our computational mesh are shown in *bold white* and finite elements in thin white. Plate motions are computed at long 86°E, lat 27° N. Abbreviations of plate names as in Fig. 6. Note that a single India/

Australia plate results in a predicted convergence of 5.2 cm/year relative to EU, incompatible with the geodetic estimate. Two separate plates result in a convergence of 3.5 cm/year of IN relative to EU, similar to the present-day observation. In the latter scenario, resisting forces from the gravitational load of Tibet act only against the smaller India plate and are thus more effective in slowing the convergent motion. (Courtesy of Elsevier © 2009)

consumes a considerable amount of the driving forces available in plate tectonics, as much as 10^{13} N/m. Along the Nazca/South America plate boundary, these forces are sufficient to reduce the convergence rate over the past 10 Myrs by some 30%. This reduction is, however, not an isolated episode of a rapid plate motion slow down. Instead, many such variations are documented from the global compilation of Müller et al. (2008), which points to the importance of topography and erosion in the global tectonic system (Cloetingh et al. 2007). The fact that models accurately predict the spreading history of the Pacific/Nazca, Nazca/South America, Nazca/Antarctica, and South America/Africa plate boundaries is of equal interest. The result is not entirely surprising and arises from the kinematic constraints of plate tectonics on the sphere. This suggests that far-field effects cannot be neglected in the geologic record at least in some cases. The strong influence of mountain belts on the plate tectonic force balance could have important implications. In an influential paper, Raymo and Ruddiman (1992) advanced the notion that Cenozoic climate change may have been caused by the uplift of Tibet. In other words, the rise of large mountain plateaus may act as a tectonic force on climate (Strecker et al. 2007). Low erosion rates have been implicated as a prerequisite for the creation of large mountain plateaus (Sobel et al. 2003; Cliff and Vannucchi 2004). This implication suggests conversely that climate can act – through large topography – as a force in plate tectonics. Overall, a significant portion of recent

changes in global plate motions can be attributed to topography-related forcing along plate boundaries rather than to mantle buoyancy. These findings are summarized in Fig. 8, where the relative plate motion changes observed globally over the past 10 Myrs are plotted. Green and red bars show variations in plate motion that are related, respectively, to the growth of the high Andes or to the presumed recent separation between India and Australia, and amount to about 35% of the total change over the Earth surface. This remarkable first-order result clearly demonstrates the ability of plate boundary forces to affect the global plate velocity field. The level of maturity achieved by neo-tectonic simulations coupled with 3-D MCMs thus allows geodynamicists to make explicit predictions of the plate tectonic force balance that can be tested against the geologic record of present and past plate motions.

Cross-References

- [Continental Drift](#)
- [Geodesy, Ground Positioning, and Levelling](#)
- [Geodesy, Networks, and Reference Systems](#)
- [GPS, Tectonic Geodesy](#)
- [Lithosphere, Oceanic](#)
- [Plate-Driving Forces](#)
- [Plates and Paleoreconstructions](#)

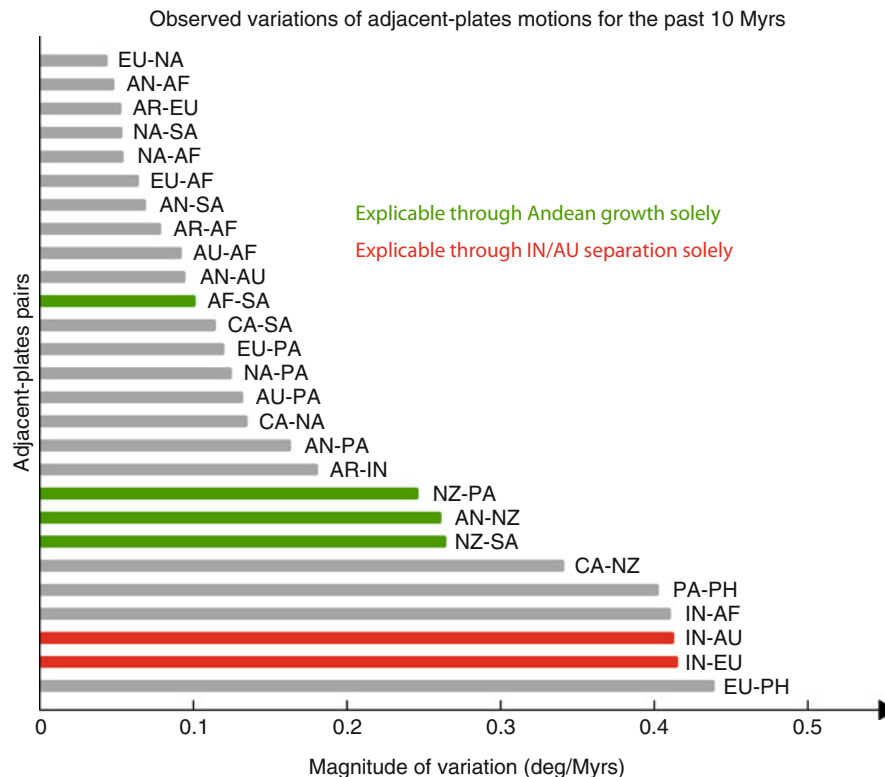


Plate Motions in Time: Inferences on Driving and Resisting Forces, Fig. 8 Observed variations of adjacent-plates motions over the past 10 Myrs. Abbreviations of plate names as in Fig. 6. For each couple of adjacent plates, variations are computed as magnitude of difference between relative rotation poles at 10 Myrs, derived from paleomagnetic reconstruction, and at present day, obtained through geodetic techniques (GPS). The Cocos oceanic plate is not considered, since a geodetic estimate for its rotation pole is not available. Variations of AF/SA, AN/NZ, NZ/SA, and NZ/PA adjacent-plates systems (*green bars*) can be entirely explained through the effect of Andean growth

(see Figs. 6 and 7). They amount to about 18% of the global relative motions changes over the past 10 Myrs. Variations of IN/EU as well as IN/AU relative motions (*red bars*) can be entirely explained through the effect of separation between India and Australia (see Fig. 7). They amount to about 17% of the global relative motions changes over the past 10 Myrs. Thus, our models of mantle/lithosphere dynamics explicitly predict about 35% of the global plate motion changes observed over the past 10 Myrs from two well-identified tectonic variations. (Courtesy of Elsevier © 2009)

Bibliography

- Allmendinger RW, Jordan TE, Kay SM, Isacks BL (1997) The evolution of the Altiplano-Puna plateau of the Central Andes. *Annu Rev Earth Planet Sci* 25:139–174
- Bercovici D (1995) A source-sink model of the generation of plate tectonics from non-Newtonian mantle flow. *J Geophys Res* 100:2013–2030
- Bercovici D (2003) The generation of plate tectonics from mantle convection. *Earth Planet Sci Lett* 205:107–121
- Bird P (1998) Testing hypotheses on plate-driving mechanisms with global lithosphere models including topography, thermal structure, and faults. *J Geophys Res* 103:10115–10129
- Bunge H-P (2005) Low plume excess temperature and high core heat flux inferred from non-adiabatic geotherms in internally heated mantle circulation models. *Phys Earth Planet Inter* 153:3–10
- Bunge H-P, Richards MA, Baumgardner JR (1996) The effect of depth dependent viscosity on the planform of mantle convection. *Nature* 379:436–438
- Bunge H-P, Richards MA, Baumgardner JR (1997) A sensitivity study of 3-D spherical mantle convection at 10^{28} Rayleigh number: effects of depth dependent viscosity, heating mode and an endothermic phase change. *J Geophys Res* 102:11991–12007
- Bunge H-P, Richards MA, Lithgow-Bertelloni C, Baumgardner JR, Grand SP, Romanowicz BA (1998) Time scales and heterogeneous structure in geodynamic earth models. *Science* 280:91–95
- Byerlee JD (1978) Friction of rocks. *Pure Appl Geophys* 116:1189–1198
- Cande SC, Stock JM (2004) Pacific-Antarctic-Australia motion and the formation of the Macquarie plate. *Geophys J Int* 157:399–414
- Clift P, Vannucchi P (2004) Controls on tectonic accretion versus erosion in subduction zones: implications for the origin and recycling of the continental crust. *Rev Geophys* 42:RG2001
- Cloetingh SAPL, Ziegler PA, Bogaard PJF, Andriessen PAM, Artemieva IM, Bada G, van Balen RT, Ben-Avraham Z, Brun J-P, Bunge H-P, Burov EB, Carbonell R, Facenna C, Gallart J, Green AG, Heidbach O, Jones AG, Matenco L, Mosar J, Oncken O, Pascal C, Peters G, Sli-aupa S, Soesoo A, Spakman WRS, Thybo H, Torsvik TH, de Vicente G, Wenzel F, Wortel MJR, the TOPO-EUROPE Working Group (2007) TOPO-EUROPE: the geoscience of coupled deep earth – surface processes. *Glob Planet Chang* 58:1–118
- Davies GF, Richards MA (1992) Mantle convection. *J Geol* 100:151–206
- DeMets C, Gordon RG, Argus DF, Stein S (1994) Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions. *Geophys Res Lett* 21:2191–2194

- Dixon TH (1991) An introduction to the Global Positioning System and some geological applications. *Rev Geophys* 29:249–276
- Forsyth DW, Uyeda S (1975) Relative importance of driving forces of plate motion. *Geophys J R Astron Soc* 43:163–200
- Forte AM (2010) Plate driving forces. In: *Encyclopedia of solid earth geophysics*. Springer, 2011
- Forte AM, Moucha R, Rowley DB, Quere S, Mitrovica JX, Simmons NA, Grand SP (2009) Recent tectonic plate decelerations driven by mantle convection. *Geophys Res Lett* 36:L23301
- Ghosh A, Holt WE, Wen L, Haines AJ, Flesch LM (2008) Joint modeling of lithosphere and mantle dynamics elucidating lithosphere-mantle coupling. *Geophys Res Lett* 35:L16309
- Gordon RG, Jurdy DM (1986) Cenozoic global plate motions. *J Geophys Res* 91:12389–12406
- Gordon RG, DeMets C, Royer J-Y (1998) Evidence for long-term diffuse deformation of the lithosphere of the equatorial Indian Ocean. *Nature* 395:370–374
- Gregory-Wodzicki KM (2000) Uplift history of the central and northern Andes: a review. *Geol Soc Am Bull* 112:1091–1105
- Hager BF, O'Connell RJ (1981) A simple global-model of plate dynamics and mantle convection. *J Geophys Res* 86:4843–4867
- Hickman SH (1991) Stress in the lithosphere and the strength of active faults. *Rev Geophys* 29:759–775
- Humphreys ED, Coblenz DD (2007) North American dynamics and Western US tectonics. *Rev Geophys* 45(2):RG3001
- Iaffaldano G, Bunge H-P (2008) Strong plate coupling along the Nazca/South America convergent margin. *Geology* 36:443–446
- Iaffaldano G, Bunge H-P (2009) Relating rapid plate-motion variations to plate-boundary forces in global coupled models of the mantle/lithosphere system: effects of topography and friction. *Tectonophysics* 474:393–404
- Iaffaldano G, Bunge H-P, Dixon TH (2006) Feedback between mountain belt growth and plate convergence. *Geology* 34:893–896
- Karato S, Jung H (2003) Effects of pressure on high-temperature dislocation creep in olivine. *Philos Mag* 83:401–414
- Kirby SH (1983) Rheology of the lithosphere. *Rev Geophys* 21:1458–1487
- Koeppen W, Wegener A (1924) *Die Klimate der geologischen Vorzeit*. Borntraeger, Berlin
- Kohlstedt DL, Evans B, Mackwell SJ (1995) Strength of the lithosphere: constraints imposed by laboratory experiments. *J Geophys Res* 100:17587–17602
- Kong X, Bird P (1995) SHELLS: a thin-shell program for modeling neotectonics of regional or global lithosphere with faults. *J Geophys Res* 100:22129–22132
- Merkouriev S, DeMets C (2006) Constraints on India plate motion since 20 Ma from dense Russian magnetic data: implications for Indian plate dynamics. *Geochem Geophys Geosyst* 7:Q02002
- Moresi L, Solomatov V (1998) Mantle convection with a brittle lithosphere: thoughts on the global tectonic styles of the Earth and Venus. *Geophys J Int* 133:669–682
- Morgan WJ (1968) Rises, trenches, great faults and crustal blocks. *J Geophys Res* 73:1959
- Müller RD, Sdrolias M, Gaina C, Roest WR (2008) Age, spreading and spreading asymmetry of the world's ocean crust. *Geochem Geophys Geosyst* 9:Q04006
- Norabuena EO, Dixon TH, Stein S, Harrison CGA (1999) Decelerating Nazca-South America and Nazca-Pacific plate motions. *Geophys Res Lett* 26:3405–3408
- Oeser J, Bunge H-P, Mohr M (2006) Cluster design in the earth sciences. In: *TETHYS, high performance computing and communications – second international conference, HPCC 2006, Munich. Lecture notes in computer science, vol 4208, pp 31–40*
- Quere S, Forte AM (2006) Influence of past and present-day plate motions on spherical models of mantle convection: implications for mantle plumes and hotspots. *Geophys J Int* 165:1041–1057
- Raymo ME, Ruddiman WF (1992) Tectonic forcing of late Cenozoic climate. *Nature* 359:117–122
- Sella GF, Dixon TH, Mao A (2002) REVEL: a model for recent plate velocities from space geodesy. *J Geophys Res* 107:2081–2111
- Sobel ER, Hilley GE, Strecker MR (2003) Formation of internally drained contractional basins by aridity-limited bedrock incision. *J Geophys Res* 108:2344
- Stein RS (1987) Contemporary plate motion and crustal deformation. *Rev Geophys* 25(5):855–863
- Stein S, Okal E (1978) Seismicity and tectonics of the Ninetyeast Ridge area: evidence for internal deformation of the Indian plate. *J Geophys Res* 83:2233–2246
- Strecker MR, Slonso RN, Bookhagen B, Carrapa B, Hilley GE, Sobel ER, Trauth MH (2007) Tectonics and climate of the Southern Central Andes. *Annu Rev Earth Planet Sci* 35:747–787
- Suppe J (2007) Absolute fault and crustal strength from wedge tapers. *Geology* 35(12):1127–1130
- Tackley PJ, Stevenson DJ, Glatzmaier GA, Schubert G (1994) Effects of multiple phase transitions in a three-dimensional spherical model of convection in Earth's mantle. *J Geophys Res* 99:15877–15902
- Tapponnier P, Zhiqin X, Roger F, Meyer B, Arnaud N, Wittlinger G, Jingsui Y (2001) Oblique stepwise rise and growth of the Tibet plateau. *Science* 294:1671–1677
- van der Hilst RD, de Hoop MV, Wang P, Shim S-H, Ma P, Tenorio L (2007) Seismostratigraphy and thermal structure of Earth's core-mantle boundary region. *Science* 315:1813–1817
- Weissel JK, Anderson RN, Geller CA (1980) Deformation of the Indo-Australian plate. *Nature* 287:284–291
- Wiens DA, DeMets C, Gordon RG, Stein S, Argus D, Engeln JF, Lundgren P, Quible D, Stein C, Weinstein S, Woods DF (1985) A diffuse plate boundary model for Indian Ocean tectonics. *Geophys Res Lett* 12:429–432
- Wilson JT (1965) A new class of faults and their bearing on continental drift. *Nature* 207:343–347
- Zhong S, Zuber MT, Moresi L, Gurnis M (2000) Role of temperature-dependent viscosity and surface plates in spherical shell models of mantle convection. *J Geophys Res* 105:11063–11082

Plate Tectonics, Precambrian

Y. J. Bhaskar Rao, T. Vijaya Kumar and E. V. S. S. K. Babu
Geochronology Division, CSIR-National Geophysical
Research Institute, Hyderabad, India

Definition

Precambrian (>541 million years, Myr) is divisible into the Hadean Eon (>4.0 billion years, Gyr), Archean Eon (4.0–2.5 Gyr), and the Proterozoic Eon (2.5–0.54 Gyr). The Archean and Proterozoic Eons are further divisible into Eras: Eo-archean (4.0–3.6 Gyr), Paleo-archean (3.6–3.2 Gyr), Meso-archean (3.2–2.8 Gyr), Neo-archean (2.8–2.5 Gyr), Paleo-proterozoic (2.5–1.6 Gyr), Meso-proterozoic (1.6–1.0 Gyr), and Neo-proterozoic (1.0–0.541 Gyr). The Phanerozoic Eon corresponds to <541 Myr. These Era names are also used with no hyphen.

Zircon (ZrSiO_4) is a mesosilicate found as an accessory phase in a wide compositional range of igneous, metamorphic, and sedimentary rocks of all ages. Zircon crystallizes in silicate melts and incorporates a wide range of high field strength incompatible elements; most common is Hafnium (Hf) with abundance ranging generally up to 4% by weight. Zircon is used widely in geochronology by the U-Th-Pb method of radiometric dating.

Introduction

Compared to other terrestrial planets that have basaltic (mafic) crusts, Earth is unique in having a prominent felsic (continental) crust alongside a mafic (oceanic) crust and most interestingly, a long sustained “Plate Tectonics” geodynamic regime (Taylor and McLennan 1985; Stevenson 2003; Rudnick and Gao 2014). In general, Earth’s continental crust is significantly thicker than the oceanic crust and is the principal archive of a rock record that dates back to ~ 4.0 Gyr. Therefore, understanding the geodynamic processes that underpin the formation of continents is fundamental to decoding the complex evolutionary history of the Earth and the terrestrial planets. In general, the geodynamic styles reflect how the planets lose their internal heat and plate tectonics represents a rather unusual mode of cooling (Sleep 2005; Lenardic 2018). The basic tenet of the plate tectonics paradigm is the observation that Earth’s lithosphere, a torsionally rigid thermal boundary conductive layer overlying the convective asthenosphere, comprises more than a dozen fragments, the “Plates”. Plate tectonics describes the essentially horizontal relative motions between the plates on the nearly-spherical surface of the Earth, in terms of the rotational pole and angular velocity. The loci of orogenic (mountain building) processes are plate boundaries that are mainly of three types: (1) divergent, (2) convergent, and (3) transform. The oceanic crust and the underlying depleted lithospheric mantle is created at divergent boundaries such as the Mid Ocean Ridge (MOR) system. This lithosphere is transported away from the MOR by the transform fault system. Formation of continental crust is a consequence of orogenic processes associated with long lived magmatic arcs along convergent plate boundaries, where the oceanic lithosphere subducts beneath oceanic or continental lithospheric plates. Along these subduction zones, calc-alkaline magmatism, due to hydrous melting of the mantle wedge overlying the subducting lithosphere, is a key product of the “subduction factory” (Keleman et al. 2004; Tatsumi 2005). Accretion of magmatic arcs and a variety of other terranes results in the development of continents and supercontinents (Cawood and Buchan 2007; Cawood et al. 2009; Santosh 2010; Nance et al. 2014). Because steep subduction of the oceanic lithosphere is the principal driver of plate tectonics on modern (Phanerozoic) Earth, the process is commonly referred to as “subduction tectonics,” the hallmark of modern style plate

tectonics (Forsyth and Uyeda 1975; Stern 2005; Hamilton 2007). Two contrasting end-member types of orogenic systems operate on modern Earth: (1) the dominantly accretionary or the external type characterized by a significant juvenile crustal component, e.g., the Circum Pacific orogen (arc) system, and (2) the dominantly collisional or interior orogens that are characterized by composite terranes of juvenile and recycled crustal components, e.g., the Indonesian-Himalayan-Alpine mountain belt where terrane accretion involving “Wilson-cycle” (Wilson 1968) processes results in the formation of large continental masses (Stern 2005; Cawood and Buchan 2007; Santosh 2010; Collins et al. 2011; Nance et al. 2014).

Unlike in the Phanerozoic Eon, where the paradigm of plate tectonics and Wilson-cycle processes constitutes a robust and unifying mechanism in which to explain not only the orogenic processes of crust formation, recycling, preservation but also the complex interactions between the solid earth, biosphere, and atmosphere relevant to evolution of life and climates, there is yet no consensus on a global geodynamic framework for the early Earth, especially during the Hadean and Archean Eons (Harrison 2009; Kemp et al. 2010, 2015; Collins et al. 2011; Bédard 2013, 2018; Cawood et al. 2013; O’Neill and Debaille 2014; Roberts and Spencer 2015; Mueller and Nutman 2017; Harrison et al. 2017; Hawkesworth et al. 2017; O’Neill et al. 2018). Many fundamental and long standing questions, such as *why did the Earth adopt plate tectonics?* and *since when was plate tectonics operative on Earth?* remain challenging and unresolved problems in the evolution of the planet (de Wit and Ashwal 1995; Condie and Pease 2008; Ernst 2009; Cawood et al. 2018). While some authors propose that plate tectonics, *in some form*, operated during the Hadean Eon (Harrison 2009; Tarduno et al. 2015; Furnes et al. 2014; Kusky et al. 2018), and some others favored an Archean onset (Sleep 2005; Parman et al. 2001; Smithies et al. 2003, 2005; Cawood et al. 2006, 2013; Condie and Kroner 2008; Dilek and Polat 2008; Naeraa et al. 2012; Korenaga 2013), others favor a much later beginning in the Earth’s history, viz., during the Neoproterozoic (Davies 1992; Hamilton 1998, 2007; Stern 2005, 2007; Dewey 2007; Brown 2008). Currently the debate focuses on the onset of modern style plate tectonics, clearly implying steep geotherms rather than *any* horizontal plate motion.

One other important mechanism by which planetary bodies lose internal heat to the surface is volcanism. For instance, Jupiter’s moon, Io is characterized by a much higher surface heat flow than modern Earth. Io cools through intrusive magmatism and volcanism with “heat pipes” as the plumbing conduits (Moore and Webb 2013) rather than plate tectonics. Away from the plate boundaries on modern Earth, “mantle plumes” and “hot spots” contribute to intense volcanism leading to the development of oceanic plateaux and large igneous provinces. Mantle plumes that are posited to originate from deep thermo-chemical boundary layers,

such as either the core-mantle boundary or the D'' layer (~660 km), account for a minor part of the planet's heat loss (Morgan 1971; Lay et al. 2008). Although not mutually exclusive, the relative importance of plate and plume tectonics (broadly, horizontal and vertical tectonic styles) in terms of the growth of continental crust through geologic time has long been a matter of debate (de Wit and Ashwal 1995; Bédard 2006; Van Kranendonk 2010; Gerya 2014; Griffin et al. 2014; Harris and Bédard 2014; Korenaga 2018). There is no dispute that Earth has been on a cooling trajectory since its very early history and that even during the Archean, the mantle was hotter by ~250 °C and the radiogenic heat production was three to five times higher than the present (Abbott et al. 1994; Korenaga 2006; Labrosse and Jaupart 2007; Herzberg et al. 2010). This has a direct bearing on the thickness, composition, and rheology of early crust and forms the basis for rationalizing the observed secular changes in the global geologic record and the geodynamic regime, especially in the context of the onset of plate tectonics. Demonstration that plate tectonics operated at any given time requires evidence for lithospheric subduction and independent plate motions. However, since the oldest oceans on present Earth are not older than Jurassic (201–145 Myr), one is forced to consider the geological record on continents preserved in the Precambrian shields, which include Archean cratons that archive Earth's most ancient geologic record.

Over the last three decades, research on early Earth processes has gained immensely from: (1) petrological and geochemical constraints on the genesis of predominant lithologies of Archean cratons; (2) in-situ zircon U-Pb geochronology and Hf±O isotopic compositions, especially on Hadean and Archean detrital zircon grains from sources such as the Jack Hills metaconglomerate/quartzite, Yilgarn craton, Western Australia (Wilde et al. 2001; Harrison 2009; Belousova et al. 2010; Kemp et al. 2010, 2015; Voice et al. 2011; Condie et al. 2011; Griffin et al. 2014; Roberts and Spencer 2015; Harrison et al. 2017; Hawkesworth et al. 2017; Mueller and Nutman 2017); (3) numerical modeling of geodynamic processes (Korenaga 2013; O'Neill and Debaille 2014; Gerya 2014; Sizova et al. 2014, 2015); and (4) remote observation of terrestrial planets and study of extra-terrestrial materials (e.g., Moore and Webb 2013; Harris and Bédard 2014; O'Neill et al. 2017; Lenardic 2018). Drawing from this global endeavor, we summarize here the different perspectives on early Earth geodynamics and the debate on the onset of modern-style plate tectonics on the planet.

Modern-Style Plate Tectonics and Criteria for Recognizing Plate Tectonics in the Precambrian Record

In its initial conceptualization, plate tectonics offered a purely kinematic description of how lithospheric plates move and

interact with one another on the nearly spherical surface of the Earth. There was little clarity on the forces that cause plate motions. Notwithstanding the traditional view that the asthenospheric convection powers contemporary plate motions, modern geodynamicists find an increasing consensus for predominance of “top-down tectonics” (Anderson 2001), at least during some stages in the Earth's history, and suggest that plates drive themselves as they organize the overall mantle convection (Davies 1999; Hynes 2005). The lithosphere is buoyant when new oceanic crust is created at the MOR. As the lithosphere ages and moves away from the MOR, it cools (conductively), thickens, and increases in density leading to a buoyancy-crossover in the lithosphere–asthenosphere system. On the present Earth, such a crossover occurs in 20–40 million year intervals. This “ridge-push” coupled with “gravitational descent” of the lithosphere in the subduction zones contributes to much of the plate motion. Based on seismic tomography, it is evident that the lithospheric slabs on Earth reach great depths, typically 1100–1300 km, and possibly down to the core-mantle boundary (Van der Hilst et al. 1997). While the “slab pull” of the sinking lithosphere and the associated “slab suction” of the entrained ambient mantle at subduction zones move the plates (Conrad and Lithgow-Bertelloni 2004), while the viscous resistance of the asthenosphere impedes plate motion. Resistance to the sinking lithosphere is also from dissipation forces that are manifest as deformation (mainly faulting and bending) in the subducting slab and phenomena such as regional mantle down swelling. The nature and extent of equilibrium between the driving and resisting forces is indicated by the limited range of the observed plate velocities, which possibly did not exceed ~17 cm/year, for instance, during the Mesozoic drift of the Indian lithospheric plate, a case of extremely rapid plate motion.

Recognizing the on-set of modern-style plate tectonics on planet Earth involves tracking the manifestation of the processes in the geological record. For instance, evidences for the creation of new oceanic lithosphere, the existence of “ocean-plate stratigraphy”, subduction processes where the lithosphere is returned to the mantle, transform and transcurrent faulting and so on at any given time. The problem of preservation of the diagnostic criteria figures prominently, as the evidence in all likelihood, can be fragmentary and incomplete owing to obliteration by tectonic and surface processes. From the theoretical considerations presented in the foregoing sections, the modern-style plate tectonics is not only unusual in a planetary perspective, but also in the Earth's evolutionary history. Thus, the Huttonian uniformitarian approach “The present is the key to the past” may not be entirely helpful to understand the Earth's earliest tectonic history (Stern 2008). In any case, best approaches should consider multiple working hypotheses and methodological uniformitarianism emphasizing upon constancy of natural laws rather than substantive uniformitarianism invoking uniform geological processes through time (Stern 2008).

Lithotectonic elements of plate convergence could be quite diverse and combined interpretation of data from different disciplines assumes importance. Thus, a combination of information including geological structures, lithology, geochronology and isotope geochemistry, magma-type, metamorphic conditions, P-T-t paths, paleomagnetic evidence for relative motion between terranes and 2-D/3-D geometry based on geophysical data provide insights into geodynamic processes in space and time. Some of the most robust criteria for recognizing the operation of the plate tectonics in the past that were at the core of the debate are as follows. In general, different authors apply the criteria in different ways. For example, while some consider when a particular assemblage or property first appeared, others emphasize on when it became common. Thus, the inferences on the onset of plate tectonics on the Earth could be quite subjective, even by the same criterion.

Ophiolites: Ophiolites are slices of the oceanic lithosphere tectonically emplaced onto the continental margins or incorporated into accretionary prisms and could be reliable indicators of past plate tectonics. As indicated in the early work (e.g., Coleman et al. 1965; Penrose Conference 1972), a complete ophiolite sequence ocean-plate stratigraphy would include a succession (from top downward) of pelagic sediments, pillow basalt, sheeted dykes, isotropic gabbros, ultramafic cumulates, and tectonized ultramafites (lherzolite and harzburgite), the latter representing components of the depleted mantle. However, generally, fragmentary and dismembered sequences are encountered in Precambrian terrains. Well preserved ophiolites are known from many Neoproterozoic (up to ca. 1.0 Gyr old) and younger terrains (Stern 2008; Ernst 2007, 2009). Some of the better-preserved Neoproterozoic ophiolites include Purtuneni ophiolites, Canada (Scott et al. 1992), and Jormua ophiolites, Finland (Peltonen et al. 1996), although ophiolites as old as ca. 3.8 Gyr (the Isua suite, Greenland; Furnes et al. 2007) have been proposed, they remain controversial (also see Bickle et al. 1994). Different types of ophiolites have been proposed and described over the last two decades based on a better characterization of the petrology, structure, and composition of the oceanic crustal sections. These include the magma-poor and magma-rich types and discrete ophiolite types that represent the near-MOR segments as against those at supra-subduction settings (Furnes et al. 2014; Kusky et al. 2018 and references therein). Current views on ocean plate stratigraphy also appreciate the spatio-temporal variations in the sedimentary and magmatic deposits on to the ocean floor that are distinct from the MOR and the subduction zone environments. This is important because eventually they all get transported, scrapped-off into the accretionary wedge, where the packages get imbricated and strongly deformed. Thus, an ophiolite may include pillow lava grading upward into pelagic chert, limestone, banded iron formation, shales, and turbidites from both proximal and distal sources and not comply strictly with the

Penrose definition of ophiolites. These suites may form structurally complex tectonic melanges marked by thrusts and nappe complexes. Such packages currently described from ancient terranes of all ages support the view that plate tectonics may have operated since the late Hadean Eon. Some examples across the timescale include the Archean Nulliak greenstone belt, Labrador, the 3.8 Gyr Isua greenstone belt, W. Greenland, and the ~3.3 Gyr Barbarton greenstone belt, South Africa, which are considered to be examples of Paleoconvergent plate boundaries or interpreted as relics of ancient accretionary orogens (Komiya et al. 2015; de Wit et al. 2018; Kusky et al. 2018). Some authors, e.g., Condie (2015), favored a definition restricting usage of the term “ophiolite” in Archean terranes to fragments of oceanic crust $\pm$ mantle, i.e., formed at an oceanic ridge, but not at the arc. By such a filter, ophiolites begin to appear in Paleoproterozoic, around 2.0 Gyr although some dismembered older (ca 2.5 Gyr) examples may also qualify.

Blueschists, low-T eclogites, and paired metamorphic belts: In modern subduction systems, the down going oceanic lithospheric slab depresses isotherms creating a low- dT/dP environment. Fluids and melts from the slab generate high- dT/dP environments in the overlying mantle wedge. This leads to the formation of “paired metamorphic belts” with the characteristic metamorphic assemblages (Oxburgh and Turcotte 1971; Miyashiro 1972; Brown 2010). Typically, they form on the overriding plate with the high-P low-T (blueschist facies) terrane next to the trench, whereas the low-P high-T terrane (greenschist facies) develops near to the arc. Paired metamorphic belts represent the thermal structure of a mature subduction zone (Van Keken et al. 2002; Kelemen et al. 2004). As these belts are usually very wide, tens of hundreds of kilometers, they could be robust features that survive several episodes of orogeny and erosion (Stern 2008). In a paired metamorphic belt, accretionary convergent margins are preserved as broad zones, whereas the erosional boundaries could be much narrower (Stern 2008 and references cited therein).

Blueschists are metamorphosed mafic rocks bearing predominantly a sodic amphibole that stabilizes at high-P, low-T typical of Pacific-type of orogenic belts. Best studied examples are the Franciscan and Sanbagwa terrains (Ernst 2003). The oldest known blueschists are Neoproterozoic (800–700 Myr; Maruyama et al. 1996) and undisputed pre-Neoproterozoic blueschists are virtually unknown. For many authors, the absence of blueschists in Archean terrains implies non-operation of plate tectonics (e.g., Stern 2005; Bédard 2006; Hamilton 2011), while for others, it implies subduction in hotter mantle temperature gradients and/or that the high-Mg content of Archean magmas inhabits the development of blueschist assemblages (e.g., Palin and White 2016).

Eclogites (s.s) are metabasalts and metagabbros that equilibrated at $P > 1.2$ GPa (>45 km) and $T (>600\text{--}650^\circ\text{C})$ with a characteristic pyropic-garnet and omphacitic-clinopyroxene-

bearing assemblage. Different types of eclogites have been described, but Coleman's group C (associated with glaucophane schists) and the "low-T eclogites" (Carswell 1990) could be reliable indicators of subduction process. Archean eclogites are rare and become abundant post Archean, mostly as mantle or crustal xenoliths in kimberlites and rarely in outcrop (Rollinson 1997). Study of eclogite inclusions in diamonds that originate in thick Archean SCLM suggests that the modern style subduction may not have occurred prior to ca 3.0 Gyr (Shirey and Richardson 2011).

UHP (Ultra High Pressure) terrains form when continental crust initially subducted to greater than 100 km returns to the surface through subsequent tectonic processes. Typical assemblages include coesite and diamond, indicative of 700–900 °C and 3–4 GPa, or more (Liou et al. 2004). Some of the oldest known UHP terranes are in Mali, ca. 620 Myr (Jahn et al. 2001) and Kazakhstan, 530 Myr (Maruyama and Liou 1998).

Batholiths: Mere presence of magmatic "island arc" features suggests modern-style plate tectonics with deep subduction (100–250 km). In such a setting from older terrains, volcanic rocks are usually removed by erosion, leaving behind well-preserved mid-crustal batholiths (Hamilton and Myers 1967; Stern 2002; Hamilton 2011). Felsic plutonic lithologies, well represented by Tonalite-Trondhjemite-Granodiorite (TTG) suites, dominate Precambrian plutons coalesced into batholiths.

Composition of igneous rocks: The relationship between tectonic setting and composition of igneous rocks has been a subject of considerable interest (see Pearce and Peate 1995; Pearce 2008). However, the diversity of sources, magmatic processes, and the problem of element mobility in secondary environments pose a serious challenge to "geochemical fingerprinting" of tectonic settings (Condie 2015). In general, the distribution of immobile trace elements and ratios (e.g., incompatible elements including Rare Earths, Y, Th, and High Field Strength Elements, such as Zr, Ta, Hf, and Nb) with similarity to modern arc magmas and the identification of certain magmatic associations like boninites and adakites in Archean granite-greenstone terranes have been used widely as evidence for operation of plate tectonics in the Archean (Foley et al. 2003; Kerrich and Polat 2006; Shirey et al. 2008).

The petrogenesis and tectonic setting of the Archean sodic TTG complexes, the most voluminous lithology of the cratons, especially the early Archean (>3.2 Gyr); has been a subject of considerable discussion (Moyen and Martin 2012; Laurent et al. 2014; Hoffmann et al. 2014). Recent models based on geological and geochemical observations and thermomechanical-numerical modeling at conditions similar to Eo-Mesoarchean times invoke melting of hydrated basalt at garnet amphibolite, granulite, or eclogite facies conditions in different tectonic scenarios including subduction, as well as the range of nonplate tectonic (no subduction) scenarios

and/or a combination of both (Rapp et al. 1991; Hill et al. 1992; Martin 1993; Abbott et al. 1994; Van Thienen et al. 2004; Smithies et al. 2005; Bédard 2006; Van Kranendonk et al. 2007; Van Kranendonk 2010; Moyen and Martin 2012; Sizova et al. 2015).

Seismic reflections of deep crust: In general, deep-crustal seismic reflection studies in many cratons indicated: (1) gentle to moderately dipping crustal detachments identified based on the geometry of reflections and truncations; and (2) reflections projecting from the lower crust into the upper mantle suggesting that the crust may have been thrust into the mantle (e.g., Yilgarn Craton, Australia, Drummond et al. 2000; Goleby et al. 2004; Kaapvaal craton, South Africa, de Wit and Tinker 2004; Dharwar craton, India, Vijaya Rao et al. 2015; Mandal et al. 2018). Neoproterozoic (2.7 Gyr) plate tectonics was inferred in the Superior craton, Canada, by the discovery of "mantle reflections" and interpretation that these reflections reflect "subduction scars" or "fossil subduction" (Calvert et al. 1995; Calvert and Ludden 1999; van der Velden and Cook 2005).

Palaeomagnetic data: The palaeomagnetic method offers a quantitative means to test relative motions between plates and the operation of Plate Tectonics during the Precambrian. In general, there is evidence in the Palaeomagnetic record that a modern-like geocentric axial dipole geomagnetic field with a weaker or comparable intensity to the present existed during the Phanerozoic and Proterozoic Eons. High quality palaeomagnetic data especially on magmatic associations like mafic dyke swarms could help in analyzing differential movements of continents at least since the Neoproterozoic (~3.0 Gyr) (McElhinny and Senanayake 1980; Pesonen et al. 2003; Evans and Pisarevsky 2008 and references cited therein). However, in the absence of precise geochronological data and poor constraints of field stability tests of remanence, the application of palaeomagnetic method to Archean terranes has been problematic.

Evolution of Earth's Geodynamic Regime and the Onset of Plate Tectonics

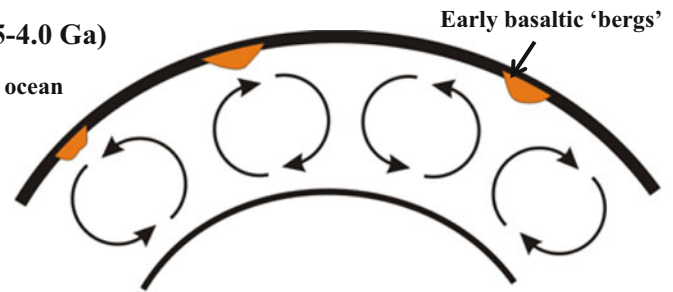
The Hadean Eon

Thermal models of planetary evolution suggest three principal modes of nonradiative planetary heat transfer (Sleep 2000): (1) magma mush oceans, (2) plume and plate tectonics, and (3) stagnant-lid convection (illustrated in Fig. 1 in corresponding stages). The magma mush ocean mode is, by far, the most energetic mode and probably characterizes the early Hadean tectonic style. The stagnant-lid convection, the least energetic mode, characterizes the present Venus and Mars and was possibly present on Earth episodically. In general, several simulations of Hadean geodynamic environments suggest the difficulty of sustained plate motion and

Plate Tectonics, Precambrian,
Fig. 1 Cartoon illustrating the evolution of global tectonic styles from the Hadean to Present involving three main stages. See text for more details. (Modified after Ernst 2007; Stern 2008 and references cited therein)

Stage I: Hadean (4.55–4.0 Ga)

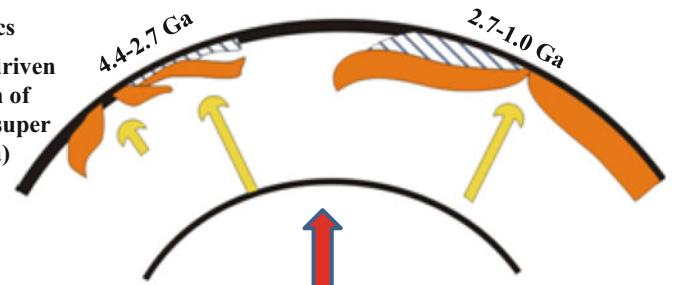
Convecting magma/mush ocean



Stage II: Archaean-Proterozoic (4.0–1.0 Ga)

II.A. Proto-plate tectonics

Plume and deep mantle-driven circulation and formation of platelet (4.4–2.7 Ga) and supercratonic stage (2.7–1.0 Ga)



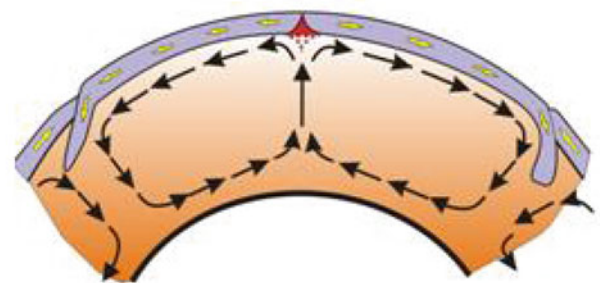
II. B. Unstable Stagnant Lithospheric Lid

Formation of oceanic and continental crust through mantle upwelling processes



Stage III: Modern-style plate tectonics <1.0 Ga

Passive mantle upwelling at mid-oceanic ridges and sinking of dense lithospheric plates at subduction zones.



subduction processes for the hotter mantle/lithosphere (Sizova et al. 2010; Rozel et al. 2017; O'Neill et al. 2017) during the early Hadean, plume induced subduction (Gerya et al. 2015), and impact triggered short-lived subduction events have been proposed (O'Neill et al. 2017). Nonplate tectonic scenarios such as stagnant-lid, heat pipes, and mantle plumes were also inferred in numerical simulations and thermo mechanical models (Touboul et al. 2012; Moore and Webb 2013; O'Neill and Debaille 2014; Gerya 2014). Following the condensation of the accreting hot Earth by

~4.57 Gyr, the planet may have witnessed further rapid increase in its internal heat ascribed to core-formation, radioactive decay of long-lived elements (U, Th, ^{40}K) and short-lived isotopes (e.g., ^{26}Al), bolide impacts, and very early impact of a Mars-sized body and such other factors (Wetherill 1976; Cameron and Benz 1991; Kleine et al. 2004; Elkins-Tanton 2008). As illustrated in Fig. 1 (stage I), this phase was succeeded by the formation of a near-surface magma mush ocean, several 100 km thick (Abe 1997; Davies 1999; Boyet and Carlson 2005), prior to 4.4 Gyr.

Early studies on the Jack Hills Hadean detrital zircon grains suggested their origin in felsic crust at ~4.4 Gyr and interaction with liquid water (Wilde et al. 2001; Valley et al. 2002). Thermodynamic considerations of mineral inclusions within Hadean zircons indicated conditions similar to modern subduction zones for the felsic hosts of the early Hadean zircons (Hopkins et al. 2008). Paleomagnetic considerations based on paleo-intensity of Hadean Jack hills zircon grains support plate tectonics prior to 4.1 Gyr (Tarduno et al. 2015). On the contrary, re-evaluation of the Jack Hills zircon U-Pb age-Hf and O isotopic data suggested the possibility of existence of a long lived mafic protolith for the felsic melts in which the zircons formed (Kemp et al. 2010, 2015). These isotopic systematics of the >3.8 Gyr zircons indicate that the early enriched mafic proto-crust may have resided deep beneath the surface, much similar to the highly enriched silicate reservoir, the UrKREEP on the Moon, which eventually contributed to the generation of KREEP-rich lunar igneous rocks with ages between 4.4 and 3.9 Gyr (reviewed by Iizuka et al. 2017). Such an interpretation supports the stagnant-lid tectonics (Debaille et al. 2013), a single lid with no relative motion for the early Earth. Debaille et al. (2013) infer a dominantly stagnant-lid regime for about ~700 million years (between 4.5 and 3.8 Gyr). Thus, combined interpretation of zircon isotopic data and numerical models suggests the relevance of dominantly stagnant-lid with episodic transient subduction-like events for the Hadean crust.

The Archean-Proterozoic Eons

Significant cooling of the Earth's crust (below the boiling point of water) by ~4.0 Gyr is inferred from preservation of the ~3.9 Gyr Isua volcano-sedimentary assemblages, western Greenland, which suggest the possibility that water-oceans marked the planet's outer rind by the Late Hadean times (Fig. 1, stage II a, b). No record of the Hadean crust older than the 4.03 Gyr component of the Acasta gneiss (Slave craton, Canada) (e.g., Bowring and Williams 1999; Mojzsis et al. 2014) is preserved and Eoarchean crustal remnants are scarce. Archean cratons consist predominantly of crust generated in the Neoproterozoic, while Paleoproterozoic to Mesoproterozoic crust is documented from few cratons (e.g., Nutman et al. 2002; Bleeker 2003; Van Kranendonk et al. 2007; Wu et al. 2008; Zeh et al. 2011; Arndt 2013; Turner et al. 2014; Wan et al. 2015; Komiya et al. 2015; Santosh et al. 2016; Mueller and Nutman 2017; Jayananda et al. 2018; Sreenivas et al. 2019; Bhaskar Rao et al. 2020). Since the 1960s, Archean geologists recognize that the large-scale granite-greenstone structures in the Archean cratons are divisible into two broad types: the dome and keel type and the orogenic linear belt type and the geodynamic models comprise two broad scenarios: "vertical" and "horizontal" tectonic regimes (Windley 1984; de Wit and Ashwal 1995 for a review). The former has been generally ascribed to mantle upwelling

(mantle plumes) and in some regions models invoking gravity driven Rayleigh-Taylor type instabilities and density inversions leading to sinking of dense volcanic-rich greenstone sequences and rising of felsic diapirs from the underlying crust producing dome and keel structures were proposed, typified by examples in Rhodesian and Pilbara cratons (Bédard 2006; Van Kranendonk et al. 2007; Wiemer et al. 2018). Horizontal tectonics in the Archean context is broadly analogous to Phanerozoic Wilson-type plate tectonics, but invokes subtle differences mainly due to higher mantle potential temperature and crustal heat flow during the Archean (Windley 1984; de Wit and Ashwal 1995; Kerrich and Polat 2006; Dewey 2007; Condie and Pease 2008; Ernst 2009). However, these "end-member" models are neither adequate nor mutually exclusive and a variety of complex mixed-mode scenarios involving, for example, multiple subduction zones and intermittent mantle plumes have been proposed (e.g., Wyman et al. 2002; Bédard 2006 and references therein).

It is now widely appreciated that as much as 73% of the present volume of Earth's continental crust may have been extracted from the mantle by the end of the Archean (Belousova et al. 2010; Dhuime et al. 2012; Cawood et al. 2013) and major peaks in the accretion of Archean juvenile crust centered around 3.0, 2.7 and 1.9 Gyr (Condie et al. 2011; Voice et al. 2011). Many researchers favor a view that the late Paleoproterozoic/early Mesoproterozoic (ca. 3.2–3.0 Gyr) marked a transition in the planet's geodynamic regime from a dominantly vertical (driven by mantle upwelling) towards dominantly horizontal (some form of plate tectonics) geodynamic regime (summarized by Hawkesworth et al. 2017; Condie 2018). This view is increasingly in favor in the light of the convergence of many different datasets including field geological observations, paleomagnetic, geochemical, ore deposit studies, and global-scale zircon U-Pb age-Hf±O datasets (Dhuime et al. 2012, 2018; Cawood et al. 2013; Tang et al. 2016; Hawkesworth et al. 2017 and references therein). The numerous geological and lithotectonic changes documented during this interval are linked directly or indirectly to the cooling of Earth's mantle and corresponding changes in convective style and the strength of the lithosphere. These data suggest a distinct change in the character, composition, and thickness of the continental lithosphere around ~3.2–3.0 Gyr. Some of the important changes include: (1) increasing proportion of basalts with "arc-like" mantle sources and the preponderance of K-rich granites; (2) the development of a stable continental lithosphere (Carlson et al. 2005; Condie and Kroner 2008; Moyen and Martin 2012); (3) the onset of a Great Thermal Divergence in the mantle (Condie et al. 2016); (4) appearance of eclogite inclusions in diamonds (Shirey and Richardson 2011); (5) an increase in the volume of crust generation marked by a change in the slope of the continental growth curve at ~3.0 Gyr estimated from the zircon age-Hf data (Belousova et al.

Plate Tectonics, Precambrian, Table 1 Earth’s geodynamic transition towards plate tectonics. (Summarized from Hawkesworth et al. 2017)

Stage-1 (the early ~15 Myr)	Initial accretion of Earth, giant impactors, core formation, radioactive decay of long and short-lived nuclides, formation and partial crystallization of deep mantle silicate magma ocean, development of an undifferentiated mafic protocrust
Stage-2 (4.5–3.0 Gyr)	Generation of crust in a pre-plate tectonic regime with episodic short lived plate tectonics alternating with nonplate tectonic mechanisms. Little or no lithospheric record for the first 600 Myr until the termination of the late heavy bombardment event around 3.9 Gyr. Potential mantle temperatures ~250 °C hotter than present day upper mantle and asthenospheric temperatures ~1350 °C at around 3.9 Gyr, which imply reduction in lithospheric rigidity inhibiting deep level lithospheric subduction. Generation of lithosphere in zones of mantle upwelling and melting, accompanied by lithospheric delamination. Start of oxygenation of atmosphere by ~3.5 Gyr
Stage-3 (3.0–1.7 Gyr)	Cooling of mantle to an extent enabling subduction of semi rigid plates. Increased generation of differentiated juvenile crust through hot subduction at relatively high geothermal gradients. Stabilization of pre 3.2 Gyr cratons with a change in crustal growth rates around 3.0 Gyr
Stage-4 (1.7–0.75 Gyr)	Significant increase in the atmospheric oxygen levels during the early Proterozoic, “the Great Oxidation Event,” along with significant increase in the volume and thickness of continental lithosphere, change to more felsic continental crust and its widespread exposure. An increase in geologic record of continental collisions, assembly and break-up of Supercontinents; Nuna and Rodinia
Stage-5 (<0.75 Gyr)	Initiation of modern cold subduction at relatively low geothermal gradients, common appearance of high- to ultrahigh-pressure metamorphic rocks. The episodic development and break-up events related to Supercontinents Gondwana and Pangea. Increased in the destruction (recycling) of continental crust. Increase in the oxygen levels of atmosphere and deep oceans, widespread development of phosphate and evaporates promoting metazoan evolution

2010; Dhuime et al. 2012, 2018); (6) a progressive increase in the appearance of collisional orogens following 3.0 Gyr, as well as ophiolites and global large igneous provinces in the geologic record (Condie 2018). On the other hand, as stated earlier many authors continue to opine that truly modern-style subduction was initiated around the Neoproterozoic (~1.0 Gyr, Fig. 1, stage III) (also see Stern 2018). Hawkesworth et al. (2017) summarized the temporal changes in the tectonic processes controlling the evolution of Earth’s lithosphere in a model comprising 5 stages (Table 1). The boundary between each stage corresponds to a change in the character of the lithosphere and dominant tectonic process.

Summary

Plate tectonics is a characteristic geodynamic feature of the Phanerozoic Earth. However, its onset on the planet remains controversial with estimates ranging from the Hadean to the Neoproterozoic. There is an increasing consensus for the view that the change from a nonplate tectonic Earth to one in which plate tectonics dominated was not abrupt and may have extended over hundreds of millions of years about 3.0 Gyr. Several changes in the behavior and character of the lithosphere around 3.0 Gyr are interpreted in terms of a geodynamic transition from a nonplate tectonic mode to the birth of plates and followed by sustained plate tectonics through the Proterozoic Era. The dynamics of such a transition are believed to be distinctly nonlinear. The predominant tectonic regime in the Hadean was stagnant-lid convection, with intermittent recycling events implying multiple episodic subduction initiation events with progressively increasing preservation potential of the manifestation of plate tectonic processes.

Cross-References

- ▶ Absolute Age Determinations: Radiometric
- ▶ Continental Crustal Structure
- ▶ Continental Drift
- ▶ Earth’s Structure, Core
- ▶ Earth’s Structure, Global
- ▶ Geodynamics
- ▶ Lithosphere, Continental
- ▶ Mantle Convection
- ▶ Mantle Plumes
- ▶ Ocean, Spreading Center
- ▶ Orogenic Belts
- ▶ Plate-Driving Forces
- ▶ Sedimentary Basins
- ▶ Seismic Structure at Mid-Ocean Ridges
- ▶ Subduction Zones

Bibliography

Abbott D, Drury R, Smith WHF (1994) Flat to steep transition in subduction style. *Geology* 22:937–940

Abe Y (1997) Thermal and chemical evolution of the terrestrial magma ocean. *Phys Earth Planet Inter* 100:27–39

Anderson DL (2001) Top-down tectonics? *Science* 293:2016–2018

Arndt NT (2013) Formation and evolution of the continental crust. *Geochem Perspect* 2(3):405–533

Bédard JH (2006) A catalytic delamination-driven model for coupled genesis of Archaean crust and sub-continental lithospheric mantle. *Geochim Cosmochim Acta* 70:1188–1214

Bédard JH (2013) How many arcs can dance on the head of a plume? A ‘comment’ on: a critical assessment of Neoarchean ‘plume only’ geodynamics: evidence from the Superior province, by Derek Wyman, PR, 2013. *Precambrian Res* 229:189–197

Bédard JH (2018) Stagnant lids and mantle overturns: implications for Archaean tectonics, magma genesis, crustal growth, mantle evolution, and the start of plate tectonics. *Geosci Front* 9:19–49

- Belousova EA, Kostitsyn YA, Griffin WL, Begg GC, O'Reilly SY, Pearson NJ (2010) The growth of the continental crust: constraints from zircon Hf isotope data. *Lithos* 119:457–466
- Bhaskar Rao YJ, Vijaya Kumar T, Sreenivas B, Babu EVSSK (2020) A review of Paleo- to Neoproterozoic crustal evolution in the Dharwar craton, Southern India and the transition towards a plate tectonic regime. *Episodes* v. 43(1):51–68
- Bickle MJ, Nisbet EG, Martin A (1994) Archean greenstone belts are not oceanic crust. *J Geol* 102:121–138
- Bleeker W (2003) The late Archean record: a puzzle in ca. 35 pieces. *Lithos* 71(2–4):99–134
- Bowring SA, Williams IS (1999) Proterozoic (4.00–4.03 Ga) orthogneisses from northwestern Canada. *Contrib Mineral Petrol* 134:3–16
- Boyett M, Carlson RW (2005) ^{142}Nd evidence for early (>4.53) global differentiation of the silicate Earth. *Science* 309:576–589
- Brown M (2008) Characteristic thermal regimes of plate tectonics and their metamorphic imprint through Earth history. In: Condie KC, Pease V (eds) When did plate tectonics begin on planet Earth? Geological Society of America special paper 440. Geological Society of America, Boulder, pp 97–128
- Brown M (2010) Paired metamorphic belts revisited. *Gondwana Res* 18:46–59
- Calvert AJ, Ludden JN (1999) Archean continental assembly in the southeastern Superior Province of Canada. *Tectonics* 18:412–429
- Calvert AJ, Sawyer EW, Davis WJ (1995) Archean subduction inferred from seismic images of a mantle suture in the Superior Province. *Nature* 375(6533):670–674
- Cameron AGW, Benz W (1991) Origin of the Moon and the single impact hypothesis IV. *Icarus* 92:204–216
- Carlson RW, Pearson DG, James DE (2005) Physical, chemical, and chronological characteristics of continental mantle. *Rev Geophys* 43(1):RG1001
- Carswell DA (1990) Eclogites and the eclogite facies; definitions and classification. In: Carswell DA (ed) Eclogite facies rocks. Blackie, Glasgow, pp 1–13
- Cawood PA, Buchan C (2007) Linking accretionary orogenesis with supercontinent assembly. *Earth Sci Rev* 82:217–256
- Cawood PA, Hawkesworth CJ, Pisarevsky SA, Dhuime B, Capitanio FA, Nebel O (2018) Geological archive of the onset of plate tectonics. *Phil Trans R Soc A* 376:20170405. <https://doi.org/10.1098/rsta.2017.0405>
- Cawood PA, Kroner A, Pisarevsky S (2006) Precambrian plate tectonics: criteria and evidence. *GSA Today* 16:4–11
- Cawood PA, Kröner A, Collins WJ, Kusky TM, Mooney WD, Windley BF (2009) Accretionary orogens through Earth history. In: Cawood PA, Kröner A (eds) Earth accretionary systems in space and time. Geological Society of London special publication 318. Geological Society of London, London, pp 1–36
- Cawood PA, Hawkesworth CJ, Dhuime B (2013) The continental record and the generation of continental crust. *Geol Soc Am Bull* 125(1–2):14–32
- Coleman RG, Lee DE, Beatty LB, Brannock WE (1965) Eclogites and eclogites: their differences and similarities. *GSA Bull* 76(5):438–508
- Collins WJ, Belousova EA, Kemp AIS, Murphy JB (2011) Two contrasting Phanerozoic orogenic systems revealed by hafnium isotope data. *Nat Geosci* 4:333–337
- Condie KC (2015) Changing tectonic settings through time: indiscriminate use of geochemical discriminant diagrams. *Precambrian Res* 266:587–559
- Condie KC (2018) A planet in transition: the onset of plate tectonics on Earth between 3 and 2 Ga? *Geosci Front* 9:51–60
- Condie KC, Kroner A (2008) When did plate tectonics begin? Evidence from the geologic record. In: Condie KC, Pease V (eds) When did plate tectonics begin on planet Earth? Geological Society of America special paper 440. Geological Society of America, Boulder, pp 281–294
- Condie KC, Pease V (eds) (2008) When did plate tectonics begin on planet Earth? Geological Society of America special paper 440. Geological Society of America, Boulder
- Condie KC, Bickford ME, Aster RC, Belousova E, Scholl DW (2011) Episodic zircon ages, Hf isotopic composition, and the preservation rate of continental crust. *Geol Soc Am Bull* 123:951–957. <https://doi.org/10.1130/B30344.1>
- Condie KC, Aster RC, van Hunen J (2016) A great thermal divergence in the mantle beginning 2.5 Ga: geochemical constraints from greenstone basalts and komatiites. *Geosci Front* 7:543–553
- Conrad CD, Lithgow-Bertelloni C (2004) The temporal evolution of plate driving forces: importance of 'slab suction' versus 'slab pull' during the Cenozoic. *J Geophys Res* 109:B10407
- Davies GF (1992) On the emergence of plate tectonics. *Geology* 20:963–966
- Davies GF (1999) Dynamic Earth: plate, plumes, and mantle convection. Cambridge University Press, New York
- de Wit MJ, Ashwal LD (1995) Greenstone belts. Oxford monograph on geology and geophysics 35. Clarendon Press, Oxford, p 809
- de Wit MJ, Tinker J (2004) Crustal structures across the central Kaapvaal craton from deep-seismic reflection data. *S Afr J Geol* 107:185–206
- de Wit MJ, Furnes H, MacLennan S et al (2018) Paleoproterozoic bedrock lithologies across the Makhonjwa Mountains of South Africa and Swaziland linked to geochemical, magnetic and tectonic data reveal early plate tectonic genes flanking subduction margins. *Geosci Front* 9(3):603–665. <https://doi.org/10.1016/j.gsf.2017.10.005>
- Debaille V, O'Neill C, Brandon AD, Haenecour P, Yin Q-Z, Mattielli N, Treiman AH (2013) Stagnant-lid tectonics in early Earth revealed by ^{142}Nd variations in late Archean rocks. *Earth Planet Sci Lett* 373:83–92
- Dewey JF (2007) The secular evolution of plate tectonics and the continental crust: an outline. In: Hatcher RD Jr, Carlson MP, McBride JH, Catalan M (eds) 4-D framework of continental crust. Geological Society of America memoir 200. Geological Society of America, Boulder, pp 1–7
- Dhuime B, Hawkesworth CJ, Cawood PA, Storey CD (2012) A change in the geodynamics of continental growth 3 billion years ago. *Science* 80–335:1334–1336
- Dhuime B, Hawkesworth CJ, Delavault H, Cawood PA (2018) Rates of generation and destruction of the continental crust: implications for continental growth. *Philos Trans R Soc A* 376:20170403. <https://doi.org/10.1098/rsta.2017.0403>
- Dilek Y, Polat A (2008) Suprasubduction zone ophiolites and Archean tectonics. *Geology* 36:431–432
- Drummond BJ, Goleby BR, Swager CP (2000) Crustal signature of Late Archean tectonic episodes in the Yilgarn craton, Western Australia: evidence from deep seismic sounding. *Tectonophysics* 329:193–221
- Elkins-Tanton LT (2008) Linked magma ocean solidification and atmospheric growth for Earth and Mars. *Earth Planet Sci Lett* 271(1–4):181–191
- Ernst WG (2003) High-pressure and ultrahigh-pressure metamorphic belts-subduction, recrystallization, exhumation, and significance for ophiolite studies. In: Dilek Y, Newcombe S (eds) Ophiolite concept and evolution of geological thought. Geological Society of America special paper 373. Geological Society of America, Boulder, pp 365–384
- Ernst WG (2007) Speculations on evolution of the terrestrial lithosphere-aesthenosphere system-plumes and plates. *Gondwana Res* 11:38–49
- Ernst WG (2009) Archean plate tectonics, rise of Proterozoic supercontinentality and onset of regional, episodic stagnant-lid behaviour. *Gondwana Res* 15:243–253
- Evans DAD, Pisarevsky SA (2008) Plate tectonics on early Earth? Weighing the paleomagnetic evidence. In: Condie KC, Pease V (eds) When did plate tectonics begin on planet Earth? Geological Society of America special paper 440. Geological Society of America, Boulder, pp 249–263. [https://doi.org/10.1130/2008.2440\(12\)](https://doi.org/10.1130/2008.2440(12))

- Foley SF, Buhre S, Jacob DE (2003) Evolution of the Archean crust by shallow subduction and recycling. *Nature* 421:249–252
- Forsyth D, Uyeda S (1975) On the relative importance of the driving forces of plate motion. *Geophys J R Astron Soc* 43:163–200
- Furnes H, de Witt M, Staudigel H, Rosing M, Muehlenbachs K (2007) A vestige of Earth's oldest ophiolite. *Science* 315:1704–1707
- Furnes H, de Wit M, Dilek Y (2014) Four billion years of ophiolites reveal secular trends in oceanic crust formation. *Geosci Front* 5(4):571–603. <https://doi.org/10.1016/j.gsf.2014.02.002>
- Gerya T (2014) Precambrian geodynamics: concepts and models. *Gondwana Res* 25:442–463
- Gerya T, Stern R, Baes M, Sobolev S, Whattam S (2015) Plate tectonics on the Earth triggered by plume-induced subduction initiation. *Nature* 527:221–225
- Goleby BR, Blewett RS, Rorsch RJ, Champion DC, Cassidy KF, Jones LEA, Groenewald PB, Henson P (2004) Deep seismic reflection profiling in the Archean northeastern Yilgarn craton, Western Australia: implications for crustal architecture and mineral potential. *Tectonophysics* 388:119–133. <https://doi.org/10.1016/j.tecto.2004.04.032>
- Griffin WL, Belousova EA, O'Neill C, O'Reilly SY, Malkovets V, Pearson NJ (2014) The world turns over: Hadean-Archean crust-mantle evolution. *Lithos* 189:2–15
- Hamilton WB (1998) Archean magmatism and deformation were not products of plate tectonics. *Precambrian Res* 91:143–179
- Hamilton WB (2007) Earth's first two billion years—the era of internally mobile crust. In: Hatcher RD Jr, Carlson MP, McBride JH, Martinez JR (eds) 4-D framework of continental crust, Geological Society of America memoir 200. Geological Society of America, Boulder, pp 233–296
- Hamilton WB (2011) Plate tectonics began in Neoproterozoic time and plumes from deep mantle have never operated. *Lithos* 123:1–20
- Hamilton WB, Myers WB (1967) The nature of batholiths. *US Geological Survey* 554-C. pp C1–C30
- Harris LB, Bédard JH (2014) Crustal evolution and deformation in a non plate-tectonic Archean Earth: comparisons with Venus. In: Dilek Y, Furnes H (eds) Evolution of Archean crust and early life. Modern approaches in solid Earth sciences 7. Springer, Dordrecht, pp 215–291
- Harrison TM (2009) The Hadean crust: evidence from >4 Ga zircons. *Annu Rev Earth Planet Sci* 37:479–505
- Harrison TM, Bell EA, Boehnke P (2017) Hadean zircon petrochronology. *Rev Mineral Geochem* 83(1):329–363
- Hawkesworth CJ, Cawood PA, Dhuime B, Kemp AIS (2017) Earth's continental lithosphere through time. *Annu Rev Earth Planet Sci* 45:169–198
- Herzberg C, Condie K, Korenaga J (2010) Thermal history of the Earth and its petrological expression. *Earth Planet Sci Lett* 292:79–88. <https://doi.org/10.1016/j.epsl.2010.01.022>
- Hill RI, Chappell BW, Campbell IH (1992) Late Archean granites of the southeastern Yilgarn Block, Western Australia: age, geochemistry, and origin. *Trans R Soc Edinb Earth Sci* 83:211–226
- Hoffmann JE, Nagel TJ, Munker C, Naeraa T, Rosing MT (2014) Constraining the process of Eoarchean TTG formation in the Itsaq Gneiss Complex, southern West Greenland. *Earth Planet Sci Lett* 388:374–386
- Hopkins M, Harrison TM, Manning CE (2008) Low heat flow inferred from >4 Gyr zircons suggests Hadean plate boundary interactions. *Nature* 456(7221):493–496
- Hynes A (2005) Buoyancy of the oceanic lithosphere and subduction initiation. *Int Geol Rev* 47:938–951
- Iizuka T, Yamaguchi T, Itano K, Suzuki K (2017) What Hf isotopes in zircon tell us about crust-mantle evolution. *Lithos* 274–275: 304–327
- Jahn BM, Caby R, Monie P (2001) The oldest UHP eclogites of the world: age of UHP metamorphism, nature of protoliths and tectonic implications. *Chem Geol* 178:143–158
- Jayananda M, Santosh M, Aadhisesan KR (2018) Formation of Archean (3600–2500 Ma) continental crust in the Dharwar Craton, southern India. *Earth Sci Rev* 181:12–42
- Keleman PB, Yogodzinski GM, Scholl DW (2004) In: Eiler J (ed) Inside the subduction factory. American Geophysical Union, Washington, DC, pp 223–276
- Kemp AIS, Wilde SA, Hawkesworth CJ, Coath CD, Nemchin A, Pidgeon RT, Vervoort JD, DuFrane SA (2010) Hadean crustal evolution revisited: new constraints from Pb-Hf isotope systematics of the Jack Hills zircons. *Earth Planet Sci Lett* 296:45–56
- Kemp AIS, Hickman AH, Kirkland CL, Vervoort JD (2015) Hf isotopes in detrital and inherited zircons of the Pilbara Craton provide no evidence for Hadean continents. *Precambrian Res* 261:112–126
- Kerrick R, Polat A (2006) Archean greenstone-tonalite duality: thermochemical mantle convection models or plate tectonics in the early Earth global geodynamics? *Tectonophysics* 415:141–165
- Kleine T, Mezger K, Palme H, Munker C (2004) The W isotope evolution of the bulk silicate Earth – constraints on the timing and mechanisms of core formation and accretion. *Earth Planet Sci Lett* 228:109–123
- Komiya T, Yamamoto S, Aoki S (2015) Geology of the Eoarchean, >3.95 Ga, Nulliak supracrustal rocks in the Saglek Block, Northern Labrador, Canada: the oldest geological evidence for plate tectonics. *Tectonophysics* 662:40–66. <https://doi.org/10.1016/j.tecto.2015.05.003>
- Korenaga J (2006) Archean geodynamics and the thermal evolution of Earth. In: Benn K, Mareschal J-C, Condie KC (eds) Archean geodynamic processes. American Geophysical Union geophysical monograph 164. American Geophysical Union, Washington, DC, pp 7–32
- Korenaga J (2013) Initiation and evolution of plate tectonics on Earth: theories and observations. *Annu Rev Earth Planet Sci* 41(1):117–151
- Korenaga J (2018) Crustal evolution and mantle dynamics through Earth history. *Phil Trans R Soc A* 376:20170408. <https://doi.org/10.1098/rsta.2017.0408>
- Kusky TM, Windley BF, Polat A (2018) Geological evidence for the operation of plate tectonics throughout the Archean: records from Archean paleo-plate boundaries. *J Earth Sci* 29(6):1291–1303. <https://doi.org/10.1007/s12583-018-0999-6>. <http://en.earth-science.net>
- Labrosse S, Jaupart C (2007) Thermal evolution of the Earth: secular changes and fluctuations of plate characteristics. *Earth Planet Sci Lett* 260:465–481
- Laurent O, Martin H, Moya J-F, Doucelance R (2014) The diversity and evolution of late-Archean granitoids: evidence for the onset of “modern-style” plate tectonics between 3.0 and 2.5 Ga. *Lithos* 205:208–235
- Lay T, Hernlund J, Buffett BA (2008) Core-mantle boundary heat flow. *Nat Geosci* 1(1):25
- Lenardic A (2018) The diversity of tectonic modes and thoughts about transitions between them. *Philos Trans R Soc A* 376:20170416. <https://doi.org/10.1098/rsta.2017.0416>
- Liou JG, Tsujimori T, Zhang RY, Katayama I, Maruyama S (2004) Global UHP metamorphism and continental subduction/collision: the Himalayan model. *Int Geol Rev* 46:1–27
- Mandal B, Vijaya Rao V, Sarkar D, Bhaskar Rao YJ, Raju S, Karuppannan P, Sen MK (2018) Deep crustal seismic reflection images from the Dharwar craton, Southern India – evidence for the Neoproterozoic subduction. *Geophys J Int* 212:777–794
- Martin H (1993) The mechanisms of petrogenesis of the Archean continental crust – comparison with modern processes. *Lithos* 30:373–388
- Maruyama S, Liou JG (1998) Ultrahigh-pressure metamorphism and its significance on the Proterozoic-Phanerozoic boundary. *Island Arc* 7:6–35
- Maruyama S, Liou JG, Terabayashi M (1996) Blueschists and eclogites of the world and their exhumation. *Int Geol Rev* 38:485–594
- McElhinny MW, Senanayake WE (1980) Paleomagnetic evidence for the existence of the geomagnetic field 3.5 Ga ago. *J Geophys Res* 85:3523–3528

- Miyashiro A (1972) Metamorphism and related magmatism in plate tectonics. *Am J Sci* 272:629–656
- Mojzsis SJ, Cates NL, Caro G, Trail D, Abramov O, Guitreau M, Blichert-Toft J, Hopkins MD, Bleeker W (2014) Component geochronology in the polyphase ca. 3920 Ma Acasta Gneiss. *Geochim Cosmochim Acta* 133:68–96
- Moore WB, Webb AG (2013) Heat-pipe Earth. *Nature* 501:501–505
- Morgan WJ (1971) Convection plumes in the lower mantle. *Nature* 230:42–43
- Moyen J-F, Martin H (2012) Forty years of TTG research. *Lithos* 148:312–336
- Mueller PA, Nutman AP (2017) The Archean–Hadean Earth: Modern paradigms and ancient processes, in Bickford, M.E., ed., *The Web of Geological Sciences: Advances, Impacts, and Interactions II: Geological Society of America Special Paper 523:75–237*, [https://doi.org/10.1130/2017.2523\(03\)](https://doi.org/10.1130/2017.2523(03))
- Næraa T, Schersten A, Rosing MT, Kemp AIS, Hoffmann JE, Kokfelt TF, Whitehouse J (2012) Hafnium isotope evidence for a transition in the dynamics of continental growth 3.2 Gyr ago. *Nature* 485(7400):627–630
- Nance RD, Murphy JB, Santosh M (2014) The supercontinent cycle: a retrospective essay. *Gondwana Res* 25(1):4–29
- Nutman AP, Friend CRL, Bennett VC (2002) Evidence for 3650–3600 Ma assembly of the northern end of the Itsaq Gneiss Complex, Greenland: implication for early Archean tectonics. *Tectonics* 21. <https://doi.org/10.1029/2000TC001203>
- O'Neill C, Debaille V (2014) The evolution of Hadean–Eoarchean geodynamics. *Earth Planet Sci Lett* 406:49–58
- O'Neill C, Marchi S, Zhang S, Bottke W (2017) Impact-driven subduction on the Hadean Earth. *Nat Geosci* 10:793–797
- O'Neill C, Turner S, Rushmer T (2018) The inception of plate tectonics: a record of failure. *Philos Trans R Soc A* 376:20170414. <https://doi.org/10.1098/rsta.2017.0414>
- Oxburgh ER, Turcotte DL (1971) Origin of paired metamorphic belts and crustal dilation in island arc regions. *J Geophys Res* 76:1315–1327
- Palin RM, White RW (2016) Emergence of blueschists on Earth linked to secular changes in oceanic crust composition. *Nat Geosci* 9:60–64
- Parman SW, Grove TL, Dann JC (2001) The production of Barberton komatiites in an Archean subduction zone. *Geophys Res Lett* 28:2513–2516
- Pearce JA (2008) Geochemical fingerprinting of oceanic basalts with applications to ophiolite classification and the search for Archean oceanic crust. *Lithos* 100:14–48
- Pearce JA, Peate DW (1995) Tectonic implications of the composition of volcanic arc magmas. *Annu Rev Earth Planet Sci* 23:251–285
- Peltonen P, Kontinen A, Huhma H (1996) Petrology and geochemistry of metabasalts from the 1.95 Ga Jormua ophiolite, northeastern Finland. *J Petrol* 37:1359–1383
- Penrose conference participants (1972) Ophiolites: *Geotimes*, 17 (12):24–25
- Pesonen LJ, Elming S-A, Metanen S, Pisarevsky SA, D'Agrella-Filho MS, Meert J, Schmidt PW, Abrahamsen N, Bulund G (2003) Paleomagnetic configuration of continents during the Proterozoic. *Tectonophysics* 375:289–324
- Rapp R, Watson EB, Miller CF (1991) Partial melting of amphibolite/eclogite and the origin of Archean trondhjemites and tonalites. *Precambrian Res* 51:1–25
- Roberts NMW, Spencer CJ (2015) The zircon archive of continent formation through time. In: Roberts NMW, Van Kranendonk MP, Shirey S, Clift PD (eds) *Continent formation through time*. Geological Society, special publications, 389. Geological Society, London, pp 197–225
- Rollinson H (1997) Eclogitic xenoliths in west African kimberlites as residues from Archean granitoid crust formation. *Nature* 389: 173–176
- Rozel AB, Golabek GJ, Jain C, Tackley PJ, Gerya T (2017) Continental crust formation on early Earth controlled by intrusive magmatism. *Nature* 545:332–335
- Rudnick RL, Gao S (2014) *Composition of the continental crust*. Treatise on geochemistry, 2nd edn. Elsevier, pp 1–51
- Santosh M (2010) A synopsis of recent conceptual models on supercontinent tectonics in relation to mantle dynamics, life evolution and surface environment. *J Geodyn* 50:116–133
- Santosh M, Yang QY, Shaji E, Mohan MR, Tsunogae T, Satyanarayanan M (2016) Oldest rocks from Peninsular India: evidence for Hadean to Neoproterozoic crustal evolution. *Gondwana Res* 29:105–135
- Scott DJ, Helmstaedt H, Bickle MJ (1992) Purtuniq Ophiolite, Cape Smith Belt, northern Quebec, Canada: a reconstructed section of early Proterozoic oceanic crust. *Geology* 20:173–176
- Shirey SB, Richardson SH (2011) Start of the Wilson Cycle at 3 Ga shown by diamonds from subcontinental mantle. *Science* 333:434–436
- Shirey SB, Kamber BS, Whitehouse MJ, Mueller PA, Basu AR (2008) A review of the isotopic and trace element evidence for mantle and crustal processes in the Hadean and Archean: implications for the onset of plate tectonic subduction. In: Condie KC, Pease V (eds) *When did plate tectonics begin on planet Earth?* Geological Society of America special paper 440. Geological Society of America, Boulder, pp 1–29
- Sizova E, Gerya T, Brown M, Perchuk LL (2010) Subduction styles in the Precambrian: insight from numerical experiments. *Lithos* 116:209–229
- Sizova E, Gerya T, Brown M (2014) Contrasting styles of Phanerozoic and Precambrian continental collision. *Gondwana Res* 25:522–545
- Sizova E, Gerya T, Stuwe K, Brown M (2015) Generation of felsic crust in the Archean: a geodynamic modelling perspective. *Precambrian Res* 271:198–224
- Sleep NH (2000) Evolution of the mode of convection within terrestrial planets. *J Geophys Res* 105:17563–17578
- Sleep NH (2005) Evolution of the continental lithosphere. *Annu Rev Earth Planet Sci* 33:369–393
- Smithies RH, Champion DC, Cassidy KF (2003) Formation of Earth's early Archean continental crust. *Precambrian Res* 127:89–101
- Smithies RH, Van Kranendonk MJ, Champion DC (2005) It started with a plume – early Archean basaltic proto-continental crust. *Earth Planet Sci Lett* 238:284–297
- Sreenivas B, Sukanta D, Bhaskar Rao YJ, Vijaya Kumar T, Babu EVSSK, Williams I (2019) A new cache of Eoarchean detrital zircons from the Singhbhum craton, Eastern India and constraints on early Earth geodynamics. *Geosci Front* 10:1359–1370. <https://doi.org/10.1016/j.gsf.2019.02.001>
- Stern RJ (2002) Subduction zones. *Rev Geophys* 40:1012
- Stern RJ (2005) Evidence from ophiolites, blueschists, and ultrahigh-pressure metamorphic terranes that the modern episode of subduction tectonics began in Neoproterozoic time. *Geology* 33:557–560
- Stern RJ (2007) When did plate tectonics begin on Earth? Theoretical and empirical constraints. *Chin Bull Sci* 52(5):578–591
- Stern RJ (2008) Modern-style plate tectonics began in Neoproterozoic time: an alternate interpretation of Earth dynamics and history. In: Condie KC, Pease V (eds) *When did plate tectonics begin on planet Earth?* Geological society of America special paper 440. Geological Society of America, Boulder, pp 265–280
- Stern RJ (2018) The evolution of plate tectonics. *Phil Trans R Soc A* 376:20170406. <https://doi.org/10.1098/rsta.2017.0406>
- Stevenson DJ (2003) Styles of mantle convection and their influence on planetary evolution. *C R Geosci* 335:99–111
- Tang M, Chen K, Rudnick RL (2016) Archean upper crust transition from mafic to felsic marks the onset of plate tectonics. *Science* 351:372–375

- Tarduno JA, Cottrell RD, Davis WJ, Nimmo F, Bono RK (2015) A Hadean to Paleoproterozoic geodynamo recorded by single zircon crystals. *Science* 349:521–524
- Tatsumi Y (2005) The subduction factory: how it operates in the evolving Earth. *GSA Today* 15(7):4–10
- Taylor SR, McLennan SM (1985) The continental crust: its composition and evolution. Blackwell, Oxford. 312 p
- Touboul M, Puchtel IS, Walker RJ (2012) 182W evidence for long-term resorption of early mantle differentiation products. *Science* 335:1065–1069
- Turner S, Rushmer T, Reagan M, Moya JF (2014) Heading down early on? Start of subduction on Earth. *Geology* 42(2):139–142
- Valley JW, Peck WH, King EM, Wilde SA (2002) A cool early Earth. *Geology* 30:351–354
- Van der Hilst RD, Widiantoro S, Engdahl ER (1997) Evidence for deep mantle circulation from global tomography. *Nature* 228:657–659
- van der Velden AJ, Cook FA (2005) Relict subduction zones in Canada. *J Geophys Res* 110:B08403. <https://doi.org/10.1029/2004JB003333>
- Van Keken PE, Kiefer B, Peacock SM (2002) High-resolution models of subduction zones: implications for mineral dehydration reactions and the transport of water into the deep mantle. *Geochem Geophys Geosyst* 3(10):1056
- Van Kranendonk MJ (2010) Two types of Archean continental crust: plume and plate tectonics on early Earth. *Am J Sci* 310:1187–1209
- Van Kranendonk MJ, Smithies RH, Hickman AH, Champion DC (2007) Secular tectonic evolution of Archean continental crust: interplay between horizontal and vertical processes. *Terra Nova* 19:1–38
- Van Thienen P, Vlaar NJ, van den Berg AP (2004) Plate tectonics on the terrestrial planets. *Phys Earth Planet Inter* 142:61–74
- Vijaya Rao V, Murthy ASN, Sarkar D, Bhaskar Rao YJ, Khare P, Prasad ASSRS, Sridhar V, Raju S, Rao GSP, Karuppannan N, Kumar P, Sen MK (2015) Crustal velocity structure of the Neoproterozoic convergence zone between the eastern and western blocks of Dharwar Craton, India from seismic wide-angle studies. *Precambrian Res* 266:282–295
- Voice PJ, Kowalewski M, Eriksson KA (2011) Quantifying the timing and rate of crustal evolution: global compilation of radiometrically dated detrital zircon grains. *J Geol* 119(2):109–126
- Wan Y-S, Liu D-Y, Dong C-Y, Xie H-Q, Kröner A, Ma M-Z, Liu S-J, Xie S, Ren P (2015) Formation and evolution of Archean continental crust of the North China craton. In: Zhai M (ed) *Precambrian geology of China*. Springer, pp 59–135
- Wetherill GW (1976) The role of large bodies in the formation of the Earth and Moon. In: *Proceedings of the lunar science conference*, vol 7, pp 3245–3257
- Wiemer D, Schrank CE, Murphy DT, Wenham L, Allen CM (2018) Earth's oldest stable crust in the Pilbara craton formed by cyclic gravitational overturns. *Nat Geosci*. <https://doi.org/10.1038/s41561-018-0105-9>
- Wilde SA, Valley JW, Peck WH, Graham CM (2001) Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Ga ago. *Nature* 409:175–178
- Wilson JT (1968) Static or mobile Earth: the current scientific revolution. *Proc Am Philos Soc* 112:309
- Windley B (1984) *The evolving continents*, 2nd edn. Wiley, Chichester, 399 p
- Wu F-Y, Zhang Y-B, Yang J-H, Xie L-W, Yang Y-H (2008) Zircon U-Pb and Hf isotopic constraints on the early Archean crustal evolution in Anshan of the north China craton. *Precambrian Res* 167:339–362
- Wyman DA, Kerrich R, Polat A (2002) Assembly of Archean cratonic mantle lithosphere and crust: plume–arc interaction in the Abitibi–Wawa subduction-accretion complex. *Precambrian Res* 115:37–62
- Zeh A, Gerdes A, Millonig L (2011) Hafnium isotope record of the Ancient Gneiss Complex, Swaziland, southern Africa: evidence for Archean crustal mantle formation and crust reworking between 3.66 and 2.73 Gyr. *Geol Soc Lond* 168:953–963

Plate-Driving Forces

Alessandro M. Forte

Department of Geological Sciences, University of Florida, Gainesville, FL, USA

Definition

Plate-driving forces. The forces exerted on Earth's tectonic plates that produce their observable horizontal motions.

Introduction

Since the advent of the theory of plate tectonics in the 1960s, it has been postulated that the forces that drive the observed horizontal motions of Earth's tectonic plates are produced by the process of thermal convection in the mantle. The proposition that such motions could be produced and sustained over very long geological time scales in a rocky shell, comprising the lithosphere and underlying deep mantle, that behaves as an imperfect elastic solid through which seismic waves travel, was initially viewed as a serious obstacle to the acceptance of the mantle convection hypothesis of plate motions (e.g., Jeffreys 1972). This obstacle was eliminated with the recognition, amply confirmed by high-temperature deformation experiments on minerals (e.g., Evans and Kohlstedt 1995), that the rheology of the mantle allows steady-state creeping flow when subjected to long-term shearing stresses, provided the ambient temperature is sufficiently high (Poirier 1985).

The relationship between the mantle shear stress and deformation (measured in terms of strain rate) is characterized by an effective viscosity. Experimental and theoretical studies of the rheology of mantle minerals indicate that their effective viscosity is strongly dependent on pressure, temperature, and chemical composition (e.g., hydration). For sufficiently low ambient stress, mantle viscosity also depends on grain size, yielding linear (Newtonian) diffusion creep, whereas for high ambient stress, the viscosity depends on stress magnitude and is independent of grain size, yielding a nonlinear (power-law) dislocation creep (Karato and Wu 1993). The temperature dependence of mantle viscosity gives rise to a lithospheric layer that is stiffer and less deformable than the much hotter, less viscous upper mantle. Indeed, the very existence of tectonic plates, behaving as effectively rigid bodies, with little surface deformation is generally explained in terms of the strong temperature dependence of mantle viscosity. Nevertheless, plate-tectonic motions with observed speeds of several cm/yr could not exist unless localized rheological weakening mechanisms (e.g., plasticity, grain-size reduction, partial melting) also operate, yielding interconnected linear

zones of weakness (i.e., plate boundaries) that continuously evolve in response to changing flow and stress distributions in the lithosphere and underlying mantle (e.g., Moresi and Solomatov 1998; Tackley 2000; Ghosh and Holt 2012; Rolf et al. 2018; Coltice et al. 2019). A rheological weakening mechanism that has been extensively employed to simulate plate boundary creation and evolution is viscoplasticity (e.g., Moresi and Solomatov 1998; Karato and Barbot 2018). The effective viscosity, η_{eff} , representing the combined operation of diffusion/dislocation creep and viscoplastic weakening may be mathematically expressed as follows:

$$\eta_{eff} = \left(\eta_{cr}^{-1} + \eta_{pl}^{-1} \right)^{-1}, \quad \text{where } \eta_{pl} = \frac{\tau_y}{2\dot{\epsilon}_{II}} \quad (1)$$

in which η_{cr} is the Arrhenius temperature-dependent creep viscosity (e.g., Karato and Wu 1993) and η_{pl} is the viscoplastic viscosity that is defined in terms of a yield stress, τ_y , and the square root of the second invariant of the strain-rate tensor, $\dot{\epsilon}_{II}$ (e.g., Moresi and Solomatov 1998). Expression (1) implies that when η_{pl} is much smaller than η_{cr} , then the effective viscosity of the lithosphere will be locally dominated by viscoplastic weakening: $\eta_{eff} \approx \eta_{pl}$ (e.g., Spiegelman et al. 2016).

The strong rheological contrast between the lithosphere and the underlying mantle is often used as the basis for models of plate-driving forces that treat the lithospheric plates and their descending limbs (subducted slabs) under the oceanic trenches (Fig. 1a) separately from the mantle (Elsasser 1969; McKenzie 1969; Solomon and Sleep 1974; Harper 1975; Forsyth and Uyeda 1975; Richardson et al. 1976; Chapple and Tullis 1977; Richter 1977; Davies 1978; Bird et al. 2008). Alternatively, the oceanic lithosphere can be regarded as an integral part of the convecting mantle. In the latter case, the lithosphere is modelled as the upper thermal boundary layer of mantle convection, such that the heat transported across the deep mantle by rising and sinking convective flow is finally transmitted to the Earth's surface by vertical conduction in balance with horizontal advection of heat in the lithosphere (Turcotte and Oxburgh 1967; Jarvis and Peltier 1982). The vertical viscosity gradient across the lithospheric plates, from an effectively undeformable (high-viscosity) surface to a more deformable (low-viscosity) base, is therefore a consequence of the thermal boundary-layer structure of the lithosphere.

This boundary-layer model provides an alternative context for interpreting plate-driving forces, in which the horizontal motions of Earth's oceanic lithospheric plates (Fig. 1a) are directly coupled to, and driven by, the mantle convective circulation (Turcotte and Oxburgh 1967; Richter 1973; Jarvis and Peltier 1982; Davies 1988). The horizontal plate motions are thus understood to be coupled to the underlying mantle flow generated by descending and ascending thermal plumes

(Hager and O'Connell 1981; Forte and Peltier 1987, 1994; Ricard and Vigny 1989; Gable et al. 1991; Vigny et al. 1991; Lithgow-Bertelloni and Richards 1998; Becker and O'Connell 2001; Lowman et al. 2008; Forte et al. 2009).

Balance Between Buoyancy Forces and Viscous Dissipation

Two distinct approaches have therefore been adopted to develop models of plate-driving forces where, on the one hand, the plate-driving forces are analyzed separately from the dynamics of the underlying mantle and, on the other, the plate motions are modelled as an integral part of the convecting mantle. Both approaches must satisfy a fundamental constraint that expresses how work done by the body forces throughout the mantle is dissipated by the viscous deformation associated with the tectonic plate motions and flow in the underlying mantle.

A mathematical expression of the balance between buoyancy and dissipation (e.g., Malvern 1969) over the entire volume, V_m , of the mantle (including the lithosphere) is:

$$-\int_{V_m} u_r \delta \rho g_o dV = \int_{V_m} \tau_{ij} \epsilon_{ij} dV - \int_{S_d} t_i \Delta u_i dS \quad (2)$$

in which τ_{ij} and ϵ_{ij} are the deviatoric stress and strain-rate tensors, respectively, t_i is a traction vector acting on surfaces S_d across which there is a discontinuity in motion Δu_i , u_r is the vertical (radial) component of mantle flow, $\delta \rho$ are lateral perturbations in density, and g_o is the radial gravitational field. All indices i and j represent components in each of the three Cartesian coordinate directions, and repeated indices (in the terms on the right-hand side of Eq. 2) denote summation over all values (1–3) of the indices. The traction vector t_i acting on the discontinuity surface S_d depends on the deviatoric stress as follows: $t_i = \tau_{ij} n_j$, in which the n_j are the Cartesian components of the local normal vector on S_d . Although the motion along S_d is discontinuous, the traction t_i must be continuous by Newton's third law (equal and opposite actions and reactions).

The left-hand side of Eq. (2) is the (positive) rate of work done by the buoyancy forces $\delta \rho g_o$ where vertical flow occurs, as in hot plumes where $\delta \rho < 0$ & $u_r > 0$, and in cold descending slabs where $\delta \rho > 0$ & $u_r < 0$ (Fig. 1a). The first term on the right-hand side of Eq. (2) is the rate at which the energy released by the buoyancy forces is dissipated by the internal stresses as they continuously deform the mantle. Since the mantle rheology on plate-tectonic time scales is viscous, this energy dissipation is simply the viscous friction that converts mechanical work into heat. The second term on the right-hand side represents the rate of frictional dissipation of energy on internal discontinuity surfaces or faults, and

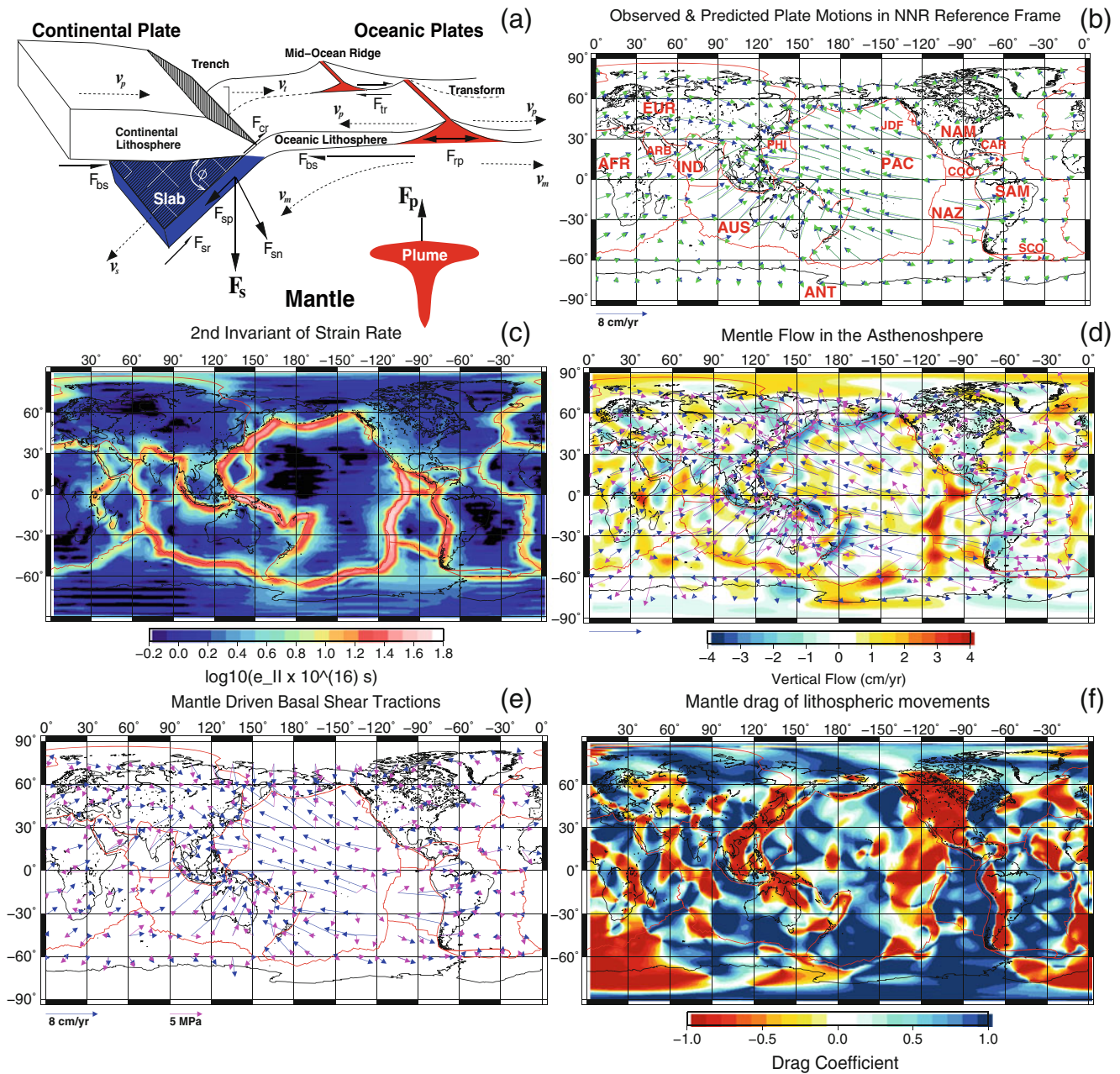


Plate-Driving Forces, Fig. 1 (a) Schematic illustration of forces acting on lithospheric plates. All plate-driving forces ultimately arise from the buoyancy forces in the convecting mantle, namely, the force F_p due to hotter and less dense upwelling plumes and the force F_s due to the colder and denser sinking slabs. These sublithospheric buoyancy forces generate the plate-driving forces F_{sp} (slab pull) and F_{rp} (ridge push). Whether the force F_{bs} (basal shear) is a plate-driving or plate-resisting force depends on the relationship of the surface plate velocity v_p to the underlying mantle flow velocity v_m (see main text for discussion). The purely resistive plate forces are F_{sr} (slab resistance), F_{cr} (collision resistance), and F_{tr} (transform resistance). The balance between the slab-normal force F_{sn} and the opposing normal force generated by mantle flow will control the trench rollback or retreat velocity v_t [Figure adapted and modified from Forsyth and Uyeda (1975)]. (b) The observed (green arrows) and predicted (blue arrows) present-day plate velocities, where the former are from the GEODVEL model (Argus et al. 2010) and the latter are obtained from a tomography-based model of mantle convection (Forte et al. 2015). The plate motions are with respect to a global No-Net-Rotation (NNR) frame of reference. The plate names are shown by the abbreviations in red font: AFR Africa, ANT

Antarctica, ARB Arabia, AUS Australia, CAR Caribbean, CCO Cocos, EUR Eurasia, IND India, JDF Juan de Fuca, NAM North America, NAZ Nazca, PAC Pacific, PHI Philippine, SAM South America, SCO Scotia. (c) A map of the second invariant of the strain-rate tensor ($\sqrt{\epsilon_{ij}\epsilon_{ij}/2}$) at the surface of the lithosphere predicted by the tomography-based mantle flow model employed in (b). (d) The mantle flow predicted at 250 km depth using the tomography-based convection model (Forte et al. 2015). The magenta arrows show the horizontal component of the predicted flow vector, and the color contours show the vertical component (scale at bottom). The blue arrows show the predicted surface plate motions from (b). (e) The horizontal traction vectors (magenta arrows) at 90 km depth (just below the mean depth of the lithosphere) exerted by the mantle flow predicted by the tomography-based convection model (Forte et al. 2015). The blue arrows show the predicted horizontal flow velocities at 90 km depth. Scale bars for vectors are at bottom. (f) A map of the drag coefficient D_b determined using the basal traction F_{bs} at 90 km depth, shown in (e), and the predicted surface plate motion v_p , shown in (b)

hence this term is only relevant in the upper mantle and especially the crust, where significant earthquake faulting occurs. Since the frictional tractions t_i will always oppose the discontinuous fault motion Δu_i , the integral over the fault surfaces S_d must be negative (e.g., Malvern 1969).

As noted above, a large number of models have evaluated the plate-driving forces by specifically focusing on the lithosphere itself, treated as distinct from the mantle. In this context, a buoyancy-dissipation balance analogous to Eq. (2) can be developed:

$$\begin{aligned} & - \int_{V_l} u_r \delta \rho g_o dV - \int_{S_l} [\delta P - \rho_o \delta \phi] u_i n_i dS \\ & + \int_{S_l} u_j \tau_{ji} n_i dS \\ & = \int_{V_l} \tau_{ij} \epsilon_{ij} dV - \int_{S_d} t_i \Delta u_i dS \end{aligned} \quad (3)$$

in which V_l represents the volume occupied by the lithospheric layer, S_l defines the lower surface of the lithosphere (where the local normal vector n_i is approximately pointing in the negative radial direction $-\hat{r}$), δP and $\rho_o \delta \phi$ are, respectively, the dynamic pressure field and self-gravitation force acting on the base of the lithosphere, and $\tau_{ji} n_i = t_j$ are the deviatoric shear tractions exerted on the base of the lithosphere by the underlying mantle flow.

If the lithosphere has a sufficiently high viscosity, owing to its relatively colder temperatures compared to the mantle, the deformation rate ϵ_{ij} in plate interiors may be assumed to be negligible. If this approximation is accepted, the first term on the right-hand side of Eq. (3) may be ignored, and the primary frictional dissipation of energy occurs on the fault surfaces S_d that define the boundaries of the plates. The vertical flow u_r in the lithospheric layer is weak, owing to the proximity of the upper bounding surface, and thus the work done by the tangential shear stresses exerted by mantle flow (third term on the left-hand side of Eq. 3) is approximately balanced by the frictional dissipation of energy on plate boundary weak zones (e.g., faults). Since the dissipation on the right-hand side of Eq. (3) is always positive, this implies the tangential stresses on the base of the lithosphere are, on average, acting in the direction of the flow and hence do positive work on lithosphere. These approximations illustrate how plate motions may be governed by a balance between basal shear stresses, acting as plate-driving forces, and resistance to motion along plate boundaries (e.g., Hanks 1977; Davies 1978).

Expressions (2) and (3) are obtained on the assumption that no external tangential stresses are applied to the upper and lower bounding surfaces of the mantle and that the local vertical flow vanishes on these bounding surfaces. Some models of plate-driving forces have employed prescribed

plate velocities as a boundary condition, thereby implying corresponding *externally* applied tangential stresses on the lithosphere. This external forcing is sometimes figuratively referred to as the “hand of God” (e.g., Karpychev and Fleitout 1996). To avoid the nonphysical possibility that such external stresses do work on the lithosphere and underlying mantle, the models should be formulated so that compensating surface stresses, generated by the flowing mantle, are able to precisely cancel the work of externally applied stresses (e.g., Hager and O’Connell 1981).

The principal idea that emerges from these general considerations of the plate-driving forces is that plate motions are produced by the buoyancy forces $\delta \rho g_o$ in the mantle (left-hand side of Eq. 2). Even when the focus is placed on the lithosphere itself, the contribution of the driving forces in the mantle is still present but “hidden” in the basal stresses that are exerted on the lithospheric plate (left-hand side of Eq. 3).

Plate-Driving Forces: Lithospheric Models

Formulating the detailed physical mechanisms by which the vertically directed buoyancy forces in the mantle ($\mathbf{F}_s$ and $\mathbf{F}_p$ in Fig. 1a) are ultimately expressed as horizontal forces acting on and driving the plates is the central challenge in developing models that predict realistic plate motions. The earliest models explored the dynamics of the lithosphere and its descent into the mantle under deep ocean trenches (e.g., Elsasser 1969, 1971; McKenzie 1969; Richter 1973), and they suggested that the negative buoyancy forces due to the cold lithospheric slabs entering the mantle ($\mathbf{F}_s$ in Fig. 1a) are the most efficient drivers of the observed tectonic plate motions.

These initial efforts set the stage for the subsequent development of *parametrized plate-force models* in which the forces driving the plate motions are represented by a limited number of discrete force vectors acting on the lithosphere (Fig. 1a). One of the most influential studies of the parametrized plate-force models was carried out by Forsyth and Uyeda (1975). The forces they considered are identified in Fig. 1a, and they include (1) the **slab-pull force** $\mathbf{F}_{sp} = \mathbf{F}_s \sin \phi$ acting in the down-dip direction where ϕ is the dip angle of the slab; (2) the **slab-resistance force** $\mathbf{F}_{sr}$ representing the opposing viscous shear stress acting on the sides of the descending slab; (3) the **collision-resistance force** $\mathbf{F}_{cr}$ representing the frictional stress acting on the upper surface of the slab as it slides past the fault surface of the shallow Benioff zone; (4) the **basal shear force** $\mathbf{F}_{bs}$ representing the tangential shear stress generated by the flowing mantle acting on the base of the lithospheric plates; (5) the **transform-resistance force** $\mathbf{F}_{tr}$ representing the frictional stresses acting on segments of the plate boundaries where there is a component of relative motion that is tangential to the boundary, such

as transform faults; and (6) the **ridge-push force** F_{rp} corresponding to the lateral force exerted on the divergent plate boundaries due to the dynamic pressure field that produces the topographic elevation of the mid-ocean ridges.

The orthogonal complement to the slab-pull force is the slab-normal force $F_{sn} = F_s \cos \phi$ (Fig. 1a). This force component was not considered in the evaluation of the plate-driving forces by Forsyth and Uyeda (1975), but the balance between this force and the opposing normal stress in the mantle will contribute to the retrograde motion (or “rollback”) of the trench (e.g., Elsasser 1971; Schellart 2008; Funicello et al. 2008) with the relative velocity v_r , illustrated in Fig. 1a.

The results obtained by Forsyth and Uyeda (1975), and earlier by Richter (1973), suggested that the balance of forces on subducting slabs is the primary control on plate motions, and this understanding has been subsequently reiterated through a long suite of published models (e.g., Lithgow-Bertelloni and Richards 1998; Conrad and Lithgow-Bertelloni 2004; Billen 2008; Funicello et al. 2008; Schellart 2008; Stadler et al. 2010) up to the present day (e.g., Rolf et al. 2018; Coltice et al. 2019).

A major source of uncertainty in the classic force-balance models of plate motions is the unknown relationship between the motion of the lithospheric plates v_p and the underlying mantle flow v_m (Fig. 1a) and hence the unknown local magnitude and direction of the basal shear force F_{bs} . The majority of models assumed the mantle is passive and that the flow below the lithosphere is entirely generated by the subducting slabs. This is equivalent to assuming that the ambient mantle temperature outside the sinking lithosphere is isothermal or adiabatic (e.g., McKenzie 1969; Hager and O’Connell 1981). For these passive-mantle models, the basal shear force acts in the opposite direction to the plate velocity and therefore opposes the plate-driving force provided by subduction. Recent global convection modelling of plate-like surface motions driven by subducting slabs (Coltice et al. 2019) indeed shows that most (in excess of 2/3 of the global surface area) of the sublithospheric mantle is passive and driven by the overlying plate motions.

The conclusions obtained from traditional force-balance modelling are nonunique and depend on which forces are assumed to be known. For example, whether basal shear forces contribute to driving or resisting the plate motions will depend on what forces (in Fig. 1a) are prescribed as known at the outset. In the study by Forsyth and Uyeda (1975), it was assumed that basal shear forces oppose the plate motions and have a direction that is antiparallel to the plate velocity. In contrast, Davies (1978) showed that if the frictional force F_{tr} acting on transform boundaries is assumed to be known in advance, on the basis of earthquake stress-drop data (Hanks 1977), then the basal shear force may act as a plate-driving force. The force balance considered by Davies

(1978) also showed that the collision-resistance force F_{cr} may strongly oppose the slab-pull force F_{sp} , greatly reducing its contribution as a plate-driving force. The significance of basal shear stresses as a plate-driving force has also been highlighted by Bird et al. (2008) who employed a finite-element model of the lithosphere that incorporates stress- and temperature-dependent viscosity as well as an explicit treatment of plate boundary faults. Ghosh and Holt (2012) similarly found that basal shear tractions generated by buoyancy-driven mantle flow are the primary drivers of surface lithospheric velocities and yield close matches to present-day plate motions. The latter model results (see also Rowley et al. 2016) showing the importance of sublithospheric basal tractions as active drivers of plate motions differ from models that show a mainly passive mantle entrained by overlying plates driven by subducting slabs (e.g., Coltice et al. 2019).

Plate-Driving Forces: Whole-Mantle Models

The popularity of parametrized plate-force models that focused on the lithosphere was based on the use of readily observed input variables derived from surface tectonics and observed plate velocities. Since the observed plate velocities were employed as an input, these models could not explain how the plate motions were generated by the internal dynamics of the mantle nor how they would evolve over time.

Hager and O’Connell (1981) introduced 3-D spherical, whole-mantle flow models that explicitly included the interaction of the moving lithospheric plates with the underlying mantle. Although these models employed the present-day plate velocities as input, they predicted the 3-D flow throughout the mantle driven by a global distribution of subducted slabs derived from seismic catalogues of earthquake hypocenters. With the advent, in the 1980s, of seismic tomographic imaging of lateral heterogeneity in the mantle, it became feasible to apply these whole-mantle flow calculations to explore the detailed connection between surface plate motions and the 3-D flow in the mantle generated by thermal convection (e.g., Forte and Peltier 1987; Ricard and Vigny 1989).

Seismic tomography provides the only direct in situ constraints on the 3-D lateral structure of the convecting mantle, and it thus provides a unique basis for developing realistic models of mantle flow that are no longer limited to an understanding of Earth dynamics derived only from surface observations. The models of mantle flow by Hager and O’Connell (1981) were therefore further extended in studies demonstrating that present-day plate motions can be successfully predicted by 3-D distributions of buoyancy forces derived from seismic tomography (e.g., Vigny et al. 1991; Forte and Peltier 1994; Becker and O’Connell 2001; Ghosh and Holt 2012; Forte et al. 2015). This confirmation that plate motions can be

predicted on the basis of stresses generated in the underlying mantle by a 3-D pattern of convection is a major step toward developing a complete understanding of the dynamics and temporal evolution of the mantle and lithosphere as a single, internally coupled system (e.g., Lithgow-Bertelloni and Richards 1998; Conrad and Lithgow-Bertelloni 2004; Lowman et al. 2008; Forte et al. 2009; Rowley et al. 2016; Rolf et al. 2018).

A tomography-based model of present-day mantle flow (Forte et al. 2015) illustrates the detailed relationship between the convective circulation in the mantle and surface plate velocities. The model incorporates coupling of viscous mantle flow to overlying plates whose motions are limited to rigid-body rotations. The buoyancy-driven plate velocities predicted by this model (Fig. 1b) match the observations well (explaining 97% of data variance). The predicted lithospheric strain rates (Fig. 1c) are characterized by values ranging from less than $3 \times 10^{-16} \text{ s}^{-1}$ in plate interiors to two orders of magnitude greater along plate boundaries. According to Eq. (1), plastic weakening would then predict viscosity reductions in plate boundary zones between one and two orders of magnitude. Below the lithosphere, in the underlying low-viscosity portion of the upper mantle referred to as the “asthenosphere,” the predicted flow velocities show significant deviations from the overlying plate motions, for example, under the western half of the North American plate and under the African plate (Fig. 1d). The lack of parallelism between the surface plate velocities and the upper-mantle flow circulation in the asthenosphere is a consequence of the complex 3-D distribution of buoyancy forces in the deep mantle that are not correlated with the surface plate geometry (e.g., subduction zones and mid-ocean ridges) in any straightforward manner. An important exception is the prediction of a mantle-wide upwelling centered directly below the East Pacific Rise (Rowley et al. 2016). The lack of parallelism between surface and sublithospheric motions is also a consequence of the strong viscosity reduction in the asthenosphere, which allows a partial decoupling between these two layers.

The complex local variations in basal shear force relative to the surface velocities show that a simple correlation between these two, as assumed in the parametrized plate-force models (e.g., Forsyth and Uyeda 1975), is not appropriate. The local correlation between basal stresses and lithospheric motions may be quantified in terms of a drag coefficient, which provides a measure of the extent to which sublithospheric mantle flow actively “drags” the overlying lithosphere or vice versa. Two possible representations for the local mantle drag coefficient are:

$$D_b = \mathbf{F}_{bs} \cdot \mathbf{v}_p / (|\mathbf{F}_{bs}| |\mathbf{v}_p|) \quad \text{or} \quad D_v = \frac{\mathbf{v}_m}{\mathbf{v}_p} \cos \theta - 1 \quad (4)$$

in which θ is the angular deviation between the sublithospheric velocity $\mathbf{v}_m$ and the overlying plate velocity $\mathbf{v}_p$

(Fig. 1a). The drag coefficients D_b and D_v are correlated, because the orientation and magnitude of the horizontal motions in the plates and sublithospheric mantle will determine the orientation and magnitude of the basal shear force $\mathbf{F}_{bs}$ (Fig. 1a). A simple 1-D analysis of purely horizontal flow shows that when the flow velocity increases with depth (i.e., the mantle “drags” the overlying plate), then both D_b and D_v will be positive. The coefficient D_v is closely related, but opposite in sign, to the drag coefficient defined in Coltice et al. (2019). The tomography-based flow prediction of plate motions (Fig. 1d, e) yields the local drag coefficient D_b shown in Fig. 1f. Positive mantle drag of the overlying plate motions occurs in 63% of the total surface area of the lithosphere. A similar calculation of D_v (not shown here) yields a positive mantle drag in 65% of the surface area. The theoretical convection-driven plate motions simulated in Coltice et al. (2019) yield positive mantle drag in 20 to 40% of the surface.

The tomography-based whole-mantle flow models thus suggest that basal shear forces generated by a fully 3-D pattern of mantle convection drive the motions of the overlying plates (e.g., Vigny et al. 1991; Forte and Peltier 1994; Becker and O’Connell 2001; Ghosh and Holt 2012; Forte et al. 2015). The role of the plates is not entirely passive, however, because their intrinsic rigidity allows them to partially resist (and hence mask) a significant fraction of the deeper mantle flow patterns (e.g., Ricard and Vigny 1989; Forte et al. 2015). Furthermore, the buoyancy forces associated with subducted lithosphere do contribute to the basal shear forces that drive the plates (e.g., Coltice et al. 2019). Indeed, the continuous descent of lithosphere into the mantle ultimately generates a global distribution of subducted slabs that provide the negative buoyancy force $\mathbf{F}_s$ (Fig. 1a).

Summary

Much progress has been achieved in our understanding of how mantle convection drives the surface plate motions since the earliest parametric force models were published over four decades ago. In particular, the increase and availability of high-performance computational resources over the past decade have permitted the development of global mantle convection simulations that incorporate increasingly Earth-like characteristics, including the strong lateral variations of viscosity needed to simulate plate-tectonic movements (e.g., Ghosh and Holt 2012; Rolf et al. 2018; Coltice et al. 2019). Furthermore, major improvements in seismic tomographic inversions for 3-D mantle structure have been incorporated into global models of mantle flow that yield very good matches to a wide suite of convection-related surface data, including the present-day plate motions (e.g., Forte et al. 2015).

The interpretation of the seismic tomography models is however imperfect because individual published tomography

models are derived from damped, smoothed inversions and they are nonunique. Moreover, the independent mineral physical modelling required to convert the seismic anomalies in the tomography models into corresponding density perturbations adds further uncertainties that may impact the reliability of the mantle flow predictions (e.g., Rowley et al. 2016). These outstanding challenges have led to the emergence of two contrasting views on how thermal convection in the mantle drives the tectonic plate motions. The long-held, predominant view is that the buoyancy forces due to subducted slabs are the primary drivers of surface plate motions (e.g., Richter 1973; Conrad and Lithgow-Bertelloni 2004). According to this view, the thermal evolution of the mantle is mainly controlled by cooling from above (i.e., subducting lithosphere), which occurs when the mantle heating is mainly internal, with little heat flux across the core-mantle boundary. In this case, the buoyancy flux associated with subducted slabs strongly dominates that of hot upwellings, and the slabs therefore determine the dynamics of surface lithospheric movements (e.g., Coltice et al. 2019).

The second view that originated with the earliest mantle convection models (e.g., Turcotte and Oxburgh 1967) and some subsequent models (e.g., Jarvis and Peltier 1982; Forte et al. 2015; Rowley et al. 2016) is that a combination of active hot upwellings (“plumes”) and subducted slabs is required to explain the surface lithospheric motions. For such upwellings to be dynamically significant, there must be a large heat flux across the core-mantle boundary (CMB) and a well-developed lower thermal boundary layer that can sustain the activity of these hot plumes (e.g., Glišović et al. 2012). Time-dependent convection modelling that integrates a mantle buoyancy distribution derived from global seismic tomography yields a constellation of strongly developed and long-lived hot upwellings that remain geographically stable in time and are sustained by a large CMB heat flux ranging between 13 and 20 Terawatts (Glišović and Forte 2015). These tomography-based convection models are characterized by a present-day buoyancy flux from hot upwellings (i.e., plumes) and cold downwellings (i.e., slabs) that are of equal magnitude, such that the hot upwellings provide a substantial contribution to plate-driving forces (e.g., Rowley et al. 2016). The issue of heat flux across the CMB thus occupies a central role in determining the dynamical contribution of mantle-wide hot upwellings on plate-driving forces. Independent constraints on this heat flux from other disciplines, such as mineral physics (e.g., Pozzo et al. 2012; Ohta et al. 2016), are important in addressing this debate.

Cross-References

- Continental Drift
- Earth's Structure, Upper Mantle
- Energy Budget of the Earth

- Lithosphere, Mechanical Properties
- Lithosphere, Oceanic
- Lithosphere, Oceanic: Thermal Structure
- Mantle Convection
- Mantle Plumes
- Mantle Viscosity
- Seismic Structure at Mid-Ocean Ridges
- Seismology, Global Earthquake Model
- Subduction Zones

Bibliography

- Argus DF, Gordon RG, Heflin MB, Ma C, Eanes RJ, Willis P, Peltier WR, Owen SE (2010) The angular velocities of the plates and the velocity of Earth's centre from space geodesy. *Geophys J Int* 180(3):913–960
- Becker TW, O'Connell RJ (2001) Predicting plate velocities with mantle circulation models. *Geochem Geophys Geosyst* 2. <https://doi.org/10.1029/2001GC000171>
- Billen MI (2008) Modeling the dynamics of subducting slabs. *Annu Rev Earth Planet Sci* 36:325–356
- Bird P, Liu Z, Rucker WK (2008) Stresses that drive the plates from below: definitions, computational path, model optimization, and error analysis. *J Geophys Res* 113:B11406. <https://doi.org/10.1029/2007JB005460>
- Chapple WM, Tullis TE (1977) Evaluation of the forces that drive the plates. *J Geophys Res* 82:1967–1984
- Coltice N, Husson L, Faccenna C, Arnould M (2019) What drives tectonic plates? *Sci Adv* 5(10):eaax4295
- Conrad CP, Lithgow-Bertelloni C (2004) The temporal evolution of plate driving forces: importance of “slab suction” versus “slab pull” during the Cenozoic. *J Geophys Res* 109:B10407. <https://doi.org/10.1029/2004JB002991>
- Davies GF (1978) The roles of boundary friction, basal shear stress and deep mantle convection in plate tectonics. *Geophys Res Lett* 5:161–164
- Davies GF (1988) Role of the lithosphere in mantle convection. *J Geophys Res* 93(10):451–10,466
- Elsasser WM (1969) Convection and stress propagation in the upper mantle. In: Runcorn SK (ed) *The applications of modern physics to the earth and planetary interiors*. Interscience, New York, pp 223–246
- Elsasser WM (1971) Sea-floor spreading as thermal convection. *J Geophys Res* 76:1101–1112
- Evans B, Kohlstedt DL (1995) Rheology of rocks, in rock physics and phase relations: a handbook of physical constants, AGU Reference Shelf 3. American Geophysical Union, Washington, DC, pp 148–165
- Forsyth D, Uyeda S (1975) On the relative importance of the driving forces of plate motion. *Geophys J R Astron Soc* 43:163–200
- Forte AM, Peltier WR (1987) Plate tectonics and aspherical earth structure: the importance of poloidal–toroidal coupling. *J Geophys Res* 92:3645–3679
- Forte AM, Peltier WR (1994) The kinematics and dynamics of poloidal–toroidal coupling in mantle flow: the importance of surface plates and lateral viscosity variations. *Adv Geophys* 36:1–119
- Forte AM, Moucha R, Rowley DB, Quéré S, Mitrovica JX, Simmons NA, Grand SP (2009) Recent tectonic plate decelerations driven by mantle convection. *Geophys Res Lett* 36:L23301. <https://doi.org/10.1029/2009GL040224>
- Forte AM, Simmons NA, Grand SP (2015) In: Romanowicz B, Dziewonski AM (eds) *Constraints on seismic models from other*

- disciplines – implications for the global mantle convective flow, in volume 1 of treatise of geophysics, 2nd edn. Elsevier (Amsterdam), pp 853–907
- Funiciello F, Faccenna C, Heuret A, Lallemand S, Di Giuseppe E, Becker TW (2008) Trench migration, net rotation and slab-mantle coupling. *Earth Planet Sci Lett* 271:233–240
- Gable CW, O'Connell RJ, Travis BJ (1991) Convection in three dimensions with surface plates. *J Geophys Res* 96:8391–8405
- Ghosh A, Holt WE (2012) Plate motions and stresses from global dynamic models. *Science* 335(6070):838–843
- Glišović P, Forte AM (2015) Importance of initial buoyancy field on evolution of mantle thermal structure: implications of surface boundary conditions. *Geosci Front* 6(1):3–22
- Glišović P, Forte AM, Moucha R (2012) Time-dependent convection models of mantle thermal structure constrained by seismic tomography and geodynamics: implications for mantle plume dynamics and CMB heat flux. *Geophys J Int* 190(2):785–815
- Hager BH, O'Connell RJ (1981) A simple global model of plate dynamics and mantle convection. *J Geophys Res* 86:4843–4867
- Hanks TC (1977) Earthquake stress drops, ambient tectonic stresses and stresses that drive plate motions. *Pure Appl Geophys* 115:441–458
- Harper JR (1975) On the driving forces of plate tectonics. *Geophys J R Astron Soc* 40:465–474
- Jarvis GT, Peltier WR (1982) Mantle convection as a boundary layer phenomenon. *Geophys J R Astron Soc* 68:385–424
- Jeffreys H (1972) Creep in the earth and planets. *Tectonophysics* 13(1–4):569–581
- Karato SI, Barbot S (2018) Dynamics of fault motion and the origin of contrasting tectonic style between earth and Venus. *Sci Rep* 8(1):11884
- Karato S-i, Wu P (1993) Rheology of the upper mantle: a synthesis. *Science* 260:771–778
- Karpychev M, Fleitout L (1996) Simple considerations on forces driving plate motion and on the plate-tectonic contribution to the long-wavelength geoid. *J Geophys Res* 127:268–282
- Lithgow-Bertelloni C, Richards MA (1998) The dynamics of Cenozoic and Mesozoic plate motions. *Rev Geophys* 36:27–78
- Lowman JP, Gait AD, Gable CW, Kukreja H (2008) Plumes anchored by a high viscosity lower mantle in a 3D mantle convection model featuring dynamically evolving plates. *Geophys Res Lett* 35: L19309. <https://doi.org/10.1029/2008GL035342>
- Malvern LE (1969) Introduction to the mechanics of a continuous medium. Prentice-Hall, Englewood Cliffs
- McKenzie DP (1969) Speculations on the consequences and causes of plate motions. *Geophys J R Astron Soc* 18:1–32
- Moresi L, Solomatov V (1998) Mantle convection with a brittle lithosphere: thoughts on the global tectonic styles of the earth and Venus. *Geophys J Int* 133(3):669–682
- Ohta K, Kuwayama Y, Hirose K, Shimizu K, Ohishi Y (2016) Experimental determination of the electrical resistivity of iron at Earth's core conditions. *Nature* 534(7605):95
- Poirier J-P (1985) Creep of crystals. Cambridge University Press, Cambridge
- Pozzo M, Davies C, Gubbins D, Alfe D (2012) Thermal and electrical conductivity of iron at Earth's core conditions. *Nature* 485(7398):355–358
- Ricard Y, Vigny C (1989) Mantle dynamics with induced plate tectonics. *J Geophys Res* 94:17543–17559
- Richardson RM, Solomon SC, Sleep NH (1976) Intraplate stress as an indicator of plate tectonic driving forces. *J Geophys Res* 81:1847–1856
- Richter F (1973) Dynamical models for sea floor spreading. *Rev Geophys Space Phys* 11:223–287
- Richter F (1977) On the driving mechanism of plate tectonics. *Tectonophysics* 38:61–88
- Rolf T, Capitanio FA, Tackley PJ (2018) Constraints on mantle viscosity structure from continental drift histories in spherical mantle convection models. *Tectonophysics* 746:339–351
- Rowley DB, Forte AM, Rowan CJ, Glišović P, Moucha R, Grand SP, Simmons NA (2016) Kinematics and dynamics of the East Pacific rise linked to a stable, deep-mantle upwelling. *Sci Adv* 2(12): e1601107
- Schellart WP (2008) Kinematics and flow patterns in deep mantle and upper mantle subduction models: influence of the mantle depth and slab to mantle viscosity ratio. *Geochem Geophys Geosyst* 9:Q03014. <https://doi.org/10.1029/2007GC001656>
- Solomon SC, Sleep NH (1974) Some simple physical models for absolute plate motions. *J Geophys Res* 79:2557–2567
- Spiegelman M, May DA, Wilson CR (2016) On the solvability of incompressible Stokes with viscoplastic rheologies in geodynamics. *Geochem Geophys Geosyst* 17:2213–2238
- Stadler G, Gurnis M, Burstedde C, Wilcox LC, Alisic L, Ghattas O (2010) The dynamics of plate tectonics and mantle flow: from local to global scales. *Science* 329(5995):1033–1038
- Tackley PJ (2000) Mantle convection and plate tectonics: toward an integrated physical and chemical theory. *Science* 288:2002–2007
- Turcotte DL, Oxburgh ER (1967) Finite amplitude convective cells and continental drift. *J Fluid Mech* 28:29–42
- Vigny C, Ricard Y, Froidevaux C (1991) The driving mechanism of plate tectonics. *Tectonophysics* 187:345–360

Plates and Paleoreconstructions

Alan G. Smith
(Deceased)

Definition

The *lithosphere* is the outer rigid part of the Earth, forming a shell whose thickness may range up to about 200 km.

A *tectonic plate* is a part of the Earth's lithosphere that is bounded by active plate margins.

A present-day *plate margin* is a seismically active zone that cuts the lithosphere.

Conservative plate boundaries are plate boundaries that separate two plates that are sliding past one another along a *transform fault* (which must cut the lithosphere). *Divergent (extensional) plate boundaries* mark where the two plates are separating from one another, generally forming *passive (or Atlantic) continental margins* (though extension can also occur behind island arcs).

Convergent plate boundaries are located between a converging oceanic and continental plate forming an *active (Pacific) continental margin*, or between a converging oceanic plate and an island arc.

A *global paleoreconstruction* shows a reassembly of the major continents (and oceans) relative to one another at some time in the past. If the reassembly is made using ocean-floor spreading magnetic anomalies and fracture zones, and derives

the geographic poles from paleomagnetism, it is a *global paleomap*. If only the continents are depicted, it is a *global paleocontinental reconstruction (or map)*. If each continent is projected separately onto a global map frame (as in pre-Mesozoic maps) it is a *global composite reconstruction*.

A *paleo-plate reconstruction* is simply a paleoreconstruction that shows the plate boundaries of the time concerned, that is, ridges, trenches, and major transform faults.

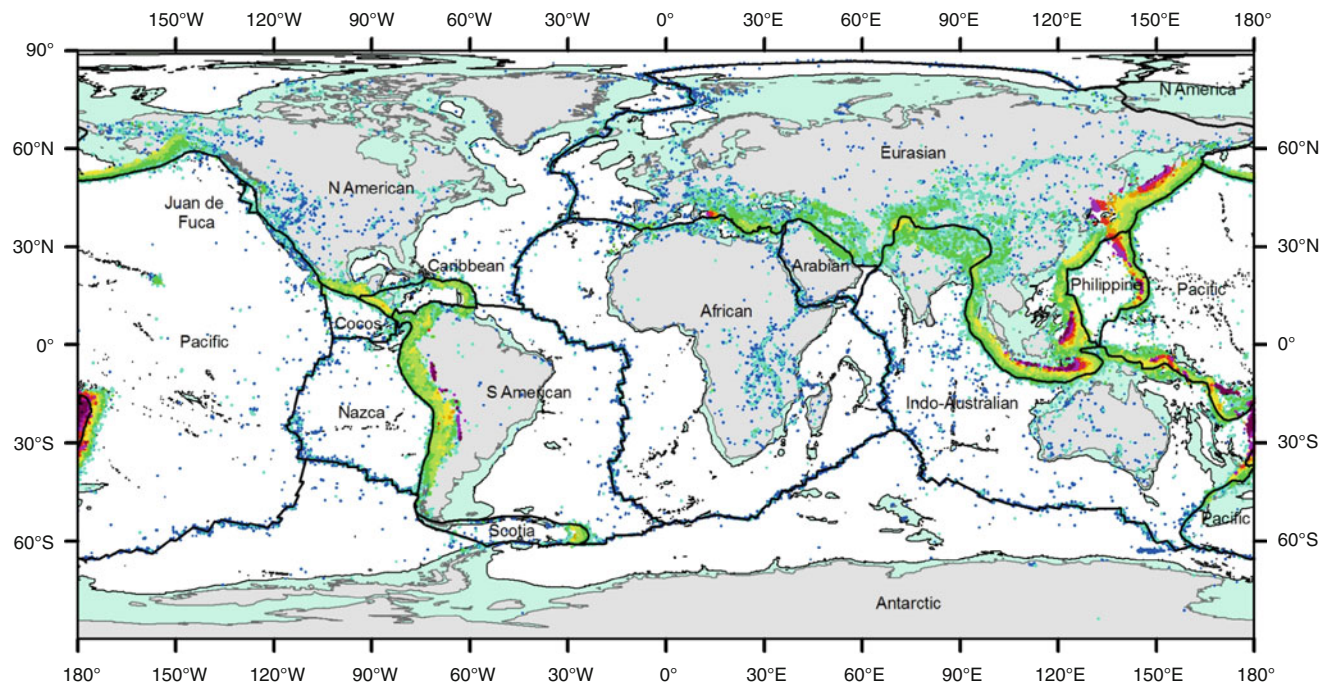
Earthquakes and Plates

Earthquakes are caused by the catastrophic release of accumulated elastic strain. The strain accumulates as a result of differential movement between parts of the lithosphere. A map of earthquakes over the past 40 years or so shows that differential movement is generally confined to narrow

seismic zones, particularly in the oceans (Fig. 1). Elsewhere there is little or no seismicity, showing that most of the lithosphere is not undergoing differential motion and is therefore rigid. The present-day surface extent of a plate is marked by zones of active seismicity that show the plate boundaries.

Plate Compositions, Sizes, and Shapes

Plates are made up of *Lithosphere, Continental*, with a thickness that probably averages about 200 km (Priestley and McKenzie 2006), and *Lithosphere, Oceanic*, whose thickness varies from a small value at a mid-ocean ridge, where it is created, to about 100 km for old oceanic lithosphere (Stein and Stein 1992; Crosby et al. 2006). Most of the plates in the Pacific Ocean are almost entirely, or entirely, made of oceanic lithosphere, but other plates are made up of varying



Plates and Paleoreconstructions, Fig. 1 Present-day earthquakes and simplified map of the present-day plates. The Philippine plate is also known as the Philippine Sea plate. The plate boundary between Eurasia and North America in northeast Asia is poorly defined. Cylindrical equidistant world map showing land areas as *light gray*. Areas between 2000 m water depth and the coastline are *blue* (colored version) or *light gray* (monochrome version). Deeper oceanic areas are uncolored. Over 400,000 earthquakes are shown. In the colored map, the color code is: 0–10 km (79,478) *blue*; 10–35 km (168,263) *light blue*; 35–100 km (79,890) *green*; 100–200 km (47,056) *yellow-green*; 200–300 km (11,144) *yellow*; 300–400 km (3874) *orange-brown*; 400–500 km (4674) *red*; 500–600 km (8427) *magenta*; 600–700 km (2477) *deep magenta*; >700 km (4) *black*. The monochrome version shows earthquakes ranging from *light gray* (shallow) to *black* (deepest). The rainbow-colored areas in the Australasia, together with smaller areas

along the western coast of the Americas, indicate zones of seismic activity of increasing depth. The continuous *black* line is the plate network showing 14 plates, all labeled. It is unclear whether the Indo-Australian plate should be considered to be two plates with a boundary in the NW Indian Ocean. The plate margins in the oceans are very narrow and conform to ideal plate margins; those in the continents are much broader. In particular, the wide belt of active deformation in the Alpine-Himalayan belt shows that the deformation is not restricted to the narrow plate margin on the map between the African, Arabian, and Indo-Australian plates on the one hand, and the southern Eurasian plate on the other. Rigid plate tectonic descriptions of such areas may be inappropriate. (Copyright A. G. Smith). The earthquakes are those of magnitude 4 or greater listed in <http://earthquake.usgs.gov/earthquakes/eqarchives/epic/>

proportions of continental and oceanic lithosphere (Fig. 1). There are no wholly continental plates. The shapes of most plates are quite irregular, reflecting their past histories of growth and destruction.

Plate Tectonics

Plate tectonics describe the motions of the lithospheric plates and the effects they cause. It is now the major branch of terrestrial tectonics, which is the study of the origins of the large-scale structures of the Earth, particularly the lithosphere. The name tectonics is derived from the Greek word *tektonikos* (τεκτονικός) “pertaining to building.”

Plate Motions and Euler’s Theorem

Plates are rigid bodies that can move relative to one another in three ways: one, they can slide past one another; two, they can move away from one another; three, they can move toward one another.

According to a theorem attributed to Euler, the relative motion of two rigid bodies on a sphere can be described as a rotation about a line passing through the Earth’s center that cuts the Earth’s surface at a point known as the rotation pole. The pole will have a latitude, λ , and a longitude, φ , with the rotation that takes place in a given time given by θ (Fig. 2). To a good approximation the Earth is a sphere and tectonic plates are rigid. Thus the motion of one plate relative to another in a

given time interval is given by their Euler rotation (λ, φ, θ). λ varies from 90° N, the N pole, conventionally given the value of +90, to 90° S, the S pole, or – 90. Longitudes are positive measured E from the Greenwich meridian, conventionally taken as 0°, or as 360°. If measured W from the Greenwich meridian, the longitudes are negative. Maps showing the whole world run from 0° E to 360° E, or 0 to –360. They could also run from 180° W, that is, –180, to 180° E, or +180. The sign of the rotation θ is found by fixing one plate as a reference and determining whether the rotation moves the second plate in an anticlockwise direction when looking down on it – a positive rotation – or moves clockwise – a negative rotation. Changing the reference plate changes the sign of the rotation.

Sliding Motions

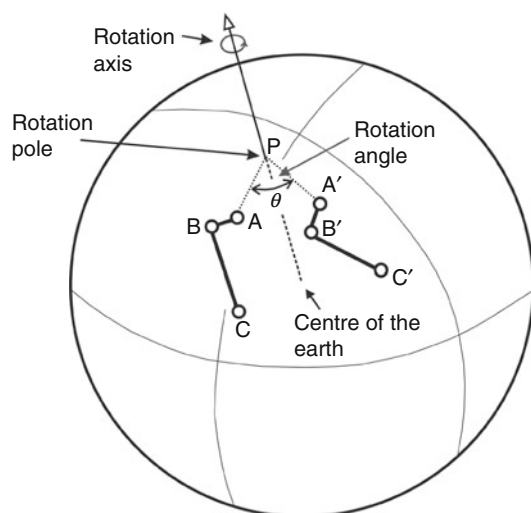
When plates slide past one another they neither create nor destroy lithosphere. They conserve the lithosphere and are described as conservative plate boundaries. On a sphere such boundaries are circular arcs forming what are known as transform faults as in the San Andreas fault. The Euler pole is the point at the center of the arc.

Divergent Motions

When a continent breaks up and the pieces move away from one another the motion is divergent. At first the margins of the continents-to-be are stretched, forming what will become passive continental margins, as in the East African Rift. Eventually the stretching reaches the stage where new ocean-floor has to form, as in the Red Sea or the Gulf of Aden. The Euler rotations for each stage are given by the rotations needed to bring the different markers (passive margins; anomalies of the same age) together.

Convergent Motions

The local motions between present-day converging plates can be found from the slip vectors of earthquakes. If the convergent margin is long enough it is also possible to use these motions to estimate the position of the Euler pole. However, the local motions and Euler poles for ancient plate margins are difficult to estimate. In favorable circumstances it is possible to calculate them from a plate circuit. For example, the Cenozoic motion along a convergent plate boundary between India and the rest of Eurasia can be calculated by successively summing the motions from Eurasia to North America, North America to Africa, and Africa to the Indo–Australian plate (Fig. 1). All of these motions are known from the ocean-floor



Plates and Paleoreconstructions, Fig. 2 The line ABC on a rigid body on the surface of a sphere (e.g., Earth) is moved to A'B'C' by a rotation through the angle θ , about the rotation axis. The rotation axis passes through the center of the Earth and cuts the surface at point P with geographic coordinates of latitude λ and longitude φ (not shown), leading to an Euler rotation of (λ, φ, θ). (Copyright A. G. Smith)

spreading patterns in the Atlantic and Indian Oceans. Their sum gives the required result. The absence of any Paleozoic and Precambrian ocean-floor means that the motions and Euler poles at all pre-Mesozoic convergent plate boundaries are highly uncertain.

Reference Frames

Three classes of reference frames are relevant to global reconstructions: any plate, the paleomagnetic reference frame, and that given by “hot-spots.”

Plates as Reference Frames

Any plate, large or small, that can be linked to all the other plates can be used as reference frame to make a global reassembly: there is nothing fundamental in the choice of a plate. These global reassemblies are not geographic maps because they do not show the geographic latitude and longitude of the time concerned. To turn a reassembly into a map one needs to know the position of the geographic pole at the time concerned, which is given by paleomagnetism.

Paleomagnetism as a Reference Frame

Over a period of several tens of thousands of years or more, the Earth’s magnetic field averages to what is known as a geocentric axisymmetric dipole field, or GADF. Virtually all paleomagnetic measurements use this field model to interpret the

results. The field is geocentric because it behaves as if it were at the Earth’s center; it is axisymmetric because the field is symmetric about the Earth’s spin (or rotation) axis; and it behaves like a bar magnet and is therefore a dipole. Such a field has a very simple relationship between the angle, I , known as the inclination that the lines of magnetic force make with the horizontal, and the latitude, λ , of the point of observation:

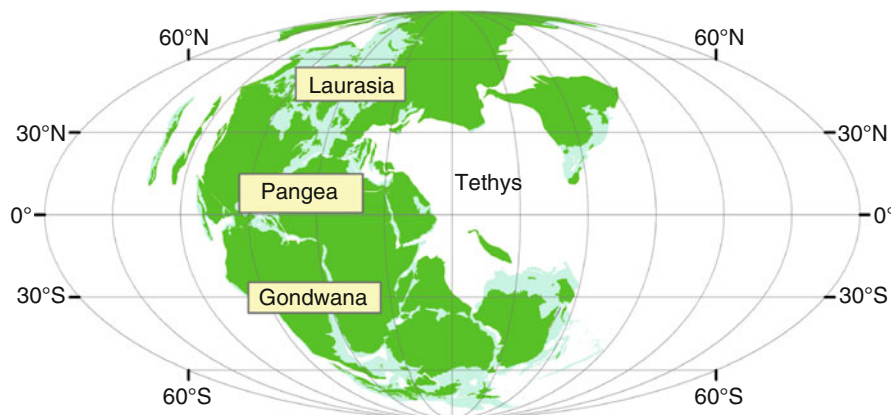
$$\tan I = 2 \tan \lambda, \quad (1)$$

where $\tan$ = tangent of the angle (I or λ).

As a plate moves, the latitude of a point on it changes and the orientation of the plate varies. Both these effects are recorded in the remanent magnetization – or *Paleomagnetism, Principles* – preserved in rocks. The actual position of the north magnetic pole at the time must lie along the line through the observation point that coincides with the direction of the declination. It is at a distance of $(90 - \lambda)$ degrees from the observation point along that line. Thus, in principle, paleomagnetic data can be used to reposition that plate at the time that the old magnetism was superimposed on the rocks.

The determination of the age and direction of magnetization in rocks becomes progressively more difficult as one goes back in time, but these problems will not be discussed here. However, any Mesozoic or Cenozoic global reassembly can be turned into a map by finding where the mean north magnetic pole lies on the reassembly and making that point the north geographic pole of the map (Smith et al. 1994). A map of Pangea just before break-up is shown in Fig. 3.

Although paleomagnetic measurements give the past latitude and orientation of a plate, they do not give its past longitude. Thus it is generally not possible to make a



Plates and Paleoreconstructions, Fig. 3 Mollweide projection of Pangea reassembled at the end of the Triassic period, ~200 Ma. The green areas are the present-day coastlines; the blue areas all lie above 2000 m water depth, showing the approximate present-day extent of the continental crust. The latitude-longitude grid is drawn at 30° and shows in particular the estimated position of the late Triassic equator. Pangea was assembled by the collision of the southern continents (Gondwana) with the northern continents (Laurasia). A large oceanic area, known as

the Tethys, separated southern Eurasia from northeast Gondwana. The distribution of continental fragments lying within the Tethys is uncertain, particularly that of China and adjacent areas (the large fragment lying east of Eurasia). Some continental slivers and/or island arcs off western North America are shown schematically. Several other slivers probably lay adjacent to other continental margins of Pangea, but no attempt has been made to depict them. (Copyright A. G. Smith)

paleoreconstruction from paleomagnetic measurements alone. However, paleo-longitude differences between any two points on different continents will be correctly given if the relative position of the two points has been determined via a plate tectonic circuit whose individual motions are known.

Hot-Spots as a Reference Frame

The origin of linear volcanic chains like the Hawaiian Islands has been attributed to the slow movement of a tectonic plate over a fixed deep plume of hot mantle, giving rise to “hot-spots” at the surface (Wilson 1963; Morgan 1981). A purely descriptive term for them is that they are examples of “large igneous provinces” created in the case of hot-spots by intense, generally basaltic, magnetism. From the point of view of making reconstructions they show an interesting property in that they move relative to one another at rates that are an order of magnitude slower than those of plates. Hot-spots have been used as a reference frame that is believed to provide longitudes as well as latitudes for a global reassembly, forming a so-called absolute reference frame (Duncan 1981; Müller et al. 1993). The paleomagnetic reference and hot-spots reference frames agree within the limits of error for much of Cenozoic time, then gradually diverge for reasons that are not clear (Livermore et al. 1983).

Paleoreconstructions

Paleoreconstructions attempt to show where the oceans and continents were in the geological past. That this is possible is due to several effects: (1) conservation of continental crust during plate motions; (2) the record of Mesozoic and Cenozoic continental separation in the intervening ocean-floor; and (3) the record of changes in continental latitude and orientation in *Paleomagnetism, Principles*.

Conservation of Continental Shapes

The extension on passive continental margins is probably about 100 km for a margin that is 200 km wide at the present-day, such as parts of the eastern North American margin (e.g., Sawyer 1985). In other areas, the extension may be much higher: Powell et al. (1988) estimated a total of 360 km extension across the Australia–Antarctica margin before the formation of ocean-floor (180 km) equivalent to 1.6° of latitude for each margin, assuming symmetric extension.

To a good approximation, continental shapes on a global scale are conserved during plate motions. In detail this is clearly not the case. For example, when continents separate

the passive margins have been stretched. When two continents converge, the continental lithosphere of one continent overrides that of the other, which is then partially subducted. But because continental lithosphere is less dense than the underlying mantle, it cannot be subducted to any great depth. Thus the continental crust in the collisional zone becomes thickened and the area of continental crust decreases. In the case of the collision between India and Asia, the Tibetan plateau, and the Himalayas may represent a loss of continental area in a strip that may be more than 500 km wide (Coward et al. 1988; Shackleton 1988). As one goes back in time, the subducted Indian crust should be pulled out from under Eurasia and added to the Indian continent. Similar adjustments, but on a generally smaller scale, are necessary to the shapes of all continents caught up in collision zones, but are rarely made.

The continental shapes used for reconstructions generally show the edges of continents at about 2000 m water depth. This shape is unfamiliar where there is a wide shallow sea between the 2 km submarine bathymetry contour and the coastline. To aid in recognition, the present-day coastline is commonly shown on continents, but the coastline of the time concerned may be quite different, as well as being very difficult to pin down.

Record of Continental Separation in the Ocean-Floor

The ocean-floor generated during the break-up of Pangea and the creation of the Atlantic, India, Antarctic, and Arctic Oceans is still mostly preserved. Successively older oceanic transform faults, fracture zones, and ocean-floor magnetic anomalies and the continental edges themselves can be fitted together to give the Euler rotations needed to reposition all the major continents to a high degree of precision for most of Mesozoic and all of Cenozoic time, or roughly back to 200 Ma. The net results are paleocontinental reconstructions, that is, a display of the relative positions of the continents in the past (Smith et al. 1973).

Paleozoic and Precambrian Composites

For pre-Mesozoic time the only ocean-floor that remains is tectonically highly deformed. Thus Paleozoic and Precambrian global “maps” are made by geographically repositioning individual continents using paleomagnetism. The uncertainties in paleomagnetic poles are considerable, particularly for Precambrian time. Although the order in which the continents are arranged on the globe is known, at least for Paleozoic time, the longitude separations of adjacent continents are not. These separations are subjectively

estimated from evidence such as the distribution of fossils, tectonic effects, and the like. The results have been called composites, rather than maps, to emphasize the fact that they include qualitative data (Smith et al. 1973). The Precambrian has few fossils that provide an accurate indication of age and only scattered age determinations. Their absence, together with the absence of ocean-floor data and the uncertainties in pole positions, means that there are no generally agreed composites for most of Precambrian time. The Precambrian is still a veritable terra incognita.

Paleo-Plate Reconstructions

As noted above, a *global paleoreconstruction* is a reassembly of the major continents (and oceans) relative to one another at some time in the past. To turn this into a *paleo-plate reconstruction*, one needs to add the plate margins of the time concerned. For Jurassic and younger periods the former positions of ocean ridges are known from the ocean-floor record in the Atlantic, Indian, and Southern Oceans, but much of the Pacific ocean-floor has been subducted and for these areas former ridge positions have to be estimated. For Triassic and older periods the positions of former ocean ridges is largely a matter of informed guesswork (e.g., Stampfli and Borel 2002, 2004).

Orogenic belts, that is, those areas made up of deformed, metamorphosed, and igneous rocks, mark the former positions of convergent margins. However, the margins are commonly difficult to locate precisely and many orogenic belts include more than one former convergent margin, together with the locations of former oceans that have since been subducted. Thus although one can join former ridges to orogenic belts to make a sketch plate boundary map for the time concerned, this map is necessarily imprecise.

Precambrian *global paleoreconstructions* are generally schematic (e.g., Collins and Pisarevsky 2005). Because paleo-plate reconstructions build on an initial global reassembly, most Paleozoic and all Precambrian paleo-plate reconstructions are essentially cartoons that show the kinds of plate boundaries that may have existed at a given time but whose locations may have considerable errors.

Conclusions

The ocean-floor magnetic anomalies, transform faults, and fracture zones, together with the paleomagnetic frame allow very precise global Cenozoic and Mesozoic maps to be made. The inferred relative longitudes of Paleozoic and Precambrian continents depend on qualitative data, giving rise to poorly defined “composites.” Fossils enable precise dating of Paleozoic features that help in the construction of composites, but

such precision is not obtainable from Precambrian fossils, giving rise to considerable uncertainties in all Precambrian reconstructions. How to convert these global paleoreconstructions into verifiable paleo-plate reconstructions is a largely unsolved problem for pre-Mesozoic time.

Cross-References

- ▶ [Lithosphere, Continental](#)
- ▶ [Lithosphere, Oceanic](#)
- ▶ [Magnetic Anomalies: Interpretation](#)
- ▶ [Magnetic Methods, Surface](#)
- ▶ [Paleomagnetism, Principles](#)
- ▶ [Subduction Zones](#)

Bibliography

- Collins AS, Pisarevsky SA (2005) Amalgamating eastern Gondwana: the evolution of the Circum-Indian Orogens. *Earth-Sci Rev* 71:229–270
- Cox A, Hart RB (1986) *Plate tectonics: how it works*. Blackwell, Oxford
- Crosby AG, McKenzie D, Sclater JG (2006) The relationship between depth, age and gravity in the oceans. *Geophys J Int* 166:553–573
- Duncan RA (1981) Hotspots in the southern ocean – an absolute frame of reference for motions of the Gondwana continents. *Tectonophysics* 74:29–42
- Isacks B, Oliver J, Sykes LR (1968) Seismology and the new global tectonics. *J Geophys Res* 73:5855–5899
- Livermore RA, Vine FJ, Smith AG (1983) Plate motions and the geomagnetic field. 1: quaternary and late tertiary. *Geophys J R Astron Soc* 73:153–171
- McKenzie DP, Parker RL (1967) The north pacific: an example of tectonics on a sphere. *Nature* 216:1276–1280
- Morgan WJ (1968) Rises, trenches, great faults and crustal blocks. *J Geophys Res* 73:1959–1982
- Morgan WJ (1981) Hotspot tracks and the opening of the Atlantic and Indian Oceans. In: Emiliani C (ed) *The oceanic lithosphere*. Wiley, New York, pp 443–487
- Müller RD, Royer JY, Lawver LA (1993) Revised plate motions relative to the hotspots from combined Atlantic and Indian Ocean hotspot tracks. *Geology* 21:275–278
- Powell CM, Roots SR, Veevers SJ (1988) Pre-breakup continental extension in East Gondwanaland and the early opening of the eastern Indian Ocean. *Tectonophysics* 155:261–283
- Priestley K, McKenzie D (2006) The thermal structure of the lithosphere from shear wave velocities. *Earth Planet Sci Lett* 244:285–301
- Sawyer DS (1985) Total tectonic subsidence: a parameter for distinguishing crust type at the U.S. continental margin. *J Geophys Res* 90:7751–7769
- Shackleton RM (1988) Tectonic evolution of the Himalayas and Tibet. In: Shackleton RM, Dewey JF, Windley BF (eds) *Philosophical transactions of the Royal Society of London, series A: mathematical and physical sciences*. The Royal Society of London, London
- Smith AG, Briden JC, Drewry GE (1973) Phanerozoic world maps. In: Hughes NF (ed) *Organisms and continents through time, volume 12: special paper*. Palaeontological Association, London, pp 1–42
- Smith AG, Smith DG, Funnell BM (1994) *Atlas of Mesozoic and Cenozoic coastlines*. Cambridge University Press, Cambridge

- Stampfli GM, Borel GD (2002) A plate tectonic model for the Paleozoic and Mesozoic constrained by dynamic plate boundaries and restored synthetic oceanic isochrons. *Earth Planet Sci Lett* 196:17–33
- Stampfli GM, Borel GD (2004) The TRANSMED transects in space and time: constraints on the paleotectonic evolution of the Mediterranean domain. In: Cavazza W, Roure FM, Spakman W, Stampfli GM, Ziegler PA (eds) *The TRANSMED atlas – the Mediterranean region from crust to mantle*. Springer, Berlin, pp 53–80
- Vine FJ, Matthews DH (1963) Magnetic anomalies over oceanic ridges. *Nature* 199:947–949
- Wilson JT (1963) A possible origin of the Hawaiian Islands. *Can J Phys* 41:863–870

Poroelasticity

Ran Bachrach

Geophysics and Planetary Sciences Department, Tel Aviv University, Tel Aviv, Israel

Definition

Poroelasticity is a joint formulation for the behavior of a solid–fluid coupled porous system. Poroelasticity describes the behavior of porous continuum and pore fluid with respect to applied displacements and stresses. Poroelasticity jointly treats the solid frame and pore fluid to calculate changes in pore pressure, solid displacements, and fluid displacements due to stresses and displacements associated with external or internal processes.

The theory of poroelasticity is based on the work of (Terzaghi 1941; Biot 1941, 1956a, b, 1962; Gassmann 1951). Poroelasticity is often called Biot theory as the work of M. A. Biot is considered to establish the complete foundation for the theory.

Basic Concepts

Consider a connected network of pores fully saturated with fluid. The following are basic concepts that form the basis for the theory of poroelasticity.

Drained and Undrained Deformation

For a fully saturated porous material, we consider two fundamental states (or thought experiments): *Drained* and *undrained* (sometimes these are called *jacketed* and *unjacketed* conditions).

Drained conditions are achieved when pore pressure in a representative elementary volume (REV) is in equilibrium with the surrounding throughout the deformation of the saturated porous medium. A special case (described as a lab experiment) is when pore pressure and fluid can freely diffuse

in and out of the sample during the deformation process. Undrained (or jacketed) conditions refer to the condition in which the boundary of the porous solid (REV) is impermeable and pore fluid cannot exit the rock during the deformation process.

Physical measurements on both states (laboratory measurements) will determine the poroelastic coefficients of the rock. Often “dry conditions” are associated with the drained experiment. In poroelasticity the term “dry” elastic modulus refers to the drained experiment and not the elastic modulus when the material is dry.

Linear Stress–Strain Formulation for Poroelastic Media

Consider the displacement vector in the solid $\mathbf{u} = [u_x, u_y, u_z]$ and the average fluid displacement vector in the pore space $\mathbf{U} = [U_x, U_y, U_z]$.

The volume of fluid displaced through a unit area normal to the x, y, z , direction is $\phi \mathbf{U}$ and the fluid displacement field is $\mathbf{w}$. The isotropic poroelastic stress–strain relations that relate stress, σ , and pore pressure, P , to displacement and strain, ϵ , can be written in an abbreviated notation as:

$$\begin{aligned}\sigma_{ij} &= 2\mu e_{ij} + \delta_{ij}(\lambda_c \theta - \alpha M \zeta) \\ P &= -\alpha M \theta + M \zeta \\ \theta &= \nabla \cdot \mathbf{u}, \zeta = \phi \nabla \cdot (\mathbf{u} - \mathbf{U})\end{aligned}$$

where λ_c is the lame’s coefficient for the closed (undrained) system and μ is the shear moduli. Note that the shear strain and stress are not affected by pore pressure and fluid presence. M , α are the two poroelastic constants associated with the poroelastic stress–strain formulation. M is known as Biot modulus, or P wave modulus, and relates dry response to saturated response as follows:

$$\begin{aligned}\lambda_c &= \lambda_{\text{dry}} + \alpha^2 M, \\ \frac{1}{M} &= \frac{\alpha - \phi}{K_0} + \frac{\phi}{K_f}.\end{aligned}$$

where K_f is the pore fluid bulk modulus and ϕ is the porosity. α is the Biot parameter and is defined as:

$$\alpha = 1 - K_{\text{dry}}/K_0$$

where K_{dry} is the bulk modulus of the rock measured under drained conditions and K_0 is the bulk modulus of the frame mineral.

Note that when the system is jacketed, that is, $\zeta = 0$ so there is no relative displacement between the fluid and matrix, the pore pressure and the total stress are related and the bulk modulus is defined in terms of “closed” elastic properties.

When the matrix is open, the assumption is that there is a constant pore pressure in the system (pore pressure is in

equilibrium with the surroundings) and thus fluid can flow freely in and out of the system. One special case of such conditions is when the pore pressure is equal to zero. The term “dry” frame conditions often refers to the states of zero pore pressure.

The poroelastic linear stress–strain relation can be written in terms of pore pressure as the coupling variable rather than the relative fluid displacement (or divergence). A popular form of the linear isotropic poroelastic stress–strain relations is given by:

$$\begin{aligned}\sigma_{ij} + \alpha P \delta_{ij} &= 2\mu e_{ij} + \lambda_{\text{dry}} \theta \delta_{ij} \\ \zeta &= (1/M)P + \alpha \theta\end{aligned}$$

Note that the pressure direction here is assumed to be negative, for example, the pore pressure resists a positive confining stress that pushes the grains together.

The above equation can be written as an equivalent linear elasticity relation by substituting the total stress with Biot effective stress $\sigma_{ij}^* = \sigma_{ij} + \alpha P \delta_{ij}$, as:

$$\begin{aligned}\sigma_{ij}^* &= 2\mu e_{ij} + \lambda_{\text{dry}} \theta \delta_{ij} \\ \zeta &= (1/M)P + \alpha \theta\end{aligned}$$

In terms of the differential pressure (also known as Terzaghi’s effective stress or just the “effective” stress) defined as $\sigma'_{ij} = \sigma_{ij} + P \delta_{ij}$ the linear stress–strain relations are given by:

$$\begin{aligned}\sigma'_{ij} - (1 - \alpha)P \delta_{ij} &= 2\mu e_{ij} + \lambda_{\text{dry}} \theta \delta_{ij} \\ \zeta &= (1/M)P + \alpha \theta\end{aligned}$$

When $\alpha = 1$ (soft materials where $K_{\text{dry}} \ll K_0$) the Biot effective stress is equal to the differential stress (often called Tarahzagi’s effective stress). Simple derivation of Biot’s effective stress law is given by Nur and Byerlee (1971).

Gassmann’s Equation and Fluid Substitution

Gassman’s equation (Gassmann 1951) relates the elastic constant of saturated material in closed or undrained state to the drained or open conditions. In its isotropic form, Gassmann’s equation can be written as:

$$\begin{aligned}\frac{K_{\text{sat}}}{K_0 - K_{\text{sat}}} &= \frac{K_{\text{dry}}}{K_0 - K_{\text{dry}}} + \frac{K_{fl}}{\varphi(K_0 - K_{fl})} \\ \mu_{\text{sat}} &= \mu_{\text{dry}}\end{aligned}$$

where K_{sat} refers to the saturated bulk modulus of the closed system. The shear modulus for the open system (dry) and the closed system are the same $\mu_{\text{sat}} = \mu_{\text{dry}}$ as fluid does not

contribute to the shear stiffness of the system. Gassmann’s equation can be derived directly from Biot’s equations or from basic elasticity considerations by accounting for mass conservation of pore fluid in a closed system. Gassmann’s equation assumes that pore fluid is in equilibrium throughout the sample and that the pore fluid does not interact with the mineral frame. For more information see Mavko et al. (1998). Anisotropic extension to Gassmann’s equation is derived by Brown and Korrington (1975) in terms of the elastic compliance tensor:

$$S_{ijkl}^{\text{dry}} - S_{ijkl}^{\text{sat}} = \frac{(S_{ij\alpha\alpha}^{\text{dry}} - S_{ij\alpha\alpha}^0)(S_{kl\alpha\alpha}^{\text{dry}} - S_{kl\alpha\alpha}^0)}{(S_{\alpha\alpha\beta\beta}^{\text{dry}} - S_{\alpha\alpha\beta\beta}^0) + (\beta_{fl} - \beta_0)}$$

where S_{ijkl}^{dry} , S_{ijkl}^{sat} are the saturated and dry 4th order compliance tensors, β_{fl} is the fluid compressibility (or inverse bulk modulus), and β_0 is the compressibility of the mineral frame.

Fluid Substitution

One of the many practical applications of Gassmann’s equation is the ability to predict the elastic modulus of a known rock saturated with different fluids. One example is the prediction of seismic velocity of a gas-saturated sediment given the seismic velocity of the same sediment saturated with brine. This fluid substitution scheme is readily derived from Gassmann’s equation:

$$\begin{aligned}\frac{K_{\text{sat}1}}{K_0 - K_{\text{sat}1}} - \frac{K_{fl1}}{\varphi(K_0 - K_{fl1})} &= \frac{K_{\text{sat}2}}{K_0 - K_{\text{sat}2}} - \frac{K_{fl2}}{\varphi(K_0 - K_{fl2})} \\ \mu_{\text{sat}2} &= \mu_{\text{sat}1}\end{aligned}$$

Skempton Coefficient

If the system is closed (jacketed), the pore pressure and the external stress are directly related as:

$$\begin{aligned}\theta &= \frac{(\bar{\sigma} - \alpha P)}{K_{\text{dry}}} \rightarrow P = \frac{1}{\alpha} (\bar{\sigma} - K_{\text{dry}} \theta), \bar{\sigma} = \sigma_{kk}/3 \\ \theta &= e_{kk} = \frac{\bar{\sigma}}{K_{\text{sat}}} \rightarrow P = \frac{1}{\alpha} (K_{\text{sat}} - K_{\text{dry}}) \theta \rightarrow P = B \bar{\sigma}, \\ B &= \frac{1}{K_{\text{sat}}} \left(\frac{K_{\text{sat}} - K_{\text{dry}}}{\alpha} \right)\end{aligned}$$

B is known as the Skempton coefficient and is one of the most measured laboratory quantity.

Three-Dimensional Consolidation

Poroelasticity provides the analytical framework for three dimensional consolidation of soils and sediments as follows:

Darcy's law relates fluid motion to pore pressure gradient and gradient of the gravitational potential G and can be written as

$$\partial \mathbf{w} / \partial t = \frac{-\boldsymbol{\kappa}}{\eta} \nabla P - \frac{-\boldsymbol{\kappa}}{\eta} \nabla G$$

where $\boldsymbol{\kappa}$ is the permeability tensor and η is the fluid viscosity.

The total stress field must satisfy the equilibrium equations (neglecting the inertia terms)

$$\sum \frac{\partial \sigma_{ij}}{\partial x_j} + \rho b_i = 0,$$

where b_i is the body force acting on the media (Gravity) and will be neglected in the following expressions. Combining Darcy's law with the equilibrium equation, we obtain the general consolidation equations that couple fluid and solid displacements of the vectors $\mathbf{u}$ and $\mathbf{w}$ as:

$$2 \sum_{ij} \frac{\partial}{\partial x_j} (\mu e_{ij}) + \frac{\partial}{\partial x_j} (\lambda_c \theta - \alpha M \zeta) = 0$$

$$\frac{\partial \mathbf{w}}{\partial t} = \frac{\boldsymbol{\kappa}}{\eta} \nabla (\alpha M \theta - M \zeta)$$

for the special case where the poroelastic coefficients are constant, the equation of motion is written (after application of the divergence operator and some algebra) as:

$$(2\mu + \lambda_c) \nabla^2 \theta - \alpha M \nabla^2 \zeta = 0$$

$$\frac{\partial \zeta}{\partial t} = \frac{\boldsymbol{\kappa}}{\eta} M_C \nabla^2 \zeta$$

$$M_C = M(2\mu + \lambda) / (2\mu + \lambda_c)$$

The latter equation can be written in terms of pore pressure as the familiar diffusion equation with an added volumetric solid deformation term:

$$\frac{\partial P}{\partial t} - \frac{\boldsymbol{\kappa}}{\eta} M \nabla^2 P = -\alpha M \frac{\partial \theta}{\partial t}$$

This equation is used extensively in the study of soil consolidation, compacting reservoirs, and stress–fluid coupled processes in the earth.

Dynamic Poroelasticity and Wave Propagation in Saturated Porous Medium

The wave equation in saturated porous media is derived by adding the inertia terms to the equation of motion. For

isotropic media with isotropic permeability, it can be written in terms of fluid and solid displacement vectors:

$$2 \sum_{ij} \frac{\partial}{\partial x_j} (\mu e_{ij}) + \frac{\partial}{\partial x_i} (\lambda_c \theta - \alpha M \zeta) = \frac{\partial^2}{\partial t^2} (\rho u_i + \rho_f w_i)$$

$$\frac{\partial}{\partial x_i} (\alpha M \theta - M \zeta) = \frac{\partial^2}{\partial t^2} (\rho_f u_i + m w_i) + \frac{\eta}{k} \frac{\partial w_i}{\partial t}$$

where $m = (\varphi \rho_f + \rho_a) / \varphi^2$ is the mass coupling parameters, and $\rho_a = \varphi \rho_f (a - 1)$ is the apparent density that relates the dynamic tortuosity to fluid density and rock porosity (Berryman 1980).

The equations of motion can be written in terms of rotation-free and dilatation-free equations similar to the equation for P and S wave in linear elasticity. However, unlike linear elasticity, in isotropic linear poroelastic media, there are two P waves that travel with two different velocities and a single shear wave velocity. The two P wave velocities are known as the fast and slow Biot wave velocities and have been observed in the laboratory (Plona 1980). When the fluid viscosity is zero, the two P waves are not attenuated but when the fluid viscosity is finite the slow Biot wave is highly attenuated and quickly decays into pore pressure diffusion. In his 1956 papers, Biot discusses many details associated with wave propagation in the low frequency range, which is the appropriate range for seismic investigation, and the high frequency range, which covers the ultrasonic case.

Bibliography

- Berryman JG (1980) Conformation of Biot's theory. *Appl Phys Lett* 37:382–384
- Biot MA (1941) General theory of three dimensional consolidation. *J Appl Phys* 12:155–164
- Biot MA (1956a) Theory of propagation of elastic waves in fluid saturated porous solid 1. Low frequency range. *J Acoust Soc Am* 28:168–178
- Biot MA (1956b) Theory of propagation of elastic waves in fluid saturated porous solid 2. Higher frequency range. *J Acoust Soc Am* 28:179–191
- Biot MA (1962) Mechanics of deformation and acoustic propagation in porous media. *J Appl Phys* 33:1482–1498
- Brown R, Korringa J (1975) On the dependence of elastic properties of porous rock on the compressibility of pore fluid. *Geophysics* 40:608–616
- Gassmann F (1951) Über die Elastizität poröser Medien. *Vierteljahresschr Naturforsch Ges Zürich* 96:1–23
- Mavko G, Mukerji T, Dvorkin J (1998) The rock physics handbook. Cambridge University Press, New York
- Nur A, Byerlee JD (1971) An exact effective stress law for elastic deformation of rock with fluids. *J Geophys Res* 76:6414–6419
- Plona TJ (1980) Observation of a second bulk compressional wave in a porous medium at ultrasonic frequencies. *Appl Phys Lett* 36:259–261
- Terzaghi K (1941) Theoretical soil mechanics. Wiley, New York

Propagation of Elastic Waves: Fundamentals

Francisco J. Sánchez-Sesma¹ and Ursula Iturrarán-Viveros²

¹Instituto de Ingeniería, Universidad Nacional Autónoma de México, Ciudad Universitaria, México City, Mexico

²Facultad de Ciencias, Departamento de Matemáticas, Universidad Nacional Autónoma de México, Ciudad Universitaria, México City, Mexico

Definition

Dynamic elasticity allows predicting and explaining the propagation of perturbations in solids. The contributions of Newton, Hooke, Cauchy, and Navier laid the basis of dynamic elasticity by the beginnings of the nineteenth century. It is possible to solve the equation of motion in terms of displacements using potentials which are solutions of the wave equations for longitudinal (or compression) and transverse (or shear) waves. The waves that propagate in fluids like water and air are called acoustic waves. Plane waves are the simplest solutions of wave equation as discovered by D'Alembert. In conjunction with Fourier transforms, plane waves are central in the solutions of a variety of problems like waves in layered media and surface guided waves.

Summary

This entry deals with dynamic elasticity fundamentals. This theory is useful to explain the propagation of elastic waves in solids. Newton's second law allowed writing the equilibrium equation within an elastic medium. Cauchy developed the concept of stress and strain while Hooke's idea to relate linearly stress and strain led to Navier's equation which is in fact Newton's equation in terms of displacements. It can be solved using solutions of wave equations for P and S waves. Both static and dynamic elasticity were developed in the nineteenth century. However, the first synthetic seismogram was computed at the beginning of the twentieth century. Plane waves are the simplest solutions of wave equation and are a key component to study elastic wave propagation in layered media and surface waves. The Green's functions in time or frequency domain are the solutions for an impulsive or harmonic point load, respectively. The canonical case of an infinite homogeneous space is due to Stokes (1849) and is given here.

It has been recently discovered that the Green's function can be retrieved from averaging correlations of a diffuse field.

This applies for a very general medium. The relationship of energy and Green function allows establishing identities between average normalized measurements (correlations or autocorrelations) and intrinsic properties of the medium (imaginary part of Green function between two points or at the source itself) and opens the door to use noise measurements to construct virtual sources and do medium characterizations of unprecedented quality.

Introduction

Waves are part of our existence and manifest themselves at many scales and with an enormous variety of effects. Perhaps the most familiar waves could be those generated in a pond after a pebble is thrown. Other waves that we cannot see are the sound waves that propagate in the air. They allow both hearing and speech. Elastic waves are of the same type of acoustic waves but can be also of transverse nature as they can propagate in solids. In what follows we establish the fundamental equations and delineate solution strategies. Recent findings are pointed out.

Newton's Second Law

During the second half of seventeenth century Isaac Newton developed groundbreaking ideas of mechanics and optics; we can credit him the invention of calculus, although this has been a matter of bitter debate with Leibnitz and his followers. Regarding the motion of bodies, his descriptions enabled the industrial revolution. Consider a particle with mass m subjected to loads. Second Newton's Law establishes that:

$$\sum \mathbf{F} = m \mathbf{a}, \quad (1)$$

which means that the sum of forces equals the product of mass and acceleration. In this equation boldface letters represent vectors. For instance, the acceleration vector $\mathbf{a} = (a_1, a_2, a_3)^T$ has components a_j for the three directions x_j (with $j = 1, 2$, and 3). The upper index T means the transpose operation.

As velocity is the change of displacement with time, acceleration is the change of velocity with time. Therefore, acceleration is the second derivative of displacement $\mathbf{d}$ with respect to time and is written as:

$$\mathbf{a} = \frac{d^2 \mathbf{d}}{dt^2}. \quad (2)$$

Newtonian mechanics explained the motion of both celestial bodies and everyday objects. By means of the methods of calculus, developed significantly by Newton himself, these ideas were applied to deformable solids. For a successful

application, a relationship between forces and displacements, a constitutive law, is needed.

Cauchy's Stress and Strain

Assuming Cartesian coordinates and studying equilibrium within a continuum we can state Newton's second law per unit volume in modern form using both index notation and partial derivatives with respect to particular variables. In the index notation a repeated index implies summation. Thanks to the concept of stress tensor developed by Cauchy we can write the equations of equilibrium in the three Cartesian coordinates by means of:

$$\frac{\partial \sigma_{ji}}{\partial x_j} + f_i = \rho \frac{\partial^2 u_i}{\partial t^2}, \quad (3)$$

where $u_i(\mathbf{x}, t)$ = displacement vector of point $\mathbf{x}$ (represented by x_j , with $j = 1, 2$, and 3 , respectively, or by x, y , and z) and time t , $\sigma_{ji}(\mathbf{x}, t)$ = stress tensor components (force per unit area) on the faces of a particle oriented along the axis j (see Fig. 1), f_i = body force per unit volume, ρ = mass density, and the right-hand term is precisely the mass per unit volume multiplied by acceleration. Moreover, it is possible to show that in order to guarantee equilibrium on moments the stress tensor must be symmetric (i.e., $\sigma_{ji} = \sigma_{ij}$). The equilibrium at a given boundary subjected to tractions or forces per unit area leads to Cauchy equation $\sigma_{ij}n_j = t_i^{(n)}$ which relates the stress tensor at a point with traction associated to a given normal vector as depicted in Fig. 2.

The motion at point $\mathbf{x} + d\mathbf{x}$ allows for a first order description if it is within a small neighborhood of $\mathbf{x}$ and can be expressed in terms of values at $\mathbf{x}$ by means of:

$$\begin{aligned} u_i(\mathbf{x} + d\mathbf{x}, t) &= u_i(\mathbf{x}, t) + \frac{\partial u_i}{\partial x_k} dx_k \\ &= u_i(\mathbf{x}, t) + \frac{1}{2} \left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} \right) dx_k + \frac{1}{2} \\ &\quad \times \left(\frac{\partial u_i}{\partial x_k} - \frac{\partial u_k}{\partial x_i} \right) dx_k. \end{aligned} \quad (4)$$

The first and third terms of the last equality exhibit the motion at $\mathbf{x} + d\mathbf{x}$ as the composition of a rigid body translation and rotation, respectively, which do not produce any strains. The middle term really implies strains and is the product of the Cauchy infinitesimal strain tensor,

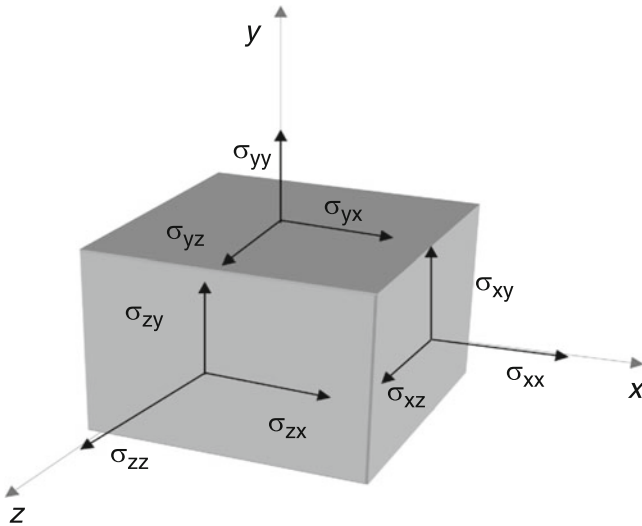
$$e_{ik} = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} \right), \quad (5)$$

and the spatial differentials dx_k . This tensor is a natural candidate for a constitutive law.

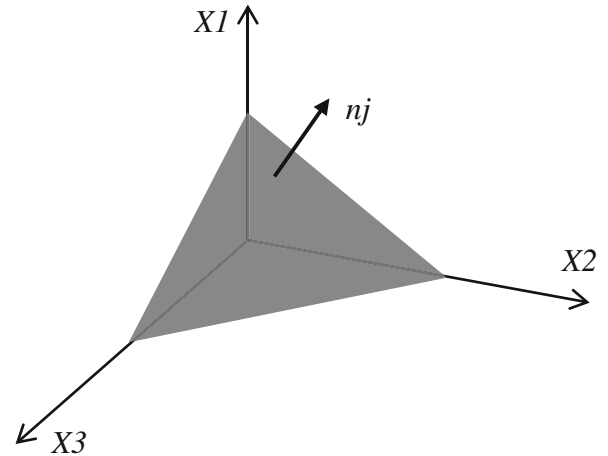
Hooke's Law

Solids undergo deformations when subjected to loads. If deformations are small, a linear relationship can be assumed. This is what Robert Hooke did when studying springs. To secure his finding he even wrote an anagram: *ceiinossttuv* to hide the Latin sentence: *ut tensio sic vis* which literally means that the displacement (*vis*) is proportional to the force (*tensio*).

Nowadays, we express Hooke's law simply by writing $F = k \times \delta$ where F = force, k = spring constant or stiffness, and δ = displacement. Hooke and Newton were contemporary, but not very friendly to each other. Hooke claimed Newton stole him the notion of gravitational force as proportional to inverse square of distance; Newton ignored him. Actually, such an idea was implicit in Kepler's laws and many suspected it, but only Newton gave a precise quantitative account.



Propagation of Elastic Waves: Fundamentals, Fig. 1 Stress components on the faces of a particle oriented along the axes x, y , and z



Propagation of Elastic Waves: Fundamentals, Fig. 2 Traction vector on a plane with normal n_j

For linear elastic materials, the more general constitutive law is

$$\sigma_{ij} = c_{ijpq} e_{pq}. \quad (6)$$

This is Hooke's law expressing a linear relationship between stresses σ_{ij} and strain e_{pq} at a point. Very likely, Hooke never imagined such an extension. Technically speaking these are second-order tensors. Both are due to Cauchy.

Elastic properties appear in the tensor c_{ijpq} which is of fourth order. At a given point, it represents 81 entries, but after recognizing the symmetry of the stress and strain tensors, we have 36 independent constants. Energy considerations allow having only 21 constants for a general anisotropic solid. If the medium is isotropic we have

$$c_{ijpq} = \lambda \delta_{ij} \delta_{pq} + \mu (\delta_{ip} \delta_{jq} + \delta_{iq} \delta_{jp}), \quad (7)$$

where δ_{ij} = Kronecker delta ($=1$ if $i=j$; $=0$ if $i \neq j$; note that $\delta_{jj} = 3$) and the parameters λ, μ = Lamé constants. Equation 7 shows that only two independent constants are needed to describe a linear isotropic solid.

Navier's Equation

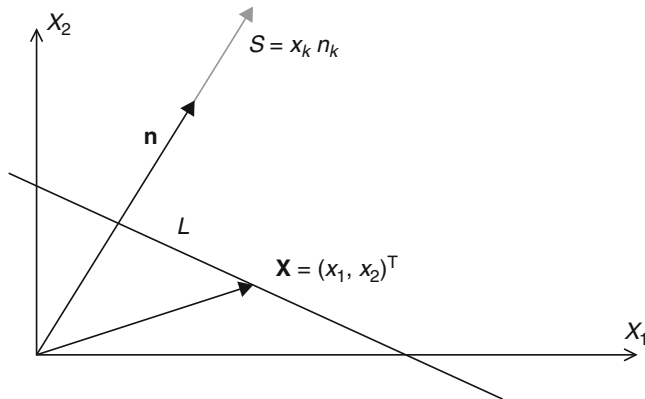
For isotropic materials, Hooke's law can be written as

$$\sigma_{ij} = \lambda e_{kk} \delta_{ij} + 2\mu e_{ij} \quad (8)$$

Substituting Eq. 8 in Eq. 3 and considering the definition of strain in Eq. 5, we can write the governing equation for motion within an elastic medium:

$$\mu \frac{\partial^2 u_i}{\partial x_j \partial x_j} + (\lambda + \mu) \frac{\partial^2 u_j}{\partial x_j \partial x_j} + f_i = \rho \frac{\partial^2 u_i}{\partial t^2}, \quad (9)$$

This is Navier's equation. It is possible to show that under very general circumstances the homogeneous solution (for $f_i = 0$) can be expressed by means of:



Propagation of Elastic Waves: Fundamentals, Fig. 3 Coordinate s along the vector $\mathbf{n}$. Line L corresponds to a constant value for s

$$u_i = \frac{\partial \phi}{\partial x_i} + \varepsilon_{ijk} \frac{\partial \psi_k}{\partial x_j} \text{ with } \frac{\partial \psi_k}{\partial x_k} = 0. \quad (10)$$

Here $\phi = \phi(\mathbf{x}, t)$ and $\psi_k = \psi_k(\mathbf{x}, t)$ are functions called potentials (see Achenbach 1973). The first is a scalar, while the second is a vector. These functions must be solutions of the homogeneous wave equations:

$$\frac{\partial^2 \phi}{\partial x_j \partial x_j} = \frac{1}{\alpha^2} \frac{\partial^2 \phi}{\partial t^2} \text{ and } \frac{\partial^2 \psi_k}{\partial x_j \partial x_j} = \frac{1}{\beta^2} \frac{\partial^2 \psi_k}{\partial t^2}, \quad (11)$$

respectively. This is supported by Lamé's theorem (for details see Aki and Richards 1980). Here $\alpha = \sqrt{(\lambda + 2\mu)/\rho}$ and $\beta = \sqrt{\mu/\rho}$ are the velocities of compressional (or P) and shear (or S) waves, respectively.

D'Alembert Solution of Wave Equation

For the linearly elastic, isotropic, homogeneous elastic medium, Navier's equation can be solved through the superposition of solutions of wave equation called potentials (see Aki and Richards 1980). Therefore, our capability to solve the canonical equation of dynamic elasticity rests upon the ability to solve the wave equation. This is subject of significant mathematical interest and many methods exist to look for solutions. Here we follow instead a quick, intuitive approach to solve the wave equation. In our opinion, the simplest solution is the one proposed by D'Alembert, a child of the French illustration of the eighteenth century. Extending his idea we can write

$$\phi(\mathbf{x}, t) = f\left(t - \frac{x_k n_k}{\alpha}\right) + g\left(t + \frac{x_k n_k}{\alpha}\right), \quad (12)$$

which is a solution of a wave equation for almost any function f and g , propagating along the unit vectors $\mathbf{n}_k$ and $-\mathbf{n}_k$, respectively. Equation 12 displays the beauty and power of D'Alembert solution. One powerful strategy is present in many methods and consists in sending explicitly plane waves in many directions.

This is what comes out after an exam of the arguments in Eq. 12. Consider that $x_k n_k = s$ is the equation of a plane in 3D space with a distance s from the origin, measured along $\mathbf{n}_k$. In vector notation we can write $\mathbf{x} \cdot \mathbf{n} = s$. This is the definition of a new coordinate s and is illustrated using the Cartesian system in 2D depicted in Fig. 3. The line L corresponds to a constant value for s .

Fourier Transforms and Plane Waves

In order to describe and generalize the power of the plane wave concept, consider the representation of a function

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \exp(+i\omega t) d\omega, \quad (13)$$

where $F(\omega)$ = Fourier transform of $f(t)$ given by

$$F(\omega) = \int_{-\infty}^{\infty} f(t) \exp(-i\omega t) dt. \quad (14)$$

Equation 13 exhibits an arbitrary function in terms of continuous superposition of sines and cosines. This fundamental property can be practically exploited under the assumptions that signals are discrete and periodic. We have then a powerful tool: the Discrete Fourier Transform. Moreover, the Fast Fourier Transform (FFT) algorithm allows for efficient computation. With this powerful tool we are effectively holding infinity in our hands. As the variable t in Eq. 13 is a dummy variable, we can substitute it by the D'Alembert argument $(t-s/c)$. Therefore, it is possible to write

$$\begin{aligned} f(t-s/c) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \exp[+i\omega(t-s/c)] d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \exp[-i\omega s/c] \exp[+i\omega t] d\omega. \end{aligned} \quad (15)$$

We first see D'Alembert argument at the complex exponential, but then the contribution of the plane wave is given by the transfer function that multiplies the Fourier transform of signal. Let us consider the following transfer function

$$\exp[-i\omega s/c] = \exp[-i(\omega/c)\mathbf{n} \cdot \mathbf{x}] = \exp[-i\mathbf{k} \cdot \mathbf{x}], \quad (16)$$

that represents a harmonic plane wave propagating and identify that the wave vector $\mathbf{k} = (\omega/c)\mathbf{n}$ that gives the direction of propagation of the plane wave, where $\mathbf{n}$ is a unit vector ($n_1^2 + n_2^2 + n_3^2 = 1$).

It is possible to extend the idea that the unit vector $\mathbf{n}$ may have complex values then we have inhomogeneous plane waves. As an example of the significant representation power of generalized plane waves, we can express a harmonic spherical wave as a superposition of both homogeneous and inhomogeneous plane waves using the Weyl integral (see Aki and Richards 1980):

$$\begin{aligned} \frac{1}{R} \exp\left(-i\omega \frac{R}{c}\right) &= \frac{1}{2\pi} \\ &\times \iint \frac{\exp(-ik_x x - ik_y y - \gamma|z|)}{\gamma} dk_x dk_y. \end{aligned} \quad (17)$$

Here $\gamma = \sqrt{\omega^2/c^2 - k_x^2 - k_y^2}$ with the sign chosen so that $\text{Re } \gamma > 0$ and $R = \sqrt{x^2 + y^2 + z^2}$. Plane waves are then building blocks to construct solutions for the half-space problems and layered media. They play a central role in seismology. Certainly, it is the success of a simple, yet profound idea.

Green's Function for Homogeneous Space

The Green's function is the displacement field produced within an elastic medium by a concentrated, impulsive unit load at a given point. For an infinite, homogeneous, isotropic, elastic medium Stokes found the analytical solution at the middle of the nineteenth century.

If the body force is a unit concentrated impulsive, it can be represented with Dirac's delta functions for space and time. Moreover, to specify the force direction, we add Kronecker delta δ_{ij} ($=1$ if $i=j$, $=0$ for $i \neq j$). Therefore, we can define $f_i \equiv \delta_{ij} \delta(|\mathbf{x} - \boldsymbol{\xi}|) \delta(t)$ and $u_i(\mathbf{x}, t) \equiv G_{ij}(\mathbf{x}, \boldsymbol{\xi}; t)$ = Green's function for the infinite elastic domain. It is the displacement in direction i at $\mathbf{x}$ when a unit impulse is applied in direction j at point $\boldsymbol{\xi}$.

The Green's function should satisfy Navier's equation, which can be written as

$$\begin{aligned} \mu \frac{\partial^2 G_{ij}}{\partial x_k \partial x_k} + (\lambda + \mu) \frac{\partial^2 G_{kj}}{\partial x_i \partial x_k} - \rho \frac{\partial^2 G_{ij}}{\partial t^2} \\ = -\delta_{ij} \delta(|\mathbf{x} - \boldsymbol{\xi}|) \delta(t). \end{aligned} \quad (18)$$

The solution of this problem is due to Stokes (1849) and can be expressed by means of

$$G_{ij}(\mathbf{x}, \boldsymbol{\xi}; t) = \frac{1}{4\pi\mu r} \{ f_1 \gamma_i \gamma_j + f_2 (\delta_{ij} - \gamma_i \gamma_j) \}, \quad (19)$$

where $r = |\mathbf{x} - \boldsymbol{\xi}| = \sqrt{(x_1 - \xi_1)^2 + (x_2 - \xi_2)^2 + (x_3 - \xi_3)^2}$ and $\gamma_j = (x_j - \xi_j)/r$ is the unit vector from $\boldsymbol{\xi}$ to $\mathbf{x}$. The functions f_1 and f_2 are given by

$$f_1(r, t) = \frac{\beta^2}{\alpha^2} \delta(t - r/\alpha) + 2\beta^2 \int_{1/\alpha}^{1/\beta} \delta(t - r\kappa) \kappa d\kappa, \quad (20)$$

and

$$f_2(r, t) = \delta(t - r/\beta) - \beta^2 \int_{1/\alpha}^{1/\beta} \delta(t - \kappa r) \kappa d\kappa. \quad (21)$$

We notice in Eq. 19 that Stokes' (1849) solution decays as $1/r$ and is modulated angularly by sines and cosines. This solution clearly shows up the emergence of both dilatational and distortional (P and S) spherical waves with their characteristic radiation patterns. Equations 20 and 21 show that both the dominant P and S pulses share the time dependence with the source. The S pulse has a larger amplitude. Between these P and S arrivals we find a disturbance decaying more rapidly away from the source.

In the frequency domain (after applying Fourier transform) these functions can be written as follows:

$$f_1(r, \omega) = (\beta^2/\alpha^2) [1 - i2/q\alpha - 2/q^2 r^2] \exp(-iqr) + [i2/k\alpha + 2/k^2 r^2] \exp(-ikr), \quad (22)$$

and

$$f_2(r, \omega) = (\beta^2/\alpha^2) [i/q\alpha + 1/q^2 r^2] \exp(-iqr) + [1 - i/k\alpha - 1/k^2 r^2] \exp(-ikr). \quad (23)$$

Here $q = \omega/\alpha$ and $k = \omega/\beta$ are the wave numbers of P and S waves, respectively. If ω equals zero, the static case, we obtain the constants $f_1 = 1$ and $f_2 = (1 + \beta^2/\alpha^2)/2$ that correspond to the Kelvin solution for a unit static load in the full elastic space. The Stokes solution for the infinite space is frequently used in the integral BEM or IBEM formulations (see Bouchon and Sánchez-Sesma 2007). Other Green's functions in two and three dimensions can be used to solve a variety of problems. A remarkable compendium of fundamental solutions is the one due to Kausel (2006).

Fifty-five years after the Stokes' (1849) solution, at the eve of the twentieth century, Lamb (1904) found the solution for an impulsive vertical point load at the surface of a half-space. This solution is based on integral representations, and the passage to time domain became premonitory of methods developed some decades later. A complete account of Lamb's problem and summary can be found in Kausel (2013) who pointed out that the first truly complete solutions to Lamb's problem were obtained by Pekeris (1955) and Chao (1960). They provided closed-form expressions for displacements engendered by a vertical and a horizontal load, respectively, but only when Poisson's ratio is $1/4$. Significant advances are due to Mooney (1974) and Richards (1979). Sundry numerical computations of the Green's function associated to Lamb's problem are due to Johnson (1974). Explicit formulae have been recently given for the displacement at the

surface of an elastic half-space for a buried point source in what can be called the 3D Lamb's problem (Feng and Zhang 2018). Their results are expressed in terms of elementary algebraic expressions and elliptic integrals as well and are validated against Johnson's (1974) solution.

For a homogeneous or layered half space, the classical Fourier methods allow obtaining the elastodynamic Green's function in integral form. Typically, the fields constructed in each layer by a set of plane waves with unknown coefficients that are found after boundary conditions are enforced. The strategies to solve this problem are diverse and we count among them the global matrix approach (Knopoff 1964), the Thomson-Haskell method (Haskell 1953), the stiffness matrix (Kausel 2006), and the Kennett reflection-transmission method (Kennett 2001, 2002). The computation of the displacement and stress fields requires integration along the horizontal wavenumber domain. Usually, in the discrete wavenumber method (DWN) the horizontal wavenumber is discretized and Fourier analysis permits to go to time domain (see Bouchon and Aki 1977; Bouchon 2003).

Theory allows solving the problem of the homogeneous boundary conditions in terms of an eigenvalue problem. This led to the astounding subject of surface waves (See Aki and Richards 2002).

Green's Function Retrieval from Correlations

The pioneering studies of Aki (1957) on seismic ambient noise and coda are the basis of significant developments. Various scattering formulations have been developed in order to explain coda features (see Aki and Chouet 1975; Sato and Fehler 1998). When multiple scattering takes place, the intensities, which are related to energy densities, follow diffusion-like equations. These waves arrive at the receiver from different directions sampling the medium along their paths.

It has been found that the imaginary part of the elastodynamic Green's function is proportional to the average cross correlations within a diffuse field (for recent reviews, see Weaver 2010 and Perton and Sánchez-Sesma 2016). A diffuse field develops within a system if there is isotropic, equipartitioned illumination. One simple way to achieve this is using a set of elastic plane waves that fulfill the Principle of Equipartition (EQP) of Energy (Weaver 1982; Sánchez-Sesma and Campillo 2006; Sánchez-Sesma et al. 2006; Weaver 2010). On the other hand, simple illumination from isolated sources may give rise to diffuse fields after multiple scattering takes place (see Hennino et al. 2001; Paul et al. 2005).

The connection between this deterministic problem, namely, the computation of elastic Green's function and the

diffuse field theory, has been pointed out by Sánchez-Sesma et al. (2011a) in the framework of the partition of the energy injected into a half-space by surface loads (see Miller and Pursey 1955; Weaver 1985). In recent works, the computation of Green's functions for 2D layered systems was achieved using a cocktail of equipartitioned homogeneous plane waves (P , SV , SH) and surface waves (Rayleigh and Love). Pertion and Sánchez-Sesma (2016) discussed the case of a layered medium, while Baena-Rivera et al. (2016) considered an alluvial basin.

In reality, for elastic inhomogeneous, anisotropic, irregular medium the role of multiple scattering should be invoked to assert that under very general circumstances the resulting seismic ambient field is very likely diffuse. Under these circumstances, the Green's function can be approximately retrieved from averaging cross correlations of the recorded motions of such diffuse field (e.g., Weaver and Lobkis 2004; Wapenaar 2004; Gouédard et al. 2008).

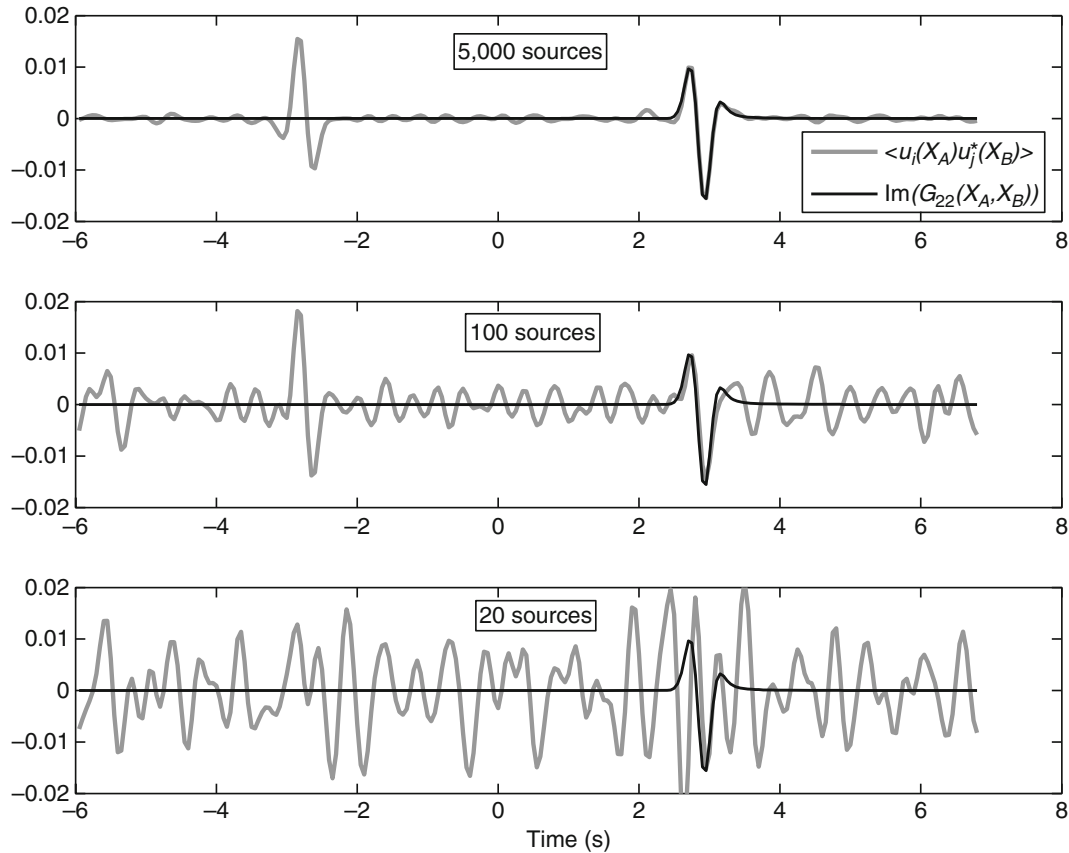
Seismic noise is produced by nearby, superficial low strength sources that we will generically call microtremors. Noise sources are 3D in nature and have significant geometrical spreading. This attenuation make that the relative sampling of the medium in depth is also much reduced.

In what follows, we review the Green's function retrieval from correlations of field fluctuations. For the layered medium with free surface overlaying a half-space, we compute the imaginary part of the Green function when source and receiver coincide at the free surface.

It has been demonstrated (e.g., Sánchez-Sesma et al. 2008) that if a 3D diffuse harmonic displacement vector field $u_i(\mathbf{x}, \omega)$ is established within an elastic medium, the average cross-correlations of such motions at points $\mathbf{x}_A$ and $\mathbf{x}_B$ can be written as:

$$\langle u_i(\mathbf{x}_A, \omega) u_j^*(\mathbf{x}_B, \omega) \rangle = -2\pi E_S k^{-3} \text{Im}[G_{ij}(\mathbf{x}_A, \mathbf{x}_B, \omega)]. \quad (24)$$

In this equation, the Green's function $G_{ij}(\mathbf{x}_A, \mathbf{x}_B, \omega)$ = displacement at $\mathbf{x}_A$ in direction i produced by a unit harmonic point force acting at $\mathbf{x}_B$ in direction j , ω = circular frequency, $k = \omega/\beta$ = shear wavenumber, β = shear wave propagation velocity, and $E_S = \rho\omega^2 S^2$ = average energy density of shear waves (ρ = mass density, S^2 = average spectral density). Note that the asterisk implies the complex conjugate and the angular brackets mean azimuthal average. Equation 1 is the



Propagation of Elastic Waves: Fundamentals, Fig. 4 Correlations are computed with 20, 100, and 5000 sources, for the scalar case. We can see that as the number of sources considered increases the retrieval, the G_{22} function is better achieved

analytical consequence of a correlation-type elastic representation theorem and has been verified recently in canonical examples of a full space (Sánchez-Sesma and Campillo 2006) and for an elastic inclusion embedded in a full space and for more general circumstances (Sánchez-Sesma et al. 2006, 2008). For the scalar case it is possible to compare the Green's function retrieval from correlations. In Fig. 4, we show how the G_{22} function is retrieved taking into account 20, 100, and 5000 sources. The correlations are computed with these numbers of sources, and we can see that as the number of sources considered increases the retrieval is better achieved.

Energy Densities at Given Points and Directions

When source and receiver are both at the same point, we have interesting consequences. For each component the autocorrelation is proportional to the directional energy density (DED) at a given point. On the other hand, the imaginary part of the Green function at the source is finite because the singularity of the Green's function is restricted to the real part. Therefore, the DED is proportional to the imaginary part of Green's function.

In what follows we compute the theoretical energy density at a given point $\mathbf{x}_A$. In order to do so we rewrite Eq. 24 considering that $\mathbf{x}_A = \mathbf{x}_B$:

$$\begin{aligned} E(\mathbf{x}_A) &= \rho\omega^2 \langle u_m(\mathbf{x}_A) u_m^*(\mathbf{x}_A) \rangle \\ &= -2\pi\mu E_S k^{-1} \times \text{Im}[G_{mm}(\mathbf{x}_A, \mathbf{x}_A)] \end{aligned} \quad (25)$$

where μ = shear modulus. The energy density of shear waves E_S is a measure of the strength of the diffuse illumination. We see that the total energy density at a point is proportional to the imaginary part of the trace of the Green tensor for coincident receiver and source.

The imaginary part is finite and regular and represents the rate of energy injected by the unit harmonic load at that point. This quantity “detects” the energy that goes back to the source-receiver and may be used to imaging. Equation 25 is valid even if the summation convention is ignored. In that case $E(\mathbf{x}_A) \equiv E_m(\mathbf{x}_A)$ and the energy density is associated to a particular direction (for discussions see Pertion et al. 2009). The connection of the imaginary part of the Green's function at the source with the optical theorem has been explored by Snieder et al. (2009).

To illustrate these ideas from theoretical point of view, let us consider again the Stokes' (1849) solution. After applying Fourier transform to Eqs. 20 and 21 and taking the limit $r \rightarrow 0$, i.e., making that source and receiver coincide, we can write that

$$\text{Im}[G_{ij}(\mathbf{x}, \mathbf{x})] = -\frac{\omega}{12\pi\rho} \left(\frac{1}{\alpha^3} + \frac{2}{\beta^3} \right) \delta_{ij}. \quad (26)$$

This shows that the power injected into the infinite elastic medium by the unit harmonic load grows linearly with frequency, as expected, it is an isotropic tensor and the energy is explicitly labeled as being related to P and S waves follows the principle of equipartition.

The relationships among energy densities and its partitions have been studied by Pertion et al. (2009) for a half-space and by Margerin et al. (2009) for a stratified medium. The Green's function may be useful to imaging the subsurface structure at the site.

For horizontally layered systems and 1D excitation, the relationship between reflection response and autocorrelation of surface motion was discovered by Claerbout (1968). An algorithm to identify reflection coefficients based upon Claerbout's results was proposed by Scherbaum (1987). In fact, this is a single-station inversion method, and, in principle, it allows imaging of the subsurface impedance structure. The idea was tested using small local recorded earthquakes.

The Horizontal-to-Vertical Spectral Ratio

This concept, the duality of diffuse field energy density – medium impulse response or Green's function, is being the engine of practical applications in exploration geophysics and engineering seismology. Indeed, the microtremor H/V spectral ratio (MHVSR) is frequently used to assess the dominant frequency of soil sites. It requires to measure ground motion produced by seismic ambient noise at a site and the processing is relatively simple. The diffuse field theory asserts that the average autocorrelations of motion components of a diffuse field at a given receiver give the directional energy densities (DEDs) which are proportional to the imaginary parts of the Green's function components when source, receiver, and direction coincide. Therefore, the MHVSR can be modeled in terms of the Green's function, and this opens the door to inversion of medium properties in layered systems (see Sánchez-Sesma et al. 2011b; García-Jerez et al. 2016; Sánchez-Sesma 2017; Kawase et al. 2011, 2015; Lontsi et al. 2015, 2019) and systems with lateral heterogeneity as well (Matsushima et al. 2014, 2017).

Concluding Remarks

The fundamentals of dynamic elasticity have been reviewed. The Newton's second law and the Hooke's elastic idea led to the Navier-Cauchy governing equations. D'Alembert seminal solution in conjunction with Fourier analysis led to powerful descriptions in terms of plane waves. Dynamic elasticity is a

science with almost two centuries of history. It is alive with significant impacts in seismology and engineering. It is central in the theory and applications of diffuse field concept.

Cross-References

- [Body Waves](#)
- [Poroelectricity](#)
- [Seismic Signals in Well Water Observations](#)
- [Seismic, Ambient Noise Correlation](#)
- [Seismogram Interpretation](#)
- [Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting](#)
- [Statistical Seismology](#)
- [Surface Waves](#)

Acknowledgments Thanks are given to E. Kausel for his keen and constructive remarks. This work was partially supported by DGAPA-UNAM Projects IN100917, IN121709, and CONACYT Mexico under PROINNOVA projects 251910 and 241763.

Bibliography

- Achenbach JD (1973) Wave propagation in elastic solids. North-Holland Publishing Co, Amsterdam/New York/Oxford
- Aki K (1957) Space and time spectra of stationary stochastic waves with special reference to microtremors. *Bull Earthq Res Inst* 35:415–456
- Aki K, Chouet B (1975) Origin of coda waves: source, attenuation and scattering effects. *J Geophys Res* 80:3322–3342
- Aki K, Richards PG (1980) Quantitative seismology. Theory and methods. W. H. Freeman, San Francisco
- Aki K, Richards PG (2002) Quantitative seismology. Sausalito, Calif: University Science Books
- Baena-Rivera M, Perton M, y Sánchez-Sesma FJ (2016) Surface waves retrieval from generalized diffuse fields in 2d synthetic models of alluvial valleys. *Bulletin of the Seismological Society of America* 106(6)
- Bouchon M (2003) A review of the discrete wavenumber method. *Pure Appl Geophys* 160:445–465
- Bouchon M, Aki K (1977) Discrete wave number representation of seismic source wave fields. *Bull Seismol Soc Am* 67:259–277
- Bouchon M, Sánchez-Sesma FJ (2007) Boundary integral equations and boundary elements methods in Elastodynamics. In: Wu R-S, Maupin V, Dmowska R (eds) *Advances in wave propagation in heterogeneous earth. Advances in geophysics*, vol 48. Elsevier-Academic Press, New York/ Boston, pp 157–189
- Chao CC (1960) Dynamical response of an elastic half-space to tangential surface loadings. *J Appl Mech* 27:559–567
- Claerbout JF (1968) Synthesis of a layered medium from its acoustic transmission response. *Geophysics* 33:264–269
- Feng X, Zhang H (2018) Exact closed-form solutions for Lamb's problem. *Geophys J Int* 214:444–459. <https://doi.org/10.1093/gji/ggy132>
- García-Jerez A, Piña-Flores J, Sánchez-Sesma FJ, Luzón F, Perton M (2016) A computer code for forward computation and inversion of the H/V spectral ratio under the diffuse field assumption. *Comput Geosci* 97:67–78
- Gouédard P, Stehly L, Brenguier F, Campillo M, Colin de Verdière Y, Larose E, Margerin L, Roux P, Sánchez-Sesma FJ, Shapiro NM, Weaver RL (2008) Cross-correlation of random fields: mathematical approach and applications. *Geophys Prospect* 56:375–393
- Haskell NA (1953) The dispersion of surface waves in multilayered media. *Bull Seismol Soc Am* 43:17–34
- Hennino R, Trégourès N, Shapiro NM, Margerin L, Campillo M, van Tiggelen B, Weaver RL (2001) Observation of equipartition of seismic waves in Mexico. *Phys Rev Lett* 86:3447–3450
- Johnson LR (1974) Green's function for Lamb's problem. *Geophys J Int* 37:99–131. <https://doi.org/10.1111/j.1365-246X.1974.tb02446.x>
- Kausel E (2006) *Fundamental solutions in elastodynamics. A compendium*. Cambridge University Press, New York
- Kausel E (2013) Lamb's problem at its simplest. *Proc R Soc A* 469:20120462. <https://doi.org/10.1098/rspa.2012.0462>
- Kawase H, Sánchez-Sesma FJ, Matsushima S (2011) The optimal use of horizontal-to-vertical (H/V) spectral ratios of earthquake motions for velocity structure inversions based on diffuse field theory for plane waves. *Bull Seismol Soc Am* 101:2001–2014
- Kawase H, Matsushima S, Satoh T, Sánchez-Sesma FJ (2015) Applicability of theoretical horizontal-to-vertical ratio of microtremors based on the diffuse field concept to previously observed data. *Bull Seismol Soc Am* 105:3092–3103. <https://doi.org/10.1785/0120150134>
- Kennett BLN (2001) *The seismic wavefield, volume I: introduction and theoretical development*. Cambridge University Press, Cambridge, UK
- Kennett BLN (2002) *The seismic wavefield, volume II: interpretation of seismograms on regional and global scales*. Cambridge University Press, Cambridge, UK
- Knopoff L (1964) A matrix method for elastic wave problems. *Bull Seismol Soc Am* 54:431–438
- Lamb H (1904) On the propagation of tremors over the surface of an elastic solid. *Philos Trans R Soc London A* 203:1–42
- Lontsi AM, Sánchez-Sesma FJ, Molina-Villegas JC, Ohrnberger M, Krüger F (2015) Full microtremor H/V(z,f) inversion for shallow subsurface characterization. *Geophys J Int* 202:298–312. <https://doi.org/10.1093/gji/ggv132>. *GJI Seismology*
- Lontsi AM, García-Jerez A, Molina-Villegas JC, Sánchez-Sesma FJ, Molkenthin C, Ohrnberger M, Krüger F, Wang R, Fäh D (2019) A generalized theory for full microtremor horizontal-to-vertical [H/V(z,f)] spectral ratio interpretation in offshore and onshore environments. *Geophys J Int* 218(2):1276–1297. <https://doi.org/10.1093/gji/ggz223>
- Margerin L, Campillo M, van Tiggelen BA, Hennino R (2009) Energy partition of seismic coda waves in layered media: theory and application to Pinyon Flats Observatory. *Geophys J Int* 177:571–585
- Matsushima S, Hirokawa T, de Martin F, Kawase H, Sánchez-Sesma FJ (2014) The effect of lateral heterogeneity on horizontal-to-vertical spectral ratio of microtremors inferred from observation and synthetics. *Bull Seismol Soc Am* 104:381–393
- Matsushima S, Kosaka H, Kawase H (2017) Directionally dependent horizontal-to-vertical spectral ratios of microtremors at Onahama, Fukushima, Japan. *Earth Planets Space* 69:96. <https://doi.org/10.1186/s40623-017-0680-9>
- Miller GF, Pursey H (1955) On the partition of energy between elastic waves in a semi-infinite solid. *Proc R Soc Lond A Math* 233:55–69
- Mooney HM (1974) Some numerical solutions for Lamb's problem. *Bull Seismol Soc Am* 64:473–491
- Paul A, Campillo M, Margerin L, Larose E, Derode A (2005) Empirical synthesis of time-asymmetrical Green function from the correlation of coda waves. *J Geophys Res* 110. <https://doi.org/10.1039/2004JB003521>
- Pekeris CL (1955) The seismic surface pulse. *Proc Natl Acad Scie USA* 41:469–480. <https://doi.org/10.1073/pnas.41.7.469>
- Perton M, Sánchez-Sesma FJ (2016) Green's function calculation from equipartition theorem. *J Acoust Soc Am* 140:1309–1318

- Perton M, Sánchez-Sesma FJ, Rodríguez-Castellanos A, Campillo M, Weaver RL (2009) Two perspectives on equipartition in diffuse elastic fields in three dimensions. *J Acoust Soc Am* 126:1125–1130. <https://doi.org/10.1121/1.3177262>
- Richards PG (1979) Elementary solutions to Lamb's problem for a point source and their relevance to three-dimensional studies of spontaneous crack propagation. *Bull Seismol Soc Am* 69:947–956
- Sánchez-Sesma FJ (2017) Modeling and inversion of the microtremor H/V spectral ratio: physical basis behind the diffuse field approach. *Earth Planets Space* 69:92. <https://doi.org/10.1186/s40623-017-0667-6>
- Sánchez-Sesma FJ, Campillo M (2006) Retrieval of the Green function from cross-correlation: the canonical elastic problem. *Bull Seismol Soc Am* 96:1182–1191
- Sánchez-Sesma FJ, Pérez-Ruiz JA, Campillo M, Luzón F (2006) The elastodynamic 2D Green function retrieval from cross-correlation: the canonical inclusion problem. *Geophys Res Lett* 33:L13305. <https://doi.org/10.1029/2006GL026454>
- Sánchez-Sesma FJ, Pérez-Ruiz JA, Luzón F, Campillo M, Rodríguez-Castellanos A (2008) Diffuse fields in dynamic elasticity. *Wave Motion* 45:641–654
- Sánchez-Sesma FJ, Weaver RL, Kawase H, Matsushima S, Luzón F, Campillo M (2011a) Energy partitions among elastic waves for dynamic surface loads in a semi-infinite solid. *Bull Seismol Soc Am* 101:1704–1709
- Sánchez-Sesma FJ, Rodríguez M, Iturrarán-Viveros U, Luzón F, Campillo M, Margerin L, García-Jerez A, Suárez M, Santoyo MA, Rodríguez-Castellanos A (2011b) A theory for microtremor H/V spectral ratio: application for a layered medium. *Geophys J Int* 186:221–225
- Sato H, Fehler M (1998) Wave propagation and scattering in the heterogeneous earth. Springer, New York
- Scherbaum F (1987) Seismic imaging of the site response using microearthquake recordings. Part 1. Method. *Bull Seismol Soc Am* 77:1905–1923
- Snieder R, Sánchez-Sesma FJ, Wapenaar K (2009) Field fluctuations, imaging with backscattered waves, a generalized energy theorem, and the optical theorem. *SIAM J Imag Sci* 2:763–776
- Stokes GG (1849) On the dynamical theory of diffraction. *Cambridge Phil Soc Trans* 9:1–62
- Wapenaar K (2004) Retrieving the elastodynamic Green's function of an arbitrary inhomogeneous medium by cross correlation. *Phys Rev Lett* 93:254301-1-4
- Weaver RL (1982) On diffuse waves in solid media. *J Acoust Soc Am* 71:1608–1609
- Weaver RL (1985) Diffuse elastic waves at a free surface. *J Acoust Soc Am* 78:131–136
- Weaver RL (2010) Equipartition and retrieval of Green's function. *Earthq Sci* 23:1–6
- Weaver RL, Lobkis OI (2004) Diffuse fields in open systems and the emergence of the Green's function. *J Acoust Soc Am* 116:2731–2734

Radioactivity in Earth's Core

V. Rama Murthy
(Deceased)

Definition

Lithophile	Affinity for silicates
Chalcophile	Affinity for sulfur
Siderophile	Affinity for metals
BSE	Bulk Silicate Earth refers to primitive silicate material in the Earth from which the core is separated. Corresponds to the total mantle + crust system of silicates now.
P, T	Pressure, temperature
TW	Terawatt. $1 \text{ TW} = 10^{12} \text{ W}$
D_K	Concentration of K in Fe-alloy/ Concentration of K in silicate melt.
GPa	GigaPascal $= 10^9$ Pascals (1 GigaPascal is equivalent to 10 kilobars of pressure)
Geoneutrinos	Electron antineutrinos produced inside the Earth due to β^- decay of naturally occurring radioactive elements in the Earth.
Oxygen fugacity, f_{O_2}	A measure of the oxidation state of a system irrespective of the presence or absence of a gas phase containing free oxygen.
IW buffer	Iron–Wustite buffer. A synthetic redox mineral reference buffer representing the oxygen fugacity of a system where Fe_{metal} is in equilibrium with FeO (wustite), according to the reaction $\text{Fe}_{\text{metal}} + \frac{1}{2}\text{O}_2 = \text{FeO}$ (wustite). The IW buffer corresponds to $\log f_{O_2} = -12.5$ at $1,200^\circ\text{C}$.

Introduction

The Earth is a thermal engine driven by its internal heat. The bulk of this heat is produced by the elements potassium (K), uranium (U), and thorium (Th) through the radioactive decay of their isotopes ^{40}K , ^{235}U , ^{238}U , and ^{232}Th . Table 1 lists the atomic percentages, radioactive decay constants, half-life, and heat production characteristics of these radioactive isotopes.

Because K, U, and Th are strongly lithophile, differentiation of the Earth into a metallic core and a silicate primitive mantle (BSE) will have partitioned them overwhelmingly into the BSE. A long-standing convention is that the radioactivity of the BSE represents the total radioactivity of the Earth and that the metallic core has no radioactivity (e.g., McDonough 1999, 2003). However, radioactivity in the core has come into focus due to recent developments in three areas: (1) modern high P, T experiments on the metal-silicate partitioning behavior of K and U; (2) theoretical and experimental advances in our understanding of the electronic structure of potassium under conditions of high pressure and temperature; and (3) thermal models of the core that satisfy the size and age of the inner core and the energy to sustain geomagnetic field. Also, it is known that the heat due to radioactivity in BSE, about 19 TW (e.g., compiled in Lodders and Fegley 1998), falls short of the surface heat flux estimated at 30–44 TW (Lee 1970; Hofmeister and Criss 2005; Pollack et al. 1993). This shortfall is suggestive of either a higher than BSE radioactive content of the mantle and/or radioactivity in the core (see Murthy 2006). In the discussion below, new experimental developments are emphasized, citing where necessary the models and calculations that bear upon the radioactivity in the core.

Radioactivity in Earth's Core, Table 1 Important heat-producing radioactive isotopes in the Earth: atomic percentages, decay constants, half-lives, and heat production values. (After Van Schmus 1995)

Isotope	Atomic percentage	Decay constant λ (year ⁻¹)	Half-life (years)	Specific isotopic heat production (μWkg^{-1})
⁴⁰ K	0.01167	5.54×10^{-10}	1.251×10^9	29.17
²³⁵ U	0.7200	9.85×10^{-10}	7.038×10^8	568.7
²³⁸ U	99.2743	1.551×10^{-10}	4.468×10^9	94.65
²³² Th	100.00	4.95×10^{-10}	1.401×10^{10}	26.38

Potassium Radioactivity in the Earth's Core

Potassium radioactivity in the Earth's core has been a topic of discussion for nearly four decades. Purely on geochemical grounds, Hall and Murthy (1971) and Lewis (1971) proposed that K enters the metallic core of the Earth as a consequence of the presence of sulfur in the core (Murthy and Hall 1970). Since then, the implications of K in the Earth's core for the energetics of the core, convection and core cooling rate, growth rate of the inner core, the geomagnetic field, and thermal evolution of the Earth have been noted in several studies (e.g., Verhoogen 1973; Goettel 1976; Gubbins et al. 1979; Stevenson et al. 1983). Early experiments, however, concluded against the presence of K in the core (Oversby and Ringwood 1972; Ganguly and Kennedy 1977; Murrell and Burnett 1986). These low pressure (1 bar to -1.5 GPa) and low temperature (1,030–1,350 °C) studies yielded very low distribution coefficients, D_K for potassium. The laboratory D_K values applied to core formation led to the conclusion that K cannot be present in an S-bearing core.

Since these first-generation experiments, we have made a substantial progress in understanding the partitioning behavior of elements. For example, we know now that the partitioning of an element between metal and silicate depends on a number of variables such as pressure, temperature, oxygen fugacity (fO_2), and the composition of both the metal and silicate melts (Murthy 1991; Hillgren et al. 1996; Jana and Walker 1997; Richter et al. 1997; Gessmann and Rubie 1998, 2000; Gessmann et al. 1999; Li and Agee 2001; Bouhifd and Jephcoat 2003; Chabot et al. 2005). In addition, core formation models suggest metal-silicate partitioning at a pressure range of 30–60 GPa and a temperature of 3,000–4,000 °C (e.g., see Rubie et al. 2003 and Chabot et al. 2005 and the discussion and references cited therein). In the context of these developments, several new experiments have investigated the possibility of K-radioactivity in the core.

Chabot and Drake (1999) made the first systematic study of the effect of some of the variables on K partition into metal alloys at 1.5 GPa and 1,900 °C. They showed that D_K was dependent on the S content of the metal alloy and on the composition of the silicate melt. Their experiments yielded low values of D_K ($\sim 6 \times 10^{-3}$) leading to their conclusion that the K content in an S-bearing metallic core will be trivially small (< 1 ppm). The D_K values obtained by Chabot and Drake (1999) are lower by an order of magnitude and at

variance with a number of later studies. This discrepancy may be due to their capsule choice (Gessman and Wood 2002) or due to other experimental artifacts caused by the extreme lability of K under experimental conditions (Murrell and Burnett 1986; Murthy et al. 2003).

Studies of K partitioning at high pressure and temperature (2.5–24 GPa and 1,500–1,900 °C) by Gessman and Wood (2002) found a significant entry of K into Fe-S-O alloys. K partition into the metallic liquids depended strongly on S and O contents of metal alloy and increased with temperature. K had little tendency to partition into S-free metal alloys, as was noted in many previous studies. They show clearly that the partitioning behavior of K is strongly influenced by anion/cation ratio of the Fe-sulfide liquid as well as the depolymerization level of the silicate liquid noted previously by Chabot and Drake (1999). Evaluating the effect of these and other variables on D_K , Gessman and Wood (2002) suggest the possibility of 100 ppm of K content in a core containing about 10% S and possibly up to a maximum of ~ 250 ppm for conditions relevant to core separation from a deep magma ocean.

Murthy et al. (2003) found a significant partition of K into Fe–FeS liquids in experiments at 1–3 GPa at temperatures above the liquidus of both metal and silicate phases (1,200–1,800 °C) and at $fO_2 \sim 1.5$ log units below IW buffer, corresponding to core separation conditions in a magma ocean (Walter and Thibault 1995; Chabot and Agee 2003; Chabot et al. 2005; Wade and Wood 2005). D_K seemed positively correlated with temperature as noted in previous studies but not correlated with pressure in the pressure range of the experiments. Extrapolating to deep magma ocean conditions and a core with $\sim 10\%$ S, the inferred K content is 60–130 ppm, a little lower but comparable to that inferred by Gessman and Wood (2002).

Bouhifd et al. (2007) studied the partition of K between silicates and FeS-rich alloy and pure Fe metal at 5–15 GPa and a temperature of 1,900 °C at an fO_2 between 1.5 and 3 log units below IW buffer. D_K was clearly positively correlated to S and O contents of the metal alloy and temperature as observed in previous experiments. In the pressure range of their experiments, no pressure dependence of D_K was observed both in sulfur-free and sulfur-bearing metal alloys. A parameterization of all existing D_K data showed a linear relationship between D_K and temperature, a relationship useful to predict D_K values at plausible

temperature relevant to core formation if the S and O contents of the core are known. For reasonable assumptions of S and O contents, Bouhifd et al. (2007) found K content of ~250 ppm in the core. If the core is free of S and O, the K content falls as low as ~25 ppm.

Partitioning experiments between peridotite silicate liquids and Fe-Ni-S-C-O melts at pressures of 1.0–7.7 GPa and temperatures of 1,650–2,200 °C by Corgne et al. (2007) did not find any relationship between D_K and temperature, as observed by Gessman and Wood (2002), Murthy et al. (2003), and Bouhifd et al. (2007). Nor did they find any dependence of D_K on the S content of metal alloy, as found in Chabot and Drake (1999), Gessman and Wood (2002), Murthy et al. (2003) and Bouhifd et al. (2007). A compositional control on D_K by the O content of the metal alloy was observed leading to their conclusion that a core containing S, C, and about 2 wt.% O could only have up to 25 ppm of K irrespective of the S and C levels. For a possible O content of ~5 wt.% (Badro et al. 2007), the K content of the core is ~80 ppm. However, the discrepancy between this work and the other recent studies is unresolved at present.

Quantum mechanical calculations predict a change in electronic structure of K at high pressures (from 4d-like orbitals to 3s-like orbitals), making K behave like a transition element (Bukowinski 1976; Lee et al. 2004). This prediction is experimentally confirmed by the work of Lee and Jeanloz (2003) that showed the alloying of K with a high-pressure polymorph of Fe at pressures >26 GPa and temperature 2,500 K. The change in chemical bonding behavior of K under pressure is noted in several other experiments (Ito et al. 1993; Parker et al. 1996; Hirao et al. 2006; Bouhifd et al. 2007). These experiments and calculations show that K entry into a Fe-metal core is facilitated by high temperatures and high pressures, making it siderophile at extreme conditions.

Radioactivity in the core is suggested in a number of studies concerned with the age and size of the inner core, and the energy required to sustain the geomagnetic field. The estimated core–mantle boundary (CMB) heat flux is 6–12 TW (Buffett 2003) and controls the rate of core cooling. Recent studies (Buffett 2003; Nimmo et al. 2004; Costin and Butler 2006) have examined the question of how best to reconcile the CMB heat flux with the size of inner core, the ~3.5 Ga history of the magnetic field, and the CMB temperature. These authors note that the presence of K in the core at ~200–400 ppm is in best accord with the present size of the inner core and the power needs of a geomagnetic dynamo. Using somewhat different parameters, similar conclusions have been reached by others (for example, Labrosse 2003; Buffett 2003; Butler et al. 2005; Costin and Butler 2006). Thus, it appears that a variety of thermal models of the core call for an additional heat source in the core. In view of the

recent experimental data, it is reasonable to attribute this at least in part to the radioactivity of K. However, no precise estimates are possible at present. A more definitive evaluation of the influence of temperature, pressure, and other variables on K partitioning into metal alloys is needed. There is also an uncertainty of the mantle K content at the time of core separation (Murthy 2006) although the convention is to use the model BSE content of ~240 ppm. Nor are the S and O contents of the core known. Given these caveats, a core K content of up to 250 ppm of K seems tenable. The corresponding radioactive heat production in the core, especially with any accompanying U and Th (see discussion below) would be ~2 TW, a significant fraction of the CMB heat flux. However, the uncertainty of these figures cannot be overemphasized. Table 2 summarizes the current estimates of K-radioactivity in the core from standard geochemical models, thermal modeling of the core, and recent high P, T partition experiments.

Radioactivity in Earth’s Core, Table 2 A comparison of inferred values of potassium abundance in the core from geochemical models, theoretical calculations, and recent experiments, and the corresponding heat production in terawatts (TW = 10¹² W) today

Method	K abundance in core (ppm)	Heat production in core (TW)	References
Geochemical models			
BSE models	0	0	McDonough (1999, 2003)
E-chondrite	550 ± 260	~4–5	Lodders (1995)
Geophysical and	200–400	–	Buffett (2003)
Geodynamo models	250–750	–	Labrosse (2003)
	Up to 1,420	9	Roberts et al. (2003)
	400	~3	Nimmo et al. (2004)
	300	–	Costin and Butler (2006)
Experiments			
Sulfur-bearing alloys	<1	0.01	Chabot and Drake (1999)
	100–250	0.8–2.0	Gessman and Wood (2002)
	60–130	0.4–0.8	Murthy et al. (2003)
	~250	~1.5	Bouhifd et al. (2007)
	~25–80	0.2–0.5	Corgne et al. (2007)
Fe-metal			
	~25	0.17	Bouhifd et al. (2007)
	~7,000 ^a	–	Lee and Jeanloz (2003)
	35	0.23	Hirao et al. (2006)

^aFor realistic conditions of core segregation, K in core will be much less (see discussion in Lee and Jeanloz 2003; Lee et al. 2004)

U (and Th) Radioactivity in the Earth's Core

There have been some attempts to evaluate the presence of U (and Th) in the core. Earlier experiments were largely at low pressure and low temperature. Furst et al. (1982) noted that under the highly reduced conditions relevant to core formation in the Earth, U and Th tend to become chalcophile (affinity for sulfur) and might enter a sulfur-bearing core. The chalcophile tendency of U under these conditions is confirmed by Murrell and Burnett (1986) and in the recent high P, T experiments described later. Feber et al. (1984) noted the significant solubility of UO_2 in impure Fe-metal at 1 bar pressure and at temperature $> 3,000$ K suggesting that significant U could be present in the core. In a series of detailed experiments, Murrell and Burnett (1986) measured the partitioning of K, U, and Th between silicate and sulfide liquids at low pressures (1 bar to 1.5 GPa) and found that U and Th tend to be less lithophile under reducing conditions and are markedly chalcophile and partition at greater levels than K into Fe–FeS liquids.

The high P, T partitioning experiments of U described below yield somewhat ambiguous results. Bao et al. (2005, 2006) investigated the partitioning of U between silicate melt and pure Fe liquids at pressures of 3.0–14.5 GPa and temperatures of 1,660–2,500 °C. They found that partition of U into the metal is dependent on whether the silicates were molten or solid and increases with both P and T. They found a similar P and T dependence of U in Fe-10 wt% S and Fe-35 wt% S liquids, with a positive dependence on S content of the metal phase. Inferred values for U content in core are in the range of 1–4 ppb for pure Fe and at least 10 ppb for 10 wt% S in core. At these latter levels, U alone can produce 1–2 TW of heat energy. Together with K, radioactivity in the core can then account for a significant fraction of the CMB heat flux. In contrast to these observations, Wheeler et al. (2006) found that U entry into core is not supported in their experiments for conditions up to 10 GPa, 2,300 °C and 28 wt% S at $f\text{O}_2$ about 2 log units below IW. The discrepancy between these two studies is not understood.

Malavergne et al. (2007) found that U partitioning into core metal is dependent on S content and highly reducing conditions ($f\text{O}_2$ about 4–5 log units below IW buffer) and at temperatures up to 2,400 °C and 20 GPa. These experiments permit a range of 0.0003–0.63 ppb U in core depending on whether the metal is S free or S rich. In experiments at temperatures of 2,200–2,400 °C and pressures up to 8 GPa, under very low oxygen fugacity (~ 4 log units below IW), Murthy et al. (2007) found U partitioning into metal to be positively correlated with S content. The data permit U concentration in the core of 1–2 ppb with a heat production of 0.1–0.3 TW.

In a series of papers, Herndon (see for example, Herndon 1998, 2006, and references to his work cited therein) has argued that gravitationally segregated U in the core functions

as a nuclear reactor to provide substantial energy. In order to segregate uranium into the core, the core–mantle differentiation needs to have occurred at $f\text{O}_2$ of about 4–5 log units below IW buffer (Malavergne et al. 2007). Such a scenario would have reduced most iron in the planet to metallic form, leaving little oxidized iron in the mantle. This is not supported by the fact that the mantle contains $\sim 8\%$ of oxidized iron. The mantle would also have been strongly depleted in Ta under such a low oxidation, which is not the case (Mann et al. 2009). Current ideas of core formation based on siderophile element partitioning in the mantle suggest a much higher oxygen fugacity of about 2 log units below IW. Schuiling (2006) discusses other difficulties of the Herndon proposal.

Conclusions

There have been many theoretical discussions of radioactivity in the Earth's core in the past. But it is the new high P, T experiments of metal-silicate partitioning of K and U that urge a serious consideration of radioactivity in the Earth's core. The experimental data could permit up to 250 ppm of K and up to 10 ppb of U in the core at the upper end. The corresponding heat production due to K and U (and associated Th) could be ~ 2 –4 TW, a significant fraction of the CMB heat flux. As mentioned in the section on K-radioactivity, the uncertainty in these numbers is very large. It may not be possible to precisely determine the radioactivity of the core by currently available geochemical and geophysical analyses except to note the general possibility of radioactivity in the core.

Away out of this impasse may now be available. Radioactive β^- decays of K, U, and Th in the interior of the Earth produce electron antineutrinos ($\bar{\nu}_e$).

These are termed geoneutrinos, and can serve as probes to directly measure the radioactivity of the Earth's mantle and core (e.g., Raghavan et al. 1998; Rothschild et al. 1998; Mantovani et al. 2004; Enomoto et al. 2005; Fiorentini et al. 2005; Giammarchi and Miramonti 2006). Geoneutrino flux measurements use the antineutrino inverse β -decay reaction with protons, $\bar{\nu}_e + p \rightarrow e^+ + n$ in an appropriate liquid in a scintillator chamber. The resulting positron and neutron emissions are then measured by usual scintillation techniques. Geoneutrinos from each radioactive isotope can be distinguished by their characteristic energy spectrum. Neutrino detectors exist now in several countries, and new ones less susceptible to interference by nuclear reactors and other near surface radioactive sources are planned. These developments in neutrino geophysics should be able to measure the U and Th abundances in the deep Earth and the core. Present detection techniques do not allow measurement of K due to the lower energy of its antineutrinos compared to those from U and Th decay. Further developments are needed for K determination. Geoneutrino measurements are likely to

provide a definitive answer about the level of radioactivity in the core and in addition clarify several questions regarding the Earth's formation, differentiation, chemical composition, heat budget, and geodynamics (Sleep 2006).

Cross-References

- [Core Dynamo](#)
- [Core-Mantle Coupling](#)
- [Energy Budget of the Earth](#)
- [Geodynamics](#)
- [Mantle Convection](#)
- [Radiogenic Heat Production of Rocks](#)

Bibliography

- Badro J, Fiquet G, Guyot F, Gregoryanz E, Occelli F, Antonangeli D, d'Astuto M (2007) Effect of light elements on the sound velocities in solid iron: implications for the composition of the Earth's core. *Earth Planet Sci Lett* 254:233–238
- Bao X, Secco RA, Gagnon JE, Fryer BJ (2005) Experiments of U solubility in Earth's core. *EOS Trans AGU* 86(18): Jt Assem Suppl (abstr)
- Bao X, Secco RA, Gagnon JE, Fryer BJ (2006) Uranium partitioning between liquid iron and silicate melt at high pressures: implications for uranium solubility in planetary cores. *arXiv:astro-ph/0606614*
- Bouhifd MA, Jephcoat AP (2003) The effect of pressure on partitioning of Ni and Co between silicate and iron-rich metal liquids: a diamond-anvil cell study. *Earth Planet Sci Lett* 209:245–255
- Bouhifd MA, Gautron LA, Bolfan-Casanova N, Malavergne V, Hammouda T, Andraut D, Jephcoat AP (2007) Potassium partitioning into molten iron alloys at high pressure: implications for Earth's core. *Phys Earth Planet Inter* 160:22–33
- Buffett BA (2003) The thermal state of Earth's core. *Science* 299:1675–1676
- Bukowski MST (1976) The effect of pressure on the physics and chemistry of potassium. *Geophys Res Lett* 3:491–503
- Butler SL, Peltier WR, Costin SO (2005) Numerical models of the Earth's thermal history: effects of inner-core solidification and core potassium. *Phys Earth Planet Inter* 152:22–42
- Chabot NL, Agee CB (2003) Core formation in the Earth and Moon: new experimental constraints from V, Cr, and Mn. *Geochim Cosmochim Acta* 67:2077–2091
- Chabot NL, Drake MJ (1999) Potassium solubility in metal: the effects of composition at 15 kbar and 1900°C on partitioning between iron alloys and silicate melts. *Earth Planet Sci Lett* 172:323–335
- Chabot NL, Draper DS, Agee CB (2005) Conditions of core formation in the Earth: constraints from nickel and cobalt partitioning. *Geochim Cosmochim Acta* 69:2141–2151
- Corgne A, Kesav S, Fei Y, McDonough W (2007) How much potassium is in the Earth's core? New insights from partitioning experiments. *Earth Planet Sci Lett* 256:567–576
- Costin SO, Butler SL (2006) Modelling the effects of internal heating in the core and lowermost mantle on the Earth's magnetic history. *Phys Earth Planet Inter* 157:55–71
- Enomoto S, Ohtani E, Inoue K, Suzuki A (2005) Neutrino geophysics with KamLAND and future prospects. *arXiv:hep-ph/0508049*
- Ferber RC, Wallace TC, Libby LM (1984) Uranium in the Earth's core. *EOS Trans Am Geophys Union* 65:785
- Fiorentini G, Lissia M, Mantovani F, Vanucci R (2005) Geo-neutrinos: a new probe of Earth's interior. *Earth Planet Sci Lett* 238:235–247
- Furst MJ, Stapanian MI, Burnett DS (1982) Observation of non-lithophile behavior for U. *Geophys Res Lett* 9:41–44
- Ganguly J, Kennedy GC (1977) Solubility of K in Fe-S liquid, silicate-K-(FeS)liq equilibria, and their planetary implications. *Earth Planet Sci Lett* 35:411–420
- Gessman CK, Wood BJ (2002) Potassium in the Earth's core? *Earth Planet Sci Lett* 200:63–78
- Gessmann CK, Rubie DC (1998) The effect of temperature on the partitioning of nickel, cobalt, manganese, chromium, and vanadium at 9 GPa and constraints on formation of the Earth's core. *Geochim Cosmochim Acta* 62:867–882
- Gessmann CK, Rubie DC (2000) The origin of the depletions of V, Cr, Mn in the mantles of the Earth and Moon. *Earth Planet Sci Lett* 184:95–107
- Gessmann CK, Rubie DC, McCammon CA (1999) Oxygen fugacity dependence of Ni, Co, Mn, Cr, V, and Si partitioning between liquid metal and magnesiowüstite at 9–18 GPa and 2200°C. *Geochim Cosmochim Acta* 63:1853–1863
- Giammarchi MG, Miramonti L (2006) Borexino: Geoneutrinos in Borexino. *Earth Moon Planet* 99:207–220
- Goettel KA (1976) Models for the origin and composition of the Earth, and the hypothesis of potassium in the Earth's core. *Geophys Surv* 2:369–397
- Gubbins D, Masters TG, Jacobs JA (1979) Thermal evolution of the Earth's core. *Geophys J R Astron Soc* 59:57–99
- Hall HT, Murthy VR (1971) The early chemical history of the Earth: some critical elemental fractionations. *Earth Planet Sci Lett* 11:239–244
- Herndon JM (1998) Composition of the deep interior of the Earth: divergent geophysical development with fundamentally different geophysical implications. *Phys Earth Planet Inter* 105:1–4
- Herndon JM (2006) Solar system processes underlying planetary formation, geodynamics, and the georeactor. *Earth Moon Planet* 99:53–89
- Hillgren VJ, Drake MJ, Rubie DC (1996) High pressure and high temperature metal-silicate partitioning of siderophile elements: the importance of silicate liquid composition. *Geochim Cosmochim Acta* 60:2257–2263
- Hirao N, Ohtani E, Kondo T, Endo N, Kuba T, Suzuki T, Kikegawa T (2006) Partitioning of potassium between iron and silicate at the core-mantle boundary. *Geophys Res Lett* 33:L08303. <https://doi.org/10.1029/2005GL025324>
- Hofmeister AM, Criss RE (2005) Earth's heat flux revised and linked to chemistry. *Tectonophysics* 395:159–177
- Ito E, Morooka K, Ujike O (1993) Dissolution of K in molten iron at high pressure and temperature. *Geophys Res Lett* 20:1651–1654
- Jana D, Walker D (1997) The influence of silicate melt composition on distribution of siderophile elements among metal and silicate liquids. *Earth Planet Sci Lett* 150:463–472
- Javoy M (1995) The integral enstatite chondrite model of the Earth. *Geophys Res Lett* 22:2219–2222
- Javoy M (1999) Chemical Earth models. In: *Earth and planetary sciences*, vol 329. C. R. Academy Science, pp 537–555
- Labrosse S (2003) Thermal and magnetic evolution of the Earth's core. *Phys Earth Planet Inter* 140:127–143
- Labrosse S, Macouin M (2003) The inner core and the geodynamo. *Compt Rendus Geosci* 335:37–50
- Labrosse S, Poirier J-P, Le Mouél J-L (2001) The age of the inner core. *Earth Planet Sci Lett* 190:111–123
- Lee WHK (1970) On the global variations of terrestrial heatflow. *Phys Earth Planet Inter* 2:332–341
- Lee KKM, Jeanloz R (2003) High-pressure alloying of potassium and iron: radioactivity in the Earth's core? *Geophys Res Lett* 30:2212. <https://doi.org/10.1029/2003GL018515>
- Lee KKM, Steinle-Neumann G, Jeanloz R (2004) Abinitio high-pressure alloying of iron and potassium: implications for the Earth's core. *Geophys Res Lett* 31:L11603. <https://doi.org/10.1029/2004GL019839>, 2004.

- Lewis JS (1971) Consequences on the presence of sulfur in the core of the Earth. *Earth Planet Sci Lett* 11:130–134
- Li J, Agee CB (2001) The effect of pressure, temperature, oxygen fugacity and composition on partitioning of nickel and cobalt between liquid Fe–Ni–S alloy and liquid silicate: implications for the Earth's core formation. *Geochim Cosmochim Acta* 65:1821–1832
- Lodders K (1995) Alkali elements in the Earth's core: evidence from enstatite chondrites. *Meteoritics* 30:93–101
- Lodders K, Fegley BJ Jr (1998) *The planetary scientist's companion*. Oxford University Press, Oxford
- Malavergne V, Tarrida M, Combes R, Bureau H, Jones J (2007) New high-pressure and high-temperature metal/silicate partitioning of U and Pb: implications for the cores of the Earth and Mars. *Geochim Cosmochim Acta* 71:2637–2655
- Mann U, Frost DJ, Rubie DC (2009) Evidence for high-pressure core-mantle differentiation from the metal-silicate partitioning of lithophile and weakly-siderophile elements. *Geochim Cosmochim Acta* 73:7360–7386
- Mantovani F, Carmignani GL, Fiorentini G, Lissia M (2004) Antineutrinos from the Earth: a reference model and its uncertainties. *Phys Rev D* 69:297–314
- McDonough WF (1999) Earth's core. In: Marshall CP, Fairbridge RW (eds) *Encyclopedia of geochemistry*. Kluwer Academic, Dordrecht
- McDonough WF (2003) Compositional model for the Earth's core. In: Carlson RW (ed) *Treatise on geochemistry. The mantle and the core*, vol 2. pp 547–568
- Murrell MT, Burnett DS (1986) Partitioning of K, U, and Th between sulfide and silicate liquids: implications for radioactive heating of planetary cores. *J Geophys Res* 91:8126–8136
- Murthy VR (1991) Early differentiation of the Earth and the problem of mantle siderophile elements: a new approach. *Science* 253:303–306
- Murthy VR (2006) Radioactivity of the Earth and the case for potassium in the Earth's core. *Earth Moon Planet* 99:23–32
- Murthy VR, Hall HT (1970) The chemical composition of the Earth's core: possibility of sulphur in the core. *Phys Earth Planet Inter* 2:276–282
- Murthy VR, van Westrenen W, Fei Y (2003) Experimental evidence that potassium is a substantial radioactive heat source in planetary cores. *Nature* 423:163–165
- Murthy VR, Draper D, Agee C (2007) Uranium in the Earth's core? Metal-silicate partitioning of Uranium at High Pressure and Temperature and Highly Reducing Conditions. In: *Workshop on early planetary differentiation. Lunar Planetary Institute Contribution 1355*, pp 78–79
- Nimmo F, Price GD, Brodholt J, Gubbins D (2004) The influence of potassium on core and geodynamo. *Geophys J Int* 156:363–376
- Oversby VM, Ringwood AE (1972) Potassium distribution between metal and silicate and its bearing on the occurrence of potassium in the Earth's core. *Earth Planet Sci Lett* 14:345–347
- Parker LJ, Atou T, Badding JV (1996) Transition element-like chemistry for potassium under pressure. *Science* 273:95–97
- Pollack HN, Hurter SJ, Johnson JR (1993) Heat flow from the Earth's interior: analysis of the global data set. *Rev Geophys* 31:267–280
- Raghavan RS, Schoenert S, Enomoto S, Shirai S, Suekane F, Suzuki A (1998) Measuring the global radioactivity in the Earth by multi-detector antineutrino spectroscopy. *Phys Rev Lett* 80:636–638
- Righter K, Drake MJ, Yaxley G (1997) Prediction of siderophile element metal/silicate partition coefficients to 20 GPa and 2800°C: the effects of pressure, temperature, oxygen fugacity, and silicate and metallic melt composition. *Phys Earth Planet Inter* 100:115–134
- Roberts PH, Jones CA, Calderwood AR (2003) Energy fluxes and Ohmic dissipation in the Earth's core. In: Jones CA, Soward AM, Zhang K (eds) *Earth's core and lower mantle*. Taylor, London
- Rothschild CG, Chen MC, Calaprice FP (1998) Antineutrino geophysics with liquid scintillation detectors. *Geophys Res Lett* 25:1083–1086
- Rubie DC, Melosh HJ, Reid JE, Liebske C, Righter K (2003) Mechanisms of metal–silicate equilibration in the terrestrial magma ocean. *Earth Planet Sci Lett* 205:239–255
- Schilling RD (2006) Is there a nuclear reactor at the center of the Earth? *Earth Moon Planet* 99:33–49
- Sleep NH (2006) Strategy for applying neutrino geophysics to the Earth sciences including planetary habitability. *Earth Moon Planet* 99:343–358
- Stevenson DJ, Spohn T, Schubert G (1983) Magnetism and thermal evolution of the terrestrial planets. *Icarus* 54:466–489
- Van Schmus WR (1995) Natural radioactivity of the crust and mantle. In: Ahrens TJ (ed) *Global Earth physics: a handbook of physical constants, AGU reference shelf 1*. American Geophysical Union, Washington, DC, pp 283–291
- Verhoogen J (1973) Thermal regime of the Earth's core. *Phys Earth Planet Inter* 7:47–58
- Wade J, Wood BJ (2005) Core formation and the oxidation state of the Earth. *Earth Planet Sci Lett* 236:78–95
- Walter MJ, Thibault Y (1995) Partitioning of tungsten and molybdenum between metallic liquid and silicate melt. *Science* 270:1186–1189
- Wheeler KT, Walker D, Fei Y, Minarik W, McDonough W (2006) Experimental partitioning of uranium between liquid iron sulfide and liquid silicate: implications for radioactivity in the Earth's core. *Geochim Cosmochim Acta* 70:1537–1547

Radiogenic Heat Production in the Continental Crust

Claude Jaupart¹ and Jean-Claude Mareschal²

¹Université de Paris Institut de Physique du Globe, Paris, France

²Centre GEOTOP-UQAM, University of Québec, Montréal, QC, Canada

Definition

The crust is enriched in heat producing elements (HPE). Steady-state crustal temperature depends on Moho heat flux and the vertical distribution of HPE. Crustal differentiation with upper crustal enrichment in HPE cools and stabilizes the crust and makes it mechanically stronger.

Crustal Thickness and Heat Production

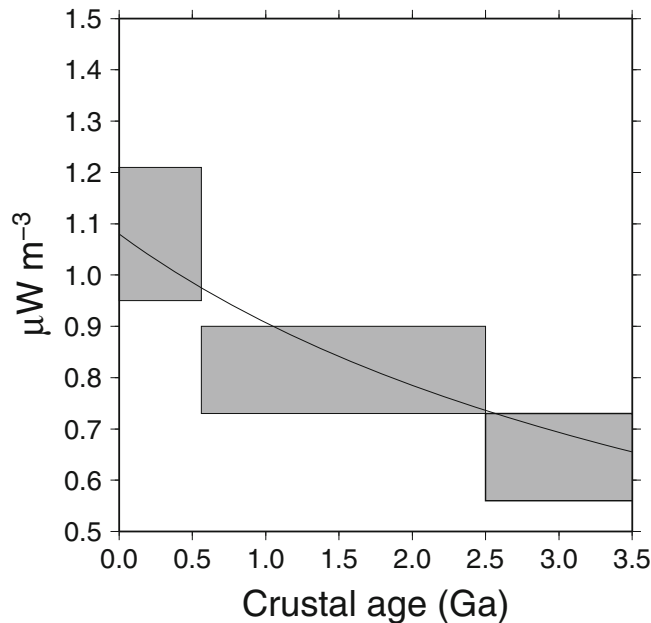
On a global average, radiogenic heat production in the continental crust accounts for about two-thirds of the surface heat flow in geological provinces that are in thermal equilibrium, that is, 30–40 mW m^{−2} over a total of 45–60 mW m^{−2} (Jaupart and Mareschal 2015). Save for a minor contribution from heat sources in the mantle lithosphere, the other component of the surface heat flow is heat supply from the convecting mantle. The key difference between these two components is in their time-dependence. The former (crustal heat production) decays with time due to the rundown of the

radioactive elements. The latter (mantle heat flux) does not vary much because continental lithosphere has remained thick over its whole history. As a consequence, the secular cooling of the mantle, which amounts to about 200 K over a total of 1600 K (see ► “Energy Budget of the Earth”), cannot induce large changes of heat flux at the base of the lithosphere. Thus, the secular evolution of the continental thermal structure is by and large a function of that of its crustal heat sources. This is consistent with the observed decrease of heat production with age (Fig. 1).

In steady state, the surface heat flux q_0 is the sum of the heat flux at the base of the crust q_m and the total crustal heat production.

$$q_0 = q_m + \int A(z)dz \quad (1)$$

with $A(z)$ the heat production. In stable continents, several independent methods indicate that the Moho heat flux lies within a narrow range of 12–18 mW m⁻² (Jaupart and Mareschal 2015), implying that variations of surface heat flux are mostly due to differences in crustal heat production. If the mantle heat flux q_m and the crustal thickness h_m can be determined independently, one can estimate the bulk average heat production of the crust, noted A_c , as follows:



Radiogenic Heat Production in the Continental Crust, Fig. 1 Average crustal heat production as a function of age, from Jaupart and Mareschal (2014). The large width of the age groups is due to the large spread of heat production values at any given age, which does not allow a fine scale separation. The curve illustrates the rundown of heat producing elements due to radioactive decay. After correction for this rundown, the crustal heat production at the time of crustal stabilization is essentially constant

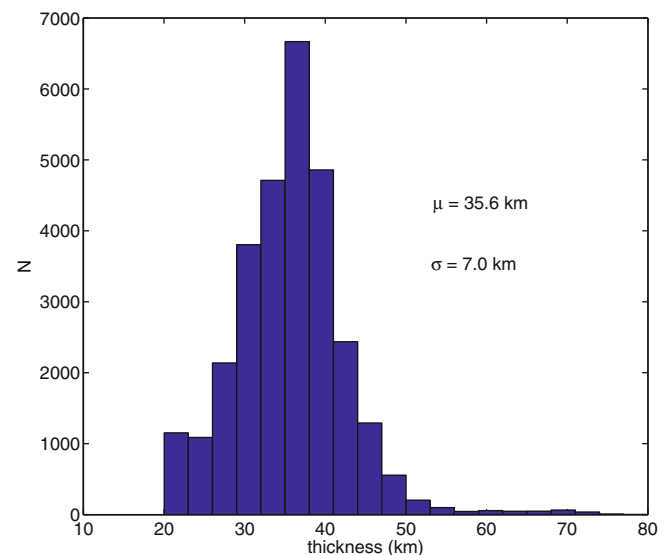
$$A_c = \frac{\overline{q_0} - q_m}{h_m} \quad (2)$$

The thickness of the continental crust varies within a 20–60 km range (Fig. 2). Thus, assuming that the average crustal heat production is constant, one should observe large variations of the surface heat flux. This is clearly not the case (Fig. 3). For the global data sets, the total lack of correlation between heat flow and crustal thickness is partly due to transient thermal effects in tectonically perturbed provinces. A recent and accurate data set in the Precambrian Canadian Shield, which is in thermal steady-state, however, does not reveal any significant correlation between the two variables, implying that the thicker the crust, the lower the mean crustal heat production. One consequence is that one must determine the amount of crustal heat sources on a case-by-case basis. Calculating the total crustal heat production is relatively straightforward with sufficient surface heat flow coverage and even approximate knowledge of lithosphere thickness.

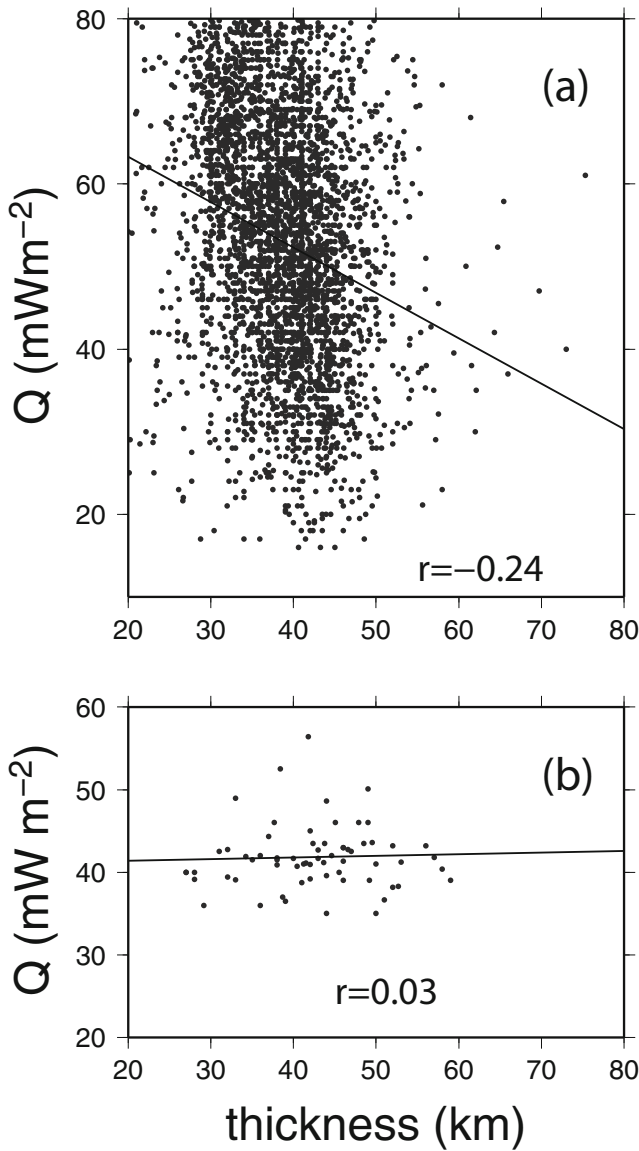
Crustal Differentiation

Constraining the vertical distribution of heat sources is fraught with a larger uncertainty than that for the average. Various estimates concur that heat producing elements are vertically differentiated with enrichment of the upper crust (Table 1).

From knowledge of the average value of heat production in surface rocks, A_s , one can calculate a “differentiation index” (Perry et al. 2006):



Radiogenic Heat Production in the Continental Crust, Fig. 2 Histogram of continental crust thicknesses sampled at 1° × 1° from Laske et al. (2013). Most of values >50 km or < 30 km are in active regions. μ and σ are the mean and standard deviation



Radiogenic Heat Production in the Continental Crust, Fig. 3 Scatter plot of heat flux and crustal thickness. (a) Global data set comparing heat flux averaged over $1^\circ \times 1^\circ$ cells vs. crustal thickness from CRUST1.0. The solid line represents the best linear “fit” to the cloud of points (with a correlation coefficient $r = -0.24$). (b) Data from eastern Canada. Heat flux data have been averaged over $1^\circ \times 1^\circ$ cells and their values are plotted against the corresponding crustal thickness values from Lithoprobe (Perry et al. 2002) and recent receiver function studies (Fiona Darbyshire, pers. comm.). The plot shows no significant trend and the correlation coefficient between the two data sets $r = 0.03$

$$D_I = \frac{A_S}{A_c} = \frac{A_S \times h_m}{\bar{q}_o - q_m} \quad (3)$$

where $\bar{q}_o$ is a regional average. D_I is equal to 1 if the crust is undifferentiated and is larger than 1 in most geological provinces because the upper crust is enriched compared to the bulk.

If the lower layer has zero heat production, an extreme case of differentiation, the upper layer thickness is:

$$h = \frac{h_m}{D_I} \quad (4)$$

and the crustal component of mantle temperature is:

$$\Delta T_c = \frac{1}{D_I} \frac{A_c h_m^2}{2\lambda} \quad (5)$$

with λ is thermal conductivity assumed constant.

A more general two-layer model has nonzero heat production in the lower layer and an upper layer of thickness $h \leq h_m/D_I$. In this case:

$$\Delta T_c = \frac{D_I A_c h^2}{2\lambda} + \left(\frac{A_c h_m (h_m + h)}{2\lambda} \right) \left(1 - \frac{D_I h}{h_m} \right) \quad (6)$$

This equation is useful to illustrate the effect of differentiation but precise calculations must include the temperature dependence of λ which may vary by a factor of 1.5 over the crustal temperature range. A value of $D_I > 1$ always results in lower value for the crustal component of Moho temperature than for an undifferentiated crust.

In general, values of D_I for different provinces are largest when heat flow is highest (Table 2). Values less than 1 are found in low heat flow subprovinces where the uppermost crustal layer has been emplaced tectonically.

Crustal Temperature and Strength

For a 40 km thick undifferentiated crust, lower crustal temperatures are near the melting temperature of the rocks at the time of crustal formation. Thermal stability of the crust can be achieved by crustal differentiation and remobilization of heat producing elements in the upper crust. For a differentiated crust, the temperature at the base of the crust increases with the thickness of the enriched layer h and decreases with increasing D_I , but the two parameters are not independent as shown by Eq. 3. In the lower crust, melting starts when temperature reaches $\approx 800^\circ \text{C}$. During the Archean with crustal heat production $\approx 1.5 \mu\text{W m}^{-3}$, lower crustal temperatures always exceeded melting conditions for a >40 km thick crust. But a crust up to 50 km thick could stay below melting if it is sufficiently differentiated ($D_I \geq 2$) (Fig. 4). The process is well illustrated for the Appalachians and the ca. 400 My Acadian orogeny, which saw the emplacement of enriched plutons in the upper crust both during and after the orogeny (Table 3). In this case, the present-day distribution of heat sources is not relevant to the Acadian thermal conditions.

Radiogenic Heat Production in the Continental Crust, Table 1 Different estimates of heat production in the continental crust (in $\mu\text{W m}^{-3}$). For each crustal model with the exception of the North

American Cordillera, the first line lists the thicknesses of the crustal layers and the total crustal thickness

Model	Upper crust	Middle crust	Lower crust	Bulk
Global geochemical models				
Rudnick and Gao (2014)	12 km	11 km	17 km	40 km
	1.6	0.96	0.18	0.89
Huang et al. (2013)	13 km	11 km	10 km	34 km
	1.6	0.73	0.17	0.94
Hacker et al. (2015) ^a	13.7 km	13 km	12.1 km	38.8 km
Model A	1.58	0.35	0.21	0.74
Model B	1.58	0.34	0.17	0.72
Model C	1.58	0.46	0.26	0.80
Model D	1.58	0.72	0.33	0.90
Exposed cross-sections				
North American cordillera				
Lee et al. (2007) ^b	1.67	0.99	0.2	0.88
Kohistan, Pakistan				
Jagoutz and Schmidt (2012) ^c	/	/	≈25 km	≈55 km
Model 1	/	/	0.18	0.69
Model 2	/	/	0.11	0.50
Model 3	/	/	0.06	0.58
Heat flow data				
Jaupart and Mareschal (2014)				0.79–0.95

^aModels A–D correspond to different end-member compositions that are compatible with geophysical characteristics

^bNo thicknesses for the three crustal layers are reported

^cModels 1–3 correspond to slightly different thicknesses and the inclusion or exclusion of the Chilas ultramafic-mafic complex & / No data for two different crustal layers at the top of the lower crust

Radiogenic Heat Production in the Continental Crust, Table 2 Average surface heat flux, $\overline{q_o}$, average crustal heat production, $\overline{A}$, crustal thickness, h_m , and differentiation index, D_I , (Eq. 3) for different provinces

Province	Age (Gy)	$\overline{q_o} \pm \sigma$ (mW m ⁻²)	$\overline{A} \pm \sigma$ ($\mu\text{W m}^{-3}$)	h_m (km)	D_I	Reference
Slave province, Can.	3.1	51	2.0	36	2.0	(1)
Superior craton core, Can.	>2.7	31.8 ± 5.2	0.78 ± 0.37	40	2.0	(2)
Superior accreted terranes, Can.	2.7	41.0 ± 8.7	0.8 ± 0.8	40	1.0	(2)
Wawa subprovince, Superior, Can.	2.7	45.1 ± 8.0	0.85 ± 0.79	40	1.0	(3)
Abitibi subprovince (all), Superior, Can.	2.7	39.9 ± 7.0	0.5 ± 0.4	38	0.7	(3)
Trans-Hudson Orogen, Can.	2.1–1.8	42 ± 11	0.7 ± 0.5	40	1.1	(2)
Flin-Flon Snow Lake Belt (THO), Can.	1.9–1.8	40 ± 5	0.32 ± 0.2	40	0.5	(3)
Wopmay Orogen, Can.	1.8	90 ± 15	4.8	32	2.0	(3, 4)
Central Shield, Aust.	1.8	72 ± 24	3.6 ± 1.9	35	3.2	(5, 6)
Eastern Gawler craton, Aust.	1.6	78 ± 19	5.0	40	2.4	(5, 6)
Grenville, Can.	1.3–1.1	41 ± 11	0.8	40	1.0	(3)
Namaqua, S. Africa	1.05	61 ± 11	2.3	43	2.0	(6, 7)
Appalachians, N. Am.	0.4	57 ± 13	2.6 ± 1.9	40	2.5	(3)

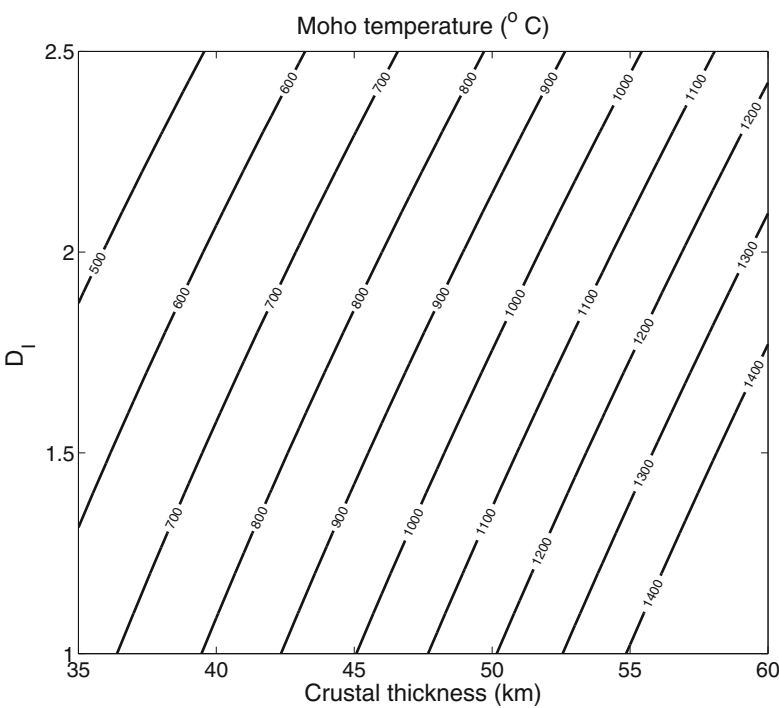
References: (1) Perry et al. (2006), (2) Jaupart et al. (2014), (3) Perry et al. (2010), (4) Lewis et al. (2003), (5) Neumann et al. (2000), (6) Mareschal and Jaupart (2013), (7) Jones (1987, 1992)

Continental lithosphere thickened on a spatial scale of >100 km has higher potential energy relative to the surrounding and experiences a tensile stress regime. Strength of

the thickened lithosphere depends on its rheology which, below the brittle region, is controlled by temperature and composition. For standard values of the rheological

Radiogenic Heat Production in the Continental Crust,

Fig. 4 Moho temperature variations in function of crustal thickness and differentiation index D_I for a mean crustal heat production $1.5 \mu\text{m}^{-3}$ representing Archean conditions. The differentiated upper crust is 10 km thick. The Moho heat flux is assumed to be 15 mW m^{-2}



Radiogenic Heat Production in the Continental Crust,
Table 3 Heat production data for the Appalachian province, USA

	A ($\mu\text{W m}^{-3}$)	Reference
Large-scale surface average	2.6 ± 0.3	(1)
Shales ^a	2.15–2.37	(2)
Syntectonic plutons (410–390 Ma) ^b	1.8–2.2	(3, 4, 5)
Post-tectonic plutons (360 Ma) ^c	4.0	(6)
Anorogenic granites (ca. 180 Ma) ^d	8.6	(6)

References: (1) Jaupart and Mareschal (2014), (2) Della Vedova and Von Herzen (1987), (3) Chamberlain and Sonder (1990), (4) Lyons (1964), (5) Jaupart et al. (1982), (6) Roy et al. (1968)

^aSediments from the US Atlantic continental margin (COST B-2 and B-3 boreholes)

^bNew Hampshire plutonic suite (three major plutons: Kinsman, Spaulding, and Bethlehem gneiss)

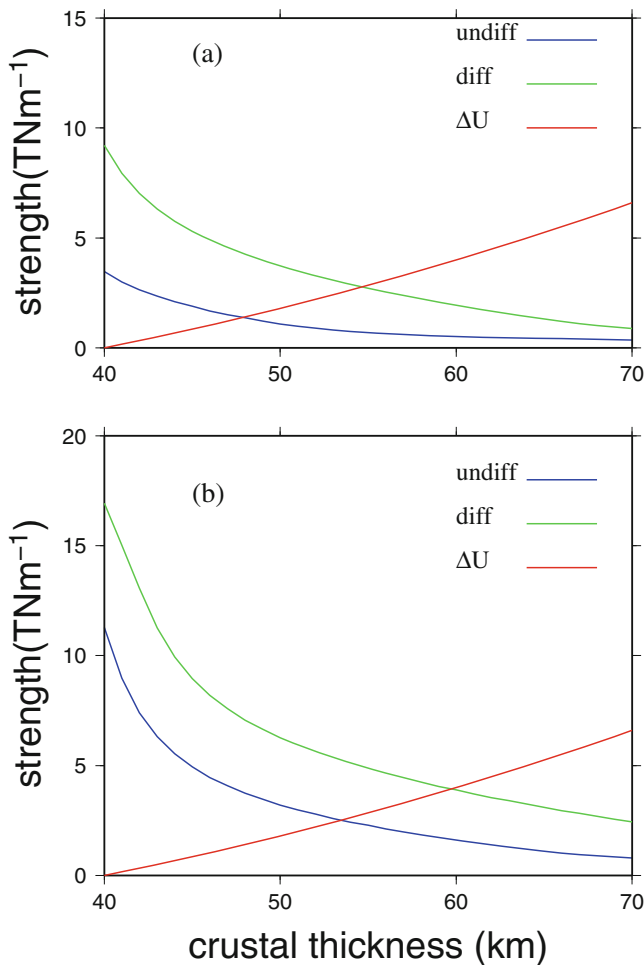
^cConcord two mica granites (several plutons scattered throughout New England)

^dWhite Mountain anorogenic plutonic series

parameters, lithospheric strength can be calculated and compared with potential energy difference. Such calculations show that: (1) for an undifferentiated crust with average crustal heat production, the lithosphere cannot withstand the stress of more than a few km crustal thickening, and (2) with increased D_I , the lithosphere can maintain greater differences in crustal thickness without collapsing (Fig. 5) (Jaupart et al. 2016).

Summary: Crustal Stabilization

In steady state, on average, temperatures in the continental crust increase with depth due to two heat flow components (crustal heat production and heat input from the mantle). It turns out that these two components account for about the same temperature rise at Moho depth. The influence of crustal heat sources is large in young provinces with high heat flow, and it happens to also be large in older ones when one accounts for radioactive decay. It also becomes large in continental collision zones with thickened crust. Over the time span of an orogenic event, 100 My, say, crustal heat production overwhelms the effect of magmatic intrusions, which are short-lived. Thus, thermal models of continents are very sensitive to crustal heat sources. In addition, the vertical distribution of crustal heat sources changes during orogenic events and is yet another variable which must be treated with care. A particularly important fact is that the distribution that is observed today is a consequence of orogeny and hence is not representative of conditions during the orogenic event itself. Usually, an orogenic event leads to the generation of crustal melts that rise toward the surface and leave a depleted residue. These processes act to generate a thermally stable and mechanically strong crust.



Radiogenic Heat Production in the Continental Crust, Fig. 5 Total strength of lithosphere vs. crustal thickness. The strength is calculated for $D_I = 1$ (undifferentiated) and $D_I = 2.5$ (differentiated crust). The potential energy difference relative to 40 km thick crust (ΔU) is given for comparison. (a) The average crustal heat production is $1.5 \mu\text{W m}^{-3}$ (i.e., the average heat production in cratons at the end of Archean). (b) The average crustal heat production is $1.1 \mu\text{W m}^{-3}$ (i.e., the average heat production of the Appalachians at the time of the Acadian orogeny)

Cross-References

- [Energy Budget of the Earth](#)
- [Heat Flow, Continental](#)
- [Lithosphere, Continental: Thermal Structure](#)
- [Lithosphere, Mechanical Properties](#)
- [Radiogenic Heat Production of Rocks](#)

Bibliography

Chamberlain CP, Sonder LJ (1990) Heat-producing elements and the thermal and baric patterns of metamorphic belts. *Science* 250:763–769

- Della Vedova B, Von Herzen RP (1987) Geothermal heat flux at the COST B-2 and B-3 wells, U.S. Atlantic continental margin. Technical report WHOI-87-27, Woods Hole Oceanographic Institution Technical Report
- Hacker BR, Kelemen PB, Behn MD (2015) Continental lower crust. *Annu Rev Earth Planet Sci* 43(6):1–6.39
- Huang Y, Chubakov V, Mantovani F, Rudnick RL, McDonough WF (2013) A reference earth model for the heat-producing elements and associated geoneutrino flux. *Geochem Geophys Geosyst* 14:2003–2029
- Jagoutz OE, Schmidt M (2012) The formation and bulk composition of modern juvenile continental crust: the Kohistan arc. *Chem Geol* 298–299:79–96
- Jaupart C, Mareschal JC (2014) Constraints on crustal heat production from heat flow data. In: Rudnick RL (ed) *Treatise on geochemistry, the crust*, vol 4, 2nd edn. Elsevier, New York, pp 53–73
- Jaupart C, Mareschal JC (2015) Heat flow and thermal structure of the lithosphere. In: Watts AB (ed) *Treatise on geophysics*, vol 6, The lithosphere, 2nd edn. Elsevier, Oxford, pp 217–253
- Jaupart C, Mann JR, Simmons G (1982) A detailed study of the distribution of heat flow and radioactivity in New Hampshire (U.S.A.). *Earth Planet Sci Lett* 59:267–287
- Jaupart C, Mareschal J-C, Bouquerel H, Phaneuf C (2014) The building and stabilization of an Archean Craton in the Superior Province, Canada, from a heat flow perspective. *J Geophys Res Solid Earth* 119:9130–9155
- Jaupart C, Mareschal J-C, Iarotsky L (2016) Radiogenic heat production in the continental crust. *Lithos* 262:398–427
- Jones MQW (1987) Heat flow and heat production in the Namaqua mobile belt, South Africa. *J Geophys Res Solid Earth* 92:6273–6289
- Jones MQW (1992) Heat flow anomaly in Lesotho: implications for the southern boundary of the Kaapvaal craton. *Geophys Res Lett* 19:2031–2034
- Laske G, Masters G, Ma Z, Pasyanos M (2013) Update on CRUST1.0 – a 1-degree global model of earth's crust. In: EGU general assembly conference abstracts. Abstract EGU2013-2658 presented at 2013 Geophys. Res. Abstracts 15, vol 15. p 2658
- Lee C-TA, Morton D, Kistler RW, Baird AK (2007) Petrology and tectonics of Phanerozoic continent formation: from island arcs to accretion and continental arc magmatism. *Earth Planet Sci Lett* 263:370–387
- Lewis TJ, Hyndman RD, Fluck P (2003) Heat flow, heat generation and crustal temperatures in the northern Canadian Cordillera: thermal control on tectonics. *J Geophys Res Solid Earth* 108:2316
- Lyons JB (1964) Distribution of thorium, uranium in three early Paleozoic plutonic suites of New Hampshire. Technical report 1144-F, U.S. Geological Survey Bulletin
- Mareschal J-C, Jaupart C (2013) Radiogenic heat production, thermal regime and evolution of continental crust. *Tectonophysics* 609:524–534
- Neumann N, Sandiford M, Foden J (2000) Regional geochemistry and continental heat flow: implications for the origin of the South Australian heat flow anomaly. *Earth Planet Sci Lett* 183:107–120
- Perry HKC, Eaton DWS, Forte AM (2002) A revised crustal model for Canada based on Lithoprobe results. *Geophys J Int* 150:285–294
- Perry HKC, Jaupart C, Mareschal JC, Bienfait G (2006) Crustal heat production in the Superior Province, Canadian Shield, and in North America inferred from heat flow data. *J Geophys Res Solid Earth* 111:B04401
- Perry H, Rosieanu C, Mareschal JC, Jaupart C (2010) Thermal regime of the lithosphere in Canada. *Can J Earth Sci* 47:389–408
- Roy RF, Blackwell DD, Birch F (1968) Heat generation of plutonic rocks and continental heat flow provinces. *Earth Planet Sci Lett* 5:1–12
- Rudnick RL, Gao S (2014) Composition of the continental crust. In: Rudnick RL (ed) *Treatise on geochemistry*, vol 4, 2nd edn. Elsevier, New York, pp 1–51

Radiogenic Heat Production of Rocks

Christoph Clauser

Institute for Applied Geophysics and Geothermal Energy,
RWTH Aachen University, Aachen, Germany

Synonyms

Radiogenic heat generation rate

Definition

Radiogenic heat production	Physical property defining the amount of heat liberated in unit time in a unit volume of rock by the decay of unstable radiogenic isotopes; dimension: W m^{-3} .
Geoneutrino	An electron antineutrino emitted in β -decay of nuclei during radiogenic heat production caused by the decay of the unstable isotopes ^{238}U , ^{232}Th , and ^{40}K .
eV (electron Volt)	A non-SI unit of energy in nuclear physics, defined as the kinetic energy gained by an electron of elementary charge when accelerating through an electric potential difference of 1 V. Thus, one electron Volt equals one Volt, which is one Joule per Coulomb, multiplied by the electron charge of $e = 1.602176487(40) \times 10^{-19} \text{ C}$. Therefore, $1 \text{ eV} = 1.602176487 \times 10^{-19} \text{ J}$.
ppm (parts per million)	A non-SI unit of relative frequency (or abundance) in 10^{-6} , similar to % (percent) or ‰ (per mil) in 10^{-2} and 10^{-3} , respectively.

Radiogenic Heat Generation

The main interior sources of heat in the Earth are the heat content of the infant Earth immediately after formation due to gravitational contraction and the decay of unstable, radioactive isotopes. The potential energy released as heat during the gravitational accretion of the infant Earth dwarfs all other heat sources. This energy is calculated from gravitational contraction. The mass accumulated in a spheroidal planet of constant density ρ and radius r equals $m = (4\pi/3) r^3 \rho$. Addition of a further shell of thickness dr and mass $dm = 4\pi r^2 \rho dr$ results in the release of additional potential gravitational energy $dE_p = G m dm/r = (16/3) \pi^2 G \rho^2 r^4 dr$, where $G = 6673 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$ is the universal gravitational constant. Integration yields:

$$E_p = \int_0^{r_E} dE_p = \int_0^{r_E} \frac{Gm}{r} dm = \frac{16\pi^2 \rho^2 G}{3} \int_0^{r_E} r^4 dr$$

$$= \frac{16\pi^2 \rho^2 G}{3} \frac{r^5}{5} = \underbrace{\left(\frac{4\pi r_E^3 \rho}{3} \right)^2}_{M_E} \frac{3G}{5r_E} = \frac{3}{5} \frac{GM_E^2}{r_E}, \quad (1)$$

where $M_E = 5.976 \times 10^{24} \text{ kg}$ and $r_E = 6.371 \times 10^6 \text{ m}$ are mass and radius of the Earth, respectively. Accordingly, the accretion of the Earth results in the release of heat on the order of $E_p = 2.24 \times 10^{32} \text{ J}$. A refined analysis considering the radial variation of density in the Earth modifies this value by only about 10 % to $E_p = 2.49 \times 10^{32} \text{ J}$ (Stacey and Davis 2008).

An additional, similarly large energy input is assumed to have occurred in the wake of an impact into the young infant Earth by a Mars-sized protoplanet orbiting the sun close to the Earth's orbit. The terrestrial Moon is generally believed to be the result of this impact, which released energy on the order of 10^{31} J (Melosh 1990; Canup and Righter 2000; Canup and Asphaug 2001). However, most of the vast energy liberated during accretion of the Earth or by this impact was radiated back into space already during accretion or immediately after the impact (Stacey and Davis 2008).

For an average specific heat capacity of $1.088 \text{ J kg}^{-1} \text{ K}^{-1}$ and a cooling by 650 K over a lifetime of 4.6 billion years, the mean rate by which the Earth is losing its original heat is 29 TW (Vacquier 1991, 1992). This corresponds to a thermal energy of $4.2 \times 10^{30} \text{ J}$. As the heat loss of the young and hot infant Earth was larger than today, the contribution of original heat to today's heat loss is certainly smaller. A larger cooling by 1175 K would result in a mean cooling rate of 52 TW corresponding to a thermal energy of $7.6 \times 10^{30} \text{ J}$ (Stacey and Davis 2008).

The annual production of radiogenic heat in the Earth, $6.3 \times 10^{20} \text{ J}$ (Jaupart et al. 2007), corresponds to more than twice the global production of primary energy in the year 2000. This huge energy source by itself clearly exceeds the world's energy annual demands predicted through the year 2030 (IEA 2008). If it were used at great scale, it may satisfy a large proportion of the primary energy demand of the entire twenty-first century.

Apart from the heat content of the infant Earth immediately after formation, the radiogenic decay of the unstable isotopes of uranium (^{238}U , ^{235}U), thorium (^{232}Th), and potassium (^{40}K) provides the largest internal source of heat. Most of these isotopes are enriched in the Earth's crust and mantle (Table 1). During radioactive decay, mass is converted into energy. Except for the tiny amount associated with the anti-neutrinos and neutrinos generated in β^- - and β^+ -decay or electron capture, respectively, all of this energy is converted into heat. Certain peaks in the corresponding γ -spectra are characteristic for the different decay series while the continuous background spectrum is due to Compton scattering and photoelectric absorption.

Radiogenic Heat Production of Rocks, Table 1 Average radiogenic heat generation rate per mass, A' , in geologic materials. (Data: McDonough and Sun 1995; Stacey and Davis 2008; Jaupart et al. 2007)

Material		Concentration (ppm by mass)				Heat generation rate (10^{-12} W kg $^{-1}$)
		C_U	C_{Th}	C_K	C_K/C_U	
Igneous rocks	Granites	4.6	18	33 000	7 000	1 050
	Alkali basalt	0.75	2.5	12 000	16 000	180
	Tholeiitic basalt	0.11	0.4	1 500	13 600	27
	Eclogite	0.035	0.15	500	14 000	9.2
	Peridotite, dunite	0.006	0.02	100	17 000	1.5
Meteorites	Carbonaceous chondrite	0.007 4–0.008 0	0.029–0.030	544–550	20 000	5.2
Moon	Apollo samples	0.23	0.85	590	2 500	47
Earth	Average crust (2.8×10^{22} kg)	1.2–1.3	4.5–5.6	15 500	13 000	293–330
	Average mantle (4.0×10^{24} kg)	0.013–0.025	0.040–0.087	70–160	2 800	2.8–5.1
	Average core	0	0	29	–	0.1
	Bulk silicate Earth (BSE)	$0.020 \pm 20\%$	$0.081 \pm 1.5\%$	$118 \pm 20\%$	5 400	4.7 ± 0.08

The two uranium isotopes, ^{238}U and ^{235}U , decay into lead, ^{206}Pb and ^{207}Pb , with a half-life of 4.5 and 0.71 billion years, respectively. Characteristic lines for the γ -spectrum of the ^{238}U decay series, for example, are produced by the daughter element bismuth, ^{214}Bi , at 609 keV, 1 120 keV, and 1 764 keV. Uranium is an abundant mobile trace element in many rocks. With a half-life of 14.1 billion years, thorium ^{232}Th decays also into lead ^{208}Pb . Characteristic lines for the γ -spectrum of the thorium decay series are produced by the daughter elements thallium, ^{208}Tl , or actinium, ^{228}Ac , at 584 keV and 2 615 keV or 912 keV and 966 keV, respectively. Thorium also occurs as a trace element, is relatively inactive chemically and frequently bound to clay minerals. The only unstable isotope of potassium is ^{40}K . It disintegrates by electron capture or emission into argon, ^{40}Ar , or calcium, ^{40}Ca , respectively, with a corresponding characteristic line in the γ -spectrum of the potassium decay at 1 460 keV. Potassium occurs in many clay minerals at concentrations of several percent.

Tabulated Data

The compilations of measured radiogenic heat generation rate which are reported in several research papers and reference books will not be duplicated here. Čermák and Rybach (1982) present a large collection of data arranged according to rock type. Van Schmus (1984, 1995) presents data on the abundance of radiogenic isotopes in various minerals and rocks of the Earth's crust and mantle and discusses the variation of radiogenic heat generation over the lifetime of the Earth. Stacey and Davis (2008) present data on average mass specific heat generation rates for various geologic materials, and Jaupart et al. (2007) discuss radiogenic heat sources in the Earth's crust and mantle. They provide an in-depth discussion on current models for the composition of a Bulk Silicate Earth (BSE), that is, the Earth's crust and mantle without the core: These are based on data from samples of either (1) meteorites considered

representative for the starting material and high-temperature processes in the early solar nebula when the Earth accreted or (2) upper mantle rocks formed by low-temperature processes. As can be expected, both show large variations in composition due to their different provenance and history.

Chondrites are considered representative of undifferentiated silicate material from the solar system prior to melting and the formation of a (mainly) iron core. Different classes of chondrites correspond to perturbations in elemental abundance in the gas state caused mainly by volatility or condensation temperature. The important elements with respect to radiogenic heat generation are uranium, thorium, and potassium. As elaborated by Jaupart et al. (2007), the first two condensate at very high temperature, why they are called 'refractory lithophile' elements. They show the same ratio in all types of chondritic meteorites, which demonstrates that they behave similarly in the early solar system. In contrast, potassium is a 'moderately volatile' element with a lower condensation temperature. The best agreement with solar concentration ratios is found for carbonaceous chondrites of the CI type of which, however, McDonough and Sun (1995) report only five finds. Compositional data exist only for three of them, in particular for the 700 g meteorite found near Ivuna in Tanzania in 1938 from which the name 'CI' is derived for this type of carbonaceous chondrite. However, chondrites are known to be more enriched in volatiles such as H_2O and CO_2 than the Earth.

A sample from the Earth's mantle is named "pyrolite," a contraction of the names of two principal mantle minerals, pyroxene and olivine, making up the complementary rocks peridotite and basalt. Because the former is the solid residue of the partial melting event which produced basalts, a mixture between the two was considered a suitable starting material, the pyrolytic mantle. Generally, pyrolite data suffer from leaching of uranium during low-temperature alteration. As most samples derive from the upper mantle, it has been questioned whether the pyrolytic composition adequately represents a bulk silicate Earth.

A third approach discussed by Jaupart et al. (2007) avoids a specific choice of a starting composition but determines it from the intersection of the two compositional trends of chondritic meteorites and peridotites. A source of error in this approach derives from the scatter in the two trends.

A fourth method proceeds from the elemental ratios of uranium and thorium and determines the primitive bulk Earth abundances from measurements on peridotites. Again, Jaupart et al. (2007) provide a concise introduction into this method. Table 1 summarizes mass-specific radiogenic heat generation rates for terrestrial rocks, meteorites, Moon surface samples, and average values for the Earth's crust, mantle, and a bulk silicate Earth (BSE).

Calculated Heat Generation Rate

The energy emitted by all of these decay processes comprises the kinetic energy of the emitted particles and the γ -radiation associated with the different decay processes. It is absorbed in the rocks and finally transformed into heat. In general, the total heat generation rate A of a rock is the sum of the individual contributions A_U , A_{Th} , and A_K by uranium, thorium, and potassium, respectively:

$$A = \rho(C_U A'_U + C_{Th} A'_{Th} + C_K A'_K), \quad (2)$$

where ρ is rock density, and A' and C are heat generation rate per mass and concentration, respectively, of the corresponding element in the rock. Table 2 shows values for A' reported by different authors. The variation is less than 3%, at most.

Thus, if a rock's density ρ and its concentrations in uranium (C_U), thorium (C_{Th}), and potassium (C_K) are known, its radiogenic heat generation rate A can be determined. Inserting the values, for example, of Rybach (1988) from Table 2 into Eq. 2 yields:

$$A(\mu W \text{ m}^{-3}) = 10^{-5} \times \rho(\text{kg m}^{-3}) \times \{9.52 \times C_U(\text{ppm}) + 2.56 \times C_{Th}(\text{ppm}) + 3.48 \times C_K(\%) \}, \quad (3)$$

where concentrations are given in weight-ppm (i.e., $10^{-6} \text{ kg kg}^{-1}$) for uranium and thorium and in weight-% for potassium.

The natural γ -radiation of rocks can be measured, for instance, by spectrometry on rock samples in the laboratory (see ► “Geomagnetic Field, Measurement Techniques”). An alternative source of γ -spectra is the natural gamma spectrometer (NGS) borehole tool which yields as output the three logs URAN (in weight-ppm uranium), THOR (in weight-ppm thorium), and POTA (in weight-% potassium). In combination with density RHOB from the compensated density log [g cm^{-3}], this information can be used directly to obtain the bulk heat generation rate A from Eq. 2 using any of the values for A'_U , A'_{Th} , and A'_K from Table 2.

Alternatively, for lack of spectra, the total γ -ray emission GR may be used, measured either in the laboratory or by suitable logging tools in boreholes. A γ -ray (GR) log is often included in most logging runs as one of the basic geophysical measurements in boreholes. Its reading reflects the combined radioactive decay of uranium, thorium, and potassium. However, it does not yield information on the individual contributions as it records the total number of γ -rays detected by the tool during a time interval. The relative gamma activities of uranium, thorium, and potassium, that is, the relative number of γ -rays emitted by the same mass during the same time interval are shown in Table 3.

Therefore, if POTA, THOR, and URAN are recorded as percent, ppm, and ppm, respectively, the gamma ray log's reading GR is proportional to the corresponding contents (Beardsmore and Cull 2001):

$$GR = X \times (POTA + 0.13 \times THOR + 0.36 \times URAN), \quad (4)$$

where the coefficient X varies with the radius of sensitivity of the log for detecting γ -rays. This varies with a number of factors, but lies within a few decimeters, at most, and can be approximated as constant for a given borehole (Beardsmore and Cull 2001). Then, bulk heat generation rate A can be related to GR by combining Eqs. 2 and 4 using the data of, for example, Rybach (1988) from Table 2:

$$\begin{aligned} \frac{A(\mu W \text{ m}^{-3})}{GR(\text{API})} &= 10^{-3} \times RHOB(\text{g cm}^{-3}) \frac{34.8 \times POTA(\%) + 25.6 \times THOR(\text{ppm}) + 95.2 \times URAN(\text{ppm})}{X(\text{API}) \times POTA(\%) + 0.13 \times THOR(\text{ppm}) + 0.36 \times URAN(\text{ppm})}, \\ \text{that is :} \\ A(\mu W \text{ m}^{-3}) &= 0.0348 \times RHOB(\text{g cm}^{-3}) \frac{GR(\text{API})}{X(\text{API})} \underbrace{\frac{POTA(\%) + 0.736 \times THOR(\text{ppm}) + 2.736 \times URAN(\text{ppm})}{POTA(\%) + 0.130 \times THOR(\text{ppm}) + 0.360 \times URAN(\text{ppm})}}_Y \\ &= 0.0348 \left(RHOB \times GR \times \frac{Y}{X} \right) \end{aligned} \quad (5)$$

Radiogenic Heat Production of Rocks, Table 2 Heat generation rate per mass, A' , for uranium, thorium, and potassium (Clauser 2009)

A'_U ($\mu\text{W kg}^{-1}$)	A'_{Th} ($\mu\text{W kg}^{-1}$)	A'_K ($\mu\text{W kg}^{-1}$)	Data source
97.0	27.0	0.0036	Birch (1954) cited in Jessop (1990)
97.7	26.3	0.0034	Hamza and Beck (1972) cited in Jessop (1990)
96.7	26.3	0.0035	Emsley (1989) and Jessop (1990) cited in Beardsmore and Cull (2001)
95.2	25.6	0.00348	Rybach (1988)

Radiogenic Heat Production of Rocks, Table 3 Relative gamma activity of uranium, thorium, and potassium. (Data source: Adams and Weaver (1958) and Emsley (1989) referred to in Beardsmore and Cull (2001))

Element	^{238}U	^{232}Th	^{40}K
Half-life (Ga)	4.47	14.1	1.28
Relative γ -activity	3 600	1 300	1

Y in Eq. 5 still contains the unknown relative elemental abundances. However, it turns out that the range for Y is limited between $1 \leq Y < 7.6$ for the two extreme cases when either $0 = \text{THOR} = \text{URAN}$ and $\text{POTA} > 0$, or $0 = \text{POTA} = \text{THOR}$ and $\text{URAN} > 0$. Beardsmore and Cull (2001) argue that in regions in which sediments have been derived from a common source through time the relative proportions of the elements in Eq. 5, and hence, Y should remain relatively constant with depth. Thus, Beardsmore and Cull (2001) maintain that a plot of A (derived, e.g., from a NGS log) versus $\text{RHOB} \times \text{GR}$ should yield a gradient, which depends only on the relative proportions of the radioactive elements in the sediments and should be similar over any region containing sediment derived from the same source.

An alternative empirical relationship was published by Bückner and Rybach (1996), which does not require information on the relative proportions of the radioactive elements in the rock. It relates bulk heat generation rate A to total γ -ray emission GR recorded by borehole tools as

$$A(\mu\text{W m}^{-3}) = 0.0158 \{ \text{GR}(\text{API}) - 0.8 \}. \quad (6)$$

This relation has been successfully tested on a number of data sets where both NGS and GR logs were available. As the relations in Eqs. 3 and 6 are based on measurements mostly on igneous rocks, caution is required when applying them to sedimentary rocks.

Measuring Techniques

A number of different analytical methods are available for determining the concentrations of uranium, thorium, and

potassium. An inter-laboratory comparison of different methods on selected materials yielded consistent results within a few percent (Rybach 1988). However, among all other methods, γ -ray spectroscopy is the only one which enables determining all three concentrations simultaneously. It implies secular equilibrium within the uranium and thorium decay series and a constant $^{40}\text{K}/\text{K}$ ratio, both of which are satisfied for most rock types (Rybach 1988). The method is based on counting the number of decays per energy channel. Commercial spectrometers will offer, for instance, 2048 channels for measuring decays in the energy range of 0 keV–3 000 keV, yielding a resolution, in this particular case, of 1.46 keV per channel. In performing measurements, it is important to realize that radioactive decay is a stochastic process. Therefore, decays need to be recorded over a sufficiently long time. In view of the minute uranium, thorium, and potassium concentrations, and depending on crystal size and sensitivity, this may imply recording times on the order of several hours. Additionally, a possible sampling bias needs to be considered when selecting samples for measurements.

The interpretation of γ -ray spectroscopy relates the elemental concentrations C linearly to the corresponding counting rates N in the energy channel x by the corresponding coefficient $h(x)$: $N(x) = h(x) C$. Thus, the total count rate due to the decay of uranium, thorium, and potassium in the energy channel x is given by:

$$N(x) = h_U(x) C_U(x) + h_{Th}(x) C_{Th}(x) + h_K(x) C_K. \quad (7)$$

The coefficients h depend mainly on the characteristic spectra of the radioactive elements, but also on the geometry of the sample, the geometric configuration of the measurement, and on the efficiency and sensitivity of the detectors. Thus, γ -spectroscopic measurements require calibration based on samples of known elemental concentrations. This enables accounting for the influence of the geometry of the detectors. Additionally, possibly variable sample geometry needs to be addressed separately. Three different effects require attention (Adams and Gasparini 1970): (1) *Variation of the continuous background spectrum* may occur if the background radiation was determined with an empty sample chamber in which no absorption of the natural background radiation occurs. This can be avoided by measuring the background radiation using a nonradioactive sample of equal size and similar density, such as water-filled plastic tubes; (2) *Deformation of the spectrum* may occur if a larger sample volume increases the likelihood for internal Compton scattering. Thus, larger samples are characterized by a decreased peak size and an increased level of the continuous spectrum; (3) *Radiation self-absorption* of a fraction of the γ -radiation emitted within the sample reduces the γ -radiation recorded at the detector. Self-absorption increases with sample thickness d and reduces the count rate N by a factor

$$F = \frac{1 - e^{-\mu d}}{\mu d}, \quad (8)$$

where μ is the absorption coefficient (Watt and Ramsden 1964). The joint effect of spectrum deformation and self-absorption is quantified by measurements on potash salt standards. While the absolute ^{40}K -concentrations are unknown, the relative variations reveal the influence of variable sample diameters. Theoretically, the total activity is proportional to the sample cross-section, that is, to the square of the sample diameter. In plotting the count rate versus sample diameter, one finds, however, a smaller exponent of only 1.5. Thus, the remainder of the radiation is either absorbed within the sample or transferred to other energy bands.

The energy peaks in the spectrum are analyzed for finding the elemental concentrations. Theory predicts a Gaussian normal distribution for the energy peaks. The count rate N at energy x in the local neighborhood of such a Gauß peak comprises a contribution of the Gauß peak itself, $G(x)$, and a linear trend $ax + b$, describing the background radiation and the Compton scattering (Adams and Gasparini 1970):

$$N(x) = G(x) + ax + b. \quad (9)$$

Such linear trends need to be removed prior to further interpretation, for example, by linear regression. Following this detrending the raw data need to be smoothed for further processing so that the energy peak's position can be clearly identified (Fig. 1). The area under the peak above the base line then corresponds to the sought total energy. In practice, the count rates within this area are summed up, or a Gaussian normal distribution function

$$G(x) = A e^{-(x-x_0)^2/2\sigma^2} \quad (10)$$

is fitted to the corrected peak, where x_0 and σ are mean energy and standard deviation of the peak, and A is a calibration constant. In a normal distribution, the peak's full width at half maximum (FWHM) is related to σ by

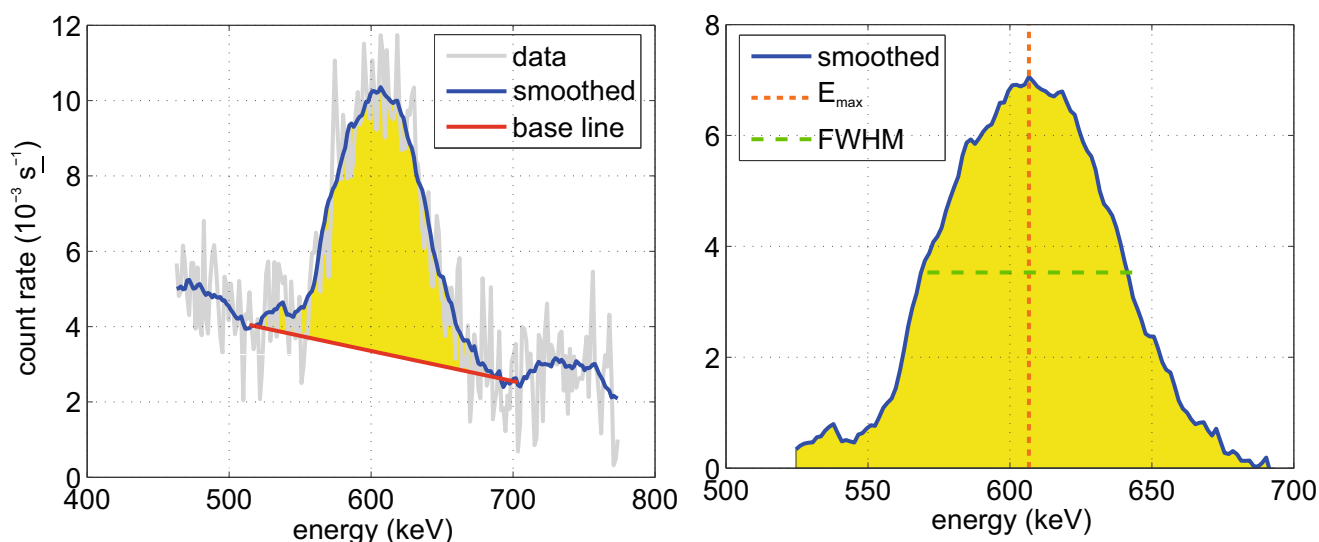
$$\text{FWHM} = 2\sigma\sqrt{\log_{10}(2)}. \quad (11)$$

The summed-up count rates analyzed this way then need to be calibrated by corresponding measurements on samples of known concentrations. Finally, this yields the concentrations of the radiogenic isotopes of uranium, thorium, and potassium.

Heat Generation and Geoneutrinos

Currently, general hypotheses on possible distributions of radiogenic heat generating elements in the mantle and core of the Earth are mainly based on analogies with chondritic material and direct evidence is lacking. However, as radiogenic sources produce heat, electron antineutrinos $\bar{\nu}$ are emitted in β^- -decay of nuclei in the decay chains of ^{238}U , ^{232}Th , and ^{40}K . Thus, the heat produced in nuclear decay is directly related to the antineutrino flux, as shown in Table 4 (Fiorentini et al. 2003).

Neutrinos are elementary particles that travel close to the speed of light, are electrically neutral, and can pass through ordinary matter almost undisturbed. They have a very small, but nonzero mass. Therefore, neutrinos are extremely difficult to detect. As the emission of geoneutrinos is coupled to radiogenic heat generation, this enables, at least theoretically,



Radiogenic Heat Production of Rocks, Fig. 1 Measurement and analysis of natural radioactivity shown here for the ^{214}Bi -peak at 609 keV; spectrum before (left) and after (right) removal of the linear base line trend (Eq. 10). FWHM computed after Eq. (11)

Radiogenic Heat Production of Rocks, Table 4 Maximum antineutrino energy and heat production rates in natural decay processes (de Meijer et al. 2006; Fiorentini et al. 2005)

Decay reaction	E_{max} (MeV)	Heat generation rate ($\mu\text{W kg}^{-1}$)
$^{238}\text{U} \rightarrow ^{206}\text{Pb} + 8\ ^4\text{He} + 6\ \text{e}^- + 6\ \bar{\nu}$	3.25	95
$^{232}\text{Th} \rightarrow ^{208}\text{Pb} + 6\ ^4\text{He} + 4\ \text{e}^- + 4\ \bar{\nu}$	2.25	27
$^{40}\text{K} \rightarrow ^{40}\text{Ca} + \text{e}^- + \bar{\nu}$	1.31	0.003 6

a geoneutrino tomography of radiogenic heat distribution within the Earth. But this concept is as difficult to implement as it is intriguing. First, there is a huge flux of cosmic neutrinos which traverses the Earth from which geoneutrinos need to be distinguished: The neutrino flux at the Earth’s surface from the sun alone is estimated at about $60 \times 10^9\ \text{s}^{-1}\ \text{cm}^{-2}$ (e.g., Bahcall 1969). The flux of geoneutrinos is three orders of magnitude smaller, but still on the order of $7.4(+2.1/-1.9) \times 10^6\ \text{s}^{-1}\ \text{cm}^{-2}$ (Kamland Collaboration 2011).

In spite of this immense flux density, the vanishing mass of neutrinos requires exceptional sensors for their detection: (1) At the Gran Sasso observatory in Italy, the detector is a tank of 100 t of liquid gallium trichloride (GaCl_3) containing the equivalent of 30 t of gallium at a subsurface laboratory. Here, antineutrinos are observed indirectly through the conversion of a gallium nucleus into a germanium isotope by neutrino capture. For all the high neutrino flux through the tank, an interaction of a neutrino and a gallium nucleus occurs only every 35 h. (2) In the *Kamioka Liquid Scintillator Antineutrino Detector (KamLAND)* experiment (Araki et al. 2005), the detector comprises of a 13 m diameter nylon balloon filled with 1 kt of liquid scintillator consisting of mineral oil, benzene, and fluorescent chemicals. KamLAND is an experiment at the Kamioka Observatory, an underground neutrino observatory near Toyama, Japan, built to detect electron antineutrinos. The experiment is located in the old Kamiokande cavity in a horizontal mine drift in the Japanese Alps. The detector is housed in an 18 m diameter stainless steel spherical vessel with 1879 photomultiplier tubes mounted on the inner surface. KamLAND is the first detector to conduct an investigation on geoneutrinos and may yield important geophysical information. It has the sensitivity to detect electron antineutrinos produced by the decay of ^{238}U and ^{232}Th within the Earth. Earth composition models suggest that the radiogenic power from these isotope decays is 16 TW, approximately half of the total measured heat dissipation rate from the Earth.

A successful antineutrino tomography requires global sampling and distinguishing between cosmic neutrinos traversing the Earth and geoneutrinos generated within the Earth. A project had been proposed some time ago (but unfortunately not funded) to deploy considerably smaller

neutrino detection sensors in a network of boreholes around the globe (de Meijer et al. 2006). Time will show whether the significant technical challenges associated with the construction of smaller sensors than those in the experiments at the Gran Sasso and Kamioka observatories can be overcome and a road opened for an antineutrino tomography of the Earth.

Summary

The heat produced in the radioactive decay of the unstable isotopes of uranium (^{238}U , ^{235}U), thorium (^{232}Th), and potassium (^{40}K) is the largest internal heat source of the Earth. During radioactive decay, mass is converted into energy. Except for the tiny amount associated with the antineutrino and neutrinos generated in β^- - and β^+ -decay or electron capture, respectively, all of this energy ends up as heat. The annual production of radiogenic heat in the Earth equals $8.6 \times 10^{20}\ \text{J}$, which is about 1½ times the global production of primary energy in the year 2017 (IEA 2020). The distribution of radiogenic isotopes in the Earth controls to a large extent the thermal regime of the Earth. Unfortunately, this distribution is known only with great uncertainty. Recently, it has become possible to detect geoneutrinos, that is, antineutrinos emitted during radioactive decay of unstable isotopes, in large detectors. With decreased size and improved accuracy of detectors together with directional resolution power, an antineutrino tomography of the Earth appears to become possible. Ideally, this would enable locating and quantifying the distribution of unstable isotopes in the Earth, thus helping to resolve a number of open questions with regard to the state and evolution of the Earth’s thermal regime.

Cross-References

- [Energy Budget of the Earth](#)
- [Heat Flow Determinations, Continental](#)
- [Heat Flow, Continental](#)
- [Heat Flow, Seafloor: Methods and Observations](#)
- [Radioactivity in Earth’s Core](#)
- [Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity](#)

Bibliography

Adams JS, Gasparini P (1970) Gamma-ray spectrometry of rocks. Elsevier, Amsterdam

Adams JAS, Weaver CE (1958) Thorium to uranium ratios as indicators of sedimentary processes: examples of the concept of geochemical facies. Bull Am Assoc Pet Geol 42:387–430

- Araki T et al (2005) Experimental investigation of geologically produced antineutrinos with KamLAND. *Nature* 436:499–503
- Bahcall JN (1969) Neutrinos from the Sun. *Sci Am* 221(1):28–37
- Beardmore GR, Cull JP (2001) *Crustal heat flow*. Cambridge University Press, Cambridge
- Birch F (1954) Heat from radioactivity. In: Faul H (ed) *Nuclear geology*. Wiley, New York, pp 148–174
- Bücker C, Rybach L (1996) A simple method to determine heat production from gamma-ray logs. *Mar Pet Geol* 13:373–375
- Canup RM, Asphaug E (2001) Origin of the Moon in a giant impact near the end of the Earth's formation. *Nature* 412:708–712
- Canup RM, Righter K (eds) (2000) *Origin of the Earth and Moon*. University of Arizona Press, Tucson
- Čermák V, Rybach L (1982) Radioactive heat generation in rocks. In: Angenheister G (ed) *Landolt-Börnstein, Group V: geophysics and space research. Physical properties of rocks, Subvol. A, vol 1*. Springer, Heidelberg/Berlin, pp 353–371
- Clauser C (2009) Heat transport processes in the Earth's crust. *Surv Geophys* 30:163–191. <https://doi.org/10.1007/s10712-009-9058-2>
- De Meijer RJ, Smit FD, Brooks FD, Fearick RW, Wörtche HJ, Mantovani F (2006) Towards Earth Antineutrino Tomography (EARTH). *Earth Moon Planet* 99(1–4):193–206
- Emsley J (1989) *The elements*. Clarendon, Oxford
- Fiorentini G, Mantovani F, Ricci B (2003) Neutrinos and energetics of the Earth. *Phys Lett B* 557:139–146
- Fiorentini G, Lissia M, Mantovani F, Vanucci R (2005) Geo-neutrinos: a new probe of Earth's interior. *Earth Planet Sci Lett* 231:139–146
- Hamza VM, Beck AE (1972) Terrestrial heat flow, the neutrino problem, and a possible energy source in the core. *Nature* 240(5380):343–344
- IEA (2008) *World energy outlook 2008*. International Energy Agency (IEA), Paris. <http://www.iea.org/textbase/nppdf/free/2008/weo2008.pdf>. Retrieved 10 July 2010
- IEA (2020) *Data and statistics*. International Energy Agency (IEA), Paris. <https://www.iea.org/data-and-statistics/data-tables?country=WORLD&energy=Balances&year=2017>. Retrieved 17 Mar 2020
- Jaupart C, Labrosse S, Mareschal J-C (2007) Temperatures, heat and energy in the mantle of the Earth. In: Bercovici D (ed) *Mantle dynamics – treatise on geophysics, vol 7*. Elsevier, Amsterdam, pp 253–303
- Jessop AM (1990) *Thermal geophysics*. Elsevier, Amsterdam
- Kamland Collaboration (2011) Partial radiogenic heat model for Earth revealed by geoneutrino measurements. *Nat Geosci* 4(9):647–651
- McDonough WF, Sun S-S (1995) The composition of the Earth. *Chem Geol* 120:223–253
- Melosh HJ (1990) Giant impacts and the thermal state of the early Earth. In: Newsom HE, Jones JH (eds) *Origin of the Earth*. Oxford University Press, New York, pp 69–83
- Rybach L (1988) Determination of heat production rate. In: Hänel R, Rybach L, Stegena L (eds) *Handbook of terrestrial heat flow density determination*. Kluwer, Dordrecht, pp 125–142
- Stacey FD, Davis PM (2008) *Physics of the Earth*, 4th edn. Cambridge University Press, Cambridge
- Vacquier V (1991) The origin of terrestrial heat flow. *Geophys J Int* 106(1):199–202
- Vacquier V (1992) Corrigendum to ‘the origin of terrestrial heat flow/prime’. *Geophys J Int* 111(3):637–638
- Van Schmus WR (1984) Radioactivity properties of minerals and rocks. In: Carmichael RS (ed) *Handbook of physical properties of rocks, vol III*. CRC Press, Boca Raton, pp 281–293
- Van Schmus WR (1995) Natural radioactivity in crust and mantle. In: Ahrens TJ (ed) *Global Earth physics – a handbook of physical constants*. AGU reference shelf, vol 1. American Geophysical Union, Washington, DC, pp 283–291
- Watt DE, Ramsden D (1964) *High sensitivity counting techniques*. Pergamon, London

Recovery of Source Magnetization Direction from Magnetic Field Data

Clive Foss

Mineral Resources, CSIRO, Lindfield, NSW, Australia

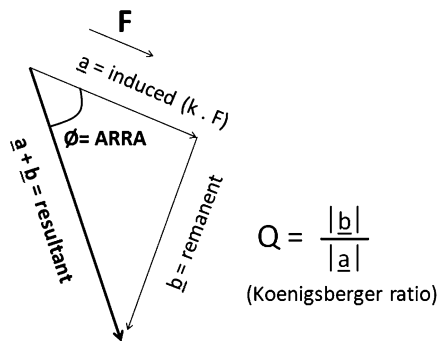
Definition

Magnetic field data are measurements of the strength, direction, or gradients of the magnetic field. Local variation between measurements is due to the external fields of nearby magnetizations. In favorable circumstances, the direction of those magnetizations can be estimated by analysis of magnetic field data.

Introduction

Most rocks at shallow depth in the earth contain ferromagnetic minerals with both induced and remanent magnetizations (see *Remanent Magnetism*). The inverse problem of determining the spatial distribution and magnetization of rocks from magnetic field analysis is a form of remote sensing with substantial challenges of non-uniqueness (see ► [“Inverse Theory, Linear”](#)). The direction of magnetization must either be known or be recovered as a part of the study. Induced magnetization is in most cases parallel to the local geomagnetic field. However, the strength of the induced magnetization is rarely known. For remanent magnetization, the strength and direction are also generally unknown. Remanent magnetization direction is of interest for age dating or correlation, global tectonic studies, and structural deformation investigations. It is also required to estimate the spatial distribution of magnetization. Analysis or inversion with an incorrect magnetization direction has long been recognized as a pitfall of magnetic field interpretation (see ► [“Magnetic Anomalies: Interpretation”](#)). A review of the many methods to estimate magnetization direction is provided by Clark (2014) and a review of the issues in deriving magnetization direction from magnetic field data is provided by Foss (2017).

The relative importance of induced and remanent magnetizations is traditionally quantified by the Koenigsberger ratio or “Q factor” (defined in Fig. 1) but this scalar statistic is insufficient to describe the vector relationship. Remanent magnetization parallel to the local geomagnetic field cannot be discriminated from induced magnetization by analysis of the static magnetic field, but also causes no error in estimation of the spatial distribution of magnetization. Many rocks contain some “soft” or “viscous” remanent magnetization in the present field direction carried by easily remagnetized



Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 1 The vector relationship between remanent, induced, and resultant magnetizations

multidomain grains (see ► [“Magnetic Domains”](#)). The diagnostic expression of remanent magnetization in magnetic field data is best represented by the angle between the resultant magnetization and the geomagnetic field (the apparent resultant rotation angle or ARRA) as shown in Fig. 1. Angular difference between resultant magnetization and the geomagnetic field is caused by “hard” remanent components generally carried in single domain or pseudo-single domain grains. Many rocks have a large assemblage of magnetic mineral grains carrying both hard and soft remanent magnetizations. If the resultant rotation angle is large, the pattern of the external magnetic field of the magnetization will deviate substantially from that due to an equivalent distribution of induced magnetization suggesting the presence of remanent magnetization. However, additional information such as the Koenigsberger ratio or magnetic susceptibility value is still required to resolve remanent magnetization direction from the resultant magnetization. An alternative approach to dealing with complications due to unknown magnetization direction is to invert transforms which have low sensitivity to magnetization direction (Paine et al. 2001). This provides an approximate mapping of the distribution of magnetization but does not recover information about the magnetization direction. Lelièvre and Oldenburg (2009) developed a method to recover magnetization direction by inverting for three orthogonal components of magnetization.

The Expression of Magnetization Direction in Magnetic Field Data

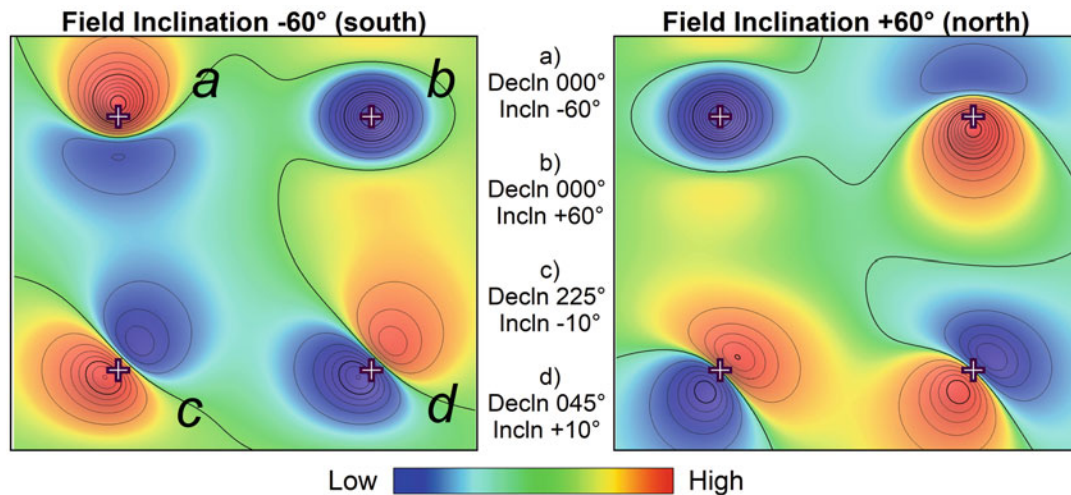
For a thin planar (“2D”) distribution of magnetization, any components of magnetization in the strike direction of the sheet do not contribute to the magnetic field, and the angle of inclination of magnetization perpendicular to the sheet combines with the dip angle of the sheet into a single term from which neither can be independently recovered (Hood 1964).

This limitation led to an incorrect but widely and long-held assumption that magnetization direction cannot be reliably recovered from magnetic field data. However, Helbig (1963) proved that the magnetization direction of a dipole (a homogeneously magnetized sphere or a source so small that its shape is irrelevant) can be recovered from magnetic field analysis provided the dipole field is isolated from other fields and the horizontal center of magnetization is known. Schmidt and Clark (1998) recognized the potential for modern computing capabilities to apply Helbig’s analysis to recovery of magnetization direction from magnetic field data and their publication led to a switch in acceptance of the ability to recover magnetization estimates from magnetic field data.

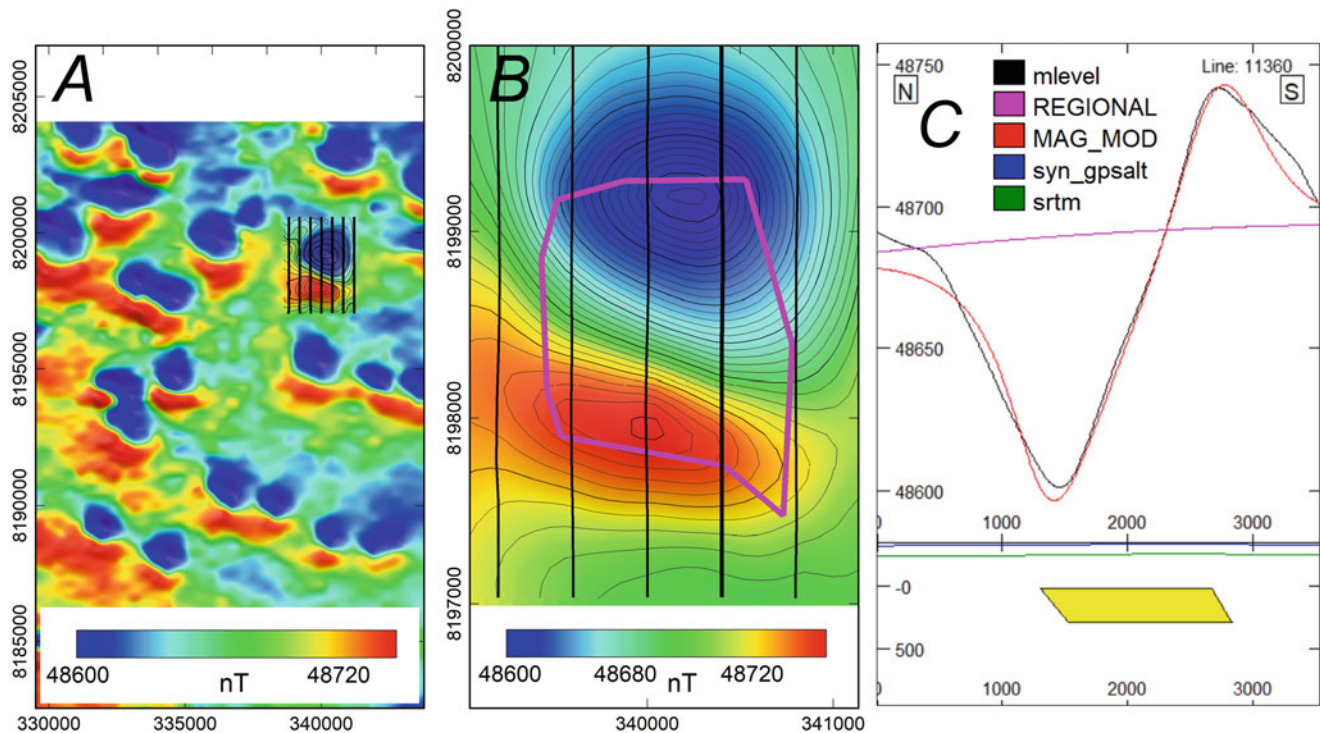
In an early three-dimensional magnetic field modelling study (see ► [“Magnetic Modeling, Theory, and Computation”](#)) Zietz and Andreasen (1967) noted that in steep inclination geomagnetic fields the declination of magnetization of their source models is indicated by the peak to trough azimuth of the magnetic anomalies, and the inclination is indicated by the peak to trough amplitude ratios. These relationships are illustrated in Fig. 2 for steep northern and southern hemisphere fields. Shallow inclination magnetizations give rise to anomalies with similar peak and trough amplitudes; steep magnetizations parallel to the field (“normal”) give rise to predominantly positive anomalies, and magnetizations anti-parallel to the field (“reverse” – see ► [“Geomagnetic Field, Polarity Reversals”](#)) give rise to predominantly negative anomalies. In the southern hemisphere, the declination of magnetization is indicated by the trough to peak azimuth, and in the northern hemisphere by the peak to trough azimuth. A specified magnetization gives rise to a consistent external field, but as shown in Fig. 2 the total magnetic intensity (TMI) anomalies vary considerably in different geomagnetic settings because of their vector addition with those differently directed fields (see ► [“Geomagnetic Field, Global Pattern”](#)). Because we can recognize the expression of magnetization direction in magnetic anomaly patterns we should expect to recover magnetization directions by magnetic field analysis or inversion. To investigate recovery of magnetization direction, we can consider the three conditions for Helbig analysis; namely, correct isolation of the magnetic field due to the magnetization, knowledge or estimation of the center of magnetization, and the influence of the shape of distribution of the magnetization.

Magnetization Contrast and Anomaly Separation

The pattern of an anomaly depends on its isolation from background and other overlapping fields in what is known as “regional-residual separation” (see ► [“Magnetic Data Enhancements and Depth Estimation”](#)). Regional-residual



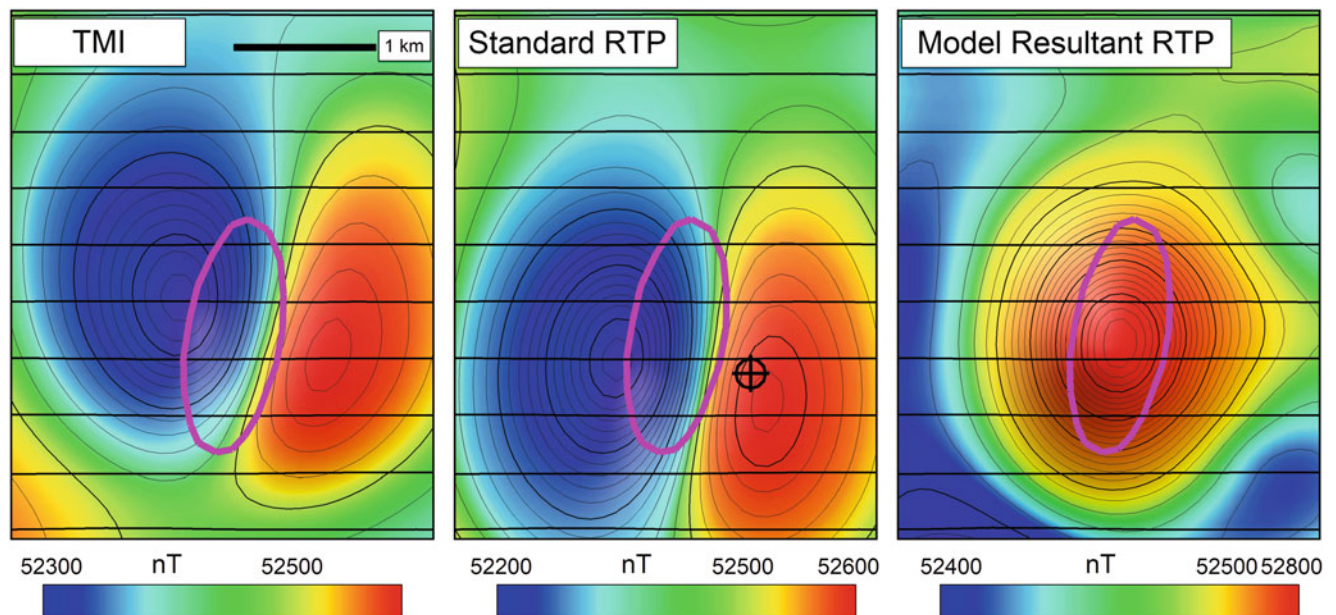
Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 2 Dipole TMI anomalies of identical magnetizations in steep southern and northern geomagnetic fields



Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 3 (a) TMI anomalies over holes in a volcanic sheet, (b) TMI computed from an inversion model of one of the holes, (c) the central section through the hole inversion model (shown in yellow)

separation is intrinsically interpretive and the output of any analytic method must be carefully checked. The choice of the regional field is also intrinsically related to the assumed background magnetization as illustrated in a study of magnetic field variations over part of the Beetaloo Basin of the Northern Territory of Australia (Foss and Dhu 2016) as shown in Fig. 3. This region is underlain by a wide volcanic

sheet delineated by a magnetic edge anomaly (beyond the area shown). The magnetic anomalies in Fig. 3 are interpreted as due to holes in the sheet; the background to each anomaly is the field above the sheet, and the direction of magnetization of the sheet is antiparallel to the magnetization derived from inversion of the hole anomalies. Because magnetic field anomalies are due to contrasts in magnetization and no



Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 4 (a) TMI anomaly at Ethabuka, Queensland, Australia, (b) Standard RTP anomaly with borehole location, (c) RTP using the

inversion magnetization direction estimate. The top of the inversion model is outlined in magenta

absolute measure of the background field is available, it is not possible to reliably map the continuous distribution of magnetization beneath an area as is sometimes claimed of voxel inversion results. The moment integrals of Helbig analysis are particularly sensitive to error in anomaly separation but an approximate extension of Helbig's theory to analysis of tensor gradients provides significant advantage in isolation of the fields for analysis (Phillips et al. 2007).

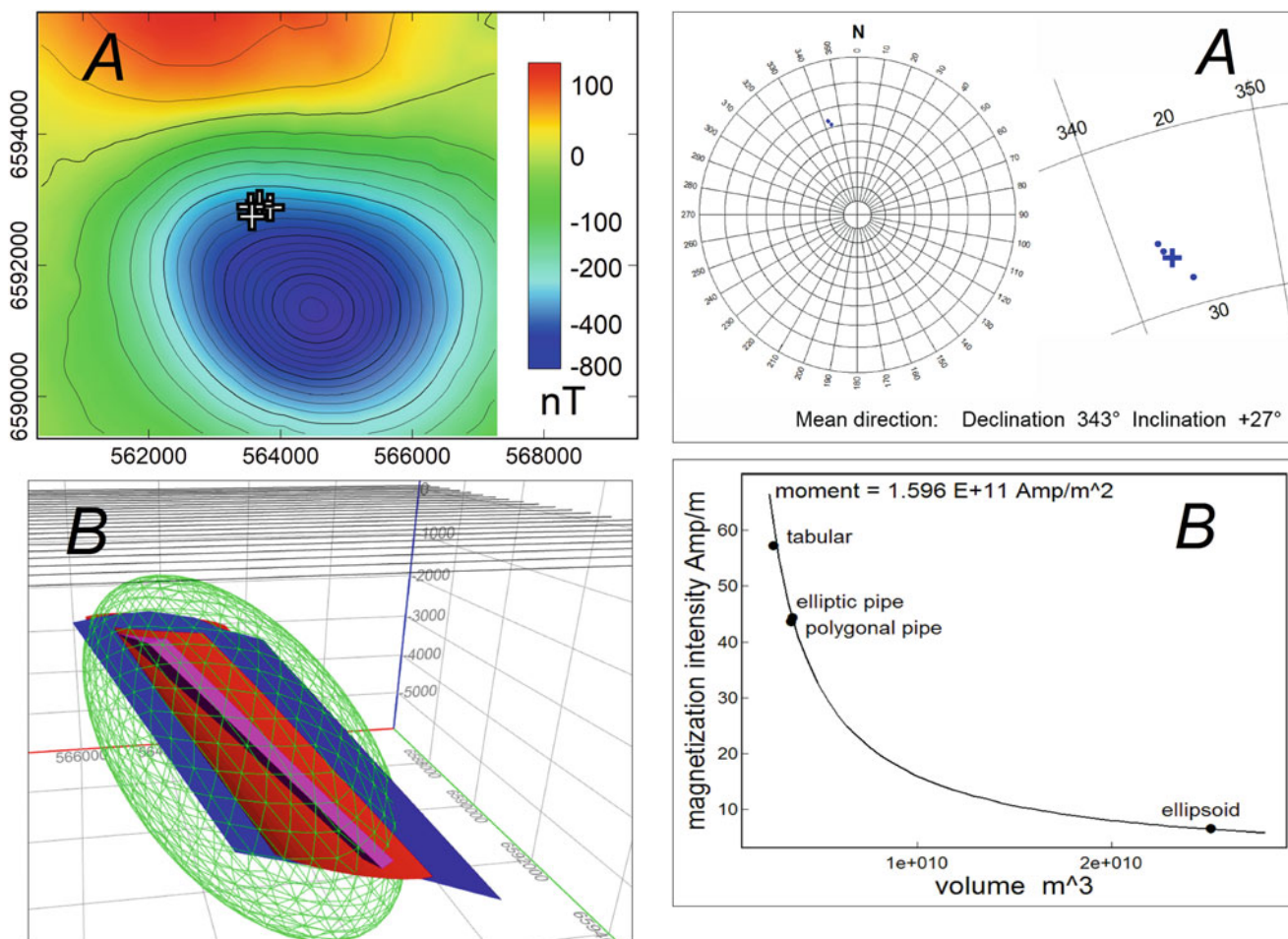
Magnetization Direction and Source Location

Helbig analysis also requires supply of the horizontal center of magnetization. Inversion can seek to simultaneously recover estimates of magnetization direction and position but error in either of these parameters gives rise to error in the other. Alternatively, various magnetic field transforms can be used to estimate the center of magnetization. Figure 4a, b show a TMI anomaly measured at Ethabuka, Queensland, Australia, and the standard RTP transform as is commonly used to locate magnetic field anomalies above their source magnetizations. The TMI anomaly pattern is markedly different to that expected of an induced anomaly, disqualifying the assumption that magnetization is in the direction of the local geomagnetic field. A borehole into the RTP peak failed to intersect the magnetization. Parametric inversion of the anomaly recovered a model of low inclination southeast-directed magnetization (consistent with the magnetization and anomaly pattern relationships described above). The top of the inversion model plotted in Fig. 4b shows that the

proposed magnetization is not tested by the borehole. Figure 4c shows the compact, essentially positive anomaly of an amended RTP transform using the resultant magnetization direction of the inversion results. Correlation of gravity anomalies and RTP transforms can provide a test of magnetization direction for bodies with co-located density and magnetization contrasts, and correlation of RTP with source-centric transforms such as the total gradient can similarly investigate whether the magnetization direction is as assumed in the RTP transform (Roest and Pilkington 1993). Dannemiller and Li (2006) used correlation between total gradient and RTP with test magnetization directions to automatically select the preferred direction. The normalized source strength (Beiki et al. 2012) provides a similar and in some cases superior central function to the total gradient transform (see ► [“Magnetic Data Enhancements and Depth Estimation”](#)).

Magnetization Direction and Source Shape

Helbig analysis strictly applies only to a dipole magnetization which is at best an approximate representation of a geological body. As already noted, recovery of magnetization direction for thin sheet magnetizations is challenging and we can expect challenges arising from the shape of distribution of magnetization to progressively increase for magnetization distributions intermediate between a dipole and a sheet. Figure 5a shows a TMI anomaly at Coompana in South Australia and Fig. 5b shows four models of different



Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 5 (a) the TMI northern anomaly at Coompana, South Australia with the center points of four alternative homogeneous magnetization models, (b) a perspective view of the models

geometry derived by inversions of the anomaly using independent forward modelling algorithms (Foss 2017). The computed fields of all the models closely match the measured field and all bodies are closely co-centered with near-identical resultant magnetization directions (Fig. 6a). Figure 6b shows a cross-plot of the volumes and intensity of magnetization of the models which both vary by a factor of 5, but their product (magnetic moment) values are consistent. These results reveal that for this body of magnetization and set of magnetic field data, estimates of the horizontal center of magnetization, mean magnetization direction, and magnetic moment are consistently recovered by inversion, but that the details of shape, intensity of magnetization, and volume are inconsistently estimated. The consistently recovered parameter values can be retained from any of the models with the models themselves discarded as improbable representations of the subsurface magnetization. Identical capabilities and limitations apply to voxel inversions. The consequence of

Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 6 (a) A stereo net plot of the four inversion model magnetization directions, (b) a cross-plot of volume and intensity of magnetization for the four models with a single magnetic moment contour

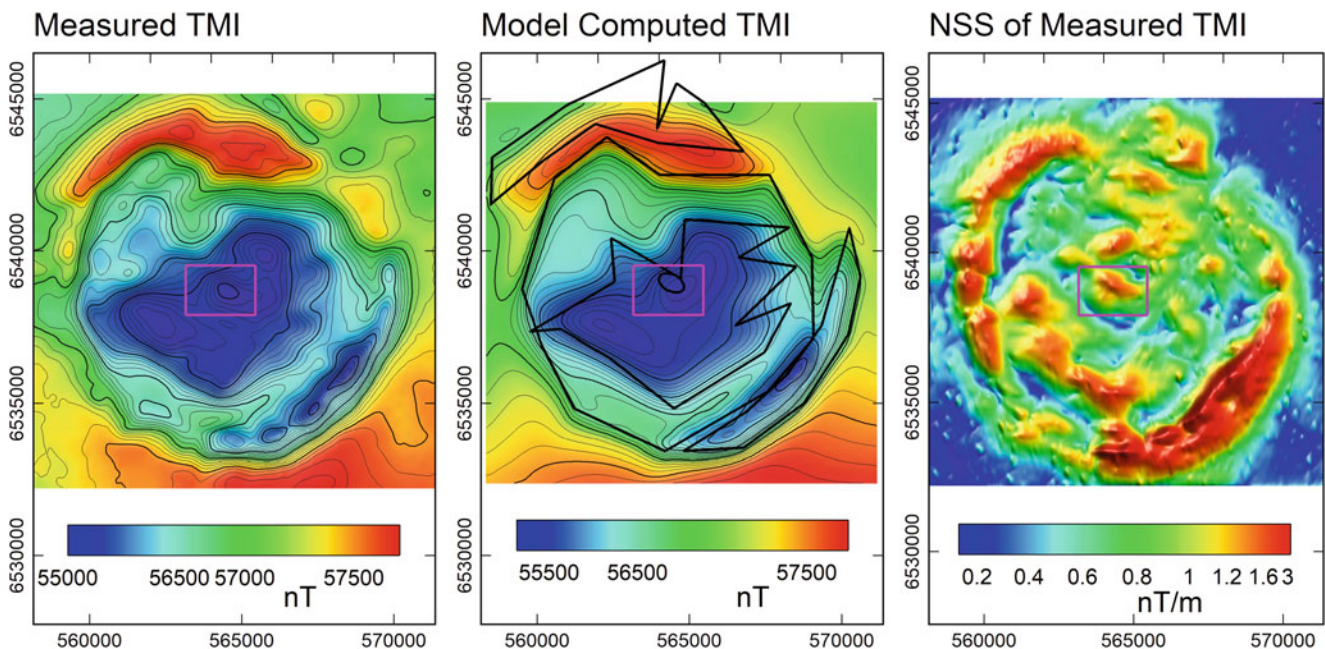
departure of shape from that of a dipole can be approximately quantified by a term nominated as “compactness” which is the ratio of the closest approach of a field measurement location to the body of the maximum extent of the body. This simple statistic does not address the full complexity of three-dimensional shape, survey extent, and magnetization direction but importantly includes the influence of distance from source. At greater distances, the magnetic field of a body appears progressively more like that of a dipole. Alternatively, for a particular body of magnetization, measured magnetic fields can be classified as near-field or far-field (analogous to definitions of electro-magnetic fields due to an antenna) with magnetization direction most easily recovered from far-fields for which source shape is less significant. Fullagar and Pears (2015) developed methods to separately address the influence of shape and magnetization direction, but these methods require either independent constraints or assumptions.

Complex Distributions of Magnetization

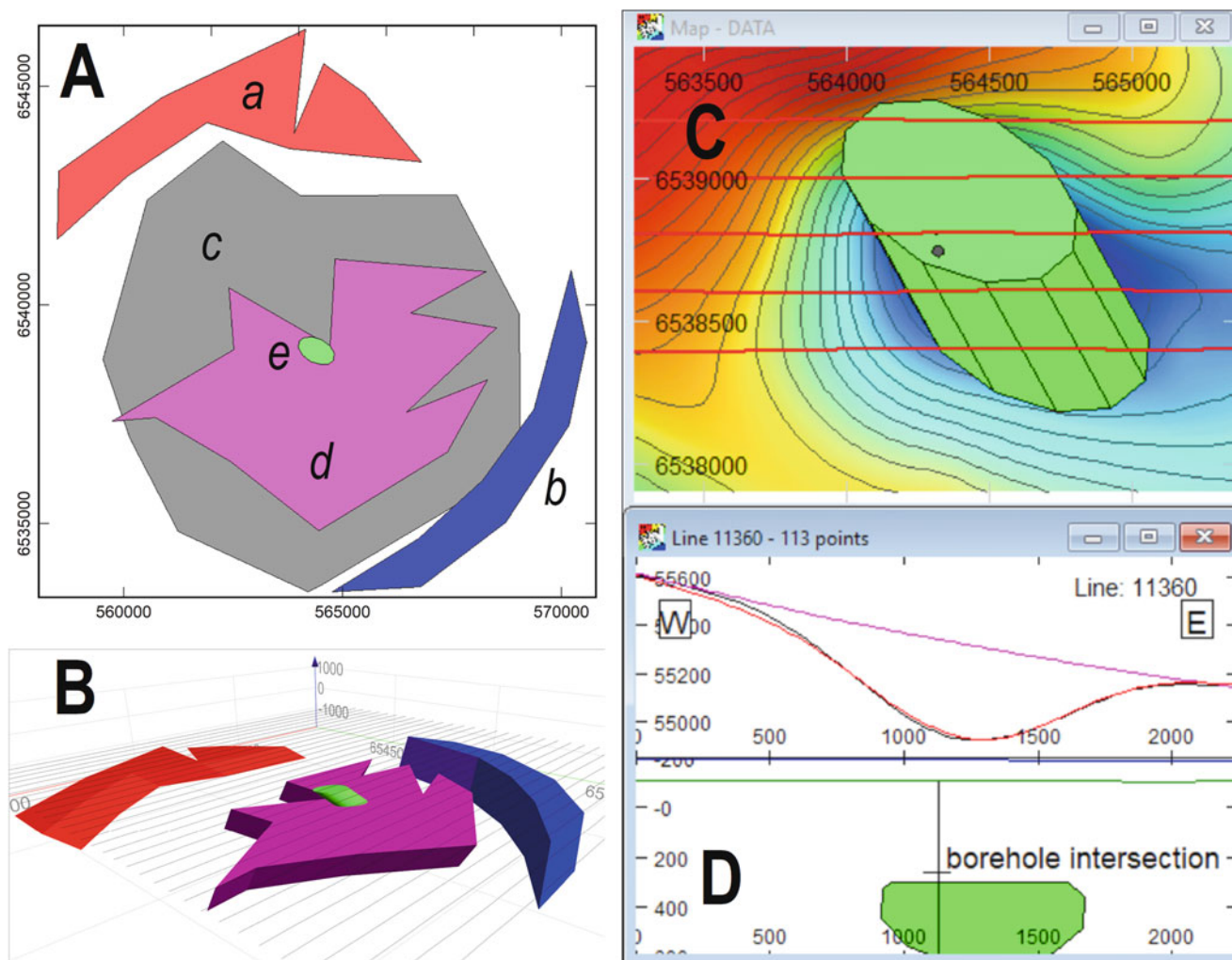
Study of complex distributions of magnetization requires application of site-specific strategies. A complex magnetic anomaly of approximately 12 km diameter measured at Coompana, South Australia (Foss 2017), and shown in Fig. 7a appears to be due to magnetizations of more than one direction. Figure 7b shows a close match to the measured field forward computed from a multi-body, multi-magnetization inversion model. The NSS transform shown in Fig. 7c highlights the circular pattern of the anomaly and a small anomaly at its Centre. In a multistage inversion, it was surmised that the anomaly can be explained with four homogeneous magnetization bodies (bodies “a” to “d” in Fig. 8a). This starting model was then inverted with gradually increased degrees of freedom. Figure 8b shows a perspective view of the resulting model which has a consistent limited depth extent and slight inward plunge, suggesting that it might represent a high-level, subvolcanic intrusive complex. To provide a drillsite depth prediction, a separate, focused inversion was performed of only the center anomaly (Fig. 8c, d) which is too small to be properly constrained in the main inversion. The main bodies of magnetization provide the background field, and the inversion model is of an excess magnetization superimposed on those magnetizations (which are themselves superimposed). This inversion successfully predicted the intersection depth.

Homogeneity and Inhomogeneity of Magnetization

If it is not feasible to reliably resolve the distribution bounds of magnetization from inversion of a far-field anomaly, it is unlikely to be feasible to determine the internal homogeneity or inhomogeneity of that magnetization. This is illustrated by two studies of the northwest Black Hill Norite anomaly as shown in Fig. 9 from the Australian Remanent Anomalies Database (<http://bit.ly/2HvhZdP>). Figure 10a shows two magnetization models which both closely match the measured anomaly. One is a homogeneous magnetization (Foss and McKenzie 2011) and the other has the same three magnetizations in similar distribution as a model by Macleod and Ellis (2016). The magnetization directions of both models are plotted in Fig. 10b. The magnetization direction of the single magnetization model is only 11° from a remanence direction measured at a nearby outcrop (the body causing the anomaly is not itself exposed). It is further from the resultant magnetization at that site but the two bodies may have different Koenigsberger ratios. The mean magnetization direction of the triple magnetization model summed from each of its component parts is close to that of the single magnetization model and there is only a small difference in their total magnetization strength. From the existing magnetic field measurements, it is difficult to discriminate between the two models and establish whether the magnetization is



Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 7 (a) the TMI southeast satellite anomaly at Coompana, South Australia, (b) TMI forward computed from the inversion model with outline of the model bodies, (c) NSS anomaly



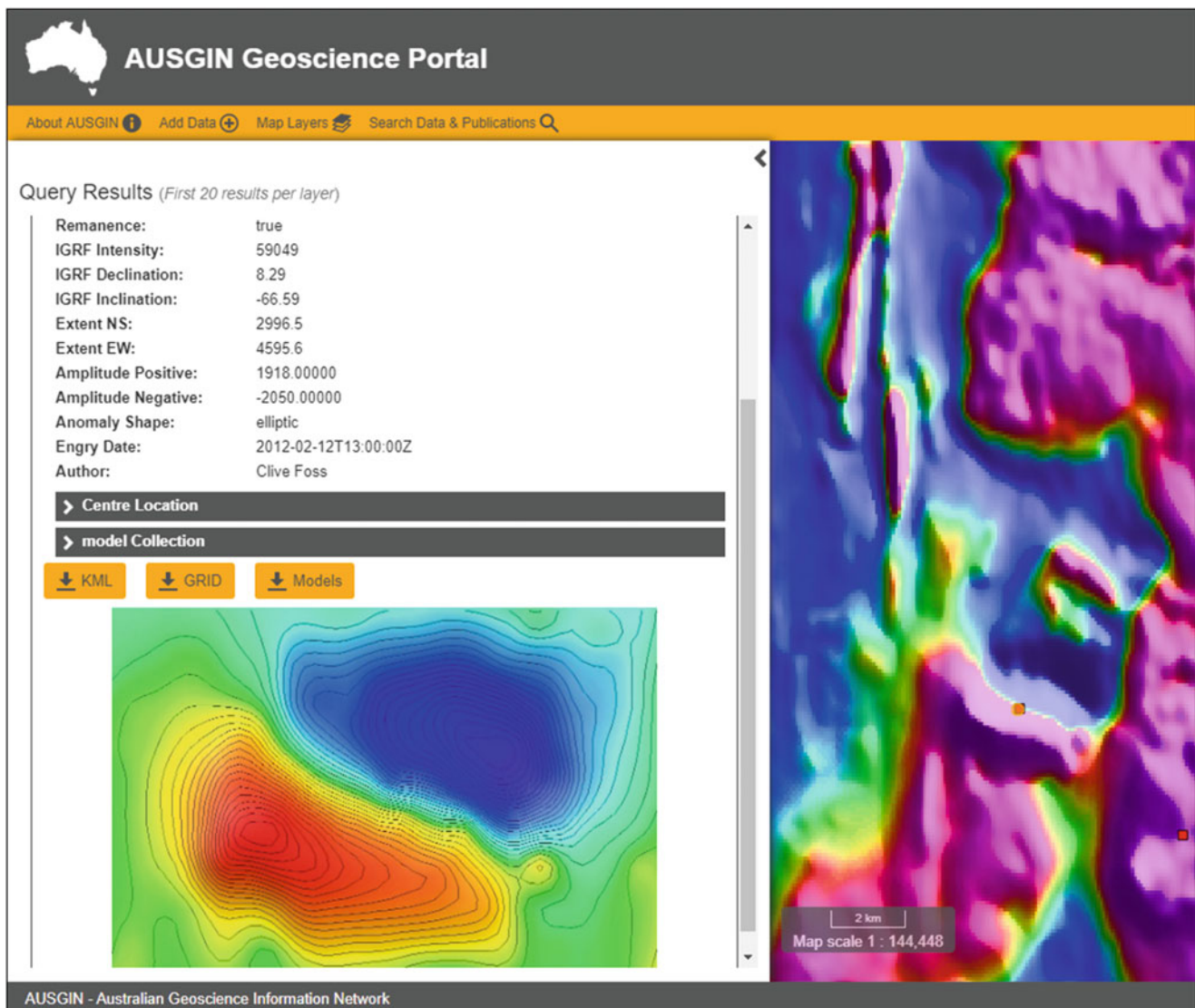
Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 8 (a) Plan view of the top of the inversion model, (b) perspective view (body “c” excluded for clarity), (c) center TMI anomaly body “e” model and drill-hole location, (d) central cross-section

homogeneous or consists of three distinct components. However, in closer proximity (with near-field measurements) it might be possible to resolve individual magnetization directions for any inhomogeneities that may be present. Sun and Li (2019) propose an analysis for multiple magnetizations after first constructing a statistical model for the proposed number of magnetizations present.

Magnetic Field Studies and Paleomagnetism

The major restrictions in application of magnetic field analysis to palaeomagnetic studies are ambiguity in recovering remanent magnetization direction from resultant magnetization and inability to resolve different remanent components from the total remanent magnetization (NRM) as is achieved in palaeomagnetic studies through alternating field and

thermal demagnetization. There are, however, opportunities for integrating magnetic field and palaeomagnetic studies arising from their very different sample volumes. Palaeomagnetic studies directly measure remanent magnetization of small volumes, in many cases collected from limited locations. Magnetic field analysis can be used to test validity in ascribing palaeomagnetic results to much larger volumes than those directly sampled. Combination of palaeomagnetic studies from outcrop and magnetic field studies of deeper magnetizations unavailable for palaeomagnetic sampling can provide more complete mapping of the distribution of magnetizations. The usefulness of magnetic field studies for mapping the distribution of magnetization depends on the precision with which magnetization directions can be recovered (in favorable circumstances to within 10° or possibly even 5°) and will be significantly enhanced if high precision can be achieved by an automated search algorithm to enable



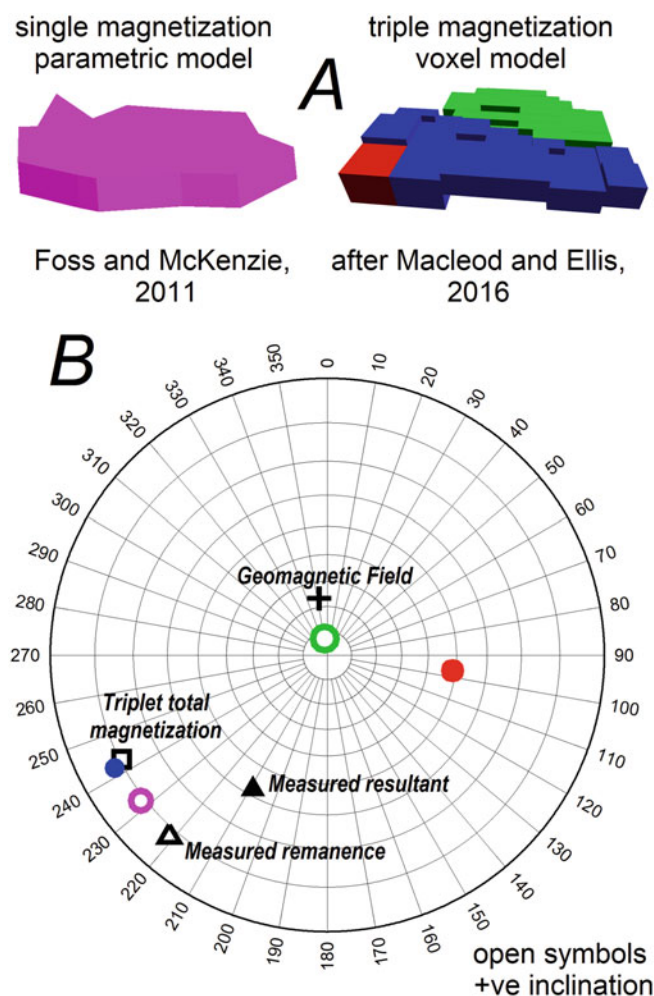
Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 9 Australian Remanent Anomaly Database image of the northwest Black Hill Norite anomaly in the AUSGIN web portal

processing of large data sets. Possible remanent magnetization directions known from a defined apparent polar wander (APW) curve (see ► [“Paleomagnetism, Polar Wander”](#)) for an area can be used to guide inversion of magnetization direction (Cordani and Shukowsky 2009; Pratt et al. 2014). As a resource for both subsurface mineral exploration and palaeomagnetic studies, the Australian Remanent Anomalies Database (<http://bit.ly/2HvhZdP>) (see ► [“Magnetic Anomaly Map, Global”](#)) has been established to compile and distribute magnetic field magnetization results, and this national coverage can be extended using global scale datasets such as EMAG2 (doi: <https://doi.org/10.7289/V5H70CVX>). At the other extreme of spatial scales, micro mapping of magnetic fields over individual mineral grains provides information

about the genesis of specific magnetizations within palaeomagnetic samples (e.g. Pastore et al. 2019).

Conclusions

In favorable circumstances, estimates of magnetization direction can be recovered from analysis or inversion of magnetic field data. Analysis is best performed on fields of compact magnetization. Analysis of magnetic fields of planar sheet magnetizations is more problematic. Inversion of magnetic field measurements close to a magnetization depends on details of the distribution of that magnetization, but at increasing distance, the magnetic field progressively approximates to



Recovery of Source Magnetization Direction from Magnetic Field Data, Fig. 10 (a) alternative single and triple magnetization inversion models for the northwest Black Hill Norite anomaly, (b) stereonet display of the inversion model magnetizations

that of a dipole and inversion is simpler (provided that the lower amplitude distal field can still be reliably isolated). More complex distributions of magnetization can also be inverted, but only where fields of individual magnetization are at least partially separated, and generally with increased scope of non-uniqueness in the results. Magnetic field analysis provides estimates of resultant magnetization. Further information is required to resolve the resultant into induced and remanent components.

Cross-References

- [Geomagnetic Field, Global Pattern](#)
- [Geomagnetic Field, Polarity Reversals](#)
- [Inverse Theory, Linear](#)
- [Magnetic Anomalies: Interpretation](#)

- [Magnetic Anomaly Map, Global](#)
- [Magnetic Data Enhancements and Depth Estimation](#)
- [Magnetic Domains](#)
- [Magnetic Modeling, Theory, and Computation](#)
- [Paleomagnetism, Polar Wander](#)
- [Remanent Magnetism](#)

Bibliography

- Beiki, Clark, Austin, Foss (2012) Estimating source location using normalized magnetic source strength calculated from magnetic gradient tensor data. *Geophysics* 77:J23–J37. <https://doi.org/10.1190/geo2011-0437.1>
- Clark (2014) Methods for determining remanent and total magnetisations of magnetic sources – a review. *Explor Geophys* 45:271–304. <https://doi.org/10.1071/EG14013>
- Cordani, Shukowsky (2009) Virtual Pole from magnetic Anomaly (VPMA): a procedure to estimate the age of a rock from its magnetic anomaly only. *J Appl Geophys* 69:96–102. <https://doi.org/10.1016/j.jappgeo.2009.07.001>
- Dannemiller, Li (2006) A new method for determination of magnetization direction. *Geophysics* 71:L69–L73. <https://doi.org/10.1190/1.2356116>
- Foss (2017) Resultant-magnetization based magnetic field interpretation. In: *Proceedings of exploration 17: sixth decennial international conference on mineral exploration* edited by Tschirhart and Thomas, paper 44, pp 637–648. <http://www.dmec.ca/Resources/Exploration-17.aspx>
- Foss, Dhu (2016) The Bark Without a dog — magnetic anomalies over holes in a volcanic sheet in the Greater McArthur Basin, NT. *ASEG Ext Abstr* 2016:1–5. <https://doi.org/10.1071/ASEG2016ab276>
- Foss, McKenzie (2011) Inversion of anomalies due to remanent magnetisation: an example from the Black Hill Norite of South Australia. *Aust J Earth Sci* 58:391–405. <https://doi.org/10.1080/08120099.2011.581310>
- Fullagar, Pears (2015) Remanent magnetisation inversion. In: *24th ASEG international geophysical conference, ASEG, extended abstracts*. <https://doi.org/10.1071/ASEG2015ab188>
- Helbig (1963) Some integrals of magnetic anomalies and their relation to the parameters of the disturbing body. *Z Geophys* 29:83–96
- Hood (1964) The Königsberger ratio and the dipping dike equation. *Geophys Prosp* 12:440–456
- Lelièvre, Oldenburg (2009) A 3D total magnetization inversion applicable when significant, complicated remanence is present. *Geophysics* 74(3):L21–L30. <https://doi.org/10.1190/1.3103249>
- MacLeod, Ellis (2016) Quantitative magnetization vector inversion. *ASEG Ext Abstr* 2016:1–6
- Paine, Haederle, Flis (2001) Using transformed TMI data to invert for remanently magnetised bodies. *Explor Geophys* 32:238–242. <https://doi.org/10.1071/EG01238>
- Pastore, Church, McEnroe (2019) Multistep parametric inversion of scanning magnetic microscopy data for modelling magnetization of multidomain magnetite. *Geochem Geophys Geosyst* 20:5334–5351. <https://doi.org/10.1029/2019GC008542>
- Phillips, Nabighian, Smith, Li (2007) Estimating locations and total magnetization vectors of compact magnetic sources from scalar, vector, or tensor magnetic measurements through combined Helbig and Euler analysis. *SEG Tech Program Ext Abstr* 26:770–774. <https://doi.org/10.1190/1.2792526>
- Pratt, McKenzie, White (2014) Remote remanence determination (RRE). *Explor Geophys* 45:314–323. <https://doi.org/10.1071/EG14031>

- Roest, Pilkington (1993) Identifying remanent magnetization effects in magnetic data. *Geophysics* 58:653–659. <https://doi.org/10.1190/1.1443449>
- Schmidt, Clark (1998) The calculation of magnetic components and moments from TMI: a case study from the tuckers igneous complex, Queensland. *Explor Geophys* 29:609–614. <https://doi.org/10.1071/EG998609>
- Sun, Li (2019) Magnetization clustering inversion – part 2: assessing the uncertainty of recovered magnetization directions. *Geophysics* 84: J17–J29. <https://doi.org/10.1190/GEO2018-0480.1>
- Zietz, Andreasen (1967) Remanent magnetization and aeromagnetic interpretation. In: *Mining geophysics*, vol 2. SEG, Tulsa, pp 569–590

Remanent Magnetism

Laurie Brown¹ and Suzanne McEnroe²

¹Department of Geosciences, University of Massachusetts, Amherst, MA, USA

²Norwegian Geological Survey, Trondheim, Norway

Definition

Remanent magnetization Permanent magnetization held in rocks and other earth materials, commonly, though not always, dating from cooling to below the Curie temperature (T_c) or Néel temperature (T_n) of the magnetic minerals in the rock, or the time of formation of the rock unit.

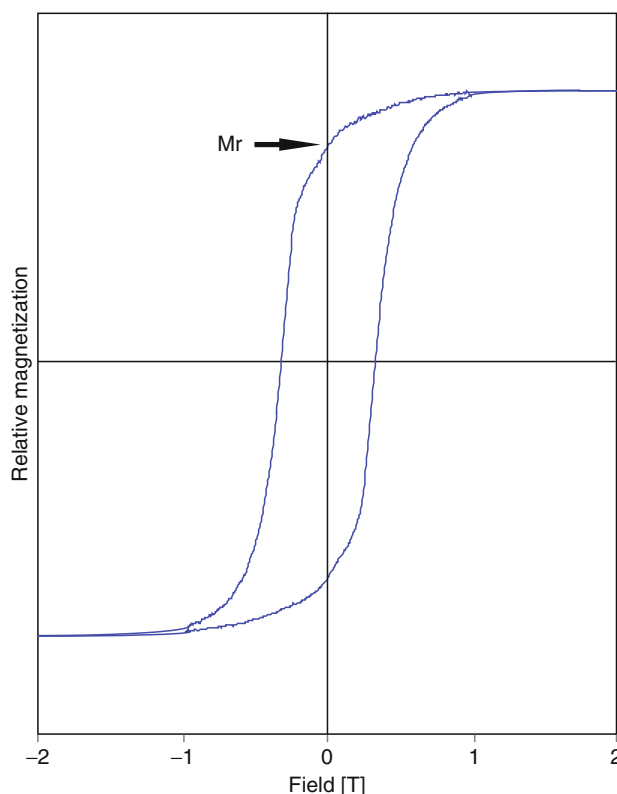
Introduction

The fact that earth materials, particularly rocks, possess a permanent magnetization has been known for some time. Early observations of rocks producing large anomalies were commonly ascribed to lightning strikes. By the mid-nineteenth century remanence was accepted as a property of some rocks, and by the early twentieth century crude measurements of the direction of magnetization were possible. During the second half of the twentieth century, the measurement of remanent magnetization (RM) became a well-established research field (see entries under ► “[Paleomagnetism, Principles](#)”). RM is a vector quantity, having both direction, usually related to the direction of the Earth’s field at the time of origin, and intensity, related to both mineral phase and the external field. Descriptions of remanent magnetization, associated paleomagnetic techniques, and results are detailed in standard paleomagnetic textbooks such as Butler (1992), McElhinny and McFadden (2000), and Tauxe (2010). More detailed discussion of the theory and development of remanent magnetization can be found in rock magnetism books, including those by Nagata (1961), Stacey and Banerjee (1974), O’Reilly (1984), and Dunlop and Özdemir (1997).

Source of RM

Remanence is a property unique to a small group of minerals possessing ferromagnetic or antiferromagnetic properties, mostly iron oxides and iron sulfides. Remanence is observable in hysteresis behavior, when, after saturation, the applied field is reversed and decreased to zero. If the magnetization does not also return to zero, but retains a measurable quantity, a remanent magnetization is present (Fig. 1). This characteristic is temperature dependent, with only some of the mineral phases possessing remanence at earth surface temperatures.

The most common and important magnetic minerals on Earth belong to the cubic oxide series magnetite ($\text{Fe}^{2+}\text{Fe}_2^{3+}\text{O}_4$) – ulvöspinel ($\text{Fe}_2^{2+}\text{Ti}^{4+}\text{O}_4$), and the rhombohedral oxide series hematite ($\text{Fe}_2^{3+}\text{O}_3$) – ilmenite ($\text{Fe}^{2+}\text{Ti}^{4+}\text{O}_3$). Both series have important solid solution versus temperature relationships, as well as magnetic complexities of key interest in the rock magnetism. Additionally, both series show remanence properties that vary with composition. Furthermore, the oxide bulk compositions in the vicinity of equilibrium tie lines between ilmenite and magnetite are found in a majority of igneous and metamorphic rocks, especially basalts. In these



Remanent Magnetism, Fig. 1 Hysteresis loop of hematite-dominated sample of the El Laco deposit, northern Chile. Horizontal axis is the external field, in Tesla; vertical axis is relative magnetization. Mr indicates the remanent magnetization left after the sample is saturated and returned to a zero field

assemblages, the slopes of the tie lines are a function of temperature, whereas tie line positions are a function of the oxygen fugacity of equilibration according to the oxygen thermobarometer (Buddington and Lindsley 1964).

End-member hematite is antiferromagnetically ordered with equal and opposite magnetic moments in alternate cation layers. However, at room temperature, a weak ferromagnetic moment is produced because the alternate moments are $\sim 0.13^\circ$ away from being perfectly antiparallel (Dzialoshinskii 1957). This effect is commonly referred to as “spin-canting” and such hematite is termed “canted--antiferromagnetic.” Many oxidized sediments contain hematite, and carry this weak, but very stable magnetization, which is mainly due to its high coercivity.

The temperature and composition phase relations of the hematite-ilmenite series are a result of a complex interaction of Fe-Ti ordering and magnetic ordering at high temperature. Slow cooling to intermediate to low temperatures produces exsolution of discrete ilmenite and hematite (Burton 1991; Harrison et al. 2000; Ghiorso and Evans 2008). Exsolved grains contain two phases, a host (hematite or ilmenite) and lamellae of the second phase. Ilmeno-hematite (hematite with ilmenite exsolution), or hemo-ilmenite (ilmenite with hematite exsolution) commonly have a strong, and very stable magnetization, referred to as “lamellar magnetism.” Lamellar magnetism is due to a defect in the magnetic moment at the interface (or contact layer) between the two phases (Robinson et al. 2002, 2004) where a ferrimagnetic substructure is present because of the different quantity of Fe^{2+} and Fe^{3+} ions in contact Fe layers, as compared to standard Fe^{3+} layers in adjacent hematite.

Intermediate members of this series rapidly cooled from high temperature produce intermediate metastable phases that are ferrimagnetic and show complicated magnetic properties near and below room temperature (Ishikawa and Akimoto 1957; Burton et al. 2008; Robinson et al. 2010) with some compositions with varied Fe-Ti ordering showing magnetic self-reversal (Harrison et al. 2005; Ishikawa and Syono 1963).

Natural Remanent Magnetization (NRM)

The magnetization held by a rock sample prior to subjection to any laboratory procedures is referred to as Natural Remanent Magnetization (NRM). It is the sum total of all remanences held in the rock and may include primary as well as secondary remanences arising from different processes. The NRM is differentiated from induced magnetization, which a rock may acquire as it sits in an external field, and loses when the field is removed. Induced magnetization is dependent on the magnetic susceptibility of the material as well as the strength of the external field. NRM is traditionally described by the manner of acquisition of the remanences involved,

leading to a number of both natural and laboratory magnetizations, as described below.

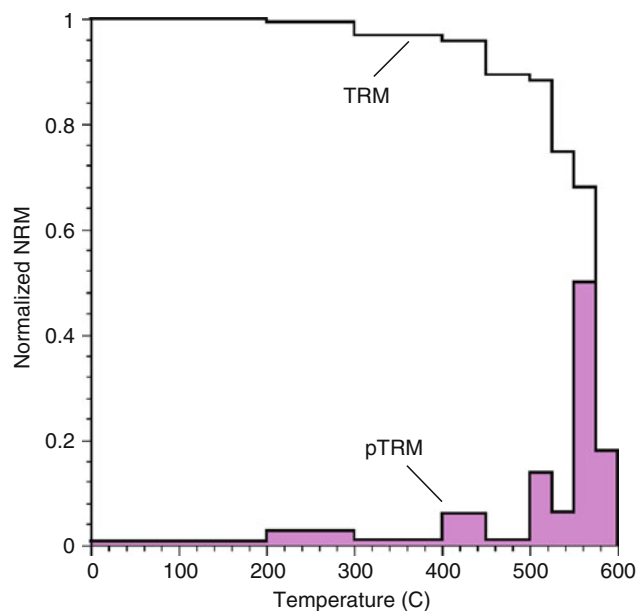
Thermoremanent Magnetization (TRM)

Igneous rocks, cooling down from a molten state through the Curie or Néel temperature (T_c or T_n), obtain a thermoremanent magnetization (TRM) parallel to the ambient field. For magnetite this temperature is 580°C while for hematite it is 675°C . At these temperatures the spontaneous nature of the ferromagnetic material occurs and remanence is blocked just below this temperature. The material must cool slightly further to the blocking temperature (T_b) before this magnetization becomes locked in with a relaxation time equivalent to geologic time spans. The direction of remanence, so important to the whole field of paleomagnetism, becomes permanent at this point, with a net direction parallel to the ambient field. The intensity of the magnetization is influenced by the size of the external field, and by the phase, composition, and grain size of the ferromagnetic minerals present.

The TRM preserved in igneous rocks is a summation of the remanences of all the ferromagnetic grains present. This means there may be a range of Curie temperatures and related blocking temperatures, giving rise to partial thermal remanence (pTRM). As a simple case, if a rock had both pure hematite and pure magnetite grains present, the hematite grains would lock in the field slightly below 675°C while the magnetite grains would provide additional remanence once the rock cooled to below 580°C . Commonly, a combination of oxide minerals of varying composition provide for remanence acquired at a range of temperatures. In Fig. 2, a plot of pTRM is shown for a basalt with low Ti-magnetite. Metamorphic rocks heated to above their Curie temperatures will have the remanence reset as the material cools down. High grade metamorphic rocks, such as granulite- and amphibolite facies rocks, heated to above $600\text{--}700^\circ\text{C}$ will have all the existing magnetic minerals remagnetized in the external field at the time of cooling, whereas rocks of lower metamorphic grade will have only minerals with T_c below the peak metamorphic temperature completely reset. Néel (1955) introduced the theory of thermoremanence; Dunlop (1977) and Dunlop and Özdemir (1997) provide detailed treatment of the theory.

Detrital Remanent Magnetization (DRM)

Sedimentary rocks obtain a remanent magnetization in a very different way. Here the detrital grains settling in a water column are influenced by the external magnetic field producing a depositional remanent magnetization (DRM).) These grains are also influenced by procedures in the sedimentary



Remanent Magnetism, Fig. 2 Thermoremanent magnetization plot from a Tertiary basalt, Meseta Lago Buenos Aires, Argentina. Upper curve labeled TRM represents the amount of magnetization left after increasing heatings up to 600 °C, with nearly 75% of the remanence being removed between 550 °C and 600 °C. Lower curve (pTRM) indicates the percentage of TRM residing within certain blocking temperatures

environment imparting a post-depositional remanence (pDRM). The general term of detrital remanent magnetization includes both processes of depositional as well as post-depositional magnetization.

When eroded ferromagnetic grains enter the sedimentary regime, they retain their magnetic moments even though the original orientation of that moment is lost. These grains, as they settle through a water column, respond to the external field, and in a mechanical response, become aligned with the influencing field over a very short time period. As redeposition experiments in the laboratory have shown, this theory is not so simple as projected (Verosub 1977). There are a number of changes that can occur once the grain is deposited, but before magnetization is locked in. These effects lead to a pDRM, or post-depositional remanent magnetization, a prevalent but complicated process (Shcherbakov and Shcherbakova 1983). Common effects are the dewatering of the sediment, compaction of the grains, and bioturbation, all of which allow for ferromagnetic particles, particularly small ones, to realign. Magnetization is not considered final until the grains are no longer able to move about or realign, occurring at a depth referred to as the lock-in depth. This gives grain size an importance in DRM, with smaller grains more susceptible to realignment, but also yielding stronger and more stable remanence. To this end fine-grained sediments, as mudstones and siltstones, retain a remanence better than coarser-grained rocks such as sandstones or conglomerates.

The notion of individual grains settling in a water body has recently been investigated (Tauxe et al. 2006) and found to be more complicated. In marine environments in particular, fine-grained sediments do not sink as discrete particles, but rather flocculate with coexisting nonmagnetic particles. The entire floc needs to be aligned with the external field, leading to variations in remanence, especially with respect to intensity, from that predicted by theory and redeposition experiments.

Chemical Remanent Magnetization (CRM)

Commonly rocks undergo chemical changes after formation, usually at temperatures below the Curie temperature of the minerals involved. These changes can be alterations to existing magnetic minerals and/or the creation of new iron oxides within the rock unit. These changes can result in new or altered remanences; this process is commonly referred to chemical remanent magnetization (CRM).

A classic example of CRM is “red beds” or sedimentary rocks with strong reddish hues due to the ubiquitous presence of hematite. Such rocks may have a DRM from the time of formation, due to the presence of detrital grains of magnetite and/or hematite. After deposition, changes brought on by differing environments, a range of aqueous solutions, or variations in oxidation state may result in weathering of existing ferromagnetic minerals, or in the deposition by precipitation of new ferromagnetic minerals. Existing magnetic minerals may break down to weaker or nonmagnetic minerals, such as hematite altering to goethite or pyrrhotite to pyrite. Ferromagnetic minerals may also alter to other magnetic minerals with different remanence characteristics. One example is the oxidation of magnetite to maghemite. Another alteration common is that of nonmagnetic grains to magnetic ones, such as the case of Fe-bearing silicate grains altering to discrete hematite grains. New ferromagnetic minerals develop, such as the formation of hematite cements and coatings, which is pervasive in red beds develops.

CRM can also be produced during progressive metamorphism, for example, when Fe-silicates breakdown to produce magnetite, in hematite-ilmenite solid solutions, or when Fe-bearing silicates react with pyrite (nonmagnetic FeS₂) to produce pyrrhotite (magnetic FeS). Lamellar magnetism mentioned earlier is a CRM because it is produced during the chemical reaction of exsolution and not at the Curie temperature of the minerals; with slow cooling this can occur hundreds of degrees below the T_c.

CRM is also recognized in igneous rocks, such as in the formation of secondary hematite in cracks and fractures in basalts or around the rims of discrete magnetite grains. In this case, the CRM is often hard to distinguish from the original TRM, as the CRM is quite stable and many of the magnetic properties of the two remanences are similar.

Viscous Remanent Magnetization (VRM)

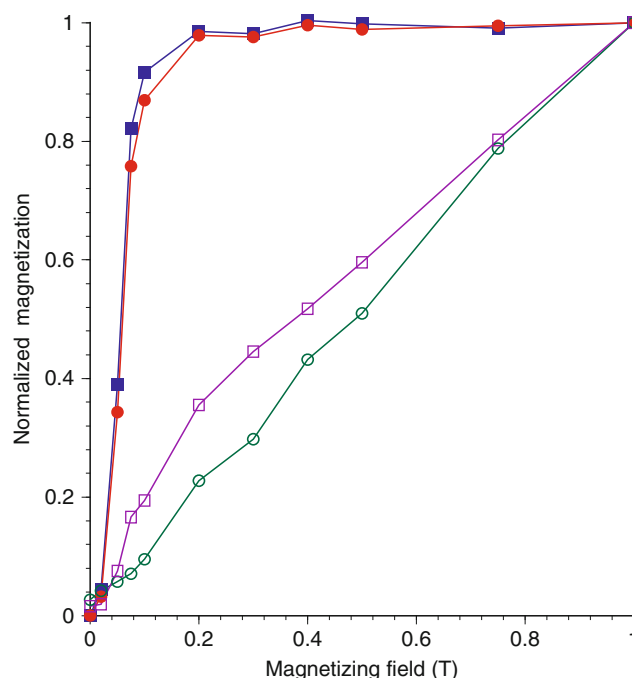
At room temperature, in low ambient fields, rocks may acquire a component of magnetization in the direction of the “present” field. This remanence acquired over long time periods in low fields, such as the Earth’s field, is referred to as viscous remanent magnetization (VRM). It is generally assumed that the strength of the VRM is related to the log of the time involved, although this is a gross simplification of the situation. The importance of other factors, including grain size, initial magnetic state, and the angle of the ambient field relative to the original remanence vector have recently been investigated and shown to also affect the strength of VRM) (Yu and Tauxe 2006). Although VRM is ubiquitous in geologic examples and is found associated with many different rocks types, in many cases it is small and can easily be removed by demagnetization techniques. Due to its soft nature, VRM is often referred to as an “overprint” commonly representing the present magnetic field and must be removed prior to investigating the primary remanence.

In natural situations where temperatures may be elevated for a considerable time, as in low-grade metamorphism, or in young oceanic lithosphere, there is commonly a thermal viscous remanent magnetization (TVRM) acquired. Here, increased temperature over extended time periods enhance the viscous remanence and produce a magnetization that may eventually replace the original NRM. Pullaiah et al. (1975) have investigated such magnetization changes over a range of temperatures and timescales, with additional discussion provided by Dunlop and Özdemir (1997).

Isothermal Remanent Magnetization (IRM)

The presence of a large magnetic field over a short time period may also produce a remanent magnetization, referred to as isothermal remanent magnetization (IRM). An IRM can be imparted in the laboratory and used to help identify properties of magnetic materials. Samples are subjected to increasingly large external fields, usually to 1 T or larger, and the resulting magnetization recorded. Differences in composition, concentration, grain size, and saturation state effect the acquisition of IRM. For example, magnetite-bearing samples will saturate in fields of 0.1–0.3 mT, while hematite-dominated rocks will require fields larger than 1 T to reach saturation (Fig. 3).

In the natural environment, a typically cited example of an IRM is the magnetization gained in a rock after being struck by lightning. The large magnetization produced by the electric currents in lightning imparts an intense, but chaotic remanence on surface rocks. First studied by Cox (1961), IRM is common in regions subject to intense thunderstorms, such as the desert southwest of the North America. In some cases, the effect of the IRM can be removed by demagnetization; in



Remanent Magnetism, Fig. 3 Normalized IRM acquisition in fields up to 1 T; closed symbols are Jurassic diabase dikes with magnetite; open circles are Jurassic mudstones with hematite

other samples, the magnetization has been totally reset, giving rise to paleo-magnetic sites with random directions (Tauxe et al. 2003).

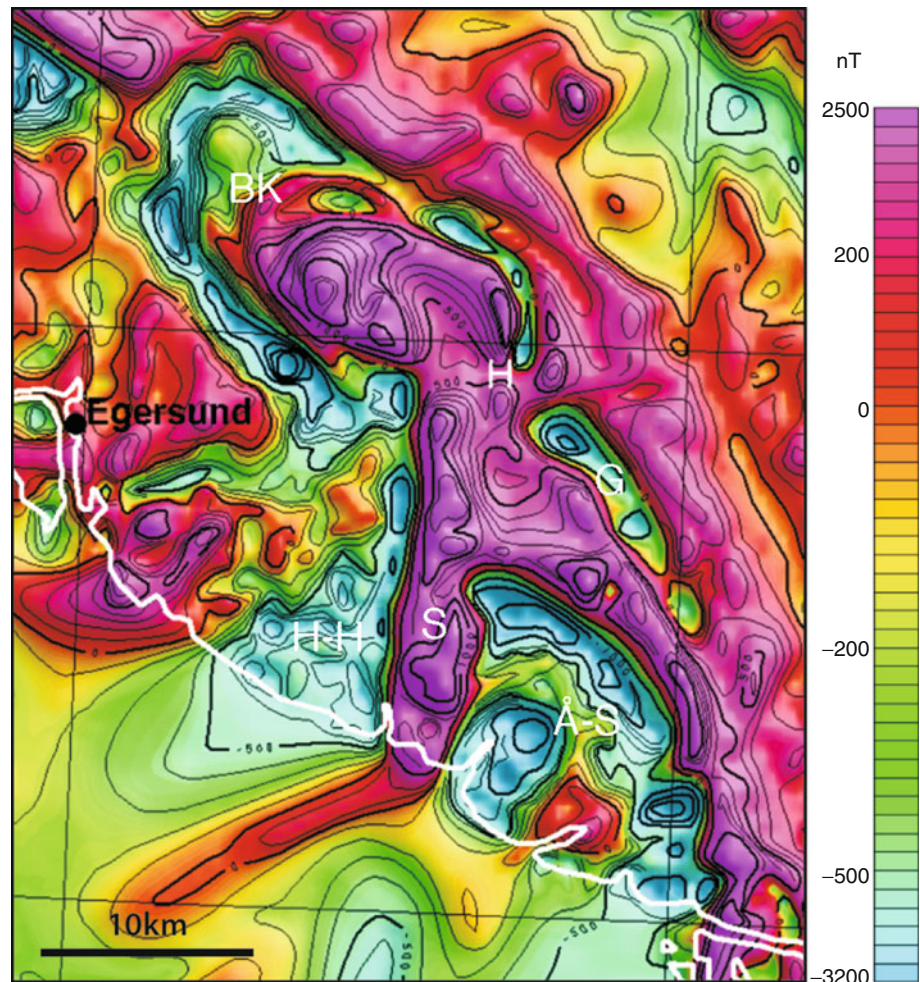
However, this type of IRM is commonly accompanied by high heat; therefore, the term is slightly misused.

Applications of Remanent Magnetization

RM is used extensively in the disciplines of magnetism, including paleomagnetism, rock magnetism, and the study of magnetic anomalies. The study of paleomagnetism centers on the identification and description of the remanent field held in rocks. Results from the measurement of RM are used widely in the study of plate tectonics, structure, regional tectonics, stratigraphy, volcanology, and archeology. Positions of the earth’s magnetic field at specific times in the geologic past provide us invaluable information on the position of continental bodies in the past, the interrelationship between different continents, and the paths they have followed in subsequent motions across the earth’s surface. RM can provide age information by comparison to established reversal timescales, as well as relative ages within a region. Regional tectonics, local rotations, and timing of folding can all be investigated using the remanence recorded in suitable rock units. Stratigraphic studies, especially in the wealth of deep-sea sediment cores from international drilling projects, lake sediments, or in thick sedimentary sequences on

Remanent Magnetism,

Fig. 4 Aeromagnetic anomaly map from a fixed-wing survey over the Rogaland region in southwestern Norway. The magnetic total field was reduced to anomaly values by subtracting the International Geomagnetic Reference Field from the total field of 1965. Color shades: pink, large positive magnetic anomalies; blue, strong negative anomalies; yellow and green, intermediate values. Letters refer to specific bodies within the Rogaland Igneous Complex. (Redrawn from McEnroe et al. 2001)



land, can be established and compared using the recorded remanence.

Experimental studies in rock magnetism make excellent use of the remanent magnetization properties of rocks and minerals. Studies of hysteresis behavior, magnetization at high and low temperatures, and responses to acquisition and removal of laboratory remanences all provide important information on oxide composition, grain size, and magnetic properties of a wide range of natural and synthetic materials.

Remanent magnetism plays an important part in the interpretation of magnetic anomalies. Well established in the 1960s is the study of marine magnetic anomalies from the seafloor, remanent polarity of extruded basalts played a pivotal role in the interpretation of observed anomalies (Vine and Matthews 1963; Heirtzler et al. 1968). The use of the polarity sequence preserved in the ocean crust allows for dating of the seafloor as well as calculations of spreading rates and investigations of plate histories. Anomalies on land show considerable complexities and many are assumed to be entirely related to induced fields due to the present magnetic field.

But, in many cases, there is a significant component of remanent magnetization interacting with the induced field, and depending on the remanent direction and intensity, enhancing or at times annihilating the induced field (McEnroe et al. 2009). An excellent case of remanent magnetization producing strong aeromagnetic anomalies comes from the Rogaland Igneous Complex in southern Norway (McEnroe et al. 2008), where anorthosites and a layered intrusion have intense remanence roughly antiparallel to the present field, producing large negative anomalies (Fig. 4).

Exploration for natural resources by aeromagnetic surveys has been widely used since the 1950s. The classic work by Balsley and Buddington (1958) and later by McEnroe and Brown (2000) in the oxide rich area of the Adirondack Mountains of New York required a detailed knowledge of the RM for accurate interpretation of the data. Understanding the contribution of RM to magnetic anomalies will be crucial for future exploration as the accuracy and resolution of these surveys increase. Magnetic surveys will also be at the forefront for geological mapping of Earth and other planets.

Summary

Remanent magnetization is a permanent magnetization residing in earth materials that is gained when the material is formed or altered in the presence of the Earth's magnetic field. Measurable RM is found in most rocks, ranging from metamorphic gneisses to lava flows to marine and lake sediments. The remanence is carried by iron oxides and sulfides, but magnetite and hematite are by far the most common and the most important remanence carriers. RM is categorized and described by various methods of acquisition; TRM for remanence produced as minerals cool through their Curie temperatures, DRM for the remanence obtained as detrital grains of magnetic minerals align during the deposition process, and CRM for the process where remanence develops from chemical changes in the rock. RM forms the basis for research in paleomagnetism, where the direction, intensity, and age of the remanence are determined and interpreted. It is also useful in experimental studies of rock magnetism, and in the investigation of anomalies in global, regional, and local magnetic field measurements.

Cross-References

- [Magnetic Anomalies: Interpretation](#)
- [Paleomagnetic Field Intensity](#)
- [Paleomagnetism, Magnetostratigraphy](#)
- [Paleomagnetism, Measurement Techniques and Instrumentation](#)
- [Paleomagnetism, Polar Wander](#)
- [Paleomagnetism, Principles](#)

Bibliography

- Balsley JR, Buddington AF (1958) Iron-titanium oxides minerals, rocks and aeromagnetic anomalies of the Adirondack area, New York. *Econ Geol* 53:777–805
- Buddington AF, Lindsley DH (1964) Iron-titanium oxide minerals and synthetic equivalents. *J Petrol* 5:310–357
- Burton BP (1991) Interplay of chemical and magnetic ordering. In: Lindsley DH (ed) *Oxide minerals*. Mineralogical Society of America, Washington, DC, pp 303–321
- Burton BP, Robinson P, McEnroe SA, Fabian K, Boffa Ballaran T (2008) A low-temperature phase diagram for ilmenite-rich compositions in the system $\text{Fe}_2\text{O}_3\text{-FeTiO}_3$. *Am Mineral* 93:1260–1272
- Butler RF (1992) *Paleomagnetism: magnetic domains to geologic terranes*. Blackwell Scientific, Boston. On-line version: <http://www.geo.arizona.edu/Paleomag/book/>
- Cox A (1961) Anomalous remanent magnetization of basalt. *US Geol Surv Bull* 1083E:131–160
- Dunlop DJ (ed) (1977) *Origin of thermoremanent magnetization*. Center for Academic Publications Japan, Tokyo
- Dunlop DJ, Özdemir Ö (1997) *Rock magnetism – fundamentals and frontiers*. Cambridge University Press, Cambridge
- Dzialoshinskii IE (1957) Thermodynamic theory of “weak” ferromagnetism in antiferromagnetic substances. *Sov J Exp Theor Phys* 5:1259–1272
- Ghiorso MS, Evans B (2008) Thermodynamics of rhombohedral oxide solid solutions and a revision of the Fe-Ti two-oxide geothermometer and oxygen-barometer. *Am J Sci* 308:957–1039
- Harrison RJ, Becker U, Redfern SAT (2000) Thermodynamics of the R to R_c phase transition in the ilmenite-hematite solid solution. *Am Mineral* 85:1694–1705
- Harrison RJ, Kasama T, White T, Simpson ET, Dunn-Borkowski RE (2005) Origin of self-reversed thermoremanent magnetization. *Phys Rev Lett* 95. <https://doi.org/10.1103/PhysRevLett95.268501>
- Heirtzler JR, Dickson GO, Herron EM, Pitman WC III, Le Pichon X (1968) Marine magnetic anomalies, geomagnetic field reversals, and motions of the ocean floor and continents. *J Geophys Res* 73:2119–2136
- Ishikawa Y, Akimoto S (1957) Magnetic properties of the $\text{FeTiO}_3\text{-Fe}_2\text{O}_3$ solid solution series. *J Phys Soc Jpn* 12:1083–1098
- Ishikawa Y, Syono Y (1963) Order-disorder transformation and reverse thermoremanent magnetization in the $\text{FeTiO}_3\text{-Fe}_2\text{O}_3$ system. *J Phys Chem Solids* 24:517–528
- McElhinny MW, McFadden PL (2000) *Paleomagnetism: continents and oceans*. Academic, San Diego
- McEnroe SA, Brown LL (2000) A closer look at remanence-dominated anomalies: Rock-magnetic properties and magnetic mineralogy of the Russell Belt microcline-sillimanite gneisses, Northwest Adirondacks Mountains, New York. *J Geophys Res* 105:16437–16456
- McEnroe SA, Robinson P, Panish P (2001) Aeromagnetic anomalies, magnetic petrology and rock magnetism of hemioilmenite- and magnetite-rich cumulates from the Sokndal Region, South Rogaland, Norway. *Am Mineral* 86:1447–1468
- McEnroe SA, Brown LL, Robinson P (2008) Remanent and induced magnetic anomalies over a layered intrusion: effects from crystal fractionation and magma recharge. *Tectonophysics*. <https://doi.org/10.1016/j.tecto.2008.11.021>
- McEnroe SA, Fabian K, Robinson P, Gaina C, Brown LL (2009) Crustal magnetism, lamellar magnetism and rocks that remember. *Elements*. <https://doi.org/10.2113/gselements.5.4.241>
- Nagata T (1961) *Rock magnetism*. Maruzen, Tokyo
- Néel L (1955) Some theoretical aspects of rock magnetism. *Adv Phys* 4:191–243
- O'Reilly W (1984) *Rock and mineral magnetism*. Blackie, Glasgow
- Pullaiah GE, Irving E, Buchan KL, Dunlop DJ (1975) Magnetization changes caused by burial and uplift. *Earth Planet Sci Lett* 28:133–143
- Robinson P, Harrison RJ, McEnroe SA, Hargraves R (2002) Lamellar magnetism in the hematite-ilmenite series as an explanation for strong remanent magnetization. *Nature* 418:517–520
- Robinson P, Harrison RJ, McEnroe SA, Hargraves R (2004) Nature and origin of lamellar magnetism in the hematite-ilmenite series. *Am Mineral* 89:725–747
- Robinson P, Fabian K, McEnroe SA (2010) The geometry of ionic arrangements and magnetic interactions in ordered ferri-ilmenite solid solutions and its effect on low-temperature magnetic behavior. *Geochem Geophys Geosyst*. <https://doi.org/10.1029/2009GC002858>
- Shcherbakov V, Shcherbakova V (1983) On the theory of depositional remanent magnetization in sedimentary rocks. *Geophys Surv* 5:369–380
- Stacey FD, Banerjee SK (1974) *The physical principles of rock magnetism*. Elsevier, Amsterdam
- Tauxe L (2010) *Essentials of paleomagnetism*. University of California Press, Berkeley. On-line version: <http://magician.ucsd.edu/Essentials/index.html>

Tauxe L, Constable C, Johnson C, Miller W, Staudigel H (2003) Paleomagnetism of the Southwestern U.S.A. recorded by 0–5 Ma igneous rocks. *Geochem Geophys Geosyst.* <https://doi.org/10.1029/2002GC000343>

Tauxe L, Steindorf JL, Harris A (2006) Depositional remanent magnetization: toward an improved theoretical and experimental foundation. *Earth Planet Sci Lett* 244:515–529

Verosub KL (1977) Depositional and post-depositional processes in the magnetization of sediments. *Rev Geophys Space Phys* 15:129–143

Vine FJ, Matthews DH (1963) Magnetic anomalies over ocean ridges. *Nature* 199:947–949

Yu Y, Tauxe L (2006) Acquisition of viscous remanent magnetization. *Phys Earth Planet Inter* 159:32–42

Remote Sensing and GIS Techniques for Tectonic Studies

Semere Solomon¹ and Woldai Ghebream²

¹Det Norske Veritas, DNV Research and Innovation, Høvik, Norway

²Department of Geology and Environmental Science, University of Akron, Akron, OH, USA

Synonyms

Remote sensing data; Space imagery

Definition

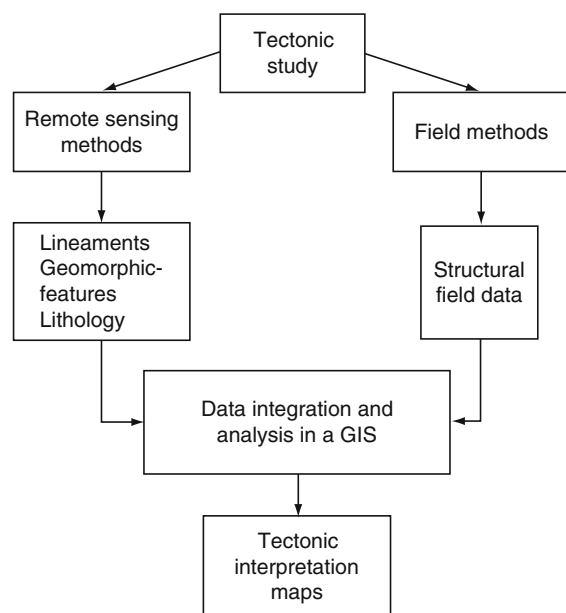
Remote sensing	Refers to small- or large-scale acquisition of information of an object or phenomenon (e.g., earth surface features), by the use of either recording or real-time sensing device (s) such as by way of aircraft, spacecraft, satellite, buoy, or ship that are not in physical contact with the object.
Geographic information system	Refers to spatial data management and analysis tools that can assist users in organizing, storing, editing, analyzing, and displaying positional and attribute information about geographical data.
Lineaments	Any linear features that can be picked out as lines (appearing as such or evident because of contrasts in terrain or ground cover, tone, pattern, size, etc.) in aerial or space imagery.
Tectonics	A field of study within geology concerned generally with the structures within the lithosphere of the Earth (or other planets) and particularly with the forces and movements that have operated in a region to create these structures.

Neotectonics	A subdiscipline of tectonics which deals with the current or recent geologic deformations.
Active fault	A planar fracture in rocks with visible displacement or seismic activity that has occurred during the geologically recent period, as opposed to passive faults that show no movements since their formation.
Lithology	The scientific study and description of rocks/rock types in terms of compositional and physical character of the rock.
Geomorphic features	Landform features such as rivers and peneplains formed by geomorphic processes (e.g., erosion and mass-wasting) that shape them.
Strain meter	An instrument used by geophysicists to measure the deformation of the Earth in the field.

Introduction

Mapping of lineaments from various space imagery are a commonly used initial step in tectonic studies. The surface expression of geological structures such as faults, joints, dykes, veins, and straight routes of rivers are often displayed or represented in the form of lineaments in aerial photographs or on remotely sensed images. In addition, remote sensing data are useful for detecting lithology and geomorphic features, which are, in turn, important in assessing the mechanisms of faulting. The most common and widely used remote sensing techniques include: Satellite Pour l’Observation de la Terre (SPOT), which is a high-resolution optical imaging Earth Observation Satellite (EOS), Landsat satellite both multispectral scanner (MSS), and thematic mapper (TM) including enhanced thematic mapper (ETM), Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), Shuttle Radar Topographic Mission (SRTM) and/or *SAR Interferometry* (InSAR), and Global Positioning System (GPS). Data acquired from optical, thermal, radar images or digital terrain elevation at the same or different sensor platforms or devices are all useful for tectonic studies.

Lineament mapping and analysis is a combined effort of technology and field evidence (Fig. 1). The availability of multispectral and multi-sensor data with synoptic coverage and use of different image enhancement techniques provides an opportunity to prepare more reliable and comprehensive lineament maps useful for tectonic studies. Field studies or ground truthing of orientation measurements of structural features and identifying rock types provide support for verifying remote sensing derived tectonic interpretation. Geographic information system (GIS) facilitates integration



Remote Sensing and GIS Techniques for Tectonic Studies, Fig. 1 A simplified flow chart showing the methodology of tectonic study using remote sensing and field data in a GIS framework

of all data types and provides a strong basis for tectonic interpretation of structural features.

Data and Method

Mapping of geological structures from various space imageries is a commonly used initial step in tectonic studies. The routine procedure for extracting tectonic features from digital remote sensing data usually involves initial digital image enhancement followed by manual interpretation (Suzen and Toprak 1998). There have been significant approaches for the evaluation and automatic detection of lineaments and curvilinear features from satellite images (Karnieli et al. 1996). However, the human expert judgment still remains to be an asset for lineament detection and interpretation.

Satellite-based remote sensing data may be optical image (e.g., SPOT and Landsat), which usually senses the behavior of visible, ultraviolet, and infrared light components of the electromagnetic energy used in imaging. The data could also be thermal image in which the devices sense the thermal energy of the electromagnetic spectrum. The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) is an example of modern remote sensing device that senses the behavior of visible, ultraviolet, infrared light, and thermal infrared components of the electromagnetic energy spectrum. The remote sensing satellite may also acquire radar images such as the Shuttle Radar Topographic Mission (SRTM) and/or Interferometric Synthetic Aperture Radar (InSAR) in which the system has a transmitter that

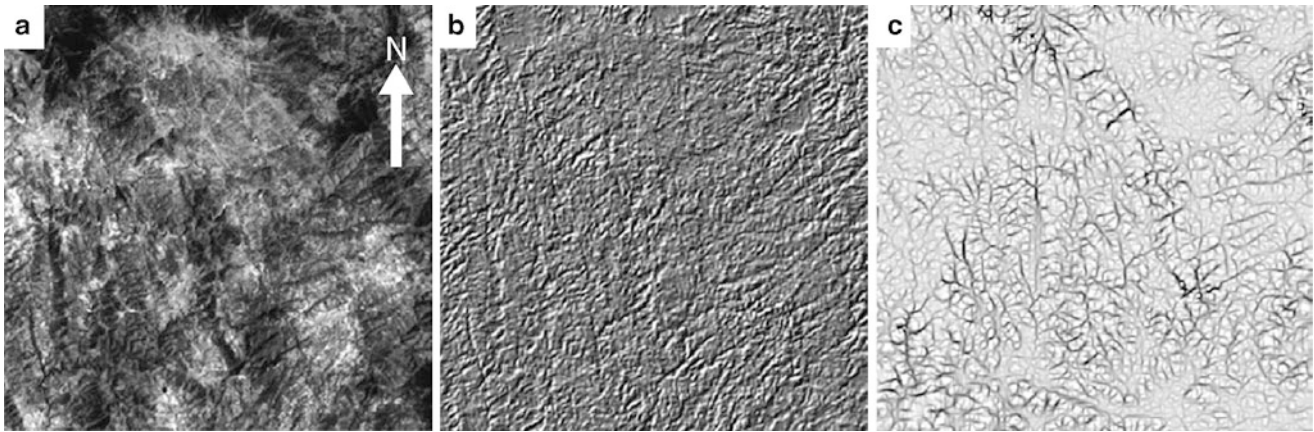
emits microwaves or radio waves of the electromagnetic energy to detect objects. The Global Positioning System (GPS) is a US space-based global navigation satellite system. GPS is made up of three parts: satellites orbiting the Earth; controlling and monitoring stations on Earth; and the GPS receiver that reads broadcasted signals from space sent by GPS satellites. Each GPS receiver then provides three-dimensional location based on latitude, longitude, and altitude, and the time that are useful for tectonic studies.

Digital elevation models (DEM) may be prepared in a number of ways, but they are frequently obtained by remote sensing rather than by direct survey. One powerful technique for generating digital elevation models is InSAR: two passes of a radar satellite (such as RADARSAT-1 or TerraSAR-X), or a single pass if the satellite is equipped with two antennas (like the SRTM instrumentation), suffice to generate a digital elevation map with extensive areal coverage and high resolution (10–30 m). Alternatively, other kinds of stereoscopic pairs can be employed using routine photogrammetric techniques, where two optical images acquired with different angles taken from the same pass of an airplane or an EOS such as the high-resolution stereo of SPOT5 or the visible and near-infrared band of ASTER (e.g., Hirano et al. 2003).

Remote Sensing

The surface expression of lineaments displayed in aerial photographs or remote sensing data (Fig. 2) may be passive faults, showing no movements since their formation or may be active faults with movements created by either earthquakes or ongoing uplifting. Earthquakes and uplifts are related to tectonic movements in the subsurface. The surface expression of these movements as structures visible on satellite imageries can be mapped and used to investigate the tectonic movements. Time series thermal satellite remote sensing datasets were used to detect pressure buildup due to tectonic activities and associated subsurface degassing that created changes in the thermal regime prior to an earthquake event (Saraf and Choudhury 2005). Such studies are useful in locating earthquakes' epicenters and understanding the fault mechanics involved in the deformation during earthquake occurrences and uplifting processes.

DEM acquired from satellites such as SRTM data or optical images are useful for derivation of topographic parameters such as slope and surface curvatures (geomorphic indices) for mapping drainage networks and lineament extraction (Zizioli 2008; Demirkesen 2008). These morphotectonic parameters can be used as indicators of active tectonics in a region. For instance, rivers are sensitive to changes in tectonic deformation, adjusting their routes over different periods of time depending on the physical properties of the host rocks, climatic effects, and tectonic



Remote Sensing and GIS Techniques for Tectonic Studies, Fig. 2 Different types of remote sensing data showing conspicuous NE-SW trending lineaments from 12 km $\times$ 12 km area (a) SPOT band

3, (b) digitized enhanced Landsat TM band 5, and (c) minimum curvature image derived from SRTM digital elevation model. The lineaments represent both extensional and strike-slip tectonics

activity. Thus, the drainage system of a region records the evolution of tectonic deformation. Quantitative measurements of a number of geomorphic indices are commonly used as a reconnaissance tool in tectonic geomorphology studies to identify areas experiencing tectonic deformation. Combinations of different geomorphic indices derived from DEM data allow the quantification of surface deformation both on maps and on stream profiles. It is possible to assess and quantify the rate of uplift by studying the influence of faults on stream geometry and changes in direction as a result of neo-tectonic activity (Gloaguen et al. 2008). Such thematic geomorphic indices can easily be incorporated into a GIS tool for further analysis and interpretation.

Interferometric Synthetic Aperture Radar (InSAR) is a radar technique used in geodesy and remote sensing. This geodetic method measures the line-of-sight displacements and uses two or more synthetic aperture radar (SAR) images to generate maps of surface deformation or digital elevation, using differences in the phase of the waves returning to the satellite, or aircraft. The technique can potentially measure millimeter-scale or even lower changes in deformation over time spans of days to years. InSAR can be used to measure ground movements due to earthquakes and to monitor creep and strain accumulation on faults (e.g., Stramondo et al. 2005). Due to its easy integration within a GIS environment, DEM data is important in such analyses. Time series data, a sequence of space imagery shots taken at different times, is a requirement in such studies. Then the changes are detected at observation points that have a fixed geographical location for later comparison and monitoring purposes.

GPS measurements, among other applications, are now in use to determine the motion of the Earth's tectonic plates and deformation around active faults and volcanoes. It provides a powerful means to directly measure the kinematic pattern of present-day crustal deformation by setting up several regional

GPS network stations and monitoring the changes in geographical locations and heights at the control points (Jouanne et al. 1998). GPS can measure movement of faults or tectonic plates to within the precision of a few millimeters to approximately 1 cm over baseline separations of hundreds of meters to thousands of kilometers (Segall and Davis 1997). The three-dimensional nature of GPS measurements allows one to determine vertical as well as horizontal displacement at the same time and place. Field methods such as the use of strain meters can provide far greater strain sensitivity than does GPS. However, strain meters cannot offer the spatial coverage and long-term stability as GPS do. InSAR measurements are tremendously exciting because of their unparalleled spatial coverage. InSAR and GPS are complementary in that GPS provides long-term stability, vector displacements, and better temporal coverage as compared to the extensive spatial coverage provided by SAR/InSAR and thus results from one system are directly relevant to the others (Segall and Davis 1997).

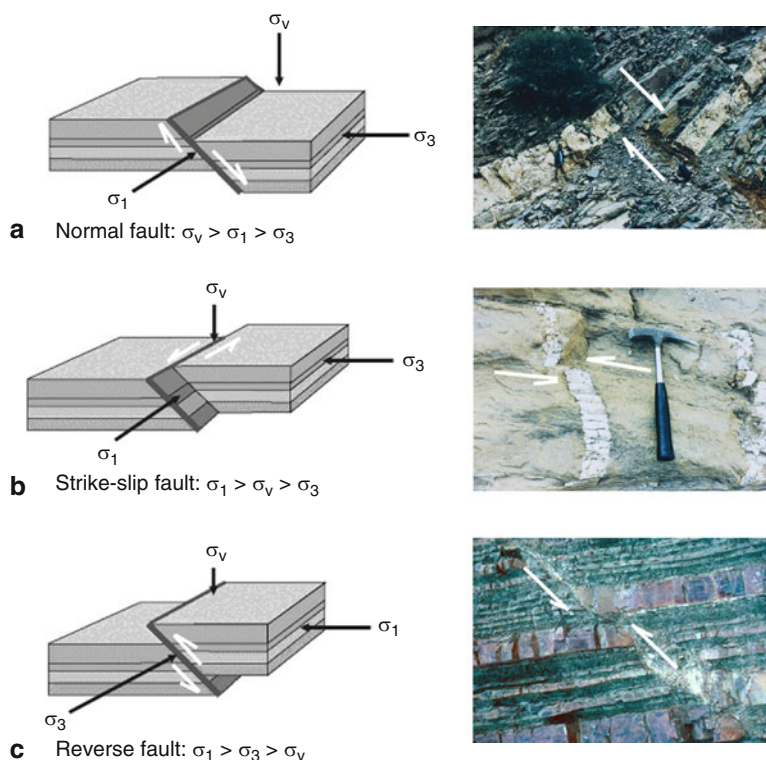
Tectonic Studies

Tectonics deals with large-scale deformations of the Earth's lithosphere. The lithosphere is broken up into several tectonic plates that move relative to one another. Plate tectonics theory explains the formation of the plates and the origin of the forces responsible for the motion of these plates. There are three types of plate boundaries characterized by their relative motion: each having different types of surface phenomena including volcanic activities, earthquakes, and faults. The tectonic styles at the different plate boundaries are:

- Extensional tectonics that results in the formation of normal faults (Fig. 3) and associated tectonic processes due to the thinning/stretching of the lithosphere.

Remote Sensing and GIS Techniques for Tectonic Studies, Fig. 3

Fault block models and their equivalent outcrop-scale structural features from field study formed by different tectonic processes: (a) Normal faults, (b) strike-slip faults, and (c) thrust faults. In the sketches, the maximum principal stress (σ_1), the vertical stress (σ_v), and minimum principal stress, (σ_3) are shown. Field structures in (a) and (c) are cross-sectional views and (b) is plain view



- Strike-slip tectonics that results in strike-slip faults (Fig. 3b) in zones of lateral displacement within the lithosphere.
- Compressional tectonics that results in the formation of thrust faults (Fig. 3c) and associated tectonic activities due to the shortening and thickening of the lithosphere.

The difference between structural geology and tectonics is the scale of observation. Structural geology deals with small-scale rock deformation, while tectonics is more concerned with larger features. Thus, remote sensing methods help in detecting large-scale tectonic features while field studies are necessary in investigating how the mapped lineaments control the tectonics of a region by looking into the details at mesoscopic-scale. All remote sensing methods used for lineament extraction cannot describe the nature of the tectonics in a region without ground truthing. It is generally recognized that lineament or fracture trace analysis used in conjunction with tectonic investigations is not credible without field verification. In regional scale studies, although field checking of each lineament is impossible due to significant cost and time involved (Mabee et al. 1994), systematic field checking is still necessary to decipher and derive confidence estimates of the nature of lineaments. The field data include observation and measurement of orientations of joints, dykes, and faults. Field evidences presented in the form of figures and sketches (e.g., Fig. 3a–c) can provide good support for tectonic

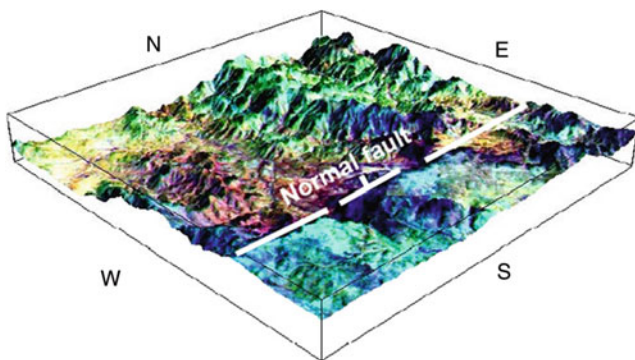
interpretation. The structural data collected from the field need to be grouped according to rock type and proximity of sampling points to take into account the geographic locations of observation points. In order to understand the structural–geological significance of fracture arrays, it is a common practice to subdivide them into separate sets on the basis of orientation. Rose diagrams are commonly used to reveal the orientations of vertical and steeply dipping joints, faults, and dykes observed at different scales. Many workers (e.g., Arlegui and Soriano 1998; Solomon and Ghebreab 2006) have effectively used remote sensing data supported by field studies to map regional spatial distributions of lineaments toward understanding their tectonic origin.

Geographic Information System (GIS)

GIS is a tool in spatial data management and analysis that can assist users in organizing, storing, editing, analyzing, and displaying positional and attribute information about geographical data. GIS facilitates integration of all data types and allows a better understanding of the tectonic interpretation of a region under study. The full potential of remote sensing and GIS can be utilized when an integrated approach is adopted. Integration of the two technologies has proven to be an efficient tool in many studies, specifically in tectonic studies (e.g., Ratzinger et al. 2006; Gloaguen et al. 2008).

For instance, different geomorphic indices derived from DEM and a number of factors influencing each index can be integrated in a GIS that will allow a better interpretation of the results. Backed with remote sensing data and ground-checking, it is possible to outline the deformed area and the local strain intensity (relative uplift rates) and thus infer the neotectonic history of a region. Jain and Verma (2006) applied a combination of remote sensing and GIS for mapping active tectonic intensity zone in India. Superimposing a lineament map over a lithological map in a GIS environment provides an insight into the nature of tectonics of a region. Visible horizontal displacement along lineaments (e.g., Fig. 3b) as evidenced from movement in rock units can be used to depict strike-slip fault tectonism.

Draping the lithological map or digitally enhanced color image over a three-dimensional DEM data in a GIS environment can also reveal the faulting mechanisms such as normal faulting due to visible subvertical displacement of lithological units (Fig. 4). In order to classify the fault mechanics involved in deformation, an overlay of historical earthquake data on mapped tectonic features or lineaments can be integrated in a model derived from remote sensing data with results such as from differential radar interferometry in a GIS environment. A GPS-derived parameters can be easily integrated in a GIS framework to provide a powerful means of directly measuring the dynamics of present-day crustal deformation by monitoring the changes in geographical (horizontal) locations and heights (vertical displacements) at well-positioned control points. All these examples demonstrate the application of remote sensing and GIS techniques for mapping both active and passive faults. However, to better understand the types of faults involved, field checking still remains a requirement.



Remote Sensing and GIS Techniques for Tectonic Studies, Fig. 4 Three-dimensional synoptic view of a digitally processed remote sensing data created by draping from ASTER bands (13, 5, and 3) in red, green, and blue order over the SRTM DEM data in a GIS environment. Note the east-west normal fault that dips to the north

Summary

Remote sensing and GIS are complementary to each other and efficient techniques for lineament mapping, and hence for tectonic studies. Advanced remote sensing techniques such as InSAR and GPS have revolutionized tectonic studies by directly measuring the motion of faults associated with earthquakes and rates of uplifts. Field studies are, however, needed to identify the types of lineaments and correlate them to the remotely acquired data. Comparison of orientations of major structures or lineaments such as joints, dykes, and faults obtained from field measurements, and plotted on rose diagrams facilitates their tectonic interpretation. GIS is a very time-saving and cost-effective tool once the database is established. Integrating data of different layers such as lineament and lithological maps, and geomorphic indices acquired, for example, from DEM, GPS measurements and field studies in a GIS environment followed by spatial analysis of the data allows correlation between different parameters for unraveling the nature of the tectonics of a region. Moreover, integrating historical dataset from past earthquake events, if any, including an understanding of the regional tectonic setting is essential in facilitating interpretation of results. Overall, combining remote sensing, GIS, and field techniques provides a powerful tool for understanding the tectonic framework of a region, and assessing the resulting fault mechanisms.

Cross-References

- [Remote Sensing, Applications to Geophysics](#)
- [SAR Interferometry](#)

Bibliography

- Arlegui LE, Soriano MA (1998) Characterizing lineaments from satellite images and field studies in the central Ebro basin (NE Spain). *Int J Remote Sens* 19:3169–3185
- Demirkesen AC (2008) Digital terrain analysis using Landsat-7 ETM+ imagery and SRTM DEM: a case study of Nevsehir province (Cappadocia), Turkey. *Int J Remote Sens* 29:4173–4188
- Gloaguen R, Käbner A, Wobbe F, Shazah F, Mahmood A (2008) Remote sensing analysis of crustal deformation using river networks. In: *IEEE international geoscience & remote sensing symposium*, paper: TH1.101.1, Boston, 6–11 July
- Hirano A, Welcha R, Langb H (2003) Mapping from ASTER stereo image data: DEM validation and accuracy assessment. *ISPRS J Photogramm Remote Sens* 57:356–370
- Jain S, Verma PK (2006) Mapping of active tectonics intensity zones using remote sensing and GIS. *J Indian Soc Remote Sens* 34:131–142
- Jouanne F, Genaudeau N, Ménard G, Darmendrail X (1998) Estimating present-day displacement fields and tectonic deformation in active mountain belts: an example from the Chartreuse Massif and

- the southern Jura Mountains, Western Alps. *Tectonophysics* 296:403–419
- Karnieli A, Meiseis A, Fisher L, Arkin Y (1996) Automatic extraction and evaluation of geological linear features from digital remote sensing data using the Hough transform. *Photogramm Eng Remote Sens* 62:525–531
- Mabee SB, Hardcastle KC, Wise DW (1994) A method of collecting and analyzing lineaments for regional-scale fractured bedrock aquifer studies. *Ground Water* 32:884–894
- Ratzinger K, Neuhäuser B, Papthoma M (2006) Hazard mapping of earthquake triggered landslides. In: First European conference on earthquake engineering and seismology, Geneva, p 1405, 3–8 Sept
- Saraf AK, Choudhury S (2005) Thermal remote sensing technique in the study of pre-earthquake thermal anomalies. *J Indian Geophys Union* 9:197–207
- Segall P, Davis JL (1997) GPS applications for geodynamics and earthquake studies. *Annu Rev Earth Planet Sci* 25:301–336
- Solomon S, Ghebreab W (2006) Lineament characterization and their tectonic significance using Landsat TM data and field studies in the central highlands of Eritrea. *J Afr Earth Sci* 46:371–378
- Stramondo S, Moro M, Tolomei C, Cinti FR, Doumaz F (2005) InSAR surface displacement field and fault modelling for the 2003 Bam earthquake (southeastern Iran). *J Geodyn* 40:347–353
- Suzen ML, Toprak V (1998) Filtering of satellite images in geological lineament analysis: application to a fault zone in Central Turkey. *Int J Remote Sens* 19:1101–1114
- Zizioli D (2008) DEM-based morphotectonics analysis of Western Ligurian Alps. *Scientifica Acta* 2:44–47

Remote Sensing, Applications to Geophysics

Hojjatollah Ranjbar

Department of Mining Engineering, Shahid Bahonar
University of Kerman, Kerman, Iran

Definition

Remote sensing is defined as the acquisition of data using a remotely located sensing device, that relies on the measurements of electromagnetic energy (EM) in the range of 0.4 μm to 1 m, and extraction of information from the data (Gupta 2003; McCloy 2006). Remote sensing can be multilevel (i.e., sensing the surface of material from a distance of few centimeters to millions of kilometers). Taking a spectrum of a leaf by using a spectroradiometer to taking an image by Hubble telescope from a star at a distance of few light years fall in the realm of remote sensing. However, here, we mean the imaging of the Earth's surface.

Introduction

The remote sensing systems can acquire images from the surface of the Earth by either passive or active methods.

Passive systems use the electromagnetic energy that is produced by the sun or any other process like anomalous heat buildup within the Earth. Active remote sensing systems create their own electromagnetic energy that (1) is transmitted from the sensor toward the terrain, (2) interacts with the terrain producing a backscatter of energy, and (3) is recorded by the remote sensor's receiver (Jensen 2000). The sending and receiving of signals are largely unaffected by atmospheric conditions. Therefore, radar satellites can record the details of the Earth's surface in cloudy weather condition and any time of the day and night. LIDAR (LIght DEtection and Ranging) and SONAR (SOund NAVigation Ranging) are also the active remote sensing systems. It should be noted that remote sensing is a geophysical method as it uses the electromagnetic spectrum as a medium for surveying the Earth's surface. It is now recognized as a separate branch of science and engineering.

The electromagnetic spectrum that is used by remote sensing systems is named according to the wavelength ranges. Gupta (2003) has classified the electromagnetic spectrum into visible (0.4–0.7 μm), near infrared (0.7–1.0 μm), Shortwave infrared (1.0–3.0 μm), mid-infrared (3.0–35 μm), far-infrared (35–1.0 mm), and RADAR (1–100 cm).

A remote sensing system that allows sensing the Earth in any wavelength and in any time makes an ideal remote sensing system. An ideal remote sensing system cannot be achieved due to the absorptions of the electromagnetic energy from the Earth's atmosphere.

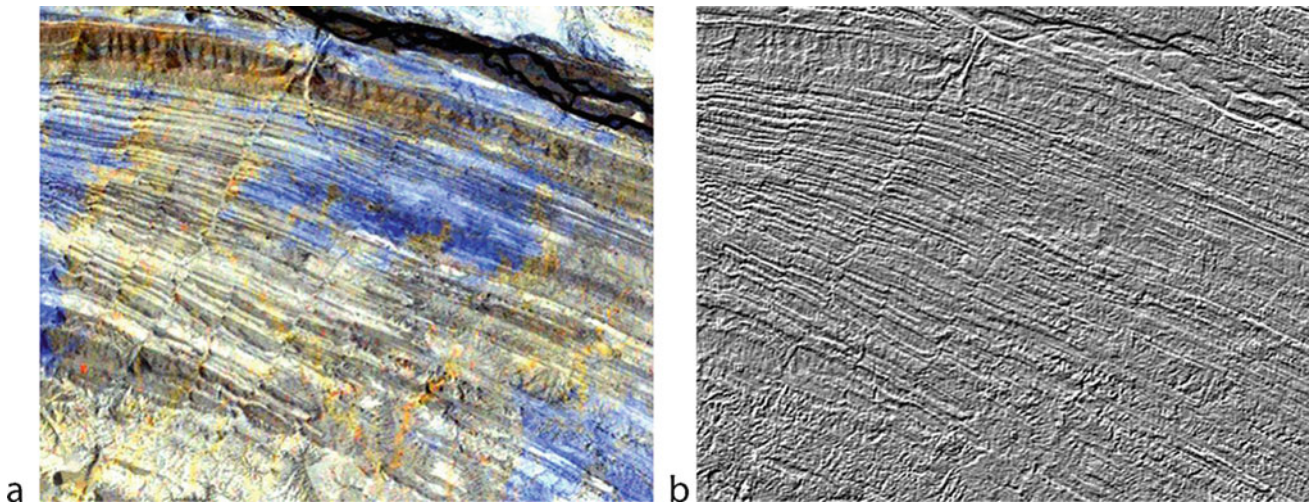
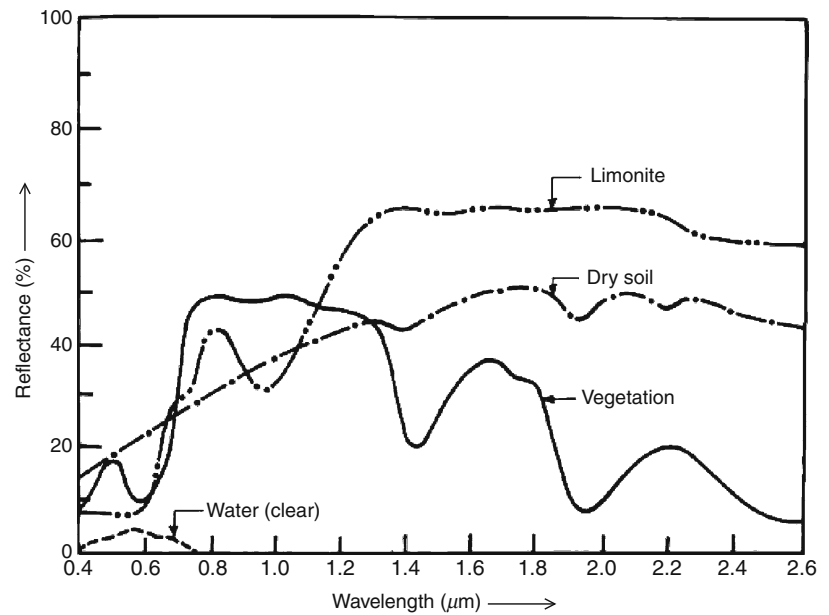
Each object has unique sets of absorption and reflectance features concerning the wavelengths. The plot is known as spectral signature. Figure 1 shows the spectra of the common surface materials. Although each material at the surface of the Earth has many absorption and reflection features, the present remote sensing systems can only image these materials in limited wavelengths because the atmosphere is not clear for all the wavelengths.

There are different sensors that are used onboard the satellites or in an aeroplane. The present remote sensing systems that orbit the Earth are imaging the Earth's surface in visible, near infrared, shortwave infrared, thermal infrared, and RADAR regions of EM spectrum. The reader can refer to Jensen (2000) and Lillesand et al. (2004) for different sensors and platforms specifications.

A digital image consists of pixels. Each pixel is representative of a surface area that defines the spatial resolution of the system. Enhanced Thematic Mapper Plus (ETM+) multispectral data has a resolution of 30 m (each pixel is equivalent to 900 m² ground area), while Quickbird has a resolution of 60 cm for its panchromatic image (each pixel is equivalent to 0.36 m² on the ground).

As these images are in digital formats, several simple or complex mathematical operations can be applied on them in order to enhance specific features from these images. This is

Remote Sensing, Applications to Geophysics, Fig. 1 Typical spectral reflectance curves for selected common natural objects (Gupta 2003)



Remote Sensing, Applications to Geophysics, Fig. 2 Enhancement of faults and other linear features by using directional filter. (a) Original satellite color image, (b) filtered image

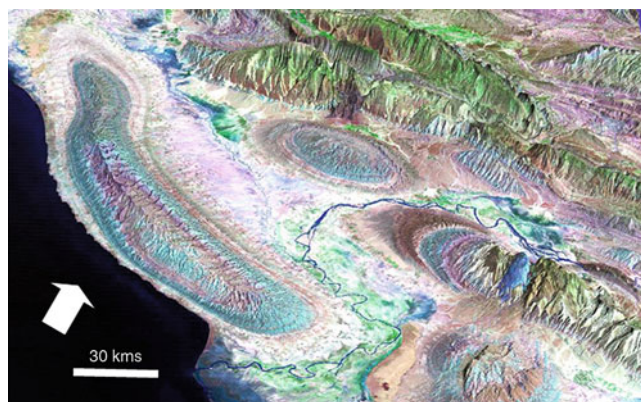
called image processing operation. Image filtering (Mather 2001) is a common image processing technique that enhances the linear features. Figure 2 shows the enhanced faults in the southern part of Iran in a filtered image.

Application of Remote Sensing to Geophysics

Remote sensing data can be used for designing the survey grid and line in geophysical surveys, verifying, and helping a better interpretation of the geophysical methods. Remote sensing images can be combined to form color images in various scales that depend on the spatial resolution of the

remote sensing system. These color images can be draped over a digital elevation model (DEM) to give a three-dimensional (3-D) model of the area (Fig. 3). These 3-D models can be used in both land and airborne geophysical surveys for designing the survey lines. The 3-D model can provide information such as the terrain ruggedness, man-made installations, surface conditions, transportation routes, etc.

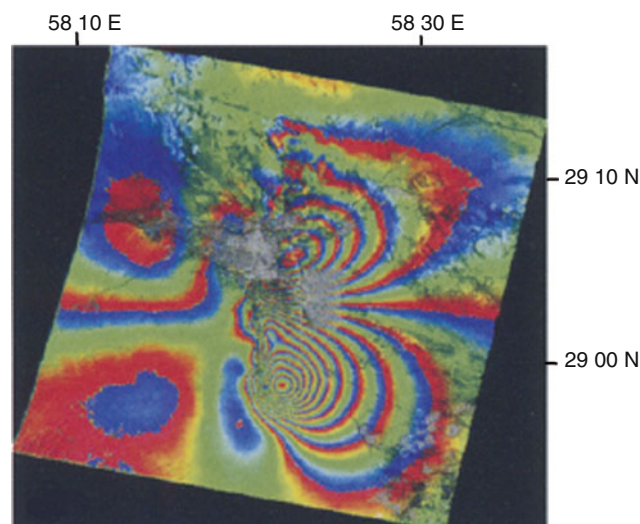
The surface of the Earth is continuously deformed due to the tectonic activities over geological time that causes the formation of structures such as fold and fault. Geophysical methods such as magnetic, magneto-telluric, gravity, and seismic methods are used for studying these structures.



Remote Sensing, Applications to Geophysics, Fig. 3 3-D view of folding due to the movement of Arabian plate against Iranian microplate. The force direction is shown by an *arrow*. Landsat color image (Band 7 in red, band 4 in green, and band 2 in blue) is draped over DEM of the area

These methods are subsurface methods. Remote sensing techniques are useful in order to study these structures and to interpret the tectonic activities in the regional and local scale at the surface of the Earth. In order to obtain a clear picture of the structures, the combined interpretation of remote sensing and geophysical data is helpful, if the geological structures have any trace in the satellite images. Remote sensing applications to tectonic and structural studies are reported by many researchers (e.g., Harris 1991; Philip 1996; Saintot et al. 1999; Chernicoff et al. 2002; Chatterjee 2003; Raharimahefa and Kusky 2006; Walker 2006). Figure 3 shows a part of Zagros simply folded belt that the folds axis directions indicate a force direction because of the subduction of Arabian plate beneath the Iranian plate. These structures are also studied by using gravitational, magnetic, and seismic methods.

Remote sensing is also useful for interpretation of geophysical data. Alteration process causes the changes in mineralogical content of rocks due to the reaction of hot solution with the rocks. Porphyry copper mineralization is associated with these alteration zones (Evans 1993). Both radiometric and remote sensing methods can be used for alteration detection (Ranjbar et al. 2001, 2010; Ranjbar and Honarmand 2004). The minerals present in sericite and k-feldspar have a higher gamma radiation. We cannot differentiate these minerals by geophysical methods. Remote sensing method is capable of recognizing the minerals responsible for the higher radiation in geophysical data. There are also minerals in the alteration zones that have very low radiation to be detected by geophysical methods. Kaolinite, Calcite, epidote, and chlorite are such minerals. These minerals are recognized by remote sensing methods. Remote sensing techniques can also be used in combination with other geophysical methods such as magnetic, electromagnetic, and electrical methods in helping with the interpretation.



Remote Sensing, Applications to Geophysics, Fig. 4 An interferogram that shows the deformation along a fault due to the Bam earthquake in Iran. The rings that are closer, indicating more deformation. (After Ye 2005)

Another area that received much attention is the field of active faults study by the combined use of interferometry and geophysical data. RADAR (Radio Detection And Ranging) images are more useful for detection of the geological structures than the images acquired in the optical ranges of EM spectrum, because the RADAR pulses are illuminating the ground from a specific direction. The RADAR wavelengths are in order of few tens of centimeters. Synthetic Aperture Radar interferometry (INSAR) is a branch of remote sensing that has emerged in the recent years that deals with the deformation in the Earth's crust. There are several papers published about INSAR method in structural and tectonic fields (e.g., Klees and Massonnet 1999; Wright et al. 2001, 2004; Ye 2005; Bürgmann et al. 2006; Taylor and Peltzer 2006). Figure 4 shows an interferogram of a fault in Bam area that triggered an earthquake in 2003. This fault is detected by remote sensing and geophysical methods.

Summary

Remote sensing images due to their multi-temporal, multi-spectral, and synoptic views are extensively used for solving problems related to mineral exploration, tectonics, vegetation cover studies, surface changes, earthquake studies, weather and volcanic eruption forecasting, etc. The images can be acquired in different wavelength ranges such as visible, near infrared, thermal, and Radar ranges of electromagnetic spectrum. We are now able to recognize many minerals from the space by using their spectral properties. Remote sensing data can be used for designing the survey grid and line in

geophysical surveys, verifying, and helping a better interpretation of the geophysical methods. Small and large structures are easily recognizable in the images, although we often need to enhance the images for recognizing these structures.

Cross-References

- Earthquake Precursors and Prediction
- Earthquake Rupture: The Inverse Problem
- Earthquakes and Crustal Deformation
- Gravity Method, Airborne
- Magnetic Anomalies: Interpretation
- Magnetic Methods, Airborne
- Plate Tectonics, Precambrian
- Remote Sensing and GIS Techniques for Tectonic Studies
- SAR Interferometry
- Seismic Data Acquisition and Processing
- Subduction Zones
- Very Long Baseline Interferometry

Bibliography

- Bürgmann R, Hilley G, Ferretti A, Novali F (2006) Resolving vertical tectonics in the San Francisco Bay Area from permanent scatterer InSAR and GPS analysis. *Geology* 34:221–224
- Chatterjee RS (2003) Structural pattern of Holenarsipur Supracrustal Belt, Karnataka, India as observed from digitally enhanced high-resolution multi-sensor optical remote sensing data aided by field survey. *Int J Appl Earth Obs Geoinf* 4:195–215
- Chernicoff CJ, Richards JP, Zappettini EO (2002) Crustal lineament control on magmatism and mineralization in northwestern Argentina: geological, geophysical, and remote sensing evidence. *Ore Geol Rev* 21:127–155
- Evans AM (1993) *Ore geology and industrial minerals: an introduction*. Blackwell, Oxford
- Gupta RP (2003) *Remote sensing geology*. Springer, Berlin
- Harris JR (1991) Mapping of regional structure of eastern Nova Scotia using remotely sensed imagery: implications for regional tectonics and gold exploration. *Can J Remote Sens* 17:122–136
- Jensen JR (2000) *Remote sensing of the environment: an earth resource perspective*. Prentice Hall, New Jersey
- Klees R, Massonnet D (1999) Deformation measurements using SAR interferometry: potential and limitations. *Geol Mijnb* 77:161–176
- Lillesand T, Keifer RW, Chipman JW (2004) *Remote sensing and image interpretation*. Wiley, New York
- Mather PM (2001) *Computer processing of remotely-sensed images. An introduction*. Wiley, Chichester
- McCloy KR (2006) *Resource management information systems: remote sensing, GIS and modelling*, 2nd edn. Taylor and Francis, Boca Raton
- Philip G (1996) Landsat thematic mapper data analysis for quaternary tectonics in parts of the Doon valley, NW Himalaya, India. *Int J Remote Sens* 17:143–153
- Raharimahefa T, Kusky TM (2006) Structural and remote sensing studies of the southern Betsimisaraka Suture, Madagascar. *Gondwana Res* 10:186–197
- Ranjbar H, Honarmand M (2004) Integration and analysis of airborne geophysical and ETM + data for exploration of porphyry type deposits in the central Iranian Volcanic Belt, using fuzzy classification. *Int J Remote Sens* 25:4729–4741
- Ranjbar H, Hassanzadeh H, Torabi M, Ilaghi O (2001) Integration and analysis of airborne geophysical data of the Darrehzar area, Kerman province, Iran, using principal component analysis. *J Appl Geophys* 48:33–41
- Ranjbar H, Masoumi F, Carranza EJM (2010) Evaluation of geophysics and spaceborne multispectral data for alteration mapping in Sar Cheshmeh mining area, Iran. *Int J Remote Sens*. <https://doi.org/10.1080/01431161003745665>
- Saintot A, Angelier J, Jean Chorowicz J (1999) Mechanical significance of structural patterns identified by remote sensing studies: a multi-scale analysis of tectonic structures in Crimea. *Tectonophysics* 313:187–218
- Taylor M, Peltzer G (2006) Current slip rates on conjugate strike-slip faults in central Tibet using synthetic aperture radar interferometry. *J Geophys Res* 111:1–16
- Walker RT (2006) A remote sensing study of active folding and faulting in southern Kerman province, S.E. Iran. *J Struct Geol* 28:654–668
- Wright T, Fielding E, Parsons B (2001a) Triggered slip: observations of the 17 August 1999 Izmit (Turkey) earthquake using radar interferometry. *Geophys Res Lett* 28:1079–1082
- Wright T, Parsons B, Fielding E (2001b) Measurement of interseismic strain accumulation across the North Anatolian Fault by satellite radar interferometry. *Geophys Res Lett* 28:2117–2120
- Wright TJ, Lu Z, Wicks C (2004) Constraining the slip distribution and fault geometry of the Mw 7.9, 3 November 2002, Denali fault earthquake with interferometric synthetic aperture radar and global positioning system data. *Bull Seismol Soc Am* 94:S175–S189
- Ye X (2005) Bam earthquake: surface deformation measurement using radar interferometry. *Acta Seismol Sin* 18:451–459

SAR Interferometry

Masato Furuya
Department of Natural History Sciences, Hokkaido
University, Sapporo, Japan

Synonyms

Differential InSAR (abbreviated as D-InSAR); Interferometric SAR (abbreviated as InSAR); Radar interferometry; SAR interferometry

Definition

- Radar** Acronym standing for Radio Detection and Ranging. A technique to detect any targets and measure the distance to them, based on the round-trip time of microwave (radio wave) pulses between the antenna and the targets. *SAR*. Acronym standing for Synthetic Aperture Radar. A technique to image any ground surfaces, using airborne or spaceborne radar sensor. Its high spatial resolution is achieved by collecting numerous return pulses from each target in sight and by effectively synthesizing large antenna size.
- InSAR** Acronym standing for Interferometric SAR. A technique to image surface topography and ground displacements, using phase values of two or more SAR images.

Introduction

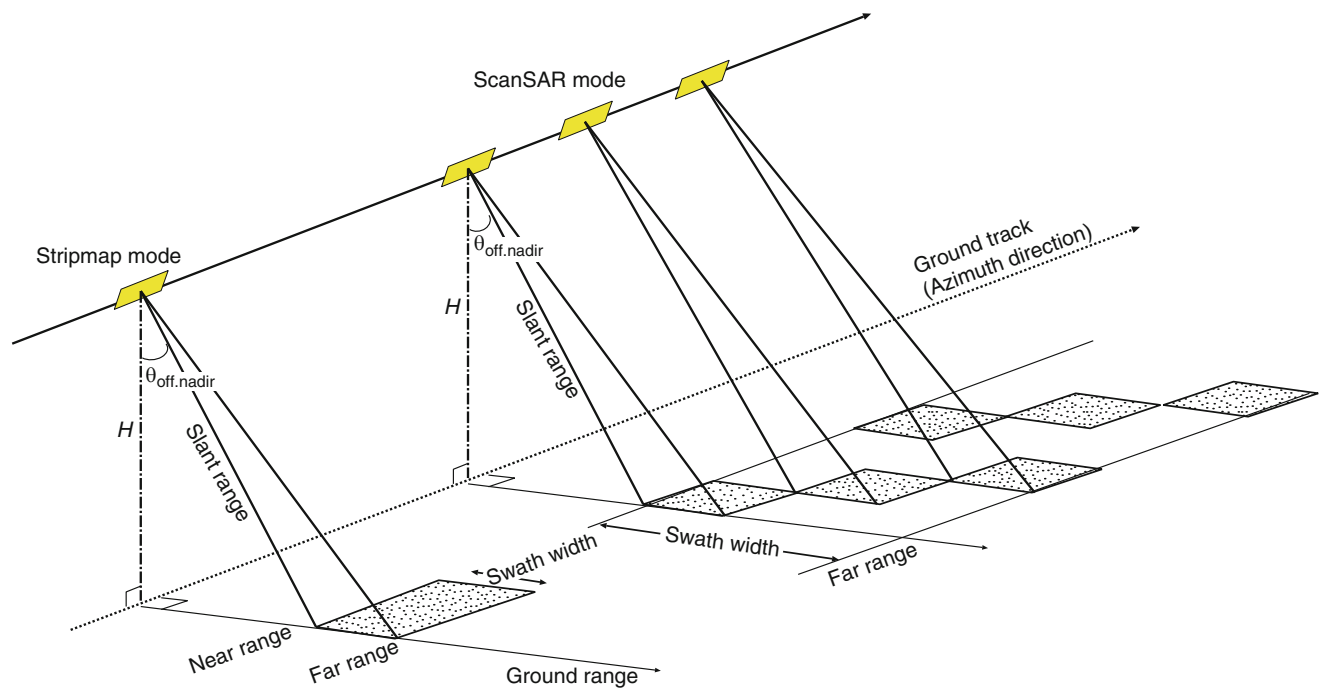
Crustal deformation data have been traditionally acquired by ground-based geodetic techniques such as leveling, triangulation, and electro-optic distance measurement. More

recently, global positioning system (GPS) has become a standard tool for high-precision crustal deformation measurement, and provided us with a wealth of data to study plate tectonics, earthquakes, volcanic activities, and atmospheric and hydrological loading deformation. All these techniques, however, require in situ benchmarks, and thus prevent us from observing inaccessible areas. Interferometric SAR (InSAR) was, therefore, regarded as a surprising and revolutionary technique when Massonnet et al. (1993) first showed an image of the co-seismic deformation associated with the 1992 M7.3 Landers earthquake, because the raw data was completely acquired on a spaceborne sensor. Another big surprise for the community was its incredibly high spatial resolution, which no other geodetic techniques were possible to achieve in practice.

Nowadays, InSAR users have proliferated in a worldwide community and applied to a variety of geophysical problems. A number of excellent and extensive reviews for advanced students and researchers are already published (Bamler and Hartl 1998; Massonnet and Feigl 1998; Burgmann et al. 2000; Hanssen 2001; Pritchard 2006; Simons and Rosen 2007; Zhou et al. 2009). I therefore tried to make this article much shorter and more introductory, but it still includes necessary and useful concepts, ranging from the fundamentals of SAR/InSAR imagery to more up-to-date topics.

Fundamentals of SAR Imaging and SAR Data

SAR satellite flies over at an altitude of hundreds of km, repeating transmission and reception of microwave pulses. The along-track and across-track axes are almost identical to the *azimuth* and *range* axis in the acquired radar image. The area illuminated on the ground is called *swath*, whose width spans roughly 50–100 km in the standard *stripmap* (or *strip*) mode with an incidence angle of 20–50° (Fig. 1). While previous SAR applications are mostly derived from the stripmap mode, another imaging mode, *ScanSAR*, is also promising because it covers much wider swath width,



SAR Interferometry, Fig. 1 Geometry of SAR imaging. SAR sensor transmits microwave pulses in slant range direction, and receives their reflected pulses. While stripmap mode achieves high spatial resolution

with a fixed off-nadir angle, ScanSAR mode achieves wider imaged area (swath) with multiple off-nadir angles at the expense of the resolution

300–500 km, by illuminating multiple swaths at the expense of reducing the resolution. ScanSAR is useful for imaging long-wavelength signals associated with, for instance, a magnitude-8-class earthquake (Motagh et al. 2008).

Although it was not strictly necessary, satellite-based SAR system has been often placed on a sun-synchronous and near-polar orbit with an inclination angle of slightly greater than 90° . When the satellite moves to the north (south), we call it is in *ascending* (*descending*) orbit.

The raw data acquired on SAR sensor are impossible to visually interpret, and require a bit involved processing algorithms; those algorithms are detailed in a couple of text books (e.g., Curlander and McDonough 1991; Cumming and Wong 2005). The first interpretable SAR image is a single-look-complex (SLC) image, whose important difference from other optical images is that each pixel consists of a complex (real and imaginary) value, i.e., amplitude and phase. This is because the waveform of each repeated pulse is precisely controlled to be identical, and hence the received pulse provides us with not only a scattering (reflection) *intensity* but also a *phase*. The phase data do contain the geometric information from the antenna to the ground targets, and are fully exploited in generating InSAR image as discussed later. However, the phase image itself is usually not as useful as the intensity image because it is again impossible to visually interpret the physical meaning. Meanwhile, the intensity image is often useful and derived from a square-root magnitude of SLC data with spatial averaging called *multi-looking*. By single-look, it means the finest spatial resolution for both

range and azimuth axis. In the standard stripmap mode, the range and azimuth resolutions are derived as,

$$\Delta r = \frac{c}{2B}, \text{ and } \Delta a = \frac{L}{2}, \quad (1)$$

respectively; the c , B , and L are the speed of light, the frequency bandwidth of the microwave pulse, and the antenna length along azimuth axis, respectively (Curlander and McDonough 1991; Cumming and Wong 2005). The waveform of each microwave pulse is called *chirp signal*, whose instantaneous frequency linearly changes by as much as the frequency bandwidth B over the duration of each pulse. It should be noted that the spatial resolution depends neither on the sensor altitude nor the carrier frequency of microwave. Intensity images are often shown in gray scale images, in which strongly (weakly) reflected objects/areas are usually colored as bright (dark). Although they simply look like black-and-white photographs, we should keep in mind that they could be acquired regardless of weather and time because SAR is actively transmitting and receiving microwaves. Also, intensity images are indispensable for high-precision image matching prior to a generation of InSAR image.

Fundamental Principles of InSAR

Interferometric SAR (InSAR) is a technique to generate a digital elevation model (DEM) or a ground displacement

SAR Interferometry, Fig. 2 (a)

Geometry of the Young's experiment. Depending on the path difference, the two coherent waves from the slit, S1 and S2, are in-phase or out-of-phase on the screen, and interference fringes are observed on the right screen. (b) Orbital fringe (flat earth fringe) can be regarded as a 3-D analogue of the Young's experiment

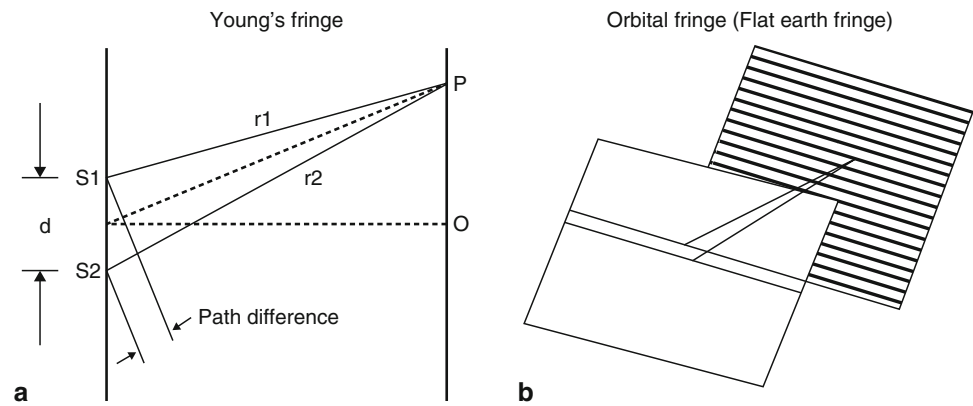
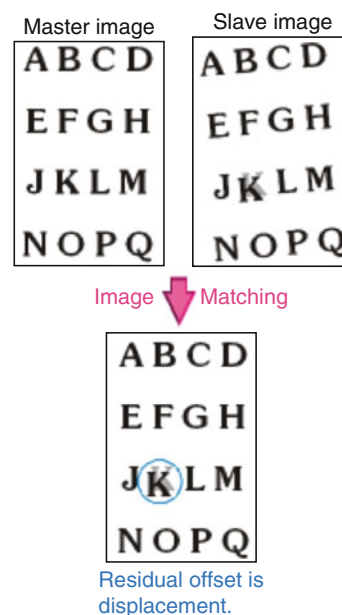


image from a *pair* of SLC images. The term *interferogram* is often used to represent InSAR image. We can understand the principle of InSAR, recalling the classical Young's experiment that is known to be a proof of the wave characteristics of the light (Ghilia and Pritt 1998). Two coherent waves out of the slits will generate "stripes" on the wall, called *interference fringe* (Fig. 2a). We can simulate the fringe if we know the separation of the slits, the distance from each slit to the wall, and the wavelength of the coherent wave. Depending on the path difference, the two coherent waves are in-phase or out-of-phase when they reach the screen. Namely, the difference of the phases generates the interference fringe. We may regard the imaging geometry of InSAR as the 3-D Young's experiment (Fig. 2b). The repeat orbit tracks, the ground surface, and the microwave correspond to the double slits, the screen, and the coherent wave, respectively. Once we get two SLC images, we can generate an initial interferogram, multiplying one SLC image with the complex conjugate of the other SLC image. We then observe similar fringes in the initial interferogram as illustrated in Fig. 2b, which is literally a map of the difference of two SLC phases. For descriptive purposes, the former SLC image is often denoted as *master*, and the latter SLC image is called *slave*. At this moment, the slave image must be precisely co-registered (or matched) to the master image (Fig. 3); we will come back to this *image co-registration* (or *image matching*) procedure later on.

While Fig. 2b shows an initial interferogram over flat areas with parallel orbits, the fringe will appear undulated if the areas are not flat. The fringe over flat areas is called *flat Earth fringe* (or, *orbital fringe*), and can be precisely simulated from the pair of orbit data. If we subtract the flat Earth fringes from the initial interferogram, we can extract *topographic fringe* that can be used to generate DEM. The Shuttle Radar Topography Mission (SRTM) was carried out along this idea in 2001, and generated 3-s resolution DEM over $\pm 60^\circ$ latitudes (Farr et al. 2007). In the case of SRTM, they carried two SAR antennas on the same platform, and thus were able to generate DEM without repeating the previous orbit track. In contrast, all the present SAR satellite systems carry only one



SAR Interferometry, Fig. 3 Image registration (matching) of the master and slave images prior to interferogram generation, and the principle of pixel-offset technique to derive large displacements. While long-wavelength distortion can be corrected, localized huge displacement remains as residual offset. Courtesy of Tobita et al. (2001a)

antenna with a repeat-pass period of several weeks, which are 11 days for TerraSAR-X, 16 days for COSMO-SkyMed, 24 days for Radarsat-1/2, 35 days for Envisat, and 46 days for ALOS. Therefore, if ground surface undergoes significant deformation during the repeat orbit cycles due, for instance, to earthquake and volcanic eruption, the interferogram will include *deformation fringe* as well. To extract deformation fringe, we must take out both orbital fringe and topographic fringe, which can be simulated from satellite orbit data and DEM. The deformation fringes represent slant range changes along the radar line-of-sight (LOS), and thus projections of the 3-D displacement vector on the ground along the unitary vectors toward the radar LOS (Fig. 4). The range changes should be interpreted as relative displacements to the

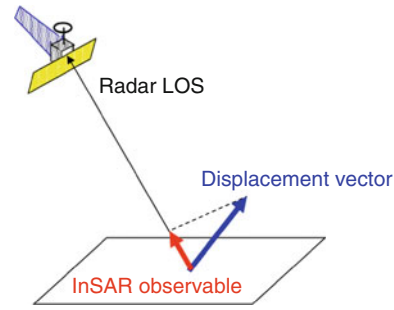
reference point(s) inside each interferogram. Depending on literatures, they denote differential interferometric SAR (D-InSAR) when the technique is used to detect deformation signals. Recently, however, the term InSAR is often and simply used to represent D-InSAR.

Even if no significant ground displacements take place during the repeat-pass period, however, we usually encounter other non-negligible fringes due to the spatial heterogeneities in the propagation delay of microwaves through the atmosphere, the errors in satellite orbit data, and those in DEM. Because these fringes limit the precision and accuracy of SAR-based crustal deformation measurement, a couple of correction approaches have been proposed. More advanced time-series analysis techniques have also been developed to overcome the issues, which will be introduced in the last section.

InSAR Processing

Image registration (Matching): Before we get an initial interferogram, we must register (or, match) each imaged target in one SLC image to the same target in the other SLC image with a sub-pixel level accuracy, because any ground objects do not usually locate at the same pixel coordinates in each SLC image. This pre-processing is called *image registration* (or *image matching*) and prerequisite to be performed prior to generating an initial interferogram. Although a simple polynomial transformation between the range and azimuth coordinates of two SLC images is sufficient in most cases, we need to take into account the effects of 3-D topography when the terrain surface is rugged to eliminate a stereoscopic effect (Michel et al. 1999).

When large ground displacements on the order of meters or more take place locally, and if we correct for the long-wavelength image distortion using the polynomial transformation, we can detect and quantify those localized displacements as a by-product of image registration without viewing InSAR image (Fig. 3; Tobita et al. 2001a). This approach to detect large displacements is called *pixel offset* or *feature tracking* technique, and has been applied to earthquakes, volcanic eruptions, and glacier movements. The advantages of pixel-offset data are twofolds. First, pixel-offset data can quantify large displacements even in such areas that completely loses interferometric coherence, where InSAR data cannot be unwrapped; we describe *coherence* and *unwrapping* later below. Secondly, in contrast to InSAR data, pixel-offset data provide us with not only range offset but also azimuth offset component. While the range offset has the same sensitivity to the 3-D displacement vector as InSAR data (Fig. 4), the azimuth offset is a projection of the displacement vector onto the unitary vector perpendicular to the LOS. Hence, the azimuth offset data are complementary to the range offset or InSAR data. Taking advantage of this property,



SAR Interferometry, Fig. 4 InSAR observable is a projection of the displacement vector along the radar line-of-sight (LOS) direction

Fialko et al. (2001) derived a full 3-D displacement map for the 1999 M7.1 Hector Mine earthquake, combining the InSAR data from both ascending and descending track with the azimuth offset data. Using pixel-offset data from both descending and ascending track, Tobita et al. (2001a, b) inferred a 3-D displacement map associated with the 2,000 eruption episode at Usu volcano.

Interferometric phase and its relation to geometry: Suppose we have two co-registered SLC images, E_1 and E_2 , acquired from different ranges r_1 and r_2 :

$$E_1 = e^{j\phi_{\text{Scatter}}} e^{-\frac{4\pi r_1}{\lambda}} \quad (2a)$$

$$E_2 = e^{j\phi_{\text{Scatter}}} e^{-\frac{4\pi r_2}{\lambda}} \quad (2b)$$

Here we assume that the reflection magnitude and scattering phase are constant during the data acquisition time. Then, the interferometric phase ϕ is derived as

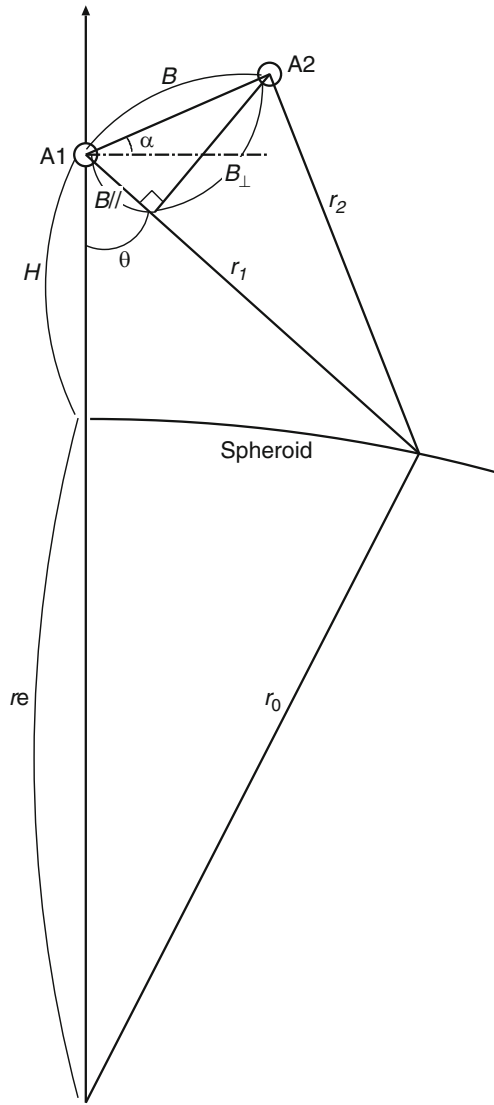
$$E_1 E_2^* = e^{-\frac{4\pi(r_1 - r_2)}{\lambda}} \quad (3)$$

or

$$\phi = \frac{4\pi}{\lambda} (r_1 - r_2) \quad (4)$$

The last one is the fundamental equation for InSAR, which describes “unwrapped” phase in the initial interferogram. The actual phase in the initial interferogram is “wrapped” into an interval $[-\pi, \pi]$, and thus has ambiguities of $2\pi N$; N is integer. In order to quantify the ground displacement along radar LOS, we have to perform 2-D *phase unwrapping* on the interferogram, which is not necessarily straightforward (Bamler and Hartl 1998; Ghilia and Pritt 1998). While the interferometric phase is strictly a phase “difference” of two SLC phases, it is conventional to simply call phase. The factor 4 is to take into account the round-trip distances.

Figure 5 is a cross section that is perpendicular to the satellite repeat tracks and passes through the Earth’s center,



SAR Interferometry, Fig. 5 Geometry of InSAR data acquisition and its relation to the baseline. The A1 and A2 are the satellite's repeat orbits, and the spatial distance between the A1 and A2 is the baseline B . The initial InSAR phase is proportional to the difference between the ranges, r_1 , and r_2 , and hence the B_{para} (Eq. 5). The fringe rate (density) along the range axis is proportional to the B_{perp} (Eq. 6)

and shows a geometry of InSAR data acquisition. The spatial separation of the repeating orbits is termed baseline (or spatial baseline), B ; the temporal separation of data acquisition is sometimes called temporal baseline. Because the baseline B is usually much shorter than the ground range distance R , a parallel ray approximation holds (Zebker et al. 1994) and the fundamental Equation 4 can be approximated as follows:

$$\phi = \frac{4\pi}{\lambda}(r_1 - r_2) \approx -\frac{4\pi}{\lambda}B_{//} = B \sin(\theta - \alpha), \quad (5)$$

where ϕ and α are defined in Fig. 5, and $B_{//}$ (or B_{para}) is a baseline component parallel to the radar LOS. The angle θ is

called *off-nadir angle*, and is identical to incidence angle if the Earth's curvature is negligible. The other baseline component $B_{\perp}$ (or B_{perp}) is perpendicular to radar LOS and gives us an important criterion for successful InSAR processing as we discuss below.

Decorrelation (Coherence): In the actual InSAR data processing, we do not necessarily get clear fringes over the entire area. Depending on the data pairs and places, it is not uncommon that no fringes are observed. To detect clear fringes, the reflected waves received at master and slave acquisitions must be more or less correlated to each other. The degree of correlation is quantified as *coherence*, and there are two independent decorrelation sources.

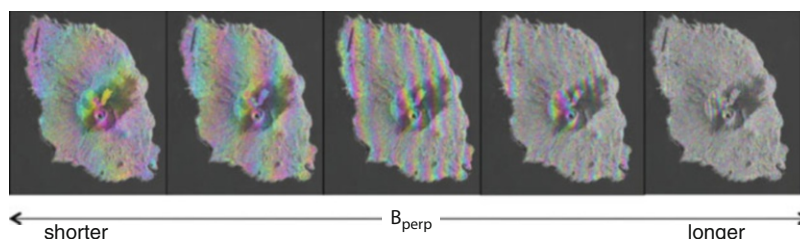
The first source of decorrelation originates in the imaging geometry. As Fig. 6 indicates, we observed higher (fewer) fringe density as becomes longer (shorter); imagine the case of zero baseline length. The fringe density can be derived from the gradient of phase (Eq. 5) along the range axis:

$$\frac{\partial \phi}{\partial R} = -\frac{4\pi B_{\perp}}{\lambda R \tan \theta} + \frac{4\pi B_{\perp}}{\lambda(r_e + H) \sin \theta} \approx -\frac{4\pi B_{\perp}}{\lambda R \tan \theta}. \quad (6)$$

Namely, the fringe density is proportional to the perpendicular baseline $B_{\perp}$, and inversely proportional to the wavelength λ ; see Simons and Rosen (2007) for the case with topography. If the fringe density becomes too high to be counted within a range resolution of SAR image, we will not be able to identify any orbital fringes. This type of decorrelation is termed *baseline decorrelation* (or *spatial decorrelation*). The *critical baseline* is given as such a perpendicular baseline that gives a phase gradient 2π over the range resolution Δr ;

$$B_{\perp}^c = \frac{\lambda R \tan \theta}{2\Delta r}.$$

For a typical value of ALOS/PALSAR with $\lambda = 23(\text{cm})$, $R = 870(\text{km})$, $\theta = 34^\circ$, $\Delta r = 5(\text{m})$, the critical baseline becomes $B_{\perp}^c = 135,000(\text{m})$, which gives an upper limit of $B_{\perp}$. However, we practically prefer much shorter $B_{\perp}$, generally less than $\sim 2,000 \text{ m}$ for ALOS/PALSAR, because in more realistic situations the effect of topography also comes in. The longer the B , the more sensitive to rugged terrain as Fig. 6 indicates. To eliminate topographic fringes, we need more accurate and higher resolution DEM if the $B_{\perp}$ becomes longer. Massonnet et al. (1996) proposed an alternative approach that could effectively reduce the $B_{\perp}$ by a combination of integer multiplied (wrapped) interferograms. For instance, if one interferogram with perpendicular baseline of 300 m is combined with the other interferogram with perpendicular baseline of 290 m with factors 1 and -1 , the effective perpendicular baseline becomes 10 m. The scaling operation,

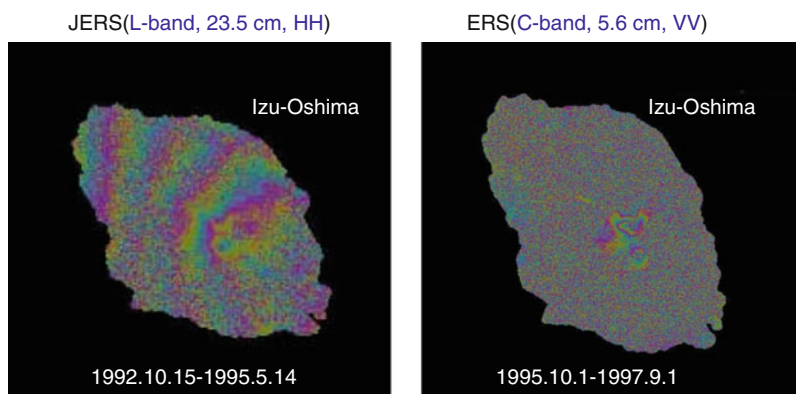


SAR Interferometry, Fig. 6 The fringe rate (density) depends on the B_{perp} ; see Eq. (6). The shorter the B_{perp} , the fewer the observed fringes, and thus better to detect deformation signals. In order words, there is a

limit in the B_{perp} over which we cannot count the number of fringes. The InSAR image is based on JERS data over Izu-Oshima volcano island, Japan. Original SAR data is copyrighted by JAXA and MITI, Japan

SAR Interferometry,

Fig. 7 Comparison of two interferograms at Izu-Oshima volcano, derived from (left) L-band HH JERS data and (right) C-band VV ERS data. While clear fringes are observed to the left even with 2.5 years temporal baseline, we can recognize the fringes only around the caldera that are covered with few vegetations



however, also scales the amount of noise, and thus the approach is limited to small integer numbers.

The second type of decorrelation is termed *temporal decorrelation*, which is related to the scattering phase in the Eq. (2a), and originates in how the microwave pulses interact with the physical objects near the ground. We often encounter the temporal decorrelation problem over vegetated areas with C-band (shorter-wavelength) SAR data and/or snow-covered areas; see Fig. 7. It should be recalled that each pixel value in SLC image is a superposition of all the reflected microwaves from all scatterers inside each resolution cell ($\sim 5 \times \sim 10$ m). Short-wavelength microwave pulses tend to be reflected on the vegetation canopies before reaching the ground surface, and their random motion will result in different scattering phases at different acquisition time, causing temporal decorrelation. On the contrary, long-wavelength microwave pulses can more easily reach the ground, which does not move as rapidly as vegetations, and thus the resulting scattering phases will be also stable over time. Besides the selection of wavelength, the polarization of microwave is also essential for better coherence over time. While, most presently, operated satellite-SAR sensors are capable of multi-polarization modes, it was shown that HH-polarization gives better coherence than VV-polarization (Cloude and Papathanassiou 1998). This is because the HH-polarized pulses can more easily penetrate through vegetations.

Outlook for InSAR Geodesy

Limitations of present InSAR: Although it has a potential to detect tens of km-scale or even larger-scale secular deformation signals on the order of mm/year, InSAR technique has been most successfully applied to detection of spatially localized signals on the order of centimeters or more, such as those associated with earthquakes, volcanic eruptions, and ground subsidence. This is because the artifacts due to inaccurate satellite orbit data and/or microwave propagation delays (advances) in the troposphere (ionosphere) can mask small-amplitude, long-wavelength deformation signals that are similar in both their amplitude and the spatial scale.

Although high-precision orbit data are indispensable to correct for the orbital fringes in the initial interferograms, their errors even on the order of 10 cm or less will generate non-negligible long-wavelength artifacts, which usually look like curved surfaces in the entire interferogram (e.g., Hanssen 2001). Conventionally, they are fitted with low-order polynomials and simply taken out unless any sort of stacking or time-series analysis discussed below is applied. While this procedure works to eliminate the artifacts due to orbit errors, it will also take out any long-wavelength geophysically interesting signals such as the inter-seismic, post-seismic, ocean tidal loading, solid-Earth tide, and post-glacial rebound signals. Alternatively, if the ground control points (GCP) are available, where the precision ground deformation data are

available, we can reestimate the baseline, based on those GCP data (e.g., Rosen et al. 1996), but such data are often unavailable in remote areas.

One approach to correct for the tropospheric delay signals is to employ the other independent estimates derived from either the GPS-based tropospheric delay estimates (e.g., Onn and Zebker 2006) or the output results from high-resolution numerical weather model (e.g., Foster et al. 2006). These so-called calibration approaches are, however, not easily applicable. The dense ground-based GPS network is limited to a few areas in the world. Also, high-resolution numerical weather model still needs significant computational resources.

Besides the tropospheric delay problem, the effects of ionosphere on both interferograms and pixel-offset images were clearly recognized in the results of the 2008 Wenchuan earthquake (Mw7.9), China, derived from ALOS/PALSAR (Kobayashi et al. 2009; Raucoules and de Michele 2010), although they were pointed out in polar region many years ago (e.g., Matter and Gray 2002). It is well known that the lower the carrier frequency is, the more significant the ionospheric dispersion impacts on the propagation delay. Thus, in many of the previous applications of C-band SAR data, the effects of ionosphere could have been neglected. While GPS also employs L-band, the high-precision GPS geodetic survey corrects for the ionospheric effect with the use of dual frequency, L1 and L2, observation data. In contrast, PALSAR is a single frequency SAR sensor and incapable of the standard ionosphere-correction approach. Empirically, however, we will encounter the ionospheric signals more frequently in the ascending data acquired in the local nighttime than in the descending data acquired in the local daytime. We also recall that the JERS, the other L-band SAR operated during 1992–1998, did not reveal any significant ionospheric signals at least in mid-latitude regions, and that most of the JERS data were acquired in the descending track. Besides the latitude, the effects of ionosphere on SAR image might, therefore, significantly depend on the data acquisition time. Like the tropospheric effects, detailed studies of ionospheric impacts on the SAR data are also currently underway.

A simple approach to eliminate those noises is *stacking*, in which several interferograms are stacked to isolate small-amplitude signals, because those noises can be regarded as temporally random, whereas the deformation signals are spatially persistent. Two important prerequisites for successful stacking are: (1) the data acquisition dates of those interferograms should not be overlapped, in order not to enhance the noises of any particular acquisition date(s), and (2) each temporal baseline should be as long as possible so that each interferogram can include as much deformation signals as possible. In reality, it is not easy to gather many independent interferograms that have desirably long temporal baselines because the available data often encounter the spatial and temporal decorrelation. Also, the simple stacking approach

inherently assumes temporally linear evolution in the ground deformation, preventing us from deriving time-series data.

Time-Series Analysis: Ferretti et al. (2000, 2001) proposed a new analysis technique called *Permanent Scatterer InSAR (PS-InSAR)*, in which they take advantage of even such data pairs whose spatial baselines are longer than the critical values. Thereby, they could expand the temporal coverage, and thus could estimate the long-term deformation signals on the order of mm/year. Key idea of PS-InSAR is to pick up only such pixels that will exhibit long-term coherence due to the existence of corner-reflector-like targets, which Ferretti et al. called “permanent scatterers.” Based on those pixels alone, they generate a stack of differential interferograms, using available DEM and orbit data. The phase values include not only deformation signals, but also such topographic signals that were not initially taken into account, because the longer spatial baseline pairs are so sensitive to the topography that the available DEM could not account for. In PS-InSAR and its variants (Werner et al. 2003; Hooper et al. 2004), they fit the differential interferogram stack to a phase model that describes not only temporal evolution of deformation but also corrections to the available DEM. Deviations from the phase model can be filtered into either non-linear deformation or atmospheric signals because the former signals are correlated and thus low-pass filtered along temporal axis, while the latter signals are temporally random; the orbit data must be assumed to be correct. A known limitation of PS-InSAR is its rather lower sampling density over non-urban areas. However, despite a lack of man-made objects, Furuya et al. (2007) succeeded in detecting active salt tectonic motion, applying a similar technique to Canyonlands National Park, Utah, presumably because the area was non-vegetated and the exposed surface rocks behaved like corner-reflector-like targets.

Another time-series analysis approach was devised and known as small baseline subset (*SBAS*) algorithm (Berardino et al. 2002). Key idea of the SBAS algorithm is *least-squares inversion* of unknown deformation at each SAR data acquisition epoch, based on the available unwrapped differential interferograms (e.g., Lundgren et al. 2001; Schmidt and Burgmann 2003). Using small baseline interferometric pairs, the SBAS approach is free from spatial decorrelation and allows us to take advantage of the fine spatial resolution of InSAR data. If the number of interferograms is greater than or equal to the number of SAR acquisitions, the inversion problem becomes an over-determined or well-determined problem, and can be easily solved in a least-squares approach. It is uncommon, however, that all the available interferometric pairs have short baselines, and accordingly the temporal sampling rate will decrease. Berardino et al. (2002) proposed to employ several groups of “small baseline subset” to overcome the lower temporal resolution issue, and solved the rank-deficient problem with the use of singular value decomposition (SVD) technique.

The SVD gives the minimum-norm least-squares solution, which is equivalent to minimizing the estimated velocities at any time intervals.

Time-series analysis of SAR data is a promising technique, but almost all previous analyses are based on the C-band ERS1/2 and Envisat data, because not only more-than-decade-long data but also high-precision, well-controlled satellite orbits are available for these satellites. As noted before, not all geophysically interesting phenomena could be detected by C-band and shorter-wavelength SAR data. If the L-band ALOS/PALSAR data are archived for a much longer time, and if the follow-on ALOS-2 and the DESDynI are launched as scheduled, the time-series analysis of SAR data will become feasible even in areas that have never been monitored before. The time-series analysis with ScanSAR data should also be possible. Long-term continuous monitoring with L-band SAR will provide us with more opportunities for new discoveries.

Cross-References

- [Earthquake Rupture: The Inverse Problem](#)
- [Earthquake Source Theory](#)
- [Earthquakes and Crustal Deformation](#)
- [Geodesy, Ground Positioning, and Levelling](#)
- [GPS, Tectonic Geodesy](#)
- [Inverse Theory, Global Optimization](#)
- [Inverse Theory, Linear](#)
- [Inverse Theory, Monte Carlo Method](#)
- [Inverse Theory, Singular Value Decomposition](#)
- [Remote Sensing and GIS Techniques for Tectonic Studies](#)
- [Remote Sensing, Applications to Geophysics](#)
- [Slow Earthquake](#)

Bibliography

- Bamler R, Hartl P (1998) Synthetic aperture radar interferometry. *Inverse Problems* 14:R1
- Berardino P, Fornaro G, Lanari R, Sansosti E (2002) A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms. *IEEE Trans Geosci Remote Sens* 40:2375
- Burgmann R, Rosen PA, Fielding EJ (2000) Synthetic aperture radar interferometry to measure Earth's surface topography and its deformation. *Annu Rev Earth Planet Sci* 28:169
- Cloude SR, Papathanassiou KP (1998) Polarimetric SAR interferometry. *IEEE Trans Geosci Remote Sens* 36:1551
- Cumming IG, Wong FH (2005) Digital processing of synthetic aperture radar data: algorithm and implementation. Artech House, Boston
- Curlander JC, McDonough RN (1991) Synthetic aperture radar: systems and signal processing. Wiley interscience, New York
- Farr TG et al (2007) The shuttle radar topography mission. *Rev Geophys* 45:RG2004
- Ferretti A, Prati C, Rocca F (2000) Nonlinear subsidence rate estimation using permanent scatterers in differential SAR interferometry. *IEEE Trans Geosci Remote Sens* 38:2202
- Ferretti A, Prati C, Rocca F (2001) Permanent scatterers in SAR interferometry. *IEEE Trans Geosci Remote Sens* 39:8
- Fialko Y, Simons M, Agnew D (2001) The complete (3-D) surface displacement field in the epicentral area of the 1999 Mw7.1 Hector mine earthquake, California, from space geodetic observations. *Geophys Res Lett* 28:3063
- Foster J, Brooks B, Cherubini T, Shacat C, Businger S, Werner CL (2006) Mitigating atmospheric noise for InSAR using a high resolution weather model. *Geophys Res Lett* 33:L16304
- Furuya M, Mueller K, Wahr J (2007) Active salt tectonics in the Needles District, Canyonlands (Utah) as detected by interferometric synthetic aperture radar and point target analysis: 1992–2002. *J Geophys Res* 112:B06418
- Ghilia DC, Pritt MD (1998) Two dimensional phase unwrapping: theory, algorithms, and software. Wiley, New York
- Hanssen RF (2001) Radar interferometry: data interpretation and error analysis. Kluwer, Dordrecht
- Hooper A, Zebker H, Segall P, Kempes B (2004) A new method for measuring deformation on volcanos and other natural terrains using InSAR persistent scatterers. *Geophys Res Lett* 31:L23611
- Kobayashi T, Takada Y, Furuya M, Murakami M (2009) Location and types of ruptures involved in the 2008 Sichuan earthquake inferred from SAR image matching. *Geophys Res Lett* 36:L07302
- Lundgren P, Usai S, Sansosti E, Lanari R, Tesauro M, Fornaro G, Berardino P (2001) Modeling surface deformation observed with synthetic aperture radar interferometry at Campi Flegrei caldera. *J Geophys Res* 106(B9):19355
- Massonnet D, Feigl KL (1998) Radar interferometry and its application to changes in the earth's surface. *Rev Geophys* 36:331
- Massonnet D, Rossi M, Carmona C, Adragna F, Peltzer G, Feigl K, Raboute T (1993) The displacement field of the landers earthquake mapped by radar interferometry. *Nature* 364:138
- Massonnet D, Vadon H, Rossi M (1996) Reduction of the need for phase unwrapping in radar interferometry. *IEEE Trans Geosci Remote Sens* 34:489
- Matter KE, Gray AL (2002) Reducing ionospheric electron density errors in satellite radar interferometry applications. *Can J Remote Sens* 28:583
- Michel R, Avouac J-P, Taboury J (1999) Measuring ground displacements from SAR amplitude images: application to the landers earthquake. *Geophys Res Lett* 26:875
- Motagh M, Wang R, Walter TR, Bürgmann R, Fielding E, Anderssohn J, Zschau J (2008) Coseismic slip model of the 2007 august pisco earthquake (Peru) as constrained by wide swath radar observations. *Geophys J Int* 174:842
- Onn F, Zebker HA (2006) Correction for interferometric synthetic aperture radar atmospheric phase artifacts using time series of zenith wet delay observations from a GPS network. *J Geophys Res* 111:B09102
- Pritchard ME (2006) InSAR, a tool for measuring Earth's surface deformation. *Phys Today* 59(7):68
- Raucoules D, de Michele M (2010) Assessing ionospheric influence on L-band SAR data: implications on Coseismic displacement measurements of the 2008 Sichuan earthquake. *IEEE Geosci Remote Sens Letters* 7:286
- Rosen PA, Hensley S, Zebker HA, Webb FH, Fielding EJ (1996) Surface deformation and coherence measurements of Kilauea volcano, Hawaii, from SIR-C radar interferometry. *J Geophys Res* 101(E10):23109
- Schmidt DA, Burgmann R (2003) Time-dependent land uplift and subsidence in the Santa Clara valley, California, from a large interferometric synthetic aperture radar data set. *J Geophys Res* 108(B9):2416
- Simons M, Rosen PA (2007) Interferometric synthetic aperture radar geodesy. In: Herring TA (ed) *Treatise on geophysics*, vol 3. Elsevier, New York, pp 391–446
- Tobita M, Murakami M, Nakagawa H, Yurai H, Fujiwara S (2001a) Two-dimensional field of three-dimensional components of deformations

- and velocities, and volume change around Usu volcano associated with the 2000 eruption by matching of SAR images (in Japanese). *J Geogr Survey Inst* 95:37
- Tobita M, Murakami M, Nakagawa H, Yarai H, Fujiwara S, Rosen PA (2001b) 3D surface deformation of the 2000 Usu Eruption measured by matching of SAR images. *Geophys Res Lett* 28:4291
- Werner CL, Wegmuller U, Strozzi T, Wiesmann A (2003) Interferometric point target analysis for deformation mapping, paper presented at IGARSS'03. Geoscience Remote Sensing Society, Toulouse
- Zebker HA, Rosen PA, Goldstein RM, Gabriel A, Werner CL (1994) On the derivation of coseismic displacement fields using differential radar interferometry: the landers earthquake. *J Geophys Res* 99(B10):19617–19634
- Zhou XB, Chang NB, Li SS (2009) Applications of SAR interferometry in Earth and environmental science research. *Sensors* 9:1876

Satellite Altimetry

Stelios P. Mertikas and Constantine Kokolakis
Geodesy and Geomatics Engineering Laboratory, School of Mineral Resources Engineering, Technical University of Crete, Chania, Greece

Synonyms

Earth observation; Marine geodesy; Remote sensing; Satellite monitoring; Sea surface height

Definition

Altimetry is a technique for measuring heights.

Satellite altimetry measures the time taken by an electromagnetic pulse to propagate from the satellite to the surface and back to the satellite

Satellite Altimetry

Satellite altimetry is a technique of measuring heights of the Earth's surface from a satellite flying approximately at an attitude of about 1000 km. The word "altimetry" originates from the Latin "altus," meaning "high" and the Greek suffix "metry" ("μετρία," as in Geo-metry, Trigono-metry, etc.) denoting the science or scientific discipline related to measuring heights. At the beginning, the objective of altimetry has been to measure height principally of sea surface to an accuracy of about ± 1 cm, continuously, objectively, and on a global scale.

Altimeters measure distances, or ranges, from a satellite to the Earth's surface and principally onto the sea, inland water levels (lakes, rivers) and ice surfaces as well (Fig. 1).

Measurements of altimetric heights are thus creating profiles of the Earth's surface along satellite orbits with respect to Earth's center of mass.

Satellite observations coming from altimeters, like those of Sentinel-3, Jason-3, CryoSat-2, IceSat-2, HY-2, and others, address several applications in solid Earth geophysics. Gravity determination, natural resources exploration, ocean circulation and variability, ice topography, hydrology, geodesy, climate change, bathymetry, sea-floor topography, etc., are a few applications of satellite altimetry.

Measurement Principles in Altimetry

Altimeters on board a satellite transmit radar or laser pulses towards the surface of the Earth. Propagated signals are then reflected by this surface and finally are received by the altimeter on the satellite. Ranges are realized by measuring time differences between satellite transmission and reception of these electromagnetic pulses. Each range value is an average of all measurements made within the footprint of the altimeter. Time series of surface height and its change can be constructed for a location along the satellite ground track during its lifetime.

Altimeter ranges require proper measurement reductions to eliminate various error sources arising from orbit determination, atmosphere delays along the propagation path of the signals (wet and dry troposphere, ionosphere), antenna gain pattern, satellite clocks, geoid height, tides (ocean, solid earth, pole), and so on.

Analysis and evaluation of return signals and specifically of their waveforms provides information upon measured range, wind speed, wave heights, and backscatter coefficient (sigma-naught) of observed targets (Fig. 2). Precise heights for oceans, lakes, rivers, sea ice, glaciers, and land surfaces are thus measured from an altitude of 800–1300 km with an accuracy of about ± 1 cm (or better).

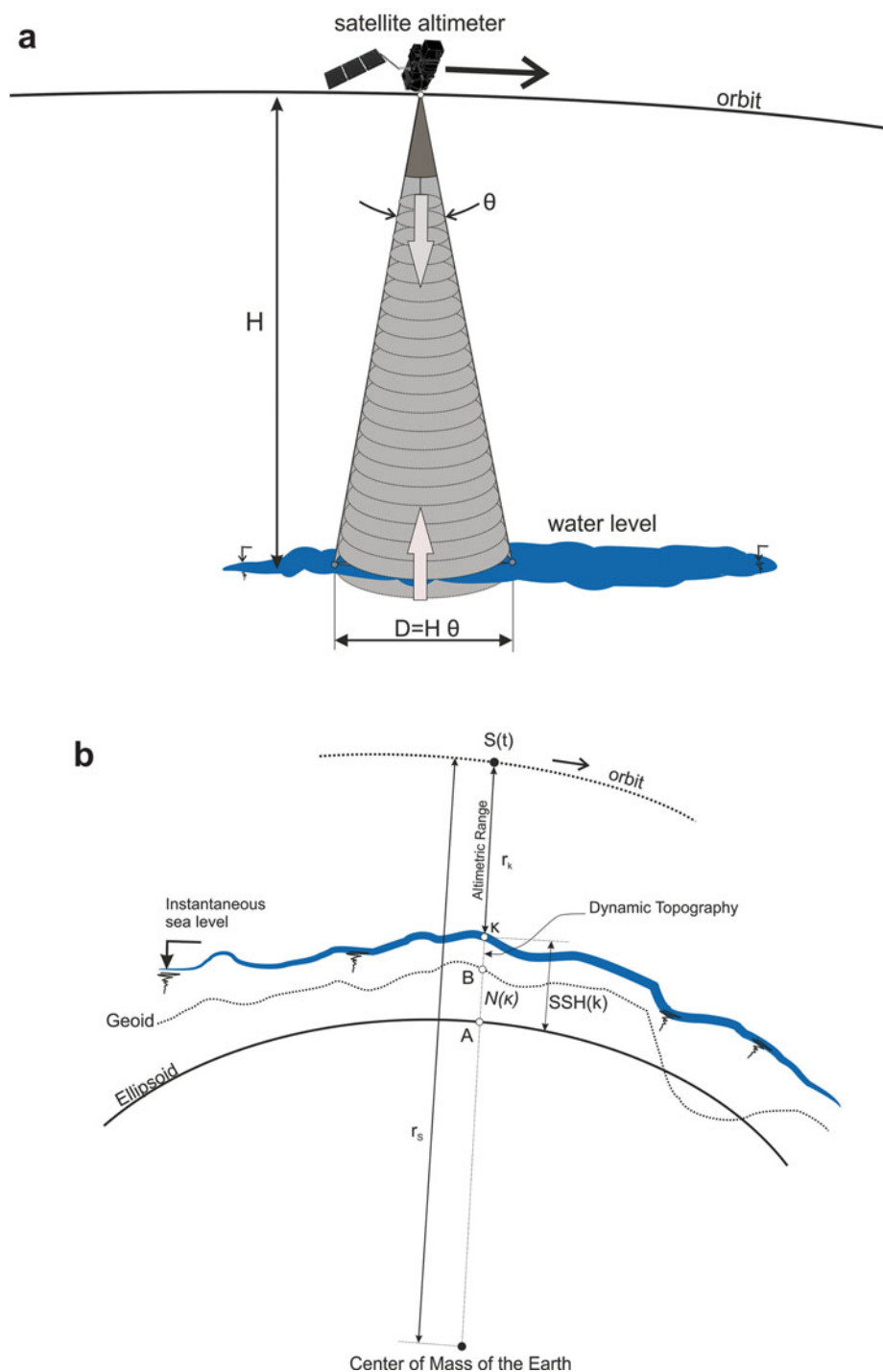
Altimetric observations are made independent of night or day, of strong winds or hurricanes, cloudy or clear weather conditions. In this manner, altimetry is producing a wealth of Earth observations continuously, accurately, objectively, repeatedly, and on a global scale.

As of 1992 with its first TOPEX/Poseidon, it has been generating precise time series of altimetric observations for building up a climate data record of ocean topography, sea level, and inland water levels. Time-averaged heights of the sea surface essentially represent the geoid, showing the variability of the gravity field at the earth's surface, caused by the uneven distribution of the density of earth's crust (Fig. 3).

The slope of ocean surface, related to changes in gravity and submarine features, is estimated by differencing height measurements along altimeter tracks. Altimetry has in that

Satellite Altimetry,

Fig. 1 Principle of satellite altimetry. The satellite altimeter transmits signals towards the Earth, which are later received after their reflection onto the Earth's surface. After orbit determination, and several measurement reductions, height profiles of the Earth's surface are established with respect to its center of mass

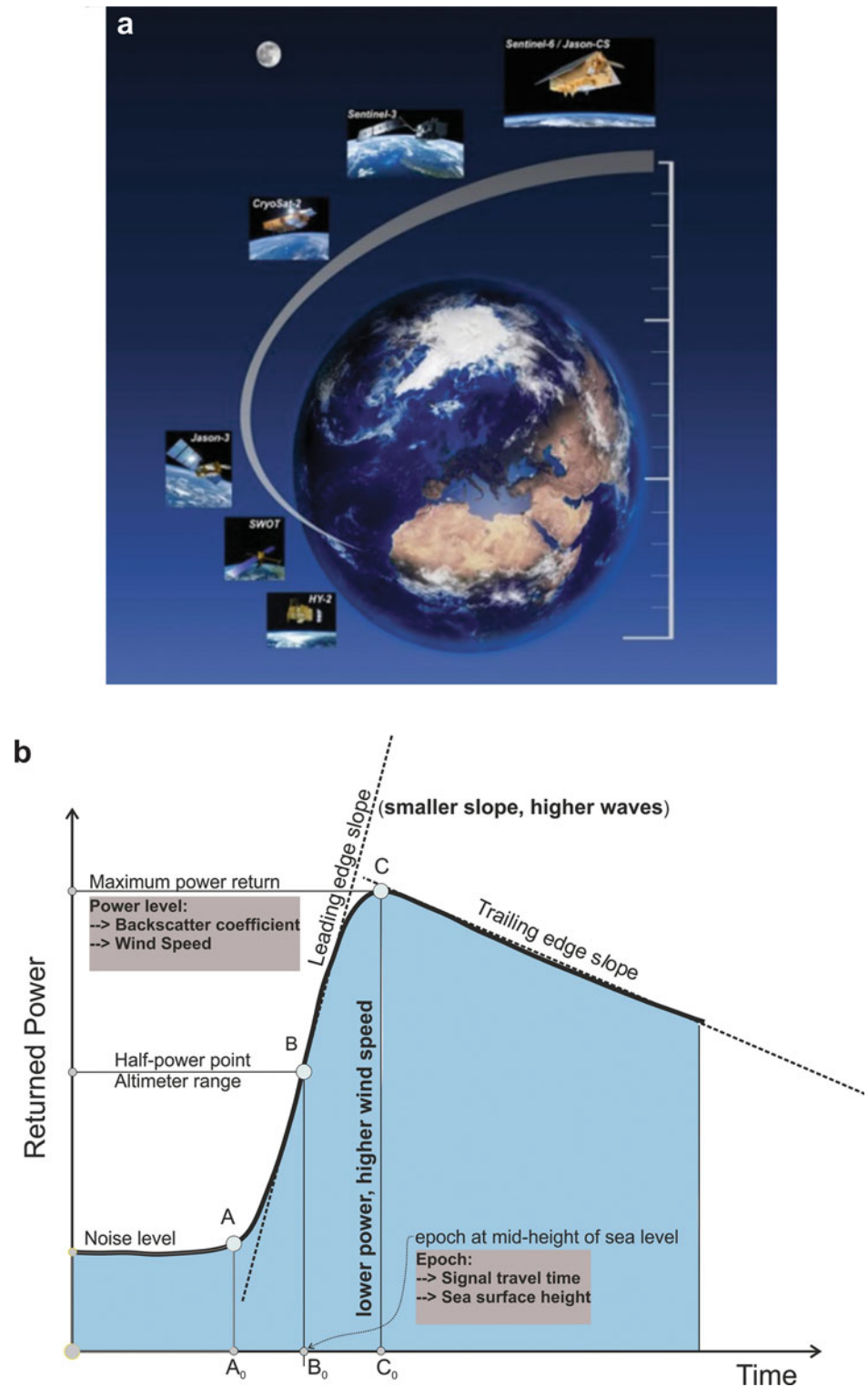


way revealed subsea mountains and features and made it possible to construct a uniform resolution map of the seafloor topography. A detailed knowledge of seafloor topography is fundamental to the understanding of several Earth processes.

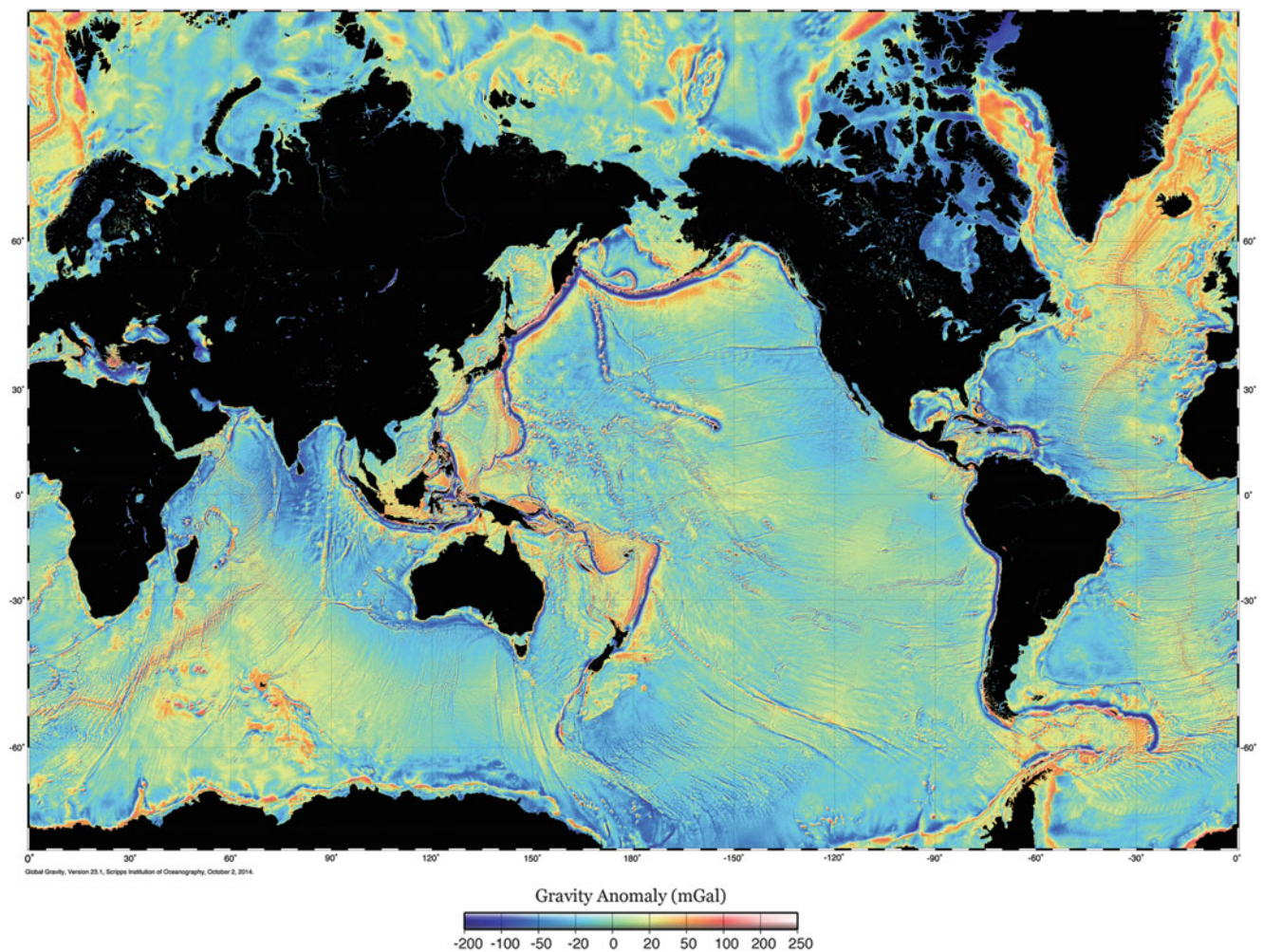
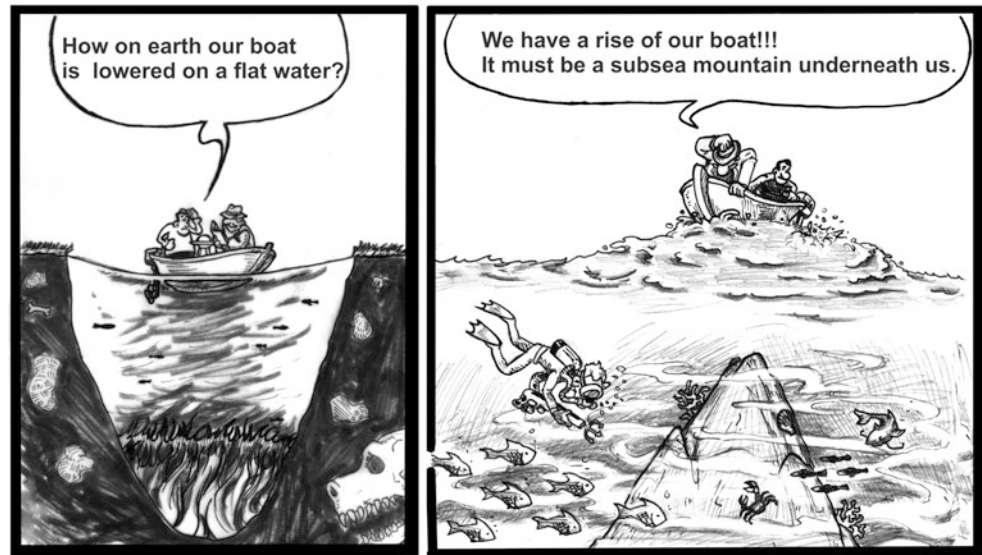
Also, gravity and bathymetry data derived from altimetry have been used to construct global models of marine gravity to a latitude of 85° with accuracies of ± 1 mGal at 7 km resolution. Such models have helped to determine gravity models for remote and inaccessible regions and also

Satellite Altimetry,

Fig. 2 Examples of satellite altimeters for Earth observation ((a) Upper image). The lower (b) diagram presents an example of a waveform for a return signal on an altimeter. Ranges are determined by measuring time difference at the half-power point (point B_0). The slope of the leading edge (AC leg) of waveform defines wind speed. The level of return power represents wave heights (Point C). The rise time of the leading edge (A_0C_0) corresponds to significant wave height



Satellite Altimetry, Fig. 3 The sea surface exhibits slopes, tiny ripples and variations (enlarged in the cartoons above), commonly invisible to human eye, but related to bathymetry and gravity changes underneath it (cartoon by Nikitas Papadoulakis)



Satellite Altimetry, Fig. 4 A global map of marine gravity as produced by combining satellite altimetry and gravity data (Sandwell and Smith 2009; Sandwell et al. 2014)

to identify submarine canyons, faults, and plate boundaries and oceanic plateaus (Fig. 4). These geomorphic features provide evidence to where to look for deposits of natural resources.

Oceanography by Altimetry

Oceans cover around 71% of the Earth's surface and contain about 97% of its water. Their role is crucial in the balance of Earth's climate. They control the amount of CO₂ in the atmosphere through plankton, absorb 90% of the thermal energy of the Earth, form microclimates, regulate fish population by ocean circulation and currents, shape and curb the greenhouse effects, etc.

Before Seasat in 1978, which was the first satellite to observe Earth's ocean, measurements for ocean monitoring were constrained to coastlines with tide gauges and/or along ship's routes. Such ocean observations on the surface are commonly biased as they are subject to and contaminated by local and dynamic effects, bathymetry, and so on. Up to that point, observations made at the Earth's surface have prevented geophysicists from understanding global oceans' mechanisms. So, satellite altimetry came to fill in this gap by providing the global picture with accuracy and objectivity.

In altimetry, information about the radiated ocean surface originates from the received echo of the transmitted satellite pulse. Over ocean, this response has a particular shape, a waveform of the Brown model as shown in Fig. 2 (bottom). The epoch at half-power point (corresponding to the mid-height of sea level) defines the altimeter range; the maximum power return provides the reflexivity coefficient of the altimeter signal on the surface (called sigma-naught and related to properties of the scattering surface); the leading edge slope of the signal return determines the amplitude of sea waves (significant wave height); and the slope of its trailing edge is connected with the off-nadir pointing of the altimeter antenna (Passaro et al. 2014).

Satellite altimetry is thus able to observe ocean surface topography, which is nothing else but the difference between the ocean surface and the geoid (the equipotential surface of the sea level). Mapping ocean surface topography with altimetry enables us to observe ocean currents and circulation. Any variation on ocean surface topography, such as a high (hill) or a low (valley) at a location, implies that currents swirl around it, primarily driven by thermodynamic processes. The Gulf Stream which flows from the Caribbean to the Grand Bank off Newfoundland is such an example. This Stream has been regularly monitored from the onset of altimetry with the TOPEX/Poseidon satellite. During its course, the Gulf Stream warms the ocean and releases heat and water vapors into the atmosphere. So it modifies and

regulates the exchanges of heat between the ocean and the atmosphere and has a straight through impact on the climate and its changes.

Altimetry has also contributed to the improvement of tidal models. Nowadays, these models have an uncertainty of ± 2 cm in open seas. This level of uncertainty allows us to examine and better understand gravity interactions between the Moon (main cause of tidal effects) and the Earth. Daily sea level measurements from altimetry have given the opportunity to adjust model parameters according to real satellite measurements in open seas where no previous observations could have been validated them (Fu and Cazenave 2000). These accurate tidal models are essential in forecasting extreme ocean events, such as tsunamis.

Satellite altimetry supports absolute and undisputable sea level monitoring over regional to global scales, accurate to some [mm/yr] and with respect to the center of mass of the Earth (Mertikas et al. 2018). Sea-level is a key indicator of the impact of climate change over oceans. It is one of the essential climate variables (Downy 2016) that needs to be systematically monitored for maintaining a global climate observing system.

Hydrology by Altimetry

Hydrology is composite word which originates from the Greek "hydro" (ὕδωρ), meaning water and the ending "logy" (λογία), corresponding to "thorough investigation." Thus, hydrology is a scientific field of Geophysics that studies the movement, distribution, and quality of water on Earth. Although the earliest altimetry missions have been dedicated to observing the open ocean and ice, recent altimetry has moved into monitoring every stretch of water on Earth (enclosed seas, lakes, rivers, flooding areas, and so on) provided satellites fly over them.

An altimeter is capable of monitoring these water levels of lakes and their variations. To our advantage altimetry monitors regions difficult or even impossible to access for in situ measurements. Levels of lakes rise and drop throughout seasons, depending upon how much water flows in (e.g., rain, snow melting) or flows out their water bodies (e.g., evaporation, pumping for irrigation). Thus, the level of water is an indicator of regional climate variations, and it is important to monitor it to be able to access quantity of water and stretch of a lake. This is achieved by altimetry (water height and volume) in combination with imagery (lake stretch). The same holds true for enclosed seas, such as the Caspian Sea.

Some altimeters with specific ground tracks are capable of monitoring adjacent lakes, as the African lakes of Victoria, Tanganyika and Malawi, and also correlate their water levels. The European Remote Sensing satellite (ERS) and its

successor Envisat had observed 215 large lakes out of 842 that exist globally. River levels and flooded areas can also be observed by altimetry. Satellite altimeters can observe heights of large rivers such as the Amazon or the Brahmaputra and monitor their changes over seasons.

The recent observation mode of synthetic aperture radar (SAR) altimetry takes advantage of Doppler effect of earth targets with respect to the moving satellite and achieves an improvement in spatial resolution, along track, of about 300 m contrary to 1–5 km in conventional altimetry (also called low resolution mode). The CryoSat-2 satellite and Sentinel-3A and Sentinel-3B implement this SAR technique which allows them to monitor narrow straits, coastal lines, and inland water resources reliably. In SAR mode, the satellite emits a series of coherent pulses (called bursts) and thus multiple signal returns (echoes) are received by the altimeter for the same target on the surface of the Earth. After processing these multiple returns, better resolution and more precise altimeter ranges are produced. SAR altimetry has applications for inland water observations, such as the CryoSat-2 SAR altimetry over the Mekong River in South-east Asia (Bercher et al. 2013). Finally new algorithms, called fully-focused algorithms, are applied for processing altimeter signals to further improve the along track resolution down to half of the antenna length, at about 0.5 m spatial resolution.

Ice and Cryosphere Monitoring

Altimetry can also measure thickness of polar sea ice, floating ice, and glacier topography as well as monitor changes in ice sheets. Altimeters of ERS-1, ERS-2, Envisat, CryoSat-2, IceSat, and IceSat-2 have observed the cryosphere and have allowed us to map Greenland and Antarctica, where about 77% of the Earth's fresh water is located and accessibility has been difficult for surface measurements. Before satellite altimetry, in situ measurements in polar caps were sparse in space and time. This limitation had not permitted us to appreciate the true extend of ice, melting progression, and variations. As of 2010, the synthetic interferometric radar altimeter of CryoSat-2 (as well as IceSat and IceSat-2) has been acquiring elevation data for glaciers and ice caps to assess the consequences of receding ice cover as a result of global warming (McMillan et al. 2014).

Also, the interferometric synthetic aperture radar of CryoSat-2 has allowed a smaller footprint (~250 m) in the along-track (fore- and aft) direction and detection of the position of the source of the echo in its across-track (left and right) direction; thus, CryoSat-2 has been able to distinguish slopes. This characteristic has produced accurate height

measurements in regions of complex topography. Signal parameters in the waveform of the echo are similar with these in Fig. 1, with the exception that the leading-edge width gives information about signal penetration into the ice and about ice roughness.

The amount of ice sheet and their variations are not only indicators of the current climate situation of the Earth but are reliable predictors for the future. If the ice of Greenland and Antarctica melts, for example, the sea level will rise about 80 m (Gardner et al. 2013).

Summary

Satellite altimetry is a powerful observation tool for Geophysics. It has proven extremely useful for uninterrupted, global, and accurate observation of marine geodesy, oceans, ice caps, lakes, rivers, coastlines, natural resources, and many more. It has revealed mechanisms of the Earth, such as global bathymetry and ice melting that no other technique could do in the past. Recognizing that progress, diverse satellite altimeters along with advanced measuring techniques (Nadir, Delay-Doppler, interferometry altimeters, wide swath, Ku-band, Ka-band frequencies, etc.) are placed into orbits and some of them such as Sentinel-6/Jason-CS are to become operational missions. New missions SWOT (Morrow et al. 2019), Quanlan (Chen et al. 2019), CRISTAL (Kern et al. 2020), etc.) are planned by various nations, organizations, and space agencies and are on their way to support Earth observation and drive the advancement of Geophysics.

Acknowledgments We thank Professor David Sandwell of Scripps Institute of Oceanography, USA, for giving us permission to publish the map of Fig. 4.

Bibliography

- Bercher N, Dinardo S, Lucas BM, Fleury S, Calmant S, Crétaux JF, Benveniste J (2013, March) Applications of CryoSat-2 SAR & SARIn modes for the monitoring of river water levels. In: Proceedings of the CryoSat third user workshop, Dresden, Germany, pp 12–14
- Chen G, Tang J, Zhao C, Wu S, Yu F, Ma C, Xu Y, Chen W, Zhang Y, Liu J, Wu L (2019) *Front Mar Sci*. <https://doi.org/10.3389/fmars.2019.00194>
- Downy C (2016) What is an ECV? ESA-Climate Change Initiative. Retrieved 16 Feb 2019, from <http://cci.esa.int/content/what-ecv>
- Fu LL, Cazenave A (eds) (2000) *Satellite altimetry and earth sciences: a handbook of techniques and applications*, vol 69. Elsevier, New York

- Gardner AS, Moholdt G, Cogley JG, Wouters B, Arendt AA, Wahr J, Ligtenberg SR (2013) A reconciled estimate of glacier contributions to sea level rise: 2003 to 2009. *Science* 340(6134): 852–857
- Kern M, Cullen R, Berruti B, Bouffard J, Casal T, Drinkwater MR, Gabriele A, Lecuyot A, Ludwig M, Midthassel R, Traver IN, Parrinello T, Ressler G, Andersson E, Martin-Puig C, Andersen O, Bartsch A, Farrell S, Fleury S, Gascoin S, Guillot A, Humbert A, Rinne E, Shepherd A, van den Broeke M R, Yackel J (2020) The Copernicus polar ice and snow topography altimeter (CRISTAL): expected mission contributions. *Cryosphere*. <https://doi.org/10.5194/tc-2020-3>
- McMillan M, Shepherd A, Sundal A, Briggs K, Muir A, Ridout A, Wingham D (2014) Increased ice losses from Antarctica detected by CryoSat-2. *Geophys Res Lett* 41(11):3899–3905
- Mertikas S, Donlon C, Féménias P, Mavrocordatos C, Galanakis D, Tripolitisiotis A, Guinle T (2018) Fifteen years of Cal/Val service to reference altimetry missions: calibration of satellite altimetry at the permanent facilities in Gavdos and Crete, Greece. *Remote Sens* 10(10):1557
- Morrow R, Fu LL, Arduin F, Benkiran M, Chapron B, Cosme E, d'Ovidio F, Farrar JT, Gille ST, Lapeyre G, Le Traon PY, Pascual A, Ponte A, Qiu B, Rasche N, Ubelmann C, Wang J, Zaron ED (2019) Global observations of fine-scale ocean surface topography with the surface water and ocean topography (SWOT) mission. *Front Mar Sci*. <https://doi.org/10.3389/fmars.2019.00232>
- Passaro M, Cipollini P, Vignudelli S, Quartly GD, Snaith HM (2014) ALES: a multi-mission adaptive subwaveform retracker for coastal and open ocean altimetry. *Remote Sens Environ* 145: 173–189
- Sandwell TD, Smith WHF (2009) Global marine gravity from retracked Geosat and ERS-1 altimetry: ridge segmentation versus spreading rate. *J Geophys Res* 114:B0141
- Sandwell TD, Muller RD, Smith WHF, Garcia E, Francis R (2014) New global marine gravity model from CryoSat-2 and Jason-1 reveals buried tectonic structure. *Science* 346(6205): 65–67

Seafloor Spreading

Richard N. Hey
Hawaii Institute of Geophysics and Planetology, School of
Ocean and Earth Science and Technology, University of
Hawaii, Honolulu, HI, USA

Definition

Seafloor spreading is the mechanism by which new oceanic lithosphere is created at and moves away from the divergent plate boundaries known as mid-ocean ridges. The seafloor spreading hypothesis led to one of the most important paradigm shifts in the history of the Earth sciences, the plate tectonics scientific revolution.

Introduction

The revolutionary seafloor spreading hypothesis improved and subsumed the continental drift hypothesis, and rapidly culminated in what is now known as plate tectonic theory. First hypothesized by Harry Hess in a 1960 preprint and paper (Hess 1962), he considered so speculative he called it “an essay in geopoetry,” and named in another influential early paper (Dietz 1961), it offered a simple explanation for many problems with the prevailing paradigm that the Earth was a mostly static, slowly contracting planet, with fixed continents and old ocean basins, and no large-scale horizontal displacements. This paradigm had previously been challenged, most notably by Alfred Wegener with his continental drift hypothesis (Wegener 1912), and by paleomagnetic measurements in the 1950s that were consistent with continental drift, but before the 1960s these ideas were not generally accepted.

The Revolution

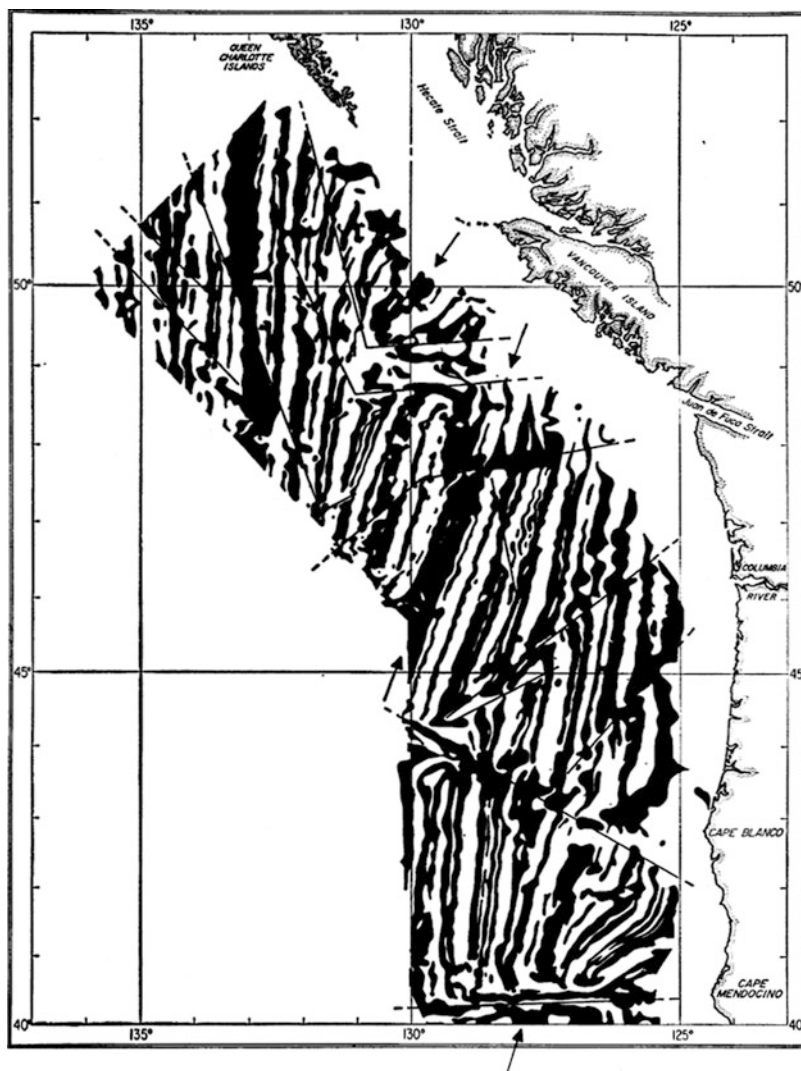
Hess realized that if mantle convection carried seafloor and continents away from seafloor spreading centers (mid-ocean ridges) toward trenches (subduction zones), with new seafloor formed at ridge axes and destroyed at trenches, this would explain the shallow bathymetry, earthquakes, high heat flow, lack of sediments, and extensional structures characterizing ridge axes, as well as the deep trenches, earthquakes, compressional structures, mountain ranges, and volcanoes characterizing subduction zones.

A key step in the confirmation of seafloor spreading was the recognition by Fred Vine and Drum Matthews (Vine and Matthews 1963) (and independently by Lawrence Morley, in a paper unfortunately turned down as too speculative by both *Nature* and the *Journal of Geophysical Research*, later published in Morley and Laroche 1964), that it offered a simple explanation for the existence of puzzling magnetic “stripes” on the seafloor. The Vine–Matthews hypothesis proposed that a combination of seafloor spreading and episodic reversals of the Earth’s magnetic field (at the time another very speculative idea) would create alternating zones of normally and reversely magnetized crust, and thus, linear positive and negative magnetic anomalies in a pattern symmetric about the spreading axis.

The next key step in the revolution occurred in 1965, when Tuzo Wilson noted that deformation of the Earth’s crust is concentrated in narrow mobile belts, and postulated that these features are all interconnected in a global network, the first qualitative model of plate tectonics (Wilson 1965). The zones

Seafloor Spreading,

Fig. 1 Raff and Mason (1961) magnetic stripes recognized as symmetric about seafloor spreading axes (*arrows*) by Vine and Wilson (1965). (From Vine 1966. Reprinted with permission from AAAS)



of extension and compression are connected by a new class of faults called transform faults by Wilson, which required relative plate motion and turned out to be the most important type of fault on Earth. He showed that seafloor spreading occurring on offset mid-ocean ridge axes would produce relative motion exactly opposite to the motion the sense of offset would predict without seafloor spreading, and that earthquakes should only occur between the offset seafloor spreading axes. These radical predictions, completely different from prevailing wisdom, were soon confirmed seismically. Furthermore, by correctly interpreting the San Andreas fault as a transform fault between the Pacific and North American plates, he predicted the existence of previously unrecognized seafloor spreading offshore western North America. Vine and Wilson (1965) showed that the predicted magnetic symmetry existed in this area, where magnetic stripes had been discovered in the mid-1950s

(Mason 1958; Mason and Raff 1961; Raff and Mason 1961), over what they recognized as the Juan de Fuca Ridge (Fig. 1), and furthermore, that the pattern of stripes corresponded perfectly with the pattern of magnetic field reversals (Cox et al. 1963; McDougall and Tarling 1963), once the Jaramillo anomaly was discovered. Vine (1966) also demonstrated similar symmetry and correlation with the reversal timescale in another important data set, the Project Magnet aeromagnetic data collected over the Reykjanes Ridge south of Iceland. His color figures of these classic data sets and correlations (Vine 1968) became iconic.

The smoking gun for many scientists was the discovery by Pitman and Heirtzler (1966) of near perfect symmetry in the Eltanin-19 profile collected over the Pacific–Antarctic Ridge. Essentially every tiny wiggle seen in the magnetic anomaly profile on the Pacific plate was mirrored on the Antarctic

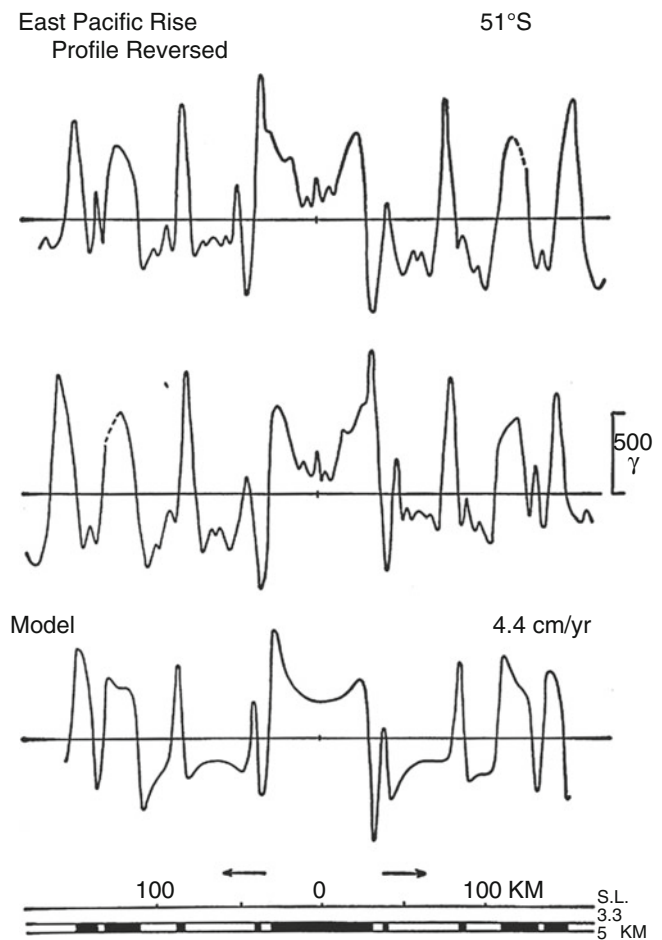
plate, and correlated perfectly with the magnetic reversal time scale (Fig. 2). The symmetry in these data required a symmetric axial process, with new seafloor carried away on both plates, and thus provided compelling evidence for both the Vine–Matthews and seafloor spreading hypotheses. Vine convincingly summarized this evidence in influential symposia and publications (Vine 1966, 1968), and, by the end of 1966, seafloor spreading was generally accepted by marine geophysicists, who quickly extrapolated the magnetic reversal time scale from ~10 Ma to ~80 Ma and worked out at least the basic recent evolutionary history of every ocean basin.

This scientific revolution culminated in plate tectonics the following year when Jason Morgan (Morgan 1968) and Dan McKenzie (McKenzie and Parker 1967) made the key assumption that plates behave rigidly, and extended Wilson's transform fault geometry to the sphere. They showed that transforms are small circles about rotation poles describing relative plate motion, that seafloor spreading rates increase as the sine of the angular distance away from these poles, and that it was possible to use known patterns of seafloor spreading to quantitatively predict other plate motions and plate boundary evolution.

Ridge Axis Geometry, Morphology, and Crustal Formation

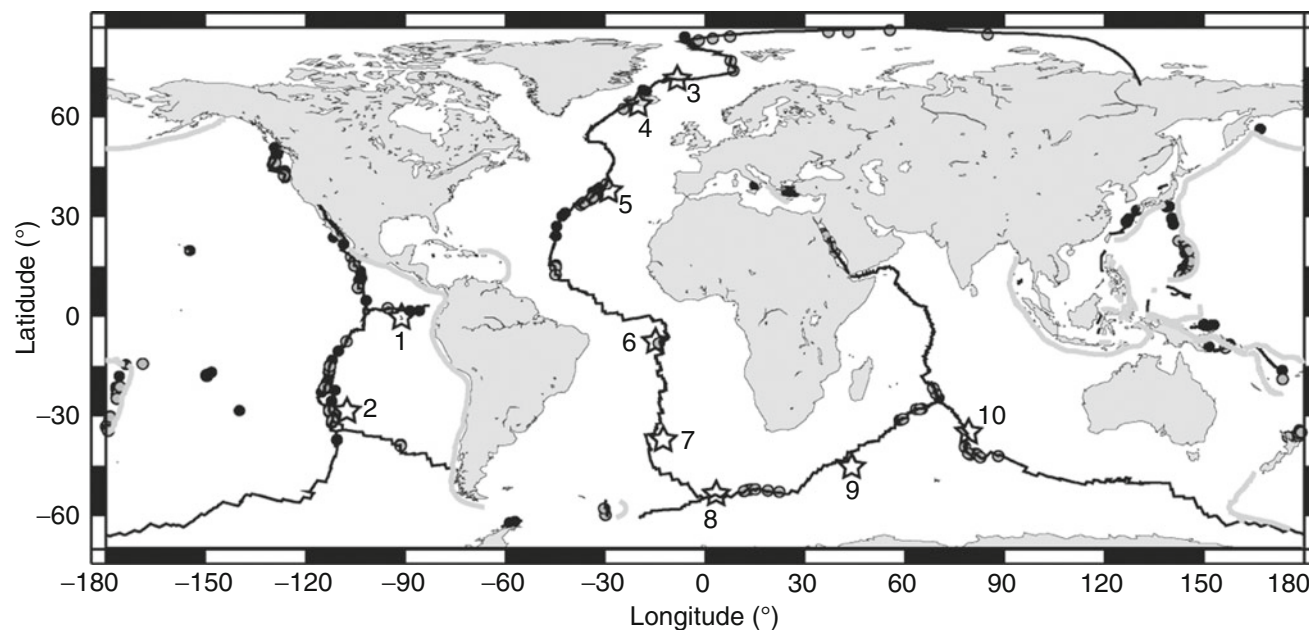
Volcanism along the mid-ocean ridge system has formed the longest mountain chain on Earth (Fig. 3). As the plates slowly (at present ~0–150 mm/yr, or km/Myr, DeMets et al. 2010, with rates remarkably similar whether using magnetic anomalies or geodetic measurements such as GPS) move apart by seafloor spreading, magma from the hot asthenosphere rises to fill the crack at mid-ocean ridge axes, elevated because of the higher temperatures. Both axial depth and morphology correlate with spreading rate (Macdonald 1982). Fast-spreading ridge axes such as the East Pacific Rise generally have shallow depths and relatively smooth morphologies, with very small extensional axial summit troughs (except where these are completely buried by the latest eruption in the neovolcanic rift zone). Slow spreading ridges such as the Mid-Atlantic Ridge generally have much deeper, rougher, and higher-amplitude axial valleys (Fig. 4), except near hotspots such as Iceland where magma supply is unusually large and even slow-spreading ridges have shallow axes with fast-spreading morphology.

The asthenosphere typically melts to become mid-ocean ridge basalt in a magma chamber under the ridge axis.



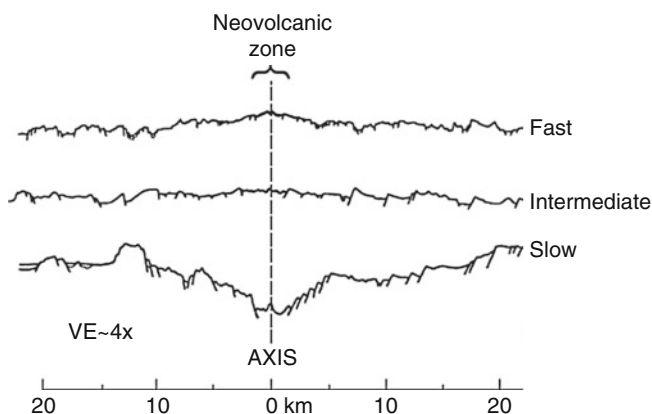
Seafloor Spreading, Fig. 2 Eltanin-19 magnetic anomaly profile (*top*) from Pitman and Heirtzler (1966), shown compared with its mirror image (*middle*) and magnetic anomaly model calculated from the magnetic reversal timescale (*bottom*), assuming seafloor spreading at a constant 44 km/Myr. (From Vine 1966. Reprinted with permission from AAAS)

Some basalt is erupted onto the seafloor to form pillow basalts and lava flows, which are progressively covered by sediments as the lithosphere ages and moves away from the axis. Below the basalts are the sheeted dikes, the paths through which the lava moved from magma chambers to the surface, and below the dikes are the layered gabbros which cooled and crystallized in place (intrusives) rather than erupting at axial volcanoes (extrusives). This typical sequence of mantle ultramafics – gabbro – sheeted dikes – extrusive basalts – sediments is observed on the seafloor in tectonic windows where existing lithosphere is rifted apart, and also where slices of old ocean lithosphere are thrust into ophiolite mountain belts as in Oman and Cyprus (Karson 2002).



Seafloor Spreading, Fig. 3 Global distribution of seafloor spreading axes (heavy black lines) with known (black dots) or inferred (gray dots) hydrothermal fields. Stars are near-ridge hotspots. (From NOAA/PMEL

Vents Program website: <https://www.pmel.noaa.gov/pubs/outstand/bake2544/images/fig01a.gif> (Baker and German 2004))



Seafloor Spreading, Fig. 4 Ridge axis fault patterns and morphology as a function of spreading rate. (Modified from Macdonald 1982)

(Fig. 5). White and clear smokers occur at lower temperatures. The chemical reactions provide energy for an unusual kind of life that does not depend on photosynthesis but flourishes as chemosynthetic communities at many of these vents, rare deep-ocean oases in an otherwise extremely barren environment (Kelley et al. 2002).

Summary

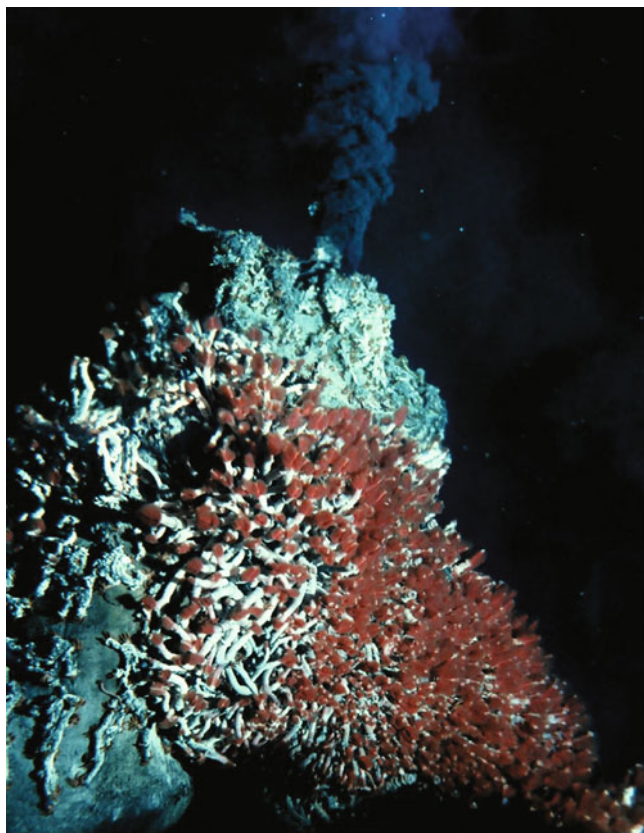
Seafloor spreading was a critical step in the contentious scientific revolution from the previous static Earth paradigm to the now universally accepted plate tectonic paradigm. Today it refers to the processes creating new oceanic lithosphere where plates move apart. Seafloor spreading replaces the lithosphere destroyed by subduction and exerts important influences on Earth's chemical and biological evolution.

Hydrothermal Vents

As the new seafloor cools it contracts and cracks. Water goes down these cracks, is heated, reacts with the surrounding rocks, and comes back up as buoyant hydrothermal vents (Baker and German 2004). These occur at a range of chemistries and temperatures, up to $\sim 400^\circ\text{C}$ in black smokers, so-called because of the sulfide-rich plumes that precipitate suddenly when the superheated water is injected into the surrounding $\sim 2^\circ\text{C}$ seawater, forming sulfide chimneys

Cross-References

- [Continental Drift](#)
- [Continental Rifts](#)
- [Lithosphere, Mechanical Properties](#)
- [Lithosphere, Oceanic](#)
- [Plate-Driving Forces](#)
- [Plates and Paleoreconstructions](#)
- [Subduction Zones](#)



Seafloor Spreading, Fig. 5 A black smoker community comprised of giant red tubeworms and hundreds of squat lobsters. This vent is located in Strawberry Fields of the Main Endeavour hydrothermal field on the Juan de Fuca Ridge. Vibrant colonies of tube worms with red gills thrive on this vent which is predominantly composed of iron- and sulfur-bearing minerals. (Credit: University of Washington; NOAA/OAR/OER)

Bibliography

- Baker ET, German CR (2004) On the global distribution of hydrothermal vent fields. In: German CR et al (eds) *Mid-ocean ridges: hydrothermal interactions between the lithosphere and oceans*. Geophysical monograph, vol 148. American Geophysical Union, Washington, DC, pp 245–266
- Cox A, Doell RR, Dalrymple GB (1963) Geomagnetic polarity epochs and Pleistocene geochronometry. *Nature* 198:1049–1051
- DeMets C, Gordon RG, Argus DF (2010) Geologically current plate motions. *Geophys J Int* 181(1):1–80. <https://doi.org/10.1111/j.1365-246X.2009.04491.x>
- Dietz RS (1961) Continent and ocean basin evolution by spreading of the sea floor. *Nature* 190:854–857
- Hess HH (1962) History of ocean basins. In: Engel AEJ, James HL, Leonard BF (eds) *Petrologic studies: a volume to Honor A.F. Buddington*. Geological Society of America, New York, pp 599–620
- Karson JA (2002) Geologic structure of the uppermost oceanic crust created at fast- to intermediate-rate spreading centers. *Annu Rev Earth Planet Sci* 30:347–384
- Kelley DS, Baross JA, Delaney JR (2002) Volcanoes, fluids and life at mid-ocean ridge spreading centers. *Annu Rev Earth Planet Sci* 30:385–491
- Macdonald KC (1982) Mid-ocean ridges: fine scale tectonic, volcanic and hydrothermal processes within the plate boundary zone. *Annu Rev Earth Planet Sci* 10:155–190
- Mason RG (1958) A magnetic survey over the west coast of the United States between latitudes 32° and 36° N, longitudes 121° and 128° W. *Geophys J R Astron Soc* 1:320–329
- Mason RG, Raff AD (1961) A magnetic survey off the west coast of North America 32° N to 42° N. *Bull Geol Soc Am* 72:1259–1265
- McDougall I, Tarling DH (1963) Dating of polarity zones in the Hawaiian islands. *Nature* 200:54–56
- McKenzie DP, Parker RL (1967) The North Pacific: an example of tectonics on a sphere. *Nature* 216:1276–1280
- Morgan WJ (1968) Rises, trenches, great faults and crustal blocks. *J Geophys Res* 73:1959–1982
- Morley LW, Laroche A (1964) Paleomagnetism as a means of dating geological events. In: *Geochronology in Canada*. Royal Society of Canada special publication, vol 8. Royal Society of Canada, Toronto, pp 39–50
- Pitman WC III, Heirtzler JR (1966) Magnetic anomalies over the Pacific–Antarctic ridge. *Science* 154:1164–1171
- Raff AD, Mason RG (1961) Magnetic survey off the west coast of the United States between 40°N latitude and 52°N latitude. *Bull Geol Soc Am* 72:1267–1270
- Vine FJ (1966) Spreading of the ocean floor: new evidence. *Science* 154:1405–1415
- Vine FJ (1968) Magnetic anomalies associated with mid-ocean ridges. In: Phinney RA (ed) *The history of the Earth's crust*. Princeton University Press, Princeton, pp 73–89
- Vine FJ, Matthews DH (1963) Magnetic anomalies over oceanic ridges. *Nature* 199:947–949
- Vine FJ, Wilson JT (1965) Magnetic anomalies over a young oceanic ridge off Vancouver Island. *Science* 150:485–489
- Wegener A (1912) Die entstehung der kontinente. *Geol Rundsch* 3:276–292
- Wilson JT (1965) A new class of faults and their bearing on continental drift. *Nature* 207:343–347

Sedimentary Basins

Magdalena Scheck-Wenderoth

Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany

RWTH Aachen University, Aachen, Germany

Definition

Sedimentary basins are regions of prolonged subsidence of the Earth's surface that provide the accommodation space for mineral and organic material (Allen and Allen 2013). These deposits – the sedimentary rocks – are the record of the past geological history including tectonic events, climatic conditions, changes in sea level, and other environmental modifications. In addition, sedimentary basins are long-lived, low-temperature geo-reactors in which the accumulated material experiences a variety of transformations (Bjorlykke 2010; Littke et al. 2008; Roure et al. 2009; Welte et al. 1997). As a result of these processes, basins contain our resources of fossil

fuels, groundwater, and inorganic commodities. Moreover, they are important reservoirs of heat and provide repositories for different socioeconomically relevant fluids such as CO₂, H₂, and CH₄.

Basin Types

Basins can be classified in terms of their plate-tectonic setting. The plate-tectonic Wilson cycle (Fig. 1) describes that the relative movements of plates on the sphere of the Earth result in a closed loop of *continental rifting* and breakup, ocean basin development, closure of oceans at subduction zones, and, in consequence of plate convergence, continental growth.

Basins thus can form as continental rifts that may either evolve into intracontinental basins or lead to continental breakup and ocean basin formation. New *oceanic lithosphere* (see entry ► “[Lithosphere, Oceanic](#)”) is produced along the mid-ocean ridges due to oceanic spreading. If oceanic lithosphere cools, it becomes denser and subsides in the oceanic basins that are bordered by passive continental margins. Of the latter “hot” or volcanic passive margins are distinguished from “cold” or nonvolcanic passive margins (White et al. 2003). If oceanic lithosphere cools beyond a specific threshold, it becomes too heavy to be sustained by the less dense asthenosphere and will finally descend back to the mantle at subduction zones. In these convergent settings, the descending plate is flexed downward and deep oceanic trenches develop above the down-going plate. Subduction may culminate in continent–continent collision if the oceanic lithosphere is entirely subducted and collisional fold and

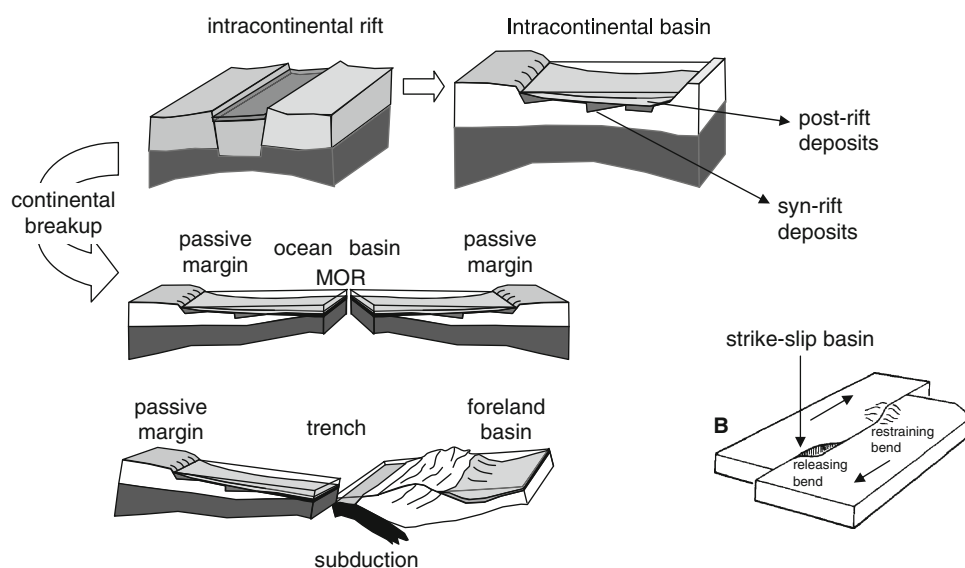
thrust belts form. Loading by collision-related fold and thrust belts also causes a downward flexure of the lithosphere and foreland basins to form. The Wilson cycle may stop at any evolutionary step, because the causative forces cease to be effective. Accordingly, a continental rift not necessarily develops into an ocean basin, but may survive for hundreds of millions of years as an intracontinental basin (Heine et al. 2008; Littke et al. 2008; Cacace and Scheck-Wenderoth 2016). Finally, horizontal movements along strike-slip faults may also cause local extension and related pull-apart basins (Allen and Allen 2013; Gürbüz 2014; Petrunin and Sobolev 2008; Smit et al. 2010; Weber and DESERT Working Group 2009).

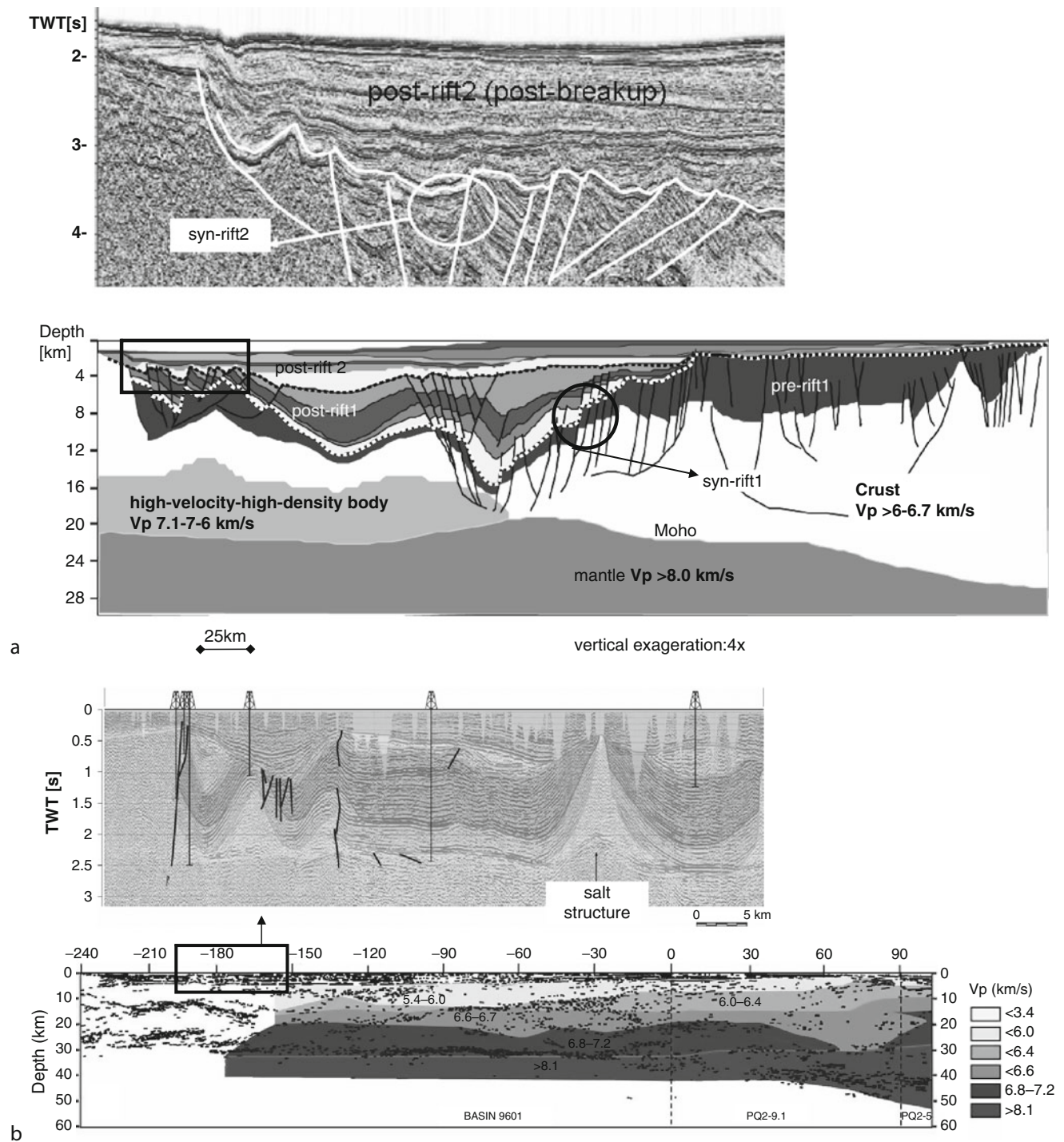
These different types of basins not only have a specific subsidence history (Xie and Heller 2009) but also a characteristic structure of the sediment fill as well as of the underlying crust and mantle lithosphere. This concerns the geometric configuration, the distribution of physical properties, and the resulting *isostatic* (see entry ► “[Isostasy](#)”) and thermal configuration. To assess the configuration of a basin, a wide spectrum of methods has to be integrated. Observations obtained from field measurements, from *deep seismic imaging* (see entry ► “[Deep Seismic Reflection and Refraction Profiling](#)”), and from wells drilled into the basin fill, potential field data, and *heat flow* (see entry ► “[Heat Flow Determinations, Continental](#)”) data, and new data sets from remote sensing need to be integrated with numerical models that simulate processes in basins at different scales.

Figure 2 shows exemplary crustal sections across the Norwegian passive margin and across the intracontinental Central European Basin System.

Sedimentary Basins,

Fig. 1 Basins in their plate-tectonic setting





Sedimentary Basins, Fig. 2 (a) *Lower panel:* Line drawing of a crustal-scale integrated cross section with three generations of basins from the Norwegian passive margin modified after Faleide et al. 2008. *Upper panel:* Seismic close-up showing the breakup unconformity and pre-breakup extensional faulting during phase syn-rift 2. (b) *Lower panel:* Line drawing with superimposed seismic p-wave velocities

from the intracontinental North German Basin modified after Bayer et al. 2002. *Upper panel:* Close-up with reflection seismic example illustrating polyphase subsidence with several sequences separated by unconformities and additional salt-tectonic deformation. (Modified after Mazur et al. 2005)

Basin-Forming Mechanisms

There are three main mechanisms leading to the formation of sedimentary basins (Allen and Allen 2005, 2013):

1. Extensional stresses causing the lithosphere to thin
2. Flexure of the lithosphere in response to vertical loading or compressive stresses
3. Viscous flow of the mantle causing positive or negative topography

All these mechanisms may act separately or in concert and may affect the lithosphere over a range of spatial and temporal scales.

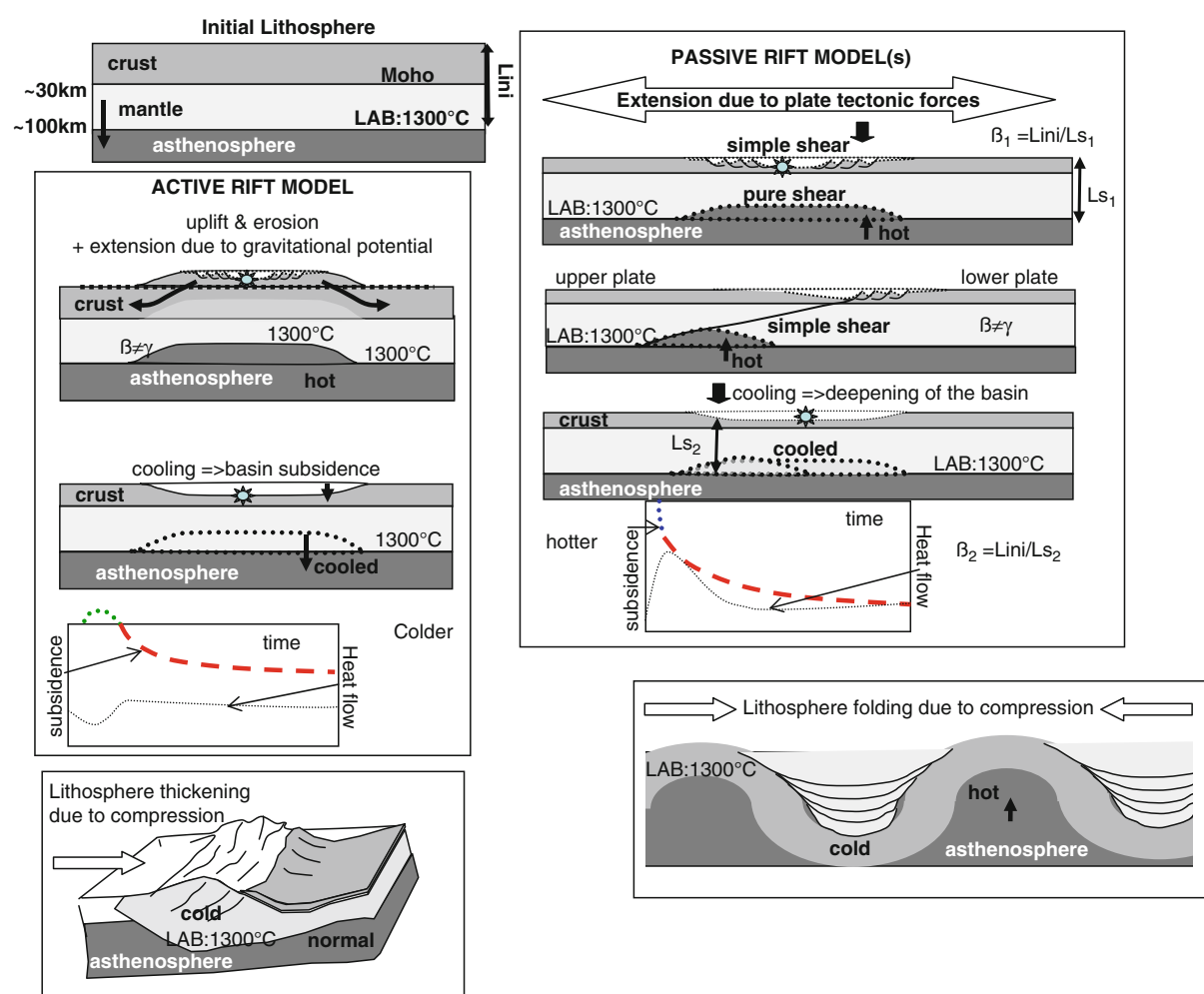
Conceptual Models of Basin Formation

Conceptual models of basin formation (Fig. 3) attempt to integrate geophysical and geological observations describing

the geometry and the physical state of a basin into evolutionary schemes that explain basin evolution. These conceptual models also need to be consistent with data on the rheological behavior of rocks known from lab experiments as well as with data on the thermal and deformation history. All these conceptual models represent end-member scenarios and explain a large fraction of observed phenomena in specific basins.

In the concept of *passive rifting* (Fig. 3), the lithosphere is stretched in response to extensional tectonic stress. This may result in pure shear thinning (McKenzie 1978), in simple shear thinning (Wernicke 1981), or in a mixture of the two (Lister et al. 1986).

The *uniform stretching model* (McKenzie 1978) predicts a crustal and lithosphere thickness proportional to the stretching factor β where β is the ratio between initial lithosphere thickness and the stretched lithosphere thickness. Lithosphere thinning results in a passive rise of the isotherm at the thermal lithosphere–asthenosphere boundary LAB (Turcotte and Schubert 2002). This, in turn, causes additional thermal subsidence after the stress ceases to be effective, and cooling



Sedimentary Basins, Fig. 3 Conceptual models of basin formation

of the previously heated lithosphere causes a density increase. In this model, uniform and symmetric lithospheric thinning by β is accommodated by brittle faulting in the upper crust as well as by ductile flow in the lower crust and upper mantle. The uniform stretching model predicts a two-stage subsidence: (1) as a result of tectonic stretching, on a short timescale (10–20 my), and (2) as a result of thermal subsidence – on a longer timescale (50–100 my). Structurally, the uniform stretching model predicts syn-rift extensional faults in the upper crust with syn-rift sediments overlain by onlapping post-rift sediments of the thermal subsidence phase. The thermal history predicted by the uniform stretching model includes an initial heat peak due to the rise of the isotherm at the LAB to a shallower position (and thus an increase of the thermal gradient within the lithosphere) followed by continuous cooling during thermal re-equilibration.

Depth-dependent stretching (Sclater and Christie 1980) is a modification of the uniform stretching model and accounts for different stretching magnitudes in the upper brittle part of the crust by a factor β and in the deeper lithosphere by a factor γ . In the *asymmetric stretching models* (Lister et al. 1986; Wernicke 1981), crust- to lithosphere-scale listric faults accommodate large offsets in response to extensional stress and finally may lead to continental breakup with an “upper plate” margin developing from the hanging wall of the master detachment fault and a “lower plate” margin in the foot wall (Osmundsen and Ebbing 2008; Sibuet et al. 2007; Zuber et al. 1986). In particular, magma-poor margins like the Galicia-Newfoundland (Sibuet et al. 2007) are commonly explained by this mechanism.

In the concept of *active rifting*, basin initiation is caused by the impingement of a rising hot asthenospheric *mantle plume* (Allen and Allen 2005; Burov and Cloetingh 2009; Haxby et al. 1976; Turcotte and Schubert 2002). The concept of active rifting is in fact a special case of basins initiated by *dynamic topography*. The emplacement of the *mantle plume* (see entry ► “Mantle Plumes”) causes uplift and induces extensional stresses at the flanks of the uplifted area due to gravitational potential. The active rise of the isotherm may lead to surface uplift above the erosional level, and the eroded material may be removed from the system. Density increase in the lithosphere due to subsequent cooling causes surface subsidence and the formation of a sedimentary basin in the previously uplifted area. As the process is transient, the system would re-equilibrate after the thermal anomaly and the related dynamic topography stop to exist. For *isostatic* reasons, net subsidence will only occur if either igneous underplating of the crust has taken place (light crust replaced by dense underplate) or if material is removed from the system by erosion during uplift. Systematic work assessing the contribution of the three processes to subsidence (extension due to gravity potential, crustal underplating, and erosion-related mass loss) is still missing. The active rift

model predicts weak to missing syn-rift faulting, a syn-rift erosional unconformity, and a subsidence history corresponding predominantly to thermal subsidence. Actually, the active rift model predicts a subsidence history identical to a specific end-member case of (passive) depth-dependent stretching in which the mantle lithosphere is severely thinned, whereas the crust is not. However, the active rift model predicts a colder heat flow scenario than the passive rift model, as the thermal gradient is not disturbed by an initial thinning of the lithosphere.

According to the concept of dynamic topography, basins can also be initiated in response to downward flow of high-density material in the convective mantle, a process causing *negative dynamic topography*. This mechanism could be related to detached subduction slabs sinking into the asthenospheric mantle and is discussed as an explanation for the subsidence of intracontinental basins that subside over several hundreds of millions of years at a very slow rate (Allen and Allen 2005; Heine et al. 2008; Ritzmann and Faleide 2009).

Two main conceptual models exist for *flexural basins* that form as response of the flexural rigidity of the lithosphere (Watts 2001) to external loads: (1) downward flexure of the lithosphere due to vertical loading (Zhang and Bott 2000) and (2) formation of crustal- to lithosphere-scale folds in response to horizontal compressive stress, a process also called *lithosphere buckling* (Burov and Cloetingh 2009; Burov and Diament 1992; Cloetingh and Burov 2010). Both concepts predict a lack of syn-rift extensional faulting, a cold thermal scenario, and subsidence lasting only as long as the stress is effective (either vertical as load or horizontal as tectonic stress). Due to structural similarity, lithosphere buckling in response to far-field compressive tectonic stress (Burov and Cloetingh 2009; Burov and Diament 1992) can be mistaken as an expression of active rifting if the analyzed region is too small to capture the full wavelength of the buckling. Again, the different predicted subsidence and thermal histories can help to distinguish the two.

Vertical loading by sediments enhances subsidence in basins of all geodynamic settings (Watts 2001). It has recently been proposed that mineral phase transitions may also play a role for subsidence history in different tectonic settings (Kaus et al. 2005).

Geodynamic modelling studies (Braun and van der Beek 2004; Brune 2018; Buck 1991; Buitier 2014; Cacace and Scheck-Wenderoth 2016; Cloetingh et al. 1998; Huismans et al. 2005; Huismans and Beaumont 2008; Kusznir and Ziegler 1992; Lavier et al. 1999; Le Pourhiet et al. 2018; Pascal and Cloetingh 2002; Pérez-Gussinyé et al. 2006; Royden and Keen 1980; Sclater and Christie 1980; Simon et al. 2009; van Wijk 2005; Wijns et al. 2005; Zuber et al. 1986) indicate that the nature of rheological coupling between the crust and the lithospheric mantle, local strain weakening and strengthening, and the rate of extension are the dominant factors controlling not only the subsidence evolution of basins

but also if they develop in a symmetric or asymmetric style and in a narrow or wide rift mode. Recent progress toward 3D thermodynamic models illustrates that some concepts for basin formation and evolution need to be reconsidered (Burov and Gerya 2014; Le Pourhiet et al. 2018).

The Configuration of the Sediment Fill

The sediment fill can vary in thickness between a few to up to 20 km and is analyzed using stratigraphic methods (Allen and Allen 2005; Catuneanu 2006; Einsele et al. 1991; Vail et al. 1991) to assess the nature of sediments (sand, silt, shale, carbonates, evaporites), the depositional environment (continental or marine), and the climatic conditions prevailing during the time of deposition. Subsidence and deposition may alternate with nondeposition due to sea level fall or tectonic surface uplift and lead to stratigraphic gaps separating different sequences (Fig. 2). In addition, the rate of subsidence and sedimentation as well as of uplift and erosion may change with time. Stratigraphic gaps and changes in depositional conditions are expressed as unconformities separating sedimentary sequences, the analysis of which is required to understand the history of a basin.

Progressive sedimentation and subsidence goes along with changes in physical properties of the deposited sediments. The two most prominent changes are (1) an increase in temperature and (2) a decrease in porosity with increasing burial depth. The initially high porosity of each layer decreases as the load of the covering deposits increases and fluids filling the initial pores are expelled. Also, chemical compaction may lead to a loss in porosity. With decreasing porosity, physical properties like density, seismic velocity, or thermal conductivity increase.

Stratigraphic analysis is performed in field studies if the fill of a basin is accessible in outcrops due to post-depositional uplift. Most of the sedimentary basins, however, extend by their very nature (areas of enduring subsidence) into the subsurface and to large depths. Accordingly, subsurface analytical methods are required to assess their configuration. The most important of these methods is *seismic imaging*. *Seismic reflections* (see entry ► [“Deep Seismic Reflection and Refraction Profiling”](#)) originate from the contrast in impedance at discontinuities in the subsurface that can represent stratigraphic boundaries and unconformities or changes in attributes of a specific layer. The reflection seismic technique provides a picture of these interfaces in a two-way travel time that is constrained by wells. Wells provide information on the absolute depth position of stratigraphic horizons and enable the conversion of seismic travel times to depth. Furthermore, wells yield data on the thermal gradient with depth as well as on changes in lithology and petrophysical properties. Properties such as

composition, compaction, porosity, permeability, density, thermal and hydraulic conductivity, *radiogenic heat production* (see entry ► [“Radiogenic Heat Production of Rocks”](#)), and content and maturation of organic matter are measured using ► [“Geophysical Well Logging”](#) methods.

Apart from recording changes in relative sea level or climatic conditions, the configuration of the sediment fill may be influenced by deformation (Kley et al. 2008). Accordingly, sediments can be faulted, folded in response to tectonic forces, or deformed due to halokinetic movements (Hudec and Jackson 2007; Scheck-Wenderoth et al. 2008; Vendeville and Jackson 1992). This deformation can take place pre-, syn-, or postdepositionally, each of which is expressed by characteristic structural styles.

In particular, basin initiation in extensional plate-tectonic conditions is typically associated with normal faulting in the upper crust leading to a thinning of the lithosphere and to the accumulation of syn-rift deposits. After the tectonic stresses have declined, basins experience a phase of post-rift thermal subsidence and may even experience late basin inversion in response to compressive stress (Cooper and Williams 1989; Mazur et al. 2005; Sandiford 1999) (Fig. 2).

Finally, there is a feedback between surface processes and deep mechanisms in the sense that *isostatic* loading by sediments will increase total subsidence and erosion of basin flanks will cause flexural rebound (Braun 2005; Braun et al. 2008; Braun and van der Beek 2004; Garcia-Castellanos 2002; Kounov et al. 2007; Simoes et al. 2010; Van der Beek 2007; Willenbring and von Blanckenburg 2010).

The Configuration of the Crust Beneath Sedimentary Basins

The different mechanisms responsible for the thinning of the lithosphere leave characteristic traces that can be detected with geophysical observations, which, however, are not always leading to unique interpretations. The most prominent attribute of the crust beneath sedimentary basins is that it is thinned compared to neighboring regions. Deep seismic data imaging of the preserved thickness of the crust helps to determine the amount of the stretching factor β as the ratio between the initial crustal thickness as often preserved outside the basin margins and the crustal thickness beneath the basin. Crustal thinning can vary considerably across a basin with values of β largest in the area of strongest thinning. The average initial thickness of the crust may reach up to 50 km in cratonic areas, around 30 km in Phanerozoic continental domains, and less than 10 km below oceans.

As the crystalline crust is characterized by a higher density and higher velocities of seismic waves than the sediments, the interface between the two can be detected. Accordingly,

the *interpretation of gravity anomalies* (see entry ► “[Gravity Anomalies, Interpretation](#)”) and the analysis of the observed variation of seismic velocities are key to evaluate the crustal structure. Refraction seismic techniques are used to detect variations in seismic p-wave velocities at depth, whereas reflection seismic methods are useful to evaluate the changes in reflectivity of the deeper crust.

Seismic p-wave velocities (V_p) increase to values >6 km/s in the upper crystalline crust and can range up to 7.4 km/s (Christensen and Mooney 1995) depending on the composition of the crust. There is a negative correlation between the quartz content of crystalline rocks and seismic p-wave velocity as well as density. Silicic (quartz-rich) rocks are characterized by smaller velocities and lower densities than mafic rocks (Table 1). Seismic shear wave velocities are also higher for crystalline than for sedimentary rocks but are also dependent of additional parameters (Christensen and Mooney 1995). Accordingly, the V_p/V_s ratio is an additional parameter helping to interpret crustal structure (Afonso et al. 2010; Mjelde et al. 2003). As also the magnetic properties are different for mafic (Fe-rich) and silicic (Fe-poor) rocks, the *interpretation of magnetic anomalies* is an additional technique for the evaluation of the crustal structure beneath basins.

At the crust–mantle boundary (Moho), often a characteristic reflection and a sudden increase of seismic p-wave

velocity to values >8 km/s are observed in deep seismic data marking the transition to the lithospheric mantle (Bauer et al. 2000; DEKORP-BASIN Research Group 1999; Fowler 1996; Hirsch et al. 2009; Levander et al. 2006; Meissner and DEKORP Research Group 1991; Thybo and Nielsen 2009; Turcotte and Schubert 2002; Weber and DESERT Working Group 2009). This is related to the change in composition from various crustal rocks to the mantle consisting mainly of peridotite.

As less dense sediments replace a denser crystalline crust, sedimentary basins should be characterized by a negative Bouguer anomaly with highest amplitude in the basin center after re-equilibration of the isotherm. Many basins, however, are characterized by a long-wavelength negative Bouguer anomaly with positive anomalies superimposed. This observation, together with strong variations in the crustal velocity structure beneath sedimentary basins, challenges the classical concepts of crustal thinning and related Moho uplift in that they commonly occur in concert with a flat Moho (Thybo and Nielsen 2009).

Crustal bodies with seismic velocities >7 km/s, generally referred to as “high-velocity bodies,” are observed in the lower crust beneath many intracontinental rift structures, for example, the Baikal and the Kenya Rifts (Thybo and Nielsen 2009); beneath intracontinental basins, for example, in the

Sedimentary Basins, Table 1 Overview on physical properties of rocks relevant to sedimentary basins: *seismic properties* (After Christensen and Mooney 1995), *thermal properties*. (After Cermak and

Rybach 1982; Fernández et al. 2005; Förster and Förster 2000; Ritter et al. 2004; Scheck-Wenderoth and Maystrenko 2008)

	Dominant lithology	Thermal conductivity λ [W/mK]	Heat production S [mW/m ³]	Density [kg/m ³]	p-wave velocity [km/s]
	Sea water	0.563	0	1,030	1.48
S	Uncompacted fine-grained siliciclastics	1.2 (2.0)	1	1,950	2.05
E	Slightly compacted fine-grained siliciclastics	1.8 (2.2)	1	2,200	2.2
D	Medium compacted fine-grained siliciclastics	2.0 (2.2)	1	2,250	2.2
I	Compacted fine-grained siliciclastics	2.1 (2.3)	1	2,550	3.5
M	Strongly compacted fine-grained siliciclastics	2.1 (2.3)	1	2,630	4.3
E	Compacted siliciclastics with carbonates	2.1 (2.5)	0.9	2,640	5.5–6.0
N	Carbonates: chalk	1.7	0.6	2,000	3.4
T	Carbonates: limestones, dolomites	2.4	0.6	2,600	6.0
S	Rock salt	3.5 (5.4)	0.1	2,150	4.0–5.5
Crust	Granites and gneisses	2.7 (3.2)	0.8	2,790	6.0–6.7
Continent	Mafic granulites/gabbros/eclogites	2.6 (3.0)	0.3	3,120	6.8–7.0
	Gabbros	2.6 (3.0)	0.5	3,150	7.1–7.6
	Basalt	1.8 (2.1)	0.4	2,580	4.0–5.0
Crust	Sheeted dikes/gabbroic intrusions	2.3 (2.6)	0.2	2,890	5.0–6.7
Ocean	Gabbro	2.3 (3.0)	0.2	3,150	7.1–7.6
Mantle	Peridotite continent	3.95 (3.2)	0.03	3,330	8.0–8.3
	Peridotite ocean	3.95 (3.2)	0.03	3,180	7.8

Danish Basin (Nielsen and Thybo 2006); and beneath extended passive margins, for example, at the conjugate margins of the North and South Atlantic (Bauer et al. 2000; Contrucci et al. 2004; Faleide et al. 2008; Franke et al. 2007; Hirsch et al. 2009; Mjelde et al. 2002, 2005; Sibuet et al. 2007). Gravity analysis indicates that these bodies are also characterized by higher densities than the average crust (Fernández et al. 2005, 2010; Franke et al. 2007; Hirsch et al. 2009; Maystrenko and Scheck-Wenderoth 2009; Osmundsen and Ebbing 2008).

Differential thinning with depth (Huisman et al. 2005; Huisman and Beaumont 2008; Kuszniir and Ziegler 1992; Lavier et al. 1999; Lavier and Steckler 1997; Pascal and Cloetingh 2002; Sclater and Christie 1980; Steckler and Watts 1978), emplacement of magmatic material in the crust during the thinning process, and magmatic underplating (Thybo and Nielsen 2009; Burg and Gerya 2008) are some possible processes during basin formation and evolution that can be responsible for the actual configuration of the crust beneath a specific basin. The relative contributions of the processes vary strongly for different basins, and, accordingly, their relative importance in general terms is still under debate.

Moreover, structural inheritance may have a fundamental role in defining the crustal and lithosphere structure. Successive suturing of different plates in Earth history may result in a mosaic of crustal domains with contrasting physical properties, possibly also including lower crustal bodies predating the rifting process (Ebbing et al. 2009; Faleide et al. 2008; Plomerová and Babuska 2010; Scheck-Wenderoth and Lamarche 2005; van Wijk 2005; Vauchez et al. 1998; Sippel et al. 2017). In basins developing on such a substrate, such older rheological discontinuities may be reactivated to localize deformation in areas of reduced strength, thus also facilitating discontinuous stretching with depth.

The Configuration of the Mantle Lithosphere

Much less is known on the configuration of the mantle lithosphere below sedimentary basins due to the limited amount of direct observations. Though reflections and lateral variations in seismic wave velocities in the lithospheric mantle have been observed in deep seismic data, the energy used in active seismic experiments is generally not sufficient to reach depth intervals below the Moho. Also, the gravity signal from the lithospheric mantle is difficult to isolate from the cumulative signal of the entire upper mantle. Recent development in the acquisition and evaluation of passive *seismological* data (Dalton and Faul 2010; Fichtner et al. 2018; Fishwick 2010; Geissler et al. 2010; Heintz and Kennett 2005; Hieronymus and Goes 2010; Plomerová and Babuska 2010; Priestley and

McKenzie 2013; Priestley et al. 2006; Ritzmann and Faleide 2009; Rychert and Shearer 2009; Schaeffer and Lebedev 2013; Zhang and Lay 1996); electromagnetic and *magnetotelluric* methods (see entries ► [“Magnetotelluric Interpretation”](#) and ► [“Magnetotelluric Data Processing”](#)) (Jones et al. 2010), together with remote sensing gravity observation (Kaban et al. 2003; Schotman et al. 2009); as well as geochemical data (O'Reilly and Griffin 2010; Trumbull et al. 2002; Wang 2010; Kovács et al. 2018) and thermal studies (Artemieva 2009; Hasterok and Chapman 2007; Hieronymus and Goes 2010; Schaeffer et al. 2016; Lucazeau 2019) indicates that the lithospheric mantle is less homogenous than previously thought. Apart from strong differences in thickness, also considerable lateral variations in surface *heat flow* (see entry ► [“Heat Flow Determinations, Continental”](#)) are observed, with a positive correlation between the two. The LAB may be located deeper than 250 km in cratonic areas, lies at around 100 km beneath Phanerozoic continental domains and old oceans, and is close to the seafloor at mid-ocean ridges.

Agreement is established that the LAB is a fundamental boundary in plate-tectonic theory that separates the rigid plates from ductile convecting material below the plates (Artemieva 2009; Eaton et al. 2009; Jones et al. 2010). Three broad definitions in terms of a mechanical boundary layer, a thermal boundary layer, and a chemical boundary layer are based on different types of data and partly in geometrical conflict. For the evolution of sedimentary basins, the depth of the thermal LAB is especially relevant as it determines the thermal and mechanical state of the lithosphere subjected to any of the basin-forming mechanisms. Systematic mapping of lithosphere thickness beneath sedimentary basins is, however, still patchy and partly contradictory depending on the method used to derive it. The thermal LAB is interpreted as an isotherm of about 1,300 °C, and its depth corresponds to the depth where an average continental conductive geotherm intersects the mantle adiabat. This is corroborated by heat flow inversion studies in *continental lithosphere* (see entry ► [“Lithosphere, Continental”](#)) (Artemieva 2009), by combined thermal and gravity modeling (Fernández et al. 2005, 2010; Freymark et al. 2017; Hasterok and Chapman 2007; Hyndman et al. 2009; Maystrenko et al. 2013; Przybycin et al. 2015; Sippel et al. 2017; Scheck-Wenderoth and Maystrenko 2008) and by cooling models in *oceanic lithosphere* (see entry ► [“Lithosphere, Oceanic”](#)) (Crosby et al. 2006; McKenzie 1978; Sclater 2003; Stein and Stein 1992).

Isostatically, thinning of the lithospheric mantle results in net surface uplift as heavier lithospheric mantle is replaced by less dense asthenospheric material (active rift model). Only if the interaction with surface processes takes place (erosion), net subsidence takes place due to subsequent cooling of the asthenospheric material.

Heat Flow in Sedimentary Basins

The thermal field in sedimentary basins has been of primary interest in the exploration of fossil fuels (Welte et al. 1997) and gains increasing importance for the use of geothermal energy (Huenges 2010). Moreover, together with composition, temperature is a main controlling factor for the rheological behavior of the lithosphere and accordingly its deformation (Burov 2011; Violay et al. 2015a, b). The hotter the lithosphere, the weaker is its rheology and the easier it is thinned to form a sedimentary basin. Accordingly, the main controlling factors for the thermal state of a basin are its plate-tectonic setting and its evolutionary stage.

In terms of observables, surface *heat flow* measurements and temperatures measured in wells characterize the present-day thermal state of a basin, whereas the maturity of organic matter (Welte et al. 1997) and thermochronological data (Andriessen 1995; Braun 2005; Braun and van der Beek 2004; Kounov et al. 2007; Simoes et al. 2010; Van der Beek 2007; Willenbring and von Blanckenburg 2010) provide a record of the thermal history.

Different families of integrated process-oriented models attempt to reproduce these observables and indicate that several processes contribute to the heat flow in sedimentary basins. There is a first-order contrast in thermal properties between sedimentary and crystalline rocks, in that sediments are thermally less conductive and produce less radiogenic heat than crystalline crustal rocks (Table 1). The amount of heat entering the basin at the base of the sediment fill depends on the thickness and composition of the crystalline crust as well as on the depth of the thermal LAB. The mafic components of the crust and the lithospheric mantle are characterized by a high thermal conductivity but low *radiogenic heat production*. Accordingly, the shallower the LAB and the thicker and radiogenic the crystalline crust, the more heat arrives at the base of the sediments. Due to their higher porosities, the thermally low-conductive sediments act as a thermal blanket, causing heat storage in the basin (Cacace et al. 2010; Theissen and Rüpke 2009; Van Wees et al. 2009). In addition, the sediments contribute a modest but, in the sum, considerable amount of radiogenic heat to the system. Also within the sediment fill, the thermal properties may vary (Table 1) with the thermal conductivity of salt being two times larger than that of clastic sediments. The upper part of sedimentary basins may store paleoclimatic signals of previous glaciations as present-day permafrost down to more than 700 m depth attest (Majorowicz and Šafanda 2018; Szweczyk and Nawrocki 2011).

In response to the distribution of thermal properties, long-wavelength variations in temperatures in sedimentary basins (scale of hundreds of kilometers) are determined by the crustal structure and composition as well as by the thermal thickness of the lithosphere (Cacace et al. 2010; Hasterok and

Chapman 2007; Hyndman et al. 2009; Scheck-Wenderoth and Maystrenko 2008, 2013; Sclater 2003). In contrast, the short-wavelength pattern of temperature distribution (scale of kilometers and less) is controlled by the internal configuration of the sediment fill.

Heat transfer by conduction is assumed to dominate the transport of heat in the lithosphere in contrast to the *convecting mantle* (see entry ► “Mantle Convection”). Geological and geochemical data together with models of coupled fluid and heat transport (Scheck-Wenderoth et al. 2014) show that the thermal field in sedimentary basins is additionally influenced by heat transport due to moving pore fluids. Advective (fluid pressure-driven) heat transport can influence the thermal field of the sedimentary fill on the regional scale. In contrast, buoyancy-driven free convective heat transport and heat transport along faults is only locally important but needs consideration for exploration on the reservoir scale.

Summarizing, the heat flow in a specific basin may vary spatially, though some ranges can be given for different types of basins (Allen and Allen 2005, 2013). Typically, the surface *heat flow* in intracontinental basins varies between 40 and 70 mW/m², can be up to 150 mW/m² in active rifts or close to volcanic arcs, and can reach values higher than 180 mW/m² at mid-ocean ridges and oceanic rifts. In oceanic basins, surface *heat flow* decreases with increasing distance from the mid-ocean ridges according to the cooling of the lithosphere with age (Parsons and Sclater 1977; Stein and Stein 1992). While magma-poor passive margins are characterized by a rather low surface heat flow, the range of surface *heat flow* at magma-rich passive margins depends on the age of the adjacent oceanic lithosphere (Gholamrezaie et al. 2018). At young margins as in the 55-my-old North Atlantic, the surface heat flow is still controlled by the cooling of the oceanic lithosphere that is considerably thinner than the lithosphere of the continental margin. This step in the thermal LAB is consistent with observed heat flow values increasing from 45 mW/m² at the continental side to 80 mW/m² at the oceanic side of the margin (Ritter et al. 2004; Scheck-Wenderoth and Maystrenko 2008). At older passive margins as in the 130-my-old South Atlantic, an opposite trend is observed (Goutorbe and Bonneville 2008; Maystrenko et al. 2013) with about 45 mW/m² at the oceanic side of the margin to up 65 mW/m² at the continental side. As the oceanic lithosphere had sufficient time to cool and thicken, the depth of the thermal LAB is continuous and similar between the continental and the oceanic part of the system. Instead, the radiogenic heat contribution of the *continental crust* (see entry ► “Earth’s Structure, Core”) thickening toward the continent becomes the controlling factor for heat input from the lithosphere. Flexural basins are generally colder than extensional basins with average *heat flow* values around 40 mW/m² in oceanic trenches and foreland basins. The thermal signature of strike-slip basins is highly variable as it depends on the

attributes of the lithosphere on which they develop and if hydrothermal heat transport takes place along the fault zones.

Summary

The *geodynamics* of sedimentary basins results from the interaction of a variety of processes acting on different spatial and temporal scales. Geophysical and geological observations indicate that intra-plate stress and heat input from the asthenospheric mantle are first-order controlling factors. On the *continental lithosphere* (see entry ► “[Lithosphere, Continental](#)”), structural inheritance and its impact on rheology is a third major player controlling the development of sedimentary basins. Intra-plate stress determines if an extensional or a compressional basin is formed. Temperature dominantly controls the rheological behavior of the lithosphere in addition to composition. These two parameters determine if the lithosphere is thinned uniformly, depth dependently, and in a symmetric or asymmetric mode. The more layered the lithosphere rheology is, the more discontinuous and asymmetric the stretching process will be. In contrast, a lithosphere deforms in a uniform stretching mode if the crust is strongly coupled to the mantle and no significant vertical or horizontal rheology contrasts are present. In both cases, the magnitude and rates of effective stress as well as the magnitude of the heat anomaly determine if stretching takes place slow enough to allow for cooling related strain hardening or fast enough to result in continental breakup.

These deep factors interact continuously with surface processes such as deposition or erosion of sediments influenced by climatic conditions. Deposition leads to *isostatic* loading and enhanced subsidence, whereas erosion results in isostatic unloading and enhanced uplift.

Cross-References

- [Continental Crustal Structure](#)
- [Continental Rifts](#)
- [Deep Seismic Reflection and Refraction Profiling](#)
- [Geodynamics](#)
- [Geophysical Well Logging](#)
- [Gravity Anomalies, Interpretation](#)
- [Heat Flow, Continental](#)
- [Heat Flow, Seafloor: Methods and Observations](#)
- [Isostasy](#)
- [Lithosphere, Continental](#)
- [Lithosphere, Mechanical Properties](#)
- [Lithosphere, Oceanic](#)
- [Lithosphere, Oceanic: Thermal Structure](#)
- [Magnetic Anomalies: Interpretation](#)
- [Magnetotelluric Interpretation](#)

- [Mantle Convection](#)
- [Mantle Plumes](#)
- [Radiogenic Heat Production of Rocks](#)
- [Seafloor Spreading](#)
- [Seismic Imaging, Overview](#)
- [Seismic Properties of Rocks](#)
- [Seismology, Global Earthquake Model](#)
- [Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat](#)
- [Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity](#)

Bibliography

- Afonso JC et al (2010) On the Vp/Vs-Mg# correlation in mantle peridotites: implications for the identification of thermal and compositional anomalies in the upper mantle. *Earth Planet Sci Lett* 289(3–4): 606–618
- Allen P, Allen J (2005) *Basin analysis. Principles and applications*. Blackwell Science, Oxford, p 560
- Allen PA, Allen JR (2013) *Basin analysis: principles and application to petroleum play assessment*, 3rd edn. Wiley-Blackwell Science, Oxford
- Andriessen PAM (1995) Fission-track analysis: principles, methodology and implications for tectono-thermal histories of sedimentary basins, orogenic belts, and continental margins. *Geol Mijnb* 74:1–12
- Artemieva IM (2009) The continental lithosphere: reconciling thermal, seismic, and petrologic data. *Lithos* 109:23–46
- Bauer K et al (2000) Deep structure of the Namibia continental margin as derived from integrated geophysical studies. *J Geophys Res* 105(B11):25,829–25,853
- Bayer U, Grad M, Pharaoh TC, Thybo H, Guterch A, Banka D, Lamarche J, Lassen A, Lewerenz B, Scheck M, Marotta AM (2002) The southern margin of the East European Craton: new results from seismic sounding and potential fields between the North Sea and Poland. *Tectonophysics* 360:301–314
- Bjorlykke K (2010) *Heat transport in sedimentary basins, petroleum geoscience*. Springer, Berlin, pp 253–257
- Braun J (2005) Quantitative constraints on the rate of landform evolution derived from low-temperature thermochronology, low-temperature thermochronology: techniques, interpretations, and applications. *Rev Mineral Geochem* 58:351–374
- Braun J, van der Beek P (2004) Evolution of passive margin escarpments: what can we learn from low-temperature thermochronology? *J Geophys Res Earth Surf* 109:F04009
- Braun J et al (2008) DOUAR: a new three-dimensional creeping flow numerical model for the solution of geological problems. *Phys Earth Planet Inter* 171(1–4):76–91
- Brune S (2018) Forces within continental and oceanic rifts: numerical modeling elucidates the impact of asthenospheric flow on surface stress. *Geology* 46(2):191–192
- Buck WR (1991) Modes of continental lithospheric extension. *J Geophys Res* 96(B12):20,161–20,178
- Buiter S (2014) How plumes help to break plates. *Nature (News & Views)* 513:36–37
- Burg JP, Gerya TV (2008) Modelling intrusion of mafic and ultramafic magma into the continental crust: numerical methodology and results. *Boll Soc Geol Ital* 127(2):199–203
- Burov EB (2011) Rheology and strength of the lithosphere. *Mar Pet Geol* 28:1402–1443
- Burov E, Gerya T (2014) Asymmetric three-dimensional topography over mantle plumes. *Nature* 513. <https://doi.org/10.1038/nature13703>

- Burov E, Cloetingh S (2009) Controls of mantle plumes and lithospheric folding on modes of intraplate continental tectonics: differences and similarities. *Geophys J Int* 178(3):1691–1722
- Burov EB, Diament M (1992) Flexure of the continental lithosphere with multilayered rheology. *Geophys J Int* 109:449–468
- Cacace M, Scheck-Wenderoth M (2016) Why intracontinental basins subside longer: 3-D feedback effects of lithospheric cooling and sedimentation on the flexural strength of the lithosphere. *J Geophys Res* 121(5):3742–3761
- Cacace M, Kaiser BO, Lewerenz B, Scheck-Wenderoth M (2010) Geothermal energy in sedimentary basins: what we can learn from regional numerical models. *Chem Erde-Geochem* 70:33–46
- Catuneanu O (2006) Principles of sequence stratigraphy. Elsevier, New York, p 386
- Cermak V, Rybach L (1982) Thermal properties: thermal conductivity and specific heat of minerals and rocks. In: Angenheister G (ed) Landolt-Börnstein, new series, geophysics and space research. Springer, Berlin, pp 305–343
- Christensen NI, Mooney WD (1995) Seismic velocity structure and composition of the continental crust: a global view. *J Geophys Res* 100(B7):9761–9788
- Cloetingh S, Burov E (2010) Lithospheric folding and sedimentary basin evolution: a review and analysis of formation mechanisms. *Basin Res.* <https://doi.org/10.1111/j.1365-2117.2010.00490.x>
- Cloetingh S, Boldreel LO, Larsen BT, Heinesen M, Mortensen L (1998) Tectonics of sedimentary basin formation: models and constraints. *Tectonophysics* 300(I–IV):1–11
- Contrucci I et al (2004) Deep structure of the West African continental margin (Congo, Zaire, Angola), between 5 degrees S and 8 degrees S, from reflection/refraction seismics and gravity data. *Geophys J Int* 158(2):529–553
- Cooper MA, Williams GD (1989) Inversion tectonics. Geological society special publication classics. The Geological Society of London, London
- Crosby AG, McKenzie D, Sclater JG (2006) The relationship between depth, age and gravity in the oceans. *Geophys J Int* 166(2):553–573
- Dalton CA, Faul UH (2010) The oceanic and cratonic upper mantle: clues from joint interpretation of global velocity and attenuation models. *Lithos* 120(1–2):160–172
- DEKORP-BASIN Research Group (1999) Deep crustal structure of the Northeast German basin: new DEKORP-BASIN'96 deep-profiling results. *Geology* 27(1):55–58
- Eaton DW et al (2009) The elusive lithosphere-asthenosphere boundary (LAB) beneath cratons. *Lithos* 109(1–2):1–22
- Ebbing J, Gernigon L, Pascal C, Olesen O, Osmundsen PT (2009) A discussion of structural and thermal control of magnetic anomalies on the mid-Norwegian margin. *Geophys Prospect* 57(4):665–681
- Einsele G, Ricken W, Seilacher A (1991) Cycles and events in stratigraphy – basic concepts and terms. In: Einsele G, Ricken W, Seilacher A (eds) Cycles and events in stratigraphy. Springer, Berlin, pp 1–19
- Faleide JJ et al (2008) Structure and evolution of the continental margin off Norway and the Barents Sea. *Episodes* 31(1):82–91
- Fernández M et al (2005) Lithospheric structure of the mid-Norwegian margin; comparison between the More and Voring margins. *J Geol Soc Lond* 162(6):1005–1012
- Fernández M, Afonso JC, Ranalli G (2010) The deep lithospheric structure of the Namibian volcanic margin. *Tectonophysics* 481(1–4):68–81
- Fichtner A, van Herwaarden DP, Afanasiev M, Simute S, Krischer L, Cubuk-Sabuncu Y, Taymaz T, Colli, L, Saygin E, Villasenor A, Trampert J, Cupillard P, Bunge H-P, Igel H (2018) The collaborative seismic earth model: generation I. *Geophys Res Lett* 45. <https://doi.org/10.1029/2018GL077338>
- Fishwick S (2010) Surface wave tomography: imaging of the lithosphere-asthenosphere boundary beneath central and southern Africa? *Lithos* 120(1–2):63–73
- Förster A, Förster H-J (2000) Crustal composition and mantle heat flow: implications from surface heat flow and radiogenic heat production in the Variscian Erzgebirge. *J Geophys Res* 105(B12):917–938
- Fowler CMR (1996) The solid earth. Cambridge University Press, Cambridge, p 472
- Franke D, Neben S, Ladage S, Schreckenberger B, Hinz K (2007) Margin segmentation and volcano-tectonic architecture along the volcanic margin off Argentina/Uruguay, South Atlantic. *Mar Geol* 244(1–4):46–67
- Freyermark J, Sippel J, Scheck-Wenderoth M, Bär K, Stiller M, Fritsche J, Kracht M (2017) The deep thermal field of the Upper Rhine Graben. *Tectonophysics* 694:114–129
- Garcia-Castellanos D (2002) Interplay between lithospheric flexure and river transport in foreland basins. *Basin Res* 14(2):89–104
- Geissler WH, Sodoudi F, Kind R (2010) Thickness of the central and eastern European lithosphere as seen by S receiver functions. *Geophys J Int* 181(2):604–634
- Gholamrezaie E, Scheck-Wenderoth M, Sippel J, Strecker MR (2018) Variability of the geothermal gradient across two differently aged magma-rich continental rifted margins of the Atlantic Ocean: the southwest African and the Norwegian margins. *Solid Earth* 9:139–158
- Goutorbe BLF, Bonneville A (2008) The thermal regime of South African continental margins. *Earth Planet Sci Lett* 267(1–2):256–265
- Gürbüz A (2014) Pull-apart basin. In: Harff J, Meschede M, Petersen S, Thiede J (eds) Encyclopedia of marine geosciences. Springer, Dordrecht
- Hasterok D, Chapman DS (2007) Continental thermal isostasy: 1. Methods and sensitivity. *J Geophys Res* 112(B6):B06414
- Haxby WF, Turcotte DL, Bird JM (1976) Thermal and mechanical evolution of the Michigan Basin. *Tectonophysics* 36(1–3):57–75
- Heine C, Dietmar Müller R, Steinberger B, Torsvik TH (2008) Subsidence in intracontinental basins due to dynamic topography. *Phys Earth Planet Inter* 171(1–4):252–264
- Heintz M, Kennett BLN (2005) Continental scale shear wave splitting analysis: investigation of seismic anisotropy underneath the Australian continent. *Earth Planet Sci Lett* 236(1–2):106–119
- Hieronymus CF, Goes S (2010) Complex cratonic seismic structure from thermal models of the lithosphere: effects of variations in deep radiogenic heating. *Geophys J Int* 180(3):999–1012
- Hirsch KK, Bauer K, Scheck-Wenderoth M (2009) Deep structure of the western South African passive margin – results of a combined approach of seismic, gravity and isostatic investigations. *Tectonophysics* 470(1–2):57–70
- Hudec MR, Jackson MPA (2007) Terra infirma: understanding salt tectonics. *Earth-Sci Rev* 82(1–2):1–28
- Huenges E (2010) Geothermal energy systems: exploration, development and utilization. Wiley-VCH, Weinheim, p 464
- Huismans RS, Beaumont C (2008) Complex rifted continental margins explained by dynamical models of depth-dependent lithospheric extension. *Geology* 36(2):163–166
- Huismans RS, Buiter SJH, Beaumont C (2005) Effect of plastic-viscous layering and strain softening on mode selection during lithospheric extension. *J Geophys Res* 110:B02406
- Hyndman RD, Currie CA, Mazzotti S, Frederiksen A (2009) Temperature control of continental lithosphere elastic thickness, T_e vs V_s . *Earth Planet Sci Lett* 277(3–4):539–548
- Jones AG, Plomerova J, Korja T, Sodoudi F, Spakman W (2010) Europe from the bottom up: a statistical examination of the central and northern European lithosphere-asthenosphere boundary from comparing seismological and electromagnetic observations. *Lithos* 120(1–2):14–29

- Kaban MK, Schwintzer P, Artemieva IM, Mooney WD (2003) Density of the continental roots; compositional and thermal contributions. *Earth Planet Sci Lett* 209(1–2):53–69
- Kaus BJP, Connolly JAD, Podladchikov YY, Schmalholz SM (2005) Effect of mineral phase transitions on sedimentary basin subsidence and uplift. *Earth Planet Sci Lett* 233(1–2):213–228
- Kley J, Franzke H-J, Jähne F, Krawczyk C, Lohr T, Reicherter K, Scheck-Wenderoth M, Sippel J, Tanner D, van Gent H (2008) Strain and stress. In: Littke RB, Baeyer U, Gajewski D, Nelskamp S (eds) *Dynamics of complex intracontinental basins. The Central European Basin System*. Springer, Berlin, pp 97–124
- Kounov A et al (2007) Present denudation rates at selected sections of the South African escarpment and the elevated continental interior based on cosmogenic ^3He and ^{21}Ne . *S Afr J Geol* 110(2/3):235–248
- Kovács I, Lenkey L, Green DH, Fancsik T, Falus G, Kiss J, Orosz L, Angyal J, Viktor Z (2018) The role of pargasitic amphibole in the formation of major geophysical discontinuities in the shallow upper mantle. *Acta Geodaet Geophys* 52(2):183–204
- Kusznir NJ, Ziegler PA (1992) The mechanics of continental extension and sedimentary basin formation: a simple-shear/pure-shear flexural cantilever model. *Tectonophysics* 215:117–131
- Lavier LL, Steckler MS (1997) The effect of sedimentary cover on the flexural strength of continental lithosphere. *Nature* 389:476–479
- Lavier LL, Buck WR, Poliakov ANB (1999) Self-consistent rolling-hinge model for the evolution of large-offset low-angle normal faults. *Geology* 27(12):1127–1130
- Le Pourhiet L, Chamot-Rooke N, Delescluse M, May DA, Watremez L, Manuel Pubellier M (2018) Continental break-up of the South China Sea stalled by far-field compression. *Nat Geosci* 11:605–609
- Levander A, Niu F, Lee C-TA, Cheng X (2006) Imag(in)ing the continental lithosphere. *Tectonophysics* 416(1–4):167–185
- Lister GS, Etheridge MA, Symonds PA (1986) Detachment faulting and the evolution of passive margins. *Geology* 14:246–250
- Littke R, Bayer U, Gajewski D, Nelskamp S (2008) *Dynamics of complex intracontinental basins: the Central European Basin System*. Springer, Berlin
- Lucazeau F (2019) Analysis and mapping of an updated terrestrial heat flow data set. *Geochem Geophys Geosyst* 20:4001–4024. <https://doi.org/10.1029/2019GC008389>
- Majorowicz J, Šafanda J (2018) Large regional variability of recent climatic change driven sub-surface temperature changes as derived from temperature logs-central Canada example. *Int J Earth Sci* 107(1):123–135
- Maystrenko Y, Scheck-Wenderoth M (2009) Density contrasts in the upper mantle and lower crust across the continent-ocean transition: constraints from 3-D gravity modelling at the Norwegian margin. *Geophys J Int* 179(1):536–548
- Maystrenko YP, Scheck-Wenderoth M, Hartwig A, Anka Z, Watts AB, Hirsch KK (2013) Structural features of the Southwest African continental margin according to results of lithosphere-scale 3D gravity and thermal modelling. *Tectonophysics* 604:104–121
- Mazur S, Scheck-Wenderoth M, Krzywiec P (2005) Different modes of the late cretaceous-early tertiary inversion in the North German and Polish basins. *Int J Earth Sci* 94(5–6):782–798
- McKenzie D (1978) Some remarks on the development of sedimentary basins. *Earth Planet Sci Lett* 40:25–32
- Meissner R, DEKORP Research Group (1991) The DEKORP surveys: major results in tectonic and reflective styles. In: Meissner R et al (eds) *Continental lithosphere: deep seismic reflections*. American Geophysical Union, Washington, DC, pp 69–76
- Mjelde R et al (2002) Lower crustal seismic velocity-anomalies; magmatic underplating or serpentinized peridotite? Evidence from the Voring Margin, NE Atlantic. *Marine Geophysical Researches* 23(2):169–183
- Mjelde R et al (2003) V (sub p)/V (sub s) ratio along the Voring Margin, NE Atlantic, derived from OBS data; implications on lithology and stress field. *Tectonophysics* 369(3–4):175–197
- Mjelde R et al (2005) Continent-ocean transition on the Voring Plateau, NE Atlantic, derived from densely sampled ocean bottom seismometer data. *J Geophys Res* 110:B05101. <https://doi.org/10.1029/2004JB003026>
- Nielsen L, Thybo H (2006) Identification of crustal and upper mantle heterogeneity by modelling of controlled-source seismic data. *Tectonophysics* 416(1–4):209–228
- O'Reilly SY, Griffin WL (2010) The continental lithosphere-aesthenosphere boundary: can we sample it? *Lithos* 120(1–2):1–13
- Osmundsen PT, Ebbing J (2008) Styles of extension offshore mid-Norway and implications for mechanisms of crustal thinning at passive margins. *Tectonics* 27(6):TC6016
- Parsons B, Sclater JG (1977) An analysis of the variation of ocean floor bathymetry and heat flow with age. *J Geophys Res* 82(5):803–827
- Pascal C, Cloetingh SAPL (2002) Rifting in heterogeneous lithosphere: inferences from numerical modelling of the northern North Sea and the Oslo Graben. *Tectonics* 21(6):1060–1071
- Pérez-Gussinyé M, Morgan JP, Reston TJ, Ranero CR (2006) The rift to drift transition at non-volcanic margins: insights from numerical modelling. *Earth Planet Sci Lett* 244(1–2):458–473
- Petrinin AG, Sobolev SV (2008) Three-dimensional numerical models of the evolution of pull-apart basins. *Phys Earth Planet Inter* 171(1–4):387–399
- Plomerová J, Babuska V (2010) Long memory of mantle lithosphere fabric – European LAB constrained from seismic anisotropy. *Lithos* 120(1–2):131–143
- Priestley K, McKenzie DP (2013) The relationship between shear wave velocity, temperature, attenuation and viscosity in the shallow part of the mantle. *Earth Planet Sci Lett* 381:78–91
- Priestley K, McKenzie D, Debayle E (2006) The state of the upper mantle beneath southern Africa. *Tectonophysics* 416(1–4):101–112
- Przybycin AM, Scheck-Wenderoth M, Schneider M (2015) The 3D conductive thermal field of the North Alpine Foreland Basin: influence of the deep structure and the adjacent European Alps. *Geotherm Energy* 3:17
- Ritter U, Zielinski GW, Weiss HM, Zielinski RLB, Sættem J (2004) Heat flow in the Voring basin, mid-Norwegian shelf. *Pet Geosci* 10:353–365
- Ritzmann O, Faleide JJ (2009) The crust and mantle lithosphere in the Barents Sea/Kara Sea region. *Tectonophysics* 470(1–2):89–104
- Roure F, Cloetingh S, Scheck-Wenderoth M, Ziegler PA (2009) Achievements and challenges in sedimentary basin dynamics: a review. In: Cloetingh S, Negendank J (eds) *New frontiers in integrated solid earth sciences. International year of planet earth*. Springer, Dordrecht, pp 145–233
- Royden L, Keen CE (1980) Rifting process and thermal evolution of the continental margin of eastern Canada determined from subsidence curves. *Earth Planet Sci Lett* 51:343–361
- Rychert CA, Shearer PM (2009) A global view of the lithosphere-aesthenosphere boundary. *Science* 324(5926):495–498. ISSN:0036-8075, 1095-9203
- Sandiford M (1999) Mechanics of basin inversion. *Tectonophysics* 305:109–120
- Schaeffer AJ, Lebedev S (2013) Global shear speed structure of the upper mantle and transition zone. *Geophys J Int* 194(1):417–449
- Schaeffer AJ, Lebedev S, Becker TW (2016) Azimuthal seismic anisotropy in the Earth's upper mantle and the thickness of tectonic plates. *Geophys J Int* 207:901–933
- Scheck-Wenderoth M, Lamarche J (2005) Crustal memory and basin evolution in the Central European Basin System – new insights from a 3D structural model. *Tectonophysics* 397(1–2):143–165
- Scheck-Wenderoth M, Maystrenko Y (2008) How warm are passive continental margins? A 3-D lithosphere-scale study from the Norwegian margin. *Geology* 36(5):419–422
- Scheck-Wenderoth M, Maystrenko YP (2013) Deep control on shallow heat in sedimentary basins. *Energy Procedia* 40:226–275

- Scheck-Wenderoth M, Maystrenko Y, Huebscher C, Hansen M, Mazur S (2008) Dynamics of salt basins. Dynamics of complex intracontinental basins; the Central European Basin System. Springer, Berlin, pp 307–322
- Scheck-Wenderoth M, Cacace M, Maystrenko Y, Cherubini Y, Noack V, Kaiser B, Sippel J, Lewerenz B (2014) Models of heat transport in the Central European Basin System: effective mechanisms at different scales. *Mar Pet Geol* 55:315–331
- Schotman HHA, Vermeersen LLA, Wu P, Drury MR, De Bresser JHP (2009) Constraints on shallow low-viscosity zones in Northern Europe from future GOCE gravity data. *Geophys J Int* 178(1):65–84
- Sclater JG (2003) Heat flow under the oceans, plate tectonics; an insider's history of modern theory of the Earth. Westview Press, Boulder
- Sclater JG, Christie PAF (1980) Continental stretching: an explanation of the post-mid-cretaceous subsidence of the central North Sea basin. *J Geophys Res* 85(B7):3711–3739
- Sibuet J-C, Srivastava S, Manatschal G (2007) Exhumed mantle-forming transitional crust in the Newfoundland-Iberia rift and associated magnetic anomalies. *J Geophys Res* 112(B6):B06105
- Simoes M, Braun J, Bonnet S (2010) Continental-scale erosion and transport laws: a new approach to quantitatively investigate macro-scale landscapes and associated sediment fluxes over the geological past. *Geochem Geophys Geosyst* 11:Q09001
- Simon K, Huismans RS, Beaumont C (2009) Dynamical modelling of lithospheric extension and small-scale convection: implications for magmatism during the formation of volcanic rifted margins. *Geophys J Int* 176(Suppl 1):327–350
- Sippel J, Meeßen C, Cacace M, Mechie J, Fishwick S, Heine C, Scheck-Wenderoth M, Strecker MR (2017) The Kenya Rift revisited: insights into lithospheric strength through data-driven 3D gravity and thermal modelling. *Solid Earth* 8:45–81
- Smit J, Brun JP, Cloetingh S, Ben-Avraham Z (2010) The rift-like structure and asymmetry of the Dead Sea Fault. *Earth Planet Sci Lett* 290(1–2):74–82
- Steckler MS, Watts AB (1978) Subsidence history and tectonic evolution of atlantic-type continental margins. *Earth Planet Sci Lett* 41:1–13
- Stein CA, Stein S (1992) A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature (London)* 359(6391):123–129
- Szewczyk J, Nawrocki J (2011) Deep-seated relict permafrost in north-eastern Poland. *Boreas* 40:385–388
- Theissen S, Rüpké LH (2009) Feedbacks of sedimentation on crustal heat flow: new insights from the Vøring Basin, Norwegian Sea. *Basin Res* 22(6):976–990
- Thybo H, Nielsen CA (2009) Magma-compensated crustal thinning in continental rift zones. *Nature* 457(7231):873–876
- Torsvik TH, Rouse S, Labails C, Smethurst MA (2009) A new scheme for the opening of the South Atlantic Ocean and the dissection of an Aptian salt basin. *Geophys J Int* 177(3):1315–1333
- Trumbull RB, Sobolev SV, Bauer K (2002) Petrophysical modeling of high seismic velocity crust at the Namibian volcanic margin. *Spec Pap Geol Soc Am* 362:221–230
- Turcotte DL, Schubert G (2002) *Geodynamics*, 2nd edn. Cambridge University Press, Cambridge, p 456
- Vail PR, Audemard F, Bowman SA, Eisner PN, Perez-Cruz C (1991) The stratigraphic signatures of tectonics, eustasy and sedimentology – an overview. In: Einsele G, Ricken W, Seilacher A (eds) *Cycles and events in stratigraphy*. Springer, Berlin/Heidelberg, pp 617–659
- Van der Beek P (2007) Thermochronological age – elevation profiles, denudation rates and relief development. *Geochim Cosmochim Acta* 71(15):A1055–A1055
- Van Wees JD et al (2009) Probabilistic tectonic heat flow modeling for basin maturation: assessment method and applications. *Mar Pet Geol* 26(4):536–551
- van Wijk JW (2005) Role of weak zone orientation in continental lithosphere extension. *Geophys Res Lett* 32:L02303
- Vaucher A, Tommasi A, Barruol G (1998) Rheological heterogeneity, mechanical anisotropy and deformation of the continental lithosphere. *Tectonophysics* 296:61–86
- Vendeville BC, Jackson MPA (1992) The fall of diapirs during thin-skinned extension. *Mar Pet Geol* 9(4):354–371
- Violay M, Toro GD, Nielsen S, Spagnuolo E, Burg JP (2015a) Thermo-mechanical pressurization of experimental faults in cohesive rocks during seismic slip. *Earth Planet Sci Lett* 429:1–10
- Violay M, Gibert B, Mainprice D, Burg JP (2015b) Brittle versus ductile deformation as the main control of the deep fluid circulation in oceanic crust. *Geophys Res Lett* 42(8):2767–2773
- Wang Q (2010) A review of water contents and ductile deformation mechanisms of olivine: implications for the lithosphere-aesthenosphere boundary of continents. *Lithos* 120(1–2):30–41
- Watts AB (2001) *Isostasy and flexure of the lithosphere*. Cambridge University Press, Cambridge, p 458
- Weber M, DESERT Working Group (2009) Anatomy of the Dead Sea Transform from lithospheric to microscopic scale. *Rev Geophys* 47:RG2002. <https://doi.org/10.1029/2008RG000264>
- Welte DH, Horsfield B, Baker DR (eds) (1997) *Petroleum and basin evolution*. Springer, Berlin, p 535
- Wernicke B (1981) Low-angle normal faults in the Basin and Range Province: nappe tectonics in an extending orogen. *Nature* 291:645–648
- White N, Thompson M, Barwise T (2003) Understanding the thermal evolution of deep-water continental margins. *Nature* 426(6964):334–343
- Wijns C, Weinberg R, Gessner K, Moresi L (2005) Mode of crustal extension determined by rheological layering. *Earth Planet Sci Lett* 236(1–2):120–134
- Willenbring J, von Blanckenburg F (2010) Long-term stability of global erosion rates and weathering during late-Cenozoic cooling. *Nature* 465:211–214
- Xie X, Heller PL (2009) Plate tectonics and basin subsidence history. *GSA Bull* 121:55–64
- Zhang G-B, Bott MHP (2000) Modelling the evolution of asymmetrical basins bounded by high-angle reverse faults with application to foreland basins. *Tectonophysics* 322:203–218
- Zhang YS, Lay T (1996) Global surface wave phase velocity variations. *J Geophys Res Solid Earth Planets* 101(B4):8415–8436
- Zuber MT, Parmentier EM, Fletcher RC (1986) Extension of continental lithosphere: a model for two scales of basin and range deformation. *J Geophys Res* 91(B5):4826–4838

Seismic Data Acquisition and Processing

Kabir Roy Chowdhury

Department of Earth Sciences, Utrecht University, Utrecht, The Netherlands

Definition

Seismic data acquisition Generation of (artificial) seismic signals on land (on surface, or, buried) or in water, reception of the signals after their travel through the interior of the earth, and their (digital) recording for later analysis. Analysis of recorded seismic signals to filter (reduce/eliminate) unwanted components

Seismic data processing (noise) and create an image of the subsurface to enable geological interpretation and eventually to obtain an estimate of the distribution of material properties in the subsurface (inversion).

Introduction

Reflection seismics is akin to the “echo-in-the-well” experiment, it involves calculating the depth of the geological boundary from the two-way traveltimes (TWT) of the seismic signal and its speed.

Seismic data acquisition and processing aims mainly to obtain an image of the sedimentary basins in interior of the earth, using waves generated by “artificial” earthquakes. These images can then be used to identify locations favorable for accumulation of hydrocarbons (oil and gas), which may then be drilled to determine the ground truth – and eventually to exploit the resources. Since the first known reflection seismic experiment in 1921 near Oklahoma City, USA (Fig. 1.3, Sheriff and Geldart (1995)), reflection seismics has established itself as the most accurate technique to image the sedimentary basins for the exploration of hydrocarbons.

The phrase “seismic” instead of “seismological” in the following stresses the “man-made” nature of the waves used. Both seismics and seismology use the basic theory of wave-propagation through the earth, for which Aki and Richards (2002) is a good resource. Table 1 summarizes the important differences between the two approaches though; let us briefly look at two.

Frequency vs. period: Due to the spectral range of the signals involved, seismology traditionally uses period (s) to

describe the waves, whereas in seismics, frequency (Hz.) is used. Waves provide information about the medium through which they propagate at the scale of their wave-length, and use of higher frequencies in seismics (shorter wavelengths) leads therefore to a greater resolution (of the structure) compared to seismology.

Wave-propagation: Seismology – again historically – mostly uses refracted energy, whereas exploration seismics is often synonymous with reflection seismics, although refraction seismic exploration predates the latter.

It may be noted, however, that recently there has been a blurring of the boundaries between the two fields. This has been driven by progress both in the instrumentation (acquisition) and in theory (processing). Seismologists are increasingly using higher frequencies and reflected wavefields, and practitioners of reflection seismics are using lower frequencies and refracted phases – both with the aim to improve the imaging of their respective targets. Italicized items in Table 1 are thus changing with time.

A similar comment applies also to “noise.” Considered earlier to be a bane of imaging, it is being increasingly used in innovative ways to optimize acquisition and improve imaging; this aspect will be briefly dealt with later.

This essay will be mainly concerned with acquisition and processing of reflection seismic data. Note, however, that seismics is being increasingly applied to both shallower depths (high resolution seismics) and crustal-scale studies down to Moho and beyond (deep seismics), see ► [“Deep Seismic Reflection and Refraction Profiling”](#) for details of the latter. Seismic Data Acquisition and Processing is a broad subject, the treatment here will have to make choices based upon space constraints, etc. Some subtopics, for example, ► [“Seismic, Migration”](#) are, however, dealt with in separate essays in this volume. This essay has been updated for the 2nd

Seismic Data Acquisition and Processing, Table 1 Imaging the earth using natural/artificial earthquakes

Keyword	Seismics	Seismology
Wave source	Explosions, vibrations	Natural earthquakes
Energy penetration	Shallow	Deep
Max imaging depth	Base of crust	Whole earth
Location of source	Precisely known	Estimated postfacto
Time of occurrence	Precisely known	Estimated postfacto
Energy involved	Small-medium	Can be huge
<i>Wave propagation</i>	Mostly vertical	Mostly horizontal
<i>Frequencies mostly excited/used</i>	1–100 Hz	0.01–1 Hz.
Receivers	Geophones	Seismometers
Wave-field sampling	Dense	Sparse (getting better)
Data volume	Terrabytes	Gigabytes
Accuracy	Large	Small-medium
Main application	Oil and gas	Earth-structure
Other applications	Civil engg., crustal	Civil engg.
Investment	\$\$\$\$	\$\$

\$\$\$\$ Very → expensive; \$\$ → Expensive

edition by briefly indicating relevant advances and including references for the same. The adjective “classical” has been used at places to indicate possible recent changes of the subject/term under discussion.

The reader is assumed to be familiar with the basic theory of elasticity and wave-propagation and the related concepts of reflection, refraction, and scattering. The concept of rays will be frequently used – especially in illustrations – for convenience; real seismic signals are of course associated with wave-fronts. Similarly, the figures will depict a 2-D section of the 3-D earth.

There are many good resources available even for the narrower field of Reflection Seismic Data Acquisition and (Signal) Processing, for example, Vermeer (2002), Yilmaz (2001), Menke (1989), Liner (2004), and Sheriff and Geldart (1995); the last one also contains some historical background and material over refraction seismics. Recently, some resources have also been made available for downloading on the Internet, for example, Claerbout (1985a, b).

In this entry, all-capitals will be used to denote acronyms for jargons, of which there are quite a few (e.g., TWT above), and italicized phrases within double quotes will refer to articles elsewhere in this volume, for example, ► [“Propagation of Elastic Waves: Fundamentals”](#).

Seismic Data Acquisition

Before seismic signals can be processed, an artificial wavefield has to be generated using suitable *sources* at appropriate locations, measured by suitable *receivers* at other locations after getting reflected back from within the earth, and stored using *recorders*. Several technical and financial parameters have to be considered to obtain optimal data, for example, dense spatio-temporal sampling, power-supply, timing accuracy (GPS?), and – increasingly – HSE (Health–Safety–Environment) aspects. Design of a seismic survey (*geometry*) needs inputs from regional geology, exploration objectives, and logistical considerations.

At first confined to land, seismic surveys are now-a-days carried out mostly in marine environments in round-the-clock operations using large vessels and a lot of instrumentation; single-channel seismics has faded away in favor of multi-channel acquisition, allowing much more information to be obtained (see ► [“Single and Multichannel Seismics”](#)). Table 2 gives an overview of the equipment used under different field environments.

Several interesting theoretical- and practical-advances have been made recently in the field of seismic data acquisition and processing. These include *simultaneous* recording of wavefields *overlapping* in time and use of *noise as signal*. These will be briefly described towards the end of this entry.

Seismic Data Acquisition and Processing, Table 2 sources and receivers used in seismic surveys

Environment	Sources	Receivers
Land	Explosives/vibrators/ impact/noise	Geophones
Marine	Air/water-guns	Hydrophones/nodes
Water-bottom	Explosives/guns	Nodes/geo/hydro- phones
Onshore- offshore	Explosives/guns	Geo/hydro-phones

Seismic Sources

One needs a signal that is high in energy (amplitude) to ensure a good depth penetration and short in duration to ensure accurate determination and *separation* of the travel-times – a Dirac-Delta spike would be ideal, possessing perfect resolution, it is however a-causal. In practice, a sharp, compact, and *repeatable* signal is preferred. This quasi-idealized waveform, possessing finite temporal duration and frequency band-width (both with respect to the ambient noise), is called a *wavelet*. The source wavelet changes form as it travels through the earth due to several physical processes to be briefly discussed below – much of the later data processing aims to undo these changes.

Repeatability of the source wavelet – that is, that of its *amplitude* and *phase* content is an important pre-requisite for the later processing steps. Explosives were the initial choice for source on land, providing large energy (good depth penetration) but having non repeatable signal shape and negative environmental impact. Development of large, truck-mounted electro-mechanical vibrators have led since 1960s to their increasing use in land-seismics, with both the above disadvantages of impulsive sources reduced significantly.

In marine environment, compressed air is used – with explosion (*air gun*) or implosion (*water gun*) – to create the acoustic waves. The sources are towed by a ship together with the receivers (*single vessel seismic*) or by a separate ship (*two-ship seismics*).

There have also been experiments with shear-wave sources – both impact-type and vibratory. These – either alone or together with compressive sources – can provide extra information about the subsurface medium. For investigating shallower structures in engineering, environmental, and archeological applications, small impact-based sources, for example, weight-drops, rifles, and even portable vibrators are being frequently used and provide the required higher resolution.

Seismic Receivers

All land-seismic (and seismological) receivers (*geophones*, *seismometers*) are electro-mechanical devices that transform the relative motion of the medium in which they are embedded, into electrical voltages. Fidelity of this transformation –

both in *amplitude & phase* is important to ensure maximum information retention for later retrieval – a flat amplitude response, with no phase distortion within the band of frequencies that are of interest, would be ideal. The output of such devices can be made to be proportional to the displacement/velocity/acceleration associated with the causative motion. Dennison (1953) provides an overview of the physico-mathematical underpinnings of geophone design; see also ► [“Seismic Instrumentation”](#).

Originally, geophones were designed to move – and hence record information – only in the vertical direction. Later, the importance of recording and analyzing the entire three-dimensional elastic wavefield came to be realized. Multi-component receivers, enabling recording/analysis of both horizontal components, or all three spatial components of the ground movement are being increasingly used even in large scale surveys.

For use in water, small piezo-electric elements – hydrophones – are employed to record pressure variations – modern deployments typically consist of thousands of such elements being towed near the water surface by *streamers*, several kilometers long, which are liquid-filled plastic tubes, fitted with fins (for buoyancy), gps receivers (for location information), and fiber-optic cables (to transfer the data) to the ship.

Three-component receivers may be deployed together with hydrophones at the water bottom (4C), to record the wavefield across it, see ► [“Ocean Bottom Seismics”](#). Advances in this area have resulted in cable-free autonomous “nodes” – for both marine- and land-environments, for example, Dean et al. (2018). Besides saving cables and manpower (read: *cost*) for their deployment, they allow for complex acquisition geometries, large recording distances, and are also environmentally friendly.

Note that both sources and receivers may be deployed in *groups*, using specific patterns, which affect the generation and sampling of the wavefield due to their direction-dependent radiation/reception characteristics.

Seismic Recorder

The time-varying electrical signals output by the receivers represent *arrivals* back-scattered from different depths, all juxtaposed in time, and embedded within the ever present background noise and require storage for later processing. In the beginning, photographic films and later magnetic tapes were used for this purpose. The digital revolution starting in the 1960s, itself partly driven by the needs of seismic data-acquisition and processing, caused a complete shift to in situ digitization and digital storage. Similarly, the wires connecting the receivers to the recorder have been mostly replaced by fiber-optic cables or wireless. Modern 3-D seismic surveys record four-dimensional data – two each for sources and receivers; repeat surveys (*time-lapse seismics*) even add a fifth dimension.

Preserving the frequency, amplitude, and phase of the signal and the desired *dynamic range* are important considerations in designing the digitizing (*sampling*) unit. Digitization of this vast amount of data (wavefield) has traditionally been done honoring the *Nyquist Sampling Criterion*, which requires a uniform sampling rate of more than twice the highest frequency – or wavenumber – depending upon the domain (time/space) – present in the data. Sampling below the Nyquist rate produces artifacts (called aliases) in the digitized data at lower frequencies (or wavenumbers) and can eventually interfere with real signals that may be present there – see Claerbout (1985a) or Menke (1989) for details.

Theoretically, Nyquist rate sampling allows the reconstruction of the uniformly sampled wavefield using an infinite integral – which, clearly, is not practical. Hence, the signal received at each receiver location is usually over-sampled (e.g., at twice- or higher-Nyquist rate) to maintain high-fidelity while reconstructing the wavefield from the traces with a finite effort, without loss of information. Spatial sampling of the seismic wavefield is seldom regular and uniform and produces “gaps,” which need to be filled by interpolation; this will be briefly discussed in a later section.

Recently, the necessity of acquiring seismic data at- or above-the Nyquist criterion has come under critical review – especially in view of the resultant massive increase in the data volume. An interesting development in this field has been included towards the end of this entry.

As each *receiver (group)* corresponds to a different *channel* in the recorder, digitization in such systems (typically consisting of thousands of channels) must preserve the time-base, to enable comparison of the arrival-times between different *traces*. Also, the actual instant of the *shot* – t_0 – must be transferred from the source to the recorder and recorded; it is often used to start the recording process itself, as seismics is only interested in travel-times, that is, arrivals-times with respect to t_0 . The digitized data – uniformly sampled time-series – from each individual experiment (*shot*), consisting of multiple *traces* (output of receivers) is called a *seismic record*.

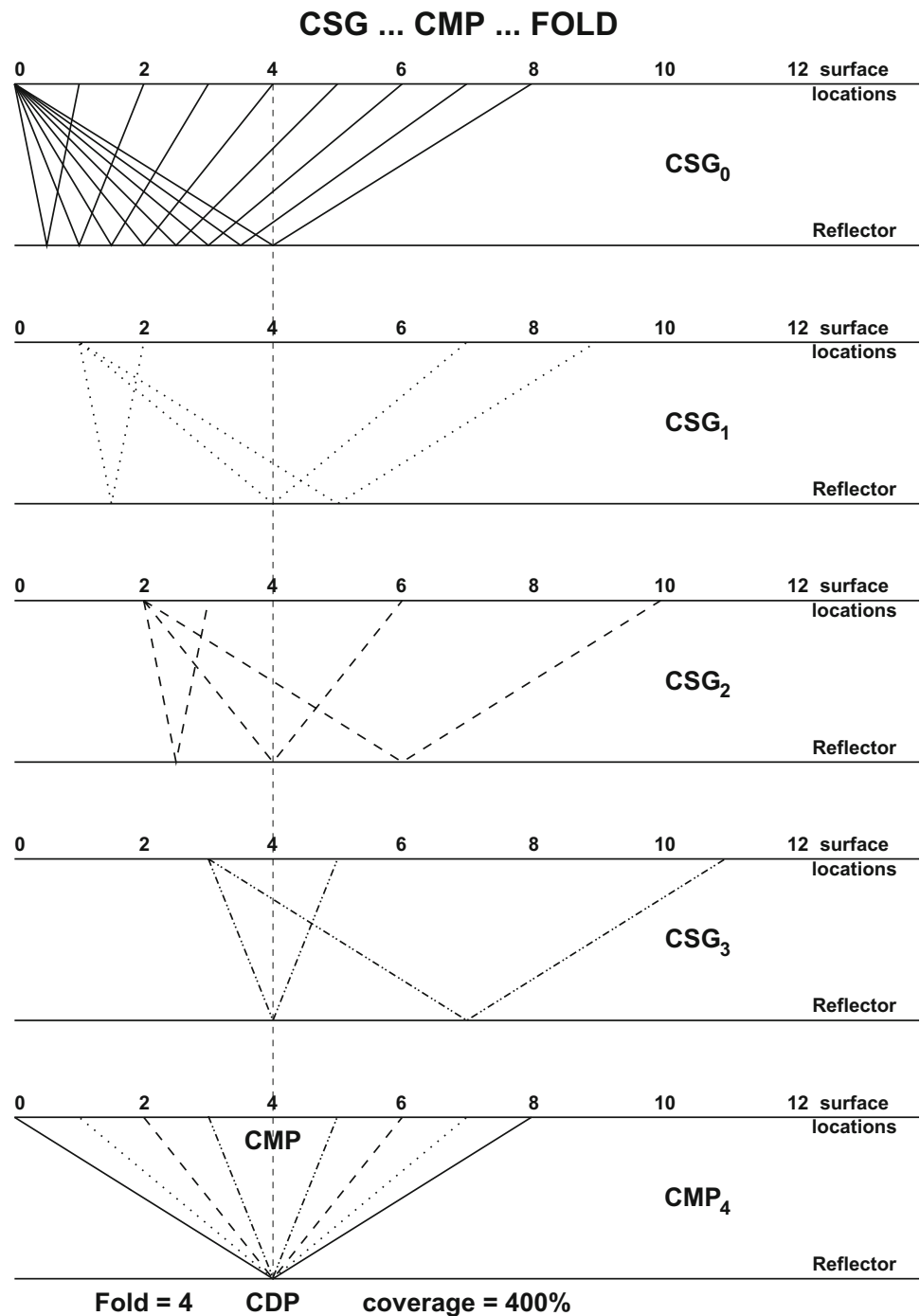
Acquisition Geometry

Assuming a *layer-cake* model (sedimentary beds parallel to the surface), early surveys deployed a number of sources and receivers along a straight line on the earth-surface to obtain a vertical cross-section of the geology below this line (2-D).

Figure 1 shows schematically the approach in such a survey and is – in spite of its simplifications – useful in understanding several basic ideas. All the five panels show the earth-surface at the top, and a reflecting boundary (*target*), parallel to it, at some depth. Numerals on the surface represent surveyed equi-distant flag-positions to denote locations. The top panel shows the first measurement, with the source at “0” and eight receivers at locations “1” thru “8”. Assuming a

Seismic Data Acquisition and Processing, Fig. 1

Schematics of seismic data acquisition by common-mid-point (CMP) profiling. Panels CSG_0 through CSG_3 represent common-source-gathers (CSG); CMP_4 is a common-mid-point gather for one common-depth-point (CDP). See text for details



homogeneous and isotropic medium, the paths predicted by Snell's law for a part of the source energy to first travel downwards to the target and then reflect upwards from it to reach the receivers are indicated by the oblique lines.

The signals output from the receivers are digitized in the recorder to yield a *seismic record*, that is, a collection of *seismic traces*. Such an ordered collection of seismic traces

is called a *gather*; having a common *source*, the record resulting from our first measurement is a *common source gather* (CSG_0), the suffix denoting source position.

Under the twin-idealizations of no background noise, and a spike-like source signal, each channel in the recorder (*seismic trace*) will consist of one single blip corresponding to the arrival time of the signal; in reality, the arrivals will have

random background oscillations due to noise, and one *wavelet* corresponding to the single reflection arrival. Assuming constant speed of propagation v and depth to the target H , it is trivial to show (e.g., Sheriff and Geldart 1995) that the travel-times to the receivers can be written as $t_x^2 = (x^2 + 4H^2)/v^2 = t_0^2 + x^2/v^2$, t_x being the arrival-time recorded by a receiver at a source-receiver *offset* of x . The travel-time curve for such a situation is thus a hyperbola – this simple relationship underlies much of seismic processing. t^2 plotted against x^2 thus yields a straight line, the slope being v^{-2} , that is, square of the *slowness* of the medium. Note that in seismics the *velocity*, which sensu-stricto is a vector, is almost always used to denote the local wave *speed* (a scalar), which is a property of the medium (rocks) ... we shall follow this usage.

Our aim is to find H , the depth to the target ($= t_0/2v$). We have thus to estimate t_0 from the rest of the reflection hyperbola. Note that the reflection points on the target for the different receivers are different.

Hence, in what has become almost universal practice, the measurement is repeated after shifting the whole set-up laterally along the measurement line, keeping all the relative distances the same. In panel 2 of Fig. 1, the source and the receivers have been shifted right by one unit; only a few ray-paths are shown for this gather (CSG_1). Similarly, gathers CSG_2 and CSG_3 are also measured and recorded. During these measurements, the same receiver locations recorded signals from different sources, so that a postmeasurement re-arrangement of the traces could also yield *common receiver gathers* (CRG), in our case we would obtain CRG_1 – CRG_{11} ; these are useful for certain processing situations.

The lowest panel of this figure shows a special kind of resorting, collecting the traces from the four *shots* with one *common reflection point* (CRP). Four traces corresponding to source/receiver combinations of 0/8, 1/7, 2/6, and 3/5 were selected respectively from the four gathers. For our simple geometry, the four ray-paths shown share two things – a *common mid-point* (CMP_4) between their respective source and receiver locations and the *common depth point* (CDP) at the target depth, the latter being the same as CRP. Such a gather is called a CMP-gather and indexed by the position of the CMP. The travel-time plot of the reflection arrivals in a CMP-gather is also a hyperbola.

The four ray-paths shown for the gather CMP_4 all have the same reflection point and thus contain information about the same subsurface geology. The arrival times of the reflection signal in the four traces are of course different, as the travel paths are different. If this difference is corrected for, then adding the four traces should increase the coherent signal (information regarding the CRP) with respect to the *random* noise. The improvement of S/N by adding N traces is given by

$$\frac{\sum(\text{N traces with identical signal})}{\sum(\text{N traces with random signal})} \approx \frac{N}{\sqrt{N}} = \sqrt{N}.$$

The improvement of the signal-to-noise (S/N) ratio is thus roughly proportional to the square-root of the number of traces added. This number (4 in our case) depends upon the survey geometry and is called the *fold* of the survey. Starting from fold 1 for $CMP_{0.5}$ (not shown), it gradually builds up to its nominal value (4 in this case) and again drops off at the other end of the survey.

Acquisition configuration can be specified by expressions describing the position of the source relative to the receivers viz. *end-on*, *split-spread*, *broad-side*, etc. Depending upon the geology and the noise regime, these configurations, as also varying *fold*, leave subtle but important *footprints* on the data.

In reality, the geology is of course not as in Fig. 1, and presence of structure (*dips*, *faults*, *folds*, etc.) is what makes hydrocarbon accumulation possible in the first place. Processing of 2-D data can remedy this situation – though only partially. Availability of more equipment and data-processing power led therefore to development of 3-D acquisition, with receivers laid out on the surface in a 2-D pattern, and sources also positioned in a different 2-D pattern, thus causing a better illumination of the subsurface by the seismic waves. Here too, the basic concept of adding *fold* number of traces in a CMP-gather holds sway – point-shaped CMPs and CDPs being replaced by finite *bins*, their sizes depending upon the survey design and objectives (see Vermeer 1990, 2002 for further insight into acquisition design).

In areas with structural complexity, the simplifying assumptions of CMP- processing break down, and the availability of computer power may make it possible – nay desirable – to process each trace of the recorded CSG separately, to try to obtain a better image (see also ► “[Seismic Imaging, Overview](#)” and ► “[Seismic, Migration](#)”).

Restricting the deployment to the surface of the earth implies – as we shall see later – a bias for horizontal structures; this was eventually removed by carrying out measurements inside bore-holes called VSP; see ► “[Vertical Seismic Profiling](#)” for details. Finally, better recording instrumentation coupled with the need to detect changes in the hydrocarbon reservoirs resulting from exploitation have given rise to *time lapse seismic* (4D), whereby repeat imaging of the same area carried out after several years of production is used to validate/improve production models for reservoirs.

Seismic Data Processing

Introduction

Reflection seismic data, acquired in the field, has to be taken through several processing steps, before it can be interpreted

in terms of the subsurface structure. The source signal, on its way down, and back up to the receivers is modified by many factors; the aim of processing is to *undo* (i.e., correct for) as many/much of these effects as possible, leaving only the effects due to the causative structure of interest (geology) to be interpreted.

Seismic data is a spatio-temporal sampling of the back-scattered seismic wavefield, an ordered collection of traces, and can be considered to be a 2/3-D data matrix along with some auxiliary information regarding location, etc. As indicated before, analysis of the recorded wavefield requires it to have been uniformly and properly sampled in all relevant dimensions. Field logistics, however, frequently causes the spatial dimension(s) of the wavefield to violate this criterion. The resulting data will then be irregularly sampled in that dimension and will need resampling. Gaps may exist even within areas of otherwise equidistant traces due to problems associated with some receiver locations (houses, factories, etc.).

Interpolating the recorded data to fill up these gaps to enable further processing requires theoretical underpinnings, for example, Gülünay (2003), Zwartjes and Sacchi (2007), and Liu and Fomel (2011). These are regularly employed to preprocess the seismic data before the real work (processing) can begin. It may be mentioned here that even after recovering uniformly sampled (spatial) wavefield obtained from an irregular acquisition geometry, the data may be aliased for coherent noise, for example, groundroll and multiples; see below for a brief explanation of these terms and ► “Seismic Noise” for details.

Recently there have been some interesting advances in this field, which economize by intentionally undercutting the Nyquist criterion during acquisition before recovering the wavefield by suitable preprocessing – these will be briefly mentioned at the end of this entry. Presently, the spatial dimension too will be assumed to have been uniformly- and properly sampled.

The traces themselves are an ordered collection of uniformly sampled amplitude values (*time-series*), with relevant information contained in their respective headers in (internationally) agreed formats. All processing steps aim to improve the spatio-temporal S/N ratio of the data by reducing the noise and/or by sharpening the wave-form (to improve the resolution).

Signal vs. Noise

Before proceeding further, it is useful to reflect on the terms *signal* & *noise*. That it is a matter of perspective is clear from this relative definition: *signal is useful noise and noise is useless signal*. In other words, someone’s noise is someone else’s signal, and vice versa. For example, the *ground-roll*, hated in reflection seismics, is useful in surface-wave

seismology and shallow-seismics. Amazingly, using noise for seismic imaging has now become a field of active research (see the section “Noise as a Seismic Source” for references).

In classical reflection seismics, signal is synonymous with *primary reflection*. *Primaries*, as these are often referred to, represent seismic energy reflected only once during its travel from source to receiver. Everything else, present in the traces, is taken to be *noise*. This includes multiply-reflected energy (*multiples*), diffractions (caused by *sharp* structures in the subsurface, for example, faults, pinch-outs), refracted arrivals, surface-waves (*ground-roll*). Nongeological noise sources include nature (wind, waves, animals) and man (traffic, drilling, industry, etc.). From processing point of view, noise could be coherent (ground-roll, water-pump, multiples), or, incoherent, each needing a different strategy. See ► “Seismic Noise” for details.

Kinematics of the Seismic Signal (Primaries)

Starting with some simple (but inaccurate) assumptions, for example, horizontal layering, constant speed, useful structural information can be extracted – a large data volume contributing to the robustness of the processing algorithms (also see ► “Seismic, Migration”). In this section, we focus on the travel-times of the waves (visualized as *rays*), see also additional information in ► “Seismic Ray Theory”.

NMO

The travel-time for a primary-reflection from a horizontal reflector, shown earlier to be hyperbolic, can be rewritten as:

$$t_x - t_0 = \frac{\sqrt{x^2 + 4H^2} - 2H}{v}. \quad (1)$$

The quantity on the left is the difference (see Fig. 2) between the oblique reflection-time at source-receiver offset (distance) x and the vertical TWT and leads to the relation:

$$\Delta t_x = \frac{2H}{v} \left(\sqrt{1 + \frac{x^2}{4H^2}} - 1 \right) = t_0 \left(\sqrt{1 + \frac{x^2}{4H^2}} - 1 \right). \quad (2)$$

Expanding the expression under square-root, and recognizing that in most seismic measurements offset $\ll$ target-depth, we obtain the approximate relation 3, which could be improved by retaining additional higher order terms.

$$\Delta t_x \approx \frac{x^2}{4vH} = \frac{x^2}{2v^2 t_0} \quad (3)$$

$\Delta t_x (= t_x - t_0)$ is called the *normal moveout* (NMO) associated with the reflection travel-time. NMO can be used to

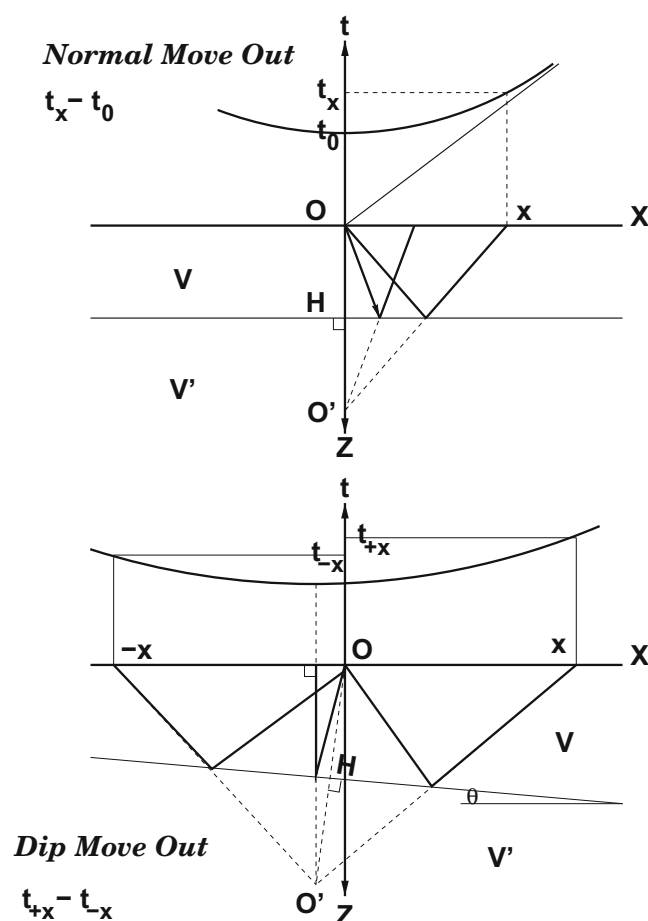
align the primary reflection in all traces at t_0 (TWT) by removing the effect of source-receiver distance (offset), that is, by *flattening the reflector*. NMO, an important concept in seismics, is used both to first identify *primaries* and later to align them for imaging the reflector. Note that to use 3 we need to know x (source-receiver offset), v (speed), and H (target depth); in practice, x is known and iteration is used to obtain optimal values for v and H .

Dipping Bed

For a dipping reflector (Fig. 2, below), travel-time for the primary reflection is still hyperbolic, given by

$$v^2 t_0^2 = x^2 + 4H^2 + 4Hx \sin \theta. \quad (4)$$

The minimum of the hyperbola is now shifted *updip*; the quantity $t_{+x} - t_{-x}$ is a measure of the asymmetry and can be used to estimate the dip.



Seismic Data Acquisition and Processing, Fig. 2 NMO of a primary reflector; horizontal (above), dipping (below). See text for details

Many Reflectors: Layer-Cake

Dix (1955) considered the case of many reflectors parallel to the surface – a good starting model for sedimentary sequences – and showed, that here too, the travel-time curve can be approximated at short offsets by a hyperbola:

$$t_x^2 \approx t_0^2 + \frac{x^2}{v_{rms}^2}, \text{ with } v_{rms} = \sqrt{\frac{\sum v_i^2 \Delta t_i}{\sum \Delta t_i}}. \quad (5)$$

The homogeneous velocity v ($= v_{nmo}$) is now replaced by v_{rms} (root-mean-square velocity), which depends upon the velocities of the layers v_i and the vertical transit times t_i through them. V_{rms} plays a role similar to v_{nmo} in flattening the primaries in the multilayer case. Individual layer-velocities may then be computed from the Dix' equation:

$$v_n = \sqrt{\frac{v_{rmsn}^2 t_n - v_{rmsn-1}^2 t_{n-1}}{t_n - t_{n-1}}}. \quad (6)$$

Velocities in Seismics

In seismics different terms are used to denote “velocity” depending upon the context. Table 3 lists a few, along with brief explanations, some of these will be elaborated later.

NMO Stretch

After NMO correction, a time-interval Δt_x , say corresponding to a period of the wavelet recorded on a trace at an offset x , becomes Δt_0 ; the wavelet is thus distorted (*stretched*). The expression $100 (\Delta t_0 - \Delta t_x) / \Delta t_x$ is a percentage measure of this stretch – 0% implying no distortion. In practice, a threshold percentage is specified to exclude parts of data – relatively large offsets and small arrival times – from being NMO-corrected (and taking part in further processing).

Semblance: A Measure of Signal Alignment

To apply optimal NMO correction, a quantitative measure of alignment of amplitudes, across several traces, is useful. Such a measure of similarity between n (amplitude) values, called *semblance*, is defined by

Seismic Data Acquisition and Processing, Table 3 Jargons associated with the term velocity in seismics

Jargon	Brief description
v_{int}	Speed in a geological interval (assumed constant)
v_{av}	Average speed between two points along a ray path
v_{app}	Apparent speed measured by receivers in field ($= dx/dt$)
v_{nmo}	Speed used for NMO-correction (strictly, only for one layer)
v_{rms}	Dix' root-mean-square NMO velocity for layer-cake situation
v_{stk}	best velocity to stack CMP gathers
v_{mig}	best velocity to migrate the seismic data

$$S = \frac{(\sum_n val)^2}{n \sum_n val^2}, \quad \text{and } S_{\text{gate}} = \frac{\sum_{\text{gate}} (\sum_n val)^2}{\sum_{\text{gate}} (\sum_n val^2)}. \quad (7)$$

Note that semblance is a dimensionless number between 1 (perfect match) and 0 (perfect mismatch). The second form uses a time gate along the traces, generally having the width of the dominant period of the signal, for increased robustness. Semblance is used extensively in modern *reflection velocity analysis*, to evaluate the goodness of alignments of primary reflections along move-out curves computed for a range of trial velocities (Fig. 3).

Velocity: Processing Point of View

Wave-speed (called velocity in seismics) in the medium is the missing link needed to convert the travel-time information to depths required for structural interpretation – and eventually drilling. Note that *velocity* is needed to find the structure (geology), but structure is needed to find the velocity. This catch-22 situation is solved iteratively – shown schematically in Fig. 4.

Velocity is a macroscopic (wave-length-scale average) property of the rock depending upon the density and elastic properties of the minerals making up the lithology (see ► “Seismic Properties of Rocks”). In rocks of interest in seismics (*sandstone, shale, limestone*), velocity is not a good indicator of lithology, with considerable overlap in

values, with some exceptions, for example, salt, anhydrite (relatively higher velocity). Presence of *porosity and pore-fluids (water, oil, gas)* are the most important factors for this overlap and are in turn caused by the *burial history* of the rocks. Wave-propagation in fluid-filled porous media is described by Biot-Gassman theory, see Lee (2008) for references and recent developments.

Propagation velocity (the missing link) can be estimated by direct measurements (see, e.g., Sheriff and Geldart 1995 for details), which have shortcomings though (see Table 4). The velocity used for processing seismic reflection data is usually determined iteratively from the data itself and will be described later.

Amplitude Changes along the Propagation Path

Several factors cause the amplitude of the seismic waves to change as they travel from source to receiver. These can be corrected for, so as not to mask the weaker changes (signals) of interest.

Geometrical Spreading

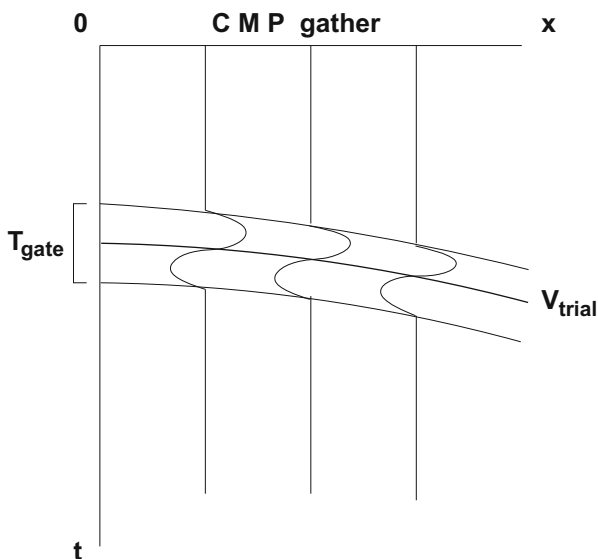
Conservation of energy requires a continuous reduction of amplitude, as a seismic wave-front spreads through a medium – hence the term *geometrical spreading*. The loss depends upon the mode of spreading and the distance travelled (r). For primaries (body-waves), amplitude ($\propto \sqrt{\text{energy}}$) decreases $\propto r^{-1}$, whereas for ground-roll (surface-wave), the decrease is $\propto r^{-1/2}$, the latter shows why ground-rolls, with their (relatively) large amplitudes, are a big problem in seismics.

Absorption

The propagating wave continuously loses energy due to absorption too, which is a physical property of the medium, and can be described by several equivalent parameters, for example, *absorption coefficient, damping factor*, the most common being the *quality factor* of the material, $Q = 2\pi/[\text{fractional energy lost per cycle}]$. It is a dimensionless quantity, with a value of 0 implying perfect absorption and ∞ implying perfect elasticity. Absorption, with Q considered to be frequency-independent within the band-width of interest in seismics, causes relatively greater attenuation of higher frequencies – leading to a change in the waveform during propagation. See ► “Seismic Viscoelastic Attenuation” for more details.

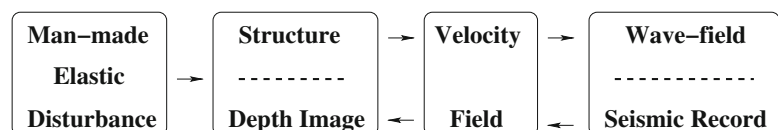
Energy Partitioning at Interfaces

Boundaries of geological heterogeneities (layering, faults, etc.) also cause changes in the amplitude of the wavelet; such



Seismic Data Acquisition and Processing, Fig. 3 Schematic drawing showing calculation of multichannel semblance. Curved bold line represents the move-out curve for a trial velocity, and the two surrounding lines represent the boundaries of the time-gate; see text for details

Seismic Data Acquisition and Processing, Fig. 4 Iteratively solving for both structure and velocity in seismics



Seismic Data Acquisition and Processing, Table 4 Direct determination of seismic velocities and their shortcomings

Method	Shortcoming
Uphole-time	Useful only for the weathering layer
Check-shots, well-shoot	Limited depth-range, destructive
VSP	Available late in exploration, expensive
Sonic log	Available late, noisy (high-frequency)
Lab measurements	Limited availability

changes are, indeed, of prime interest in seismics. As in optics, the interaction between the wave-fronts and the geological structure depends upon their relative dimensions, that is, their radii of curvature – with specular reflections and point-scattering building the two end-members, both of which are encountered in seismics. Another concept from optics, *diffraction*, is useful to understand the complexity of the interaction between the wave-front and the medium. See ► “[Seismic Diffraction](#)”, ► “[Seismic Waves, Scattering](#)”, and ► “[Energy Partitioning of Seismic Waves](#)” for additional details.

Waveforms: Convolution, Deconvolution

Factors modifying the source signal along the path of the seismic wave may be divided as: near-source (*ns*), that is, weathering layer, earth (*e*), that is, the target geology, near-receiver (*nr*), receiver (*r*) and recorder (*rec*), with the output trace (*o*) as the final result. Each of these, denoted in Eq. 8 below by the expression in parentheses, affects (*filters*) the source wavelet (*s*). In a series of papers/reports (Robinson 2005; Treitel 2005), the MIT geophysical analysis group (GAG) laid the foundation of the *digital revolution* in seismic data processing, by examining the nature of these filters and developing methods to undo their effects. These resulted in major advances in *time-series analysis and digital filtering* (Robinson and Treitel 1964) and a critical evaluation of the (statistical) nature of earth’s reflectivity (*target geology*).

Convolutional Model of the Seismic Trace

As the source – and recorded – signals are both time-series (uniformly sampled, ordered collection of amplitudes), it is useful to represent all the other elements mentioned in the above-paragraph also as such. For a column of vertically layered reflectivity, such a time series would correspond to values equal to *RCs* placed at times converted from depths using velocities. Now, making the crucial assumption that all these filter elements are linear systems, the recorded trace can be expressed as:

$$o(t) = s(t) * ns(t) * e(t) * nr(t) * r(t) * rec(t) + n(t). \quad (8)$$

In Eq. 8, $*$ (*star*) is the convolution operator, well-known in the theory of linear systems; $n(t)$ represents some additive noise which does not follow this model, hopefully, it is mostly removed early in the processing. The time-series that transform $s(t)$ into $o(t)$ can also be interpreted as the *impulse*

response of the corresponding elements, for example, $r(t)$ is the response of the receiver to a sudden spike signal. Using *Fourier-Transforms* to change the time-series into their *spectra*, and remembering that convolution in time-domain corresponds to multiplication in frequency domain, one obtains:

$$O(\omega) = S(\omega) \cdot NS(\omega) \cdot E(\omega) \cdot NR(\omega) \cdot R(\omega) \cdot REC(\omega), \quad (9)$$

where the noise term has been neglected (see Sheriff and Geldart 1995 for introduction to linear operators and Fourier theory). Equation 9 clearly shows the *filtering* effect of the different elements, each one modifying the spectrum of the incoming signal by modifying/removing a part of its frequencies. Our aim, in seismic data processing, is to extract $e(t)$, the geological structure from the recorded signal $o(t)$.

Deconvolution as Inverse Filtering

Undoing the act of the filterings implied in Eqs. 8 and 9 is called *deconvolution* (decon), or, *inverse filtering*. Equation 9 can be rewritten as $O(\omega) = E(\omega) \cdot REST(\omega)$, where $REST(\omega)$ groups together all the elements on the right besides the geology. Then, $E(\omega)$, or $e(t)$, can be *estimated* from

$$E(\omega) \cong O(\omega) / REST(\omega), \quad \text{or,} \quad (10)$$

$$e(t) \cong o(t) * rest(t)^{-1}.$$

The approximation sign, for both forms of Eq. 10 – in frequency-domain (first), or, in time-domain (second) – is necessary, even in the *noise-free* case. Spectral division needs precautions to avoid zero-division in parts of the spectrum, where frequencies have been weakened/removed. Fortunately, addition of *noise* helps, since signals of interest in seismics exist – by definition – only above the ambient noise level. See Liner (2004), Sheriff and Geldart (1995) and Yilmaz (2001) for the stabilizing role of *spectral whitening* in decon.

Wavelet Processing

Wavelets: Let’s take a closer look at *seismic wavelet*, introduced in the section about “[Seismic Sources](#),” as a signal of finite frequency band-width and temporal duration. Using standard concepts from time-series analysis (Sheriff and Geldart 1995; Yilmaz 2001), simple examples of wavelets are:

$$a : (3, -2, 1), \quad b : (2, 3, -1) \quad \text{and} \quad c : (-1, 2, 3),$$

the numbers representing uniformly sampled amplitudes starting from $t = 0$. Remembering that squares of the amplitudes in a wave(let) are measures of energy, we see that these three wavelets, while looking very different, have the same total energy. Depending upon the energy build-up, wavelet *a* is called *minimum-delay* (energy is front loaded), *b* is *mixed-delay*, and *c* is *maximum delay*; physical (causal) wavelets are

minimum delay, although in the example, a is not strictly causal, due to the instantaneous build-up of energy at $t = 0$. In frequency domain, the expressions minimum/mixed/maximum-phase are used instead.

Wavelet estimation: *Auto-correlation* of the wavelets a - c are all symmetrical about $t = 0$, that is, have no phase information, for example, $\varphi_{bb} = (-2, 3, 14, 3, -2)$; these are Fourier Transforms of the respective power-spectra. In seismics, an estimate of the power spectrum is often available from the data. The question then arises whether an estimate of the wavelet may be obtained from it – an outline follows. Using Z -transform notation, one can write the wavelet, say c , and its auto-correlation as polynomials:

$$C(Z) = -1 + 2Z + 3Z^2, \text{ and,} \\ \Phi_{cc}(Z) = -3Z^{-2} + 4Z^{-1} + 14 + 4Z - 3Z^2,$$

Z being the *unit-delay* operator, its powers denoting time-shifts with respect to $t = 0$. According to the fundamental theorem of algebra, a polynomial of degree n in Z must have n roots, that is, it can be expressed as a product of n factors of the form: $(Z - Z_1)(Z - Z_2) \dots (Z - Z_n)$, each factor representing a basic wavelet (*doublet*). Half the doublets of an auto-correlation polynomial are minimum-delay; their product represents the Z -transform of the unique minimum-delay causative wavelet. See Yilmaz (2001) and Sheriff and Geldart (1995) for details, assumptions, critical remarks, and alternate approaches (e.g., *homomorphic deconvolution*) to deconvolution of time-series.

Wavelet manipulation: Much of seismic processing is involved with manipulating the wavelet (deconvolution in a general sense). While very powerful, it contains potential for pitfalls, if applied without a proper understanding of the suitability of the particular technique, as each decon step also causes artifacts.

Spiking decon aims to sharpen the shape of the signal, to improve temporal resolution – and interpretation. Ideally, it involves convolving the wavelet with its inverse operator, to yield a spike, that is, perfect resolution.

Zero-phasing converts the signal to one with zero-phase; the result is a symmetrical signal (a-causal) and is primarily useful for interpretation if the peak can be made to coincide with the reflecting boundary.

Any-phasing is used in merging seismic datasets of different vintages and with differing source wavelets.

General shaping groups methods to convert the signal to any desired shape optimally – using some statistical criteria.

Depending upon whether a model is available for decon, the methods could also be divided in *deterministic*, that is, model-based and *statistical*.

Deterministic Deconvolution

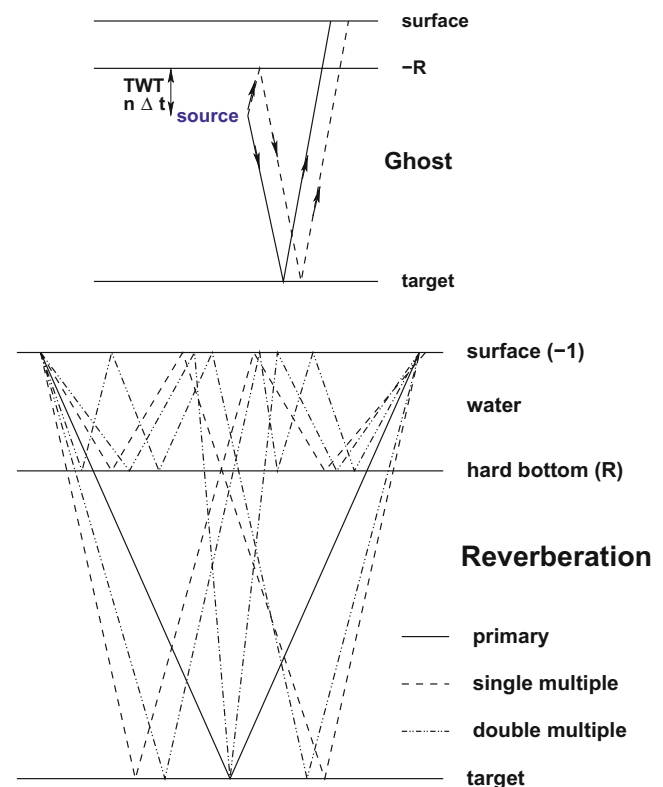
Vibroiseis processing: Vibrators (see the section on “[Seismic Sources](#)”) use a repeatable source signal, called *sweep*. It is a

time-limited (typically, 10–20s long) signal with the frequency continuously varying between given start- and end-values and comes in many flavors, for example, up-, down-, linear-, nonlinear-sweeps. Neglecting other terms, one could write from Eqs. 8 and 9: $o(t) = s(t) * e(t)$. The recorded signal is thus the convolution of *earth reflectivity* with the sweep signal. We could remove its effect (deconvolve) by *cross-correlating* the observed signal with the sweep (which we know precisely), a process, which is equivalent to convolving with its time-reversed version, and get

$$s(-t) * o(t) = s(-t) * s(t) * e(t) \approx \delta(t) * e(t) \approx e(t). \quad (11)$$

Due to the sweep-signal being time-limited, its auto-correlation is not a Delta-spike (ideal), but is a symmetrical (zero-phase) signal called *Klauder wavelet*. The result is thus not quite the desired earth reflectivity (although it has the correct phase) and needs further processing for improvement (see Yilmaz 2001; Liner 2004; Sheriff and Geldart 1995).

De-ghosting: The effect of large RC s in the shallow subsurface has been mentioned earlier. Figure 5 (above) shows one such situation; here the source is placed below the weathering layer, for better energy transmission towards the deeper target (ray going directly downwards). A part of the



Seismic Data Acquisition and Processing, Fig. 5 deterministic deconvolution applied to ghost (above) and reverberation (below). The near vertical ray-paths are shown obliquely for better visualization, see text for details

wave-energy also travels upwards and gets reflected down from the base of the weathering layer (ray going first up, and then down). In certain cases, the $RC_{weathering}$ could be quite large and negative. The energy reflected downwards follows with a short delay behind the direct wave and is called a *ghost*; the observed record is thus corrupted by that caused by a delayed ghost. Removing the latter from the recorded trace is called *deghosting* and is an example of model-based decon. Assuming the TWT between the source and the base of the weathering to be n samples ($= n\Delta t$), one can write: $o(t) = s(t) - Rs(t - n\Delta t)$, or, using Z-transforms,

$$O(Z) = S(Z) - RS(Z) \quad Z^n = S(Z)(1 - RZ^n).$$

$(1 - RZ^n)$ is, clearly, the Z-transform of the ghost-operator. Hence,

$$S(Z) = R(Z)(1 - RZ^n)^{-1}, \quad \text{or,} \quad s(t) = o(t) + s(t - n).$$

The last form above implies recursive filtering in the time-domain to achieve deghosting. Alternately, expanding $(1 - RZ^n)^{-1}$, the inverse-filter operator in time-domain can be written as

$$g(t)^{-1} = (1, 0, 0, \dots + R, 0, 0, \dots + R^2, 0, 0, \dots)$$

De-reverberation The lower part of Fig. 5 shows another situation, where strong reflectivity associated with the water-bottom causes long trains of high-amplitude *reverberation* of signals in the water-layer. The ray-paths shown schematically are: one primary reflection from the target, *two* multiples reflected once in the water layer, and *three* multiples reflected twice; there could be many more, posing a serious problem in marine seismics. Depending upon the depth of water, certain frequencies will, as a result, experience severe distortion (enhancement or suppression). In the simplified case of a water-column with a TWT equal to the sampling interval, and remembering that the negative reflectivity causes phase-change, the total operator (signal + reverberation) can be written as:

$$\begin{aligned} w(t) &= (1, -2R, +3R, \dots) \rightarrow W(Z) \\ &= 1 - 2RZ + 3R^2Z^2 - \dots = (1 + RZ)^{-2}. \end{aligned}$$

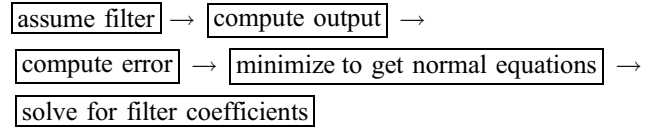
It follows that the deconvolution in this case can be achieved by the operator

$$W(Z)^{-1} = (1 + RZ)^2, \quad \text{or,} \quad w(t)^{-1} = (1, 2R, R^2).$$

This elegant operator is called the Backus filter (see Backus 1959).

Statistical Deconvolution

In the absence of a deterministic model, one could attempt to change the signal wavelet to any desired shape, by designing filters that are optimal in a statistical sense. Based upon work in information theory by Norbert Wiener and others, the applications in seismics were pioneered by the MIT-GAG group, for example, Robinson (1967). Schematically, the basic approach is:



If errors are assumed to be Gaussian, and l_2 norms are used, the operators obtained are called Wiener filters. Such *optimum filters* are used widely, for example, in

- Zero-lag spiking – to increase resolution
- Zero-phasing – to ease interpretation
- Prediction Filtering – to remove multiples which are predictable, the remnant being the prediction error, corresponding to the deeper signal

Wiener optimum filter The normal equations for the filter-coefficients f are given by the matrix equation shown in Eq. 12 in its compact form

$$\phi_{input,input} * f = \phi_{input,output}, \quad (12)$$

which relates the auto-correlation of the recorded (*input*) wavelet to its cross-correlation with the desired (*output*) wavelet. For the derivation of Eq. 12, and a detailed treatment of statistical deconvolution, see, for example, Yilmaz (2001) or Sheriff and Geldart (1995) – an example is shown below to illustrate the approach.

Spiking filter If the wavelets are all n -sample long, the Matrix Eq. 12 can be expanded as

$$\begin{pmatrix} \phi_{i,i}(0) & \phi_{i,i}(1) & \dots & \phi_{i,i}(n-1) \\ \phi_{i,i}(1) & \phi_{i,i}(0) & \dots & \phi_{i,i}(n-2) \\ \phi_{i,i}(n-1) & \phi_{i,i}(n-2) & \dots & \phi_{i,i}(0) \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{pmatrix} = \begin{pmatrix} \phi_{i,d}(0) \\ \phi_{i,d}(1) \\ \vdots \\ \phi_{i,d}(n-1) \end{pmatrix} \quad (13)$$

The auto-correlation matrix $\phi_{i,i}$ in Eq. 12, with the same element in each diagonal descending from left to right, is a

Toeplitz matrix; f is a column vector with the filter-coefficients to be determined and $\phi_{i,o}$ is a column-vector with elements from the cross-correlation matrix. Equations with Toeplitz matrices can be efficiently solved by a procedure called *Levinson recursion*.

Wiener filter: a simple example Given the input wavelet $i_t = (1, -1/2)$, let us find the optimum two-element Wiener-operator to transform it to the desired wavelet $d_t = (1, 0)$, that is, a zero-delay unit-spike. We get, $\varphi_{i,i} = (5/4, -1/2)$, and, $\varphi_{i,d} = (1, 0)$. Equation 13 then becomes

$$\begin{pmatrix} 5/4 & -1/2 \\ -1/2 & 5/4 \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$\text{yielding } f_{\text{Wiener}} = \left(\frac{20}{21}, \frac{20}{21} \right).$$

Applying this filter to the input, we obtain the output $(20/21, -2/21, -4/21)$, which compared to the desired output, gives a squared-error of $1/21$. The ideal filter for this decon is the *inverse filter* for the input wavelet. Writing $I(Z) = 1 - Z/2$ for the Z-transform of the input, the Z-transform of the inverse-filter (which will convert the input to an ideal unit-spike) is $(1 - Z/2)^{-1} = 1 + Z/2 + Z^2/4 + \dots$, which is an infinitely long operator! For an honest comparison of its performance with that of the Wiener-filter, we apply its first two terms to the input, getting the filtered version as $(1, 0, -1/4)$; although looking better at the first glance, its squared error is $1/16$, that is, larger than that of the Wiener-filter! It can be shown that the Wiener filter is the *best* two-element filter for this problem.

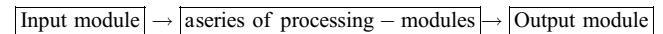
Suppose the input wavelet is $(-1/2, 1)$, that is, not minimum delay, which we want to transform to a zero-delay spike. Normal equations now give the Wiener-filter as $(-10/21, -4/21)$, with the output $(5/21, -8/21, -4/21)$ and the squared error as $6/21$. Inverse filter is now $(-2, -4, -8, \dots)$, which is extremely unstable! Its first two filter-elements give the output $(1, 0, -4)$ with 16 as error! Wiener filter performs here worse than in the first case, because it was trying to convert a maximum-delay wavelet to a minimum-delay spike, but it still does better than the (finite) inverse filter. In this case, if a maximum delay spike $(0, 1)$ was desired, Wiener-filter coefficients would be $(16/21, -2/21)$, giving a filtered output of $(-8/21, 17/21, -2/21)$ with a squared error $4/21$, which is better than that for a zero-lag spike output. Table 5 summarizes the results.

Seismic Data Acquisition and Processing, Table 5 Performance of Wiener- and Inverse-filters

Input wavelet	Desired wavelet	2-point Wiener		2-point inverse	
		Filter	Error	Filter	Error
$(1, -0.5)$	$(1, 0)$	$(20/21, 8/21)$	$1/21$	$(1, 0.5)$	$1/16$
$(-1/2, 1)$	$(1, 0)$	$(-10/21, -4/21)$	$6/21$	$(-2, -4)$	16
$(-1/2, 1)$	$(0, 1)$	$(16/21, -2/21)$	$4/21$		

The Processing Flow: Putting it all Together

Most of the processing *modules (filters)* operate on the data (time-series) sequentially, the entire process resembling a *flow*, though there are a few stand-alone modules too. The operations could be on individual traces (*single-channel*), or on a gather of traces (*multichannel*). Schematically, a seismic processing flow looks like:



Modules have been developed for carrying out specific tasks within the flow, for example, static correction, band-pass filtering, stacking, migration. Usually, there is a choice of modules (algorithms) available for a specific step – each with slightly different characteristics (and artifacts) – and the proper selection of the modules for a flow needs both expertise and experience. This point is illustrated in the Fig. 6, which shows six *different* results of processing the same data.

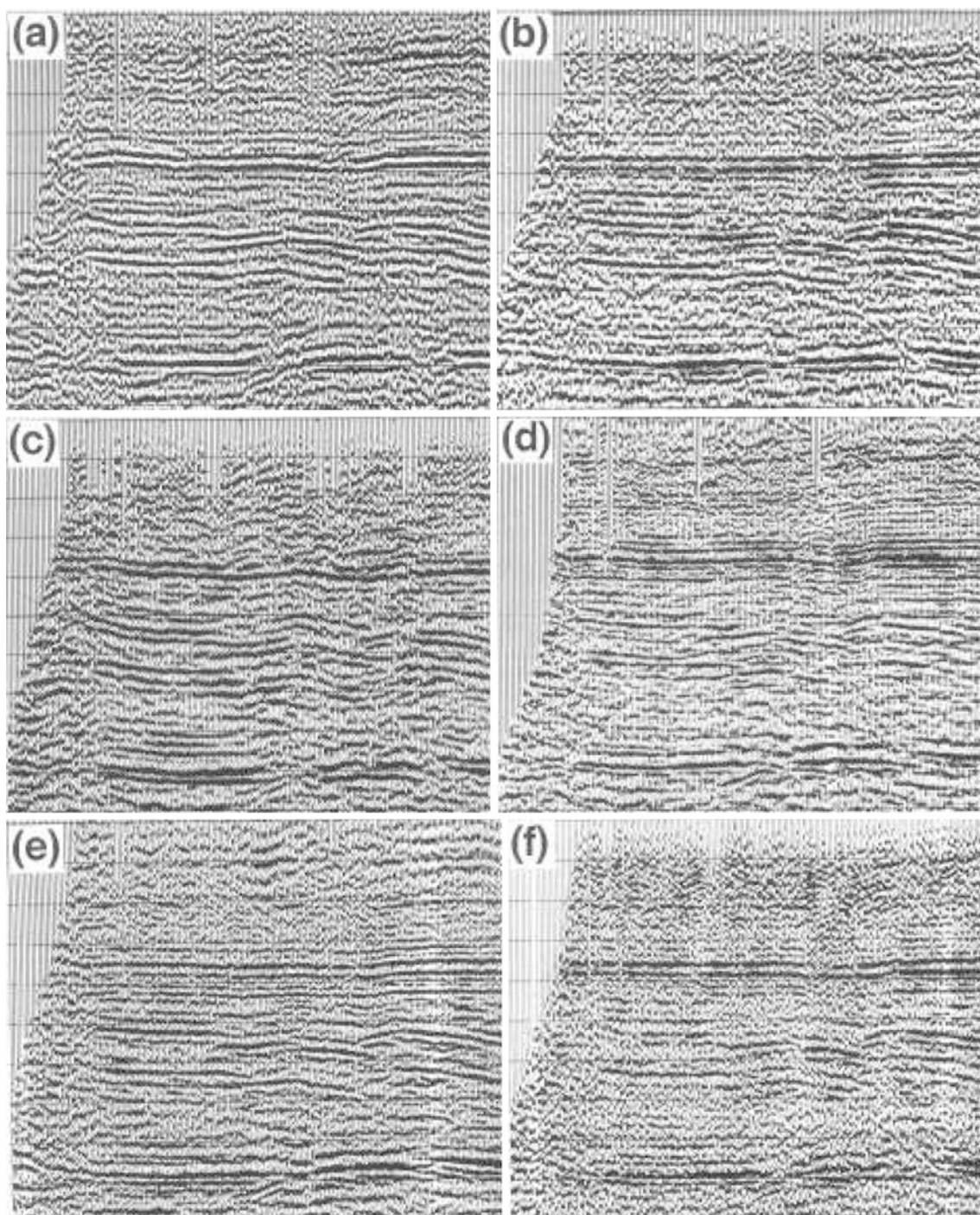
An overview of commonly applied *corrections* (processing module) is shown in Fig. 7. Space constraints will permit us to briefly describe only selected items from this list, which itself is not exhaustive; see Yilmaz (2001) for a more detailed treatment, and ► “[Seismic Imaging, Overview](#)” for additional information. Note that some modules may be applied more than once in the flow, and also, that parts of the flow may be iteratively repeated, till a reasonable result is obtained. The latter shows the importance of quality control (Q/C), by means of visual-display and other (quantitative) tools. The decision as to whether the processing of a dataset is finished depends often on the geological objectives, technical possibilities, and managerial constraints of time and money.

Preprocessing

Editing of seismic traces is an important first step, in view of the largely automated processing sequences later. *Geometry assignment* is also an essential step at this stage and attaches acquisition information to the traces, for example, source- and receiver-coordinates. Each seismic trace is assigned a header, to store such and other information to enable efficient inter-process communication.

Prestack Processing

Static corrections These are time-invariant corrections applied to the traces, due to, for example, elevation differences, and involve up-, or, down-shifting the entire trace in time; an example of their dramatic effect can be seen in Fig. 8. The static effects due to slow lateral changes (long wavelength statics) are particularly difficult to model and can cause imaging problems. *Residual statics* involves *small* time-shifts applied late in the flow to improve the result; it uses the powerful concept of *surface-consistency* to try to correct for near-surface errors that were not modeled properly in the earlier stages. Its implementation by Rothman (1985)



Seismic Data Acquisition and Processing, Fig. 6 Seismic Data Processing has no perfect answer. Seismic cross-sections produced from the same data processed by six different contractors. (Figure from Yilmaz 2001, courtesy SEG and the author)

heralded the use of nonlinear optimization (*simulated annealing, genetic algorithm*) in seismics.

Amplitude corrections Loss of amplitude due to *geometrical spreading* and *absorption* can be corrected for using the theory described earlier; the latter needs a Q-model, in the absence of which empirical relationships based on the total travel path/time are used. A part of the record may be removed

from processing due to the presence of noise, or suspected nonprimaries; depending upon the part of the data-volume removed, one then talks about *top-mute*, *bottom-mute*, or a *generalized mute*. Similarly, a *balancing* (amplitude equalization) may be applied to several adjacent traces to compensate, in an ad hoc manner, for local variations, for example, bad receiver coupling.

Seismic Data Acquisition and Processing, Fig. 7

Components of seismic processing flow

Pre-processing

QC: edit bad traces – assign field-geometry – resample – notch-filter

Pre-stack processing

statics: elevation – datum – residual

amplitude: spreading – Q-compensation – mute – balance

filtering: band-pass – f-k – f-x – τ -p – median – notch

deconvolution: deterministic – predictive – statistical

velocity analysis: CMP sort – const. velocity stack – semblance analysis

stack: NMO correction – DMO correction – CMP stack – AGC – display

Post-stack processing

migration velocity analysis – migration – time-to-depth – display

No-stack processing

migration velocity analysis – prestack depth migration – display

Special processing

True Amplitude (AVO, DHI) – VSP – Anisotropy – image rays

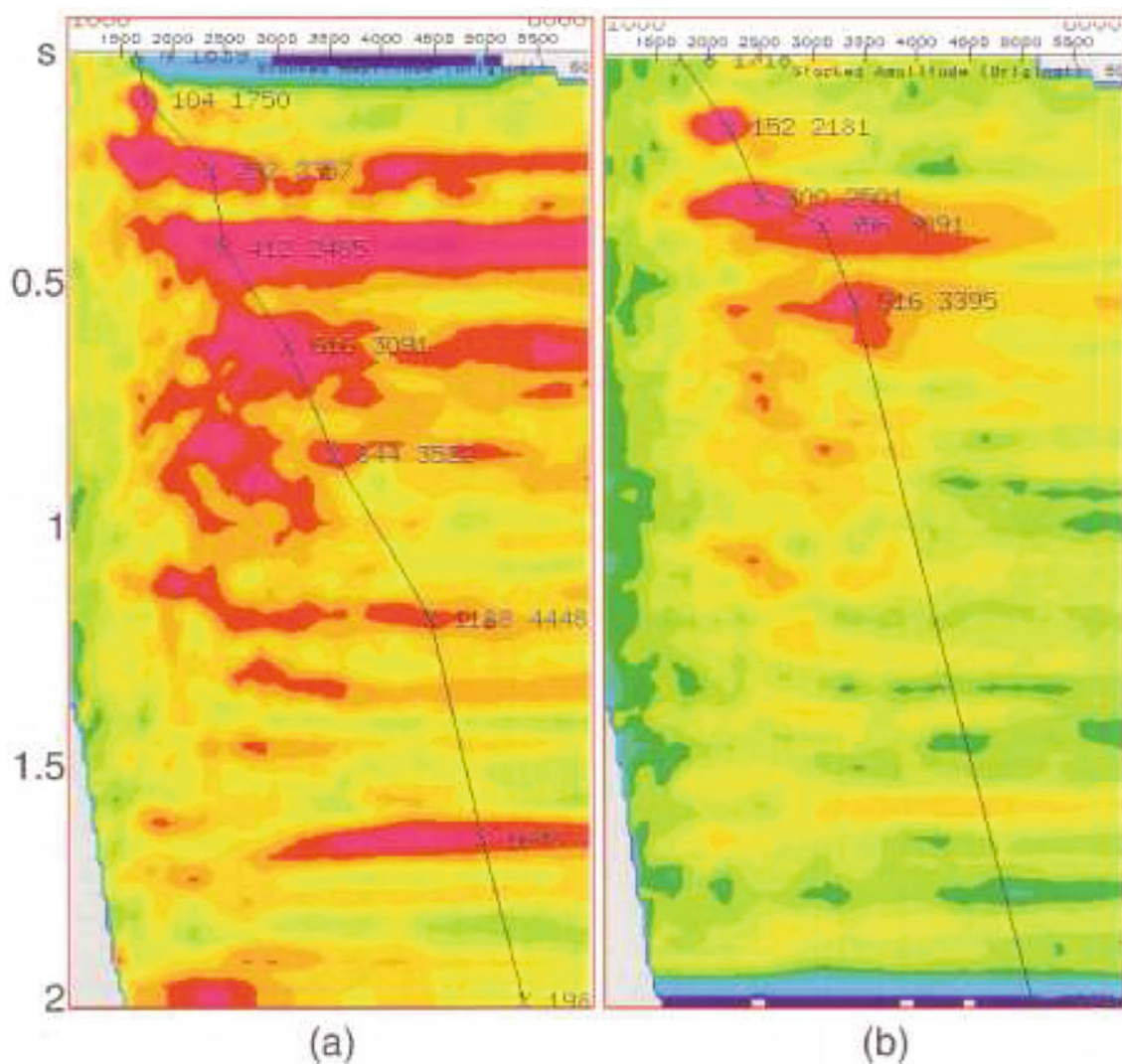
Filtering, sharpness, taper Any process that removes/reduces a part of the suspected *noise* from the data is a *filter*. Frequency-filters (high-cut, low-cut, band-pass) are the simplest examples. Data $f(t)$ is transformed using Fourier theory to its spectrum $F(\omega) = A(\omega) \exp^{-i\omega t}$ in the frequency domain, the amplitudes mainly corresponding to noise are zeroed-out, and the data is transformed back to the time-domain. Development of algorithms for fast and efficient Fourier transform of time-series (FFT) have caused large scale application of digital filters.

Multichannel data enables double-transformation of $f(x, t)$ to $F(\omega, k)$, making filtering possible based upon slopes (apparent velocities) in the ω - k plane; this is particularly effective in eliminating, for example, slow travelling ground-roll (large amplitude surface waves), which often mask the primaries. An example of such filtering is shown in Fig. 9. *Notch* filters are used to remove a narrow band of frequencies, for example, a 50 Hz noise from overhead transmission line. τ -p transforms are useful in filtering multiples, and in untangling far-offset data for velocity analysis, these use the Radon domain for the decomposition (Phinney et al. 1981). A few general comments apply to all filters:

- Filtering is effective only to the extent of signal-noise separation in the transformed domain.
- For any filtering, there is a trade-off between sharp cut-offs in the transform-domain and oscillatory artifacts in time-domain – and vice versa. A compromise solution to this unavoidable problem is to apply *tapers* to smoothen the cut-off and thus minimize edge-effects.

Deconvolution This important aspect has been dealt with in some detail in an earlier section.

Stacking velocity analysis This is almost always carried out in the CMP-domain, after *re-sorting* the data. The aim is to determine the velocity model, that is, $v_{rms}(TWT)$, to be used for computing the best move-out correction for the CMP-gathers. For a suite of velocity models, hyperbolic move-out curves are computed for the range of TWTs of interest; semblances are then computed to determine how well the arrivals in the gather line-up along these curves and displayed in a contour plot in the v_{trial} -TWT domain, allowing interactive picking of an improved velocity model (Fig. 8). The process is repeated at as many CMPs as possible – sometimes grouping neighboring CMPs together for averaging, velocity being a macroscopic property. The result is a laterally varying



Seismic Data Acquisition and Processing, Fig. 8 Stacking velocity analysis using semblance (color contours). Semblance values are shown for a dataset for a range of trial velocities (horizontal- axis) and enable interactive velocity picking as a function of TWT (vertical- axis). The

right panel shows a dramatic improvement in resolution as a result of proper static correction. (Figure from Yilmaz 2001, courtesy SEG and the author)

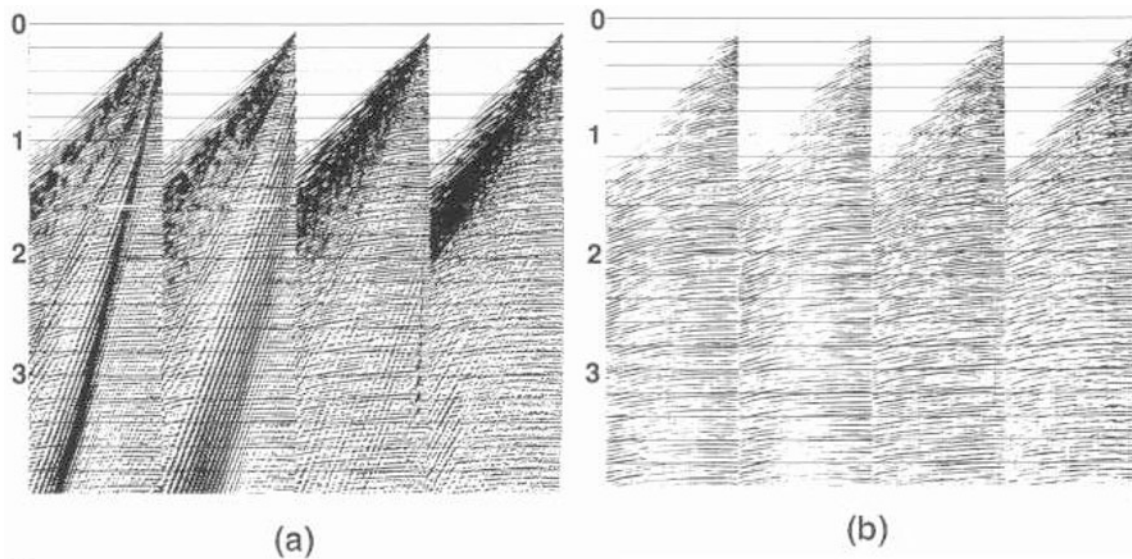
model of v_{rms} . Equation (6) can now be used to infer interval velocities.

CMP-stack Once a reasonable velocity function has been determined, each trace in the CMP-gather (say, CMP_4 in Fig. 1) is shifted in time by subtracting the corresponding move-out corrections. The move-out corrected traces in the CMP-gather are then added (*stacked*) together to produce one trace. This process, *CMP-stack*, reduces random noise – which does not line-up, while strengthening the reflection signal – which does, and thus improves S/N ratio of the data. Note that stacking reduces the data volume too – by a factor of *fold*! Much of the power of the seismic imaging derives from this simple step, which enhances the *primary* reflections (those only once reflected) at the expense of everything else.

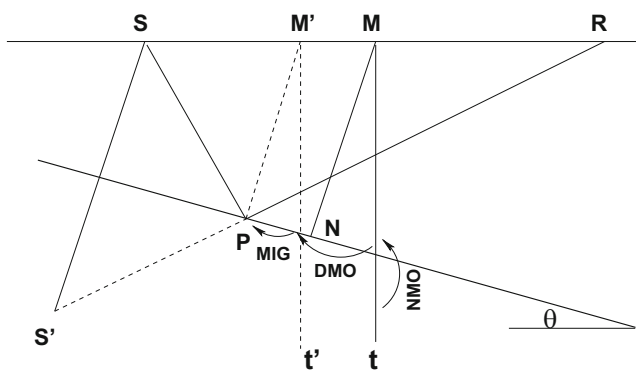
Zero-offset traces/sections The stack traces are also called *zero-offset* traces, the move-out correction having made the source and receiver coincident. A collection of stack traces is a stack- or zero-offset section and represents the first (albeit approximate) 2-D cross-section of the subsurface. For display purposes, CMP-stack sections may be subjected to *automatic gain control* (AGC), an extremely nonlinear time-variant amplitude scaling, to balance weaker/deeper signals and stronger/shallower ones.

Poststack Processing: Positioning Properly

The CMP-stack has one big drawback, and dips were neglected throughout, which is what we are really after. This results in many artifacts in the section, for example, crossing



Seismic Data Acquisition and Processing, Fig. 9 Use of two-dimensional Fourier Transform as an apparent-velocity filter for four marine-seismic records brings out (weaker) reflections. (Figure from Yilmaz 2001, courtesy SEG and the author)



Seismic Data Acquisition and Processing, Fig. 10 Effect of dip in positioning the reflector

layering, diffraction tails. Anticlinal structures are somewhat flattened, and synclinal structures could give rise to *bow-ties*.

Migration Fig. 10 shows the problem schematically in the CMP-domain, for the case of a trace recorded from source S at receiver R from a reflector with a dip θ . After conventional prestack processing, the zero-offset trace would be plotted on the t axis below the mid-point M . This is clearly an error, as the zero-offset ray for M should be incident normally on the reflector – at N , this correction is called *migration* (Yilmaz 2001; Sheriff and Geldart 1995).

Migration steepens and shortens energy alignments and moves these up- dip, clarifying the tangled image. Figure 11 shows an example of a successful migration. For further details, please see ► “[Seismic, Migration](#)”.

DMO Fig. 10 shows yet another error to be considered – the actual reflection point for the source-receiver combination S – R is P , and not N . Even worse, for the different S – R pairs

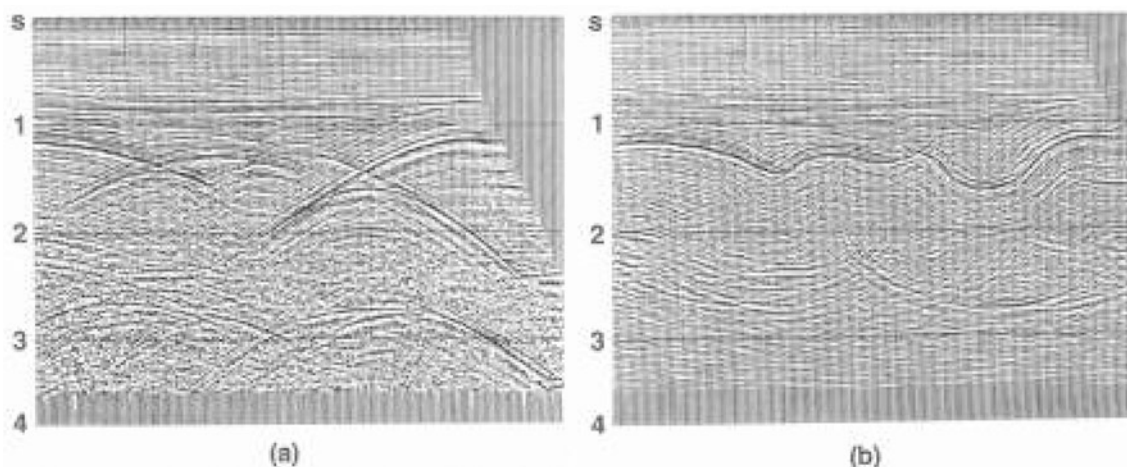
making-up the CMP-gather with mid-point M , the reflection points are all different, that is, are *smeared* along the reflector, the amount of smear being dip-dependent. The process used to correct for this dip-dependent part of the move-out correction is called *DMO*. In practice, this step is applied before migration as indicated in Fig. 7; Figure 10 shows the sequence:

- Reflection time is NMO corrected and plotted *below* M .
- NMO corrected time is DMO corrected and plotted below M' , the *true* zero-offset point.
- NMO + DMO corrected time is MIGRATED and plotted at P , the *reflection point*.

Time-to-depth conversion For final structural interpretation, the TWTs in the seismic section (stacked, or, migrated) need to be converted to depths, with velocity again playing the key role. For a homogeneous medium, this is just a rescaling of the vertical axis; with the velocity varying smoothly only in vertical direction (e.g., for flat sedimentary sequences), a non-uniform stretch of the vertical axis may suffice. Laterally varying velocities present depth-conversion problems though, increasing with the strength of the heterogeneity; ray-bending now needs to be taken into consideration.

No-Stack Processing: Imaging Complex Structures

In the presence of strong lateral velocity variations (e.g., below salt structures), the conceptual model used to process CMP-gathers breaks down. Removing the simplifying assumptions makes the imaging physically more reasonable, albeit at the cost of substantially increased computational effort.



Seismic Data Acquisition and Processing, Fig. 11 Migration positions energy from dipping structures properly. Here bow-tie like artifacts in the top part of panel (a) are imaged back into causative synclinal

structures in panel (b). The artifacts persisting in the bottom of panel (b) probably point to lack of interest in imaging deeper structures in this case. (Figure from Yilmaz 2001, courtesy SEG and the author)

Prestack depth migration and Migration velocity analysis

Simply put, *prestack depth migration* involves tracing the seismic energy from the source to the receiver for every recorded trace, with the philosophy that every seismic trace should be computable *if the structure and the velocity model were both known*. A detailed velocity model is essential for the success of PSDM; often a simplified model is assumed and iteratively improved using *migration velocity analysis* (MVA). For details/issues regarding 2-D vs. 3-D, time- vs. depth- and post-stack vs. prestack migration, see ► “Seismic, Migration” and Yilmaz (2001).

Special Processing

True amplitude – AVO, DHI Observed variations of the *RC* with respect to angle of incidence may be interpreted in terms of changes in lithology across the reflecting boundary (*amplitude* versus *offset*, or, *AVO*) and may even indicate the nature of the pore-fluids. Such *direct hydrocarbon indicators* (DHI) include *bright-spots*, *flat-spots*, *polarity-reversals*, etc. (see Yilmaz 2001; Sheriff and Geldart 1995). A prerequisite for such analyses is *true amplitude processing*, avoiding modules that remove differential amplitude information, for example, balancing, stacking, AGC.

Converted waves Using multicomponent receivers, it is possible to identify waves that have been converted at the reflection boundary and hence possess asymmetrical up- and down-ray-paths. Proper processing of such data, with CCP (*common conversion point*) replacing CDP, provides a better constraint for imaging.

VSP & Cross-well tomography Bore-holes can be used for placing receivers (and sources), resulting in significant noise-reduction. The first processing step now is to separate up- and down-going wavefields, for details, see ► “Vertical Seismic Profiling”.

Anisotropy Many seismic media are *anisotropic*, a common example being shales, which exhibit faster speeds parallel to the layering than across it, and require modification of procedures for proper imaging, for example, the move-out curve would no more be hyperbolic. This field is proving important for reservoir studies too, see Helbig and Thomsen (2005) for an overview and also ► “Magnetic Anisotropy”.

Some Recent Developments

Recording overlapping wavefields Several efforts have been initiated recently to gather more (denser) data without significantly adding to the recording time. Commonly referred to as *simultaneous/continuous sources*, these use rapidly firing sources, for example, airguns/vibrators in marine surveys, and record wavefields overlapping in time (*blending*), which are later separated using various flavors of *deblending*. This results in a considerable increase in spatial bandwidth due to much closer spacing of the sources. Not having to repeatedly reshoot the lines later from the intermediate source locations also leads to significant savings of time (read: *costs*) besides improving data quality by reducing/eliminating unavoidable location errors. Another variation of this approach is “simultaneously” recording sources from different azimuths for both land- and marine-surveys.

Blending/deblending in seismic acquisition mentioned above involves coding/decoding different parts of the recorded wavefield in a manner enabling their later separation. The subject is well known in communications theory; the “Cocktail Party Effect” uses differential characteristics of the desired signal (voice of a single person at a distance) to make it understandable from out of a cacophony of many others including some (loud) near ones (Cherry (1953)).

The related but different issue of intentionally recording overlapping signals in seismics is not new (Garotta 1983), although it is currently undergoing rapid development – both in theoretical and practical aspects (Beasley 2008). Carrying the idea farther, an interesting case has been reported, wherein wavefields generated by individual elements of the airgun firing at rapid but randomized intervals are being continually recorded (Hegna et al. 2019).

Simultaneous sources could be a game changer in the field of seismics due to its positive impact on several fronts – faster, cheaper, safer, and better, for example, Nakayama et al. (2019).

Randomly undersampling the wavefield The importance of properly sampling the seismic wavefield in both temporal and spatial dimensions has been stressed before, as the first step for deriving accurate structural information regarding the subsurface. Sampling in the time domain is usually not a problem albeit at the cost of the increased data volume (by the factor) associated with oversampling with respect to the Nyquist criterion.

Proper sampling in the spatial domain à la Nyquist, however, is a different story and is seldom achieved in a modern acquisition setting, owing partly to the scale of operations – the necessary preprocessing step of interpolation of traces to fill-up the “gaps” has already been mentioned.

The necessity – or otherwise – to adhere to the Nyquist criterion while sampling analog signals has been studied also by signal processing engineers. Recent work in information theory has shown that signals can indeed be recovered even from (severely) undersampled data under certain circumstances. If the signal of interest exhibits a certain structure in a transformed domain, it can be sampled below the Nyquist criterion and recovered in that domain – fortunately, this characteristic is shown by reflection seismic data too.

One of the earliest contributions in this field, specifically for the case of spatial undersampling in reflection seismics, was by Bednar (1996). In it, the possibility of *randomly* sampling the wavefield spatially was discussed and was hinted to be more efficient and economical than uniform undersampling. This is perhaps not surprising, given that imaging reflectors is essentially summing up (integration) of energy along certain specified trajectories (see ► “[Seismic, Migration](#)”), and efficient evaluation of integrals by summing values at random points is known in applied mathematics.

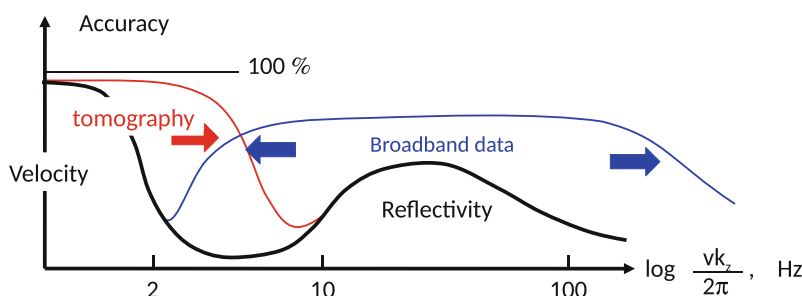
In practice, *random* undersampling of the signal is supposed to perform superior to *regular* undersampling as it will convert *aliased* energy (due to undersampling) into incoherent noise, which can then be easily filtered out. Hennenfent and Herrmann (2008) showed *jittery* undersampling (with respect to the Nyquist criterion) to perform even better, as it also enables control of the omnipresent gaps that need to be filled by interpolation for later reconstruction. Herrmann (2010) discusses this idea of randomized undersampling

further. Developed from the technique of *compressive sampling*, commonly used in modern electronics, e.g., digital camera and magnetic resonance imaging, it is shown to lead to a significant reduction in data volume and hence to be able to master the “curse of dimensionality” – a phrase which points to the exponential increase in data volume with each added dimension of seismic acquisition. It is again shown to be particularly effective, if the signal is sparse in a suitable transform-domain.

Further studies about the applicability of this technique to related fields, for example, ► “[Seismic, Waveform Modeling and Tomography](#)” are subjects of active current research. One may dare to state that the Nyquist sampling criterion has at last been conquered – at least as far as seismics is concerned!

Noise as a seismic source As hinted in the first edition, it is now an exciting and established field of research under the general banner of *seismic interferometry & daylight imaging*. Starting from early conjectures (Claerbout 1968) and later breakthroughs (Fink 1993, 1997), there have been both theoretical and practical advances understanding its possibilities and limits. Although this topic belongs to some other essays in this volume (► “[Seismic, Ambient Noise Correlation](#)”, ► “[Seismic Noise](#)”), a very brief introduction follows. The basic idea is that “noise” – also in seismics – contains useful information, which can be extracted by – and this is somewhat surprising – some pretty straightforward processing. Noise recorded at two locations can be used to obtain relevant medium properties in the intervening medium. This approach can also be used to place virtual sources anywhere in the medium. For an introduction to this fascinating subject, see Curtis et al. (2006); Wapenaar and Snieder (2007); Wapenaar et al. (2010a, b); Snieder and Wapenaar (2010).

Imaging vis-a-vis Inversion Imaging tries to obtain useful (drillable) structural information using large data-redundancy and simple conceptual models, whereas inversion aims at getting values for the physical parameters of the medium, using more involved theory, see ► “[Seismic Imaging, Overview](#)”. Buske et al. (2009) give a good introduction to the two approaches, whereas Weglein et al. (2009) discuss some critical underlying questions. Till recently, resolution in seismic imaging suffered from an information-gap as pointed out by Claerbout (1985b): velocity information was being obtained by kinematic analyses for apparent frequencies below ca. 2 Hz and reflectivity was being derived for frequencies above ca. 10 Hz. Recent improvements in tomographic velocity analysis and broadband data acquisition have resulted in increasingly narrowing this gap, which results in improved resolution (Nichols 2012). As shown in the Fig. 12, there is now even an information overlap between the results obtained from these two processing techniques. This, in turn, gives rise to interesting questions regarding the consistency of the high- and low-frequency results that have finally to be combined.



Seismic Data Acquisition and Processing, Fig. 12 Sketch showing the improvement in resolution in reflection seismics. In black, the famous sketch by Claerbout (1985b), showing the gap between information (accuracy) from velocity analysis and reflectivity. Advances in

tomography (red) and broadband acquisition (blue) have changed the situation essentially to that of an informational overlap. (Figure from Lambare et al. (2014), courtesy EAGE and the author)

Full wave-form inversion Obtaining the visco-elastic properties of the medium so as to be able to reproduce each seismic/seismological trace *completely* remains the ultimate goal of inversion of seismic/seismological data (see ► “Seismic, Waveform Modeling and Tomography”). This subject is currently the subject of major theoretical and applied research, the latter also profiting from considerable increase in available computational power. In the beginning, efforts were limited to achieving a reasonable match between a selected part (phase) of the observed waveform and the computed synthetic. Given improvements in acquisition, theory, and the computer power, progress is being made to reproduce larger parts of the entire wavetrain – a complete reproduction may though be never achievable. For a quick introduction to this topic, including underlying problems, and related developments in exploration geophysics, see Virieux and Operto (2009); Virieux et al. (2017); Brittan and Jones (2019). Note that *perfect inversion implies perfect imaging – and vice versa!*

Summary

The simple *echo-in-the-well* experiment mentioned at the start needs many physico-mathematical supports to analyze data obtained from the earth’s subsurface. Starting at data acquisition, these acquisition/processing modules yielding the final image resemble a pipeline (flow). Several of these have been explained briefly; for others, cross-references elsewhere in this volume have been provided.

Cross-References

- Energy Partitioning of Seismic Waves
- Magnetic Anisotropy
- Ocean Bottom Seismics
- Propagation of Elastic Waves: Fundamentals
- Seismic Diffraction

- Seismic Imaging, Overview
- Seismic Instrumentation
- Seismic Noise
- Seismic Properties of Rocks
- Seismic Ray Theory
- Seismic Viscoelastic Attenuation
- Seismic Waves, Scattering
- Seismic, Ambient Noise Correlation
- Seismic, Migration
- Seismic, Waveform Modeling and Tomography
- Single and Multichannel Seismics
- Vertical Seismic Profiling

Bibliography

- Aki K, Richards PG (2002) Quantitative seismology, 2nd edn. University Science Books, CA, USA
- Backus M (1959) Water reverberations – their nature and elimination. *Geophysics* 24(2):233–261
- Beasley C (2008) Simultaneous sources: a technology whose time has come. *Soc Expl Geophys. Expanded abstracts*, 2796–2800
- Bednar J (1996) Coarse is coarse of course unless *The Leading Edge* 15(6):763–764
- Brittan J, Jones I (2019) FWI evolution – from a monolith to a toolkit. *The Leading Edge* 38(3):179–184
- Buske S, Lecomte I, Nemeth T, Operto S, Sallares V (2009) Imaging and inversion – introduction. *Geophysics* 74(6):WCA1–WCA4
- Cherry E (1953) Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am* 25(5):975–979
- Claerbout JF (1968) Synthesis of a layered medium from its acoustic transmission response. *Geophysics* 33(2):264–269
- Claerbout JF (1985a) Fundamentals of geophysical data processing. Blackwell. <http://sepwww.stanford.edu/sep/prof/fgdp5.pdf>
- Claerbout JF (1985b) Imaging the earth’s interior. Blackwell. <http://sepwww.stanford.edu/sep/prof/iei2/>
- Curtis A, Gerstoft P, Sato H, Snieder R, Wapenaar K (2006) Seismic interferometry – turning noise into signal. *The Leading Edge* 25(9):1082–1092
- Dean T, Tulett J, Barnwell R (2018) Nodal land seismic acquisition: the next generation. *First Break* 36:47–52
- Dennison AT (1953) The design of electromagnetic geophones. *Geophys Prosp* 1(1):3–28
- Dix C (1955) Seismic velocities from surface measurements. *Geophysics* 20:68–86

- Fink M (1993) Time reversal mirrors. *J Phys D Appl Phys* 26:1333–135. <http://iopscience.iop.org/0022-3727/26/9/001>
- Fink M (1997) Time reversed acoustics. *Phys Today* 50:34–40. <https://doi.org/10.1063/1.881692>
- Garotta R (1983) Simultaneous recording of several vibroseis® seismic lines. *Soc Expl Geophys. Expanded abstracts*, 308–310
- Gülünay N (2003) Seismic trace interpolation in the Fourier transform domain. *Geophysics* 68(1):355–369
- Hegna S, Klüver T, Lima J, Wisloff J (2019) Making the transition from discrete shot records to continuous seismic records and source wavefields, and its potential impact on survey efficiency and environmental footprint. *Geophys Prospect* 67(6):1472–1485
- Helbig K, Thomsen L (2005) 75-plus years of anisotropy in exploration and reservoir seismics: a historical review of concepts and methods. *Geophysics* 70:9ND–23ND
- Hennenfent G, Herrmann F (2008) Simply denoise: Wavefield reconstruction via jittered undersampling. *Geophysics* 73(3):V19–V28
- Herrmann F (2010) Randomized sampling and sparsity: getting more information from fewer samples. *Geophysics* 75(6):WB173–WB187
- Lambare G, Guillaume P, Montel J (2014) Recent advances in ray-based tomography. In: *Extended abstracts, 76th annual meeting, EAGE*
- Lee M (2008) Comparison of the modified Biot-Gassmann theory and the Kuster-Toksöz theory in predicting elastic velocities of sediments. U.S. Geological survey scientific investigations Report
- Liner C (2004) *Elements of 3D seismology*. PennWell corporation, OK
- Liu Y, Fomel S (2011) Seismic data interpolation beyond aliasing using regularized nonstationary autoregression. *Geophysics* 76(5):V69–V77
- Menke W (1989) *Geophysical data analysis*, rev edn. Academic, New York
- Nakayama S, Blacquièrre G, Ishiyama T, Ishikawa S (2019) Blended-acquisition design of irregular geometries towards faster, cheaper, safer and better seismic surveying. *Geophys Prospect* 67(6):1498–1521
- Nichols D (2012) Resolution in seismic inversion. In: *Expanded abstracts, 74th annual meeting – workshop, EAGE*
- Phinney R, Roy Chowdhury K, Frazer LN (1981) Transformation and analysis of record sections. *J Geophys Res* 86(B1):359–377
- Robinson E (1967) Predictive decomposition of time series with application to seismic exploration. *Geophysics* 32:418–484
- Robinson E (2005) The MIT Geophysical Analysis Group (GAG): 1954 and beyond. *Geophysics*. <https://doi.org/10.1190/1.2000287>
- Robinson E, Treitel S (1964) Principles of digital filtering. *Geophysics* 29:395–404
- Rothman D (1985) Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics* 50(12):2784–2796
- Sheriff RE, Geldart LP (1995) *Exploration seismology*, 2nd edn. Cambridge University Press, Cambridge, UK
- Snieder R, Wapenaar K (2010) Imaging with ambient noise. *Phys Today* 63(9):44–49
- Treitel S (2005) The MIT Geophysical Analysis Group (GAG): 1954 and beyond. *Geophysics*. <https://doi.org/10.1190/1.1993707>
- Vermeer G (1990) Seismic wavefield sampling. Society of Exploration Geophysicists, Tulsa
- Vermeer GJ (2002) 3-D seismic survey design. Society of Exploration Geophysicists, Tulsa
- Virieux J, Operto S (2009) An overview of full-waveform inversion in exploration geophysics. *Geophysics* 74(6):wcc1–wcc26
- Virieux J, Asnaashari A, Brossier R, Métivier L, Ribodetti A, Zhou W (2017) An overview of full-waveform inversion in exploration geophysics. In: *Encyclopedia of exploration geophysics, SEG*. pp R1–R1–40. <https://doi.org/10.1190/1.9789560803027.entry6>
- Wapenaar K, Snieder R (2007) Chaos tamed. *Nature* 447:643
- Wapenaar K, Draganov D, Snieder R, Campman X, Verdel A (2010a) Tutorial on seismic interferometry, part I. *Geophysics* 75(5):75A195–75A209
- Wapenaar K, Slob E, Snieder R, Curtis A (2010b) Tutorial on seismic interferometry, part II. *Geophysics* 75(5):75A211–75A227
- Weglein A, Zhang H, Ramírez A, Liu F, Lira J (2009) Clarifying the underlying and fundamental meaning of the approximate linear inversion of seismic data. *Geophysics* 74(6):WCD1–WCD13
- Yilmaz O (2001) *Seismic data analysis, processing, inversion and interpretation of seismic data*, 2nd edn. Investigations in geophysics, vol I. Society of Exploration Geophysicists, Tulsa
- Zwartjes P, Sacchi M (2007) Fourier reconstruction of nonuniformly sampled, aliased seismic data. *Geophysics* 72(1):V21–V32

Seismic Diffraction

Enru Liu

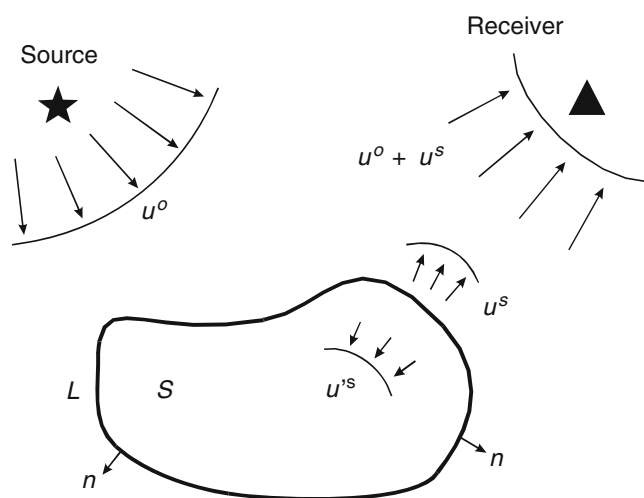
China University of Mining and Technology, Xuzhou, China

Definition

- Diffraction** Redistribution in space of the intensity of waves resulting from the presence of an object. It is also referred as the penetration of wave energy into areas forbidden by geometrical optics, e.g., the bending of wave energy around obstacles without obeying Snell's law as explained in Huygens' principle (generation of secondary sources).
- Diffraction wave-field** An event observed on seismic data produced by diffracted energy, resulting at the termination of reflectors (as at faults and other abrupt changes in seismic impedance), and it is characterized on seismic records and sections by a distinctive alignment.
- Diffraction tomography** An inverse technique that is used in seismic exploration to reconstruct the physical properties under investigation using wave-equation propagation.

Introduction

When a wave (elastic wave, electromagnetic wave, or sound wave) meets an object, particle, or obstacle, it is diffracted due to scattering of energy of the propagating wave (Fig. 1). However, in the literature, the terms of diffraction and scattering are often used interchangeably, and it can be sometimes confusing. Scattering and diffraction are two physical phenomena that any kind of waves can experience, but they are not the same thing. Scattering is understood in terms of particles, and behaves similarly for waves. Scattering is effectively bouncing off something. For waves, it is being absorbed and then almost immediately released in another direction. Scattering occurs because an object gets in the way of the wave. The part of the wave that strikes the object



Seismic Diffraction, Fig. 1 Problem configuration: a scattering object S bounded by the curve L with outward normal n . Upon an incidence of u_0 located at source, the total wave-field received at receiver is the superposition of the incident wave-field u_0 and the scattered wave-field u_s .

must either pass through (e.g., light through glass), be absorbed (sunburn), or be scattered (light bouncing off the wall, so we can see the wall). Diffraction is due to part of a wave being removed. It is an action taken by the part of the wave that does not strike an object. Here is an example of diffraction: Imagine a straight wave traveling forward along the surface of water. If you block the left half of the wave, the right half will not just keep moving forward, and it will expand toward the left, toward where the blocked wave would have been. A wave seems to continuously regenerate itself, always pushing forward on itself. When a section is removed, parts of the wave get pushed into the empty spot. This, in some ways, correlates to your pushing a wide row of blocks. If many people push on a few blocks each, a straight line can be maintained. If one person tries to do so, the blocks in front will tend to spread out.

In a sense, diffraction and scattering refer to a wave being redirected as a result of interacting with objects. However, a more precise definition used in optics distinguishes these two. In optics, scattering generally implies interaction of waves (or photons) with spatially uncoordinated (unordered) atoms, i.e., wave being scattered off the particles and surfaces. This means that if one looks at a picture of the scattered radiation, one would see a spatially continuous footprint. Diffraction, on the other hand, occurs when the object or part of the object is made up of ordered atoms. These atoms, being neatly arranged, “scatter” the waves or photons in a coordinated way, i.e., in specific directions, giving rise to what we can see on a film as bright spots rather. In other words, diffraction is a special type of scattering that leads to large-scale interference effects. Usually this is because the surface causing the scattering has some overall organization, such as a ruled diffraction grating or the knife-edge of a slit.

Although scattering and diffraction are not logically separate, the treatments tend to be separated, with diffraction being associated with departures from geometrical optics caused by the finite wavelength of the waves. Thus, diffraction traditionally involves apertures or obstacles whose dimensions are large compared to a wavelength. To lowest approximation the interaction of waves is described by ray tracing (geometrical optics). The next approximation involves the diffraction of the waves around the obstacles or through the apertures with a consequent spreading of the waves. Simple arguments show that the angle of diffraction of the waves are confined to the region $\theta \leq \lambda/a$ where λ is the wavelength and a is linear dimension of the aperture or obstacle (approximations considered work well if $\lambda/a \leq 1$). Note that diffraction may cause the localization of seismic waves – a phenomenon that is similar to the localization of lights in crystal (Larose et al. 2004). This phenomenon is caused by the focusing and defocusing of energy when seismic wave propagates through media with distributions of periodical or random distribution of scattering bodies.

In order to gain some perspective on these two extremely complex phenomena (diffraction and scattering), various theories and models have been developed in physics, and the limitations and validity of these theories are controlled by two ratios: object dimension a to wavelength (a/λ) and path-length L to wavelength $\lambda(L/\lambda)$. In contrast to other branches of physics, in geophysics path-length is also important as we are interested both in near field as well as far field (and often near and far fields are treated differently). We also often use the dimensionless parameters ka and kL (where $k = \omega/\lambda$ is the wave number, ω is frequency).

Diffraction Theories

Application of any diffraction model can be divided into two separate tasks. First, one must obtain the fields exiting a diffracting object (i.e., the near fields, or the boundary field values), or a reasonable approximation thereof. The second step involves propagating those fields to the desired observation point. These are distinct and separate parts of the diffraction models. Most texts do not make this separation clear. Instead, the boundary value assumptions and subsequent propagation into the far field are lumped together into one theoretical treatment. If the resulting diffraction pattern is at all inaccurate, it is difficult to determine how much of that error is due to incorrect boundary fields and how much is the result of the propagation calculation. Because of this, it is often difficult to know which model is appropriate for a particular problem.

There are a number of different models to compute diffraction wave-field due to wave scattering, including the Huygens’ Principle, the Rayleigh-Sommerfeld theory, the

Kirchhoff's diffraction theory, Taylor series perturbation theory or the high-order Born approximation, Rytov phase approximation, and a model referred to as angular spectrum of plane waves. The well-known Fraunhofer and Fresnel approximations, as they appear in most introductory texts, are derived from the Kirchhoff model. Several other methods are also available for treating diffraction problems: discrete wave-number techniques; generalized ray techniques; and various numerical methods (finite difference methods; finite element method; and boundary integral or element methods). Each of these theoretical methods and models is based on some assumptions and has its strengths and weaknesses, and each can be satisfactorily employed for some ranges of problems. The choice of an appropriate model is based on what is known about a specific problem.

If there are several objects or particles, the scattering from one object will induce further scattered fields from all the other objects, which will induce further scattered fields, from all the other objects, and so on. This process is called multiple scattering. Not all theories are applicable to multiple scattering problems (in physics, this is called many-body problem). The Kirchhoff approximation ignores multiple scatterings between any two surface points. In general, it has been considered valid for the large-scale objects. Perturbation theory based on the Taylor series expansion, also sometimes called the high-order Born approximation, is valid for the small-scale objects whose dimensions are less than a wavelength (or objects whose physical properties are not too different from background solids). The Rytov phase approximation to large-scale object is not subject to the stringent restrictions that apply to the Kirchhoff approximation. Studies have shown that the Rytov approximation improves the Kirchhoff approximation in both amplitude and phase. To some degrees, the high-order Born series approximation can account for multiple scattering between closely-spaced objects. For instance, the second-order Born approximation might be sufficient to guarantee the accuracy for general rough surfaces without infinite gradients and extremely large surface heights. In contrast to other branch of field, e.g., optics, in seismology two kinds of waves exist, compressional and shear waves. These two waves can convert to each other when one meets an object. When multiple objects exist, the conversion and interaction between different wave types due to multiple scattering can be very complex. Therefore, care must be taken when one uses any diffraction theory to solve specific geophysical problems.

Geometrical Theory of Diffraction

The geometrical theory of diffraction (GTD) is an extension of geometrical optics that accounts for wave diffraction by edges. It was introduced in 1953 by Keller (the most

commonly cited reference was published by Keller in 1962). The geometrical theory of diffraction was devised to eliminate many of the problems associated with geometrical optics. The strongest diffracted fields arise from edges, but ones of lesser strength originate from point discontinuities (tips and corners). The total field $u = (u_1, u_2, u_3)$ at an observation point $\vec{x}$ is decomposed into geometrical optic rays (the incident or reference field) u_i^0 and diffracted components u_i^d

$$u_i(\vec{x}) = u_i^0(\vec{x}) + u_i^d(\vec{x}). \quad (1)$$

The behavior of the diffracted field is based on the following postulates of GTD:

1. Wavefronts are locally plane waves.
2. Diffracted rays emerge radially from an edge.
3. Rays travel in straight lines in a homogeneous medium.
4. Polarization is constant along a ray in an isotropic medium.
5. The diffracted field strength is inversely proportional to the cross sectional area of the flux tube.
6. The diffracted field is linearly related to the incident field at the diffraction point by a diffraction coefficient (see Achenbach et al. 1982, for various analytic solutions).

GTD is a high frequency method for solving wave scattering problems from large-scale discontinuities or discontinuities in more than one dimension at the same point, and it uses ray diffraction to determine diffraction coefficients for each diffracting object-source combination. These coefficients are then used to calculate the field strength and phase for each direction away from the diffracting point. These fields are then added to the incident fields and reflected fields to obtain a total solution. Multiple scattering wave-fields cannot be easily computed using GTD.

Kirchhoff Approximation

In the Kirchhoff representation of diffracted wave-fields, the i th component of diffracted wave-field u_i is computed using

$$u_i^d(\vec{x}) = - \int_{\Sigma} [u_k(\vec{x})] c_{kjpq}^0 \left[\frac{\partial G_i^p(\vec{x}, \vec{X})}{\partial x_q} \right] n_j dS_{\vec{x}}, \quad (2)$$

where n_j is the j th component of the normal n to the surface of scattering object, $\vec{X}$ is a point on the face of the scattering object, $[u_k]$ is the k th displacement discontinuity across the object in the direction of n (object normal), and c_{kjpq}^0 is the elastic tensor of the background, which are often assumed to be isotropic. Equation 2 provides a means of evaluating the diffracted field so long as the displacement discontinuity $[u]$

on the object can be estimated accurately. Since the amplitudes and phases of $[u]$ are unknown, in the Kirchhoff approximation, these are taken to be the same as if the object were infinitely long, that is, the effect of the boundary is ignored (see Douglas and Hudson 1990; Liu et al. 1997). Therefore, the Kirchhoff approximation is a high frequency approximation, which is only applicable to diffraction from objects whose dimension is larger than the wavelength.

Perturbation Theory: Born and Rytov Approximation

The diffracted wave-field in Eq. 1 can be written as an infinite series of the Taylor series expansion and is derived by assuming that the physical property of scattering objects is written as a perturbation of background media. The Born approximation consists of taking the incident field in place of the total field as the driving field at each point in the scatterer. It is the perturbation method applied to scattering by an extended body. It is accurate if the scattered field is small, compared to the incident field, in the scatterer. It is only valid for weak scattering or when the obstacles are small compared to both the wavelength and the spacing between the objects. Clearly, it has serious limitations when dealing with large-scale objects. The simplest approximation, called single scattering or the first-order Born approximation, is to ignore the multiply scattered field between objects. This approximation has been widely used in geophysics (Hudson and Heritage 1981; Wu 1982, 1989).

The Rytov approximation linearly perturbs the phase of a wave-field with respect to model parameters such as velocity whereas the Born approximation perturbs the amplitude. When the Green's functions for point sources are replaced by Eikonal approximations, the Rytov perturbed wave-field becomes a scaled, differentiated, time-delayed version of the reference or incident wave-field.

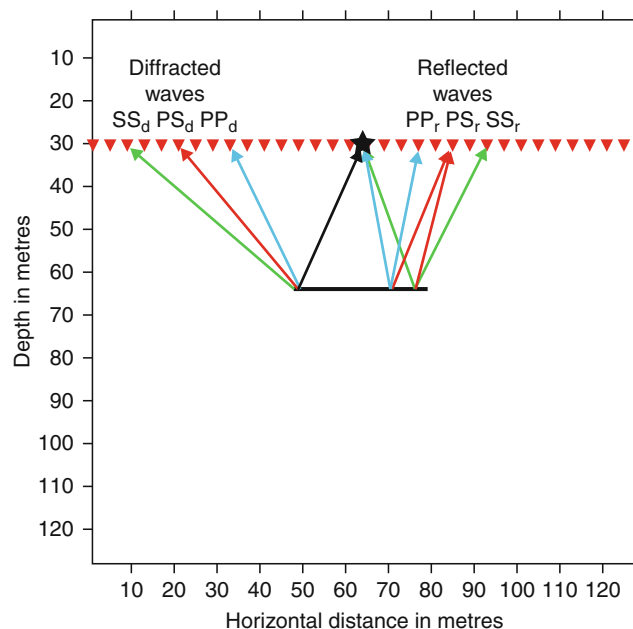
Numerical Methods Used to Compute Diffraction Wavefield

Various numerical methods can also be used to compute diffracted wave-fields, particularly now, when highperformance computers are widely available. In geophysics, finite difference methods have been used widely in the study of scattering of elastic waves by crustal heterogeneities with continuous variation of physical properties and they have also been used to model scattering by thin cracks and large fractures (see Coutant 1989; Fehler and Aki 1978; Coates and Schoenberg 1995; Vlastos et al. 2003, 2007). Elastodynamic boundary integral equation or boundary element method has also been widely used to compute wave-fields from discrete inclusions with various spatial distributions as well as rough-surface topographic variations

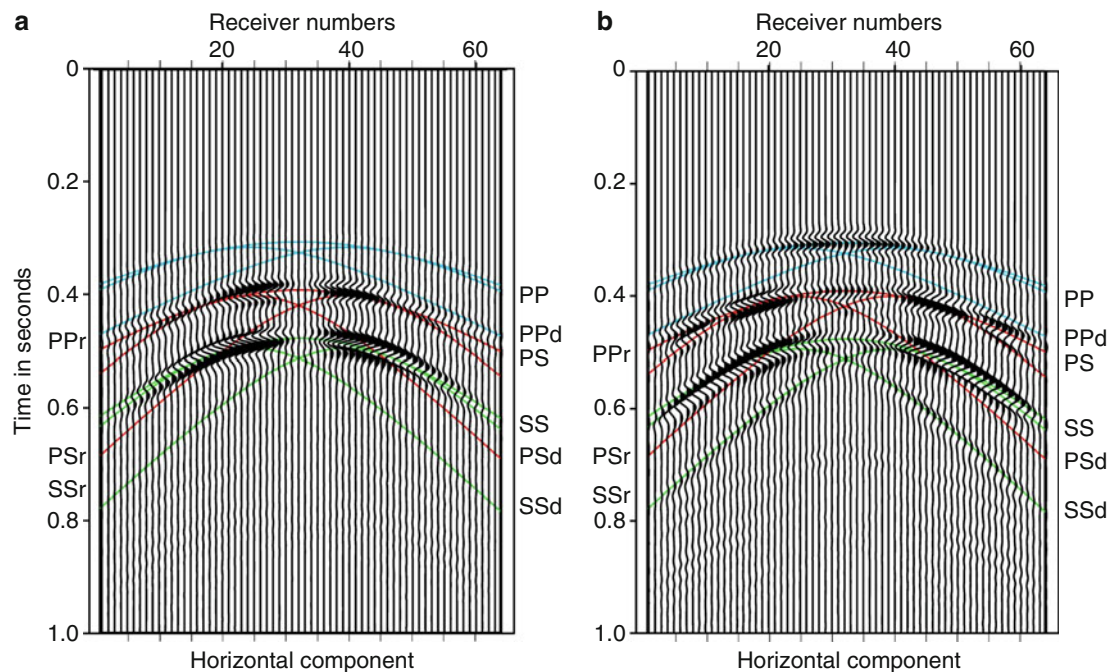
(see Bouchon 1987; Pointer et al. 1998; Liu and Zhang 2001; Sanchez-Sesma and Campillo 1991, 1993).

An Example of Seismic Diffraction from a Single Fracture

Some practical applications of diffraction theories in seismology include scattering from cavities; topographic variation, fractures, cracks (Fig. 2). Here we give three examples to demonstrate the application of various theories in tackling diffraction problems. An example is given here for diffraction from a single fracture as computed using the finite difference method (Vlastos et al. 2003, 2007). The model geometry is shown in Fig. 2. The source, receivers, and fracture are situated in an ideal elastic full space ($V_p = 3300$ m/s, $V_s = 1800$ m/s, density $\rho = 2.2$ g/m³). The receiver array at which vertical and horizontal particle displacements are recorded is horizontal and 340 m above the fracture. The fracture is 300 m long. The source is located at the center of the receiver array. The source type is a vertical force. The source signal is a Ricker wavelet with a peak frequency of 25 Hz and a pulse initial time of 0.1 s. Figure 3 also shows the different kinds of waves generated by the interaction of the waves generated by the source and the fracture. The source generates both P and S waves. When they reach the fracture boundary those waves are reflected and we have PPr, PSr, SPr, and SSr waves. We calculate the theoretical ray travel-times and overlap them on the synthetic seismograms. Figure 3a, b



Seismic Diffraction, Fig. 2 Schematic representation showing diffraction from a fracture, and representation of the ray paths of the different kind of waves generated by the source that interact with the fracture



Seismic Diffraction, Fig. 3 Comparison between the theoretical ray travel-times and the synthetic seismograms generated by the modeling method: (a) the horizontal (x) and (b) the vertical (z) components

show the horizontal (x) and the vertical (z) components, respectively, of the synthetic seismograms together with the theoretical ray travel-times. As we can see from both figures, we have very good agreement between the theoretical ray travel-times and the synthetic seismograms. All types of waves are accurately represented in the synthetic seismograms. Owing to the type of source that we implement, we have strong arrivals at short offsets on the horizontal component and strong arrivals at long offsets on the vertical component. In addition to that, the diffracted waves from the tips of the fracture and the PPr and PPd waves are not visible in the horizontal component, but they are very clearly demonstrated in the vertical component and follow the theoretical travel-times. This is expected because the source causes vertical displacements on the medium, so very close to the source and very far away from it, the horizontal displacement is negligible. Another aspect of the comparison between the theoretical and the modeled data is that they give us further insight into the waveform patterns. For instance, we can see in both Fig. 3a, b that in the areas of superposition between the reflected waves from the fractures and the diffracted waves from the tips we have maximum amplitude in the wave-field, as a result of constructive interference. This gives us valuable information concerning the medium we are examining.

Diffraction Tomography

Seismic tomography is emerging as an imaging method for determining subsurface structure. When the viewangle

coverage is limited and the scale of the medium inhomogeneities is comparable with the wavelength, as is often true in geophysical applications, the performance of ordinary ray tomography becomes poor. Other tomographic methods are needed to improve the imaging process, e.g., diffraction tomography. It has been widely used in surface reflection profiling (SRP), vertical seismic profiling (VSP), and cross-hole measurements. Theoretical formulations are derived by Wu and Toksoz (1987) for two-dimensional geometry in terms of line sources along a source line and line receivers along a receiver line. The theory for diffraction tomography is based on the Born or Rytov approximation. Multisource holography, which is similar to Kirchhoff-type migration, often gives distorted images of the object. This distortion causes long tails of the image in the case of SRP and a strong noise belt in the case of VSP and is due to incomplete and nonuniform coverage of the object spectrum. The filtering operation of diffraction tomography helps in correcting the nonuniform coverage (including duplication) of the object spectrum in the reconstruction process and therefore reduces the distortions. On the other hand, multisource holography is better suited for imaging sharp boundaries with large acoustic impedance contrasts since diffraction tomography is restricted, to weak inhomogeneities. In addition, multisource holography has the flexibility to be used with an arbitrary number of sources (including a single source). Its sampling interval is not restricted by the Nyquist frequency. Numerical examples show that combined data sets (such as surface reflection data combined with VSP data or cross-hole data combined with surface data) improve the image quality.

Summary

Diffraction refers to the spatial distribution of the intensity of seismic waves resulting from the presence of an object (e.g., a hill- or valley-like topographic feature on the surface, or a fracture, cavity, or cave in the subsurface). It is also referred as the penetration of wave energy into areas forbidden by geometrical optics, e.g., the bending of wave energy around obstacles without obeying Snell's law as explained in Huygens' principle (as secondary sources). In the geophysical literature, the words of diffraction and scattering are often used interchangeably and it can be confusing. Diffraction and scattering are two different physical phenomena, but they are related to each other. Several analytic diffraction theories have been developed, e.g., geometrical theory of diffraction and Kirchhoff diffraction theory. More recently, numerical methods, such as finite difference and boundary element or boundary integral methods, are becoming increasingly used by geophysicists to simulate wave diffractions by complex variation of Earth's topography or subsurface cavities (cavities), fractures, irregular layers, etc. Geophysicists now often use the diffracted wave-field to reconstruct the subsurface physical properties (diffraction tomography) to solve the so-called inverse diffraction problem.

Cross-References

- [Deep Seismic Reflection and Refraction Profiling](#)
- [Earthquake Rupture: The Inverse Problem](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Ray Theory](#)
- [Seismic Waves, Scattering](#)
- [Seismic, Migration](#)

Bibliography

- Achenbach JD, Gantesen AK, McMaken H (1982) Ray methods for waves in elastic solids, with application to scattering by cracks. Pitman Learning, London
- Bouchon M (1987) Diffraction of elastic waves by cracks or cavities using the discrete wave-number method. *J Acoust Soc Am* 81:1671
- Coates RT, Schoenberg M (1995) Finite-difference modelling of faults and fractures. *Geophysics* 60:1514
- Coutant O (1989) Numerical study of the diffraction of elastic waves by fluid-filled cracks. *J Geophys Res* 94:17805
- Douglas A, Hudson JA (1990) The effect on teleseismic P of the zone of damage created by an explosion. *Geophys J Int* 103:111
- Fehler M, Aki K (1978) Numerical study of diffraction of plane elastic waves by a finite crack with application to location of a magma lens. *Bull Seismol Soc Am* 68:573
- Hudson JA, Heritage JR (1981) The use of the Born approximation in seismic scattering problems. *Geophys J Int* 66:221
- Keller JB (1962) Geometrical theory of diffraction. *J Opt Soc Am* 52:116
- Larose E, Margerin L, van Tiggelen BA, Campillo M (2004) Weak localization of seismic waves. *Phys Rev Lett* 93:048501-1–048501-4

- Liu E, Zhang ZJ (2001) Numerical study of elastic wave scattering by distributed cracks or cavities using the boundary integral method. *J Comput Acoust* 9:1039
- Liu E, Crampin S, Hudson JA (1997) Diffraction of seismic waves by cracks with application to hydraulic fracturing. *Geophysics* 62:253
- Pointer T, Liu E, Hudson JA (1998) Numerical modelling of seismic waves scattered by hydrofractures: application of the indirect boundary element method. *Geophys J Int* 135:289
- Sanchez-Seama FJ, Campillo M (1993) Topographic effects for incident P, SV, and Rayleigh waves. *Tectonophysics* 218:113
- Sanchez-Sesma FJ, Campillo M (1991) Diffraction of P, SV, and Rayleigh waves by topographic features: a boundary integral formulation. *Bull Seismol Soc Am* 81:2234
- Vlastos S, Liu E, Main IG, Li XY (2003) Numerical simulation of wave propagation in media with discrete distributions of fractures: effects of fracture sizes and spatial distributions. *Geophys J Int* 152:649
- Vlastos S, Liu E, Main IG, Narteau C (2007) Numerical simulation of wave propagation in fractured media: scattering attenuation at different stages of the growth of a fracture population. *Geophys J Int* 171:865
- Wu RS (1982) Attenuation of short period seismic waves due to scattering. *Geophys Res Lett* 9:9
- Wu RS (1989) The perturbation method in elastic wave scattering. *Pure Appl Geophys* 131:605
- Wu RS, Toksoz MN (1987) Diffraction tomography and multisource holography applied to seismic imaging. *Geophysics* 52:11

Seismic Discontinuities in the Transition Zone

Lev P. Vinnik

Institute of Physics of the Earth, Moscow, Russia

Definition

The transition zone (TZ) is the mantle layer bounded by the 410- and 660-km seismic boundaries. The high P-wave and S-wave velocity gradients within the TZ are caused by a series of polymorphic phase transitions, the depths (pressures) of which are controlled by temperature and composition. Structure of the TZ plays an important role in the heat/mass transfer between the upper and the lower mantle.

Mineral Physics Data on the Phase Transitions in the TZ

The most frequently used model of mantle composition is pyrolite which contains ~60% of olivine (Mg,Fe)₂SiO₄. At a depth of ~410 km olivine (α) transforms to wadsleyite (β , modified spinel). The Clapeyron slope of this transition is positive (4.0 MPa/K, Katsura et al. 2004); the increase of the S-wave velocity is ~12%. At a depth of ~550 km wadsleyite transforms to ringwoodite (γ , silicate spinel). At a depth of ~660 km ringwoodite transforms to a mixture of

perovskite (Mg,Fe)SiO₃ and magnesiowüstite (Mg,Fe)O. The S velocity contrast of this transition is comparable to that of the α/β transition. The post-spinel transition is sharp and has a negative Clapeyron slope of -2.8 MPa/K (e.g., Hirose 2002). This value is disputed and more recent estimates are in a range from -0.4 to -1 MPa/K (Ohtani and Sakai 2008).

The other components of pyrolite are pyroxenes and garnet. These components experience in the TZ more gradual transformations: the pyroxenes dissolve into garnet (majorite) and majorite transforms to perovskite near the bottom of the TZ. The Clapeyron slope of the post-majorite transition is 1.3 MPa/K (Hirose 2002). The post-spinel and post-majorite phase boundaries intersect at $1,700$ – $1,800$ °C, and the corresponding seismic discontinuity may be an effect of both transitions. At the temperatures less than $1,700$ °C the discontinuity is formed mainly by the post-spinel transition, while at the higher temperatures the post-majorite transition becomes dominant (Hirose 2002).

Seismic Methods

The models of composition and temperature of the TZ are constrained by seismic data. An increase of the seismic velocity with depth is mirrored by the increase of the apparent velocity of the P and S seismic arrivals with distance. Details can be obtained from the analysis of the related loop (triplication) in the travel times and modeling the related waveforms with synthetic seismograms.

The triplications corresponding to discontinuities in the TZ are observed at epicentral distances between $\sim 1,700$ and $\sim 3,200$ km, whereas properties of the mantle may change significantly on a scale of a few hundred kilometers. Observations of reflected and mode-converted phases of distant earthquakes provide better lateral resolution. In the neighboring field of exploration seismology, accurate mapping of discontinuities is based on observations of reflected phases. These signals are small relative to noise, and the detection is performed by using receiver arrays and stacking the recordings of different sensors with move-out time corrections. In seismology adequate receiver arrays still are few, especially in the oceans. The idea of receiver function approach in seismology is to replace receiver arrays by seismic-source arrays that can be very large and dense. The differences in the waveforms of individual earthquakes that are recorded at the same station are eliminated by appropriate frequency filtering (deconvolution).

P-wave receiver functions (PRFs) present the oldest and most usable variety of receiver functions. This technique is based mainly on observations of Ps (converted from P to S) phases. The delay of the Ps phase relative to P depends on the depth of the discontinuity and velocities along the wave-propagation paths. The amplitude of the Ps phase is

proportional to the S velocity contrast at the discontinuity. The detection of small seismic phases in PRFs is accompanied by measurements of their slowness. In a laterally homogeneous Earth, Ps phases from the TZ discontinuities differ by slowness from lithospheric reverberations that arrive in the same time window, and this difference helps to separate the signals from noise. The best results can be obtained by combining source and receiver arrays. Unfortunately, practical detection of the TZ seismic phases in noise is often based on relaxed criteria. Some of the widely cited results are obtained from noisy receiver functions which show a continuous train of positive “swells” and negative “troughs.” This interference pattern may be mistaken for a sequence of separate seismic arrivals.

S-wave receiver functions (SRFs) are complementary to PRFs and deal with the Sp phases converted from S to P. Multiple scattering at shallow discontinuities, which presents the major source of noise in PRFs, is practically absent in SRFs, because the Sp phases from large depths arrive much earlier than the scattered phases from shallow discontinuities. Another useful method is based on observations of ScS reverberations. Detection of the phases reflected from TZ discontinuities in this method is in principle similar to that employed in receiver functions and this technique can be viewed as a variety of receiver function techniques. A shortcoming of this method is its low (in comparison to PRFs and SRFs) lateral resolution. Precursors to the seismic phases SS (SH component), PP, and P'P' (SH, P, and P' waves reflected from the Earth's surface between the source and the receiver) include phases reflected from TZ discontinuities. These phases can be detected by using the receiver function approach. Lateral resolution of the SS-precursor technique (a few thousand kilometers) is by an order of magnitude lower than of P410s and P660s phases in PRFs, but the SS precursors are useful in the studies of the TZ on a global scale.

Topography and Sharpness of TZ Discontinuities

The amplitudes of seismic phases converted or reflected from the TZ discontinuities at 410- and 660-km depths vary laterally but on the average are approximately two times lower than predicted for olivine. This relationship implies that the actual composition of the TZ is broadly similar to pyrolite which contains $\sim 60\%$ of olivine.

In the literature there are numerous reports on observations of reflected phases from a discontinuity at a depth around 520 km (e.g., Flanagan and Shearer 1998). The most popular interpretation relates this discontinuity to the phase transition from wadsleyite to ringwoodite. However, the theoretical amplitudes of reflected and converted phases from the 520-km discontinuity in pyrolite should be several times lower than the observed ones and, for this reason, practically

unobservable. There are indications of a low S-wave velocity in a depth range from 450 to 500 km in the vicinities of several hotspots. Hence the seismically detectable 520-km discontinuity in hotspot regions may in reality present the base of this low-velocity layer (e.g., Vinnik et al. 2012).

The width of the 410-km discontinuity has been estimated by seismic studies since about 1993. The estimates range from 4 km to 35 km. According to Katsura et al. (2004) the width of the olivine-wadsleyite transition is from 7 to 13 km. The largest width is possible in regions of rapid mantle flow. The smallest value is characteristic of stagnant mantle with a slow vertical movement. The width of 4 km and less has no accepted explanation.

There are many observations of short-period P waves reflected from the 660-km discontinuity, which indicate that this discontinuity is sharp. However, the spectra of P660s converted phase are shifted to longer periods relative to P with implication that the 660-km boundary is a few tens of kilometers wide. This complexity may be explained with a model of two phase transitions (post-spinel and post-majorite) of different width in about the same depth range.

Analysis of PRFs demonstrates that the large, in a range of a few seconds, lateral variations of travel times of the TZ seismic phases are caused mainly by lateral heterogeneity of the upper mantle outside the TZ. These variations are similar for P660s and P410s seismic phases and can be removed by taking the difference between the P660s and P410s arrival times at the same station. Lateral variations of this difference are sensitive to changes in the TZ thickness. The data on the P660s-P410s differential times suggest that the depths to the 410- and 660-km discontinuities usually vary in a range of not more than a few kilometers or a fraction of a second in time almost everywhere, except in the anomalously hot and cold regions: hotspots and subduction zones, respectively. These normal depths of the major boundaries practically coincide with those in the IASP91 model (410 and 660 km), and correspond to the normal TZ thickness of 250 km.

The results for hotspots, most of which are located in oceans, are controversial. The related data are highly variable in quality. Therefore, instead of relying on statistics, I prefer to single out the results for Iceland, which is one of the most thoroughly investigated hotspots. The network of seismograph stations in Iceland is large, and there are several independent tomographic studies of this region. Seismic tomography and receiver functions reveal beneath Iceland a narrow columnar low-velocity body extending into the TZ, in which the 410-km discontinuity is depressed by ~ 15 km, while the 660-km discontinuity is flat. The conclusion on the flat 660-km boundary is by Du et al. (2006). The depressed 410-km discontinuity suggests that the TZ temperature is elevated by ~ 150 °C. The flat 660 km discontinuity means that either the temperature at this depth is close to normal or,

as suggested by high-pressure data (Hirose 2002), a sensitivity of depth of the transition to temperature is low.

This model is disputed by Jenkins et al. (2016) who propose that the 660-km boundary is depressed like the 410-km boundary. This depression implies that the Clapeyron slope at a depth of 660 km is positive. The positive slope is possible if the 660-km discontinuity is related with the phase transition in garnet. It should be noted that the dual interpretation of the data on topography of the 660-km discontinuity beneath Iceland depends on some poorly resolved details. The seismic data for another hotspot in the Atlantic (Cape Verde, Vinnik et al. 2012) provide clear evidence of a depression on the 410-km discontinuity and uplift of the 660-km discontinuity of comparable amplitude. Note that for the ray paths outside the columnar body beneath Iceland, the time difference between the P660s and P410s seismic phases is close to the standard time with implication that the TZ of the normal oceanic mantle has the same 250-km thickness as beneath the continents.

Broadly similar results are obtained for several other hotspots. For example, similar measurements at several stations in the region of the South Pacific superswell (Suetsugu et al. 2007) show that beneath one station the thickness of the TZ is 216 ± 19 km. The anomalous region cannot be larger than a few hundred kilometers. The average for all other stations is 248 ± 5 km, very close to the standard value of 250 km. By comparison, the SS-precursor data portray in the South Pacific a large (several thousand kilometers) region, where the TZ thickness is reduced to ~ 230 km (e.g., Flanagan and Shearer 1998). A similar region is found in the Atlantic. These data, if accepted at face value, imply that the TZ beneath oceans differs from that beneath continents, and the TZ beneath the hotspots is the normal oceanic TZ, contrary to the receiver function data. Substantial discrepancies between the results of the two methods exist in continental regions where numerous seismograph stations provide good-quality data.

Subduction zones demonstrate another kind of complexity. Temperature in subducted slabs is anomalously low and the equilibrium depth of the olivine-wadsleyite phase transition should be ~ 100 km less than in ambient mantle. However, the 410-km discontinuity in most of the presently active subduction zones apparently cannot be detected with either PRFs or other methods, most likely because the olivine-wadsleyite phase transformation is kinetically hindered and a wedge-shaped zone of olivine may persist at depths greatly exceeding 410 km. The signals from the 410-km discontinuity that are sometimes obtained in seismic studies of these regions are most likely formed outside the cold subducted lithosphere. Observations of short-period phases converted from S to P near the hypocenters of deep earthquakes show that the 660-km discontinuity in subducted slabs varies in depth but sometimes is depressed by up to ~ 50 km. This may be

explained by the negative Clapeyron slope of the post-spinel transition and a low temperature of the subducted slab. The temperature anomalies may be up to -600°C .

Seismic observations sometimes reveal a low S-wave velocity layer over the 660-km discontinuity. This feature can be explained by accumulation of oceanic crust in subduction zones above the 660-km discontinuity.

Robust evidence for an elevated differential time P660s-P410s and a depression of the 660-km discontinuity is provided by PRFs in the Alps where the active phase of Cenozoic subduction is over, but remnants of the oceanic lithosphere are present in the TZ. Similar anomalies are either found or can be expected in other regions of the Alpine belt.

The Issue of Water in the TZ

The cited results of mineral physics were obtained for the dry TZ. Many researchers examined solubility of water in minerals of the TZ (see, e.g., Ohtani and Sakai 2008). Hydration means modification of structures of the TZ minerals by incorporation of hydroxyl (OH). It appears that wadsleyite and ringwoodite may incorporate up to ~ 2.0 wt% of water, and the TZ is thought to be the main reservoir of water in Earth (e.g., Karato 2011). Pressure of the olivine-wadsleyite transition decreases by hydration by up to about 1 GPa (~ 30 km in depth), whereas the pressure of the post-spinel transition increases. The width of the two-phase loop between olivine and wadsleyite increases with increasing water content, and may reach a few tens of kilometers in depth relative to several kilometers for dry conditions. Hydration of 1 wt% lowers S velocity by 1–2%, whereas P velocities remain practically the same.

The predicted anomalies in the depth (30 km) and thickness (a few tens of kilometers) of the 410-km discontinuity are sufficiently large to be detected by PRFs. These predictions stimulated attempts to detect the effects of water in seismic data. So far the results are controversial. Karato (2011) stressed striking incompatibility between the results of different studies and concluded that frequently used seismological observations are more sensitive to the temperature and major element chemistry than to water content. The estimates of the Clapeyron slope for the post-spinel transition (between -0.4 and -1.3 MPa/K) in the dry TZ may explain only part of the observed topography on the 660-km discontinuity, but the discrepancy can be reduced in the wet TZ (Litasov et al. 2005).

A low S velocity layer atop the 410-km discontinuity may present another possible effect of hydration of the TZ. Owing to the large water solubility, the TZ may have higher water concentration than water storage capacity of the upper mantle. Then upwelling mantle material entering the upper mantle from the TZ may undergo dehydration melting.

Evidence of the low-velocity layer atop the 410-km discontinuity is obtained in PRFs and SRFs at a number of locations that include Antarctica, Africa, Siberia, North America, China, and Arabian Peninsula. Most locations of this layer seem to be associated with Cenozoic and Mesozoic mantle upwellings.

Summary

Seismic observations of the major TZ discontinuities near 410- and 660-km depths on a global scale are broadly consistent with the pyrolite mantle model. The 410-km discontinuity is related to the olivine-wadsleyite phase transition with a positive Clapeyron slope. The 660-km discontinuity is related to the post-spinel transition in the olivine component and post-majorite transition in the other components, with a negative and positive Clapeyron slope, respectively. In the normal mantle, the depths of the discontinuities are stable and in good agreement with the IASP91 model, where the thickness of the TZ is 250 km. Anomalous topography of the TZ discontinuities is related to hot and cold regions (hotspots and subduction zones, respectively). The anomalies beneath oceans are related mainly to hotspots, where the 410-km discontinuity is depressed by ~ 20 km. The corresponding temperature anomalies are up to $\sim 200^{\circ}\text{C}$. Beneath continents, except Africa and Antarctica, the anomalies are related mainly to Cenozoic subduction zones, where the 660-km discontinuity can be depressed by up to ~ 50 km. The related temperature anomalies are in a range of several hundred degrees Celsius. In spite of high water solubility in wadsleyite and ringwoodite, credible observations of seismic effects of hydration in the TZ are few. The thin low-S-velocity layer atop the 410-km discontinuity, found at a number of locations, is the exception. The low velocity can be an effect of a hydrous melt. On a regional scale the TZ may contain less well-understood complexities such as the 520-km discontinuity or a low S velocity layer between 450- and 500-km depths.

Cross-References

- [Body Waves](#)
- [Earth's Structure, Global](#)
- [Earth's Structure, Upper Mantle](#)
- [Geodynamics](#)
- [Mantle Convection](#)
- [Mantle Plumes](#)
- [Seismic Phase Nomenclature: The IASPEI Standard](#)
- [Seismic, Receiver Function Technique](#)
- [Seismology, Global Earthquake Model](#)
- [Subduction Zones](#)

Bibliography

- Du Z, Vinnik LP, Foulger GR (2006) Evidence from P-to-S mantle converted waves for a flat “660-km” discontinuity beneath Iceland. *Earth Planet Sci Lett* 241:271–280
- Flanagan MP, Shearer PM (1998) Global mapping of topography of transition zone velocity discontinuities by stacking SS precursors. *J Geophys Res* 103:2673–2692
- Hirose K (2002) Phase transitions in pyrolitic mantle around 670-km depth: implications for upwelling of plumes from the lower mantle. *J Geophys Res* 107(B4):2078. <https://doi.org/10.1029/2001JB000597>
- Jenkins J, Cottaar S, White RS, Deuss A (2016) Depressed mantle discontinuities beneath Iceland: evidence of a garnet controlled 660 km discontinuity? *Earth Planet Sci Lett* 433:159–168
- Karato SI (2011) Water distribution across the mantle transition zone and its implications for global material circulation. *Earth Planet Sci Lett* 301(3–4):413–423
- Katsura T, Yamada H, Nishikawa O, Song M, Kubo A, Shinmei T, Yokoshi S, Aizawa Y, Yoshino T, Walter MJ, Ito E, Funakoshi K (2004) Olivine-wadsleyite transformation in the system (Mg,Fe)₂SiO₄. *J Geophys Res* 109:B02209. <https://doi.org/10.1029/2003JB002439>
- Litasov KD, Ohtani E, Sano A, Suzuki A, Funakoshi K (2005) Wet subduction versus cold subduction. *Geophys Res Lett* 32(13). <https://doi.org/10.1029/2005GL022921>
- Ohtani E, Sakai T (2008) Recent advances in the study of mantle phase transitions. *Phys Earth Planet Inter* 170:240–247
- Suetsugu D, Shiobara H, Sigioka H, Fukao Y, Kanazawa T (2007) Topography of the mantle discontinuities beneath the South Pacific superswell as inferred from broadband waveforms on seafloor. *Phys Earth Planet Inter* 160(3–4):310–318
- Vinnik L, Silveira G, Kiselev S, Farra V, Weber M, Stutzmann E (2012) Cape Verde hotspot from the upper crust to the top of the lower mantle. *Earth Planet Sci Lett* 319:259–268

Seismic Hazard

Andrzej Kijko

Natural Hazard Centre, Africa, University of Pretoria, Pretoria, Republic of South Africa

Definition

Seismic hazard	Any physical phenomenon associated with an earthquake (e.g., ground motion, ground failure, liquefaction, and tsunami) and its effects on land, man-made structures, and socioeconomic systems that have the potential to produce a loss. The terms are used also without regard to a loss to indicate the probable level of ground shaking occurring at a given point within a certain period of time.
Seismic hazard analysis	Quantification of the ground motion expected at a particular site or selection of sites.

Deterministic seismic hazard analysis	Quantification of a single or relatively small number of individual earthquake scenarios.
Probabilistic seismic hazard analysis	Quantification of the rate (or probability) that a specified level of ground motion will be exceeded at least once at a site (or a map of sites) given all possible earthquakes during the specified exposure time.
Ground motion model (known also as ground motion prediction equation)	A mathematical equation that provides a means of predicting the level of ground shaking and its associated uncertainty at any given site or location, based on earthquake magnitude, source-to-site distance, local soil conditions, and focal mechanism.

Introduction

Estimating the ground motion that could occur at a particular site is crucial in the design of both vital structures, such as nuclear power plants, bridges, and dams, and ordinary structures, such as houses and commercial buildings. The process of assessing future earthquake ground motion parameters is called seismic hazard assessment or seismic hazard analysis. Seismologists and earthquake engineers distinguish between seismic hazard and seismic risk, although the two phrases have the same meaning in everyday usage. Seismic hazard is used to characterize the severity of the ground motion at a site regardless of the consequences, whereas risk refers exclusively to the consequences to humans and infrastructure from the occurrence of the hazard. Therefore, even a strong earthquake could have negligible risk potential if it occurred in sparsely populated and undeveloped areas. In contrast, a relatively weak seismic event in a highly developed and densely populated location poses a high risk owing to the possible loss of life and extensive infrastructure damage.

Seismic hazard analysis is performed *deterministically* when a particular earthquake scenario is considered and *probabilistically* when the likelihood or frequency of specified earthquake size and location are evaluated.

Deterministic seismic hazard analysis (DSHA) involves an initial assessment of the maximum possible earthquake magnitude $m_{\max}$ for each of the seismic sources, such as active faults or seismic source zones (SSHAC 1997). The maximum earthquake magnitude for a particular seismic source is often referred to as the maximum credible earthquake (MCE). An area with a radius of up to 450 km around the site of interest can be investigated. Assuming that each of these earthquakes will occur at the minimum possible distance from the site, the ground motion is calculated using approximate attenuation equations, known as the ground motion model (GMM).

However, this straightforward and intuitive procedure is overshadowed by the complexity and uncertainty in selecting the appropriate earthquake scenario. This creates the need for an alternative, *probabilistic* methodology free from selecting an earthquake occurrence scenario. Probabilistic seismic hazard analysis (PSHA) quantifies the rate (or probability) of exceeding various ground motion levels at a site (or a map of sites) from all earthquakes of all possible magnitudes and at all significant distances from the site of interest. Therefore, deterministic earthquake scenarios are special cases of the probabilistic approach. Depending on the scope of the project, DSHA and PSHA could be complementary, i.e., providing additional insights into the seismic hazard. The current study discusses PSHA.

In principle, any natural hazard caused by seismic activity can be described and quantified by the formalism of PSHA. As the damage caused by ground shaking quite often results in the largest economic losses, this presentation of the basic concepts of PSHA is illustrated by quantifying the likelihood of ground shaking generated by earthquakes.

The classic (Cornell 1968; McGuire 2004) procedure, known as the Cornell–McGuire procedure for PSHA, includes four steps (Fig. 1).

Step 1. Identification and parameterization of the *seismic sources* (also known as *source zones*, *seismogenic zones*, or *seismic zones*) that can affect the relevant site. These can be represented as area, fault, or point sources. Area sources are often used when a specific fault cannot be identified. In classic PSHA, uniform distribution of seismicity is assigned to each seismic source, implying that earthquakes are equally likely

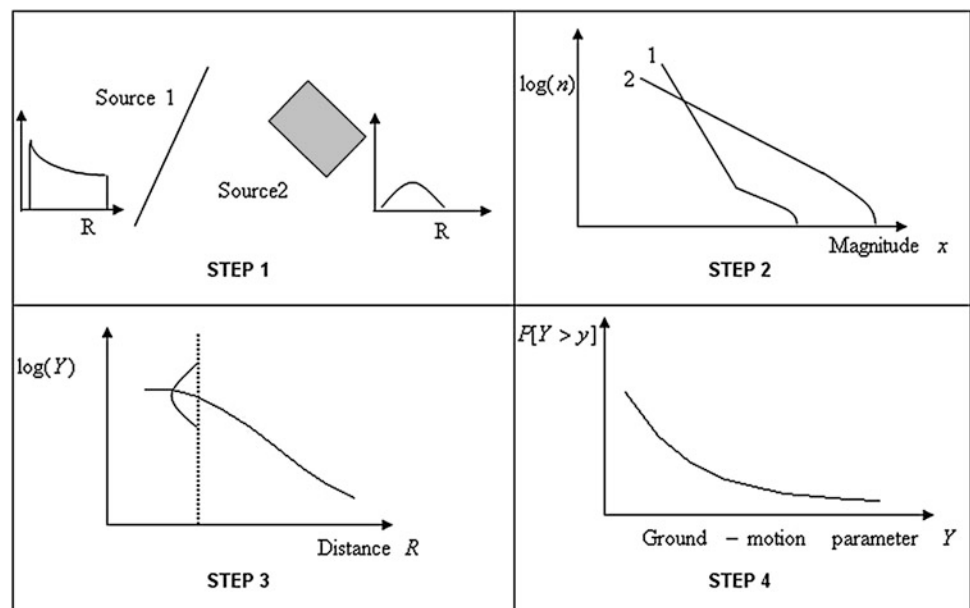
to occur at any point within the source zone. Combining earthquake occurrence distributions with the source geometry results in the space, time, and magnitude distributions of earthquake occurrences. Seismic source models can be interpreted as a list of potential scenarios, each with associated earthquake magnitude, location, and seismic activity rate.

Step 2. Specification of the temporal and magnitude distributions of seismicity for each seismic source. The Cornell–McGuire approach assumes that earthquake occurrence in time is random and follows the Poissonian process. This implies that earthquake occurrences in time are statistically independent and occur at a constant rate, i.e., the occurrence of future earthquakes does not depend on the occurrence of past earthquakes. The model of earthquake magnitude recurrence used most often is the Gutenberg–Richter frequency–magnitude relation

$$\log(n) = a - bm, \quad (1)$$

where n is the number of earthquakes with a magnitude of m and a and b are parameters. It is assumed that earthquake magnitude m belongs to the domain $\langle m_{\min}, m_{\max} \rangle$, where $m_{\min}$ is the level of completeness of the earthquake catalogue and magnitude $m_{\max}$ is the upper limit of the earthquake magnitude for a given seismic source. Parameter a is the measure of the level of seismicity, whereas parameter b describes the ratio between the number of small and large events. The Gutenberg–Richter relations can be interpreted either as a cumulative relationship where n is the number of

Seismic Hazard, Fig. 1 Four steps of PSHA



events with magnitude equal or larger than m or as a density law, i.e., n is the number of earthquakes in a specific, small magnitude interval around m . Under these assumptions, the seismicity of each seismic source is described by four parameters, i.e., the (annual) rate of seismicity λ , which is equal to the parameter of the Poissonian distribution, the lower and upper limits of earthquake magnitude $m_{\min}$ and $m_{\max}$, and the b -value of the Gutenberg–Richter relation.

Step 3. The relevant parameters in PSHA include peak ground acceleration, peak ground velocity, peak ground displacement, spectral acceleration, macroseismic intensity, strong ground motion duration, and the like. Most ground motion models available today are empirical and depend on the earthquake magnitude, source-to-site distance, type of faulting, and local site conditions. Therefore, the third step, i.e., choosing an appropriate ground motion model, is crucial, as, quite often, it is a major contributor to the formal uncertainty in the estimated PSHA.

Step 4. Calculations of the ultimate result of PSHA in terms of a *seismic hazard curve*, i.e., the annual probability of exceeding a specified ground motion parameter at least once. An alternative definition of the hazard curve is the frequency of exceedance vs ground motion amplitude. This calculation integrates the uncertainties in the earthquake location, earthquake magnitude, and ground motion model into the probability that the relevant ground motion parameter will be exceeded at the specified site during the specified time interval.

Cornell–McGuire PSHA Methodology

Conceptually, the computation of a seismic hazard curve is fairly simple (Kramer 1996). Assuming that seismic hazard is characterized by ground motion parameter Y , the probability of exceeding a specified value y , $P[Y \geq y]$, is calculated for an earthquake of particular magnitude located at a possible source and then multiplied by the probability that such a particular earthquake will occur. The computations are repeated and summed for the entire range of possible magnitudes and earthquake locations. The resulting probability $P[Y \geq y]$ is calculated by utilizing the total probability theorem

$$P[Y \geq y] = \sum P[Y \geq y|E_i]P[E_i] \quad (2)$$

where

$$P[Y \geq y|E_i] = \int \cdots \int P[Y \geq y|x_1, x_2, x_3, \dots] \times f_i(x_1)f_i(x_2|x_1)f_i(x_3|x_1, x_2) \cdots dx_3 dx_2 dx_1 \quad (3)$$

$P[Y \geq y|E_i]$ denotes the probability of ground motion parameter $Y \geq y$ at the relevant site when an earthquake

occurs within seismic source i . The variables x_i ($i = 1, 2, \dots$) are uncertainty parameters that influence Y . In the classic approach, developed by Cornell (1968) and later extended to accommodate ground motion uncertainty, the parameters of ground motion are earthquake magnitude M and earthquake distance R . Functions $f(\cdot)$ are probability density functions (PDFs) of parameters x_i . Assuming that $x_1 \equiv M$ and $x_2 \equiv R$, the probability of exceedance (3) takes the form:

$$P[Y \geq y|E_i] = \int_{m_0}^{m_{\max}} \int_{R|M} P[Y \geq y|m, r] f_M(m) f_{R|M}(r|m) dr dm, \quad (4)$$

where $P[Y \geq y|m, r]$ denotes the conditional probability that the chosen ground motion level y is exceeded for a given magnitude and distance, $f_M(m)$ is the probability density function (PDF) of earthquake magnitude, and $f_{R|M}(r|m)$ is the conditional PDF of the distance from the earthquake for a given magnitude. The conditional PDF of the distance $f_{R|M}(r|m)$ arises in specific instances, such as a seismic source being represented by a fault rupture. As the earthquake magnitude depends on the length of the fault rupture, the distance to the rupture and the resulting magnitude are correlated. In Eq. (4) the magnitude m_0 is chosen on the basis of the minimum magnitude that will cause damage or loss. In most PSHA, the m_0 is set in the range of 4.0–5.0.

The standard choice for the probability $P[Y \geq y|m, r]$ is a normal, complementary, cumulative distribution function (CDF), based on the assumption that the ground motion parameter y is a log-normal random variable, $\ln(y) = g(m, r) + \varepsilon$, where ε is a random error. The mean value of $\ln(y)$ and its standard deviation are known and are defined as $\overline{\ln(y)}$ and $\sigma_{\ln(y)}$, respectively. In most engineering applications of PSHA, it is assumed that earthquake magnitudes follow the Gutenberg–Richter relation (1), which implies that $f_M(m)$ is a negative exponential distribution, shifted from zero to level of completeness $m_{\min}$, and truncated from the top by $m_{\max}$

$$f_M(m) = \frac{\beta \exp[-\beta(m - m_{\min})]}{1 - \exp[-\beta(m_{\max} - m_{\min})]} \quad (5)$$

In Eq. (5), $\beta = b \ln 10$, where b is the parameter of the Gutenberg–Richter frequency–magnitude relation (1). It should be noted that the truncation of the classic Gutenberg–Richter frequency–magnitude relation is obligatory to ensure that the seismic energy released by the seismogenic process is finite.

If n_S seismic sources $\{E_i, i = 1, \dots, n_S\}$ can be distinguished in the vicinity of the relevant site, each with an annual average rate of earthquake occurrence λ_i , the total average annual rate of events with a site ground motion level y or more takes the form

$$\lambda(y) = \sum_{i=1}^{n_s} \lambda_i P[Y \geq y | E_i], \quad (6)$$

After assuming that in every seismic source, earthquake occurrences in time follow a Poissonian distribution, the probability that y , a specified level of ground motion at a given site, will be exceeded at least once within any time interval t is

$$P[Y > y; t] = 1 - \exp[-\lambda(y)t]. \quad (7)$$

For $t = 1$ year, the plot $P[Y > y; t]$ vs ground motion parameter y is the *hazard curve* – the ultimate product of PSHA (Fig. 2). For small probabilities, less than 0.05,

$$\begin{aligned} P[Y > y; t = 1] &= 1 - \exp[-\lambda] \\ &\cong 1 - \left(1 - \lambda + \frac{1}{2}\lambda^2 - \dots\right) \cong \lambda, \end{aligned} \quad (8)$$

which means that the probability (7) is approximately equal to $\lambda(y)$. This shows that PSHA can be characterized interchangeably by the annual probability (7) or by the rate of seismicity (6). Often, practitioners prefer to describe seismic hazard in terms of frequency rather than probability. Frequencies have a clear physical meaning and, following Eq. (6), they are additive. Therefore, frequencies facilitate assessing the individual hazard contribution from each seismic source/area and support the optimal design decision and choice of alternative sites for structure location.

In the Cornell–McGuire procedure for PSHA, it is assumed that the earthquakes in the catalogue are independent events. As clusters of seismicity, or foreshocks and

aftershocks, violate this assumption, these dependent events are removed from the catalogue before computing PSHA. However, removing such dependent events is one of the most controversial practices in current PSHA.

Estimation of Seismic Source Parameters

Following the Cornell–McGuire PSHA procedure, each seismic source is characterized by four parameters:

- Level of completeness of the seismic data, $m_{\min}$.
- The annual rate of seismic activity λ , corresponding to the magnitude $m_{\min}$
- B -value of the Gutenberg–Richter frequency–magnitude relation
- The upper limit of earthquake magnitude $m_{\max}$

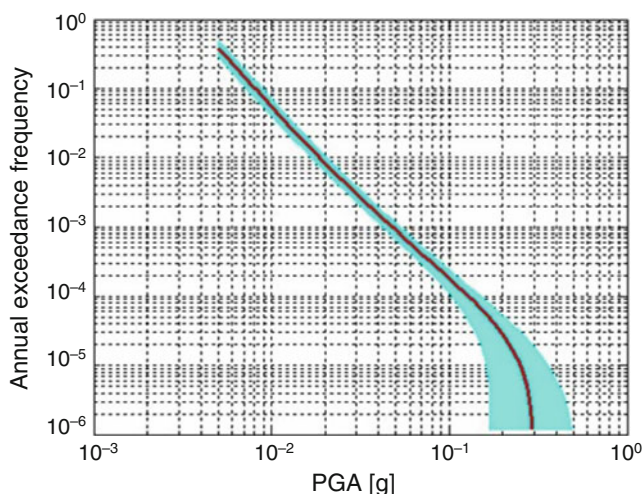
Estimation of $m_{\min}$. The level of completeness of the seismic event catalogue, $m_{\min}$, can be estimated in at least two different ways.

The first approach is based on information provided by the seismic event catalogue, where $m_{\min}$ is determined from the deviation of the empirical and assumed earthquake magnitude distribution model. In most instances, the model is based on the Gutenberg–Richter relation (1). Occasionally, $m_{\min}$ is estimated from a comparison of the day-to-night ratio of events. Despite being used widely, the approach based on the seismic event catalogue has several weak points. One of these is the prerequisite knowledge of the spatial and temporal earthquake occurrence model.

The second approach utilizes information on the detection capabilities and signal-to-noise ratio of the seismic stations that record seismic events. The approach is free from assumptions that seismic events are statistically independent and that they form a stationary process. Choosing the most appropriate procedure for estimating $m_{\min}$ depends on several factors, such as knowledge of the time history and evolution of the seismic network and data collection and processing.

Estimation of the rate of seismic activity λ and the b -value of Gutenberg–Richter. The maximum likelihood procedure is the conventional approach to estimate the seismic source recurrence parameters λ and b . If successive earthquakes are independent in time, the number of earthquakes with magnitude equal to or exceeding a level of completeness $m_{\min}$ follows the Poissonian distribution, with the parameter equal to the annual rate of seismic activity λ . The maximum likelihood estimator of λ is then equal to n/t , where n is the number of events occurring within time interval t .

For specified $m_{\max}$, the maximum likelihood estimator of the b -value of the Gutenberg–Richter relation can be obtained from the recursive solution of the equation:



Seismic Hazard, Fig. 2 Example of a peak ground acceleration (PGA) seismic hazard curve and its confidence intervals

$$1/\beta = \bar{m} - m_{\min} + \frac{(m_{\max} - m_{\min}) \exp[-\beta(m_{\max} - m_{\min})]}{1 - \exp[-\beta(m_{\max} - m_{\min})]}, \quad (9)$$

where $\beta = b \ln 10$ and $\bar{m}$ is the sample mean of earthquake magnitude. If the range of earthquake magnitudes $\langle m_{\max} - m_{\min} \rangle$ exceeds two magnitude units, the solution of Eq. (9) can be approximated by the well-known Aki–Utsu estimator

$$\beta = 1/(\bar{m} - m_{\min}). \quad (10)$$

In most instances, estimating λ and the b -value by the above simple formulas cannot be performed because seismic event catalogues are incomplete. Generally, seismic event catalogues contain three types of data, namely, prehistoric (paleo-) events, the incomplete largest historical events that usually occurred over a period of a few hundred years, and the instrumental data, recorded over the last 100 years. The completeness of instrumental data often varies with time (Fig. 3).

Obviously, combining and utilizing all three types of data would be the optimal procedure to apply in such an instance. Additionally, such a procedure should accommodate information on highly uncertain sizes (magnitudes/focal intensities) of paleo- and historical earthquakes and unavoidable uncertainties associated with the applied earthquake occurrence models. The procedure developed by Smit et al. (2019) extends the well-known approach developed by Weichert (1980), which does not allow for the paleo- and historic part of the catalogue.

Estimation of $m_{\max}$. The maximum magnitude $m_{\max}$ is defined as the upper limit of magnitude for a given seismic source. Often, the magnitude of the largest possible earthquake or maximum credible earthquake is synonymous with the upper limit of earthquake magnitude. This terminology assumes a sharp cutoff earthquake magnitude distribution at a maximum magnitude $m_{\max}$.

As a rule, $m_{\max}$ is crucial in PSHA, particularly in assessing long return periods, although, currently, there is no generally

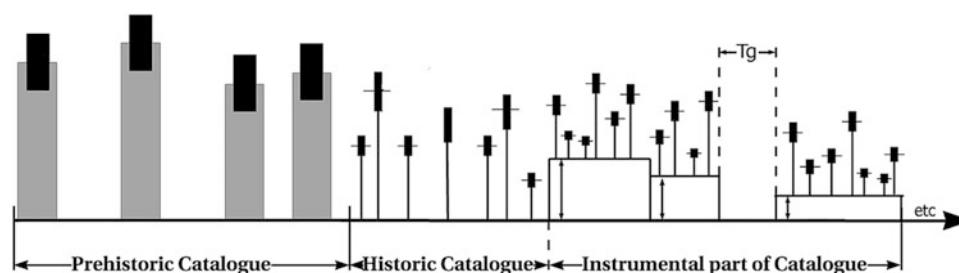
accepted method for estimating $m_{\max}$. It is estimated by combining several factors, based on two types of information, namely, the seismicity of the area and the geological, geophysical, and structural information of the seismic source. Using The seismological information focuses on the maximum observed earthquake magnitude within a seismic source and the statistical analysis of the available seismic event catalogue. The geological and geophysical information is used to identify distinctive tectonic features, which control the value of $m_{\max}$.

Current assessments of $m_{\max}$ are divided between deterministic and probabilistic procedures, based on the nature of the tools applied.

Deterministic procedures. The deterministic procedure most often applied is based on the tectonic features of the investigated area, revealed by empirical relationships between magnitude and various parameters, such as the fault rupture length, rupture dimension, or slip rate. Relationships differ for different seismic areas and different types of faults. Despite empirical relationships between magnitudes and fault rupture parameters being used extensively in PSHA (particularly for assessing the maximum possible magnitude generated by fault-type seismic sources), the weak point of the approach is the requirement to specify the highly uncertain length of a future rupture. An alternative approach to determining earthquake recurrence on singular faults with a segment-specific slip rate is the so-called cascade model, where segment rupture is defined by the individual cascade-characteristic rupture dimension.

Another deterministic procedure that has strong intuitive appeal is based on records of the largest historical or paleo-earthquakes. This approach is particularly applicable to areas of low seismicity, where large events have long return periods. In the absence of any tectonic and geological indications, it is assumed that the maximum possible earthquake magnitude is equal to the largest magnitude observed $m_{\max}^{\text{obs}}$, plus an increment. Typically, the increment varies from $1/4$ to 1 magnitude unit.

Another deterministic procedure for $m_{\max}$ evaluation, particularly for area-type seismic sources, is based on extrapolation of the Gutenberg–Richter frequency–magnitude relation



Seismic Hazard, Fig. 3 Illustration of a typical seismic event catalogue. Prehistoric (paleo-) data suffer from uncertainty related to the time of occurrence and the event size. The historic part of the catalogue contains only the largest events, the size of which is also highly

uncertain. The third part of the catalogue contains complete data, with different levels of completeness. T_g denotes gaps in the catalogue where records are absent or the seismic networks were out of operation. (After Smit et al. 2019)

when the frequency–magnitude curve is truncated at the specified value of the annual probability of exceedance (e.g., 0.001). However, currently, the procedures based on such extrapolation are not used often, and they are replaced by procedures based on the physical parameters of fault rupture.

The value of $m_{\max}$ can be estimated also from tectonic and geological features, such as the strain rate or the rate of seismic moment release. Similar approaches have also been applied in evaluating the maximum possible magnitude of anthropogenic seismicity, particularly that generated by fluid injection or deep mining. However, mostly, the uncertainty of $m_{\max}$ as determined by any deterministic procedure is large, often reaching a value of one unit on the Richter scale.

Probabilistic procedures. The first probabilistic procedure for maximum regional magnitude was developed in the late 1960s and is based on the formalism of the extreme values of random variables. Statistical tools required for the estimation of the end point of distribution functions have only been used recently (Kijko 2004, and references therein).

Available statistical tools for estimating $m_{\max}$ vary significantly. Selecting the most suitable procedure depends on the assumptions of the statistical distribution model and/or the information available on past seismicity. Some of the procedures can be applied in extreme cases when no information is available on the nature of the earthquake magnitude distribution. Some of the procedures can also be used when the earthquake catalogue is incomplete, i.e., when only a limited number of the largest magnitudes are known. Two estimators are presented here. Broadly speaking, the first estimator is straightforward and simple in application, whereas the second requires a more computational effort but provides more accurate results. It is assumed that both the analytical form and the parameters of the distribution functions of earthquake magnitude are known.

Based on the distribution of the largest among n observations, and on the condition that the largest observed magnitude $m_{\max}^{\text{obs}}$ is equal to the largest magnitude to be expected, the “simple” estimate of $m_{\max}$ is of the form (Pisarenko et al. 1996)

$$\hat{m}_{\max} = m_{\max}^{\text{obs}} + \frac{1}{nf_M(m_{\max}^{\text{obs}})}, \quad (11)$$

where $f_M(m_{\max}^{\text{obs}})$ is the PDF of the earthquake magnitude distribution. If applied to the Gutenberg–Richter relation with PDF (5), it takes the simple form

$$\hat{m}_{\max} = m_{\max}^{\text{obs}} + \frac{1 - \exp[-\beta(m_{\max}^{\text{obs}} - m_{\min})]}{n\beta \exp[-\beta(m_{\max}^{\text{obs}} - m_{\min})]}. \quad (12)$$

The second (“advanced”) procedure often used in assessing $m_{\max}$ is based on the formalism derived by Cooke (1979)

$$\hat{m}_{\max} = m_{\max}^{\text{obs}} + \int_{m_{\min}}^{m_{\max}^{\text{obs}}} [F_M(m)]^n dm, \quad (13)$$

where $F_M(m)$ denotes the CDF of random variable m . If applied to the Gutenberg–Richter frequency–magnitude relation (1), the respective CDF is

$$F_M(m) = \begin{cases} 0 & m < m_{\min} \\ \frac{1 - \exp(-\beta(m - m_{\min}))}{1 - \exp(-\beta(m_{\max} - m_{\min}))} & m_{\min} \leq m \leq m_{\max} \\ 1 & m \geq m_{\max} \end{cases}, \quad (14)$$

and the $m_{\max}$ estimator (13) takes the form

$$\hat{m}_{\max} = m_{\max}^{\text{obs}} + \frac{E_1(n_2) - E_1(n_1)}{\beta \exp(-n_2)} + m_{\min} \exp(-n), \quad (15)$$

where $n_1 = n/(1 - \exp[-\beta(m_{\max}^{\text{obs}} - m_{\min})])$, $n_2 = n_1 \exp[-\beta(m_{\max}^{\text{obs}} - m_{\min})]$, and $E_1(\cdot)$ denotes an exponential integral function.

Despite these probabilistic procedures providing powerful tools for evaluating $m_{\max}$, they have a significant weak point, namely, most of the catalogues of tectonic origin events are too short and provide insufficient information for reliable estimations of $m_{\max}$. Therefore, the Bayesian extension of statistical procedures, allowing the inclusion of alternative and independent information, such as local geological conditions, tectonic environment, geophysical data, paleo-events, a similarity with another seismic area, and the like, is superior to the procedures based only on knowledge of the seismicity of the area.

Finally, the recent trend in $m_{\max}$ assessment must be mentioned. As most of the current techniques provide quite uncertain point estimates of $m_{\max}$, several researchers suggest using instead the assessment of the confidence level of the maximum magnitude to be observed in the future time interval T . However, many engineering applications require knowledge of the point estimates of $m_{\max}$. Therefore, the formalism of the quantile assessment T -maximum cannot be used.

Numerical Computation of PSHA

Excepting a number of special cases, the hazard curve cannot be computed analytically. For the most realistic distributions, the integrations can only be evaluated numerically. The common practice is to divide the possible ranges of magnitude and distance into n_M and n_R intervals, respectively. The average annual rate (6) is then estimated as

$$\lambda(Y > y) \cong \sum_{i=1}^{n_S} \sum_{j=1}^{n_M} \sum_{k=1}^{n_R} \lambda_i P[Y > y | m_j, r_k] f_{M_j}(m_j) f_{R_k}(r_k) \Delta m \Delta r, \quad (16)$$

where $m_j = m_{\min} + (j - 0.5)(m_{\max} - m_{\min})/n_M$, $r_k = r_{\min} + (k - 0.5)(r_{\max} - r_{\min})/n_R$, $\Delta m = (m_{\max} - m_{\min})/n_M$, and $\Delta r = (r_{\max} - r_{\min})/n_R$.

Applying the procedure to a grid of points will result in a map of PSHA, on which the contours of the expected ground motion parameter during the specified time interval can be mapped.

An example of applying the above formalism is shown in Fig. 4. This figure shows the worldwide distribution of seismic hazard expressed in terms of the PGA, with a 10% probability of being exceeded at least once in 50 years. The maps were created in the framework of the Global Earthquake Model (GEM) project.

Deaggregation of Seismic Hazard

By definition, the PSHA aggregates ground motion contributions from earthquake magnitudes and distances of significance to a site of engineering interest. Therefore, the PSHA results are not representative of a single earthquake. However, an integral part of the design procedure of any critical structure is examining the most relevant earthquake acceleration time series, which are generated by earthquakes at specific magnitudes and distances. Such earthquakes are called “controlling earthquakes,” and they are used to determine the shapes of the response spectral acceleration at the site.

Controlling earthquakes are characterized by mean magnitudes and distances derived from so-called deaggregation analysis (McGuire 2004). During the deaggregation procedure, the results of PSHA are separated to determine the dominant magnitudes and the distances that contribute to the hazard curve at a specified reference probability.

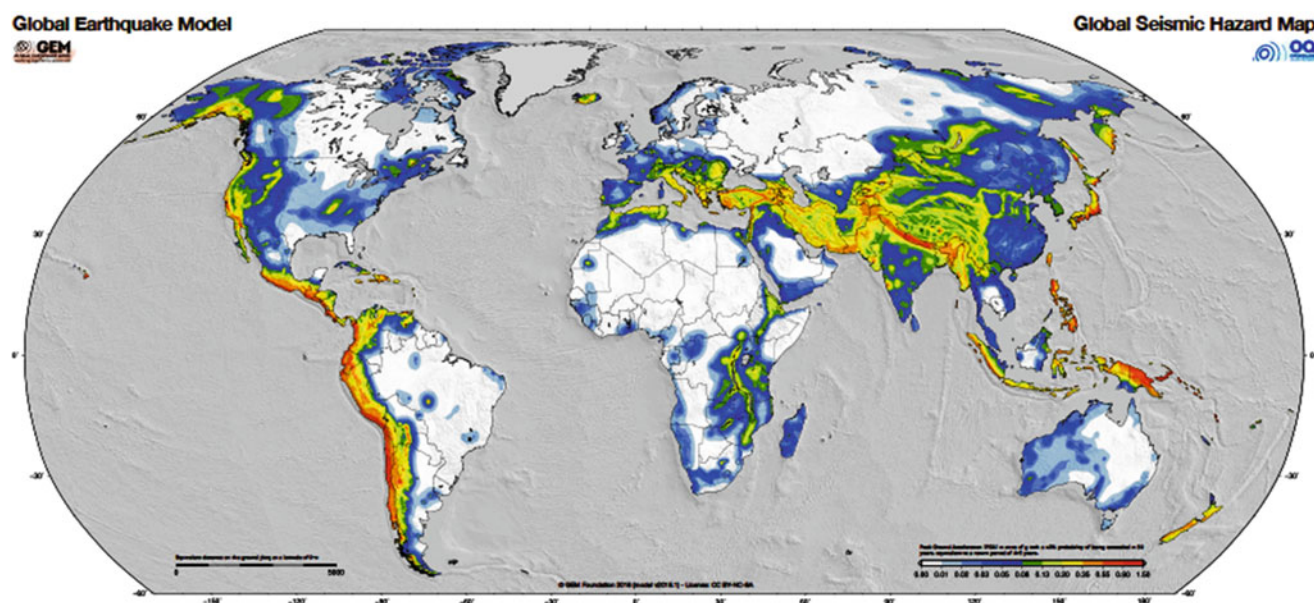
Controlling earthquakes are calculated for different structural frequency vibrations, typically for the fundamental frequency of a structure. In the deaggregation process, the hazard for a reference probability of exceedance of specified ground motion is portioned into magnitude and distance bins. The relative contribution to the hazard for each bin is calculated. The bins with the largest relative contribution identify those earthquakes that contribute the most to the total seismic hazard.

Modifications of Cornell–McGuire PSHA Procedure and Alternative Models

Source-Free PSHA Procedures

The concept of seismic sources is the core element of the Cornell–McGuire PSHA. However, seismic sources or specific faults often cannot be identified and mapped, and the causes of seismicity are not understood. In such instances, the delineation of seismic sources is highly subjective and is a matter of an expert opinion. In addition, seismicity within the seismic sources is often not distributed uniformly, as is required by the classic Cornell–McGuire procedure. The difficulties in dealing with seismic sources have stimulated the development of alternative techniques, which are free from the delineation of seismic sources (Veneziano et al. 1984; Woo 1996; Molina et al. 2001).

The core of source-free procedures is calculation of the empirical distribution of the specified seismic hazard parameter, normalized for the duration of the seismic event catalogue. Alternatively, the nonparametric PSHA is based on



Seismic Hazard, Fig. 4 Example of the product of PSHA. Map of seismic hazard of the world. Peak ground acceleration expected at 10% probability of exceedance at least once in 50 years. The Global Earthquake Model (GEM) project

seismicity distributions, which are assessed by data-based kernel functions.

By their nature, nonparametric procedures function well in areas of high seismic activity and when the record of past seismicity is considerably complete. However, the nonparametric approach has significant weak points, as it provides an unreliable estimate of small probabilities for areas of low seismicity. Additionally, the procedure is not recommended for an area where the seismic event catalogues are incomplete. Moreover, in its present form, the procedure is not capable of using any additional geophysical or geological information to supplement the seismological data. Therefore, a procedure that accommodates the incompleteness of the seismic event catalogues and, at the same time, does not require the specification of seismic sources would be an ideal tool for analyzing and assessing the seismic hazard. Such a procedure, which can be classified as a *parametric-historic* procedure for PSHA, has been used successfully in several parts of the world (see, e.g., Giardini 1999). In some instances, during the computation process of seismic hazard, the Gutenberg–Richter frequency–magnitude relation is extended to account for the contribution of tectonic faults and characteristic earthquakes. The procedure accepts the potential contribution of seismicity from active faults and compensates for the incompleteness of seismic event catalogues. The final maps of seismic hazard are smoothed by a Gaussian-type kernel function. This conceptually simple and intuitive parametric–historic approach combines the best of the deductive and nonparametric historical procedures and, in many instances, is free from the disadvantages characteristic of each of the procedures.

Another approach to assessing seismic hazard is based on Monte Carlo simulations of long-term seismic catalogues. Despite this approach having several advantages and being straightforward and easy to apply, its role in PSHA is highly underestimated. The Monte Carlo-based procedure allows easy incorporation of alternative earthquake space–time–size occurrence models, such as Markovian, non-Poissonian, and non-Gutenberg–Richter frequency–magnitude relation. In addition, it allows for the assessment of uncertainties associated with the input data and model parameters. The potential disadvantage of the Monte Carlo-based procedure is the requirement for extensive computation power.

Alternative Earthquake Recurrence Models

Time-dependent models. In addition to the classical assumption that earthquake occurrence in time follows a Poissonian process, alternative approaches are used occasionally. These procedures attempt to assess the temporal or spatiotemporal dependence of seismicity. Time-dependent earthquake occurrence models specify the distribution of the time to the next earthquake, where this distribution depends on the magnitude of the most recent earthquake. In order to incorporate the memory of past events, non-Poissonian distributions or

Markov chains are applied. In this approach, the seismogenic zones that recently produced strong earthquakes become less hazardous than those that did not rupture in recent history.

Time-dependent models are also applied to quantify the occurrence of large earthquakes on segments of active faults. Kramer (1996) provides an overview of non-Poissonian models. Several time-dependent models play an important role in PSHA, of which the best-known are the two models by Shimazaki and Nakata (1980). These two models of earthquake occurrence, namely, a *time-predictable* and a *slip-predictable* model, are based on the correlation of seismic activity with earthquake-related coastal uplift in Japan.

The time-predictable model states that earthquakes occur when the accumulated stress on a fault reaches a critical level. However, the stress drops, and the magnitudes of subsequent earthquakes vary among seismic cycles. Therefore, assuming a constant fault slip rate, the time to the next earthquake can be estimated from the slip of the previous earthquake. The slip-predictable model is based on the assumption that irrespective of the initial stress on the fault, an earthquake occurrence always causes a reduction in stress to the same level. Therefore, the fault slip in the next earthquake can be estimated from the time since the previous earthquake (Shimazaki and Nakata 1980; Scholz 1990; Thenhaus and Campbell 2003).

The second group of time-dependent models is based less firmly on the physical considerations of earthquake occurrence, and these models attempt to describe intervals between the consecutive events by specified statistical distributions. At least two such models play a significant role in the current practice of PSHA, namely, the log-normal and Brownian passage time (BPT) renewal model.

Employing a log-normal model was justified by the discovery that normalized intervals between the consecutive large earthquakes in the Circum-Pacific region follow a log-normal distribution, with an almost constant standard deviation. Following this discovery, the log-normal model became a key component of time-dependent PSHA procedures.

More often applied nowadays than the log-normal model is the Brownian passage time (BPT) distribution. This model is described by two parameters, μ and σ , which, respectively, represent the mean time interval between the consecutive earthquakes and the standard deviation. The aperiodicity of earthquake occurrence, as described by the BPT model, is controlled by the variation coefficient $\alpha = \sigma/\mu$. For a small α , the aperiodicity of earthquake occurrence is small, and the shape of the distribution is almost symmetrical. For a large α , the shape of the distribution is similar to the log-normal model, i.e., skewed to the right and peaked at a smaller value than the mean. The straightforward control of aperiodicity of earthquake occurrence by parameter α makes the BPT model extremely attractive, and it has been used to model earthquake occurrence in many parts of the world.

Several comparisons between time-dependent and time-independent earthquake occurrence models have shown that the time-independent (Poissonian) model is applicable to most engineering applications of the PSHA. The exception is when the time interval from the occurrence of the last earthquake exceeds the mean time interval between consecutive events related to a seismic hazard dominated by a single seismic source with a significant component of characteristic earthquakes. However, in many instances, the information on strong seismic events provided by current earthquake databases is insufficient to distinguish between different models. Therefore, using non-Poissonian models is justified only if sufficient data are available.

Clearly, such time-dependent models can result in a more realistic PSHA, but many of them have not yet reached the level of development required by routine engineering applications.

Spatiotemporal models. Often, a sequence of foreshocks and a sequence of aftershocks after the main event can be observed. The fore- and aftershock sequences cannot be described by the homogenous Poissonian process, as their rate depends on time, the magnitude of the main event, and its location. Some areas are notorious for the generation of earthquake swarms. Frequently, individual tectonic faults interact with each other, and such seismic activity leads to earthquakes being clustered in both time and space.

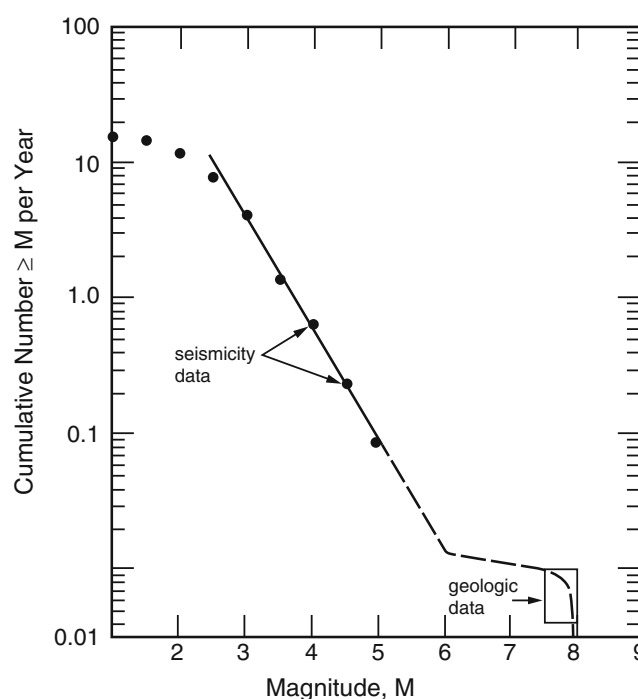
The best-known spatiotemporal model of earthquake occurrence is ETAS (epidemic-type aftershock sequence) and its mutations. The core assumption of the ETAS model is that each earthquake, however small, triggers its own aftershocks (offspring) independently of the others. The probability distributions of location and magnitude of the triggered event are dependent on the magnitude and location of the triggering event. The rate of aftershock occurrence decays with time according to an inverse power law known as the modified Omori law. The magnitudes of all the events, including background events and their offspring, are independent random variables following the same probability distribution. An alternative spatiotemporal model is one in which each mainshock and its aftershocks are treated as spatially and temporally independent clusters. However, the mathematical formulation does not have a closed form, which restricts the usage of the model. An elegant, straightforward, and statistically consistent incorporation of foreshocks and aftershocks into standard PSHA is done by Iervolino et al. (2014). The proposed formalism provides a remedy to one of the most controversial current practices of PSHA, known as declustering (i.e., removal) of “dependent” events from the seismic event catalogue.

Alternative frequency–magnitude models. In the classic Cornell–McGuire procedure for PSHA assessment, it is assumed that earthquake magnitudes follow the Gutenberg–Richter relation truncated from the top by a seismic source

characteristic, the maximum possible earthquake magnitude m_{max} . The PDF of this distribution is given by Eq. (5).

Generally, the Gutenberg–Richter relation describes magnitude distributions within seismic source zones adequately; however, in some instances, it appears not to apply. In such an instance, the relationship (5) must be modified. In some places, particularly in areas known for paleo-earthquakes, seismic belts, and large faults, the Gutenberg–Richter relation underestimates the occurrence of large magnitude events. Consequently, the continuity of the distribution (5) breaks down. The distribution is adequate only for events up to a magnitude of 6.0–7.0. Larger events tend to occur within a relatively narrow range of magnitudes (7.5–8.0) but with a frequency higher than that predicted by the Gutenberg–Richter relation (5). These events are known as *characteristic earthquakes* (Youngs and Coppersmith 1985, Fig. 5). However, it must be noted that the characteristic earthquake model is not free from controversy. Some researchers believe that the frequency–magnitude Gutenberg–Richter relation satisfies the earthquake occurrence characteristics equally well as does the earthquake characteristic model. The discrepancy between the two models would disappear if a sufficiently long record of seismicity were available.

When a seismic hazard is caused by seismicity of anthropogenic origin, an alternative to the Gutenberg–Richter



Seismic Hazard, Fig. 5 Gutenberg–Richter characteristic earthquake magnitude distribution. The model combines the Gutenberg–Richter frequency–magnitude relation with a uniform distribution of characteristic earthquakes. The model predicts higher rates of exceedance at magnitudes near the characteristic earthquake magnitude. (After Youngs and Coppersmith 1985)

frequency–magnitude relation (1) is often required. For example, the magnitude distributions of tremors generated by mining activity are multimodal and change their shape in time (Gibowicz and Kijko 1994). Often, the only possible method that could lead to a satisfactory PSHA for mining or fracking areas is replacing the analytical, parametric frequency–magnitude distribution by its model-free, nonparametric counterpart.

Two additional modifications of the recurrence models are introduced regularly: one is when earthquake magnitudes are uncertain and the other when the seismic occurrence process comprises temporal trends, cycles, short-term oscillations, and purely random fluctuations. The effect of error in earthquake magnitude determination (particularly significant for historical events) can be minimized by the simple procedure of correcting the earthquake magnitudes in a catalogue (e.g., McGuire 2004). Modelling random fluctuations in earthquake occurrence is often done by introducing compound (mixed) distributions, in which the parameters of earthquake recurrence models are treated as random variables (Kijko et al. 2016).

Ground Motion Models

The assessment of seismic hazard at a site requires knowledge of the GMM, i.e., the prediction equation of strong-motion parameter as a function of distance, earthquake magnitude, faulting mechanism, and, often, the local site condition. The most common form of the GMM is

$$\ln(y) = \text{const} + g_1(m) + g_2(r) + g_3(\text{faulting system}) + g_4(\text{soil}) + \varepsilon. \quad (17)$$

In Eq. (17), y is the amplitude of the designated ground motion parameter. This parameter can be the peak ground acceleration (PGA), spectral acceleration, displacement, the Modified Mercalli (MM) Intensity, seismic record duration, and the like. Functions $g_1(m)$ and $g_2(r)$ are the function of earthquake magnitude and the function of the shortest earthquake distance from the site to the earthquake source, respectively. Function $g_3(\cdot)$ represents the faulting mechanism, $g_4(\cdot)$ describes the site effect, and ε is the random error with zero mean and standard deviation $\sigma_{\ln(y)}$. Assuming that $\ln(y)$ has a normal distribution, and following the properties of log-normal distribution, regression of (17) provides the mean value of $\ln(y)$, the exponent of which corresponds to the median value of y , $\bar{y}$. As the log-normal distribution is positively skewed, the mean value of y , $\bar{y}$, exceeds the median value $\bar{y}$ by a factor of $\exp\left(-0.5\sigma_{\ln(y)}^2\right)$. This indicates that the seismic hazard for a particular site is higher when expressed in terms of $\bar{y}$ when the hazard for the same site is expressed in terms of $\bar{y}$.

In general, GMMs are built in three distinct ways: (a) empirically, by application of observed ground motion characteristics; (b) by stochastic methods, which are based on the physical properties of the earthquake source and seismic wave travel path; and (c) by the hybrid empirical method.

Empirical methods are applied in areas of high seismicity in the presence of dense instrumental networks, where a significant number of ground motion records are available. Examples of such empirical GMMs are those developed as part of the Next Generation Attenuation Phase 2 (NGA-West2) project, managed by the Pacific Earthquake Engineering Research Center (PEER). Stochastic methods are used mainly in areas of low seismicity, where only a limited number of observations are available. The hybrid empirical method, like the stochastic method, is used to develop GMM for areas with limited ground motion records; however, the technique used is different. In the hybrid empirical method, the ground motion in the target area is predicted from the empirical GMM in the host region with the help of adjustment factors between the two regions. These factors consider the differences in earthquake source, wave propagation, and site response characteristics between the two regions.

The ground motion models remain a particularly important component of PSHA, as ground motion uncertainty is a major contributor to the uncertainty of the PSHA. An excellent discussion on all the aspects of the estimation, application, and role of GMMs in PSHA is provided by McGuire (2004).

Uncertainties in PSHA

Contemporary PSHA distinguishes between two types of uncertainty, namely, aleatory and epistemic.

Aleatory uncertainty derives from the randomness in nature and is inherent for any random phenomenon. It represents unique details of an earthquake, such as its source, path, and site, and cannot be known before the earthquake's occurrence. It is sometimes referred to as "randomness," "stochastic uncertainty," or "inherent variability" (SSHAC 1997). Typical examples of aleatory uncertainties are the number of future earthquakes in a specified area, parameters of future earthquakes as origin times, epicenter coordinates, depths and their magnitudes, size of the fault rupture, and associated stress drop. Additionally, the uncertainty of ground motion parameters at the given site is also aleatory by their nature. The aleatory uncertainties are characteristic of the current model and cannot be reduced by the incorporation of additional data. Aleatory uncertainties are included in the PSHA by means of integration over these uncertainties and are represented by the hazard curve.

Epistemic uncertainty is the uncertainty deriving from insufficient knowledge about the model or its parameters. In

the past, epistemic uncertainty was known as statistical or professional uncertainty (McGuire 2004). Examples of epistemic uncertainties are the boundaries of seismic sources, distributions of seismic source parameters (e.g., annual rate of seismic activity λ , b -value, and $m_{\max}$), or median value of the ground motion parameter, given the source properties. The model (in the broad sense of its meaning, as, e.g., a particular statistical distribution) can be inexact and, therefore, predicts values that differ from the observed values by a fixed, but unknown, amount. When uncertainties are associated with the numerical values of the parameters, they are also epistemic by nature. Epistemic uncertainty can be reduced by incorporating additional information or data. The epistemic distributions of model parameters can be updated using the Bayes' formalism. Epistemic uncertainties are included in the PSHA through the use of alternative hypothesis, different sets of parameters with different numerical values, different models, or a *logic tree*.

The major benefit of the separation of uncertainties into aleatory and epistemic is potential guidance in the preparation of input for PSHA and the interpretation of the results. However, the division of uncertainties into aleatory and epistemic is model dependent and, to a large extent, arbitrary, indefinite, and often confusing.

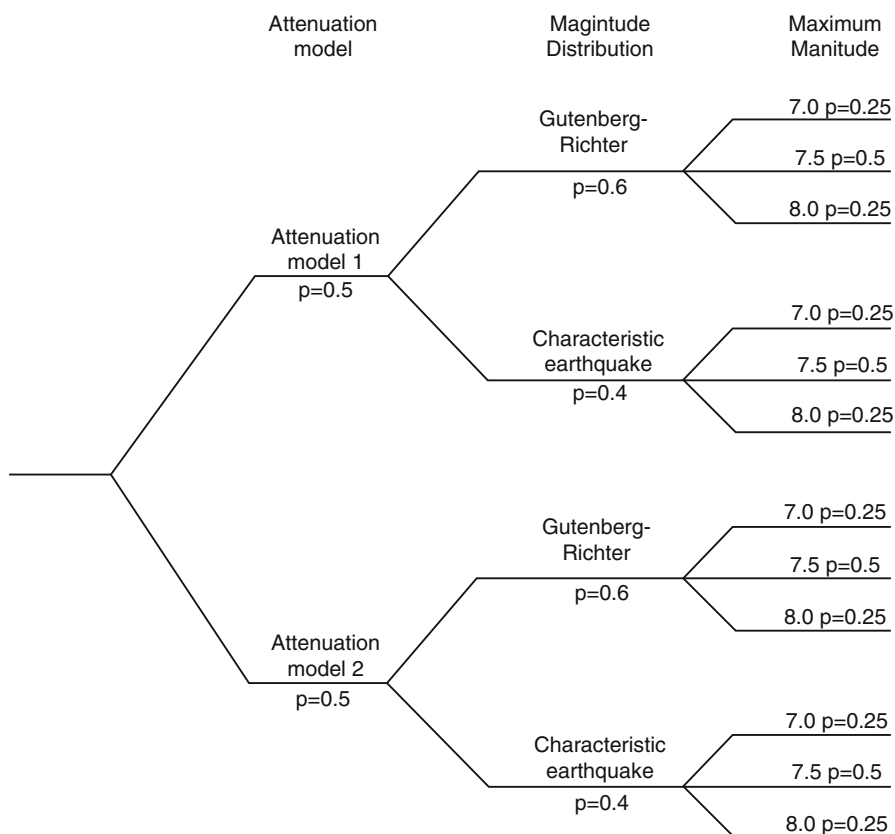
Logic Tree

The mathematical formalism of PSHA computation integrates over all the random (aleatory) uncertainties of a particular seismic hazard model. In many instances, however, because of the lack of information on the factors controlling the mechanism of the earthquake generation and wave propagation processes, the best choices for the elements of the seismic hazard model are not clear. The uncertainty could originate from the choice of alternative seismic sources and earthquake generation mechanisms, rupture type and directivity, competitive earthquake recurrence models and their parameters, particular site and its topography, settings of the recording instruments, and the chosen ground motion model. The standard approach to the explicit treatment of alternative hypotheses, models, and parameters is applying the *logic tree* formalism (Coppersmith and Youngs 1989). Although not without controversy, the logic tree formalism provides a convenient tool for the quantitative treatment of any alternatives. Each node of the logic tree (Fig. 6) represents uncertain assumptions, models, or parameters, and the branches extending from each node are the discrete uncertainty alternatives (McGuire 2004).

In logic tree analysis, each branch is weighted according to its probability of being correct. As a result, each end branch

Seismic Hazard,

Fig. 6 Example of a logic tree. The alternative hypothesis accounts for uncertainty in the ground motion attenuation relation, magnitude distribution model, and the assigned maximum magnitude $m_{\max}$



represents a hazard curve with an assigned weight, where the sum of weights of all the hazard curves is equal to 1.0. The derived hazard curves are used to compute the final (e.g., mean) hazard curve and its confidence intervals. Figure 6 shows an example of a logic tree, accounting for uncertainty in the ground motion attenuation model, magnitude distribution model, and assigned maximum magnitude $m_{\max}$.

It must be noted that the choice of the structure of the logic tree and its respective weights is critical and often extremely subjective.

Controversy

Although the PSHA procedure was formulated more than 50 years ago, controversy still surrounds it. Among others, the disagreements are related to (1) the requirement of declustering of the seismic event catalogue, (2) absence of the upper limit of ground motion parameters, (3) division of uncertainties between aleatory and epistemic, (4) correctness of the characteristic earthquakes model, and (5) accounting for epistemic uncertainty by applying the logic tree formalism.

Declustering, i.e., removing earthquakes that occur in clusters, such as foreshocks, aftershocks, and swarms, from the seismic catalogue (in order to create a catalogue of “independent” events) is probably the most controversial factor in current PSHA practice. The procedure violates the fundamental rule of physics taught to first-year students, i.e., “never twist data to suit theory; instead, design such theory, which suits the data.” However, declustering does the exact opposite. Declustering affects the estimation of the earthquake occurrence parameters (the magnitude activity rate and the b -value), and, thereby, it affects (usually underestimates) the true hazard and risk. Iervolino et al. (2014) show that in terms of the annual rate of exceedance of PGA, declustering leads to 30% underestimation of the true hazard.

Among others, Cornell and Winterstein (1988) contend that, in most instances, correct seismic hazard analysis can be performed without declustering. They show that the classic, time-independent (Poissonian) model is applicable even when the temporal clustering of earthquake occurrence is close to 30%. In addition, declustering can be avoided by replacing the Poissonian model (as well as the exponential Gutenberg–Richter magnitude distribution) with their compounded (mixed) counterparts.

Further criticism of the current PSHA procedure concerns the absence of the upper limit of the ground motion parameter in the hazard curve. The core of the classic Cornell–McGuire PSHA is the assumption that the logarithm of the ground motion parameter is distributed normally. As the domain of the normal distribution is unlimited from the right side, it results in a nonzero probability of unrealistically high values

for the ground motion parameter. Several solutions have been suggested; however, none of these is physically justifiable. The absence of the upper bound of the earthquake-generated ground motion parameter has been identified as the “missing piece” in current seismic hazard formalism. However, the problem of estimating the site-characteristic upper limit of PGA has been solved by Pavlenko (2017), who contends that, as a maxima of acceleration time series, PGA should follow an extreme value distribution, not the normal distribution. Based on extensive statistical tests of PGA records by K-NET and KiK-net networks, Pavlenko demonstrates that the distribution of $\ln(\text{PGA})$ converges to the generalized extreme value (GEV) distribution. Replacing the normal distribution in Eqs. (4) and (5) by the GEV distribution allows for the assessment of the upper bound value of PGA. Therefore, the GEV distribution provides a model that is physically justifiable and describes the scatter of the logarithm of PGA significantly better than does the currently used normal distribution, and its application provides the required upper bound of the seismic hazard curve.

Furthermore, criticism surrounds the division of uncertainties into aleatory and epistemic in the current PSHA procedure. As noted in section “[Uncertainties in PSHA](#),” the division remains an unresolved issue.

Different criticism is directed at the ergodic assumptions that underlie the formalism of the PSHA procedure. The ergodic process is a random process in which the distribution of a random variable in space is the same as the distribution of that variable at a single point when sampled as a function of time. It has been shown that the major contribution to PSHA uncertainty derives from the uncertainty of the ground motion model. The uncertainty of the ground motion parameter y is characterized by its standard deviation, $\sigma_{\ln(y)}$, which is calculated as the misfit between the observed and the predicted ground motions at several seismic stations for a small number of recorded earthquakes. Therefore, $\sigma_{\ln(y)}$ mainly characterizes the spatial and not the temporal uncertainty of ground motion at a single point. This violates the ergodic assumption of the PSHA procedure, leading to an overestimation of seismic hazard, particularly when the exposure times are longer than the earthquake return times.

Castanos and Lomnitz (2002) and Mulargia et al. (2017) have voiced equally strong criticism about the application of the logic tree formalism. They describe the application of the logic tree formalism as a misunderstanding in probability and statistics, as it is fundamentally wrong to admit “expert opinion as evidence on the same level as hard earthquake data.”

The science of seismic hazard assessment is, therefore, subject to much debate. Currently, PSHA represents a best-effort approach to quantify an issue of which not enough is known to provide definitive results and, as indicated by the numerous estimations, considerably more time and measurements will be needed before these issues can be resolved.

Further reading: Several excellent studies are available that describe all aspects of modern PSHA. Thenhaus and Campbell (2003) and McGuire (2004) present outstanding overviews covering theoretical, methodological, and procedural issues of modern PSHA. The most comprehensive treatment to date of all aspects of PSHA, including the treatment of aleatory and epistemic uncertainties, is provided by the SSHAC (1997) report. Seismic hazard from a geologist's perspective is described in a book by Yeats et al. (1997). Kramer (1996) provides an elegant, coherent, and understandable description of the mathematical aspects of both DSHA and PSHA. The information presented here benefited from all the sources quoted above, particularly the exceptional work by Kramer (1996).

Summary

The term seismic hazard refers to any physical phenomenon associated with an earthquake (e.g., ground motion, ground failure, liquefaction, and tsunami) and its effect on the land, infrastructure, and socioeconomic systems that, potentially, can produce a loss. The term is also used, without regard to any loss, to indicate the probable level of ground shaking occurring at a given point within a certain period of time. Seismic hazard analysis involves the quantification of the expected ground motion at particular sites. Seismic hazard analysis can be performed deterministically when a particular earthquake scenario is considered or probabilistically when the likelihood or frequency of a specified level of ground motion at a site during a specified exposure time is evaluated. In principle, any natural hazard caused by seismic activity can be described and quantified in terms of the probabilistic methodology. Classical probabilistic seismic hazard analysis includes four steps, which are (1) identification and parameterization of the seismic sources, (2) specification of temporal and magnitude distributions of earthquake occurrence, (3) calculation of ground motion prediction equations and their uncertainty, and (4) integration of uncertainties in earthquake location, earthquake magnitude, and ground motion prediction equations into the hazard curve.

An integral part of PSHA is the assessment of uncertainties. Contemporary PSHA distinguishes between two types of uncertainty, namely, aleatory and epistemic. Aleatory uncertainty stems from the randomness in nature and is the probabilistic uncertainty inherent in any random phenomenon. Aleatory uncertainties are characteristic of the currently used model and cannot be reduced by the incorporation of additional data. Epistemic uncertainty derives from insufficient knowledge about the model or its parameters. Epistemic uncertainty can be reduced by incorporating additional information or data. Aleatory uncertainties are included in the probabilistic seismic hazard analysis because of the integration over these uncertainties, and

they are represented by the hazard curve. In contrast, epistemic uncertainties are included through the use of alternative models, different sets of parameters with different numerical values, or a logic tree.

The PSHA procedure as known in its current form is not without controversy. The controversy arises from issues such as (1) the absence of the upper limit of the ground motion parameter, (2) division of uncertainties between aleatory and epistemic, and (3) the methodology itself, particularly the application of the logic tree formalism.

An extended version of this article is available at the author's ResearchGate profile https://www.researchgate.net/profile/Andrzej_Kijko/research.

Cross-References

- [Characteristic Earthquakes and Seismic Gaps](#)
- [Earthquake, Magnitude](#)
- [Earthquakes, Early and Strong Motion Warning](#)
- [Earthquakes, Intensity](#)
- [Earthquakes, ShakeMap](#)
- [Earthquakes, Strong-Ground Motion](#)
- [Seismic Zonation](#)
- [Statistical Seismology](#)

Bibliography

- Castanos H, Lomnitz C (2002) PSHA: Is it Science? Opinion Paper. *Eng Geol* 66:315–317
- Cooke P (1979) Statistical inference for bounds of random variables. *Biometrika* 66:367–374
- Coppersmith K, Youngs RR (1989) Issues regarding earthquake source characterization and seismic hazard analysis with passive margins and stable continental interiors. In: Gregersen S, Basham PW (eds) *Earthquakes at North Atlantic passive margins: neotectonics and postglacial rebound*. Kluwer, Dordrecht, pp 601–631
- Cornell CA (1968) Engineering seismic risk analysis. *Bull Seismol Soc Am* 58:1583–1606
- Cornell CA, Winterstein SR (1988) Temporal and magnitude dependence in earthquake recurrence models. *Bull Seismol Soc Am* 78:1522–1537
- Giardini D (1999) The global seismic hazard assessment program (GSHAP) 1992–1999. *Ann Geofis* 42:957–1230
- Gibowicz SJ, Kijko A (1994) *An introduction to mining seismology*. Academic, San Diego
- Iervolino I, Giorgio M, Polidoro B (2014) Sequence-based probabilistic seismic hazard analysis. *Bull Seismol Soc Am* 104:1006–1012
- Kijko A (2004) Estimation of the Maximum Earthquake Magnitude $m_{\max}$. *Pure Appl. Geophys* 161:1655–1681
- Kijko A, Smit A, Sellevoll MA (2016) Estimation of earthquake hazard parameters from incomplete data files. Part III. Incorporation of uncertainty of earthquake-occurrence model. *Bull Seismol Soc Am* 106:1210–1222
- Kramer SL (1996) *Geotechnical earthquake engineering*. Prentice-Hall, Englewood Cliffs
- McGuire RK (2004) *Seismic hazard and risk analysis*. Earthquake Engineering Research Institute, Oakland. MNO-10

- Molina S, Lindholm CD, Bungum H (2001) Probabilistic seismic hazard analysis: zoning free versus zoning methodology. *Boll Geofis Teor Appl* 42:19–39
- Mulargia F, Stark PB, Geller RJ (2017) Why is probabilistic seismic hazard analysis (PSHA) still used? *Phys Earth Planet Inter* 264:63–75
- Pavlenko VA (2017) Estimation of the upper bound of seismic hazard curve by using the generalized extreme value distribution. *Nat Hazards* 89:19–33. <https://doi.org/10.1007/s11069-017-2950-z>
- Pisarenko VF, Lyubushin AA, Lysenko VB, Golubieva TV (1996) Statistical Estimation of Seismic Hazard Parameters: Maximum Possible Magnitude and Related Parameters. *Bull. Seism. Soc. Am.* 86:691–700
- Scholz CH (1990) *The Mechanics of Earthquakes and Faulting*. Cambridge: Cambridge University Press
- Shimazaki K, Nakata T (1980) Time-predictable recurrence model for large earthquakes. *Geophys Res Lett* 7:279–282
- Smit A, Stein A, Kijko A (2019) Bayesian Inference in Natural Hazard Analysis for Incomplete and Uncertain Data. *Environmetrics*, p. e2566. <https://doi.org/10.1002/env.2566>
- SSHAC – Senior Seismic Hazard Committee (1997) Recommendations for probabilistic seismic hazard analysis: guidance on uncertainty and use of experts. NUREG/CR-6372, UCR-ID-122160, Main Report 1. Prepared for Lawrence Livermore National Laboratory
- Thenhaus PC, Campbell KW (2003) Seismic hazard analysis. In: Chen WF, Scawthorn C (eds) *Earthquake engineering handbook*. CRC Press, Boca Raton, pp 8-1–8-50
- Veneziano D, Cornell CA, O'Hara T (1984) Historic method for seismic hazard analysis. Report, NP-3438. Electric Power Research Institute, Palo Alto
- Weichert DH (1980) Estimation of the earthquake recurrence parameters for unequal observation periods for different magnitudes. *Bull Seismol Soc Am* 70:1337–1346
- Woo G (1996) Kernel estimation methods for seismic hazard area source modeling. *Bull Seismol Soc Am* 86:353–362
- Yeats RS, Sieh K, Allen CR (1997) *The geology of earthquakes*. Oxford University Press, New York
- Youngs RR, Coppersmith KJ (1985) Implications of fault slip rates and earthquake recurrence models to probabilistic seismic hazard estimates. *Bull Seismol Soc Am* 75:939–964

Seismic Imaging, Overview

Gerard T. Schuster
Division of Environmental and Earth Sciences, King
Abdullah University of Science and Technology, Thule,
Saudi Arabia

Definition

Seismic imaging or tomography (tomo = slice and graph = picture) is a procedure for estimating the earth's rock parameters from seismic data. These rock parameters can be represented by the spatial distribution of, e.g., P-wave velocity, S-wave velocity, porosity, density, or anisotropic parameters. The result of inversion is graphically presented as a 2-D or 3-D grid of pixels, where each pixel contains the value of the model parameter of interest, which is P velocity in

Fig. 1g–h. Such tomograms are used to estimate the geometry and lithology of geologic layers, and can help exploration geophysicists and earthquake seismologists understand the evolution of the earth's interior.

There are five main types of seismic imaging: seismic migration, least squares migration, full waveform inversion (FWI), phase-like inversion, and migration velocity analysis. Four of these methods can be derived as special cases of finding the optimal model that minimizes a waveform or a phase-related misfit function. This entry will present these methods as they are applied to body wave events in seismic data, and overview the current best practices in waveform tomography.

Inversion Theory

There are four steps to inverting for the model $\mathbf{m}$ from seismic data $\mathbf{d}$. We will assume that the input data are either seismic traces, generated by man-made or earthquake sources, or some skeletonized part of the traces, such as first arrival traveltimes or the phase of an event at some frequency.

Step 1: $\mathbf{L}(\mathbf{m}) = \mathbf{d}$. Establish the mathematical relationship $\mathbf{L}(\mathbf{m}) = \mathbf{d}$ between the seismic data $\mathbf{d}$ and model $\mathbf{m}$.

$$\mathbf{d} = \mathbf{L}(\mathbf{m}), \quad (1)$$

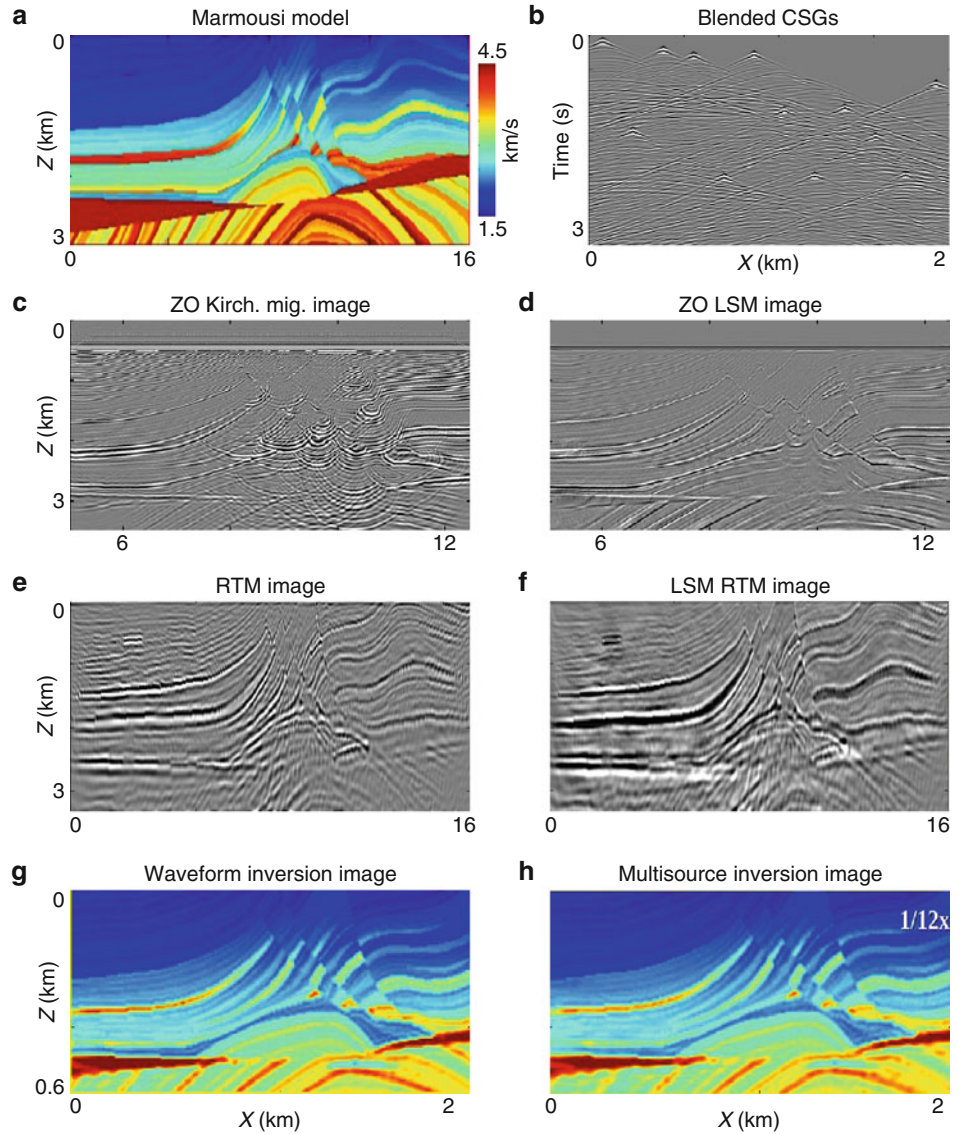
where $\mathbf{L}$ represents the forward modeling operator for the actual model. Equations A1–A3 in Table 1 show the three steps in establishing a linearized version of Eq. 1 for the acoustic wave equation, where the extensions to the elastic (Mora 1987) and anisotropic (Barnes et al. 2008; Operto et al. 2009) wave equations are tedious but straightforward.

Step 2: Discretize $\mathbf{m}$, $\mathbf{d}$, and $\mathbf{L}$. Discretize the 3D earth model into a 3D grid of N physical parameters (e.g., unknown slowness in each cell) and assemble the unknowns into the $N \times 1$ vector $\mathbf{m}$. Discretize the seismic traces in both space and time (or frequency) into an $M \times 1$ vector $\mathbf{d}$ of data measurements. In this case, $\mathbf{L}$ reduces to a $M \times N$ matrix. The forward modeling for FWI requires knowledge of the source wavelet, which can be estimated in a variety of ways: stacking of the direct arrival at different near-offset hydrophones in marine data, iterative inversion of the source wavelet (Mora 1987; Zhou et al. 1997; Pratt 1999), or deconvolution of the shot gather with a time-windowed near-offset trace (Sheng, personal communication). The time window can be a few periods long centered about the direct arrival.

Step 3: Linearize $\mathbf{L}\delta\mathbf{m} \approx \delta\mathbf{d}$. Linearize the nonlinear relationship between the data and model. Expanding the i th data measurement $d_i(\mathbf{m})$ to first order in $\delta\mathbf{m}$ by a Taylor series about a first-guess model $\mathbf{m}_0$ (close to the true model) gives the linearized estimate:

Seismic Imaging, Overview,

Fig. 1 (a) Marmousi velocity model, (b) time-shifted and blended shot gathers, (c) zero-offset Kirchhoff migration image, (d) zero-offset least squares migration image after 30 iterations (courtesy of Naoshi Aoki), (e) reverse time migration (RTM) image, (f) least squares RTM image (Dai and Schuster 2009), (g) waveform tomogram after 50 iterations using the prestack gathers, and (h) waveform tomogram where the inputs are eight supergathers, each with 12 phase-encoded shot gathers (Zhan and Schuster in press); images are at different scales from one another



$$\begin{aligned}
 d_i(\mathbf{m}) &\approx d_i(\mathbf{m}_0) + \sum_j \frac{\delta d_i(\mathbf{m}_0)}{\delta m_j} \delta m_j \rightarrow \delta d_i(\mathbf{m}) \\
 &= \sum_j \overbrace{[\partial d_i(\mathbf{m}_0) / \partial m_j]}^{\text{wavepathfunction}} \delta m_j,
 \end{aligned} \tag{2}$$

or in matrix-vector notation

$$\delta \mathbf{d} = \mathbf{L} \delta \mathbf{m}. \tag{3}$$

Here, $\partial d_i(\mathbf{m}_0) / \partial m_j = \lim_{\Delta m_j \rightarrow 0} [d_i(\mathbf{m}_0 + \hat{\mathbf{j}} \Delta m_j) - d_i(\mathbf{m}_0)] / \Delta m_j$ is the Fréchet derivative with respect to the j th parameter; the data residual $\delta d_i = [d_i(\mathbf{m}) - d_i(\mathbf{m}_0)]$ is the difference between the i th components of the predicted data vector $\mathbf{d}(\mathbf{m}_0)$ and the observed data vector $\mathbf{d}(\mathbf{m})$; and $\hat{\mathbf{j}}$ is the j th unit

vector in the finite-dimensional model space. The model perturbation $\delta \mathbf{m} = \mathbf{m} - \mathbf{m}_0$ is the difference between the actual model $\mathbf{m}$ and the guessed model $\mathbf{m}_0$, and $\mathbf{L}$ is now interpreted as the Jacobian matrix. Its elements $[\mathbf{L}]_{ij} = \partial d_i(\mathbf{m}_0) / \partial m_j$ determine the sensitivity of the data to the perturbations in the model. For a windowed arrival, the Jacobian plots out in model space as a wavepath for a single source-receiver pair (Woodward 1992; Luo 1992; Dahlen et al. 2002; Marquering et al. 2002; Montelli et al. 2004; van der Hilst and de Hoop 2006; Xu and Xie 2009); and the velocity variations within its first Fresnel zone mostly influence the event of interest. Eq. A3 shows that the linearized equations take the form of the Born approximation (Stolt and Benson 1986) to the Lippmann-Schwinger equation, and its kernel is the Jacobian given by Eq. A4.

For a single source and a single receiver in a smoothly increasing velocity medium, Eq. A4 plots out as a curved

Seismic Imaging, Overview, Table 1 Symbols and specific formulas for inverting the Helmholtz equation for the perturbed slowness distribution $\delta s(\mathbf{x})$, where $\delta s(\mathbf{x})$ is assumed to be small compared to the background slowness model $s(\mathbf{x})$. The harmonic pressure field is denoted by $p(\mathbf{x}|\mathbf{s})$ for a source localized about $\mathbf{s}$, and the Green's function for the Helmholtz equation is given by $G(\mathbf{x}|\mathbf{s})$ for a point source at $\mathbf{s}$ and observer

at $\mathbf{x}$. The source wavelet spectrum is $W(\omega)$ at the angular frequency ω and the body force term localized about the point $\mathbf{s}$ is denoted by $f(\mathbf{x}|\mathbf{s})$. The predicted (observed) traveltime of an event received at $\mathbf{g}$ for a source at $\mathbf{s}$ is denoted as $\tau(\mathbf{g}|\mathbf{s})$ ($\tau(\mathbf{g}|\mathbf{s})^{\text{obs.}}$); and $\varepsilon = \|\mathbf{d} - \mathbf{d}^{\text{obs.}}\|^2$ is the misfit function for waveform inversion

Symbol	Mathematical formula
(A1) Helmholtz equation: $\mathbf{A}(\mathbf{d}, \mathbf{m}) = \mathbf{f}$	$(\nabla^2 + \omega^2 s(\mathbf{x}))p(\mathbf{x} \mathbf{s}) = f(\mathbf{x} \mathbf{s})$ $\mathbf{m} \rightarrow s(\mathbf{x}); \mathbf{d} \rightarrow p(\mathbf{x} \mathbf{s})$
(A2) linearized Helmholtz equation: $\delta \mathbf{A}(\mathbf{d}, \mathbf{m}) = 0$	$(\nabla^2 + \omega^2 s(\mathbf{x}))\delta p(\mathbf{x} \mathbf{s}) = -2\omega^2 s(\mathbf{x})\delta s(\mathbf{x})p(\mathbf{x} \mathbf{s})$ $\delta \mathbf{m} \rightarrow \delta s(\mathbf{x}); \delta \mathbf{d} \rightarrow \delta p(\mathbf{x} \mathbf{s})$
(A3) Lippmann-Schwinger equation with born approx.: $\delta \mathbf{d} = \mathbf{L}\delta \mathbf{m}$	$\delta p(\mathbf{g} \mathbf{s}) = -2\omega^2 \int G(\mathbf{g} \mathbf{x})s(\mathbf{x})\delta s(\mathbf{x})W(\omega)G(\mathbf{x} \mathbf{s})d\mathbf{x}^3$ where $p(\mathbf{x} \mathbf{s}) = W(\omega)G(\mathbf{x} \mathbf{s})$
(A4) Jacobian or wavepath function: $\delta \mathbf{d}/\delta \mathbf{m}$	$\delta p(\mathbf{g} \mathbf{s})/\delta s(\mathbf{x}) = -2s(\mathbf{x})\omega^2 W(\omega)G(\mathbf{g} \mathbf{x})G(\mathbf{x} \mathbf{s})$
(A5) misfit gradient or reverse time migration or $\mathbf{L}^\dagger[\mathbf{L}\delta \mathbf{m} - \delta \mathbf{d}]$: $\delta \varepsilon/\delta \mathbf{m}$	$\delta \varepsilon/\delta s(\mathbf{x}) = \gamma \int G(\mathbf{g} \mathbf{x}) * \Delta d(\mathbf{g} \mathbf{s})G(\mathbf{x} \mathbf{s}) * dgds$ where $\gamma = -2\omega^2 s(\mathbf{x})W(\omega)$ and $\Delta d(\mathbf{g} \mathbf{s}) = p(\mathbf{g} \mathbf{s}) - p(\mathbf{g} \mathbf{s})^{\text{obs.}}$
(A6) least squares migration or linearized inversion: $\mathbf{m}^{(k)} - \delta \varepsilon/\delta \mathbf{m}$	$m(\mathbf{x})^{(k)} - \gamma \int G(\mathbf{g} \mathbf{x}) * \Delta d(\mathbf{g} \mathbf{s})^{(k)}G(\mathbf{x} \mathbf{s}) * dgds$
(A7) waveform inversion: $\mathbf{m}^{(k)} - \delta \varepsilon/\delta \mathbf{m}$	$m(\mathbf{x})^{(k)} - \gamma \int G^{(k)}(\mathbf{g} \mathbf{x}) * \Delta d(\mathbf{g} \mathbf{s})^{(k)}G^{(k)}(\mathbf{x} \mathbf{s}) * dgds$
(A8) wave equation traveltime inversion: $\mathbf{m}^{(k)} - \delta \varepsilon/\delta \mathbf{m}$	$m(\mathbf{x})^{(k)} - \gamma \int G^{(k)}(\mathbf{g} \mathbf{x}) * \Delta d(\mathbf{g} \mathbf{s})^{(k)}G^{(k)}(\mathbf{x} \mathbf{s}) * dgds$ where $\Delta d(\mathbf{g} \mathbf{s}) \approx p(\mathbf{g} \mathbf{s})^{\text{obs.}}(\tau(\mathbf{g} \mathbf{s}) - \tau(\mathbf{g} \mathbf{s})^{\text{obs.}})$

“fat” ray (trace a ray that connects the source and receiver and honors Snell’s law. Surround this ray with a “fat finite-frequency ray” such that the propagation of events from the source to receiver within the “fat” ray differs in traveltime by no more than 1/2 the period of the source wavelet. This fat ray region is that portion of the earth which mostly influences the traveltime of the event of interest) that connects the source and receiver points. This fat ray is denoted as a wavepath by Woodward (1992) and forms the basis of finite-frequency traveltime tomography. There seems to be a general (Montelli et al. 2004), but not a universal (van der Hilst and de Hoop 2006), agreement that finite-frequency tomography can be superior to that of ray-based tomography. As an example, earthquake seismologists use these wavepaths (renamed as banana-doughnuts) with finite-frequency tomography to invert earthquake traveltimes and surface wave data for deep mantle velocity variations attributed to, for example, ascending plumes of rocks (Dahlen et al. 2002; Marquering et al. 2002; Montelli et al. 2006).

Step 4: Solve $\mathbf{L}\delta \mathbf{m} = \delta \mathbf{d}$ by an iterative gradient method. Eq. 3 is typically an overdetermined, inconsistent, and poorly conditioned system of equations. Therefore, the solution we seek is the one that minimizes the sum of the data misfit $\|\mathbf{L}\mathbf{m} - \mathbf{d}\|_K^2$ and model penalty $\lambda^2 \|\mathbf{C}\mathbf{m}\|_W^2$ functions in the K and W norms (Clinthorne et al. 1993):

$$\varepsilon = \frac{1}{2} \|\mathbf{L}\delta \mathbf{m} - \delta \mathbf{d}\|_K^2 + \frac{\lambda^2}{2} \|\mathbf{C}\delta \mathbf{m}\|_W^2. \quad (4)$$

where $\|\mathbf{L}\delta \mathbf{m} - \delta \mathbf{d}\|_K^2 = (\mathbf{L}\delta \mathbf{m} - \delta \mathbf{d})^\dagger \mathbf{K}(\mathbf{L}\delta \mathbf{m} - \delta \mathbf{d})$ and $\mathbf{K}$ dimes taken as the inverse of the data covariance matrix (Tarantola 1987) that is real and symmetric. The penalty function (also known as the regularization term) is defined as $\|\delta \mathbf{C}\mathbf{m}\|_W^2 = (\mathbf{C}\mathbf{m})^\dagger \mathbf{W}(\mathbf{C}\mathbf{m})$, where the real symmetric matrix $\mathbf{W}$ might strongly weight certain regions in the model because they mostly influence the data. The matrix $\mathbf{C}$ might be a spatial second-derivative operator in one or several directions so as to encourage solutions with smoothly varying model parameters along selected directions.

A simplified steepest descent solution for the i th model parameter m_i is given by

$$m_i^{(k+1)} = m_i^{(k)} - \alpha \frac{\overbrace{\delta \varepsilon}^{\text{gradient Equation A5}}}{\delta m_i} \rightarrow \mathbf{m}^{(k+1)} \quad (5)$$

$$= \mathbf{m}^{(k)} - \underbrace{\alpha \mathbf{L}^\dagger \delta \mathbf{d}^{(k)}}_{\text{migration of data residual}},$$

where $\lambda = 0$; $\mathbf{K} = \mathbf{I}$; α is the step length; and preconditioning is used (Beydoun and Mendes 1989; Clinthorne et al. 1993; Causse et al. 1995) to accelerate convergence. A sequence of models is generated until the data residual falls below some acceptable level. In practice, a preconditioned conjugate gradient method (Mora 1987; Luo and Schuster 1991; Epanomeritakis et al. 2008), a Gauss-Newton Krylov solver (Erlangga and Hermann 2009), or a limited memory quasi-

Newton method (Pratt et al. 1998; Plessix 2006) for different implementations of the adjoint method (Plessix 2009) is implemented; and sometimes a direct matrix solver is used to find the Hessian inverse if the problem size is small enough (Pratt and Goulty 1991). Approximations to the Hessian by a systematic procedure can be found in Thierry et al. (1999). There is strong evidence (Cruse et al. 1990; Brossier et al. 2010) that using the L_1 norm, misfit function is noticeably more resistant to data noise than the L_2 norm misfit function, and the combination of the two is sometimes the best choice.

Five Types of Seismic Imaging Methods

Five types of seismic imaging methods and their resolution properties will be discussed (there are many seismic inversion methods, but the five types discussed here are often used in the geophysical community): migration (modest resolution of reflectivity), least squares migration (high resolution of reflectivity), full waveform inversion (high resolution of velocity), phase-like inversion (modest resolution of velocity), and migration velocity analysis (modest resolution of velocity). The high-resolution methods typically pay the price of increased computational cost and decreased robustness, compared to the moderate-resolution methods with relatively low cost and desirable robustness. For reflection migration and inversion, the spatial resolution limits are approximately defined by a generalized Radon transform analysis (Beylkin et al. 1985); and the resolution limits for ray-based transmission tomography (Williamson 1991) can be estimated by considering the width of the transmission Fresnel zone.

Migration is the first iterate solution of Eq. 5, and least squares migration, sometimes known as linearized inversion, is the final iterative solution where the operator $\mathbf{L}$ is not updated after each iteration; also, $\mathbf{m}(\mathbf{x})$ represents the reflectivity model at the position $\mathbf{x}$. In contrast to migration, the waveform inversion tomogram is the final iterative solution, where the velocity model $\mathbf{m}$ and $\mathbf{L}$ are updated after every iteration; waveform inversion falls under the class of non-linear optimization methods. It can be shown under certain assumptions that waveform inversion reduces to either wave equation traveltimes tomography (Luo and Schuster 1991) or ray-based tomography. These last two methods are classified as phase-like inversion methods. Unlike minimizing the data misfit function in Eq. 4, migration velocity analysis updates the velocity model to minimize a *model misfit function*, which is the normed difference between the predicted migration image and the actual migration image in the, e.g., common image gather. Convergence problems associated with local minima in the misfit function are reduced by emphasizing *flatness* in the CIG misfit function (Symes and Carazone 1991; Shen et al. 1993).

Migration

If the starting model is a smoothly varying velocity distribution that only generates the accurate first arrival, then the input data $\mathbf{d} - \mathbf{d}^{(0)}$ residual becomes the scattered data. In this case, the desired model is the reflectivity distribution, which is similar to the slowness perturbation function $\delta s(\mathbf{x})$, and the background velocity model is the inverse of slowness $s(\mathbf{x})$. If a finite-difference solution to the wave equation, for example, is used to calculate $G(\mathbf{x}|\mathbf{s})$ and $G(\mathbf{g}|\mathbf{x})$ in Eq. A5, then the first model update $\delta \mathbf{m}^{(1)}$ in Eq. 5 is known as the reverse time migration image (Whitmore 1983; McMechan 1983). If a oneway wave equation method is used to generate the Green's function, then the migration method is a phase-shift or FX-type algorithm (Stolt and Benson 1986; Claerbout 1992; Etgen et al. 2009). A diffraction-stack migration method results if the Green's function is replaced by its asymptotic approximation $G(\mathbf{x}|\mathbf{s}) = A(\mathbf{x}, \mathbf{s})e^{i\omega\tau_{xs}}$ (here, $A(\mathbf{x}, \mathbf{s})$ accounts for geometric spreading losses, and τ_{xs} is the first arrival time for a ray that connects the source at $\mathbf{s}$ with the observer at $\mathbf{x}$). A ray-tracing method can be used to compute these traveltimes for a sufficiently high frequency and smoothly varying medium [Bleistein et al. 2001]). An example of standard poststack migration is depicted in Fig. 1c, where the velocity model is shown in Fig. 1a. The migration image depicts the reflectivity distribution computed by a Kirchhoff migration method with an eikonal traveltimes solver.

Seismic migration is also used to migrate teleseismic body waves processed to form so-called receiver functions. Examples include Ryberg and Weber (2000), Sheehan et al. (2000), and also Bostock et al. (2001) who used a ray-Born inversion approach. More recently, Nowack et al. (2007) applied the Gaussian beam migration approach of Hill (2001) to the migration of teleseismic body waves. Despite its widespread use and its robustness, migration is considered to be a moderate-resolution method because it approximates the inverse Hessian matrix $[\mathbf{L}^\dagger \mathbf{L}]^{-1}$ by a diagonal matrix.

Least Squares Migration

If the background model $1/s(\mathbf{x})$ is not updated after each iteration (i.e., $\mathbf{L}$ in Eq. 5 is independent of the k index), then $\mathbf{m}^{(k+1)}$ for large k is known as the least squares migration image (Nemeth et al. 1999; Duquet et al. 2000). As in standard migration, the model to be iteratively updated is the reflectivity distribution and not the velocity model. The iterative least squares migration Eq. A6 is interpreted as a sequence of standard migrations, where the data residual is back projected into the earth model by the migration operator $\mathbf{L}^\dagger$. Least squares migration (LSM) is also known as linearized waveform inversion (Lailly 1984; Tarantola 1986, 1987; Jin et al. 1992; Lambaré et al. 1992) and is superior to standard migration by reducing migration artifacts caused by a poor acquisition geometry; it also can provide a spatial resolution that is more than twice (Nemeth et al. 1999; Yu

et al. 2006) that of standard migration if the migration velocity is sufficiently accurate. Its main drawbacks are that its effectiveness is very sensitive to the accuracy of the migration velocity model, and it can be more than an order of magnitude more expensive than standard migration. (Recent developments [Dai and Schuster 2009] in phase-encoded migration suggest a great reduction in the cost of LSM.) As an example, Fig. 1d depicts the LSM image obtained from zero offset (ZO) data, which is more accurate than the ZO standard migration image in Fig. 1c. Figure 1e–f depict the RTM and LSM RTM images obtained from the prestack shot gathers.

Full Waveform Inversion

If the background model is updated after each iteration, then Eq. A7 is known as nonlinear waveform inversion (Tarantola 1987; Mora 1987; Mora 1989; Song et al. 1995; and many others); a common designation for waveform inversion is FWI or full waveform inversion. Unlike LSM or standard migration, Eq. A7 for FWI iteratively updates the velocity and reflectivity models so that the final velocity image can be much improved in both resolution and accuracy. Its main drawbacks are a tendency to get stuck in local minima, it is computationally expensive compared to standard migration, and there might be more than one model that can explain the same data, i.e., a nonunique solution. For example, elastic isotropic modeling codes can sometimes generate predicted traces that adequately fit the observed data, but the estimated isotropic velocity model is inconsistent with the actual anisotropic rocks in the real earth.

Successful examples of waveform tomography images are shown in Fig. 1g–h and were obtained from the same shot gathers used for migration. The final tomogram shows the velocity distribution that is almost identical to that of the actual model. One of the keys to success in waveform inversion is that a good starting model is often required for an accurate result. This starting model can be obtained by migration velocity analysis (Stork 1992; Jousset et al. 2009), reflection traveltime tomography (Langan et al. 1985; Bishop et al. 1985; Nolet 1987; Zelt 2002), or refraction tomography (Pratt and Goult 1991; Luo and Schuster 1991; Min and Shin 2006; Sheng et al. 2006).

A major challenge to the success of waveform inversion is the limited offset range between sources and receivers and the lack of low-frequency information in the recorded data (Sirgue and Pratt 2004; Barnes et al. 2008; Boonyasiriwat et al. 2009; Kelly et al. 2009). Such deficiencies can prevent waveform inversion from reconstructing the low wave-number parts of the model, and sometimes prevent convergence to any type of reasonable model. Remedies being explored include the possibility of recording data at much lower frequencies with more capable recording devices, and

obtaining very wide-offset data. Other challenges address the validity of the acoustic approximation versus the reality that the recorded data are largely of elastic (Vigh et al. 2009) or viscoelastic nature (Causse et al. 1999; Askan 2006). Elastic inversion of elastic seismograms (Mora 1987; Mora 1989; Zhou et al. 1997; Brossier et al. 2009) have been presented, but the acoustic approximation can still provide useful results. One approach to viscoelastic inversion is to invert for the acoustic velocity first, then follow this with inversion for attenuation parameters (Kamei and Pratt 2008); another approach is to use a ray-based method to invert for the attenuation factor Q and then use this Q to correct for the attenuation in the data (Pratt et al. 2005; Sheng et al. 2007; Boonyasiriwat et al. 2009).

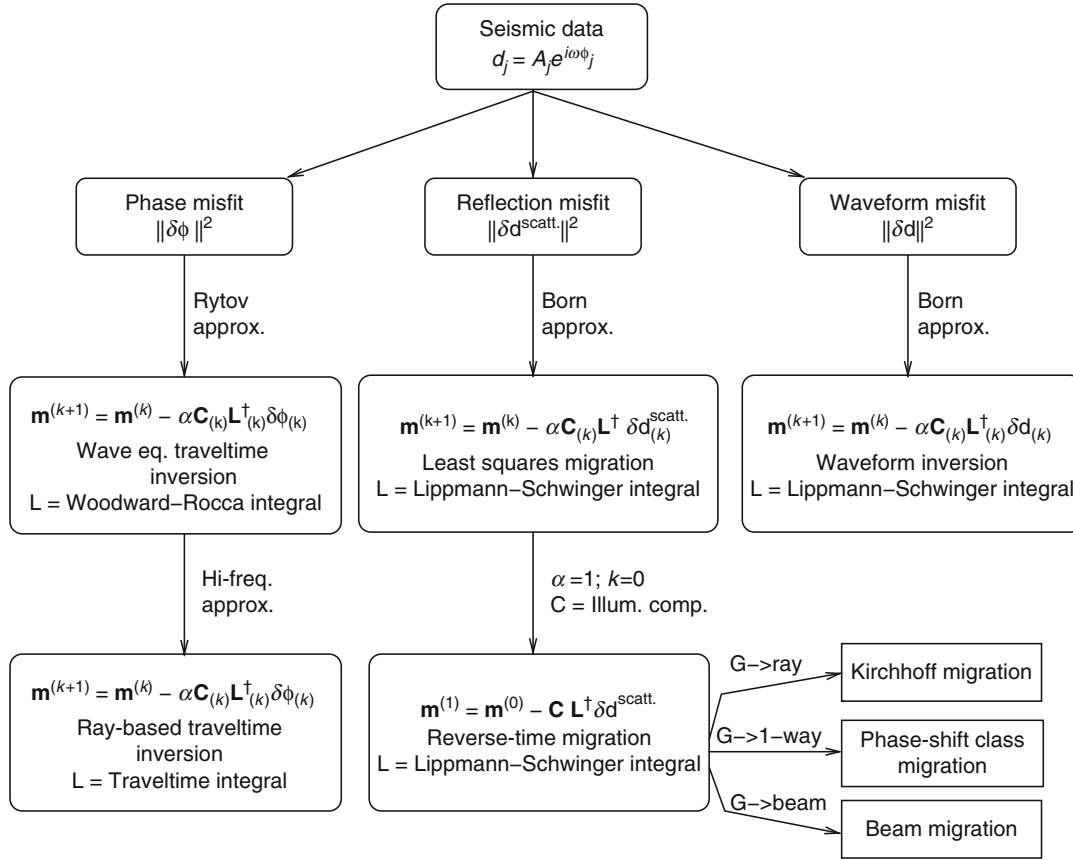
Wave Equation Traveltime Inversion

If the data residual $\delta d(\mathbf{g}|\mathbf{s})$ is replaced by the traveltime residual $\delta \tau(\mathbf{g}|\mathbf{s})$ weighted by the recorded trace $d(\mathbf{g}|\mathbf{s})^{\text{obs}}$, then this is known as wave equation traveltime (Luo and Schuster 1991) tomography (WT); it is a variant of Rytov inversion (Woodward 1992) and updates the velocity model by smearing weighted traveltime (not waveform) residuals over the associated wavepaths (Woodward 1992; Luo 1992). In the high-frequency limit, it reduces to ray-based traveltime tomography (RT). The advantage of WT over RT is that it does not require a high-frequency approximation and accounts for the band-limited nature of waves as they propagate through the earth. Shin et al. (2002, 2007) use a modified logarithm norm to invert for phase data, which bears a close relationship to the Rytov inversion method; and Effelsen (2009) compares phase inversion to traveltime tomography for inverting refraction events. Figure 2 illustrates how the above seismic imaging methods are related to one another.

The main disadvantage of WT is that it is at least an order of magnitude more expensive than RT because it requires a finite-difference solution of the wave equation rather than a ray-traced approximation. Earthquake seismologists often see wave equation traveltime as a major improvement in estimating the earth's velocity parameters from earthquake records (Dahlen et al. 2002; Marquering et al. 2002; Montelli et al. 2006; van der Hilst and de Hoop 2006).

Migration Velocity Analysis

The previous imaging methods can be described as estimating the earth model by predicting synthetic data that *best matches the observed data in the data domain*. In contrast, migration velocity analysis (MVA) finds the velocity model that *best flattens or focuses the migration sections in the image domain*; here, the image domain is the migration cube in the (x, y, z) indices and some other index s such as shot index, receiver index, source-receiver offset, (Yilmaz and Chambers 1984; Faye and Jeannot 1986; Al-Yahya 1989; Toldi 1989; Stork 1992; Lafond and Levander 1993), common angle



Seismic Imaging, Overview, Fig. 2 Three types of seismic imaging methods: phase inversion, least squares migration (LSM), and nonlinear waveform inversion. Note that LSM is a linearized inversion so that the modeling operator $\mathbf{L}$ does not get updated after each iteration. This compares to nonlinear waveform inversion which updates $\mathbf{L}$ after each iteration. The Lippmann-Schwinger-type and Woodward-Roca-type integrals are displayed in Table 1, and details about various migration methods are in the Seismic Migration section in the *Encyclopedia of*

Solid Earth Geophysics. The traces that only contain scattered arrivals are obtained by muting the direct arrivals; and the ensemble of these traces is symbolized by the vector $\mathbf{d}^{\text{scatt.}}$. Instead of computing the Green's functions in the integral equations by a finite-difference solution to the wave equation, various approximations such as ray-based Green's functions (Kirchhoff migration), one-way wave equation approximations (phase-shift type migration), and Gaussian Beam (beam-like migration) are used for migration

parameter (Xu et al. 1998), or subsurface offset coordinate (Rickett and Sava 2002; Sava and Biondi 2004; Robein 2010).

To understand why we seek a velocity model that flattens events in the image domain, consider the 2D migration image $m(x, z, s_0)$ in Fig. 3a obtained by migrating one shot gather (the source is at the surface with shot index s_0). If the migration velocity is correct, then all of the single-shot migration images should be similar in appearance. This means that all of the migrated reflection events should be flat in the common image gather (CIG) given by $m(x_{\text{CIG}}, z, s)$ for all values of s, z , and a fixed value of x_{CIG} (see Fig. 3b). Note that if the migration velocity is accurate, the reflector boundary denoted by the dotted horizontal line will be flat for a common image gather. If the migration velocity is too slow, then the imaged reflector boundary will curve upward as illustrated by the curved dashed line, and if too fast, the imaged reflector will curve downward.

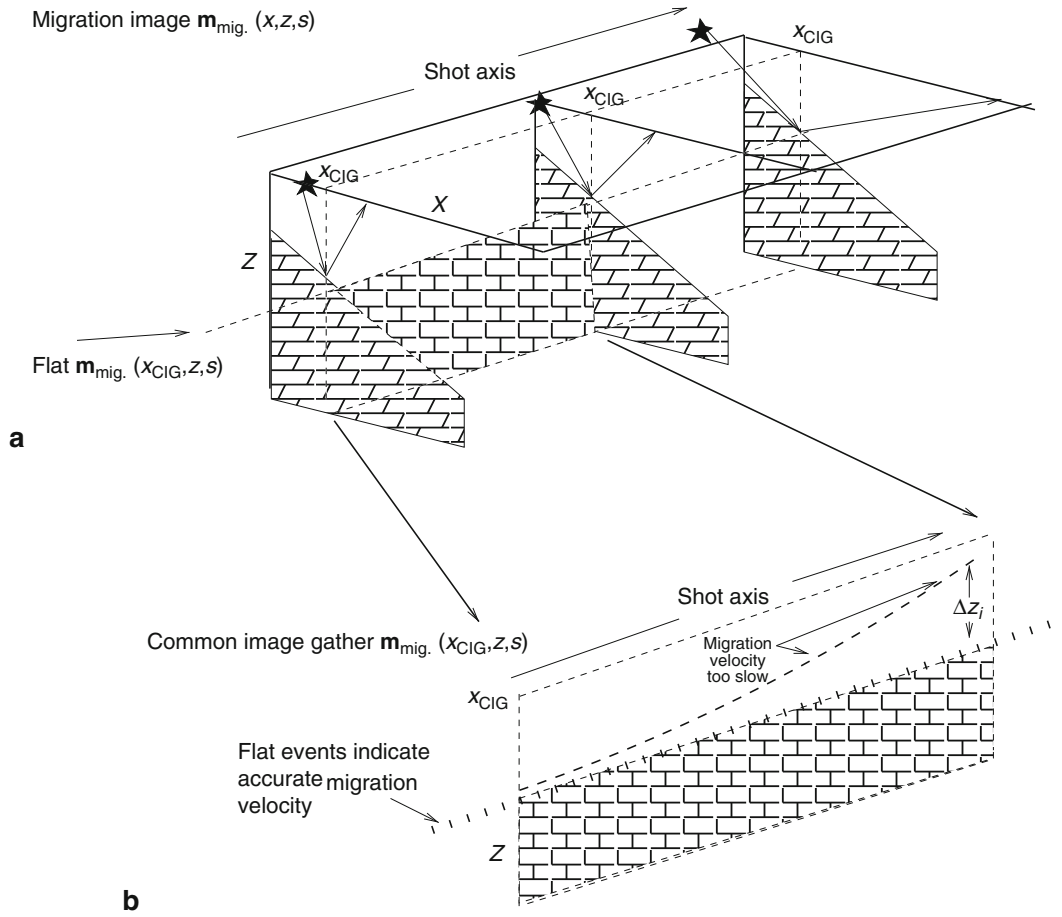
With MVA, the goal is to find the velocity model so that, ideally, the predicted migration image $\mathbf{m}_{\text{mig.}}^{\text{pred.}} = \mathbf{L}^\dagger \mathbf{L} \mathbf{m}$ best fits the actual migration image $\mathbf{m}_{\text{mig.}} = \mathbf{L}^\dagger \mathbf{d}$. The associated misfit function can be constructed so that it is similar to that of Eq. 4, except that the norm of the *migration* residual

$$\varepsilon = \frac{1}{2} \left\| \mathbf{m}_{\text{mig.}}^{\text{pred.}} - \mathbf{m}_{\text{mig.}} \right\|_K^2 + \frac{\lambda^2}{2} \left\| \mathbf{C} \mathbf{m} \right\|_W^2 \quad (6)$$

is minimized rather than the *data* residual; in this case, $\mathbf{m}$ represents the velocity model. To find the velocity model, we can use the unregularized steepest descent equation:

$$\mathbf{m}^{(k+1)} = \mathbf{m}^{(k)} - \overbrace{\alpha \mathbf{L}^\dagger \mathbf{L} \delta \mathbf{m}_{\text{mig.}}^{(k)}}^{\text{migration-modeling of migration residual}}, \quad (7)$$

where $\delta \mathbf{m}_{\text{mig.}}^{(k)} = \mathbf{m}_{\text{mig.}}^{\text{pred.}} - \mathbf{m}_{\text{mig.}}$ is the migration residual at the k th iteration. The importance of this formula is that it shows



Seismic Imaging, Overview, Fig. 3 (a) Migration image $\mathbf{m}_{\text{mig.}}(x, z, s)$ cube in the model coordinates of x and z and the surface shot offset index s ; the common image gather $\mathbf{m}_{\text{mig.}}(x_{\text{CIG}}, z, s)$ is computed by migrating the shot gathers and displaying the migration image at the fixed value of

$x = x_{\text{CIG}}$ for all shot s and depth z values. (b) Common image gather $\mathbf{m}_{\text{mig.}}(x_{\text{CIG}}, z, s)$ for the common image point at x_{CIG} . The migration image should only be non-zero along the interface between the brick region and open region shown above

that the gradient term is computed by modeling the migration residual to get the *filtered* data residual $\mathbf{L}\delta\mathbf{m}_{\text{mig.}}^{(k)}$, and then the velocity model is updated by smearing this filtered (the filtered data residual is constructed by migrating and modeling the actual data residual; hence, it is a filtered version of the actual data residual) data residual into the model by the migration operation $\mathbf{L}^\dagger[\mathbf{L}\delta\mathbf{m}_{\text{mig.}}^{(k)}]$. These filtered residuals are smeared along wavepaths for each source-receiver pair of traces, and the migration and modeling operators are updated after each iteration. This leads to a moderate resolution of the velocity model because the filtered, not unfiltered, data are kinematically fitted. Moreover, MVA is largely a curve fitting exercise that is mostly insensitive to subtle amplitude variations in the migration traces; ignoring such information will blind MVA to subtle variations in the impedance distribution.

To reduce computational costs and increase robustness at the expense of reduced model resolution, MVA is sometimes implemented with the following steps:

1. Automatically pick the depth residual $\Delta z_i = z_i - z_i^{\text{ref.}}$ of a coherent CIG event in Fig. 3b at the i th shot position; here, z is the picked depth of the targeted reflection at the x_{CIG} offset in the i th migrated shot gather. The depth $z_i^{\text{ref.}}$ of the reference event for that reflector, is estimated from the near-offset trace in that CIG. A computer algorithm can window about the near-offset reflection of interest and use cross-correlation with neighboring traces in the CIG to estimate the depth lag $z_i - z_i^{\text{ref.}}$ associated with the strongest correlation energy.
2. The misfit function is then defined as

$$\varepsilon = 1/2 \sum_i (\Delta z_i)^2 + \text{model smoothness constraints}, \quad (8)$$

for each CIG and the summation is over the shot index in Fig. 3b. Sometimes the misfit function is the sum over all CIGs.

3. The velocity model is iteratively updated by a gradient optimization method until ε is minimized.

An example of the above procedure is shown in Fig. 4, where the left column of images depict the (a) migration image obtained with an inaccurate velocity model, (b) the CIG with curved events, and (c) the migration velocity model. After seven iterations of MVA, the right column of figures is obtained. Note the flat events in the CIG, and the final velocity model is almost the same as the actual velocity model.

The Fréchet derivative $\frac{\partial \varepsilon}{\partial c(\mathbf{x})}$ associated with the gradient of ε can be numerically computed by using a ray-tracing scheme (Stork 1992; Chauris et al. 1998) to determine the change in depth z_i of a migrated reflection point with respect to a change in the velocity parameter at $\mathbf{x}$. Sometimes, the depth residual Δz_i for the i th ray is converted into a time residual $\Delta \tau_i$, and, similar to traveltime tomography, this *converted* time residual is smeared along the reflection ray to iteratively update the velocity model. For arbitrary reflector geometries in a homogeneous media, an analytic conversion formula was derived

by Al-Yahya (1989), and a ray-tracing method was used by Stork (1992) and others (Robein 2010).

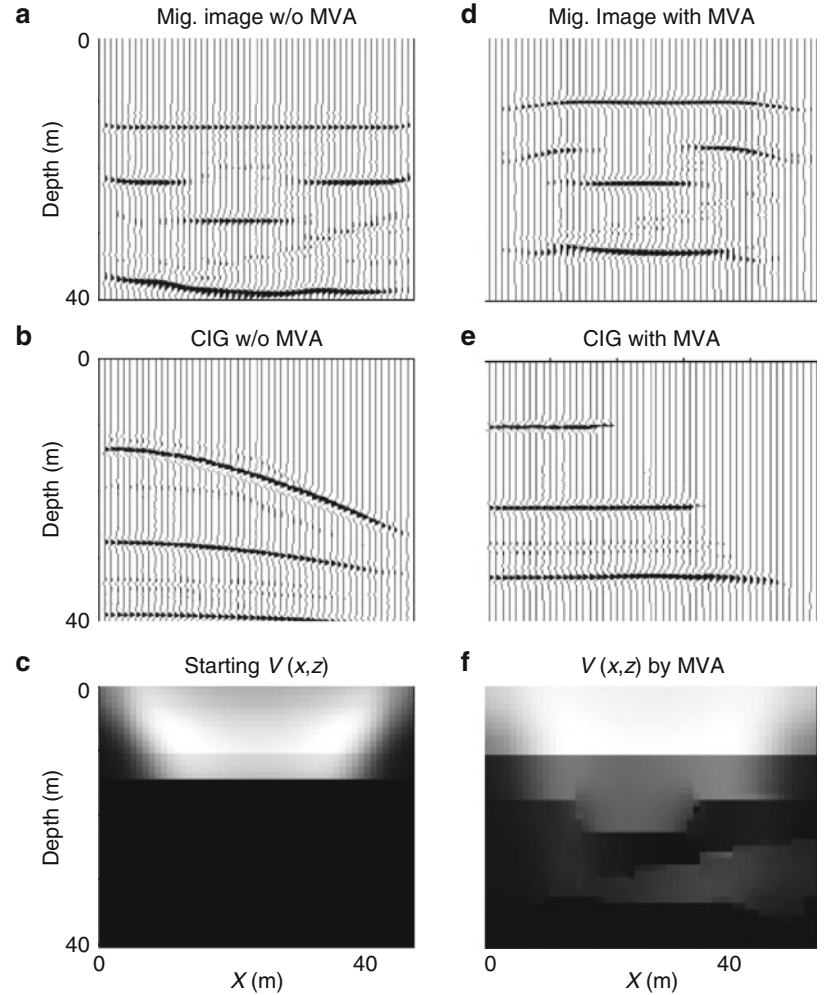
Similar to waveform inversion, MVA seeks to predict the wiggly migration traces seen in the migration image $\mathbf{L}^\dagger \mathbf{d}$ and, therefore, can easily get stuck in the many local minima of ε . To avoid this problem, Symes and Carazone (1991) proposed the smoother differential semblance misfit function (Chauris et al. 1998; Shen et al. 2003):

$$\varepsilon = 1/2 \left\| \sum_{x,y} \left[\gamma \sum_z \sum_h \partial m(x,y,z,h)_{\text{mig.}} / \partial h^2 \right] \right\|^2 + \text{constraints.} \quad (9)$$

that rewards flat events (i.e., accurate velocity models) in the CIG domain. Here, γ is a normalization term that depends on (x, y) . Empirical tests by Chauris et al. (1998) suggest that the semblance based ε is much smoother than that in Eq. 6 and is nearly devoid of local minima in misfit function. Some links between MVA and FWI are established in Symes (2008), and MVA compared to several tomography methods is presented

Seismic Imaging, Overview,

Fig. 4 (a) Migration image (Sun 2001) obtained by prestack migration using the smooth homogeneous velocity model in (c) that is far from the true model approximated in (f); (b) common image gather in the shot offset index, where the curved events indicate an incorrect velocity model; (c) incorrect migration velocity model used to compute (a); (d) migration image obtained after seven iterations of MVA; (e) CIG after seven iterations of MVA; and (f) velocity model inverted by seven iterations of MVA. This result closely resembles the true velocity model



by Le Bégat et al. (2004). The subsurface offset domain can be exploited for MVA (Sava and Biondi 2004) and extraction of scattering angle information (Rickett and Sava 2002), and an excellent summary of MVA research is given in Robein (2010).

In summary, the exploration community heavily favors MVA over many other velocity estimation methods because it is robust, efficient, and the picking of depth residuals is easily automated in the migration image domain. (Traveltimes picked from traces are notoriously difficult to automate because waveforms often interfere with one another. In comparison, migration untangles these events and focuses them to their origin along the reflectors so that automatic picking is easier in the CIG. Sometimes semblance methods are used to find the best fit hyperbolic or parabolic curve to the data [Robein 2010].) Its chief disadvantage is that an MVA image lacks the detailed resolution of waveform inversion, which suggests that MVA should be used to estimate the starting velocity models for waveform tomography.

Recent Advances in Seismic Imaging

In the last 15 years, several breakthroughs have enabled practical waveform inversion of seismic data. One of these advances is the relentless increase in computational capabilities of cluster computers and GPU-like processors, and two others are multiscale inversion and phase-encoded waveform inversion.

Multiscale Waveform Inversion

One of the main difficulties with waveform inversion is that the misfit function is plagued by many local minima. If the starting velocity model is moderately far from the actual model (an erroneous velocity might be one where the modeled events arrive by more than a period after the actual arrivals), then the iterative gradient solution gets stuck there and never reaches the global minimum or actual model.

The partial cure to this local-minima problem is a multiscale approach (Bunks et al. 1995), where the initial iterations estimate a coarse-grid velocity model from low-frequency data. For a reasonable starting model, this often prevents getting stuck in local minima because the predicted low-frequency arrivals are more likely to arrive within a period of the arrivals in the low-pass filtered data. After a number of iterations, the velocity model is refined to a finer scale (the grid interval is halved) and intermediate frequency data are iteratively inverted to update the velocity model. After suitable reduction of the data misfit, the model grid is refined again and higher-frequency data are inverted until a satisfactory model is reconstructed. One of the first relevant demonstrations of multiscale FWI applied to real data in a complex environment was performed by Ravaut et al. (2004). Other

results with both synthetic and field data (Sirgue and Pratt 2004; Sirgue et al. 2007; Plessix 2009; Vigh and Starr 2007, 2008; Sirgue et al. 2010) convincingly demonstrate the effectiveness of this approach. As mentioned earlier, improved model reconstructions can be achieved if more accurate starting models are employed, lower frequency data are recorded, and wider offset data are acquired. The new challenges are to employ modeling and inversion codes that robustly take into account the effects of viscoelasticity and anisotropy in the data.

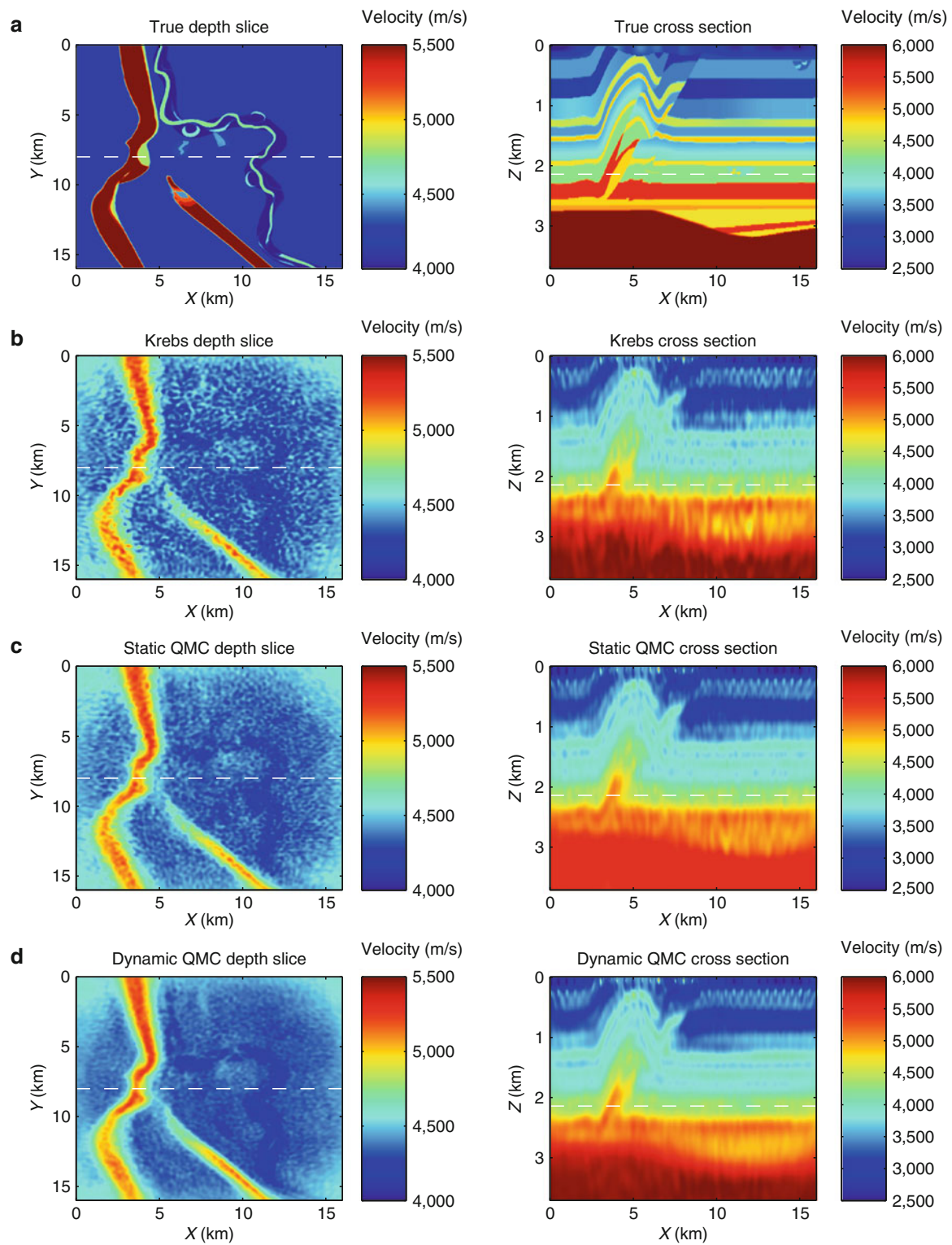
Phase-Encoded Multisource Waveform Inversion

A major difficulty, until recently, is the enormous computational expense of waveform inversion. Each shot gather of residuals must be migrated at each iteration, which for 3D problems can be too demanding even for the most powerful computers. To relieve this problem, Krebs et al. (2009), Operto et al. (2009), and Dai and Schuster (2009) proposed summing phase-encoded shot gathers into supergathers and migrating the supergathers at each iteration; a supergather is modeled with just one finite-difference simulation, where the computational cost is the same as that for one shot gather. This is similar to the phase-encoding RTM of Romero et al. (2000), except iterations are used to reduce the crosstalk noise in supergather migration. The result can be an enormous cost savings compared to conventional waveform inversion or migration.

Figure 1h shows an example of phase-encoding shot gathers, where each trace in a shot gather has the same random time shift but different shots have different time shifts. In this case, 12 shot gathers were time-shifted and blended together into one supergather; there were 192 traces per shot gather. A total of 16 nonoverlapping supergathers were input into the iterative waveform inversion code, where each supergather migration costs the same as the migration of one shot gather because 12 shots were excited at nearly the same time. Hence, the Fig. 1h tomogram costs 1/12 that of the Fig. 1g tomogram. For 3D inversion, the computational cost savings can be more than two orders of magnitude compared to conventional waveform inversion.

For a 3D example, Fig. 5a shows results after waveform inversion of 3D synthetic data, where the model size is $800 \times 800 \times 186$ grid points and the spatial sampling interval is 20 m. There are 1089 sources evenly distributed along the surface with an interval of 500 m in the inline (X) and crossline (Y) directions. Multisource waveform inversion using the Krebs method (Krebs et al. 2009), static Quasi-Monte Carlo (QMC) phase encoding, and dynamic QMC phase encoding (Boonyasiriwat and Schuster 2010) are applied to this data set.

Figure 5b–d show that dynamic QMC phase encoding provides a higher-quality tomogram than the other methods, yet the computational cost is two orders of magnitude less



Seismic Imaging, Overview, Fig. 5 Depth slices at $z = 2.1$ km and cross-sections at $y = 8$ km of (a) true model, and inverted models using (b) the Krebs method (1089 CSGs/supergather), (c) static QMC method (99 CSGs/supergather), and (d) dynamic QMC method (99 CSGs/

supergather) after 40 iterations. A greater number of CSGs/supergather require a greater number of iterations to get the same accuracy in the final tomogram; the dynamic encoding strategy is more effective than the static strategy. (Figures courtesy of C. Boonyasiriwat)

than that of conventional waveform inversion. The Krebs strategy is the most efficient because 1089 shot gathers were blended into one supergather, compared to the QMC strategies that used 99 CSGs/supergather. Dynamic phase encoding changed the time shifts of each shot gather after each iteration, while the static strategy kept the phase encoding the same.

Current Status and Future of Seismic Imaging

Reverse time migration is becoming the preferred means for seismic imaging beneath complex regions such as salt bodies. Earthquake seismologists are now recognizing the benefits of migration imaging earthquake records for tectonic structures such as subduction zones (Bostock et al. 2001) or using wide-angle seismic experiments to image the crust (Brenders and Pratt 2007) and mantle (Montelli et al. 2006). Least squares migration and waveform inversion are now being considered as viable upgrades to standard migration because of the tremendous speed up from phase-encoded multisource methodology. There is still a debate about whether waveform inversion should be computed in the frequency (there are additional advantages by formulating the problem in the Laplace transform domain [Shin and Ha 2008]) or time domains (Vigh and Starr 2007; Warner 2008), but there is no debate that we eventually need to account for viscoelastic and anisotropic effects in the data.

3D waveform inversion of earthquake records for whole earth tomograms greater than 1 Hz is still too computationally demanding except at very low frequencies, and the same can be said for 3D exploration geophysics at frequencies above 50 Hz. Challenges still remain, especially in the critically important area of anisotropic RTM (Zhang and Zhang 2009; Fowler et al. 2010) and waveform inversion; estimation of accurate anisotropic migration velocity models is an ongoing line of research. Earthquake seismologists are now testing the possibility of using earthquake records for inverting basin structures and velocity models so as to improve their simulation-based predictions of earthquake hazard. Passive seismic recordings and the use of interferometry to extract surface wave records (Shapiro and Campillo 2004; Shapiro et al. 2005), followed by inversion for S-velocity tomograms are playing an increasingly important role in earthquake seismology.

Cross-References

- [Body Waves](#)
- [Free Oscillations of the Earth](#)
- [Inverse Theory, Global Optimization](#)
- [Inverse Theory, Linear](#)
- [Inverse Theory, Monte Carlo Method](#)
- [Numerical Methods, Finite Difference](#)
- [Seismic Tomography](#)

- [Seismic, Ambient Noise Correlation](#)
- [Seismic, Migration](#)
- [Seismic, Waveform Modeling and Tomography](#)
- [Single and Multichannel Seismics](#)
- [Traveltime Tomography Using Controlled-Source Seismic Data](#)
- [Vertical Seismic Profiling](#)

Bibliography

- Al-Yahya K (1989) Velocity analysis by iterative profile migration. *Geophysics* 54:718–729
- Aoki N, Schuster GT (2009) Fast least-squares migration with a deblurring filter. *Geophysics* 74:WCA83–WCA93
- Askan A (2006) Full waveform inversion for seismic velocity and anelastic losses in heterogeneous structures. PhD dissertation, Carnegie Mellon University
- Barnes C, Charara M, Tsuchiya T (2008) Feasibility study for an anisotropic full waveform inversion of cross-well seismic data. *Geophys Prospect* 56:897–906
- Beydoun W, Mendes M (1989) Elastic ray-born L2- migration/inversion. *Geophys J Int* 97:151–160
- Beylkin G, Oristaglio M, Miller D, (1985) Spatial resolution of migration algorithms. In: Berkhout AJ, Ridder J, van der Walls LF (eds) *Proceeding of the 14th international symposium on acoustical imaging*, pp 155–167
- Bishop T, Bube KP, Cutler RT, Langan RT, Love PL, Resnick JR, Shuey RT, Spindler DA, Wyld HW (1985) Tomographic determination of velocity and depth in laterally varying media. *Geophysics* 50:903–923
- Bleistein N, Cohen JK, Stockwell JW (2001) *Mathematics of multi-dimensional seismic inversion*. Springer, New York
- Boonyasiriwat C, Schuster GT (2010) 3D multisource full-waveform inversion using dynamic QMC phase encoding. In: *Expanded Abstracts of SEG International Meeting*, p 29
- Boonyasiriwat C, Valasek P, Routh P, Cao W, Schuster GT, Macy B (2009) An efficient multiscale method for time-domain waveform tomography. *Geophysics* 74:WCC59–WCC68. <https://doi.org/10.1190/1.3151869>
- Bostock MG, Rondenay S, Shragge J (2001) Multiparameter two-dimensional inversion of scattered teleseismic body waves: theory for oblique incidence. *J Geophys Res* 106:771–782
- Brenders AJ, Pratt RG (2007) Efficient waveform tomography for lithospheric imaging: implications for realistic 2D acquisition geometries and low frequency data. *Geophys J Int* 168:152–170
- Brossier R, Operto S, Virieux J (2009) Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion. *Geophysics* 74:105–118
- Brossier R, Operto S, Virieux J (2010) Which data residual norm for robust elastic frequency-domain full waveform inversion? *Geophysics* 75:R37–R46
- Bunks C, Salek FM, Zaleski S, Chavent G (1995) Multiscale seismic waveform inversion. *Geophysics* 60:1457–1473
- Causse E, Mittet R, Ursin B (1999) Preconditioning for full-waveform inversion in viscoacoustic media. *Geophysics* 64:130–145
- Chauris H, Noble M, Podvin P (1998) Testing the behavior of differential semblance for velocity estimation. In: *Expanded abstracts of SEG international meeting*, pp i1305–1308
- Claerbout J (1992) *Earth soundings analysis: processing versus inversion*. Blackwell Scientific, Cambridge, MA
- Clinthorne N, Pan T, Chiao P, Rogers W, Stamos J (1993) Preconditioning methods for improved convergence rates in iterative reconstructions. *IEEE Trans Med Imaging* 12:78–83

- Cruse E, Pica A, Noble M, McDonald J, Tarantola A (1990) Robust elastic non-linear waveform inversion: application to real data. *Geophysics* 55:527–538
- Dahlen F, Hung S, Nolet G (2002) Frechet kernels for finite-frequency traveltimes I Theory. *Geophys J Int* 141:157–174
- Dai W, Schuster GT (2009) Least-squares migration of simultaneous sources data with a deblurring filter. In: Expanded abstracts of SEG international meeting, pp. 2990–2993
- Duquet B, Marfurt K, Dellinger J (2000) Kirchhoff modeling, inversion for reflectivity, and subsurface illumination. *Geophysics* 65:1195–1209
- Effelsen K (2009) A comparison of phase inversion and traveltime tomography for processing of near-surface refraction traveltimes. *Geophysics* 74:WCB11–WCB24
- Epanomeritakis I, Akcelik V, Ghattas O, Bielak J (2008) A Newton-CG method for large-scale three-dimensional elastic full waveform seismic inversion. *Inverse Problems* 24:975–987
- Erlangga Y, Hermann F (2009) Seismic waveform inversion with Gauss-Newton-Krylov method. In: Expanded abstracts of SEG international meeting, pp 2357–2361
- Etgen J, Gray S, Zhang Y (2009) An overview of depth imaging in exploration geophysics. *Geophysics* 74:WCA5–WCA17
- Faye J-P, Jeannot J-P (1986) Prestack migration velocities from focusing depth analysis. In: Expanded abstracts of SEG international meeting, pp 438–440
- Fowler P, Du X, Fletcher R (2010) Coupled equations for reverse time migration in transversely isotropic media. *Geophysics* 75: S11–S22
- Hill NR (2001) Prestack Gaussian beam depth migration. *Geophysics* 66:1240–1250
- Jin S, Madariaga R, Virieux J, Lambaré G (1992) Twodimensional asymptotic iterative elastic inversion. *Geophys J Int* 108:575–588
- Jousselin P, Duquet B, Audebert F, Sirgue J (2009) Bridging the gap between ray-based tomography and wave-equation migration image gathers. In: Expanded abstracts of SEG international meeting, pp 3979–3983
- Kamei R, Pratt G (2008) Waveform tomography strategies for imaging attenuation structure for cross-hole data. In: 70th conference and technical exhibition, EAGE Expanded Abstracts, p F019
- Kelly S, Ramos-Martinez J, Tsimelzon B, (2009) The effect of improved, low-frequency bandwidth in full-wave form inversion for velocity. In: Expanded abstracts of SEG international meeting, pp 3974–3977
- Krebs JR, Anderson JE, Hinkley D, Neelamani R, Lee S, Baumstein A, Lacasse MD (2009) Fast full-wavefield seismic inversion using encoded sources. *Geophysics* 74:WCC177–WCC188
- Lafond C, Levander A (1993) Migration moveout analysis and depth focusing. *Geophysics* 58:91–100
- Lailly P (1984) Migration methods: partial but efficient solutions to the seismic inverse problem. In: Santosa F, Pao YH, Symes W, Holland CH (eds) *Inverse problems of acoustic and elastic waves*. SIAM, Philadelphia
- Lambaré G, Virieux J, Madariaga R, Jin S (1992) Iterative asymptotic inversion in the acoustic approximation. *Geophysics* 57:1138–1154
- Langan R, Lerche I, Cutler RT (1985) Tracing of rays through heterogeneous media: an accurate and efficient procedure. *Geophysics* 50:1456–1465
- Le Bégat S, Chauris H, Devaux V, Nguyen S, Noble M (2004) Velocity model estimation for depth imaging: comparison of three tomography methods on a 2D real data set. *Geophys Prospect* 52:427–438
- Luo Y (1992) Calculation of wavepaths for band-limited seismic waves. In: Expanded abstracts of SEG international meeting, pp 1509–1512
- Luo Y, Schuster GT (1991) Wave-equation traveltime inversion. *Geophysics* 56:645–653
- Marquering H, Dahlen FA, Nolet G (2002) Threedimensional sensitivity kernels for finite-frequency traveltimes: the banana-doughnut paradox. *Geophys J Int* 137:805–815
- McMechan G (1983) Migration by extrapolation of time-dependent boundary values. *Geophys Prospect* 31:413–420
- Min D, Shin C (2006) Refraction tomography using a waveform-inversion back-propagation technique. *Geophysics* 71(3):R21–R30
- Montelli R, Nolet G, Masters G, Dahlen F, Hung SH (2004) Global P and PP traveltime tomography: rays versus waves. *Geophys J Int* 158:637–654
- Montelli R, Nolet G, Dahlen F (2006) Comment on “Banana-doughnut kernels and mantle tomography” by van der Hilst and de Hoop. *Geophys J Int* 167:1204–1210
- Mora P (1987) Elastic wavefield inversion. PhD dissertation, Stanford University, pp 143
- Mora P (1989) Inversion=migration+tomography. *Geophysics* 54:1575–1586
- Nemeth T, Wu C, Schuster GT (1999) Least-squares migration of incomplete reflection data. *Geophysics* 64:208–221
- Nolet G (1987) Seismic tomography: with applications in global seismology and exploration. Springer, Dordrecht
- Nowack RLWP, Chen UK, Dasgupta S (2007) Imaging offsets in the Moho: synthetic tests using Gaussian beams with teleseismic waves. *Pure Appl Geophys* 164:1921–1936
- Operto S, Virieux J, Ribodetti A, Anderson J (2009) Finite-difference frequency-domain modeling of viscoelastic wave propagation in two-dimensional tilted transversely isotropic media. *Geophysics* 74:T75–T95
- Plessix RE (2006) A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophys J Int* 167:495–503
- Plessix RE (2009) 3D frequency-domain full-waveform inversion with an iterative solver: geophysics. *Geophysics* 74:WCC149–WCC157
- Pratt G (1999) Seismic waveform inversion in the frequency domain, part I: theory and verification in a physical scale model. *Geophysics* 64:888–901
- Pratt RG, Goulty NR (1991) Combining wave-equation imaging with traveltime tomography to form high-resolution images from crosshole data. *Geophysics* 56:208–224
- Pratt RG, Shin C, Hicks GJ (1998) Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion. *Geophys J Int* 133:341–362
- Pratt RG, Hou F, Bauer K, Weber M (2005) Waveform tomography images of velocity and inelastic attenuation from the Mallik 2002 crosshole seismic surveys. In: Dallimore SR, Collett TS (eds) *Scientific results from the Mallik 2002 gas hydrate production research well program, Mackenzie Delta, Northwest Territories, Canada*. Geological Survey of Canada, Canada
- Ravaut C, Operto S, Imbrota L, Virieux J, Herrero A, dell’Aversana P (2004) Multi-scale imaging of complex structures from multi-fold wide-aperture seismic data by frequency-domain full-wavefield inversions: application to a thrust belt. *Geophys J Int* 159:1032–1056
- Rickett J, Sava P (2002) Offset and angle-domain common-image gathers for shot-profile migration. *Geophysics* 67:883–889
- Robein E (2010) Seismic Imaging. EAGE Publications, The Netherlands
- Romero L, Ghiglia D, Ober C, Morton S (2000) Phase encoding of shot records in prestack migration. *Geophysics* 65:426–436
- Ryberg T, Weber M (2000) Receiver function arrays: a reflection seismic approach. *Geophys J Int* 141:1–11
- Sava P, Biondi B (2004) Wave-equation migration velocity analysis-1: theory. *Geophys Prospect* 52:593–606
- Shapiro N, Campillo M (2004) Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise. *Geophys Res Lett* 31: L07614. <https://doi.org/10.1029/2004GL019491>.
- Shapiro N, Campillo M, Stehly L, Ritzwoller M (2005) High-resolution surface-wave tomography from ambient seismic noise. *Science* 307:1615–1618
- Sheehan AF, Shearer PM, Gilbert HJ, Dueker KG (2000) Seismic migration processing of P-SV converted phases for mantle discontinuity

- structure beneath the Snake River plain, Western United States. *J Geophys Res* 105:055–065
- Shen P, Symes W, Stolk C (2003) Differential semblance velocity analysis by wave equation migration. In: Expanded abstracts of SEG international meeting, pp 2135–2139
- Sheng J, Leeds A, Buddensiek M, Schuster GT (2006) Early arrival waveform tomography on near-surface refraction data. *Geophysics* 71(4):U47–U57
- Shin C, Ha W (2008) A comparison between the behavior of objective functions for waveform inversion in the frequency and Laplace domains. *Geophysics* 73:VE119–VE133
- Shin C, Min D-J, Marfurt KJ, Lim HY, Yang D, Cha Y, Ko S, Yoon K, Ha T, Hong S (2002) Traveltime and amplitude calculations using the damped wave solution. *Geophysics* 67:1637–1647
- Shin C, Pyun S, Bednar B (2007) Comparison of waveform inversion, part 1: conventional wavefield vs logarithmic wavefield. *Geophys Prospect* 55:449–464
- Sirgue L, Pratt RG (2004) Efficient waveform inversion and imaging: a strategy for selecting temporal frequencies. *Geophysics* 69:231–248
- Sirgue L, Etgen J, Albertin U (2007) 3D full-waveform inversion: wide-versus narrow-azimuth acquisitions. In: Expanded abstracts of SEG international meeting, pp 1760–1764
- Sirgue L, Barkved OI, Dellinger J, Etgen J, Albertin U, Kommedal JH (2010) Full waveform inversion: the next leap forward in imaging at Valhall. *First Break* 28:65–70
- Song Z, Williamson P, Pratt G (1995) Frequency-domain acoustic-wave modeling and inversion of crosshole data, part 2: inversion method, synthetic experiments and real-data results. *Geophysics* 60:786–809
- Stolt R, Benson A (1986) Seismic migration: theory and practice. In: *Handbook of geophysical exploration*, vol 5. Geophysical Press, London, UK
- Stork C (1992) Reflection tomography in the postmigrated domain. *Geophysics* 57:680–682
- Sun H (2001) Wavepath migration for depth imaging and velocity analysis. PhD dissertation, University of Utah
- Symes W (2008) Migration velocity analysis and waveform inversion. *Geophys Prospect* 56:765–790
- Symes W, Carazone J (1991) Velocity inversion by differential semblance optimization. *Geophysics* 56:654–663
- Tarantola A (1986) Linearized inversion of seismic reflection data. *Geophys Prospect* 32:998–1015
- Tarantola A (1987) Inverse problem theory: methods for data fitting and model parameter estimation. Elsevier Science, Amsterdam
- Thierry P, Operto S, Lambaré G (1999) Fast 2D ray-born inversion/migration in complex media. *Geophysics* 64:162–181
- Toldi J (1989) Velocity analysis without picking. *Geophysics* 54:191–199
- van der Hilst R, de Hoop M (2006) Reply to comment by R. Montelli, G. Nolet, and F.A. Dahlen on “Banana-doughnut kernels and mantle tomography”. *Geophys J Int* 167:1211–1214
- Vigh D, Starr EW (2007) Comparisons for waveform inversion, time domain or frequency domain? In: Expanded abstracts of SEG international meeting, pp 1890–1894
- Vigh D, Starr EW (2008) 3D prestack plane-wave full-waveform inversion. *Geophysics* 73:135–144
- Vigh D, Starr EW, Elapavuluri P (2009) Acoustic waveform inversion vs. elastic data. In: Expanded abstracts of SEG international meeting, pp 2298–2301
- Virieux J, Operto S (2009) An overview of full-waveform inversion in exploration geophysics. *Geophysics* 74:WCC1–WCC26
- Warner M (2008) 3D wavefield tomography: synthetic and field data examples. In: Expanded abstracts of SEG international meeting, pp 3330–3334
- Whitmore ND (1983) Iterative depth migration by backward time propagation. In: Expanded abstracts of SEG international meeting, pp 827–830
- Williamson P (1991) A guide to the limits of resolution imposed by scattering in ray tomography. *Geophysics* 56:202–207
- Woodward MJ (1992) Wave-equation tomography. *Geophysics* 57:15–26
- Xu S, Chauris H, Lambar G, Noble M (1998) Common angle image gather: a strategy for imaging complex media. In: Expanded abstracts of SEG international meeting, pp 1538–1541
- Xu W, Xie X (2009) How serious is the nonlinear effect on traveltime delays predicted by sensitivity kernels. In: Expanded abstracts of SEG international meeting, pp 4049–4053
- Yilmaz O, Chambers R (1984) Migration velocity analysis by wave-field extrapolation. *Geophysics* 49:1664–1674
- Yu J, Hu J, Schuster G, Estill R (2006) Prestack migration deconvolution. *Geophysics* 71:S53. <https://doi.org/10.1190/1.2187783>
- Zelt C (2002) Modelling strategies and model assessment for wide-angle seismic traveltime data. *Geophys J Int* 139:183–204
- Zhan G, Schuster GT (in press) Multisource phase-encoded waveform inversion. *J Explor Geophys*
- Zhang Y, Zhang H (2009) A stable TTI reverse time migration and its implementation. In: Expanded abstracts of SEG international meeting, pp 2794–2798
- Zhou C, Schuster GT, Hassanzadeh S, Harris JM (1997) Elastic wave-equation traveltime and waveform inversion of crosshole seismic data. *Geophysics* 62:853–868

Seismic Instrumentation

Duncan Carr Agnew

Institute of Geophysics and Planetary Physics, Scripps
Institution of Oceanography, University of California San
Diego, La Jolla, CA, USA

Synonyms

Seismometry

Definition

Datalogger. Device for recording electrical signal from a seismometer, usually in digital form and with accurate time information.

Geophone. Another name for seismometer, used in geophysical exploration.

Seismometer. Device for providing a record of ground motion, usually by converting it into an electrical signal.

Strainmeter. Device for continuously measuring the deformation of the Earth, either as extension along a line, or volume change.

Introduction

Seismic instruments give a time record of ground motion caused by elastic waves. The first instruments were built at

the end of the nineteenth century; subsequent developments have evolved towards wider coverage of both the frequency and amplitude of the waves that could be recorded. Over most of the twentieth century much of this evolution, and of the diversity of instrument types, was related not to how ground motion was sensed but to how it was recorded; the last three decades have seen such rapid development in digital technology, especially in data storage, that recording techniques unique to seismology are no longer needed. Modern seismic systems all consist of a seismometer for converting some aspect of ground motion into an electrical signal, and a datalogger for recording this signal. In exploration geophysics, the name geophone is a common synonym for seismometer; now that data recording is clearly separated from sensing ground motion, the old term seismograph could probably be abandoned. Since dataloggers are not unique to seismology, this article discusses only the general requirements for them, focusing instead on the designs of seismic sensors, or seismometry. Three recent reviews cover seismometry in more detail: Wielandt (2002), Havskov and Alguacil (2004), and Bormann (2009).

Requirements for Instruments

The frequencies to be measured by seismometers range from 0.31 mHz (the slowest free oscillation) up to 1 kHz (in near-surface geophysics): roughly six orders of magnitude. Fig. 1 shows the amplitudes that can occur over most of this frequency range, from the background noise level at quiet sites (Berger et al. 2004) to the large signals that have been observed at moderate distances from very large earthquakes, to the even larger, and damaging, strong ground motions close to an earthquake. The large signals are infrequent, but also of considerable scientific importance – but so are very small signals, which can be seen only if the seismometer has lower noise than the ground does.

Since digital dataloggers can record over a much wider range of amplitudes and frequencies than the older analog recorders could, modern seismic systems can be characterized as broadband (covering a wide band of frequencies) and high dynamic range (a wide range of amplitudes). But the range from noise to the largest signal covers over 10 orders of magnitude, too much for any single instrument and datalogger. Since most seismometer designs require practical tradeoffs between low-frequency capability and other performance goals, seismometers are still classified by what frequencies and amplitudes they cover best. Systems for recording the largest signals without distortion are called *strong-motion* seismometers, and have usually been designed to meet the needs of earthquake engineers rather than seismologists, though digital technology is blurring this distinction.

Inertial Seismometers: Basic Principles

Almost all seismometers measure motion of the ground using the inertia of an internal mass, and so are called *inertial* sensors. This mass, of amount m , is subject to five forces: (1) constraints that restrict the mass motion to be in a particular direction (denoted by a unit vector $\mathbf{e}$) making the system one with a single degree of freedom; (2) the gravitational force vector, $\mathbf{g}m$, along that direction, namely, $\mathbf{e} \cdot \mathbf{g}m$; (3) an elastic restoring force (from a spring), $-k(x - x_0)$, where k is the spring constant and x the displacement (in the direction $\mathbf{e}$) away from an equilibrium position x_0 ; (4) a viscous damping force proportional to velocity $-d\dot{x}$; (5) additional forces F_b that may be applied by a feedback system (discussed below).

The displacement of the mass relative to an inertial frame is $x + \mathbf{e} \cdot \mathbf{u}$, where x is the motion of the mass relative to the Earth, which is what we can measure, and $\mathbf{u}$ is the (vector) motion of the Earth relative to the inertial frame, which is that of the Earth in the absence of seismic waves (we can neglect non-inertial effects from the Earth's rotation). Combining forces (2), (3), and (4), and applying the momentum equation, the acceleration of the mass is given by

$$m(\ddot{x} + \mathbf{e} \cdot \ddot{\mathbf{u}}) = \mathbf{e} \cdot \mathbf{g}m - k(x - x_0) - d\dot{x}$$

where we allow the possibility of variations in both $\mathbf{e}$ and $\mathbf{g}$. With no motion, we assume that the mass position is x_0 , so that $\mathbf{e}_0 \cdot \mathbf{g}_0 m = -kx_0$. Then the above equation becomes, for small variations in $\mathbf{e}$ and $\mathbf{g}$, and after some rearrangement of terms,

$$\frac{k}{m}x + \frac{d}{m}\dot{x} + \ddot{x} = -\mathbf{e}_0 \cdot \ddot{\mathbf{u}} - \mathbf{e}_0 \cdot (\mathbf{g} - \mathbf{g}_0) - (\mathbf{e} - \mathbf{e}_0) \cdot \mathbf{g}_0$$

more usually written as

$$\ddot{x} + 2\gamma\dot{x} + \omega_0^2 x = -\mathbf{e}_0 \cdot \ddot{\mathbf{u}} - \mathbf{e}_0 \cdot (\mathbf{g} - \mathbf{g}_0) - (\mathbf{e} - \mathbf{e}_0) \cdot \mathbf{g}_0 \quad (1)$$

where $\omega_0 = \sqrt{k/M}$ is the natural frequency of the seismometer ($T_0 = 2\pi/\omega_0$ is called the free period); $\gamma = d/2M$ is the damping constant. These names describe the behavior of the seismometer if the right-hand side of the equation is zero and $x = x_0$ at $t = 0$; the subsequent motion of the seismometer mass is then

$$x(t) = x_0 \cos\left(\frac{2\pi t}{T_0}\right) e^{-\gamma t},$$

which is a decaying sinusoid with period T ; the damping γ needs to be large enough to avoid ringing (a sign of resonance effects), and is usually set to about 0.8.

The right-hand side of Eq. 1 shows that an inertial sensor responds to three inputs:

1. The acceleration of the ground, $\ddot{\mathbf{u}}$; this is why inertial sensors are often called accelerometers.
2. Changes in the gravitational vector $\mathbf{g}$ along the direction of sensitivity; if this is the dominant signal, the sensor is usually referred to as a gravimeter (see *Gravimeters*).
3. Changes in the direction of sensitivity relative to the direction of gravity. If $\mathbf{e}_0$ and $\mathbf{g}_0$ are parallel (the mass moves vertically), then to first order $\mathbf{e} - \mathbf{e}_0$ is perpendicular to $\mathbf{g}_0$, and this contribution is zero. However, if $\mathbf{e}_0$ and $\mathbf{g}_0$ are perpendicular (the mass moves horizontally), then to first order $\mathbf{e} - \mathbf{e}_0$ is parallel to $\mathbf{g}_0$, and this contribution can be significant; if this is the dominant signal, the sensor would be referred to as a tiltmeter, since changes in $\mathbf{e}$ reflect tilting of the sensor (or the ground it is attached to).

Much confusion has been created by the use of these different terms for what is the same kind of sensor, when the difference actually refers to the type of signal being measured.

Ignoring the tilt and gravity terms, suppose that the ground displacement is purely sinusoidal, $u(t) = Ue^{i\omega t}$ (U being complex, and the real part being taken). Then the mass motion will be $Xe^{i\omega t}$, where

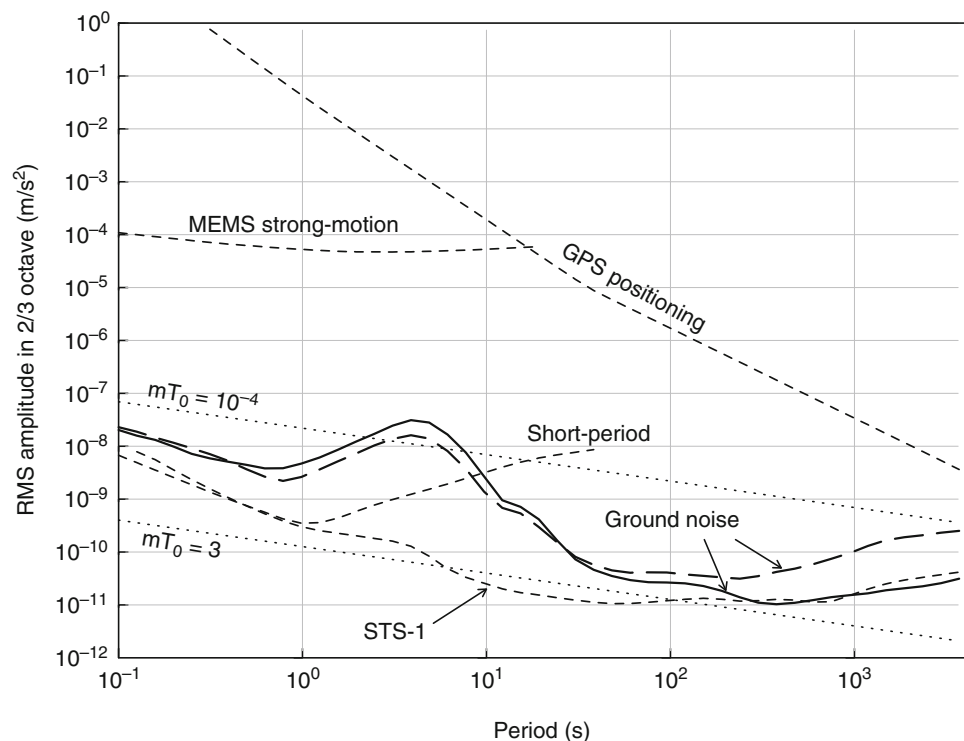
$$\frac{X}{U} = \frac{\omega^2}{\omega_0^2 - \gamma i\omega + \omega^2} \quad (2)$$

is the frequency response of the seismometer. For ground motion at frequencies much higher than the natural frequency ω_0 , $X \approx U$: the mass motion looks like ground displacement. For ground motion at much lower frequencies, $X \approx U\omega^2/\omega_0^2$: the mass motion looks like ground acceleration, scaled by the inverse of the natural frequency of the seismometer.

The minimum noise level of any inertial sensor is set by the thermal (Brownian) motion of the mass in equilibrium with its surroundings. This noise level can be expressed as an equivalent ground acceleration, with a flat power spectral density of $8\omega_0\lambda k_B\theta/m$, where k_B is Boltzmann's constant and θ is the temperature (Aki and Richards 2002). For a temperature of 310 K and $\gamma = 0.8$, this expression is $1.72 \times 10^{-19}/(mT_0) \text{ m}^4 \text{ s}^{-3}$, with m in kilograms and T_0 in seconds. The dotted lines in Fig. 1 show noise levels for two cases that cover the range of designs: $mT_0 = 10^{-4}$ (1 g mass and free period of 0.1 s), which cannot resolve ground noise at the longer periods, and $mT_0 = 3$ (300 g and 10 s), which shows that a moderately large mass and long period are necessary and sufficient for this thermal noise limit to fall below ground noise at all frequencies.

Seismic Instrumentation,

Fig. 1 Seismic and instrument noise levels. The solid and long dashed lines show the lowest levels of vertical and horizontal ground motion observed by the stations of the Global Seismographic Network. The dotted lines show the seismometer Brownian noise limit for two different values of the product of the sensor mass and free period. The dashed lines show the noise levels for a high-quality short-period system using a moving-coil transducer (Rodgers 1994), for a MEMS sensor used for measuring strong motion (Evans et al. 2005), for a sensitive broadband system (Ringler and Hutt 2009), and for GPS used at seismic frequencies



Mechanical Design

In an inertial sensor, the mass needs to be constrained to move in one direction only, with a restoring force that is both exactly linear in the mass displacement and of whatever amount needed for the sensor to have the desired free period. The free period enters into the design partly because of Eq. 2: the longer the period, the greater the displacement of the mass for a given acceleration at periods longer than the free period; that is to say, the higher the long-period sensitivity, and the lower the noise. Sensors designed to measure only high frequencies, such as the geophones used in geophysical exploration, can have a short free period, which means that the springs used can be relatively stiff (large k) and rugged. Seismometers for high frequencies use elastic elements both to constrain mass motion to a straight line and to provide a linear restoring force (Fig. 2a). The latest innovation for short-period instruments is the MEMS (Micro-Electronic Mechanical Systems) sensor, in which the mass-spring system is machined from monolithic silicon and packaged with integral electronics; these are mass-produced for non-seismological applications and so are relatively inexpensive. While not as

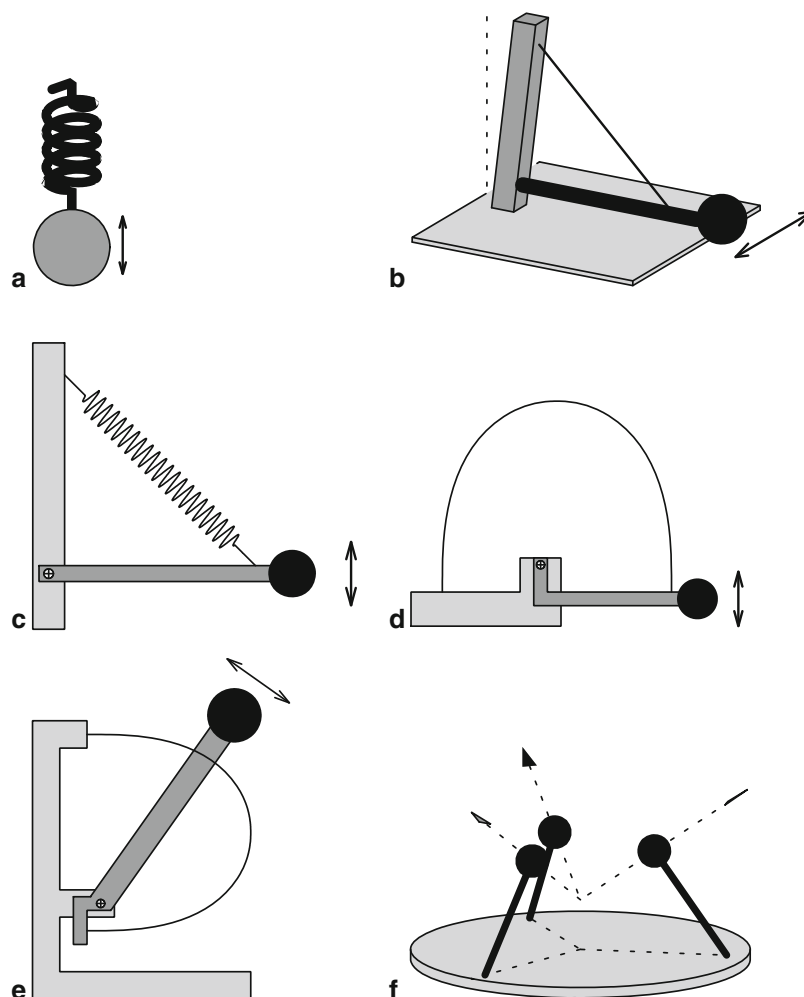
sensitive as traditional seismometers, they are already useful for measuring strong ground motions.

The high sensitivity needed for scientific purposes makes the mechanical design more challenging. If the mass moves vertically, it must be stably suspended against gravity while the restoring force is also kept low. In most long-period seismometers, the mass is constrained by hinges and stops to move along a small part of the circumference of a circle, which for most applications is an adequate approximation to a straight line (departures from this cause cross-coupling (Rodgers 1968; LaCoste 1967)). For sensing horizontal motion, the circle is slightly tilted from a horizontal plane, in which case the hinged mass is a horizontal pendulum (also called a garden-gate suspension, shown in Fig. 2b). The restoring force then comes from gravity and is easily adjusted by tilting the instrument.

Long-period sensors that measure in other directions use a spring arranged to give a weak, but still linear, restoring force. One common design is the LaCoste suspension (Fig. 2c), which uses a *zero-length spring*: a helical spring that exerts a force exactly proportional to its physical length. The

Seismic Instrumentation,

Fig. 2 Mechanical designs for seismometers, shown in cartoon form. (a) is a simple mass on a spring. (b) is the “garden-gate” design used for long-period horizontal instruments; note that the support post is not vertical. (c) shows the geometry of a LaCoste suspension; the restoring force of the spring must be equal to its physical length. (d) shows a vertical sensor that uses a leaf spring (elastica) to supply the restoring force. (e) and (f) show the design of a single sensor in a triaxial seismometer, and how three such sensors are arranged to record orthogonal motions



geometry of the spring and hinged mass is chosen to make the restoring force nearly independent of mass position, giving a very long period. An alternative method uses a single sheet of elastic material, bent to form a leaf spring (Fig. 2d), to create a long-period system with adequate linearity over the actual range of motion.

The traditional arrangement of seismometers is an orthogonal pair for horizontal motion, and a single sensor of vertical motion; if only one sensor is used, it is usually a vertical to give the best signal-to-noise ratio for P waves. Many systems now use a “triaxial” arrangement, with three identical systems measuring motion at an angle of 54.73° to the vertical, and at azimuths 120° apart; this also gives three orthogonal directions, namely those of the edges of a cube balanced on a corner (Fig. 2e, f).

Any seismometer requires careful design to ensure that the only vibration that takes place is the mass motion, or at least that other modes of vibration (called *parasitic resonances*) will not produce any signals if excited by ground motion. Long-period seismometers need special materials that do not creep much under load and whose dimensions and elastic constants are insensitive to temperature changes; even then, isolation and active environmental control may be needed to reduce the effects of changes in air temperature, air pressure (which causes varying buoyancy of the mass), and magnetic fields (which affect spring materials).

Electronic Design

Sensing the mass motion in a seismometer is now almost always done electronically, in one of two ways (Agnew 1986). In the first, electrodynamic method, permanent magnets and coils of wire are arranged so that the motion of the sensor produces an induced voltage $V = G\dot{x}$, where G is the generator constant; this voltage is then amplified and recorded. This method is simple and rugged, and provides damping if some of the induced current flows through a resistor. However, because the voltage produced depends on the velocity of the mass, and because the noise in the amplifiers rises with low frequency, electrodynamic sensing cannot resolve ground noise at periods longer than about 10 s.

The second method is active sensing of the mass motion. An oscillating voltage is input to a variable capacitor or inductor attached to the mass, which produces an output voltage with the same frequency and amplitude proportional to x ; this output voltage is then demodulated to produce $V = Sx$, where S is the sensitivity. Up to the point of demodulation, the signal is at a frequency (several kHz) at which electronics noise is relatively low, so such displacement sensors can be made sensitive enough to measure motions less than an atomic diameter.

This output voltage can be recorded directly, but more often is used in a feedback system: the integral of this voltage (over some frequency band) is used to apply an additional

force to the mass, usually with a coil and magnet, acting against the mass displacement. Because this force acts against the apparent force of ground acceleration, this is called a force-balance accelerometer. Feedback does not decrease noise, but does have three other merits. First, it can be used to vary the seismometer response much more than mechanical modifications can; for example, a mechanical system with a free period longer than 10 s is difficult to make, but this is easy to create electronically; and the response of the electronic system will be much more stable over time. Second, because the feedback force reduces the mass motion, the mechanical system needs to be linear over a smaller range. Finally, the calibration depends on the relation between voltage and feedback force, and this is very stable over time.

Calibration

Equation 2 is an example of the transfer function of a seismometer: the relation, as a function of frequency, between the ground motion and the output. In that example, the output was assumed to be mass motion x ; a more realistic example would include a displacement sensor that generates a voltage $V = Sx$ followed by a datalogger that digitizes the voltage and produces a number $N = AV$. The total transfer function of the system, from input Earth displacement to digital output, would then be

$$\frac{AS\omega^2}{\omega_0^2 - \gamma i\omega + \omega^2}. \quad (3)$$

Seismometer calibration consists of determining all the parameters in expressions like this one. Some components, such as the datalogger, are calibrated by putting a known signal in and recording the output. For the seismic sensor, this requires producing a known ground motion, which can be done with a shake table before the seismometer is installed. After installation, a seismometer calibration can be checked by putting a signal into the seismometer electronics that will apply force to the mass without otherwise disturbing the output.

Calibration information is usually provided in a form like Eq. 3: a leading constant, and the (complex) polynomials that give the response as a function of a frequency. Often the polynomials are specified by giving their roots in the complex plane, referred to as poles and zeros. The frequency response can usually be found quite accurately, though the absolute response to ground motion is usually difficult to estimate to better than 1%. Sometimes mundane aspects of the calibration, such as the actual direction of measurement, can be in error by significant amounts.

Installation

An ideal seismometer installation should maximize signals of interest whether distant earthquakes or the shaking of a building, while minimizing noise that might hide these signals.

Any actual installation will be a compromise between this ideal and such practical matters as access and cost. For earthquake signals, putting an instrument deeper below the Earth's surface will give better results, but if this requires a deep drill-hole, a shallower installation may be all that can be done. An extreme example of low-quality installations being accepted for reasons of cost is ocean-bottom seismometers, for which the standard installation is the seismometer sitting on the bottom, sometimes on very soft sediments, however it happens to land after being dropped off a ship.

On land, the shallowest installations are for temporary instruments used in geophysical exploration, with seismometers planted on the surface with an attached spike to hold them in place. More permanent installations range from instruments set on surface rock exposures or in shallow holes in soil, to purpose-built deep vaults, caves (if available), and drill-holes. Deeper installations reduce the noise from wind and the range of temperature variation, and even a meter of soil will attenuate daily fluctuations. It is advisable, though not always possible, for instruments to be installed away from noise sources such as roads, machinery, rivers, and trees; in an urban setting borehole installations at depths of a few hundred meters may be the only way to reduce the noise to reasonable levels.

Displacement Seismometry Using Satellites

Because the mass motion is proportional to acceleration for motions with periods longer than the seismometer free period, data from inertial seismometers has to be integrated twice to find long-period displacements. This integration can introduce significant error, especially if the response is slightly nonlinear or the sensor tilts.

Direct sensing of displacement can now be done using repeated distance measurements between a ground sensor and Earth satellites, since the satellites define an inertial reference frame unaffected by ground motion. Distances between satellites of the Global Positioning System (GPS) and a receiver on the ground can be measured to within less than a millimeter precision, even over short times and in the presence of large accelerations; the actual ground displacement can be found after some processing (Bilich et al. 2008). The accuracy of GPS distance measurements is lower than their precision because of propagation effects, including interference from signals reflected from nearby objects (known as multipath), but it has proved possible to determine horizontal ground displacements to within a few millimeters over time spans up to several minutes and at rates up to 10 Hz. This sensitivity is too low to measure most earthquake signals, but can provide high-quality displacement data for ground motions near large earthquakes; a combination of inertial and GPS sensors is the optimal way to measure strong motion.

Deformation Seismometers

A final class of seismometers measures, not the displacement at a point, but quantities related to the displacement gradient. For a vector displacement $\mathbf{u}$, the gradient $\nabla\mathbf{u}$ is a tensor, which can be decomposed into two parts: a symmetric part, the strain tensor, which describes local deformation; and an antisymmetric part, which is equivalent to a vector that describes local rotation, tilting about horizontal axes and rotating about the vertical axis.

Two types of instruments, both known as strainmeters, can measure components of the strain tensor (Agnew 1986). Extensometers measure relative displacement along a line between two end points; this displacement divided by the baseline (the distance between the points) gives the extensional or compressional strain. Extensometers with baselines as short as a few centimeters are installed in boreholes, and ones with baselines up to tens of meters in tunnels. In both of these, the measurement is made relative to a solid length standard. Much longer instruments use optical methods to measure strain over hundreds of meters. The other class of strainmeter is the volumetric type, in which the change in a volume of fluid is found by sensing the displacement of fluid in and out of a container cemented in a borehole: such an instrument measures the volume strain, or dilatation.

Measuring the rotation vector requires either a stable direction of reference or some way of measuring rotation directly with respect to inertial space. Until recently, no technology existed that could do this with the low noise levels required. A few instruments measure rates of rotation around a vertical axis using Sagnac interferometers, also called laser gyroscopes (Schreiber et al. 2009). These sense the difference in frequency between two light beams propagating in opposite directions around a closed loop.

Seismic Data Recording

Three specialized requirements for seismological dataloggers are: (1) a high dynamic range, to capture all the signals possible; (2) large amounts of storage, since it may be necessary to record for a long time to capture unpredictable events; and (3) accurate absolute timing, to relate seismic-wave travel times to a common system. In all three areas, the progress of electronics has meant rapidly improving performance at ever-lower cost.

Many seismic dataloggers use specialized systems to provide 24 bits of range (about 5×10^7); this is accomplished by oversampling followed by digital filtering. The amount of storage depends on the sample rate; except in limited situations, a rate of about 200 Hz will capture all seismic data without aliasing. A day of 3-component data with this resolution and sample rate comes to just over 50 Mb, an amount that in the last two decades has gone from requiring specialized and

bulky storage to something easily dealt with. Cross-correlation methods can find time delays to a precision of 0.1 of the sample interval, and so the time of each sample point should be known well: for 200 Hz sampling, this requirement would be 500 μ s. This too is a level of accuracy that until recently was not easily attained. It can be reached without much difficulty if radio signals from the GPS satellites are available, though providing it over long times in the absence of such a signal still requires expensive equipment. Lower levels of accuracy can be obtained from other radio signals, and, over intervals of a few days, from inexpensive crystal oscillators.

Often, a single system records data telemetered from multiple sensors, forming an array (if the region covered spans only a few wavelengths of the waves being recorded) or network (if larger). The most precise timing then requires corrections for the transmission time of the data (latency) unless a separate datalogger is used for each sensor.

Summary

With over a century of development, inertial seismic sensors are a mature technology, usually capable of recording ground motion much better than it can be modeled. While no single sensor can cover the full range of amplitudes and periods of seismic waves, only a few instruments are needed to provide a faithful record of ground motion. Since other developments in electronics have largely routinized digital recording of seismic data, most users treat seismometers as a “black box” system that can be acquired and used with little specific expertise: a usually justifiable assumption, though as always it is important to know enough to recognize poor performance when it does occur.

Cross-References

- [Earthquakes, Location Techniques](#)
- [Earthquakes, Strong-Ground Motion](#)
- [Free Oscillations of the Earth](#)
- [Gravimeters](#)
- [Seismic Noise](#)
- [Seismogram Interpretation](#)
- [Seismology, Rotational](#)

Bibliography

- Agnew DC (1986) Strainmeters and tiltmeters. *Rev Geophys* 24:579–624
- Aki K, Richards PG (2002) *Quantitative seismology*. University Science Books, Sausalito
- Berger J, Davis P, Ekström G (2004) Ambient earth noise: a survey of the global seismographic network. *J Geophys Res* 109:B11307. <https://doi.org/10.1029/2004JB003408>

- Bilich A, Cassidy JF, Larson KM (2008) GPS seismology: application to the 2002 M_w 7.9 Denali fault earthquake. *Bull Seismol Soc Am* 98:593–606
- Bormann P (ed) (2009) *New manual of seismological observatory practice*. <http://www.iaspei.org/projects/NMSOP.html>
- Evans JR, Hamstra RH, Kundig C, Camina P, Rogers JA (2005) TREMOR: a wireless MEMS accelerograph for dense arrays. *Earthq Spectra* 21:91–124
- Havskov J, Alguacil G (2004) *Instrumentation in earthquake seismology*. Springer, Dordrecht
- LaCoste LJB (1967) Measurement of gravity at sea and in the air. *Rev Geophys* 5:477–526
- Ringler AT, Hutt CR (2009) Self-noise models of seismic instruments. *EOS Trans Am Geophys Union Fall Meet Suppl* 90:S23A–1736
- Rodgers PW (1968) The response of the horizontal pendulum seismometer to Rayleigh and love waves, tilt, and free oscillations of the earth. *Bull Seismol Soc Am* 58:1384–1406
- Rodgers PW (1994) Self-noise spectra for 34 common electromagnetic seismometer/preamplifier pairs. *Bull Seismol Soc Am* 84:222–229
- Schreiber KU, Hautmann JN, Velikoseltsev A, Wassermann J, Igel H, Otero J, Vernon F, Wells JPR (2009) Ring laser measurements of ground rotations for seismology. *Bull Seismol Soc Am* 99:1190–1198
- Wielandt E (2002) Seismometry. In: Lee WHK (ed) *International earthquake and engineering seismology Part A*. Elsevier Academic Press, New York, pp 283–304

Seismic Microzonation

Fumio Yamazaki¹ and Yoshihisa Maruyama²

¹National Research Institute for Earth Science and Disaster Resilience, Tsukuba, Ibaraki, Japan

²Graduate School of Engineering, Chiba University, Inage-ku, Chiba, Japan

Definition

Seismic microzonation	The mapping of an area on the basis of various factors that can affect the intensity of ground shaking, such as seismic hazard, geological conditions, and topographical features, so as to account for the effects of local conditions on earthquake-induced damage.
-----------------------	---

Introduction

Local site conditions affect the intensity of ground shaking, and as a consequence, the extent of earthquake-induced damage. The amplitude, frequency content, and duration of strong ground motion are significantly influenced by local site conditions. A well-known example is the 1985 Mexico City earthquake. Although the fault rupture of the earthquake was about 350 km away from Mexico City, the city sustained

catastrophic damage due to the strong amplification of the ground motion by soft soil deposits (Seed et al. 1988). The 1989 Loma Prieta earthquake caused extensive damage in the San Francisco Bay Area. The San Francisco Bay mud significantly influenced the amplitude, frequency content, and duration of ground shaking and resulted in the collapse of the northern portion of the I-880 Cypress Viaduct (Earthquake Engineering Research Institute 1990; Kramer 1996). Seismic microzonation provides the basis for site-specific risk analysis, which can assist in the mitigation of earthquake-induced damage.

Methodology

Seismic microzonation typically involves the mapping of predominant periods, soil amplification factors, topographical conditions, liquefaction susceptibility, etc. To draft microzonation maps for a particular region, various data such as existing geological maps, borehole survey data, seismic observation data, and microtremor observation data are collected. Since seismic microzonation entails spatial classification of soil conditions in a small area (e.g., a city), geological data are required for not just a single location but for many locations. In this regard, geological classification maps are most often used as one of the data sources. However, to classify the target area in a more quantitative manner, actual soil profiles obtained from borehole survey data or seismic observation data are better sources. Unfortunately, in most cases, the borehole survey data and/or seismic observation data available for a small area are insufficient. Thus, microtremor observation data have emerged as a popular source for dense spatial information on site amplification characteristics. Three examples of seismic microzonation are described hereafter.

Example 1: Seismic Microzonation Based on Geomorphological Classification Maps

Several seismic microzonation studies in Japan have employed geomorphological and geological data from the Digital National Land Information (DNLI), which is a geospatial information system (GIS) database that covers the whole of Japan with a 1×1 km mesh, to estimate site amplification characteristics (Matsuoka and Midorikawa 1995; Fukuwa et al. 1998; Yamazaki et al. 2000).

Wakamatsu et al. (2004) drafted the Japan Engineering Geomorphologic Classification Map (JEGM) on the basis of the analysis of local geomorphological features at scales of 1:50,000, and all the attributes were digitized and stored in a GIS database. Wakamatsu and Matsuoka (2013) extended the JEGM to 250×250 m grid cells that were categorized into 24 classes on the basis of geomorphological characteristics.

The shear-wave velocity averaged over the upper 30 m (V_s^{30}) is often used as a simplified index of site conditions

(Building Seismic Safety Council 2003). Region-wide site condition maps for California were constructed on the basis of V_s^{30} and the classification of geological units (Wills et al. 2000). The Next Generation of Ground-Motion Attenuation Models (NGA) project was launched in an attempt to collect all publicly available site condition information at strong motion stations. V_s^{30} is used in the absence of site condition information (Chiou et al. 2008). Matsuoka et al. (2006) constructed a nationwide V_s^{30} distribution map using the nationwide shear-wave velocity datasets for Japan, which were obtained from 1000 K-NET and 500 KiK-net seismic stations and the JEGM.

The National Research Institute for Earth Science and Disaster Resilience (NIED), Japan, has developed an open web system that interactively provides seismic hazard maps for Japan; this system is called the Japan Seismic Hazard Information Station (J-SHIS) (Fujiwara et al. 2006). J-SHIS uses the JEGM and V_s^{30} distribution map to draw probabilistic seismic hazard maps for the whole of Japan made by the Headquarters of Earthquake Research Promotion, Japan (Fig. 1).

Example 2: Seismic Microzonation Based on Dense Borehole Data and GIS

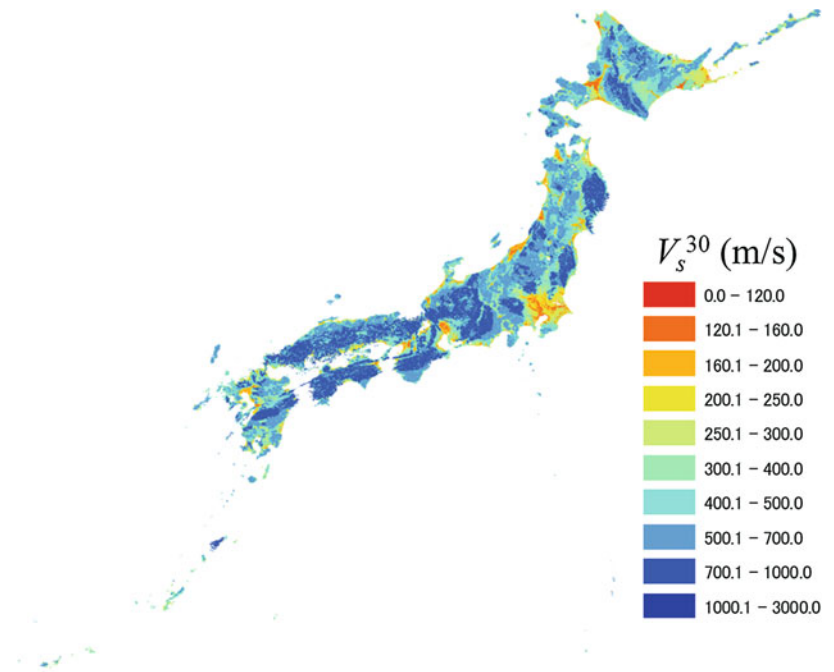
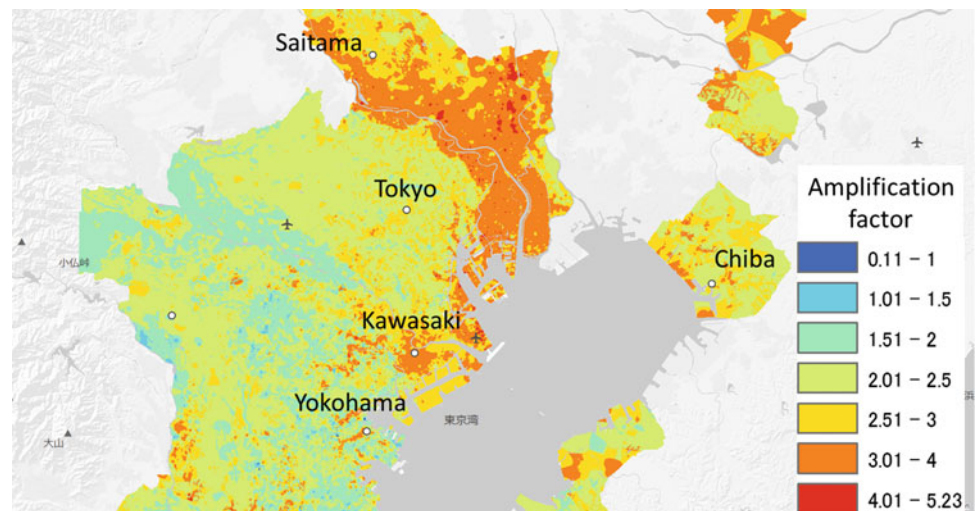
Since 2001, the Tokyo Gas Co., Ltd. has been operating the *Super-Dense Real-time Monitoring of Earthquakes* (SUPREME) system, having about 4000 seismometers (SI-sensors), in order to control natural gas supply soon after the occurrence of earthquakes (Shimizu et al. 2006).

This system employs a GIS to interpolate the monitored spectral intensity (SI) values by using subsoil data from 60,000 boreholes. The digitized borehole data specify the location, depths of soil layers, classification of subsurface soil, standard penetration test (SPT) blow counts, surface elevation, and elevation of the ground water table. Thus, microzonation of the area on the basis of individual borehole data is possible. Shear-wave velocities are estimated from an empirical relationship by using the SPT-N values; then, the average shear-wave velocities in the top 20 m of soil at a borehole site are used to estimate the amplification factors of the SI values (Fig. 2).

The accuracy of seismic microzonation can be confirmed after several years of operating a dense seismic network by evaluating the seismic records obtained for moderate to small earthquake events occurring in that period. The SUPREME system was activated during the 2011 Great East Japan Earthquake, and Tokyo Gas confirmed its effectiveness in emergency response after an earthquake (Inomata and Norito 2012).

Example 3: Seismic Microzonation Based on Microtremor Measurements

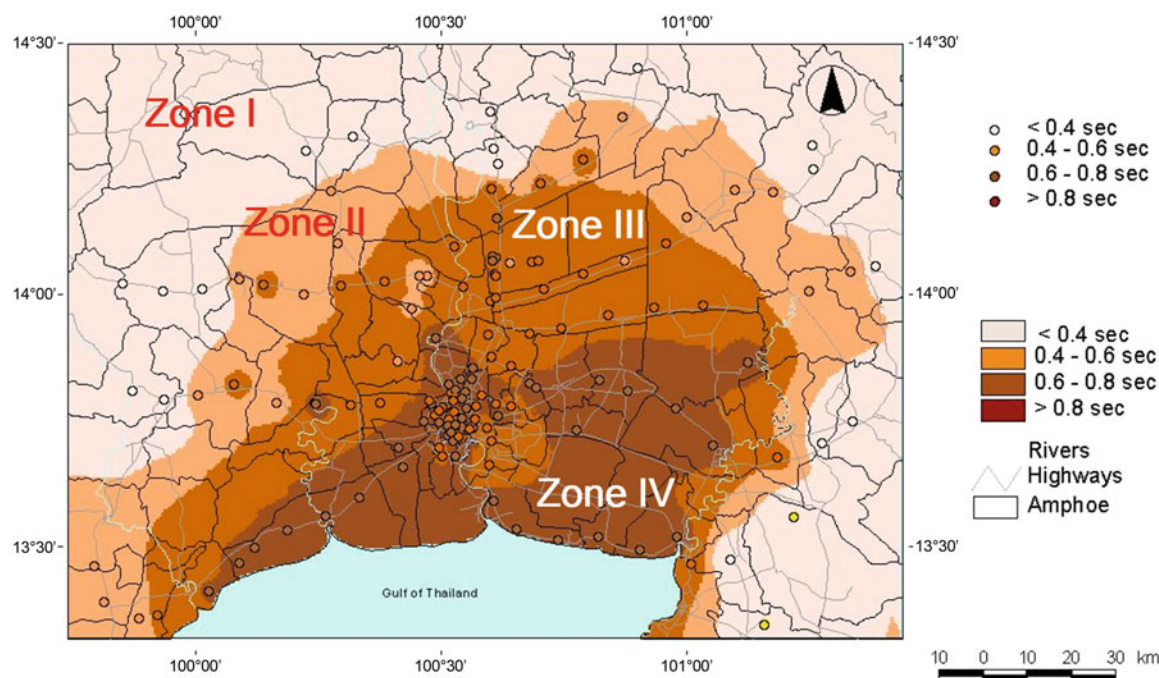
Microtremor measurements have emerged as a popular tool for determining the dynamic properties of soil layers and,

Seismic Microzonation,**Fig. 1** V_s^{30} distribution map for the whole of Japan**Seismic Microzonation,****Fig. 2** Site amplification map of Tokyo and surrounding areas, developed using dense borehole data

hence, are being widely employed for seismic microzonation. In this method, ambient vibrations (of the order of microns) on the earth's surface are measured. The main sources of these vibrations are traffic and industrial and human activities (Kanai 1983; Lermo and Chavez-Garcia 1994). Microtremor measurements can be used to determine the predominant period of vibrations at a site. Nakamura (1989) proposed the horizontal-to-vertical (H/V) spectral ratio method, in which the predominant periods of ground vibrations are determined from the ratio of horizontal and vertical Fourier spectra of the microtremors recorded at a site. Konno and Ohmachi (1998) drafted a map of fundamental periods and amplification

factors for the 23 wards of Tokyo on the basis of microtremor measurements carried out at 546 stations.

Tuladhar et al. (2004) drew a seismic microzonation map for the greater Bangkok area, Thailand, on the basis of microtremor observations carried out at 150 sites. The predominant periods of these sites were obtained by using the H/V method. The estimated predominant periods were validated by comparing them with the transfer functions obtained from one-dimensional wave-propagation analysis conducted at eight sites. According to the variation in the predominant period of the ground, the greater Bangkok area was classified into four zones as follows: Zone I (period less than 0.4 s), Zone II



Seismic Microzonation, Fig. 3 Microzonation of greater Bangkok area on the basis of variation in predominant period

(0.4–0.6 s), Zone III (0.6–0.8 s), and Zone IV (longer than 0.8 s). Figure 3 illustrates the microzonation of the greater Bangkok area on the basis of variation in the predominant period. Based on microtremor array measurement, Jirasakjamroonsri et al. (2019) also evaluated site response characteristics of soft soil deposits in the Bangkok Metropolitan region using the horizontal-to-vertical (H/V) spectral ratio method and the spatial autocorrelation (SPAC) method. Spatial variations of the predominant period and shear-wave velocity profile are presented in a seismic microzonation map.

Summary

The objectives and methodologies to perform seismic microzonation are described and some examples are presented. The three major methods introduced to achieve seismic microzonation are the uses of geomorphological classification maps, dense borehole datasets, and microtremor measurements. The results of seismic microzonation are compiled for a GIS to draft microzonation maps, and they can be used to predict ground motions during disastrous earthquakes and thus can assist in the mitigation of earthquake-induced damage.

Cross-References

- [Earthquakes, Intensity](#)
- [Earthquakes, Strong-Ground Motion](#)

- [Seismic Hazard](#)
- [Seismic Zonation](#)
- [Seismicity, Intraplate](#)
- [Seismology, Global Earthquake Model](#)

Bibliography

- Building Seismic Safety Council (BSSC) (2003) The 2003 HEHRP recommended provisions for the development of seismic regulations for new buildings and other structures. FEMA, Washington, DC
- Chiou B, Darragh R, Gregor N, Silve W (2008) NGA project strong-motion database. *Earthquake Spectra* 24(1):23–44
- Earthquake Engineering Research Institute (1990) Loma Prieta earthquake reconnaissance report. *Earthquake Spectra* 6:1–448
- Fujiwara H, Kawai S, Aoi S, Morikawa N, Senna S, Kobayashi K, Ishii T, Okumura T, Hayakawa Y (2006) National seismic hazard maps of Japan. *Bull Earthq Res Inst Univ Tokyo* 81:221–231
- Fukuwa N, Arakawa M, Nishizaka R (1998) Estimation of site amplification factor using digital national land information. *J Struct Eng* 44(B):77–84. (in Japanese)
- Inomata W, Norito Y (2012) Result of SUPREME (Super-Dense Real time Monitoring Earthquake system for city gas supply) in the great east Japan earthquake. In: *Proceedings of the international symposium on engineering lessons learned from the 2011 great east Japan earthquake*. pp 1508–1513
- Jirasakjamroonsri A, Poovarodom N, Warnitchai P (2019) Seismic site characteristics of shallow sediments in the Bangkok metropolitan region, and their inherent relations. *Bull Eng Geol Environ* 78(3):1327–1343
- Kanai K (1983) *Engineering seismology*. University of Tokyo Press, Tokyo, pp 128–139
- Konno K, Ohmachi T (1998) Ground-motion characteristics estimated from spectral ratio between horizontal and vertical components of microtremor. *Bull Seismol Soc Am* 88(1):228–241

- Kramer SL (1996) Geotechnical earthquake engineering. Prentice Hall, Upper Saddle River
- Lermo J, Chavez-Garcia FJ (1994) Are microtremors useful in site response evaluation? *Bull Seismol Soc Am* 84(5):1350–1364
- Matsuoka M, Midorikawa S (1995) GIS-based integrated seismic hazard mapping for a large metropolitan area. In: *Proceedings of the fifth international conference on seismic zonation, II*. pp 1334–1341
- Matsuoka M, Wakamatsu K, Fujimoto K, Midorikawa S (2006) Average shear-wave velocity mapping using Japan engineering geomorphologic classification map. *J Struct Eng Earthq Eng Japan Soc Civ Eng* 23(1):57s–68s
- Nakamura Y (1989) A method for dynamic characteristics estimation of subsurface using microtremor on the ground surface. *Q Rep Railw Res Inst* 30:25–33
- Seed HB, Romo MP, Sun JJ, Jaime A, Lysmer J (1988) The Mexico earthquake of September 19, 1985—relationships between soil conditions and earthquake ground motions. *Earthquake Spectra* 4:687–729
- Shimizu Y, Yamazaki F, Yasuda S, Towhata I, Suzuki T, Isoyama R, Ishida E, Suetomi I, Koganemaru K, Nakayama W (2006) Development of real-time control system for urban gas supply network. *J Geotech Geoenviron* 132(2):237–249
- Tuladhar R, Yamazaki F, Wamitchai P, Saita J (2004) Seismic microzonation of the greater Bangkok area using microtremor observations. *Earthq Eng Struct Dyn* 33(2):211–225
- Wakamatsu K, Matsuoka M (2013) Nationwide 7.5-arc-second Japan engineering geomorphologic classification map and Vs30 zoning. *J Disaster Res* 8(5):904–911
- Wakamatsu K, Matsuoka M, Hasegawa K, Kubo S, Sugiura M (2004) GIS-based engineering geomorphologic map for nationwide hazard assessment. In: *Proceedings of the 11th international conference on soil dynamics & earthquake engineering and 3rd international conference on earthquake geotechnical engineering*, vol 1. pp 879–886
- Wills CJ, Petersen M, Bryant WA, Reichle M, Saucedo GJ, Tan S, Taylor G, Treiman J (2000) A site-conditions map for California based on geology and shear-wave velocity. *Bull Seismol Soc Am* 90(6B):S187–S208
- Yamazaki F, Wakamatsu K, Onishi J, Shabestari KT (2000) Relationship between geomorphological land classification and site amplification ratio based on JMA strong motion records. *Soil Dyn Earthq Eng* 19(1):41–53

Seismic Monitoring of Nuclear Explosions

Paul G. Richards¹, Zhongliang Wu², Won-Young Kim¹ and David P. Schaff¹

¹Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA

²Institute of Earthquake Forecasting, China Earthquake Administration, Beijing, China

Acronyms

CTBT	Comprehensive Test Ban Treaty or Comprehensive Nuclear-Test-Ban Treaty (its formal name)
CTBTO	CTBT Organization
DPRK	Democratic People's Republic of Korea

IDC	International Data Centre (of the CTBTO)
IMS	International Monitoring System (of the CTBTO)
LTBT	Limited Test Ban Treaty
OSI	On-site inspection
TTBT	Threshold Test Ban Treaty
UNE	Underground Nuclear Explosion

Definition

We take this title to mean the methods of acquiring records of ground motion (seismograms) and of analyzing them for purposes of detecting and identifying those seismic signals that originate from a nuclear explosion and the work of characterizing that explosion.

Introduction

The original development of nuclear weapons, and their first use in 1945, was followed by several decades of further weapon development in which more than two thousand nuclear test explosions were conducted. About 500 of these were carried out in the atmosphere, mostly in the 1950s and 1960s. They generated radioactive fallout that was detected worldwide with some regional concentrations and aroused widespread public opposition to nuclear testing. A few nuclear tests were carried out underwater and in space. Most nuclear test explosions, about 1500, were conducted underground in ways that greatly reduced fallout – the first of them in 1957, in Nevada, USA – generating signals that have been intensively studied by seismologists. Hundreds of these individual nuclear tests consisted of multiple nuclear devices, exploded almost simultaneously.

A ban on nuclear testing in the atmosphere, underwater, or in space was negotiated and went into effect in 1963 between the USA, the USSR, and the UK. Known as the Limited Test Ban Treaty (LTBT), it has since been ratified or acceded to by more than a hundred countries. Though France and China did not sign, and China carried on with nuclear testing in the atmosphere up to 1980, eventually both these countries came to abide by its terms.

The concept of a Comprehensive Test Ban Treaty (CTBT) emerged in the 1950s, intended as a restraint upon nuclear weapon development. It was debated in many forums for more than 40 years and finalized in terms of specific treaty text in September 1996. But this treaty is not in effect (as of 2019), due to continuing debate in specific countries that have not ratified this treaty and whose ratification is needed as a condition for the CTBT to enter into force. They include the Democratic People's Republic of Korea (DPRK), India, and Pakistan (not signed or ratified) and China, Israel, and the USA (signed but not ratified). Those countries that have signed the treaty are effectively adhering to a moratorium on

nuclear testing. They include the five countries recognized as nuclear weapon states by the Non-Proliferation Treaty which went into effect in 1970. Listing them in the order in which they acquired nuclear weapons capability, these are the USA, the USSR (whose CTBT obligations have been assumed by Russia), the UK, France, and China. The two countries that by far have conducted the most nuclear test explosions – the USA with 51% of the world total and the USSR/Russia with 35% – ended nuclear testing in the early 1990s (see Yang et al. 2003, for lists of nuclear explosions conducted in the twentieth century). India and Pakistan carried out nuclear explosive tests in May 1998. To date (November 2019), only the DPRK has conducted such tests in the present century – six of them, in October 2006, May 2009, February 2013, January and September 2016, and September 2017.

Seismic monitoring of nuclear explosions has been an important activity ever since the first nuclear test in July 1945 in New Mexico. Such monitoring is driven by two different objectives that have engaged a range of different institutions and organizations. The first objective, which dominated for the early decades of nuclear testing up to the early 1990s when nuclear explosions were being conducted on average about once a week, was to acquire basic information about military weapons being tested, especially if (from the point of view of the monitoring organization) the tests were being carried out by a potential adversary. Relevant questions were: What countries had nuclear weapon programs, developed to the level of carrying out nuclear explosive tests? And how big were these explosions? The second objective has been in the context of a major initiative in nuclear arms control, namely, to achieve confidence in the capability to monitor compliance with a CTBT, recognizing that many countries – considering whether or not to support such a treaty and to be bound by its terms – would need to have confidence in the monitoring system to some adequate degree. Given that monitoring cannot be done all the way down to zero yield, evaluation of progress toward this second objective entails questions such as: Down to what small size can nuclear explosions be detected, located, identified, and attributed with high confidence? And what are the specific capabilities of different types of monitoring program, applied to different parts of the world, to catch evidence of a nuclear test, should one occur?

Seismology is the most effective technology for monitoring nuclear tests carried out underground, which is the one environment that was not covered by the LTBT and which is also the hardest environment to monitor. The importance of achieving the two objectives stated above has shaped modern seismology itself, in that much of the funding that has led to the facilities and bodies of knowledge now used widely in seismological research, including studies of seismic hazard, was stimulated by government programs intended to improve capabilities for seismic monitoring of nuclear

explosions. These facilities and methods include high-quality ► **“Seismic Instrumentation”**, global networks that monitor for earthquakes as well as explosions, quantitative methods of characterizing seismic sources (various magnitude scales, the moment tensor), theoretical understanding of seismic wave propagation in Earth models of increasing and more realistic complexity, our knowledge of the Earth’s internal structure, and methods of seismic signal detection and interpretation.

The technical capability to monitor explosions, or a perceived lack of such capability, has played a role in the development of policy options on weapons testing and/or arms control and the content of international treaties. A key technical question arising in debates has been down to what value of yield can monitoring be accomplished – and with what level of confidence? Some seismologists claim now that there is no fundamental technical problem with monitoring explosions down to 1 kt, even if determined efforts at evasion must be considered. But there have been assertions that it is possible to muffle and thus hide (or confuse the procedures for identifying) the seismic signal even from a substantial underground explosion at the level of 10 kt or more. These latter assertions do not appear plausible after review of the technical difficulties (NAS 2012); but, as assertions, one finds that they continue to survive.

Seismic monitoring for underground nuclear explosions must be done with recognition of the great variety and number of earthquakes, chemical explosions, and other nonnuclear phenomena that generate seismic signals every day. Efforts to sort out and identify signals from underground nuclear explosions in the midst of signals from these other phenomena have made great progress since they commenced in the 1950s, and improvements in monitoring capability will surely continue to be made.

Sections below describe basic properties of earthquake and explosion signals and different steps in seismic monitoring for nuclear explosions. A review is given of methods used for decades in the era when thousands of kilometers separated nuclear weapon testing activity and monitoring stations (“teleaseismic monitoring”), when nuclear weapon testing was commonplace and there was little incentive to hide testing activity. Descriptions are then given of modern methods (“regional monitoring”) that monitor for very small explosions and the possibility of tests conducted in ways intended to evade discovery. Examples are given of special events that were important in developing effective and in some cases new discriminants; and finally a brief summary is given of monitoring capabilities, as of 2019, emphasizing the utility of data and data products from the International Monitoring System and its associated International Data Centre that are operated today by the Preparatory Commission of the CTBT Organization, headquartered in Vienna, Austria (www.ctbto.org).

Basic Properties of Earthquake and Explosion Signals

Seismic monitoring for underground nuclear explosions has to face the reality of hundreds of earthquakes, chemical explosions, and other nonnuclear phenomena, generating seismic signals daily that will be recorded at multiple stations by any effective monitoring network. But after decades of effort, an extensive infrastructure of national and international agencies now sorts out and identifies the signals from earthquakes, chemical explosions, and the occasional underground nuclear explosion. Modern methods of nuclear explosion monitoring are vastly more capable than they were when this work began in the late 1950s. The improvements have mostly been steady as data quality and quantity from monitoring networks increased but with occasional jumps in capability as new types of analysis have been developed, validated, and eventually applied by different agencies.

Seismic signals are traditionally grouped into teleseismic waves and regional waves, depending on the distance at which they are observed. Teleseismic waves propagate either as ► **“Body Waves”** through the Earth’s deep interior, emerging with periods typically in the range 0.2–10 s at distances greater than about 1500 km, or as ► **“Surface Waves”**, analogous to the ripples on the surface of a pond, with periods of about 15–100 s.

Teleseismic waves were the basis of most US monitoring of foreign nuclear tests prior to 1987. Teleseismic body waves are further subdivided into *P*-waves and *S*-waves. (*P*-waves, which are the fastest-traveling seismic waves and are therefore the first to arrive, are excited efficiently by explosions. Earthquakes tend to excite *S*-waves and surface waves more efficiently.)

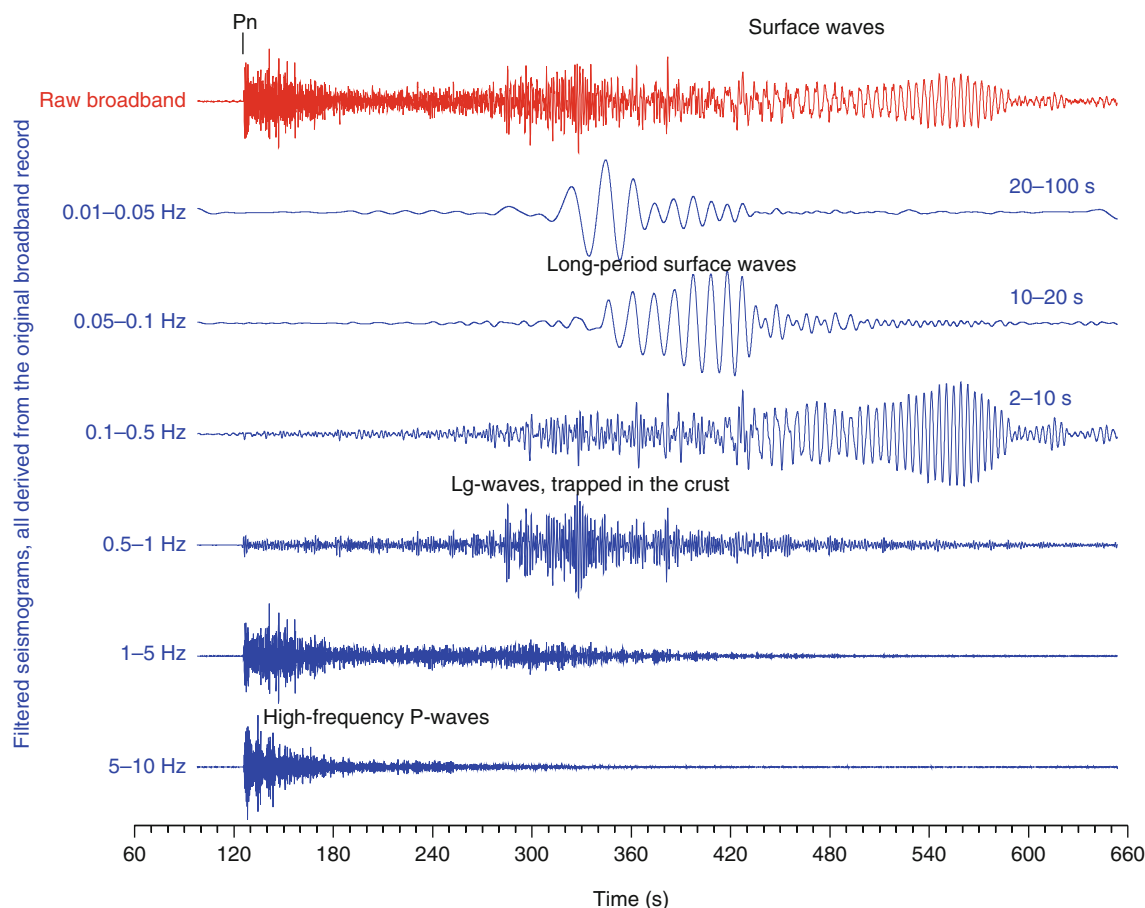
For small (subkiloton) explosions, teleseismic signals can be too weak for detection at distant stations, and monitoring then requires regional signals. Regional seismic waves are of several types, including *P*-waves and *S*-waves, all propagating only at shallow depths (less than 100 km below the Earth’s surface) with periods as short as 0.05 s (frequencies as high as 20 Hz, i.e., cycles per second). Regional waves reach distances up to 1500 km and sometimes beyond, depending on source size and whether the propagation path is an attenuating one or not. They are regional also in the sense that they have speeds and attenuation properties that vary according to details of local structures in the Earth’s crust and uppermost mantle – so they can vary from place to place within continents and oceans.

Figure 1 shows a regional seismogram of a Soviet underground nuclear explosion in Kazakhstan recorded in July 1989 at a distance of slightly less than 1,000 km by a high-quality seismographic station in northwestern China. The original recording is shown in red. Different signals derived from it are shown in blue, each of them filtered to pass information in a particular band of frequencies.

Seismologists characterize the size of seismic signals by means of logarithmic magnitude scales (see *Earthquake, Magnitude*), with each scale based on a different type of seismic wave. A magnitude scale using teleseismic surface waves was first described in the 1930s based on the logarithm (to the base 10) of amplitude of maximum ground displacement due to surface waves with periods about 20 s. It is known as the M_s scale. Another widely used magnitude scale is that based on the amplitude of teleseismic *P*-waves. Known as m_b , it entails measurement of ground motion at about 1 s period. As part of the assignment of M_s and m_b values, for a particular seismic event as recorded at a particular station, a standard correction is applied to account for the distance between the source and the receiver at which the data was obtained. Magnitudes range from about –3 for the smallest observable microearthquakes up to above 9 for the largest earthquake. A 1 kt underground explosion has an m_b roughly about 4, and each year there are about 7500 shallow earthquakes worldwide with $m_b > 4$ (Ringdal 1985). Although use of seismic moment has superseded use of m_b and M_s in much of modern seismology, and magnitude is only an empirical estimator of seismic event size, magnitude scales are still often used in discussion of seismic monitoring because this is a practical way to relate that discussion directly to properties of signal strength. For example, monitoring capability is often characterized in terms of contour maps or shaded maps indicating the magnitude levels down to which detection or identification is deemed possible with given resources, such as signals from a particular network. We conclude this article with such a map (see Fig. 8). Explosion energy is measured in kilotons. A kiloton is formally defined as a trillion calories and is roughly the energy released by exploding a thousand tons of TNT.

The Different Steps in Explosion Monitoring

Nuclear explosion monitoring entails a series of steps, beginning with *detection* of signals (did a particular station detect anything?) and *association* (can we gather all the different signals, recorded by different stations, that originate from the same “event”?). The next steps involve making a *location* estimate and an *identification* (did it have the characteristics of an earthquake, a mining blast, a nuclear weapon test?). Then follow the steps of *yield estimation* (how big was it?) and *attribution* (if it was a nuclear test, what country carried it out?). Once the CTBT is formally in effect, the possibility arises of conducting an on-site inspection (OSI) on the territory of a CTBT signatory state, at and near the location where objective evidence indicates that a nuclear test has occurred. Extensive rules govern procedures for making an OSI request, for its consideration and possible approval by a 51-member Executive Council of the CTBO, for its conduct if approved



Seismic Monitoring of Nuclear Explosions, Fig. 1 The vertical-component seismogram recorded at station WMQ in northwestern China, for an underground nuclear explosion on July 8, 1989, in Kazakhstan at a distance of almost 1,000 km, is shown in red (top). Filtered versions of the original trace in different frequency bands are shown in blue. Time in seconds at bottom is with respect to the time the explosion occurred. Different types of seismic wave propagate at different frequencies, and hence their ground motions show up in different bands. *P*-waves, in this case the regional wave called *Pn* which travels in the uppermost mantle, arrive about 120 s after the explosion at this distance,

involving short-period (high-frequency) motions. Long-period surface waves can be seen in the top two blue traces. Some surface waves arrive up to 600 s after the explosion at this distance and thus travel as much as five times slower than *P*-waves. *S*-waves (weak in this example) are shear waves, traveling slower than *P* waves. A high-frequency wave marked as *Lg*, which is often the largest wave at regional distances from an earthquake but is only weakly excited by explosions, is dominated by shearing motions and is largely trapped in the Earth's crust. The amplitude of ground motion in the longest period band is less than 2% of the amplitude in the short period band from 1 to 5 Hz

(at least 30 votes in favor are required), and for consequential actions if an OSI supports the accusation of a treaty violation. OSIs can involve seismic monitoring for explosion-induced aftershocks, which (from experience in the USA and the former Soviet Union) are known to occur after some underground nuclear test explosions.

Detection

Concerning detection, nuclear explosion monitoring is often done with arrays of sensors, deployed as a group spread out over an area roughly 10 km across, that facilitate methods to enhance signal-to-noise ratios. This is done typically by stacking signals from independent sensors, often with appropriate delays to increase signal strength and reduce noise. Array data can also give estimates of the direction from which signals are arriving.

In making assessments of detection capability, one of the key concepts widely used in seismology is the *magnitude threshold*, above which 90% of the seismic events can be detected at more than three stations (the least number of stations for making routine estimates of the location of the event). Transferring from magnitude to yield, one infers the capability for detecting nuclear tests above a certain size (NAS 2002; NAS 2012).

Association

Association is the effort to identify those sets of detected signals, from different stations, which all originate from the same seismic event. It is one of the hardest steps in practice, particularly when multiple seismic sources around the world are active at roughly the same time (within minutes), resulting in signals from different events that are interlaced in the

waveforms recorded by each station. In such cases, array data can be helpful in resolving which signals correspond to which event.

Location

To obtain a location estimate, typically the arrival times of various seismic waves are measured from the recorded waveforms such as shown in Fig. 1. They are used to find four parameters: the latitude, longitude, depth, and origin time. In this work, it is necessary to know the travel time from any hypothesized source location to any particular seismographic station for any type of seismic wave that the station might observe. In practice, locating seismic events accurately on a global basis (say, to within 10 km of their true location) using networks with stations several hundred km apart requires extensive efforts in station calibration. Thus, it is important to include path-specific travel time corrections to standard travel time models to account for lateral variations of Earth structure (Murphy et al. 2005; Myers et al. 2010). Many authors have shown that greatly improved precision of location estimates can be achieved for a given region if seismic events are located relative to each other and in large numbers – preferably thousands of them or more, all at the same time – rather than one at a time (Waldhauser and Schaff 2008; Schaff et al. 2018b).

Methods of Identification

Identification of the nature of a seismic source on the basis of its seismic signals – that is, making a determination from seismograms as to whether it could be a nuclear explosion, or a natural earthquake, or a mine blast, or something more exotic such as a bolide impacting our planet and exploding in the atmosphere – is a large subject in view of the many possibilities. See, for example, Richards (1988), OTA (1988), Dahlman et al. (2009), and Bowers and Selby (2009). Seismic events generate many different types of seismic wave, in various different frequency bands as shown in Fig. 1, and different types of seismic source generate a different mix of seismic waves. We can make an analogy here with sound waves and the capability of the human ear and brain to analyze them. A deep bass voice, a gunshot, a whistle, and rolling thunder constitute a set of sound sources that are easily distinguished from each other on the basis of their different frequencies, their emergent or impulsive nature, and their duration. It is the mix of information in both the time domain and the frequency domain that is effective.

Seismic methods for discriminating between earthquakes and explosions are based on interpretation of the event location (including its depth); on the relative excitation of a variety of body waves and surface waves; and on features of the frequency spectrum for signals associated with each of these two different types of source. Within these three broad categories, many different methods have been tried, with various degrees of success. As the capabilities of each method

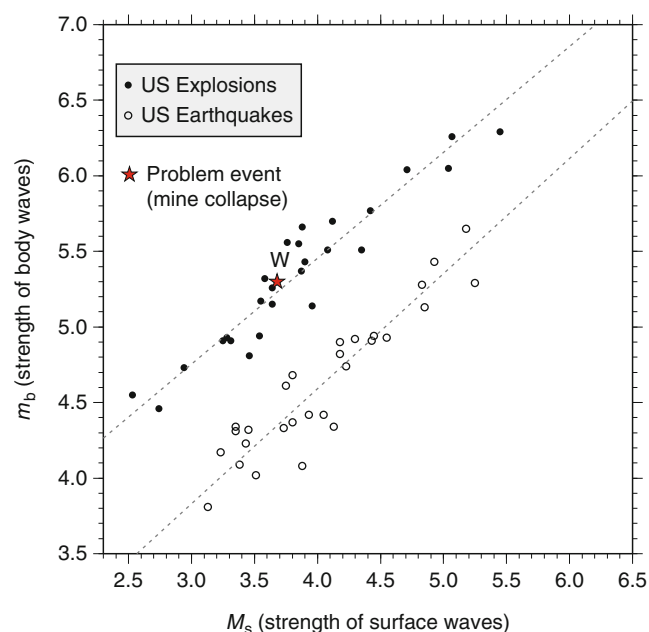
are probed, the question of interest is often: “Down to what size of seismic event does this method of discrimination work?” In some cases discrimination is unambiguous even at very small event size. (For example, however small an event, it may be presumed to be an earthquake if it is located at a depth greater than 15 km below the Earth’s surface. And, even a small event will attract attention, if it occurs in an area that is geologically stable and that for decades has had no seismic activity.)

The most useful methods for discrimination can be listed as follows:

- Interpretation of the location. Is the event in a seismic or an aseismic area? Below the floor of an ocean? At depth below a continent? There is an important role here for common sense: seismic events in Canada attract less attention from western monitoring agencies than such events in the DPRK (though a seismic event in the Canadian Shield would still attract attention and intensive study).
- Relative amplitude of body waves and surface waves. This can be studied by plotting the event of interest on an M_s : m_b diagram, as shown in Fig. 2. The surface-wave amplitude is read typically on the vertical component from signals with period about 20 s and the body-wave amplitude at about 1 s period. (Though effective for large enough events, an explosion with m_b much below 4.5 may not have large enough surface-wave signals at teleseismic distances to apply this method dependably.)
- Use of the observed “first motion” of the ground. Does the initial P -wave motion of the ground indicate compressions radiated to all directions from the source, leading to upward motions, as would be the case for a simple explosion? Or, are dilatations recorded at some azimuths, leading to downward motions – as would sometimes be expected from earthquakes but not from explosions?

The methods described so far in this section have concerned the use of teleseismic signals, which can be used to monitor effectively for high magnitudes and on down to somewhere in the magnitude range from 4.0 to 4.5. Since the late 1980s/early 1990s, there has been growing recognition of the merits of regional waves, to enable monitoring down to far lower magnitudes, often well below magnitude 3. The method is based upon the general observation that explosion signals, when compared to earthquakes, have much stronger P -waves at high frequency, whereas signals from earthquakes have stronger S -waves (and surface waves).

This modern method is being studied with frequencies in the range 0.5–20 Hz. and sometimes even higher. An example is shown in Fig. 3 comparing regional seismic signals of a very small earthquake and a small explosion. The method has been demonstrated by seismologists in China, even down to around m_b 2 (Pan et al. 2007).

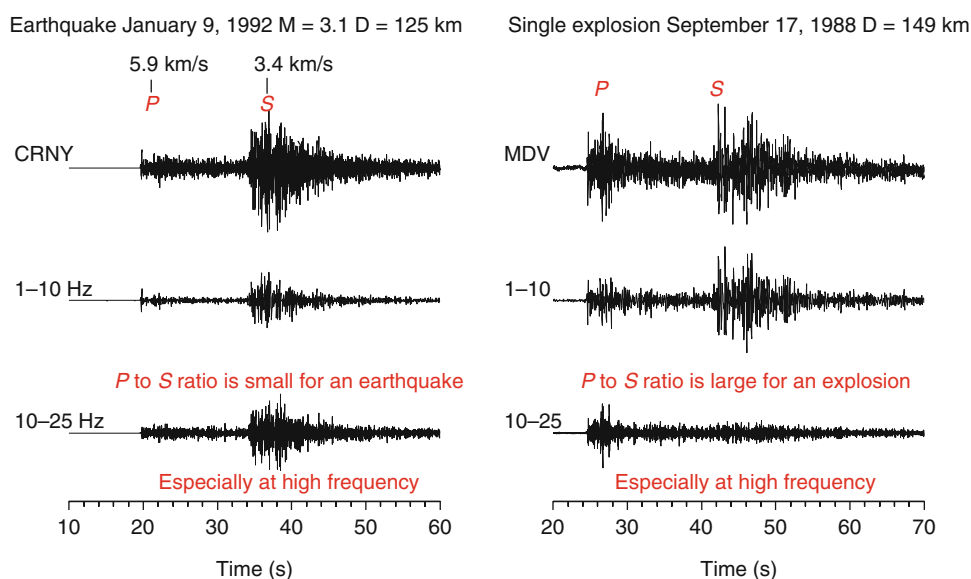


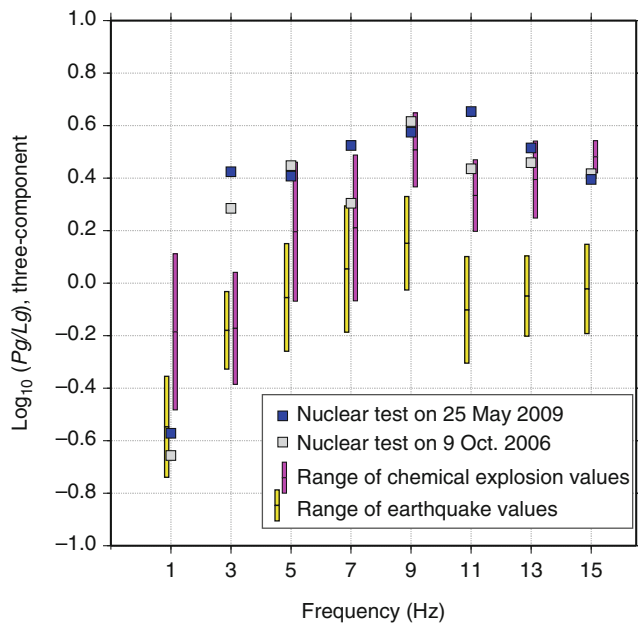
Seismic Monitoring of Nuclear Explosions, Fig. 2 An M_s : m_b diagram from Bowers and Walter (2002). It can be seen here that for seismic events of the same M_s value, earthquakes have a significantly smaller m_b value than do the explosions. The offset is about 0.8 m_b units, at $M_s = 5$. Because magnitudes are based on logarithmic scales, and $10^{0.8} \sim 6$, it follows that at frequencies near those at which body wave magnitude is measured (about 1 Hz), the P -waves from an underground nuclear explosion are typically about 6 times larger than such waves from an earthquake having the same strength of surface waves. Also, indicated by the red star are the body-wave and surface-wave magnitudes of an interesting but fortunately rare event, a large mine collapse with P -wave magnitude greater than 5. This event, which plots with the explosion population, is discussed further below – see Fig. 6

As an important example of this development, Fig. 4 shows the results of an analysis of the P -wave and S -wave spectra, pertinent to identifying the very small underground nuclear explosion conducted by the DPRK on October 9, 2006, and a larger test nearly 3 years later on May 25, 2009. The smaller explosion took place at 0135 hours (GMT), and by 0706 hours the US Geological Survey (USGS) had issued a report based on seismic signals from 20 stations around the world including sites in China, the Republic of Korea (ROK), Russia, Japan, Kazakhstan, Kyrgyzstan, Alaska, and Nevada. Its magnitude, about 4, indicated a subkiloton yield (see Koper et al. 2008, who discuss the uncertainty of estimating yield in view of the variability of seismic signal excitation for shots of different depth). From teleseismic signals, the explosive nature of such a small seismic event can be difficult to recognize (they could be signals from an earthquake). Fortunately, discrimination for events such as this is very clear providing high-quality regional data is available.

In this analysis, the original seismograms from station MDJ, located in China, are filtered to pass narrow frequency bands as illustrated in blue in Fig. 1 but this time with bands centered on each of the frequencies from 1, 3, 5, 7, 9, 11, 13, to 15 Hz as indicated for the horizontal axis in Fig. 4. Here the amplitudes of the Pg and Lg waves are measured in each narrow band. The amplitude ratio is formed (the “spectral ratio”), and the quantitative comparison can begin. Figure 4 shows how this ratio varies with frequency for the set of eight earthquakes and for the set of four small chemical explosions. The ratio differs for these two populations as frequency rises, and the separation between them is very clear at the high frequencies (from 9 to 15 Hz in this case). The spectral ratios

Seismic Monitoring of Nuclear Explosions, Fig. 3 Typical vertical-component records from an earthquake and an explosion. Traces plotted are unfiltered (top); low-frequency band-pass filtered (middle); and high-frequency band-pass filtered (bottom). (From Kim et al. 1993)





Seismic Monitoring of Nuclear Explosions, Fig. 4 Spectral ratios are shown, for nuclear explosions carried out by the DPRK in 2006 and 2009, as measured from waveforms recorded at station MDJ in China (distance, about 370 km). They are compared with these ratios for a small group of earthquakes and another group of chemical explosions, all in the vicinity of the DPRK's nuclear test site (about 370 km to the south of MDJ). Colored bars represent $\pm$ one standard deviation in the ratios for chemical explosions (yellow) and small earthquakes (magenta). The spectral ratios for these two events in the DPRK are both explosion-like

of the signals recorded for the nuclear events of 2006 and 2009 are like those of the known chemical explosions.

This successful seismic discriminant based upon regional waves is important in enabling capability to be extended down to lower magnitudes than can be monitored teleseismically. In practice, there is often very little difference between the magnitude thresholds for detection (at enough stations to enable a useful location estimate) and identification, since so many regions of the Earth are now monitored to low magnitude for earthquakes as part of investigations into seismic hazard. It may take only one regional seismogram to enable discrimination to be carried out with high confidence (provided the recording is of adequate quality and is for a station that has an archive of signals from previous known earthquakes and explosions).

Along with the use of regional seismic waves and their spectral ratios at 5 Hz and higher, another successful method for distinguishing between earthquakes and explosions is the use of observed seismic waveforms to make estimates of the set of forces that appear to be acting at the seismic source. The set of forces here is quantified by what seismologists call the *moment tensor*. As shown by Ford et al. (2009) from study of numerous earthquakes and underground explosions, seismic events separate into specific populations as determined by the way their moment tensors behave – whether they are more

representative of the all-around (isotropic) features of an explosion, or of the type of shearing motions more typical of an earthquake.

In general for underground tests, seismic data alone cannot distinguish between nuclear explosions and chemical explosions in which all the material making up the explosive was fired within less than about a tenth of a second. But chemical explosions, if large, are very rarely fired in this way. In the case of the DPRK tests of 2006 and 2009, both of which were announced as nuclear, objective evidence for the nuclear nature of the 2006 explosion came from several different detections of radionuclides that are diagnostic of a nuclear explosion. Although such radionuclides were not detected from the 2009 explosion, it was so large as to be implausible as a chemical explosion, since it would have to have consisted of literally thousands of tons of explosive material.

We return later in this article to a discussion of additional nuclear tests in the DPRK up to 2017, some of which generated small earthquakes (explosion-induced aftershocks).

Yield Estimation

Yield estimation was of particular importance in the years following 1974 when a bilateral treaty between the USA and the USSR was negotiated, intended to go into effect in 1976. This was the Threshold Test Ban Treaty (TTBT), limiting the size of underground nuclear explosions conducted by these two countries to a yield of not more than 150 kt. The TTBT proved contentious, with each side sending the other several inquiries asserting that the agreed-upon limits had possibly been exceeded (Timerbaev 2007). But this treaty, and a related treaty placing yield limitations on so-called Peaceful Nuclear Explosions, finally entered into force in 1990.

Both treaties have become less important since the CTBT was finalized and a nuclear testing moratorium by the signatory countries began in 1996. Yield estimation is however still important as an exercise in the interpretation of signals from the few underground explosions since that date, specifically those of India and Pakistan in 1998 and of the DPRK in the present century.

For several tens of underground nuclear explosions, most of them at the Nevada Test Site in the USA or at Kazakhstan's Semipalatinsk Test Site when it was part of the Soviet Union, the yield has subsequently been announced by agencies involved in the conduct of the test. It has therefore been possible to calibrate observed seismic magnitudes for these tests against the announced yields, and an example is given in Fig. 5 using m_b values and yields reported for Nevada explosions in tuff and rhyolite.

The line $m_b = 4.05 + 0.75 \log(\text{yield})$ fits the data well (yield in kilotons). Such a calibration curve can be applied to obtain a seismic yield estimate for Nevada explosions with unannounced yield. But it requires correction, prior to its use in obtaining a seismic yield estimate for an explosion at a

different site. This must be done, to allow for physical and geological differences between the sites. For example, in different rock types, there can be different efficiencies in the coupling of nuclear yield into seismic energy and differences in the propagation efficiencies as seismic waves travel out from the source of interest, as compared to seismic signals from a Nevada explosion. In this connection, it is of interest to note the m_b value and yield for the US nuclear explosion LONGSHOT (conducted in 1965 in the volcanic breccias of an Aleutian island). The m_b value is 5.9, corresponding to a yield of about 300 kt. if the Nevada curve of Fig. 5 is applied directly. But the announced yield for LONGSHOT is 80 kt. One way to obtain a calibration curve for the Aleutians is therefore to add a correction of about 0.4 m_b units to the Nevada values of m_b at a given yield, before the curve of Fig. 5 is used to supply a seismic yield estimate in this new location. This m_b correction, for a site differing from that where a calibration curve is directly available, is called the *test site bias*. If the bias correction is not applied, then a Nevada magnitude–yield curve can give too high a seismic yield estimate for a non-Nevada explosion.

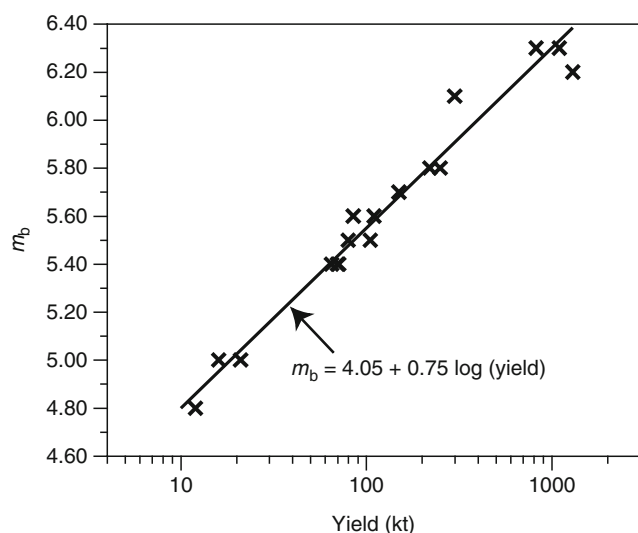
Note that the Nevada Test Site is in a region of active tectonics, with significant episodes of volcanism in the last few million years, resulting in high temperatures within the upper mantle and thus anomalous attenuation of seismic waves propagating through the hot and partially molten upper layers of the Earth, 100 or 200 km in thickness beneath the Nevada Test Site. Such propagation through an attenuating medium is presumed to be a contributing cause of bias.

The existence of m_b bias has long been known in seismology in connection with what is called “station bias.” By this

term is meant the systematic difference between mean m_b values (obtained for a particular seismic event by averaging reported m_b from seismometers all over the globe) and m_b reported by just one station. For example, the station BMO in Oregon (another region of active tectonism) has reported m_b values that for a given earthquake are typically about 0.3 units below the global average; and station KJN in Finland (in a stable shield region) reports values about 0.15 m_b units higher than the average. Their station bias values are thus -0.3 and $+0.15$, respectively. Station bias values commonly range over ± 0.3 m_b units, so it may be expected that source region bias (which is what must be applied when a standard m_b –yield curve is used for different source regions) can also range over about 0.6 m_b units.

The nuclear weapon test site of the USSR that conducted the most underground nuclear explosions was near the city of Semipalatinsk, in northeastern Kazakhstan. Several multi-megaton underground explosions were conducted on Russia’s Novaya Zemlya island test site, far to the north of continental Eurasia (see Khalturin et al. 2005). But these were all prior to the intended date of entry-into-force of the TTBT (March 1976). After that date, the magnitude of the largest underground tests at Semipalatinsk rose higher and higher over several years, with some magnitudes exceeding 6.1. Such magnitudes, according to the Nevada Test Site formula discussed above, $m_b = 4.05 + 0.75 \log(\text{yield})$, implied yields greater than 500 kt, far in excess of the TTBT limit (150 kt). Intensive discussion in political and technical areas ensued, with stronger and stronger evidence accumulating to indicate a substantial test site bias between the Nevada and Semipalatinsk test sites. For example, it was of great interest that teleseismic signals from the largest underground explosions from these two tests, if recorded at the same station in a shield region, looked significantly different. The teleseismic P -wave from a large underground explosion at the site in Kazakhstan would routinely have frequency content at the 5 Hz level and sometimes higher (Der et al. 1985). The signal from Nevada would not contain such high frequencies. It was as if the signal from Nevada had passed through some type of filter, which of course would reduce its amplitude. Correcting for that effect would mean that the appropriate relation between magnitude and yield for an underground nuclear explosion at Semipalatinsk had the form

$m_b = 4.05 + \text{bias} + 0.75 \log(\text{Yield})$, and Ringdal et al. (1992) and Murphy (1996) among many others concluded that the appropriate formula relating teleseismic P -wave magnitude and yield at Semipalatinsk should be this equation with a bias of 0.4. Support for their conclusion came from many arguments (see Richards 1988, for a review). But in the political realm, the most persuasive was the very practical one associated with a Joint Verification Experiment of September 14, 1988, in which a team from the USA at the Semipalatinsk Test Site was allowed to make close-in measurements (within



Seismic Monitoring of Nuclear Explosions, Fig. 5 Seismic magnitude m_b vs announced yield, for 17 Nevada Test Site nuclear explosions in tuff and rhyolite. The straight line here, which fits the data quite well, can be used to make a yield estimate of other events at this test site, in similar rock, if the seismic magnitude is known. (Data from Nuttli 1986)

a few tens of meters) of a large Soviet underground nuclear explosion, in particular of the speed and extent of the shock wave it sent out into rock near the source at that test site. From such shock wave measurements, a reliable non-seismic method provided an accurate yield estimate (it was in the range 100–150 kt). Stations around the world provided measurements teleseismically, giving a seismic magnitude around 6.1 – comparable with the largest magnitudes of Semipalatinsk explosions since 1976, indicating that they too had been conducted in a way that respected the 150 kt limit of the TTBT. A reciprocal Joint Verification Experiment had been conducted at the Nevada Test Site, on August 17, 1988, with a Russian team making its own close-in measurements of the shock wave from a large US underground nuclear test intended to be in the range 100–150 kt. Apparently, the yield of this explosion somewhat exceeded 150 kt. Timerbaev (2007) and news reports give it as 180 kt.

Special Events

The work of monitoring – for both earthquakes and explosions – is done in practice by hundreds of professionals who process the vast majority of seismic events routinely and who also look out for the occasional events that, in the context of monitoring for the possibility of underground nuclear explosions, exhibit interesting characteristics and which may then become the subject of special study.

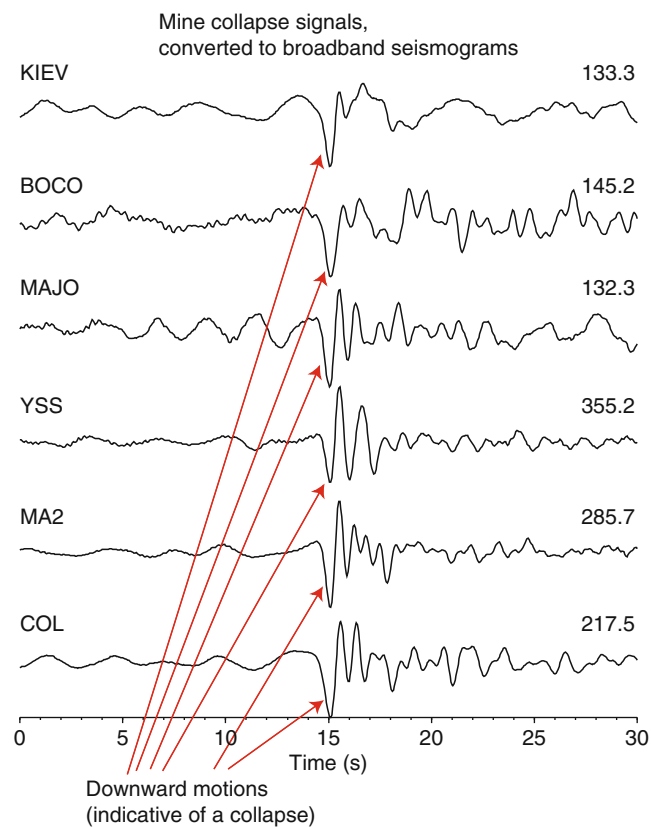
These special events have stimulated the development of effective new discrimination techniques and a better appreciation of overall monitoring capability. Examples include a mine collapse in 1989 in Germany and two such collapses in 1995, in the Urals (Russia) and in Wyoming (USA); a small earthquake of magnitude 3.5 and its smaller aftershock in 1997 beneath the Kara Sea near Russia's former nuclear test site on Novaya Zemlya; the series of underground nuclear test explosions carried out by India and Pakistan in 1998; underwater explosions in 2000 associated with the loss of a Russian submarine in the Barents Sea; and a number of small seismic events near the DPRK's nuclear test site in addition to the six announced nuclear test explosions at that location from 2006 to 2017.

The mining collapses were seismically detected all over the world. For example, teleseismic *P*-waves from the Wyoming event of 1995 were detected in South and North America, Europe, and Asia. Mining collapses such as these have caused concern because their mix of surface waves and body waves as recorded teleseismically can appear explosion-like using the classical M_s : m_b discriminant, as shown in Fig. 2 (see above). But a careful analysis of regional and teleseismic waves from these events has showed that although the surface waves were quite weak, and in this respect seemed explosion-like, they had the wrong sign. Therefore the motion at the

source was implosive (the ground had moved inward toward the source and thus downward), rather than explosive. Indeed, mining collapses are an implosion phenomenon, and it was important to learn that their implosive nature could be reliably determined from seismic recordings. Teleseismic waveforms from the Wyoming mine collapse are shown in Fig. 6. This is an example of the use of what seismologists call the “first motion” of the *P*-wave, which is clearly downward in these data.

The Kara Sea earthquake was too small to apply the M_s : m_b discriminant (the surface waves were too small to measure reliably). This event showed the importance of accurate locations and of using spectral ratios of regionally recorded *P*- and *S*-waves to discriminate small events (Richards and Kim 1997).

As we have discussed earlier, the DPRK nuclear test of 2006 was of interest as an example of a nuclear explosion that was promptly detected globally, though its yield has been estimated as less than 1 kt. This event required regional seismic data in order to determine that indeed an explosion had been carried out and that the signals were not from an



Seismic Monitoring of Nuclear Explosions, Fig. 6 The *P*-wave signals of the Wyoming mine collapse are shown at six teleseismic stations, processed to bring out the fact that the first motion of the ground at these stations is downward, indicative of an implosion rather than explosion (for which the first motion would be upward). (From Bowers and Walter 2002)

earthquake (Kim and Richards 2007). Subsequently, xenon radionuclides were detected that decisively identified this explosion as nuclear.

The DPRK at this time of writing (November 2019) has conducted six nuclear explosive tests, the first (October 9, 2006) with m_b about 4 and the last (September 3, 2017) with m_b greater than 6. The magnitude of the largest test has been assigned values lying mostly in the range from 6 to 6.3, reflecting the fact that the number of seismographic instruments providing data today is roughly a hundred times the number available in the 1980s and 1990s when magnitude–yield relations were being worked out, and the effects of station bias (for many new stations) are not yet quantified. Of interest as a demonstration of capability to monitor for even small seismic events is that the huge 2017 explosion (with yield around 200 kt) was followed by tens of seismic events that were detected and intensely analyzed by groups in neighboring countries – including China and the Republic of Korea (ROK) – as well as in distant countries and by the Vienna-based IDC using IMS data. These events have been identified as small earthquakes, down to magnitude of about 2.2, apparently induced by the very large explosion of 2017 (e.g., Kim et al. 2018; Schaff et al. 2018a). As shown in Fig. 7, thirteen of these small earthquakes occurred from Sept 2017 to April

2018 close to a northeast–southwest line about 700 m long, lying about 8 km to the north of the main mountain – Mount Mantap – in which most of the DPRK’s test explosions took place. Low-magnitude seismic activity has continued in the region, with events in January and June of 2019. The earthquakes have relative locations determined via analysis of whole waveforms recorded regionally in China and ROK (Schaff et al. 2018b). Their overall location to the north of the test site is determined from work of Tian et al. (2018), who had access to data from numerous seismic stations in China.

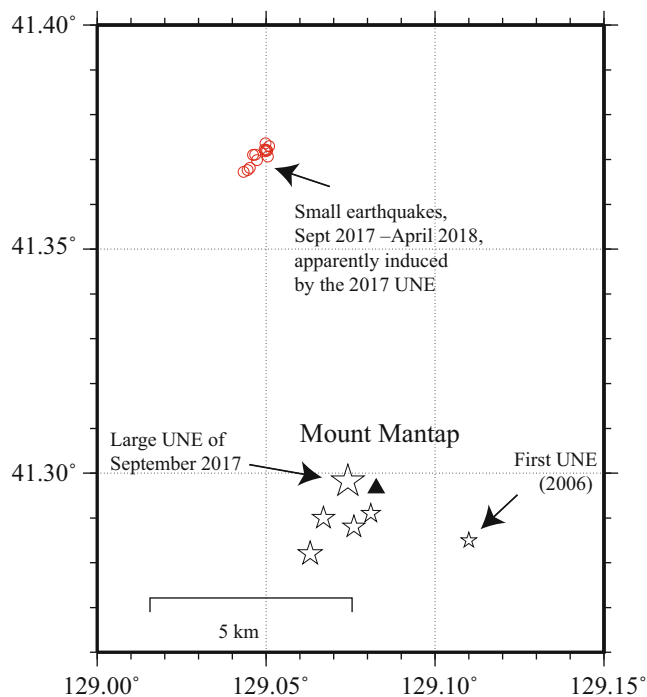
The overall point here is that the small 2006 nuclear explosion in the DPRK was promptly well-characterized; and in later years it has become possible to detect, locate, and identify seismic events in the region that are a hundred times smaller.

Evasion

Several methods have been proposed, by which underground explosions might be concealed. One method is simply to make them small enough; but then there would be relatively little to learn, from the point of view of a weapon designer. The more important methods are those which combine as many features as possible, designed to reduce seismic signal-to-noise ratios at all relevant monitoring stations. Proposed methods include emplacement of the nuclear device in material such as dry alluvium, to reduce the coupling of explosion energy into seismic signal (but that method is likely to result in leakage of detectable radioactivity); waiting until a sufficiently large natural earthquake occurs fairly near a test site (which presents the formidable challenge of identifying the event within a few minutes of its occurrence as large enough and then within a couple of minutes executing the weapon test so that its seismic signals would hopefully be swamped by the large and prolonged signals from the earthquake); and setting off a sequence of several explosions that are designed to simulate a natural earthquake signal.

Careful study of each of these methods indicates that they are relatively ineffective in comparison with the methods known as cavity decoupling and mine masking, which we next discuss and which are widely regarded as setting the practical levels down to which seismic monitoring of nuclear explosions is possible.

When an underground explosive device is tightly packed into its hole (“tamped” or “fully coupled”) and is detonated at sufficient depth to contain all radioactive products, a shock wave travels some distance from the shot point out into the surrounding rock at speeds that exceed the normal P -wave speed. This nonlinear phenomenon reduces at sufficient distance from the shot point, and thereafter the wave propagation can be regarded as elastic. The so-called elastic radius for a tamped explosion, i.e., the radius beyond which wave



Seismic Monitoring of Nuclear Explosions, Fig. 7 A map showing locations for the summit of Mount Mantap (black triangle); the first of the DPRK’s UNEs (small star), to the east and south of this summit; five subsequent UNEs conducted within the mountain (larger stars) including the large test explosion of September 3, 2017; and a series of small aftershocks aligned over several hundred meters, about 8 km to the north of Mount Mantap (red circles). (Adapted from Kim et al. 2018)

propagation is linear, is roughly 100 m times the cube root of the yield (in kilotons).

If the explosion is set off inside a large underground cavity instead of being tamped, then the shock wave set up in the rock can be weakened or even eliminated, in which case only elastic waves are radiated. The explosion is said to be fully decoupled if only elastic waves result, and theoretical work begun in 1958 has addressed the question of how much weaker the seismic signal might be made. Theoretical work has indicated that signals could thereby be reduced by factors in the range 50–100, compared to a tamped explosion. The cavity radius itself is the “elastic radius” for a fully decoupled shot. For salt, the cavity radius required for full decoupling has been estimated as about 25 m times the cube root of the yield (in kilotons). For hard rock the cavity size for full decoupling is comparable; for weak salt it is somewhat greater. See Sykes (1996) for further discussion and Denny and Goodman (1990) for estimates of the decoupling factor derived from the practical experience in 1966 of carrying out a small nuclear explosion (about 0.38 kt) in the cavity produced by a tamped shot of 5.3 kt conducted 2 years earlier in a Mississippi salt dome. They conclude that the amplitude reduction is about 70, at low frequencies, for salt. At frequencies that have conventionally been used for seismic monitoring, the seismic signal strength is proportional (very roughly) to the volume within the elastic radius. This volume is substantially reduced by fully decoupling, which is the reason why cavity decoupling has been proposed as offering the technical possibility of a clandestine program of nuclear testing. However, the signal strength is not nearly so strongly reduced, by decoupling, at frequencies above that associated with resonances of the internal surface at the elastic radius. In practice, the frequency above which decoupling is likely to be substantially less effective is around 10–20 Hz, divided by the cube root of the yield (in kt). The overall effect on the seismic signals from a fully decoupled shot of yield Y , given the results of Denny and Goodman, is to make these signals look like those from a tamped shot of yield $Y/70$.

A thorough discussion of decoupling as an evasion scenario would have to include several non-seismological considerations. These are reviewed in NAS (2012) and include the military significance of being able to carry out nuclear tests up to various different yield levels (e.g., 0.1, 1, or 10 kt.) and the political consequences if an attempted clandestine test program was uncovered. Technical considerations include methods of (clandestine) cavity construction and the capabilities of non-seismological surveillance techniques. Preventing the leakage of radioactivity from an underground cavity would be a challenge, given that much of the energy of a decoupled explosion goes into pumping up the pressure in the cavity. While some assert that clandestine use of cavity decoupling would be so difficult to execute that it belongs to the realm of fantasy, others have been persuaded that the risk

might indeed be manageable and that estimates of concealable yields, under this evasion scenario, must be made. The NAS (2002) report describes ten “layers of difficulty” with successfully hiding an underground nuclear explosion via cavity decoupling, concluding that even a nation committing significant resources to this work could not have confidence in being able to get away with tests above 1 or 2 kt.

The evasion scenario known as mine masking hypothesizes the execution of an underground nuclear weapon test explosion in a mining region, concurrently with a large mine blast. Such blasts in a big commercial operation consist of hundreds of separate charges, fired in sequence to break and/or move as much rock as possible, in a procedure known as ripple firing (Khalturin et al. 1998). Regardless of the logistical difficulties of such a scenario, estimates of the possibilities of concealment via this approach can come from taking examples of signals from large mine blasts, and signals from small underground nuclear explosions, and then adding them together before subjecting them to the methods used to discriminate between various types of seismic events. What is typically found is that the maximum size of the identifiable waves (e.g., the P -waves) from the mine blast is about that expected from individual sub-blasts (commonly called “delays”), and these amplitudes are spread out over a longer time in seismograms.

A study of mine masking possibilities by Smith (1993) used several different examples of mine blast seismograms together with single-fired explosion records and found a number of features that could be used to identify a simultaneous shot within a ripple-fired blast. He concluded that to conceal a single-fired deep detonation (depth is required for containment of radionuclides), the single explosive shot should not exceed 10% of the total explosive.

The conclusion here is that mine blasts are not effective for concealing large releases of energy at the level associated with kiloton-scale nuclear weapons tests, unless the nuclear explosion was subject to efforts at decoupling. Again non-seismic considerations arise, including an assessment of the plausibility of carrying out a complicated decoupled nuclear explosion at the same time and location as a large mine blast that would itself attract some level of monitoring attention – particularly if the seismic signals seemed unusual in comparison with those from prior blasting in the region.

Event Detection Capability of the International Monitoring System

In 1976, a group of international scientists was established at the Conference on Disarmament in Geneva, for the study of monitoring technologies and data analysis methods in the context of supporting a future test ban treaty. This group of scientific experts (GSE) played an essential role in laying

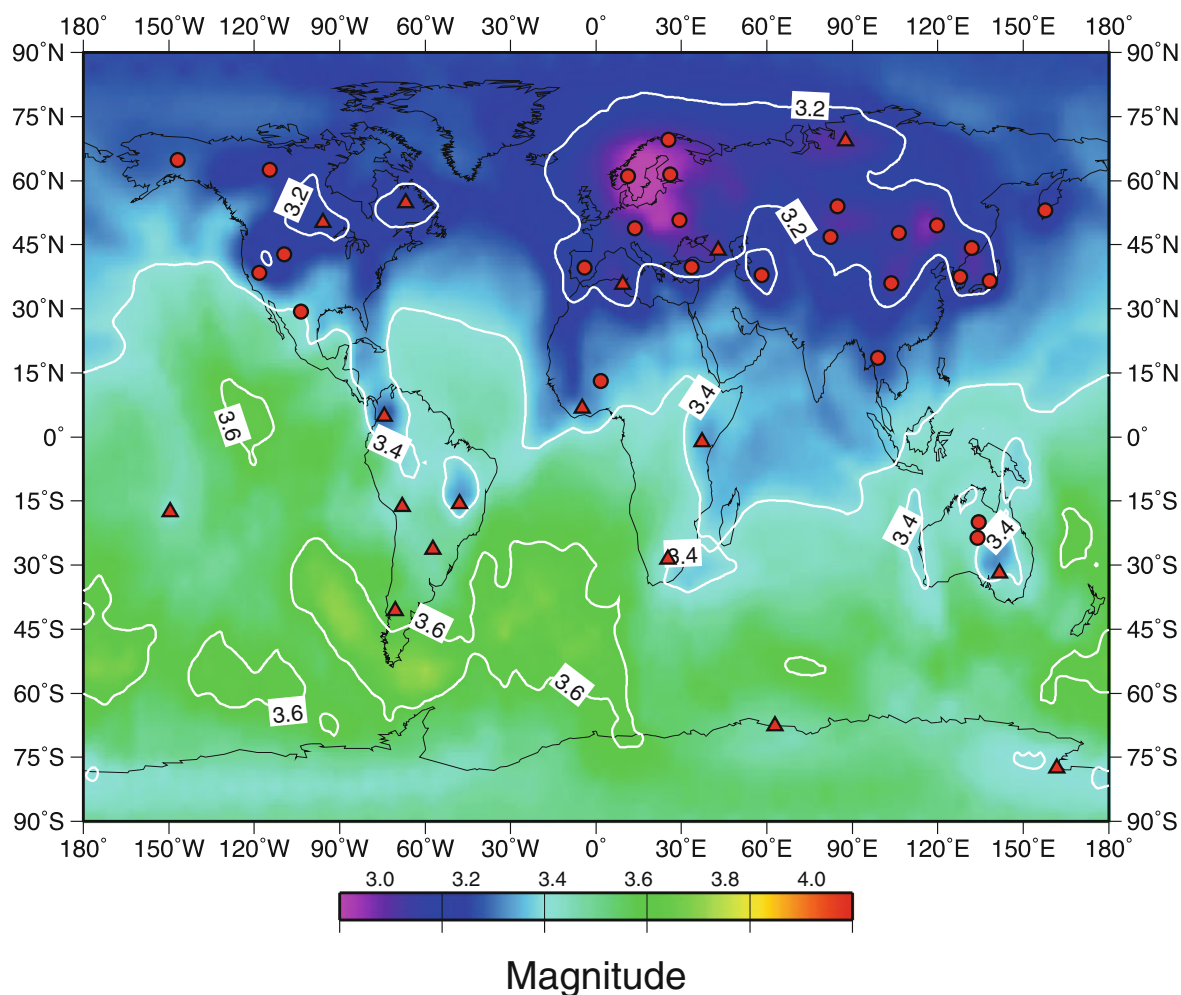
the scientific groundwork for the final stage of CTBT negotiations conducted from 1994 to 1996. Prior to the final stage, GSE organized a series of technical tests which contributed significantly to the development of the international system being built today to support treaty verification.

The text of the CTBT includes a description of networks to monitor treaty compliance using hydroacoustic, infrasound, and radionuclide technologies as well as seismological methods. The CTBT Organization (CTBTO) operates an International Monitoring System, as well as an International Data Centre to analyze signals sent via satellite to headquarters in Vienna. Extensive descriptive material on these networks is available online (see <http://www.ctbto.org>).

To implement the CTBT seismic monitoring system, a sequential four-step process is needed to build each station

(CTBTO PrepComm 2009): (1) site survey; (2) installation; (3) certification; and (4) operation. It must be demonstrated for IMS stations that data received at the International Data Centre (IDC) are *authentic*. This is achieved through a special digital “signature” embedded in the data flow from each station. The IMS station must be certified to ensure that all of its equipment, infrastructure, and settings meet the technical specifications set by the CTBTO and to ensure also that all data are transmitted to the IDC through the Global Communication Infrastructure (GCI) in a timely manner.

Here, we note that the primary seismographic network is to consist of 50 stations, many of them arrays; and that location estimates are based upon detection of signal at three stations or more. An auxiliary network of 120 continuously operating stations is to provide seismic waveform



Seismic Monitoring of Nuclear Explosions, Fig. 8 A map illustrating detection capability of the IMS primary seismographic network. This network is intended to have 50 stations, but it is incomplete as of 2019. Here, is shown the capability of the network, based on practical experience with 43 stations. The capability is represented by the magnitude of the smallest seismic event that would be detected with a 90% probability

by three stations or more. The method used to develop such maps is described by Kværna and Ringdal (2013). A complete IMS network (50 stations) may be expected to lower the illustrated magnitude threshold by about 0.1 magnitude units, in some regions. (Figure, courtesy of T. Kværna.)

data, again via satellite, in order to help characterize the events detected by the primary network. Although these two networks are not completely built (as of August 2019, 44 primary and 108 auxiliary stations have been certified), enough stations operate to provide good indications of what the detection capability will be when all stations are installed and providing data. Their performance today, although the networks are incomplete, is considerably better than what was envisaged for the IMS and IDC during the CTBT negotiations of the 1990s.

Figure 8 shows a map of the detection capability of the primary seismic network of the IMS, based on practical experience with 43 operating stations. This network is planned eventually to have 50 stations. Capability is expressed in terms of magnitude thresholds, above which 90% of the seismic events are expected to be detected at enough stations to provide a location estimate. The work of identifying events is left to member states. We note that event identification is not just a technical matter since it is a political act for one country to make an allegation that another country has committed a treaty violation. The evidence in support of such an allegation can come from the IMS and IDC, as well as from the National Technical Means of member states, and/or from a subset of the thousands of seismographic stations operated around the world for purposes not directly related to monitoring for nuclear explosions.

Summary

We have described the basic steps in monitoring for nuclear explosions and have emphasized the seismic monitoring system specified by the Comprehensive Nuclear-Test-Ban Treaty of 1996.

When the treaty was being negotiated, the goal for the International Monitoring System was that it be capable of detecting and identifying treaty violations – nuclear explosive tests – at the 1 kt level and higher, if they are not evasively tested. Recognizing that a 1 kt underground nuclear explosion has a magnitude in the range about 4–4.5, if it is conducted in the way that almost all the more than 1500 prior underground nuclear explosions were carried out (i.e., well-tamped and not with intent to reduce the signals picked up by monitoring networks), the evidence from Fig. 8 is that this design capability has been significantly exceeded. For almost all of the northern hemisphere, including Eurasia and North America, capability is good down to about magnitude 3.3. This corresponds to a yield of less than 100 t (0.1 kt) for a well-tamped explosion in hard rock. Only time will tell whether this capability, combined with other monitoring assets, is deemed adequate to support entry into force of the CTBT.

Cross-References

- [Body Waves](#)
- [Earthquake, Magnitude](#)
- [Seismic Instrumentation](#)
- [Seismological Networks](#)
- [Seismology, Monitoring of CTBT](#)
- [Surface Waves](#)

Bibliography

- Bowers D, Selby ND (2009) Forensic seismology and the comprehensive Nuclear-Test-Ban Treaty. *Annu Rev Earth Planet Sci* 37:209–236
- Bowers D, Walter WR (2002) Discriminating between large mine collapses and explosions using teleseismic *P*-waves. *Pure Appl Geophys* 159:803–830
- CTBTO PrepComm (2009) Verification regime. <http://www.ctbto.org/verification-regime/>
- Dahlman O, Mykkeltveit S, Haak H (2009) Nuclear Test Ban: converting political visions to reality. Springer, Berlin
- Denny MD, Goodman MD (1990) A case study of the seismic source function: Salmon and Sterling reevaluated. *J Geophys Res* 95:19705–19723
- Der Z, Mcelfresh T, Wagner R, Burnett J (1985) Spectral characteristics of *P* waves from nuclear explosions and yield estimation. *Bull Seismol Soc Am* 75:379–390. also erratum, 75, 1222–1223
- Ford SR, Dreger DS, Walter WR (2009) Identifying isotropic events using a regional moment tensor inversion. *J Geophys Res* 114: B01306. <https://doi.org/10.1029/2008JB005743>
- Khalturin VI, Rautian TG, Richards PG (1998) The seismic signal strength of chemical explosions. *Bull Seismol Soc Am* 88:1511–1524
- Khalturin VI, Rautian TG, Richards PG, Leith WS (2005) A review of nuclear testing by the Soviet Union at Novaya Zemlya, 1955–1990. *Sci Glob Secur* 13:1–42
- Kim W-Y, Richards PG (2007) North Korean nuclear test: seismic discrimination at low yield. *EOS Trans Am Geophys Union* 88(14):157–161
- Kim W-Y, Simpson DW, Richards PG (1993) Discrimination of Earthquakes and explosions in the Eastern United States using regional high-frequency data. *Geophys Res Lett* 20:1507–1510
- Kim W-Y, Richards PG, Jo EY, Ryoo YG (2018) Identification of seismic events on and near the North Korean test site following the underground nuclear test explosion of 2017 September 3. *Seismol Rev Lett* 89:2220–2230
- Koper KD, Herrmann RB, Benz HM (2008) Overview of open seismic data from the North Korean event of 9 October 2006. *Seismol Rev Lett* 79:178–185
- Kværna T, Ringdal F (2013) Detection Capability of the Seismic Network of the International Monitoring System for the Comprehensive Nuclear-Test-Ban Treaty. *Bull of the Seismol Soc Am* 103:759–772. <https://doi.org/10.1785/0120120248>
- Murphy JR (1996) Types of seismic events and their source descriptions. In: Husebye ES, Dainty AM (eds) *Monitoring a comprehensive Nuclear Test Ban Treaty*, NATO ASI series E, vol 303. Kluwer Academic, Dordrecht, pp 225–245
- Murphy JR, Rodi W, Johnson M, Sultanov DD, Bennett TJ, Toksöz MN, Ovtchinnikov V, Barker VW, Reiter DT, Rosca AC, Shchukin Y (2005) Calibration of International Monitoring System (IMS)

- stations in Central and Eastern Asia for improved seismic event location. *Bull Seismol Soc Am* 95:1535–1560
- Myers SC, Begnaud ML, Ballard S, Pasyanos ME, Phillips WS, Ramirez AL, Antolik MS, Hutchenson KD, Dwyer JJ, Rowe CA, Wagner GS (2010) A crust and upper mantle model of Eurasia and North Africa for Pn travel time calculation. *Bull Seismol Soc Am* 100:640–656
- NAS (2002) National Academy of Sciences report, technical issues related to the comprehensive Nuclear-Test-Ban Treaty. National Academy Press, Washington, DC
- NAS (2012) National academy of sciences report, the comprehensive Nuclear Test Ban Treaty – technical issues for the United States. National Academy Press, Washington, DC
- Nuttli OW (1986) Yield estimates of Nevada Test Site explosions obtained from seismic Lg waves. *J Geophys Res* 91:2137–2151
- OTA (1988) Office of technology assessment, congress of the United States, seismic verification of nuclear testing treaties, OTA-ISC-361. U. S. Government Printing Office, Washington, DC
- Pan C-Z, Jin P, Wang H-C (2007) The applicability of P/S amplitude ratios for the discrimination of low magnitude seismic events. *Acta Seismol Sin* 20:553–561. <https://doi.org/10.1007/s11589-007-0553-6>
- Richards PG (1988) Seismic methods for verifying test ban treaties. In: Schroeder D, Hafemeister D (eds) Chapter 4 of nuclear arms technologies in the 1990s, AIP conference proceedings, vol 178. American Institute of Physics, New York, pp 54–108
- Richards PG, Kim W-Y (1997) Testing the Nuclear Test-Ban Treaty. *Nature* 389:781–782
- Ringdal F (1985) Study of magnitudes, seismicity and earthquake detectability using a global network. In: Kerr, AU (ed) *The VELA Program: A Twenty-Five Year Review of Basic Research*, published for the Defense Advanced Research Projects Agency by Executive Graphics Services, pp 611–624
- Ringdal F, Marshall PD, Alewine RW (1992) Seismic yield determination of Soviet underground nuclear explosions at the Shagan River Test Site. *Geophys J Int* 109:65–77
- Schaff DP, Kim W-Y, Richards PG, Jo E, Ryoo Y (2018a) Using waveform cross-correlation for detection, location, and identification of aftershocks of the 2017 nuclear explosion at the North Korea test site. *Seismol Res Lett* 89:2113–2119
- Schaff DP, Richards PG, Slinkard M, Heck S, Young C (2018b) Lg-wave cross correlation and epicentral double-difference location in and near China. *Bull Seismol Soc Am* 108:1326–1345
- Smith AT (1993) Discrimination of explosions from simultaneous mining blasts. *Bull Seismol Soc Am* 83:160–179
- Sykes LR (1996) Dealing with decoupled nuclear explosions under a comprehensive test ban treaty. In: Husebye ES, Dainty AM (eds) *Monitoring a comprehensive Nuclear Test Ban Treaty*, NATO ASI series E, vol 303. Kluwer, Dordrecht, pp 247–293
- Tian D, Yao J, Wen L (2018) Collapse and earthquake swarm after North Korea's 3 September 2017 nuclear test. *Geophys Res Lett* 45:3976–3983. <https://doi.org/10.1029/2018GL077649>
- Timerbaev R (2007) On the “Threshold” Test Ban Treaties of 1974–76. *Secur Index Russ J Int Secur* 13:155–162. <https://doi.org/10.1080/19934270.2007.9756509>
- Waldhauser F, Schaff DP (2008) Large-scale relocation of two decades of Northern California seismicity using cross-correlation and double-difference methods. *J Geophys Res* 113. <https://doi.org/10.1029/2007JB005479>
- Yang X, North R, Romney C, Richards PG (2003) Worldwide nuclear explosions. Chapter 84 of *International Handbook of Earthquake and Engineering Seismology*, part B. In: Lee WHK, Kanamori H, Jennings P, Kisslinger C (eds) *On behalf of the International Association of Seismology and Physics of the Earth's Interior*. Academic Press, London and San Diego, pp 1595–1599

Seismic Noise

Dhananjay Kumar^{1,2} and Imtiaz Ahmed²

¹Chevron, Houston, TX, USA

²BP, Houston, TX, USA

Definition

Seismic noise	Noise is the undesirable part of seismic data that is not signal, and signal is what fits our conceptual model.
Multiple	A seismic event that experiences more than one reflection in the subsurface.
SNR	Signal-to-noise ratio – is a measure of strength of signal compared to noise, and it is a measure of seismic data quality.
Active seismic	Seismic data recordings of artificial manmade sources such as dynamite and vibroseis in land surveys and air gun in marine surveys. Example: exploration and engineering seismology.
Passive seismic	Seismic data recordings of natural sources such as earthquake, solar waves, and ocean waves. Example: earthquake seismology.

Introduction

Seismic noise comprises all of the unwanted recorded energy that contaminates seismic data. A part of the seismic energy is considered noise if it does not fit the conceptual model of seismic signal. Seismic noise can be random or coherent. The identification of seismic noise depends on the type of data analysis and the type of data available – a part of data treated as noise in one application can be signal in another application. For example, the S-wave energy is generally considered noise in a P-wave processing project; vice versa, the P-wave energy is considered noise in an S-wave processing project.

Historically, only the traveltime information of the seismic data was used to study the subsurface. For example, the knowledge about the earth's deep interior was primarily derived from seismic traveltime information recorded during deep earthquakes. Also, only the traveltime information was used to derive the structural image of subsurface for exploration projects. As seismic technology has advanced and the appetite for understanding complicated geological features has increased, there has been a push toward technologies using the complete waveforms (amplitude and traveltime) in seismic analysis. Full waveform inversion technology (Tarantola 1986) is one such example that uses the complete waveform to estimate properties of the subsurface from the

seismic data. The success of such technologies is understandably very dependent on clean, noise-free seismic data.

It is essential that seismic data are carefully processed to derive high quality seismic images (see Yilmaz 2001 for seismic data processing). One of the major challenges in seismic data processing is to separate noise from signal or to attenuate noise. In practice, noise cannot be completely attenuated and occasionally it is not even desired to attenuate noise but to use it as signal. The objective of the noise attenuation or noise separation process in seismic data processing is to enhance signal-to-noise ratio (SNR). There have been significant progress in data processing to improve SNR; advances have been made in random noise attenuation (Yilmaz 2001) and coherent noise attenuation (see Weglein and Dragoset 2005 for multiple attenuation methods).

Recently, there have been various efforts to use seismic noise as signal, for example: (1) using multiple reflected energy (multiples) in seismic migration and inversion to image subsurface; and, (2) using very low and very high frequency passive seismic signal for reservoir monitoring. In the following sections, we write brief descriptions about the types of seismic noise, the noise attenuation techniques, and how seismic noise can be useful.

Types of Seismic Noise

There are two types of seismic noise: random noise and coherent noise. In a multichannel seismic dataset, random noise does not correlate either with the neighboring channels (i.e., no spatial correlation) or along the same channel (i.e., no temporal correlation). Coherent noise, however, is part of the data that correlates spatially and/or temporally. Random noise is easier to attenuate during seismic data processing. Coherent noise is difficult to attenuate in processing; therefore, residual coherent noise can interfere with real signal and be misinterpreted as signal. The possible sources of these seismic noises can be placed under four categories: (1) ambient sources, (2) wave propagation related noise, (3) data acquisition related noise, and (4) data processing artifacts. The severity and types of noise can differ between marine and land acquisition environment.

Ambient noise is the noise from unwanted sources like wind, swell, power line, activities on nearby road, surface facility, marine activities like ships and marine animals, and other cultural noise. Ambient noise can be present in various forms on seismic data, such as linear features, localized very high amplitude response, and mono-frequency events. Ambient noise can be random noise and coherent noise.

Wave propagation related noise includes the surface waves, multiples, and geologic noise. Seismic response for an active source survey include primary reflection event (e.g.,

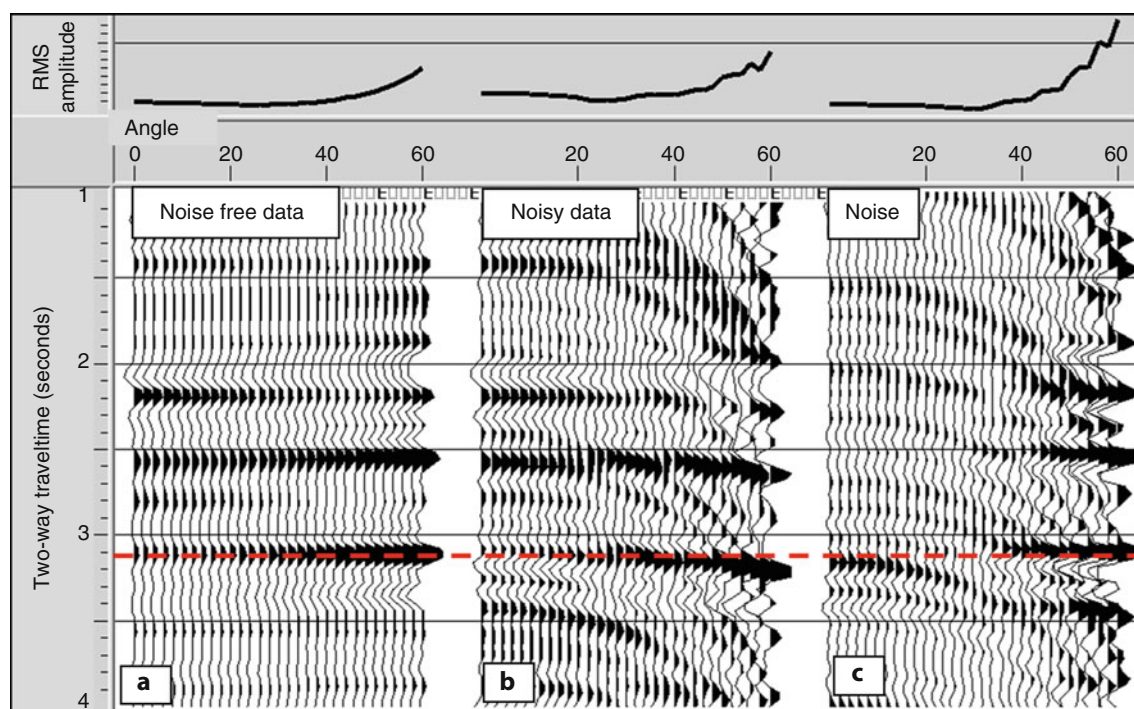
incident P-wave reflected back as P-wave), refraction, ground roll, mode converted event (e.g., incident P-wave reflected as S-wave), and several events reflected multiple times in the subsurface (multiples). In reflection seismology, we are interested in P-wave (and/or S-wave) primary reflections. All the other coherent energies recorded are considered noise. Multiples are considered the major source of coherent noise in the seismic experiment and are really difficult to attenuate. Wave propagation related noise also includes seismic response from unwanted geology (complicated stratigraphy, shallow gas, and faults) not considered in seismic analysis; they are known as geologic noise.

Data acquisition related noise is due to poor source and receiver coupling to the earth, source and recording instruments-generated noise, and acquisition footprint related to the acquisition design. Noise from poor coupling and noise related to instruments can be easily identified and attenuated. Acquisition footprint related to the acquisition design, observed as linear trend in seismic amplitude map view, is common in seismic data and can be suppressed in processing.

Seismic data processing is another source of noise in processed seismic data. There are various factors affecting seismic reflection amplitude that do not contain subsurface information (Sheriff 1975), but it is impossible to correct for all the factors affecting amplitudes. Many approximations and assumptions are made in seismic data processing depending on computer resources availability, project timeline, understanding of the physics of wave propagation, and the type of seismic data available. Some of these factors may alter the data and may introduce noise in the recorded data. Some examples are: (1) noise introduced in seismic data due to poor multiple attenuation and poor normal move-out (NMO) correction (Fig. 1); (2) noisy subsurface image due to imperfect velocity model used and/or approximate physics used for migration; (3) noise introduced due to inaccurate amplitude calibration of the raw amplitudes from the processed seismic data with synthetic seismic amplitude for quantitative AVO/AVA (amplitude variation with offset/angle) analysis; (4) artifacts introduced from the process of frequency enhancement to broaden frequency bandwidth for better depth resolution; and (5) artifacts introduced from the process of interpolation and regularization to compensate for the irregular acquisition geometry.

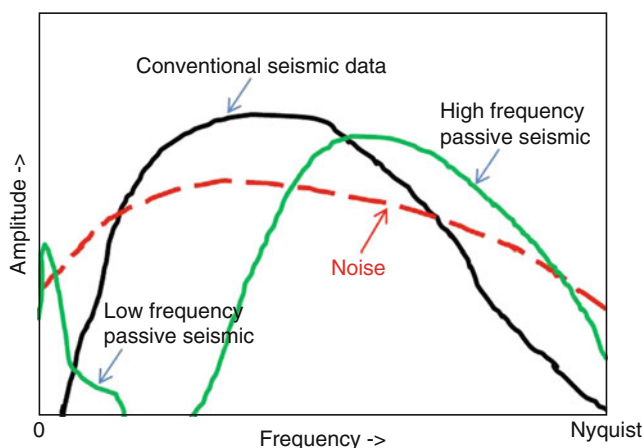
Enhancing Signal Over Noise

There is always some noise present in the seismic data. Figure 2 schematically shows amplitude spectra of various seismic signals and noise, and noise is present in the whole signal bandwidth. Thus, the objective is to enhance SNR by noise attenuation or separation so that data can be effectively



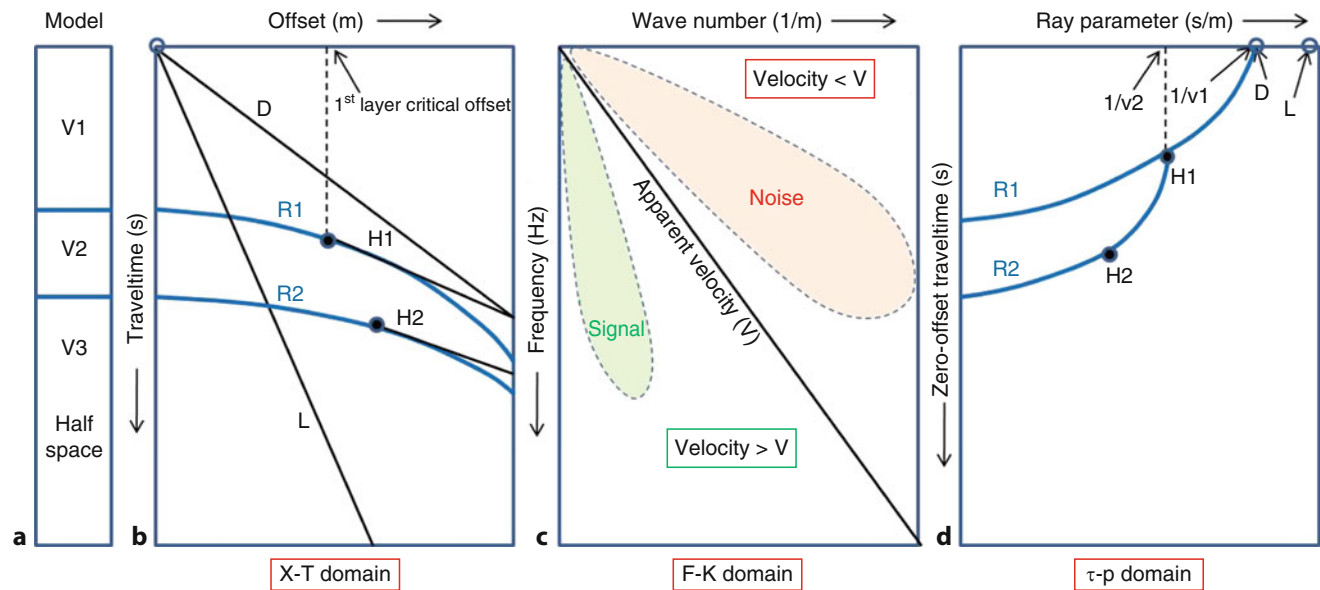
Seismic Noise, Fig. 1 Example of seismic noise. Shown in this figure is normal move-out (NMO) corrected synthetic seismic angle gather for AVA analysis with angle (in degrees) as x-axis and two-way traveltimes (in seconds) as y-axis for: noise-free case (a), with noise (b), and the noise present in data (c). The response in the third panel (c) is the difference between response in the first panel (a) and the second panel (b). On the top of angle gather is the root mean square (RMS) amplitude

value as a function of angle; the RMS value is computed in entire time gate shown here. In (c), the dipping coherent noises are multiples and the strong amplitudes at far angles are residual data after NMO. Thus, part of the data became residual data (un-flattened data after NMO in (b) is noise in AVA analysis) due to imperfect seismic velocity and/or seismic modeling method used in NMO (see red lines marked at 3,100 ms across three gathers for residual data)



Seismic Noise, Fig. 2 Schematic amplitude spectrum: for conventional broadband seismic data (active seismic for exploration range 10–60 Hz, and passive seismic for earthquake range from 10–100 Hz), low frequency passive seismic data (typically less than 10 Hz), high frequency passive seismic (microseismic) data (typically 30–300 Hz), and noise. Because noise is present at all frequencies, there is a need to do noise attenuation so that we get signal-to-noise ratio greater than 1. Spectrum is plotted for near zero frequency to Nyquist frequency. Nyquist frequency is the highest frequency without aliasing; beyond this frequency signal cannot be reliably estimated. Note that passive low and high frequency data can be treated as noise in broadband seismic data

used for analysis. An important step in noise attenuation/separation is to identify signal from noise. White (1984) discusses spectral methods for signal and noise estimation on seismic reflection data. Kanasevich (1990) and Chap. 6 of Yilmaz (2001) have good discussion on noise and multiple attenuation methods with data examples. Noise level in data compared to signal can be estimated as the ratio of autocorrelation of data to a cross correlation of data. This is because autocorrelation of data represents both signal and noise but cross correlation of data with a representative signal trace (can be local partial stack of data) represents signal in data. Signal and noise separation process can be broadly divided into two methods: prediction based methods and separation based methods. By performing seismic simulation over a known (conceptual) model and comparing synthetics to the field seismic data, signal and noise can be identified on data, and therefore noise can be attenuated. Also, signal and noise representation differs in different data domains. Therefore, a suitable data domain can be identified for the optimum separation of the two. Figure 3 schematically shows various seismic events for a three-layer model in native acquisition (X-T) domain and transformed domains: (1) X-T domain (represents spatial-time domain), (2) F-K domain (represents frequency – wave number domain), and (3) τ -p domain



Seismic Noise, Fig. 3 Seismic responses in various data domains for a three-layer model. The time domain earth model is shown in left (a) with seismic velocity for three layers as V1, V2, and V3. The second from left is the seismic response for a three-layer model in X-T domain (b), where D represents direct arrival, L represents linear noise, such as ground rolls, R1 and R2 are primary reflection events from 1st and 2nd layer interfaces, and H1 and H2 are refraction events. The third from left shows the

seismic response in F-K domain (c), where noise (D, L events) is on the upper right corner with apparent velocity less than the reference apparent velocity (V) and signal (R1, R2, H1, H2 events) are on the lower left corner corresponding to lower wave numbers. The right most plot (d) shows the seismic response in the τ -p domain, where the linear event in X-T domain plots on a point in the τ -p domain (Diebold and Stoffa 1981)

(representing zerooffset traveltime – ray parameter domain). The key to successful noise attenuation is large separation between signal and noise in a specific data domain. Different types of noise may require different domains for better separation between signal and noise. For example, a linear event on a shot gather (X-T domain) maps to a radial line in the F-K domain and it maps onto a point in the τ -p domain, and thus can be rejected by F-K filtering and τ -p filtering, respectively. Caution should be observed while performing the forward and inverse transforms, as some implementations may not be completely reversible and may also introduce artifacts.

Random noise is not correlated and can be attenuated easily during data processing. One of the most robust methods to attenuate random noise on multichannel data by a factor of square root of N is by stacking (compositing) data from N channels. If prestack data is needed, however, F-X deconvolution (predictive deconvolution) is effective in random noise attenuation, where F-X corresponds to frequency – space domain. Deconvolution operation (Yilmaz 2001) is designed in space domain to predict coherent signals, and then coherent signals can be subtracted from the seismic record.

Coherent noise is relatively more difficult to attenuate. An appropriate data domain (Fig. 3) can be selected to distinguish signal from noise using the noise characteristics, like velocity and frequency. Radon filtering in τ -p domain and wavelet transform filtering in F-K domain is very effective method for coherent noise attenuation. In many field data cases, signal

and noise are not well separated in any data domain and therefore it is very difficult to attenuate the noise. Recently a version of F-K filtering called Curvelet transform filtering has shown better results for both random and coherent noise attenuation (Neelamani et al. 2008). Multiples are example of coherent noise present in the data. There has been extensive research in the field of multiples suppression. The surface related multiple elimination (SRME) has been heavily relied upon to suppress multiples (Berkhout and Verschuur 1997). This technique is based on the stationary phase theory of the seismic signal and the multiples are predicted by convolving traces with one another and then stacking these convolved traces.

Another way to have better SNR in seismic data is via improved seismic acquisition techniques. There have been various developments in seismic acquisition on land and in marine environment. For example, over/under marine seismic acquisition promises better SNR in both low and high frequency range (Moldoveanu et al. 2007). In over/under towed-streamer acquisition pairs of streamer are deployed at two different depths in the same vertical plane. Other advances in seismic acquisition promise cleaner and better seismic images, for example, Multicomponent Ocean Bottom recording, Wide Azimuth recording, and Simultaneous Sources recording. The multiples suppression is strongly dependent on the acquisition effort. For example, the suppression of multiples in the case of Wide Azimuth acquisition is the result of a natural weighting of the traces going into the stack

because of the areal nature of the acquisition (VerWest and Lin 2007).

Some noise will always remain in the seismic data even after careful seismic processing. Also, there are some noises that are still not understandable and difficult to attenuate. Therefore, it is important to incorporate noise in the seismic data analysis – for example, in seismic inversion the data should not be over fitted as it might be fitting the noise.

Use of Noise as Signal

There have been significant efforts recently to use all of the recorded energy in the seismic data. Traditionally, only the primary reflection wavefield is used in reflection seismic imaging, while the multiple/ghost is discarded. However, the ghost/multiples can be used as signal, for example, as an input to mirror migration (Verm 1987) to produce superior shallow images for the sparse receiver ocean bottom data where illumination holes deteriorate the primary images (Clarke et al. 2006; Dash et al. 2009). Note that multiples are also generated from the same source as primary, they travel longer paths and contain more information than primary, and therefore in some circumstances multiples can be more useful than primaries. Full waveform inversion is another technique that uses both the primary and the multiples to invert for the subsurface parameters more effectively.

Multiples have also been used to interpolate for missing near offsets in seismic recording using the technique of interferometry (Schuster 2009). Seismic interferometry can also be used to extract primary signal from background noise and multiples by simply cross-correlating recorded data at two different stations (Sabra et al. 2005; Curtis et al. 2006; Draganov et al. 2007).

Some new seismic acquisition methods have been useful in using conventional multiples to improve SNR and/or extract primary. For example, over/under towed-streamer acquisition technology uses multiples to improve seismic data quality (Moldoveanu et al. 2007), and there is a possibility to estimate primaries from surface multiples from data recorded with simultaneous sources also called blended data (van Groenestijn and Verschuur 2010). In the simultaneous source acquisition (Berkhout 2008; Hampson et al. 2008) multiple sources are fired in a short time interval to speed up the acquisition.

Converted S-wave data, regarded as noise in reflected P-wave data imaging, has been quite successful in imaging gas reservoirs and areas where we have shallow gas anomalies (Tatham and McCormack 1991). The property that S-wave does not get as attenuated as P-wave when the wavefield travels through these porous medium helps

create better images through pure S-wave or converted S-wave imaging.

Low frequency earth's ambient passive response caused by natural phenomenon (such as wind, ocean waves, and human-made noise) and high frequency passive seismic response due to small earthquakes (microseisms) caused by induced fractures in a petroleum reservoirs are considered noise in a broadband active seismic data, but they can be effectively used to study subsurface. Low frequency earth's ambient noise can be used to extract signal (Draganov et al. 2007), and the low frequency passive seismic anomaly can be correlated with the presence of hydrocarbon (Saenger et al. 2009); however the research is still in its early stages. High frequency passive seismic data (also called microseismic data) are effectively used in hydraulic fracture monitoring (Warpinski 2009) by locating microseisms induced by hydrocarbon production related activities.

Summary

Seismic noise is an integral part of the seismic record and is defined as all unwanted seismic energy on data. It can be divided into two categories: random and coherent noises. Random noise is not correlated among traces and is easier to attenuate compared to coherent noise that is spatially and/or temporally correlated. Multiples and geologic noise that are coherent noise are more difficult to attenuate and often interfere with seismic signal and makes seismic analysis challenging. Strategies for seismic noise attenuation are needed to preserve the seismic signal of interest and to improve signal-to-noise ratio (SNR). The success in noise attenuation lies in identification and then separation or prediction of signal and noise. Transformation of data to different data domains (X-T, F-K, τ -p, curvelet, wavelet domains) have helped in better separating noise from signal. Wave-equation extrapolations, inverse scattering methods, surface related multiple elimination, deconvolution, etc., model the noise and/or data in the process of noise attenuation. Advances in seismic data acquisition and processing have been made to improve SNR. Recently various efforts have been made to use noise as signal and are an active topic of research. This includes using multiples as well as primaries in seismic migration and inversion, and using low and high frequency passive seismic data in imaging subsurface.

Cross-References

- [Earthquake Source Theory](#)
- [Seismic Data Acquisition and Processing](#)
- [Single and Multichannel Seismics](#)

Bibliography

- Berkhout AJ (2008) Changing the mindset in seismic acquisition. *Lead Edge* 27:924–938
- Berkhout AJ, Verschuur DJ (1997) Estimation of multiple scattering by iterative inversion, Part I: theoretical considerations. *Geophysics* 62:1586–1595
- Clarke R, Xia G, Kabir N, Sirgue L, Mitchell S (2006) Case study: a large 3D wide azimuth ocean bottom node survey in deepwater GOM. In: 76th annual international meeting, Society of Exploration Geophysicists, Expanded abstracts, pp 1128–1132
- Curtis A, Gerstoft P, Sato H, Snieder R, Wapenaar K (2006) Seismic interferometry – turning noise into signal. *Lead Edge* 25:1082–1092
- Dash R, Spence G, Hyndman R, Grion S, Wang Y, Ronen S (2009) Wide-area imaging from OBS multiples. *Geophysics* 74(4):Q41–Q47
- Diebold J, Stoffa PL (1981) The traveltime equation tau-p mapping and inversion of common midpoint data. *Geophysics* 46:238–254
- Draganov D, Wapenaar K, Mulder W, Singer J, Verdel A (2007) Retrieval of reflections from seismic background-noise measurements. *Geophys Res Lett* 34:L04305. <https://doi.org/10.1029/2006GL028735>
- Hampson G, Stefani J, Herkenhoff F (2008) Acquisition using simultaneous sources. *Lead Edge* 27:918–923
- Kanasewich ER (1990) Seismic noise attenuation. *handbook of geophysical exploration, seismic exploration*, vol 7. Pergamon, New York
- Moldoveanu N, Combee L, Egan M, Hampson G, Sudora L, Abriel W (2007) Over/under towed-streamer acquisition: a method to extend bandwidth to both higher and lower frequencies. *Lead Edge* 26:41–58
- Neelamani R, Baumstein AI, Gillard DG, Hadidi MT, Soroka W (2008) Coherent and random noise attenuation using the curvelet transformation. *Lead Edge* 27:240–248
- Sabra KG, Gerstoft P, Roux P, Kuperman WA (2005) Extracting time-domain Green's function estimates from ambient seismic noise. *Geophys Res Lett* 32:L03310. <https://doi.org/10.1029/2004GL021862>
- Saenger EH, Schmalholz SM, Lambert M-A, Nguyen TT, Torres A, Metzger S, Habiger RM, Muller T, Rentsch S, Mendez-Hernandez E (2009) A passive seismic survey over a gas field: analysis of low-frequency anomalies. *Geophysics* 74(2):029–040
- Schuster GT (2009) *Seismic interferometry*. Cambridge University Press, Cambridge, UK
- Sheriff RE (1975) Factors affecting seismic amplitudes. *Geophys Prospect* 23:125–138
- Tarantola A (1986) A strategy for nonlinear elastic inversion of seismic reflection data. *Geophysics* 51:1893–1903
- Tatham RH, McCormack MD (1991) Multicomponent seismology in petroleum exploration. Society of Exploration Geophysicists, Tulsa
- van Groenestijn GJA, Verschuur DJ (2010) Using surface multiples to estimate primaries by sparse inversion from blended data. *Geophys Prospect*. <https://doi.org/10.1111/j.1365-2478.2010.00894.x>
- Verm R (1987) Imaging VSP's 3 kilometers beyond the borehole receiver. In: Offshore technology conference proceedings, Paper 5570
- VerWest BJ, Lin D (2007) Modeling the impact of wide-azimuth acquisition on subsalt imaging. *Geophysics* 72:241–250
- Warpinski N (2009) Microseismic monitoring: inside and out. *J Pet Technol* 61:80–85
- Weglein AB, Dragoset WH (eds) (2005) Multiple attenuation, *Geophysics reprint series no. 23*. Society of Exploration Geophysicists, Tulsa
- White RE (1984) Signal and noise estimation from seismic reflection data using spectral coherence methods. *Proc IEEE* 72(10):1340–1356
- Yilmaz O (2001) *Seismic data analysis: processing, inversion, and interpretation of seismic data*. Society of Exploration Geophysicists, Tulsa

Seismic Phase Nomenclature: The IASPEI Standard

Johannes Schweitzer ¹, Dmitry A. Storchak ² and Peter Borman (Deceased)

¹NORSAR, Kjeller, Norway

²International Seismological Centre (ISC), Thatcham, UK

Definition

Different types of seismic waves can travel through the Earth. Their propagation depends on the wave type and the seismic velocities in the Earth. Due to the internal structure of the Earth (propagation velocities), the different wave types may be split at internal discontinuities (reflection, refraction, and conversion of waves) and will be recorded at seismic stations separated in time. The onsets of the different, separated waves are called seismic phases. Depending on their type and path through the Earth, seismic phases are labeled with different names, which are standardized for use by the seismological community.

Introduction

The working group (hereinafter WG) on the standard seismic phase names was set up by the IASPEI Commission on Seismological Observation and Interpretation (CoSOI) in 2001. The WG was chaired by D. A. Storchak and included R. D. Adams, P. Bormann, E. R. Engdahl, J. Havskov, B. L. N. Kennett, and J. Schweitzer. The WG put together a modified standard nomenclature of seismic phases that was meant to be concise, consistent, and self-explanatory on the basis of agreed rules. The list was not meant to satisfy specific requirements of seismologists to name various phases used in a particular type of research. Instead, it was hoped that the new list would ensure an expanded standardized data reporting and exchange by data analysts and other users. After numerous consultations with the seismological community, the standard seismic phase list was finalized and adopted by the CoSOI/IASPEI at its meeting in Sapporo on July 04, 2003.

The original list of standard seismic phase names was first published as part of the *New Manual of Seismological Observatory Practice* (Storchak et al. 2002), and then the version formally approved by the IASPEI was published in the *Seismological Research Letters* (Storchak et al. 2003). Various updates (e.g., in the revised second edition of the *New Manual of Seismological Observatory Practice* (Bormann et al. 2013)) to the list were required due to progress in observational

seismology and relevant changes in other observational standards. This entry accommodates the advances made in the nomenclature since its last publication.

The new nomenclature partially modified and complemented an earlier one published in the 1979 edition of the (Willmore 1979). It is more in tune with modern Earth and travel-time models. As opposed to former practice, the WG tried to make sure that the phase name generally reflects the type of wave and the path it has traveled. Accordingly, symbols for characterizing onset quality, polarity, etc. are no longer part of the phase name. The WG was also aware that seismic phases exist that are common in some regions but are only rarely or not found in other regions, such as Pb (P*), PnPn, PbPb, etc. The names and definitions of acoustic and hydroacoustic phases are likely to be reviewed and revised based on new analysis practices being established in the data centers.

The extended list of phase names as presented below reflects significantly increased detection capabilities of modern seismic sensors and sensor arrays, even of rather weak phases, which were rarely observed on the classical analog records. It also accounts for improved possibilities of proper phase identification by means of digital multichannel data processing such as frequency-wavenumber (f-k) analysis and polarization filtering, by modeling the observations with synthetic seismograms or by showing on the records the theoretically predicted onset times of phases. Finally, the IASPEI Seismic Format (ISF) (<http://www.isc.ac.uk/standards/isf/download/isf.pdf>) is much more flexible than the older formats previously used by the ISC, the NEIC, and other data centers. It also allows reporting, computer parsing, and archiving of phases with long or previously uncommon names. ISF also accepts complementary parameters such as onset quality, measured backazimuth and slowness, amplitudes and periods of other phases in addition to P and surface waves, for components other than vertical ones, and for instruments with nonstandard response characteristics.

This increased flexibility of the parameter-reporting format requires improved standardization, which limits the uncontrolled growth of incompatible and ambiguous parameter data. Therefore, the WG agreed on certain rules. They are outlined below prior to the listing of the standardized phase names. To facilitate the understanding of the phase names, ray diagrams are presented below. They have been calculated for local seismic sources on the basis of an average one-dimensional two-layer crustal model and for regional and teleseismic sources using the global 1D Earth model *ak135* (Kennett et al. 1995).

Before assigning abbreviated shortcut seismic phase names, one should agree first on the language to be used and its rules. As in any other language, we need a suitable alphabet (here plain Latin letters); numbers (here Arabic

numbers and $\pm$ signs); an orthography, which regulates, for example, the use of capital and lowercase letters; and a syntax, which describes the rules of correct order and mutual relationship of the language elements. One should be aware, however, that like any historically developed language, the seismological nomenclature will inevitably develop exceptions to the rules and depend on the context in which it is used. Although not fully documented below, some exceptions will be mentioned. Note that our efforts are mainly aimed at standardized names to be used in international data exchange so as to build up unique, unambiguous global databases for research. Many of the exceptions to the rules are related to specialized, mostly local research applications. The identification of related seismic phases often requires specialized procedures of data acquisition and processing that are not part of seismological routine data analysis. Also, many of these exceptional phases are rarely or never used in seismic event location, magnitude determination, source mechanism calculations, etc., which are the main tasks of international data centers. We focus, therefore, on phases that are particularly important for seismological data centers as well as for the refinement of regional and global Earth models on the basis of widely exchanged and accumulated parameter readings. In addition, we added references to the first definition of some wave types and phase names.

Standard Letters, Signs, and Syntax Used for Describing Seismic Phases

Capital Letters

Individual capital letters that stand for primary types of seismic body waves include:

P	Longitudinal wave that has traveled through the Earth's crust and mantle, from <i>undae primae</i> (Latin) = first waves (Von Dem Borne 1904)
K	Longitudinal wave that has traveled through the Earth's outer core, K, from Kern (German) = core (Sohon 1932; Bastings 1934)
I	Longitudinal wave that has traveled through the Earth's inner core (Jeffreys and Bullen 1940)
S	Transverse wave that has traveled through the Earth's crust and mantle, from <i>undae secundae</i> (Latin) = second waves (Von Dem Borne 1904)
T	Wave that has partly traveled as sound wave in the sea, from <i>undae tertiae</i> (Latin) = third waves (Linehan 1940)
J	Transverse wave that has traveled through the Earth's inner core (Bullen 1946)

Exceptions

- A capital letter *N* used in the nomenclature does not stand for a phase name but rather for the number of legs traveled

(or $N-1$ reflections made) before reaching the station. N should usually follow the phase symbol to which it applies. For examples see syntax below.

- The lowercase letters p and s may stand, in the case of seismic events below the Earth's surface, for the relatively short *upgoing leg* of P or S waves, which continue, after reflection and possible conversion at the free surface, as downgoing P or S wave. Thus seismic depth phases (e.g., pP , sP , sS , pPP , sPP , $pPKP$, etc.) are uniquely defined. The identification and reporting of such phases are of utmost importance for source depth determination (Schrade 1931; Stechschulte 1932; Gutenberg et al. 1933; Macelwane et al. 1933).
- Many researchers working on detailed investigations of crustal and upper mantle discontinuities denote both the up- and downgoing short legs of converted or multiple-reflected P and S phases as lowercase letters p and s , respectively.

Individual or double capital letters that stand for surface waves include:

L	(Relatively) long-period surface wave, unspecified, from <i>undae longae</i> (Latin) = long waves (Von Dem Borne 1904)
R	Rayleigh waves (short- to very long-period waves in crust and upper mantle) (Angenheister 1921)
Q	Love waves, from <i>Querwellen</i> (German) = transverse waves (Angenheister 1921)
G	(Very long-period) global (mantle) Love waves, firstly observed and reported by Gutenberg and Richter (1934); in honor of Gutenberg, Byerly proposed the usage of G for these waves (Richter 1958)
LR	Long-period Rayleigh waves, usually relating to the Airy phase maximum in the surface wave train
LQ	Long-period love waves
PL	Fundamental leaking mode following P onsets, firstly observed and reported by Somville (1930, 1931)

Lowercase Letters and Signs

Single lowercase letters generally specify the part of Earth's crust or upper mantle in which a phase has its turning point or at which discontinuity it has been reflected and/or eventually converted:

g	Following the phase name characterizes waves "bottoming" (i.e., having their turning point in the case of P or S body waves) or just travel (surface waves) within the upper ("granitic") Earth's crust (e.g., Pg , Sg , Rg), (Jeffreys 1926)
b	Following the phase name characterizes body waves bottoming in the lower ("basaltic") Earth's crust (Jeffreys 1926) (e.g., Pb , Sb ; alternative names for these phases are P^* , S^* (Conrad 1925)); also used for phases radiated directly to the Earth's surface from events in the lower crust
n	Following the phase name characterizes a P or S wave that is bottoming in the Earth's uppermost mantle or traveling as head wave below the Mohorovičić discontinuity (e.g., Pn , Sn),

(continued)

	introduced after Andrija Mohorovičić discovered the Earth's crust and separated the crustal travel-time curve from the normal ($=n$) mantle phase (Mohorovičić 1910); also used for phases radiated directly to the Earth's surface from events in the uppermost mantle
m	(Upward) reflections from the outer side of the Mohorovičić (Moho) discontinuity (e.g., PmP , SmS)
c	Reflections from the outer side of the core-mantle boundary (CMB), usage proposed by James B. Macelwane (see Gutenberg 1925)
i	Reflections from the outer side of the inner core boundary (ICB)
z	Reflections from a discontinuity (other than free surface, CMB or ICB) at depth z (measured in km). Upward reflections from the outer side of the discontinuity may additionally be complemented by a $+$ sign (e.g., $P410+P$; this, however, is not compulsory), while downward reflections from the inner side of the discontinuity must be complemented by a $-$ sign (e.g., $P660-P$)

An exception from these rules is the use of lowercase p or s to indicate arrivals of longitudinal or transverse waves that were first radiated to go up toward the free surface to be reflected/converted back into the Earth as normal P or S waves (see near-source surface reflections and conversions section of the phase list below).

Naming phases radiated directly to the Earth's surface in the near-source region from events in the lower crust (Conrad discontinuity) as Pb and from below the Moho as Pn is a pragmatic decision, which helps analysts and source-location programs to separate different branches of the travel-time curves.

Double lowercase letters following a capital letter phase name indicate the travel-time branch to which this phase belongs. Due to the geometry and velocity structure of the Earth, the same type of seismic wave may develop a triplication of its travel-time curve with different, in some cases, well-separated, branches. Thus, it is customary to differentiate between different branches of core phases and their multiple reflections at the free surface or the CMB. Examples are $PKPab$, $PKPbc$, $PKPdf$, $SKSac$, $SKKSac$, etc. The separation of the different PKP branches with letters ab , bc , and df was introduced by Jeffreys and Bullen (1940).

Three lowercase letters may follow a capital letter phase name to specify its character, e.g., as a forerunner (pre) to the main phase, caused by scattering (e.g., $PKPpre$) or as a diffracted wave extending the travel-time branch of the main phase into the outer core shadow (e.g., $Pdif$ in the outer core shadow for P).

Syntax of Generating Complex Phase Names

Due to refraction, reflection, and conversion in the Earth, most phases have a complex path history before they reach the station. Accordingly, most phases cannot be described by a single capital letter code in a self-explanatory way. By

combining the capital and lowercase letters as mentioned above, one can describe the character of even rather complex refracted, reflected, or converted phases. The order of symbols (syntax) regulates the sequence of phase legs due to refraction, reflection, and conversion events in time (from left to right) and in space.

Examples for Creating Complex Standard Phase Names

Traditional examples of complex phase names are as follows.

Refracted and Converted Refracted Wave

- PKP is a pure refracted longitudinal wave. It has traveled the first part of its path as P through the crust and mantle, the second through the outer core, and the third again as P through the mantle and crust. An alternative name for PKP is P' (Angenheister 1921), which should be read as “P prime.”
- PKIKP (alternative to PKPdf) is also a pure refracted longitudinal wave. It has traveled the first part of its path as P through the crust and mantle, the second through the outer core, the third through the inner core, and the fourth and fifth parts back again through the outer core and mantle/crust.
- SKS is a converted refracted wave. It has traveled as a shear wave through the crust and mantle, being converted into a longitudinal P wave when refracted into the outer core and converted back again into an S wave when entering the mantle.
- SKP or PKS are converted refracted waves with only one conversion from S to P when entering the core or from P to S when leaving the core, respectively.

Pure Reflected Waves

- In the case of (downward only) reflections at the free surface or from the inner side of the CMB, the phase symbol is just repeated, e.g., PP, SS (Geiger 1909), PPP, SSS, KK, KKK, etc.
- In the case of (upward) reflections from the outer side of the Moho, the CMB, or the ICB, this is indicated by inserting symbols m, c, or i, respectively, between the phase symbols, e.g., PmP, PcP, ScS, and PKiKP.
- Reflections from any other discontinuity in the mantle or crust at depth z may be from the inner side (–; i.e., downward back into the mantle) or from the outer side (+; i.e., back toward the surface). To differentiate between these two possibilities, the sign has to follow z (or the

respective number in km), for example, P410+P or P660–P.

- To abbreviate names of multi-leg phases due to repeated reflections, one can also write Phasename N. This type of abbreviation is customary in the case of multiple phases with long phase names such as PmP2 for PmPPmP (free-surface reflection of PmP), SKS2 for SKSSKS (the alternative name for $S'2$, the free-surface reflection of SKS), PKP3 for PKPPKPPKP (double free-surface reflection of PKP; alternative name to $P'3$), or P4KP for PKKKKP (triple reflection of P at the inner side of the CMB).

Two additional notes are to be made. First, PKP2 = PKPPKP are now alternative names for $P'2$ or $P'P'$, respectively. This should not be mistaken for the old usage of PKP2 for PKPab. Secondly, in the case of multiple reflections from the inner side of the CMB, the WG followed the established tradition of placing the number N not after but in front of the related phase symbol K.

Reflected Waves with Conversion at the Reflection Point

In the case that a phase changes its character from P to S, or vice versa, one writes:

- PS (first leg P, second leg S) or SP (first leg P, second leg S) in the case of reflection/conversion from the free surface downward into the mantle (Geiger and Gutenberg 1912a, b).
- PmS or SmP, respectively, for reflection/conversion from the outer side of the Moho.
- PcS or ScP for reflection/conversion from the outer side of the CMB.
- Pz+S or Sz–P for reflection/conversion from the outer (+) side or inner (–) side, respectively, of a discontinuity at depth z . Note that the – is compulsory, the + is not.
- pS or sP reflection/conversion at the free surface for body waves with a direct upgoing first leg.

In this context, it is worth mentioning that mode conversion is impossible for reflections from the inner side of the CMB back into the outer core because the liquid outer core does not allow the propagation of S waves.

The WG determined the new IASPEI standard phase names along these lines and rules. Where these deviate from other traditionally used names, the latter are given as well. Either the traditional names are still acceptable alternatives (alt) or they are old names (old), which should no longer be used.

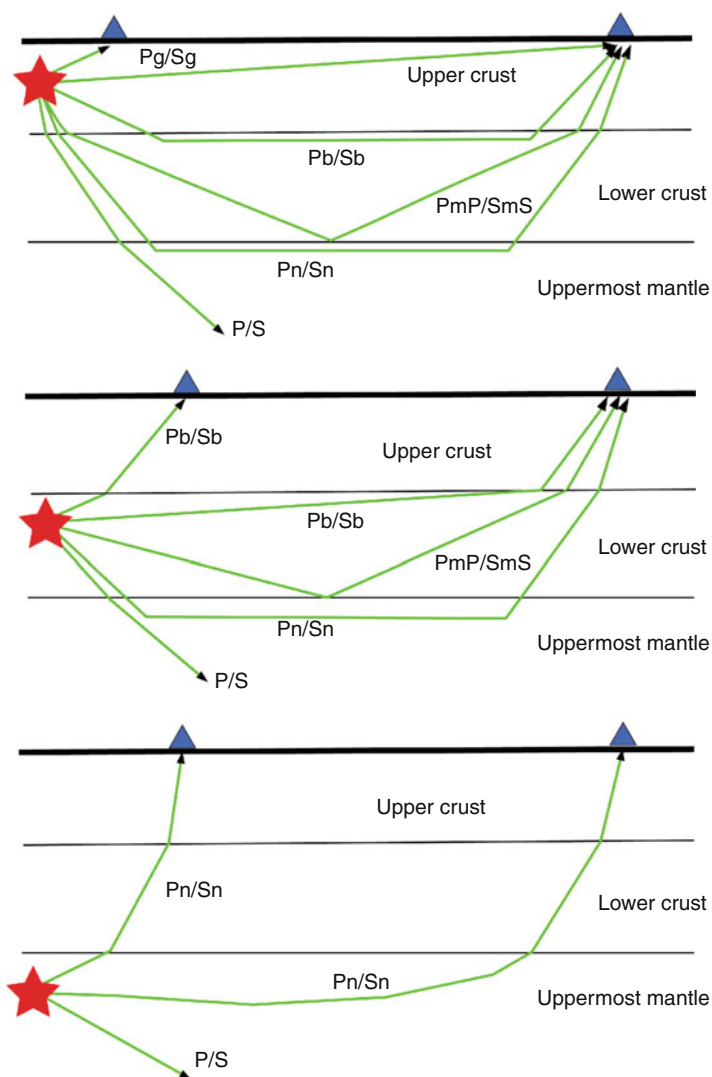
Ray-Path Diagrams for Some of the IASPEI Standard Phases

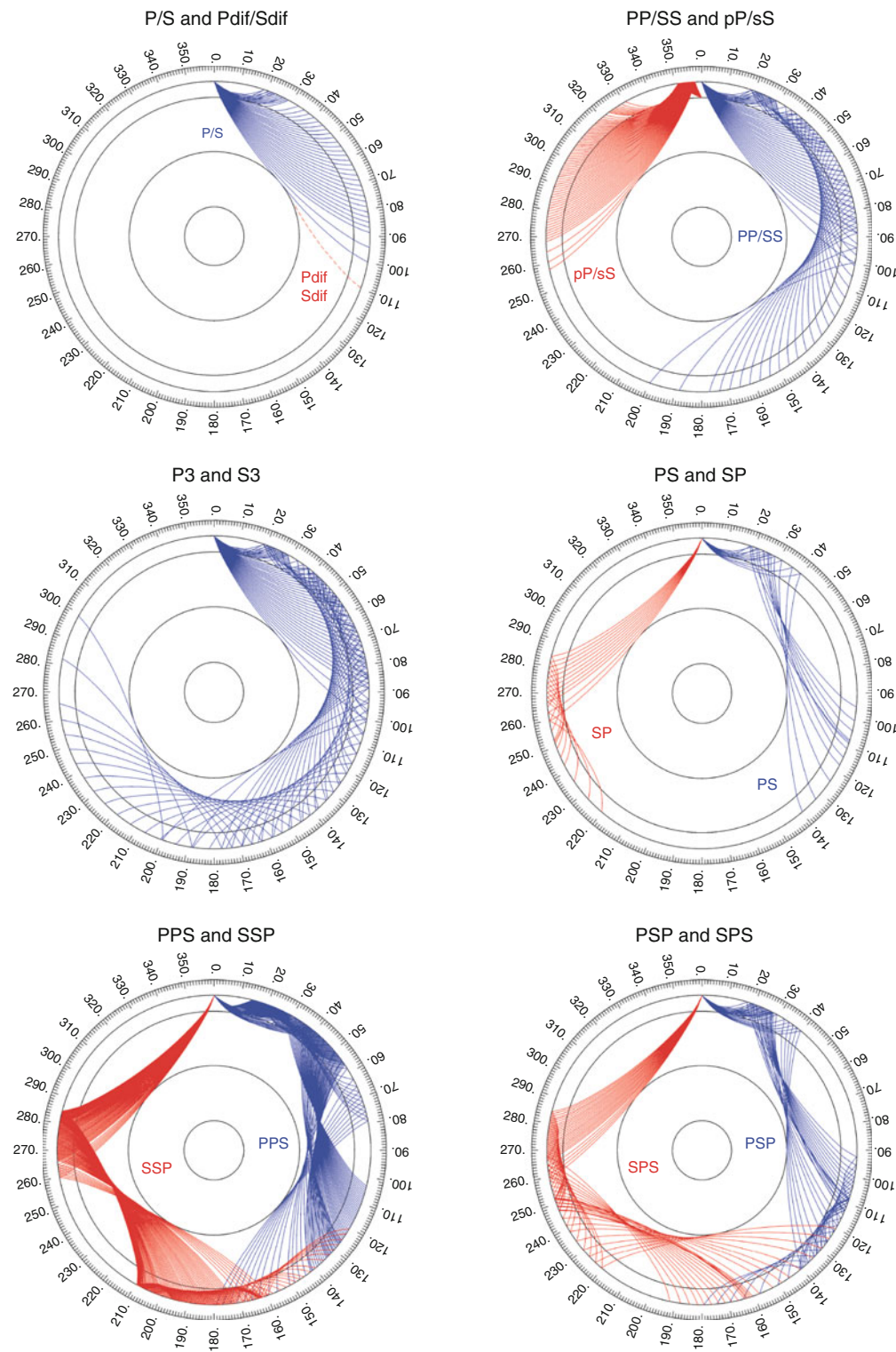
We show ray paths through the Earth for many of the mentioned phases. The three diagrams for crustal phases are sketches illustrating the principal ray paths in a two-layer crust (Fig. 1). The rays in all other figures (Figs. 2, 3, 4, 5, and 6) were calculated by using the ray picture part of the WKB3 code (Chapman 1978; Dey-Sarkar and Chapman 1978); as velocity model, we chose the standard Earth model *ak135* (Kennett et al. 1995). For some types of P and S phases, the ray paths through the Earth are very similar

because the velocity ratio V_P/V_S does not change enough to give very different ray pictures. In these cases, we calculated only the ray paths for the P-type ray (i.e., P, Pdif, pP, PP, P660P, P660–P, PcP, PcP2, and PcP4) and assumed that the corresponding ray paths of the respective S-type phases were very similar. To show the different ray paths for phases with similar phase names, we show on many figures rays leaving the source once to the left and once to the right in different colors. The three most important discontinuities inside the Earth are indicated as black circles (i.e., the border between the upper and lower mantle, the CMB, and the ICB).

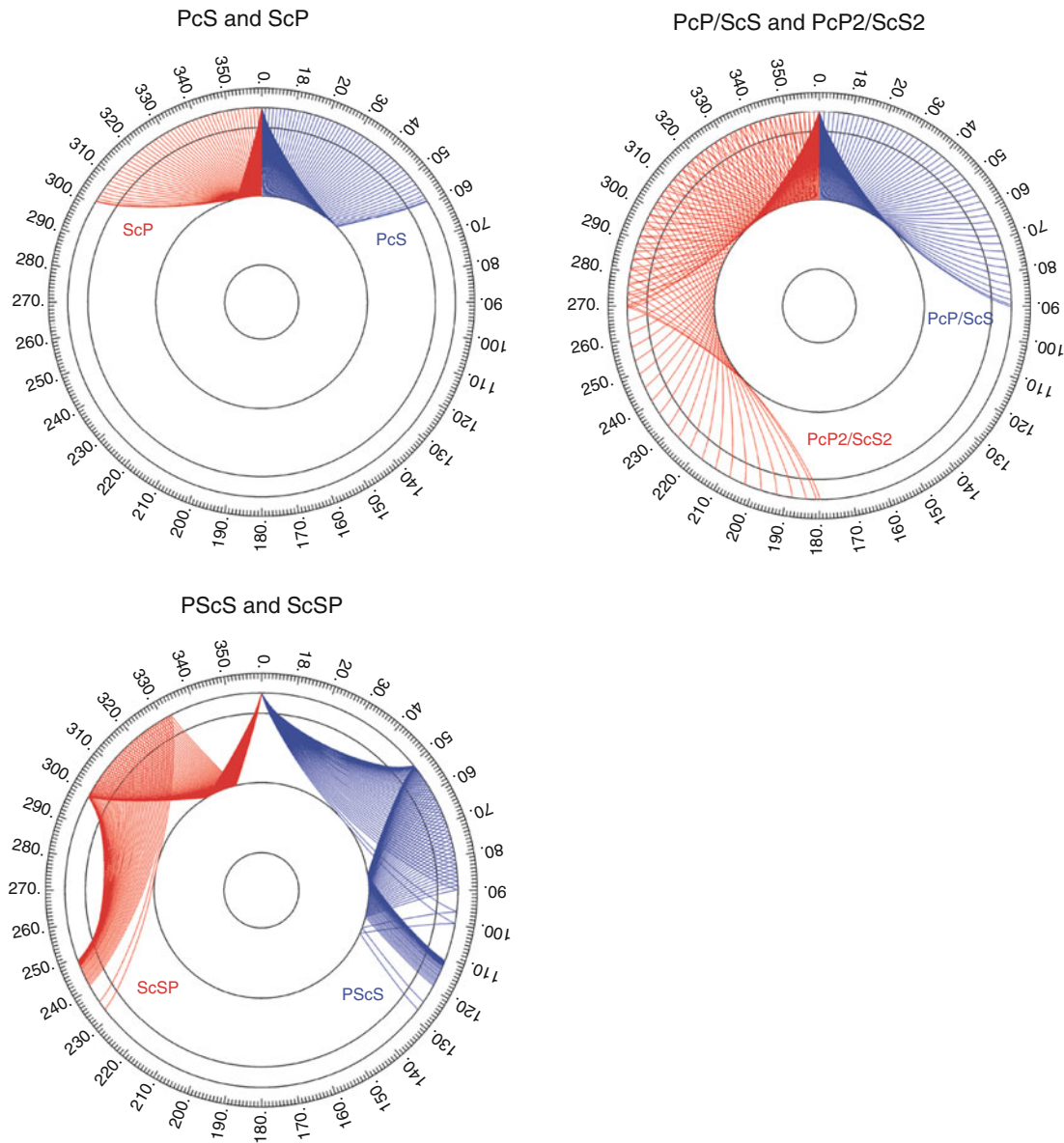
Seismic Phase Nomenclature: The IASPEI Standard,

Fig. 1 Seismic “crustal phases” observed in the case of a two-layer crust in local and regional distance ranges ($0^\circ < D < \text{approximately } 20^\circ$) from the seismic source in the (a) upper crust, (b) lower crust, and (c) uppermost mantle





Seismic Phase Nomenclature: The IASPEI Standard, Fig. 2 Mantle phases observed at teleseismic distance ranges ($D > \text{approximately } 20^\circ$)



Seismic Phase Nomenclature: The IASPEI Standard, Fig. 3 Mantle phases reflected from the Earth's core

IASPEI Standard Seismic Phase List

Crustal Phases

Pg	At short distances, either an upgoing P wave from a source in the upper crust or a P wave bottoming in the upper crust; also, at larger distances, arrivals caused by multiple P-wave reverberations inside the whole crust with a group velocity around 5.8 km/s
Pb	(alt:P*) Either an upgoing P wave from a source in the lower crust or a P wave bottoming in the lower crust
Pn	Any P wave bottoming in the uppermost mantle or an upgoing P wave from a source in the uppermost mantle
PnPn	Pn free-surface reflection

(continued)

PgPg	Pg free-surface reflection
PmP	P reflection from the outer side of the Moho
PmPN	PmP multiple free-surface reflection; <i>N</i> is a positive integer. For example, PmP2 is PmPPmP
PmS	P to S reflection/conversion from the outer side of the Moho
Sg	At short distances, either an upgoing S wave from a source in the upper crust or an S wave bottoming in the upper crust; also, at larger distances, arrivals caused by superposition of multiple S-wave reverberations and SV to P and/or P to SV conversions inside the whole crust
Sb	(alt:S*) Either an upgoing S wave from a source in the lower crust or an S wave bottoming in the lower crust
Sn	Any S wave bottoming in the uppermost mantle or an upgoing S wave from a source in the uppermost mantle
SnSn	Sn free-surface reflection

(continued)

SgSg	Sg free-surface reflection
SmS	S reflection from the outer side of the Moho
SmSN	SmS multiple free-surface reflection; N is a positive integer. For example, SmS2 is SmSSmS
SmP	S to P reflection/conversion from the outer side of the Moho
Lg	A wave group observed at larger regional distances and caused by superposition of multiple S-wave reverberations and SV to P and/or P to SV conversions inside the whole crust. The maximum energy travels with a group velocity of approximately 3.5 km/s
Rg	Short-period crustal Rayleigh wave

Mantle Phases

P	A longitudinal wave, bottoming below the uppermost mantle; also an upgoing longitudinal wave from a source below the uppermost mantle
PP	Free-surface reflection of a P wave leaving a source downward
PS	P, leaving a source downward, reflected as an S at the free surface. At shorter distances, the first leg is represented by a crustal P wave
PPP	Analogous to PP
PPS	PP which is converted to S at the second reflection point on the free surface; travel time matches that of PSP
PSS	PS reflected at the free surface
PcP	P reflection from the core-mantle boundary (CMB)
PcS	P converted to S when reflected from the CMB
PcPN	PcP reflected from the free surface $N - 1$ times; N is a positive integer. For example, PcP2 is PcPPcP
Pz+P	(alt:PzP) P reflection from the outer side of a discontinuity at depth z ; z may be a positive numerical value in km. For example, P660+P is a P reflection from the top of the 660 km discontinuity
Pz-P	P reflection from the inner side of a discontinuity at depth z . For example, P660-P is a P reflection from below the 660 km discontinuity, which means it is precursory to PP
Pz+S	(alt:PzS) P converted to S when reflected from the outer side of discontinuity at depth z
Pz-S	P converted to S when reflected from the inner side of a discontinuity at depth z
PScS	P (leaving a source downward) to ScS reflection at the free surface
Pdif	(old:Pdiff) P diffracted along the CMB in the mantle
S	Shear wave, bottoming below the uppermost mantle; also an upgoing shear wave from a source below the uppermost mantle
SS	Free-surface reflection of an S wave leaving a source downward
SP	S, leaving a source downward, reflected as P at the free surface. At shorter distances, the second leg is represented by a crustal P wave
SSS	Analogous to SS
SSP	SS converted to P when reflected from the free surface; travel time matches that of SPS
SPP	SP reflected at the free surface
ScS	S reflection from the CMB

(continued)

ScP	S converted to P when reflected from the CMB
ScSN	ScS multiple free-surface reflection; N is a positive integer. For example, ScS2 is ScSScS
Sz+S	(alt:SzS) S reflection from the outer side of a discontinuity at depth z ; z may be a positive numerical value in km. For example, S660+S is an S reflection from the top of the 660 km discontinuity
Sz-S	S reflection from the inner side of discontinuity at depth z . For example, S660-S is an S reflection from below the 660 km discontinuity, which means it is precursory to SS
Sz+P	(alt:SzP) S converted to P when reflected from the outer side of a discontinuity at depth z
Sz-P	S converted to P when reflected from the inner side of a discontinuity at depth z
ScSP	ScS to P reflection at the free surface
Sdif	(old:Sdiff) S diffracted along the CMB in the mantle

Core Phases

PKP	(alt:P') unspecified P wave bottoming in the core
PKPab	(old:PKP2) P wave bottoming in the upper outer core; ab indicates the retrograde branch of the PKP caustic
PKPbc	(old:PKP1) P wave bottoming in the lower outer core; bc indicates the prograde branch of the PKP caustic
PKPdf	(alt:PKIKP) P wave bottoming in the inner core
PKPpre	(old:PKhKP) a precursor to PKPdf due to scattering near or at the CMB
PKPdf	P wave diffracted at the inner core boundary (ICB) in the outer core
PKS	Unspecified P wave bottoming in the core and converting to S at the CMB
PKSab	PKS bottoming in the upper outer core
PKSbc	PKS bottoming in the lower outer core
PKSdf	PKS bottoming in the inner core
P'P'	(alt:PKPPKP) Free-surface reflection of PKP
P'N	(alt:PKPN) PKP reflected at the free surface $N - 1$ times; N is a positive integer. For example, P'3 is P'P'P'
P'z-P'	PKP reflected from the inner side of a discontinuity at depth z outside the core, which means it is precursory to P'P'; z may be a positive numerical value in km
P'S'	(alt:PKPSKS) PKP converted to SKS when reflected from the free surface; other examples are P'PKS and P'SKP
PS'	(alt:PSKS) P (leaving a source downward) to SKS reflection at the free surface
PKKP	Unspecified P wave reflected once from the inner side of the CMB
PKKPab	PKKP bottoming in the upper outer core
PKKPbc	PKKP bottoming in the lower outer core
PKKPdf	PKKP bottoming in the inner core
PNKP	P wave reflected $N - 1$ times from the inner side of the CMB; N is a positive integer
PKKPpre	A precursor to PKKPdf due to scattering near the CMB
PKiKP	P wave reflected from the inner core boundary (ICB)
PKMIKP	P wave reflected $N - 1$ times from the inner side of the ICB
PKJKP	P wave traversing the outer core as P and the inner core as S

(continued)

PKKS	P wave reflected once from the inner side of the CMB and converted to S at the CMB
PKKSab	PKKS bottoming in the upper outer core
PKKSbc	PKKS bottoming in the lower outer core
PKKSdf	PKKS bottoming in the inner core
PcPP'	(alt:PcPPKP) PcP to PKP reflection at the free surface; other examples are PcPS', PcSP, PcSS', PcPSKP, and PcSSKP
SKS	(alt:S') unspecified S wave traversing the core as P
SKSac	SKS bottoming in the outer core
SKSdf	(alt:SKIKS) SKS bottoming in the inner core
SPdifKS	(alt:SKPdifS) SKS wave with a segment of mantle-side Pdif at the source and/or the receiver side of the ray path
SKP	Unspecified S wave traversing the core and then the mantle as P
SKPab	SKP bottoming in the upper outer core
SKPbc	SKP bottoming in the lower outer core
SKPdf	SKP bottoming in the inner core
S'S'	(alt:SKSSKS) Free-surface reflection of SKS
S'N	SKS reflected at the free surface $N - 1$ times; N is a positive integer
S'z-S'	SKS reflected from the inner side of discontinuity at depth z outside the core, which means it is precursory to S'S'; z may be a positive numerical value in km
S'P'	(alt:SKSPKP) SKS converted to PKP when reflected from the free surface; other examples are S'SKP and S'PKS
S'P	(alt:SKSP) SKS to P reflection at the free surface
SKKS	Unspecified S wave, traversing the core as P with one reflection from the inner side of the CMB
SKKSac	SKKS bottoming in the outer core
SKKSdf	SKKS bottoming in the inner core
SNKS	S wave reflected $N - 1$ times from the inner side of the CMB; N is a positive integer
SKiKS	S wave traversing the outer core as P and reflected from the ICB
SKJKS	S wave traversing the outer core as P and the inner core as S
SKKP	S wave traversing the core as P with one reflection from the inner side of the CMB and then continuing as P in the mantle
SKKPab	SKKP bottoming in the upper outer core
SKKPbc	SKKP bottoming in the lower outer core
SKKPdf	SKKP bottoming in the inner core
ScSS'	(alt:ScSSKS) ScS to SKS reflection at the free surface; other examples are ScPS', ScSP', ScPP', ScSSKP, and ScPSKP

Near-Source Surface Reflections and Conversions (Depth Phases)

pPy	All P-type onsets (Py) as defined above, which resulted from the reflection of an upgoing P wave at the free surface or an ocean bottom. WARNING: The character "y" is only a wild
-----	--

(continued)

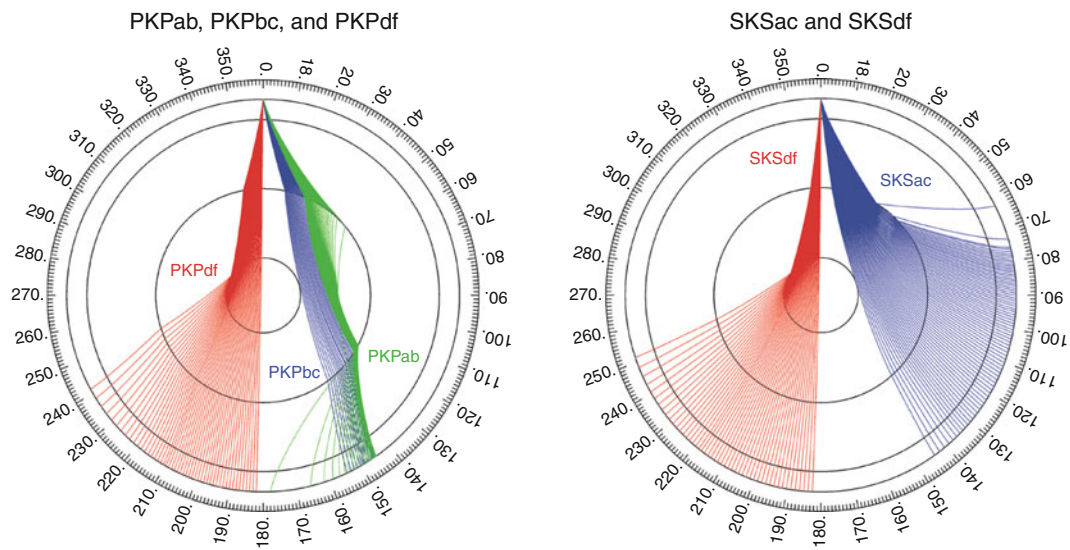
	card for any seismic phase, which could be generated at the free surface. Examples are pP, pPKP, pPP, pPcP, etc.
sPy	All Py resulting from the reflection of an upgoing S wave at the free surface or an ocean bottom, for example, sP, sPKP, sPP, sPcP, etc.
pSy	All S-type onsets (Sy) as defined above, which resulted from the reflection of an upgoing P wave at the free surface or an ocean bottom, for example, pS, pSKS, pSS, pScP, etc.
sSy	All Sy resulting from the reflection of an upgoing S wave at the free surface or an ocean bottom, for example, sSn, sSS, sScS, sSdif, etc.
pwPy	All Py resulting from the reflection of an upgoing P wave at the ocean's free surface
pmPy	All Py resulting from the reflection of an upgoing P wave from the inner side of the Moho

Surface Waves

L	Unspecified long-period surface wave
LQ	Love wave
LR	Rayleigh wave
G	Mantle wave of Love type
GN	Mantle wave of Love type; N is an integer and indicates wave packets traveling along the minor arcs (odd numbers) or major arc (even numbers) of the great circle
R	Mantle wave of Rayleigh type
RN	Mantle wave of Rayleigh type; N is an integer and indicates wave packets traveling along the minor arcs (odd numbers) or major arc (even numbers) of the great circle
PL	Fundamental leaking mode following P onsets generated by coupling of P energy into the waveguide formed by the crust and upper mantle
SPL	S wave coupling into the PL waveguide; other examples are SSPL and SSSPL

Acoustic Phases

H	A hydroacoustic wave from a source in the water, which couples in the ground
HPg	H phase converted to Pg at the receiver side
HSg	H phase converted to Sg at the receiver side
HRg	H phase converted to Rg at the receiver side
I	An atmospheric sound arrival, which couples in the ground
IPg	I phase converted to Pg at the receiver side
ISg	I phase converted to Sg at the receiver side
IRg	I phase converted to Rg at the receiver side
T	A tertiary wave. This is an acoustic wave from a source in the solid Earth, usually trapped in a low velocity oceanic water layer called the SOFAR channel (SOund Fixing And Ranging)
TPg	T phase converted to Pg at the receiver side
TSg	T phase converted to Sg at the receiver side
TRg	T phase converted to Rg at the receiver side



Seismic Phase Nomenclature: The IASPEI Standard, Fig. 4 Seismic rays of direct core phases

Amplitude Measurements

The following set of amplitude measurement names refers to the IASPEI Magnitude Standard of 2013 (see http://www.iaspei.org/commissions/commission-on-seismological-observation-and-interpretation/Summary_WG_recommendations_20130327.pdf), compliance to which is indicated by the presence of leading letter I. The absence of leading letter I indicates that a measurement is nonstandard. Letter A indicates a measurement in nm made on a displacement seismogram, whereas letter V indicates a measurement in nm/s made on a velocity seismogram. All amplitude measurements for standard magnitudes, with the exception for M_L , are made on vertical-component seismic records.

IAML	Displacement amplitude measured according to the IASPEI standard for local magnitude M_L
IAMs_20	Displacement amplitude measured according to the IASPEI standard for surface-wave magnitude $M_S(20)$
IVMs_BB	Velocity amplitude measured according to the IASPEI standard for broadband surface-wave magnitude $M_S(BB)$
IAmb	Displacement amplitude measured according to the IASPEI standard for short-period teleseismic body-wave magnitude m_b
IVmB_BB	Velocity amplitude measured according to the IASPEI standard for broadband teleseismic body-wave magnitude $m_B(BB)$
AX_IN	Displacement amplitude of phase of type X (e.g., PP, S, etc.), measured on an instrument of type IN (e.g., SP, short-period; LP, long-period; BB, broadband)
VX_IN	Velocity amplitude of phase of type X and instrument of type IN (as above)

(continued)

A	Unspecified displacement amplitude measurement
V	Unspecified velocity amplitude measurement
AML	Displacement amplitude measurement for nonstandard local magnitude
AMs	Displacement amplitude measurement for nonstandard surface-wave magnitude
Amb	Displacement amplitude measurement for nonstandard short-period body-wave magnitude
AmB	Displacement amplitude measurement for nonstandard medium to long-period body-wave magnitude
END	Time of visible end of record for duration magnitude

Unidentified Arrivals

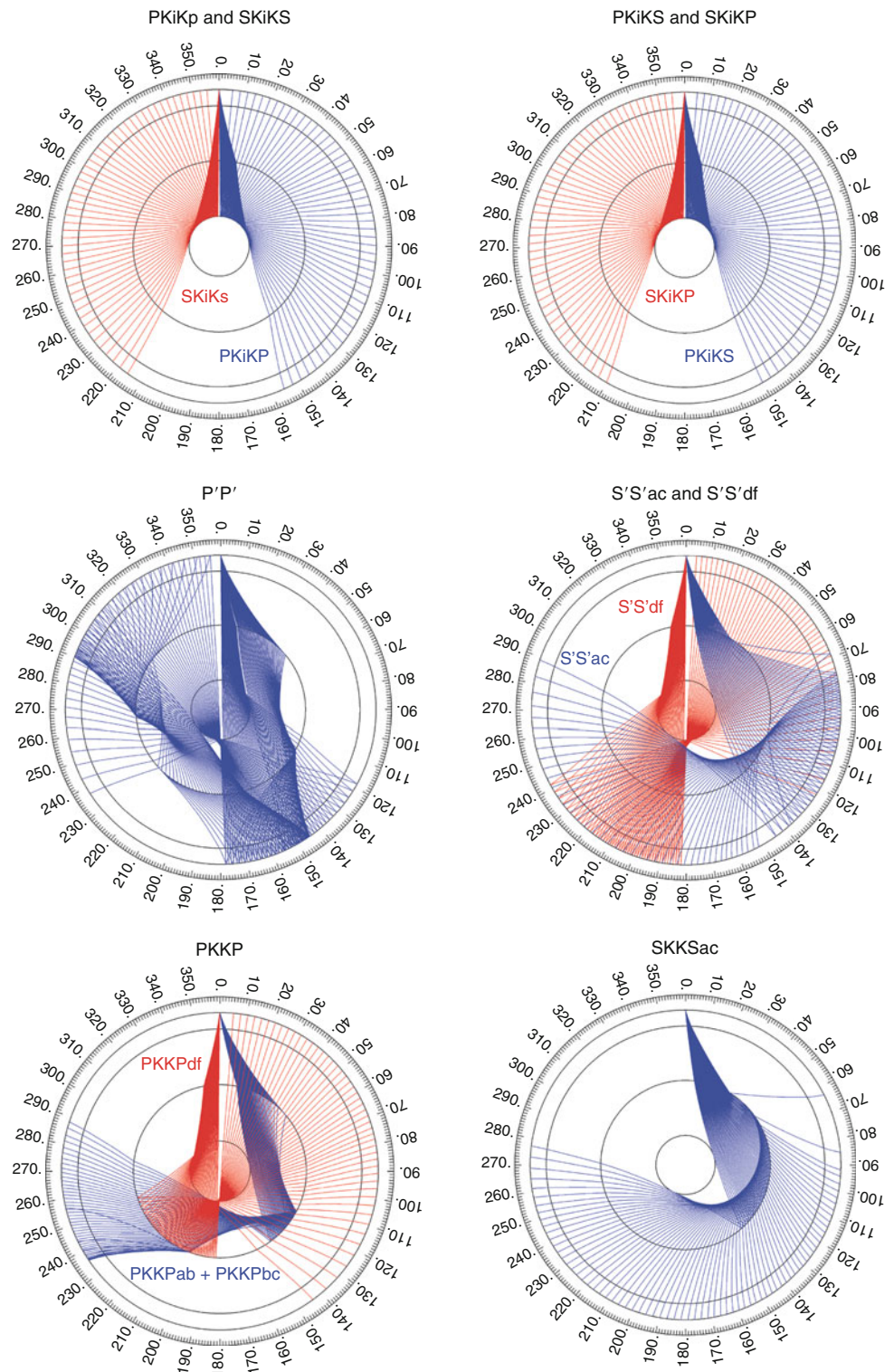
x	(old:i, e, NULL) unidentified arrival
rx	(old:i, e, NULL) unidentified regional arrival
tx	(old:i, e, NULL) unidentified teleseismic arrival
Px	(old:i, e, NULL, (P), P?) unidentified arrival of P-type
Sx	(old:i, e, NULL, (S), S?) unidentified arrival of S-type

Summary

Since the early days of instrumental seismology, seismograms showed distinguished onsets of different seismic phases with different characteristics. Seismologists developed an own nomenclature for these phases, which is basic for any seismogram interpretation. This entry includes the historical development and currently established definitions of the seismic phase nomenclature as recommended by IASPEI.

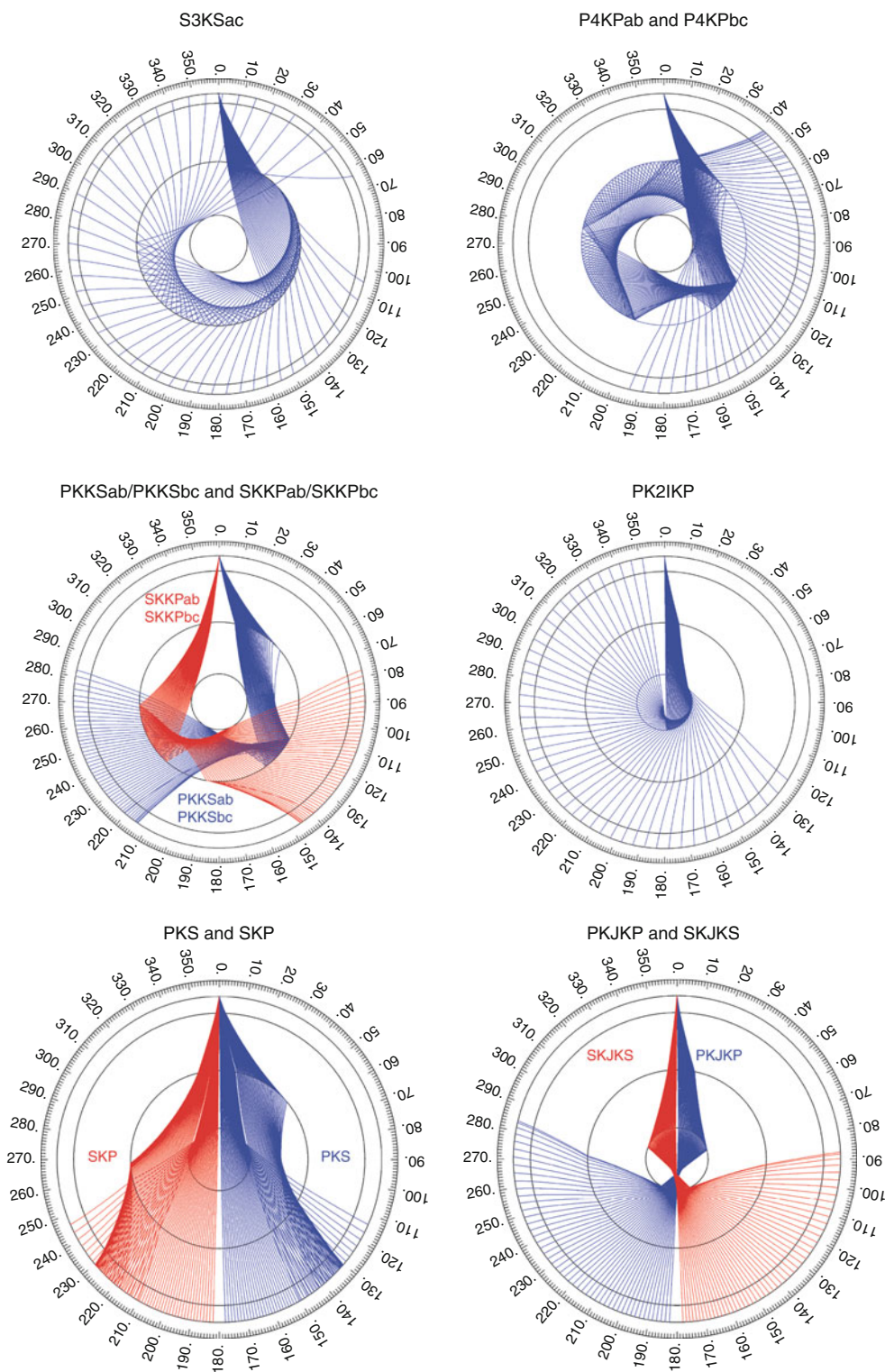
**Seismic Phase Nomenclature:
The IASPEI Standard,**

Fig. 5 Seismic rays of single-
reflected core phases



**Seismic Phase Nomenclature:
The IASPEI Standard,**

Fig. 6 Seismic rays of multiple-
reflected and converted core
phases



Cross-References

- [Body Waves](#)
- [Earth's Structure, Core](#)
- [Seismogram Interpretation](#)
- [Surface Waves](#)

Bibliography

- Angenheister GH (1921) Beobachtungen an pazifischen Beben. Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse 113–146
- Bastings L (1934) Shear waves through the Earth's core. *Nature* 134:216–217
- Bormann P, Storchak DA, Schweitzer J (2013) The IASPEI standard nomenclature of seismic phases. In: Bormann P (eds) *New manual of seismological observatory practice*, 2nd edn. GeoForschungsZentrum, Potsdam, IS2.1, pp 1–20. https://doi.org/10.2312/GFZ.NMSOP-2_IS_2.1
- Bullen KE (1946) A hypothesis on compressibility at pressures of the order of a million atmospheres. *Nature* 157:405
- Chapman CH (1978) A new method for computing synthetic seismograms. *Geophys J R Astron Soc* 54:481–518
- Conrad V (1925) Laufzeitkurven des Tauernbebens vom 28. November, 1923. *Mitteilungen der Erdbeben-Kommission der Akademie der Wissenschaften in Wien, Neue Folge* 59:23
- Dey-Sarkar SK, Chapman CH (1978) A simple method for the computation of body wave seismograms. *Bull Seismol Soc Am* 68:1577–1593
- Geiger L (1909) Seismische Registrierungen in Göttingen im Jahre 1907 mit einem Vorwort über die Bearbeitung der Erdbebendiagramme. Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse 107–151
- Geiger L, Gutenberg B (1912a) Konstitution des Erdinnern, erschlossen aus der Intensität longitudinaler und transversaler Erdbebenwellen, und einige Beobachtungen an den Vorläufern. *Phys Z* 13:115–118
- Geiger L, Gutenberg B (1912b) Ueber Erdbebenwellen. VI. Konstitution des Erdinnern, erschlossen aus der Intensität longitudinaler und transversaler Erdbebenwellen, und einige Beobachtungen an den Vorläufern. Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse 623–675
- Gutenberg B (1925) Bearbeitung von Aufzeichnungen einiger Weltbeben. *Abh Senckenb Naturforsch Ges* 40:57–88
- Gutenberg B, Richter CF (1934) On seismic waves (first paper). *Gerlands Beitr Geophys* 43:56–133
- Gutenberg B, Wood HO, Richter CF (1933) Re suggestion by Dr. Harold Jeffreys regarding $\bar{P}$ and Pg. *Gerlands Beitr Geophys* 40:97–98
- Jeffreys H (1926) On near earthquakes. *Mon Not R Astron Soc Geophys Suppl* 1:385–402
- Jeffreys H, Bullen KE (1940) (reprinted 1948; 1958; 1967; 1970). *Seismological Tables*. British Association for the Advancement of Science, Gray Milne Trust, London, p 50
- Kennett BLN, Engdahl ER, Buland R (1995) Constraints on seismic velocities in the Earth from traveltimes. *Geophys J Int* 122:108–124
- Linehan D (1940) Earthquakes in the West Indian region. *Trans Am Geophys Union* 30:229–232
- Macelwane JB, Brunner GJ, Joliat JS (1933) Re suggestion by Doctor Harold Jeffreys and others regarding $\bar{P}$ and Pg. *Gerlands Beitr Geophys* 40:98
- Mohorovičić A (1910) Potres od 8. X 1909, God. Izvjestie Zagr. met. Ops. Zag. 1909, Zagreb. (Das Beben vom 8. X 1909, Jahrbuch des meteorologischen Observatoriums in Zagreb für das Jahr 1909 9(4):1–63)
- Richter CF (1958) *Elementary seismology*. W. H. Freeman, San Francisco/London, p 768
- Scrase FJ (1931) The reflected waves from deep focus earthquakes. *Proc R Soc Lond Ser A* 132:213–235
- Sohon FW (1932) *Seismometry*. Part II of Macelwane JB, Sohon FW. *Introduction to theoretical seismology*. Wiley, New York, 149 pp
- Somville O (1930) A propos d'une onde longue dans la première phase de quelques séismogrammes. *Gerlands Beitr Geophys* 27:437–442
- Somville O (1931) A propos d'une onde longue dans la première phase de quelques séismogrammes (II). *Gerlands Beitr Geophys* 29:247–251
- Stechschulte VC (1932) The Japanese earthquake of March 29, 1928. *Bull Seismol Soc Am* 22:81–137
- Storchak DA, Bormann P, Schweitzer J (2002) Standard nomenclature of seismic phases. In: Bormann P (eds) *New manual of seismological observatory practice*, vol 2. GeoForschungsZentrum, Potsdam, IS2.1, pp 1–18
- Storchak DA, Schweitzer J, Bormann P (2003) The IASPEI standard seismic phase list. *Seismol Res Lett* 74:761–772
- Von Dem Borne G (1904) Seismische Registrierungen in Göttingen, Juli bis Dezember 1903. Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse:440–464
- Willmore PL (1979) *Manual of seismological observatory practice*, World Data Center A for Solid Earth Geophysics, Report SE-20, September 1979, Boulder, Colorado, 165 pp

Seismic Properties of Rocks

Nikolas I. Christensen

Department of Earth and Ocean Sciences,

University of British Columbia, Vancouver, BC, Canada

Synonyms

Rock P and S velocities

Definition

- | | |
|------------------------------|--|
| Compressional (P) waves | Seismic waves in which the rock particles vibrate parallel to the direction of wave propagation. |
| Shear (S) waves | Seismic waves in which the rock particles vibrate perpendicular to the direction of wave propagation. |
| Poisson's ratio (σ) | The ratio of the lateral unit strain to the longitudinal unit strain in a body that has been stressed longitudinally within its elastic limit ($2\sigma = (R^2 - 2)/(R^2 - 1)$ where $R = V_p/V_s$). |
| Transversely isotropic | Anisotropic solids with a single symmetry axis. Rock symmetry axes are usually normal to foliation, cleavage, or layering. |

Introduction

Although many disciplines have contributed significantly to our knowledge of the Earth's interior, none has a resolution comparable to seismology. For nearly seven decades, seismic studies have provided geophysicists with worldwide information on crustal and upper mantle compressional (P) and shear (S) wave velocities. Significant data have recently become available on velocity gradients, velocity reversals, compressional and shear wave velocity ratios, and anisotropy in the form of azimuthal variations of compressional wave velocities, as well as shear wave splitting. Reflections within the crust and mantle originate from contrasts of acoustic impedances, defined as products of velocity and density. The interpretation of this seismic data requires detailed knowledge of rock velocities provided by laboratory techniques to a precision at least comparable with that of seismic measurements. In particular, to infer composition of the Earth's upper 30–50 km, the “crust,” requires studies of the elasticity of rocks at conditions approaching those that exist at these depths. Of fundamental importance is the presence of mean compressive stress and temperature increasing with depth and on the average reaching about 1 GPa and 500 °C at the base of the crust. Because of this, the most relevant velocity measurements for identifying probable rock types within the crust have been measurements at elevated pressures and temperatures. These measurements often allow the seismologist to infer mineralogy, porosity, the nature of fluids occupying pore spaces, temperature at depth, and present or paleolithospheric stress based on mineral and crack orientations.

Francis Birch (Fig. 1) was the pioneer in the study of rock velocities. In addition to his laboratory work on physical properties of rocks and minerals at high pressures and temperatures, he was well-known for his studies of heat flow and theoretical work on the composition of the Earth's interior. Two of his benchmark papers on compressional wave velocities in rocks (Birch 1960, 1961) set the stage for modern experimental studies of rock elasticity and have been frequently cited during the past five decades. These papers for the first time provided information on compressional wave velocities for many common rock types, as well as major findings on their anisotropies and relations to density. It is interesting to note that these measurements were carried out to pressures of 1 GPa, a pressure at which even today only a limited number of laboratories generate for modern rock seismic velocity measurements.

Measurement Techniques

Rock velocities are usually measured in the laboratory using the pulse transmission technique. The transit time of either a



Seismic Properties of Rocks, Fig. 1 Francis Birch (1903–1992), a pioneer in rock physics research

compressional or shear wave is measured along the axis of a cylindrical rock specimen of known length. The cores are usually taken from rock samples using a 2.54 cm inner diameter diamond coring bit. The cores are trimmed and ground flat and parallel on a diamond grinding disk. The volume of each core is obtained from the length and diameter. The cores are weighed, and densities are calculated from their masses and dimensions. The cores are then fitted with a copper jacket to prevent penetration of high-pressure oil into the rock samples. For measurements at high temperatures, where gas is the pressure medium, the samples are usually encased in stainless steel. Transducers are placed on the ends of the rock core (Fig. 2). Compressional and shear waves are often generated by means of lead zirconate titanate (PZT) and AC cut quartz transducers with resonant frequencies of 1 MHz. The sending transducer converts the input, an electrical pulse of 50–500 V and 0.1–10 μ s width, to a mechanical signal, which is transmitted through the rock. The receiving transducer changes the wave to an electrical pulse, which is amplified and displayed on an oscilloscope screen (Fig. 3). Once the system is calibrated for time delays, the travel time through the specimen is determined directly by a computer or with the use of a mercury delay line. The major advantage of the delay line is that it increases the precision, especially for signals with slow rise times, because the gradual onset of the first arrival from the sample is approximated by the delay line. The velocity is the ratio of the length of the specimen to the travel time of the compressional or shear wave. The total error limits for V_p and V_s are estimated to be less than 1%.

Interfacing the pressure system with a computer for data acquisition and storage permits automatic calculations of velocities as successive readings are taken. Using a least-squares routine, the computer fits a curve to the data points and calculates velocities for selected pressures. A velocity versus pressure curve is plotted along with recorded data points. Sample length, density, measured pressure velocity pairs, traces of the waveforms at selected pressures, the curve fit equations, and calculated pressure velocity pairs are recorded and stored digitally.

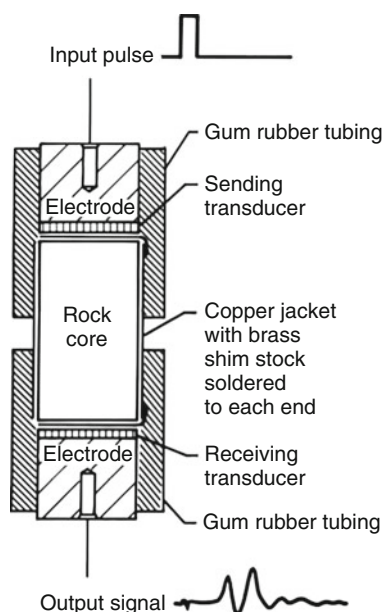
Hydrostatic pressure generating systems capable of producing true hydrostatic pressures as high as 3 GPa, equivalent to a depth of approximately 100 km, have been used for rock velocity measurements. Low-viscosity synthetic petroleum and argon for high-temperature measurements are frequently used as pressure media. An alternate technique for obtaining velocities under quasi-hydrostatic conditions has used a

triaxial press with cubic samples. Transducers are placed on the six pistons, and corrections are made for travel times through the pistons. Rock velocities obtained using this technique have provided valuable information on the effect of temperature on velocity but have been limited to pressures of 0.6 GPa.

The behavior of a rock's velocity as a function of pressure is primarily dependent upon mineralogy and porosity. Many igneous and metamorphic rocks have porosities of the order of a few tenths of 1%, which are present as thin openings between grain boundaries. As pressure is applied to the rock, the cracks close and velocities increase. Once the cracks close, any increase in velocity with increasing pressure is related to the intrinsic effects of pressure on the mineral velocities. This is illustrated in Fig. 4 for a garnet granulite. Velocities first increase rapidly over the first 100 MPa as cracks close and then increase slowly as pressure is increased. Also the velocity determined at a given pressure depends upon whether the pressure is approached from lower or higher pressure (Fig. 4). This hysteresis is usually quite small if sufficient time is taken for measurements between pressure increments.

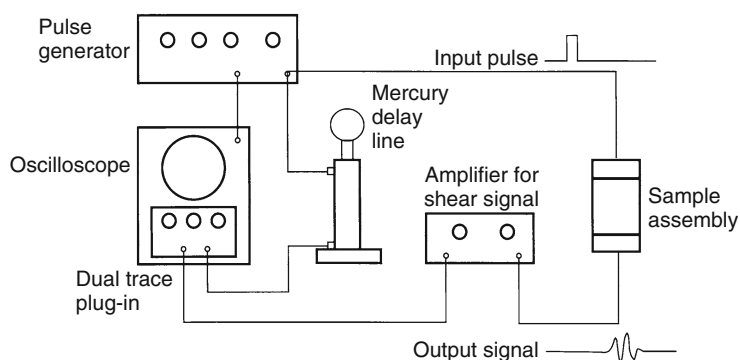
A considerable number of investigations have also focused on the influence of temperature on rock velocities. These studies have used either resonance techniques or more frequently the pulse transmission method. Early studies demonstrated that the application of temperature to rock at atmospheric pressure results in the creation of cracks that often permanently damage the rock and dramatically lower velocities. Thus reliable measurements of the temperature derivatives of velocities are obtained only at confining pressures high enough to prevent crack formation. At elevated confining pressures, $\delta V_p / \delta T$ for common rocks often ranges from -0.3×10^{-3} to -0.6×10^{-3} km/s/°C, and $\delta V_s / \delta T$ varies between -0.2×10^{-3} and -0.4×10^{-3} km/s/°C.

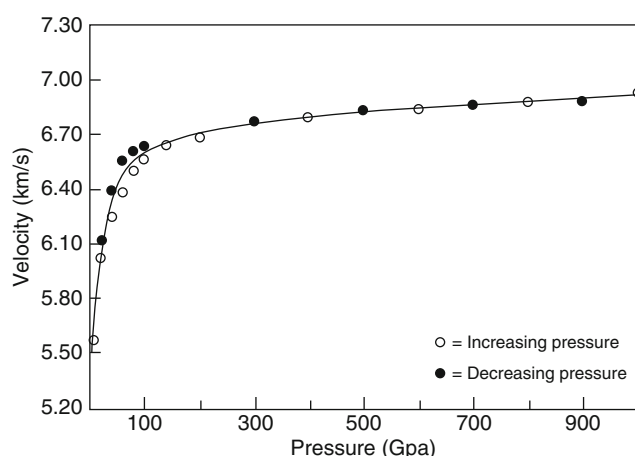
At high temperatures, usually between 650 °C and 1200 °C, rocks undergo partial melting. Several factors influence partial melting, including rock composition, pressure, and the involvement of aqueous and carbonic fluids. For rocks in which velocities are not modified by temperature-induced phase changes, it has been observed



Seismic Properties of Rocks, Fig. 2 Transducer and rock core assembly for velocity measurements at elevated pressures

Seismic Properties of Rocks, Fig. 3 Electronics for velocity measurements using a mercury delay line





Seismic Properties of Rocks, Fig. 4 Compressional wave velocity measurements as a function of confining pressure for a mafic granulite

that as temperature increases, velocities decrease linearly below partial melting temperatures and then drop significantly at the onset of partial melting.

Rock Velocities

Seismic velocities have been measured for practically all igneous and metamorphic rock types believed to be important constituents of the lithosphere. Because rock classification schemes allow for considerable variations in mineralogy for a given rock type, many rocks have wide ranges in elastic properties. However, some lithologies, such as the monomineralic rocks hornblende and dunite with little or no alteration, have fairly well-defined velocities. For detailed summaries of rock seismic properties, the reader is referred to the compilations of Birch (1960), Christensen (1982, 1996), Gerbade (1982), Ji et al. (2002), Rudnick and Fountain (1995), and Mavko et al. (1998).

Table 1 contains average velocities, in the order of increasing compressional wave velocity, for several common igneous and metamorphic rocks. Volcanic rocks usually have lower velocities than their plutonic equivalents. This is due to the presence of glass, abundant alteration products, and vesicles in volcanic rocks, all of which have lower velocities. In general, for a given composition, velocity increases with increasing metamorphic grade. For example, mica- and quartz-bearing schists have higher velocities than slates and phyllites. Low-grade metamorphosed basalts have lower velocities than higher-grade amphibolite and mafic granulite. Eclogites have the highest velocity of mafic rocks. Note that shear velocities are relatively high in quartzites and low in serpentinites.

Early attempts to infer crustal composition by comparing laboratory- and field-derived velocities relied primarily on compressional wave velocities. However correlations

Seismic Properties of Rocks, Table 1 Average compressional (V_p) and shear (V_s) wave velocities, velocity ratios (V_p/V_s), and Poisson's ratios (σ) at 1 GPa for common rock types (Christensen 1996)

Rock	V_p (km/s)	V_s (km/s)	V_p/V_s	σ
Serpentinite	5.607	2.606	2.152	0.36
Andesite	5.940	3.177	1.870	0.30
Quartzite	6.091	4.054	1.502	0.10
Basalt	6.118	3.291	1.859	0.30
Granitic gneiss	6.271	3.627	1.729	0.25
Granite-granodiorite	6.372	3.726	1.710	0.24
Tonalite gneiss	6.366	3.636	1.751	0.26
Slate	6.379	3.432	1.858	0.30
Phyllite	6.398	3.608	1.774	0.27
Mica quartz schist	6.523	3.654	1.785	0.27
Zeolite facies basalt	6.530	3.493	1.869	0.30
Diorite	6.675	3.756	1.777	0.27
Diabase	6.814	3.766	1.809	0.28
Greenschist facies basalt	6.983	3.955	1.766	0.26
Marble	6.985	3.794	1.841	0.29
Mafic granulite	7.000	3.849	1.818	0.28
Amphibolite	7.046	3.987	1.767	0.26
Anorthosite	7.124	3.717	1.917	0.31
Gabbro	7.299	3.929	1.858	0.30
Pyroxenite	7.935	4.519	1.756	0.26
Eclogite	8.198	4.594	1.785	0.27
Dunite	8.399	4.783	1.756	0.26

between compressional wave velocity and composition are limited due to the similar velocities of many common crustal rock types. Because of this nonuniqueness of compressional wave velocity laboratory and field data comparisons, many recent studies have focused on investigations of crustal composition using both compressional and shear wave velocities. In these studies the ratio V_p/V_s or Poisson's ratio (σ) calculated from V_p/V_s have resolved some of the ambiguities.

The values of V_p/V_s and σ , assuming isotropic elasticity, are given in Table 1 for several common igneous and metamorphic rocks at 1 GPa. This high pressure eliminates cracks so the values only reflect mineralogy. The relatively low σ for quartzites (0.10) agrees well with isotropic aggregate calculations based on the elastic constants of single crystal alpha quartz. Anorthosites, on the other hand, have relatively high Poisson's ratios (<0.31). As expected, values for the granites and granitic gneisses, consisting primarily of quartz and feldspar, fall in between those of quartzites and anorthosites and are relatively low. Thus, crustal regions, where field measured values of $\sigma \leq 0.25$ are observed, are likely quartz-rich. Serpentinities containing lizardite, the variety of serpentine stable at crustal PT conditions, have extremely high values of Poisson's ratio (0.36), whereas unaltered dunites and peridotites have Poisson's ratios in the range of 0.25–0.26. Partially serpentinized dunites and peridotites have Poisson's ratios that fall between these limiting values. Laboratory measurements have established

a well-defined relationship between Poisson's ratio, percent serpentinization, and density. Changes in Poisson's ratio with progressive metamorphism of mafic igneous and pelitic rocks are considerably more complicated than the above examples (Christensen 1996).

Velocity Anisotropy

Most crustal and upper mantle rocks show some degree of velocity anisotropy, which can originate from several processes. Laminar flow within magma and lava will orient elongate crystals such as feldspars along flow directions. Tabular sediments may settle preferentially, and anisotropy may be enhanced by sediment compaction. Plastic flow and recrystallization during metamorphism often produce strong mineral orientations parallel to foliation and banding. In shallow crustal rocks, oriented cracks producing anisotropy often originate from differential principle stresses. Anisotropy observed in laboratory measurements at pressures above approximately 100 MPa, where cracks are closed, often originates from preferred orientations of highly anisotropic minerals such as micas, amphiboles, pyroxenes, and olivine.

In general, for a given propagation direction in an anisotropic rock, there are three waves, one compressional, and two shear. Their vibration directions form an orthogonal set, which usually are not parallel or perpendicular to their propagation direction. Compressional and shear wave velocities vary with propagation direction, and two shear waves travel in a given direction through the rock with different velocities. This latter property of anisotropic rocks, termed shear wave splitting, was first recognized in laboratory studies and has been observed by field studies in several crustal and upper mantle regions.

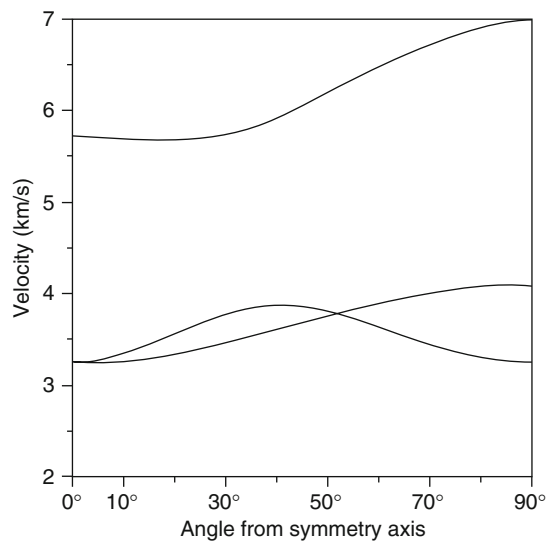
Anisotropy is usually studied in the laboratory by taking cores from a rock in different directions. Following the early investigations of Birch (1960, 1961), it is common practice to take three mutually perpendicular cores for velocity measurements. If the rock has a planar structure, such as a foliation, cleavage, or bedding, two cores are oriented with their axes within the planar structure. One of these cores is oriented parallel to a lineation, if present. In general, three velocities are measured per core: the compressional wave velocity, the velocity of the shear wave vibrating parallel to the layering, and the velocity of the shear wave vibrating in a plane perpendicular to layering. For cores taken perpendicular to layering, shear wave velocities are measured with vibration directions parallel to the axes of the cores taken in the layering.

It has been demonstrated that the above procedure provides information on maximum compressional wave anisotropy and maximum shear wave splitting. In general, the highest compressional wave velocities propagate in the

plane of the foliation and parallel to lineations. Maximum shear wave splitting is observed for propagation within the planar structures. At near-normal incidence, there is often minimal shear wave splitting.

Velocity measurements using multiple cores have provided detailed information on anisotropic velocities for non-axial propagation (e.g., Johnston and Christensen 1995; Christensen and Okaya 2007). The number of measurements necessary to completely describe wave propagation depends on the symmetry of the rock fabric. In the most general anisotropic elastic solid, 21 independent constants are required to describe the equations of motion. Examples of materials with this type of behavior are minerals possessing triclinic symmetry such as kyanite and some feldspars. For many rocks, the existence of symmetry elements in the elastic properties leads to the vanishing of some elastic constants along with simple algebraic relations between others (e.g., Auld 1990). Some crustal and upper mantle metamorphic rocks behave as elastic solids with orthorhombic symmetry, which require nine independent constants to describe the elastic tensors. Shales and many low- to medium-grade metamorphic rocks often have well-developed bedding or foliation and behave as transversely isotropic elastic solids (hexagonal symmetry with the symmetry axis normal to bedding or foliation). Transversely isotropic solids have five independent elastic constants, which can be calculated from five independent velocity measurements (two compressional wave velocities, one quasi-compressional wave velocity, and two shear wave velocities) and density. In an isotropic solid, only two independent constants are required for a complete description of elastic behavior, and wave velocities are independent of propagation direction.

To describe three-dimensional wave propagation in anisotropic rocks, phase velocity surfaces can be calculated using the Kelvin-Christoffel equations (e.g., Auld 1990), and elastic constants can be determined from velocity and density measurements. For transversely isotropic rocks, these surfaces describe variations in velocity as a function of angle to the bedding or foliation (Fig. 5). Three velocity surfaces are calculated, one for the quasi-compressional wave, one for the shear wave vibrating parallel to the planar structure, and one for the quasi-shear wave vibrating in a plane perpendicular to the foliation or bedding. For propagation parallel and perpendicular to the foliation, all wave modes are pure. The velocities shown in Fig. 5 were calculated from velocity measurements at 600 MPa for a transversely isotropic quartz-mica schist from South Island, New Zealand. They show several important features about elastic wave propagation in this rock, which are typical of many foliated rocks. First, compressional wave velocities do not increase significantly until propagation directions greater than about 45° from foliation normal are reached. At larger angles, compressional wave velocity increases rapidly and reaches a maximum for



Seismic Properties of Rocks, Fig. 5 Compressional wave anisotropy (*upper curve*) and shear wave splitting (*lower curves*) for a transversely isotropic quartz-mica schist

propagation parallel to the foliation. Shear wave singularities (directions in which two shear waves have equal velocities) occur for propagation parallel to and at approximately 50° to the symmetry axis. Shear wave splitting occurs for all other propagation directions and reaches to maximum of 90° from the normal to the foliation.

Summary

Beginning with the compressional wave velocity measurements of Birch (1960, 1961), much has been learned about elastic wave propagation in rocks. Compressional wave velocities are now available for most common rock types at pressures existing in the continental crust and uppermost mantle. Additional important contributions include laboratory measurements of shear wave velocities, velocity ratios, and the influence of temperature on velocities. Estimates of crustal composition from compressional wave velocities are nonunique, but this ambiguity is often reduced by complimentary shear wave velocity (and Poisson's ratio) observations. Recent field studies have found that seismic anisotropy is an important crustal feature. Thus systematic laboratory studies of compressional wave anisotropy and shear wave splitting will be critical in understanding crustal composition and deformation, just as they have been in investigations of the upper mantle.

Cross-References

- [Magnetic Anisotropy](#)
- [Seismic Imaging, Overview](#)

- [Seismic, Reflectivity Method](#)
- [Traveltime Tomography Using Controlled-Source Seismic Data](#)
- [Vertical Seismic Profiling](#)

Bibliography

- Auld BA (1990) *Acoustic fields and waves in solids*, vol 1. Robert E. Krieger, Malabar
- Birch F (1960) The velocity of compressional waves in rocks to 10 kilobars, 1. *J Geophys Res* 65:1083–1102
- Birch F (1961) The velocity of compressional waves in rocks to 10 kilobars, 2. *J Geophys Res* 66:2199–2224
- Christensen NI (1982) Seismic velocities. In: Carmichael RS (ed) *Handbook of physical properties of rocks*, 2nd edn. CRC, Boca Raton, pp 1–228
- Christensen NI (1996) Poisson's ratio and crustal seismology. *J Geophys Res* 101:3139–3156
- Christensen NI, Okaya DA (2007) Compressional and shear wave velocities in South Island, New Zealand rocks and their application to the interpretation of seismological models of the New Zealand crust. In: Okaya D, Stern T, Davey F (eds) *A continental plate boundary: tectonics at South Island, New Zealand*. American Geophysical Union geophysical monograph, vol 175. American Geophysical Union, Washington, DC, pp 125–155
- Gerbande H (1982) Elastic wave velocities and constants of elasticity of rocks and rock forming minerals. In: Angenheister G (ed) *Physical properties of rocks*. Springer, Landolt-Bornstein, pp 1–140
- Ji SJ, Wang Q, Xia B (2002) *Handbook of seismic properties of minerals, rocks and ores*. Presses inter Polytechnique, Montreal, pp 1–630
- Johnston JE, Christensen NI (1995) Seismic anisotropy of shales. *J Geophys Res* 100:5991–6003
- Mavko G, Mukerji T, Dvorkin J (1998) *The rock physics handbook*. Cambridge University Press, Cambridge, MA, pp 289–303
- Rudnick RL, Fountain DM (1995) Nature and composition of the continental crust; a lower crustal perspective. *Rev Geophys* 33:267–309

Seismic Quiescence and Activation

Gennady Sobolev

Institute of Physics of the Earth, Russian Academy of Sciences, Moscow, Russia

Definition

Seismic quiescence: Relative decrease in number of earthquakes or energy in any area of a seismic active region within a certain time interval in comparison with long-term observations in the same region.

Seismic activation: Relative increase in number of earthquakes or energy in any area of a seismic active region within a certain time interval in comparison with long-term observations in the same region.

Introduction

Both seismic quiescence and seismic activation before large earthquakes have been repeatedly described in the scientific literature. Earthquakes causing destruction of buildings, loss of life and economic damage are usually considered as large ones. Comparative size of different earthquakes is measured by their magnitude M (see ► “[Earthquake, Magnitude](#)”). The most destructive earthquakes fall in the range of magnitudes 7–9; and seismic energy released from an earthquake’s source (see ► “[Earthquakes, Energy](#)”) lies in the range of 10^{22} – 10^{25} ergs and the intensity of seismic shaking at the Earth surface (see ► “[Earthquakes, Intensity](#)”) exceeds the 7 level of the 12-degree-scale.

Just the first condition of detection of quiescence or activation follows from the stated above: a necessity of long-term history of seismic observations. However, for the purpose of rigorously proven detection of such occurrences one should be sure that any seismic event, magnitude of which exceeded some minimum threshold, was not missed within the observation period. Just in the second half of the twentieth century development of seismic networks of observation equipped with calibrated seismographs lead to creation of instrumental catalogues of earthquakes what enabled to judge objectively about appearance of seismic quiescence and activation.

The first sufficiently founded reports of appearance of seismic quiescence before large earthquakes were published at the beginning of the 1960s in Japanese and were practically unknown outside Japan. Then, in 1969 K. Mogi, on the basis of a visual analysis of seismicity maps, formulated a hypothesis that a seismic quiescence may be a precursor of a large earthquake (Mogi 1979). At the end of the 1980s M. Wyss and R. Habermann studied instrumental seismic catalogues of a number of regions of the Earth and determined basic rules of a formal detection of a seismic quiescence (Wyss and Habermann 1988). In brief, they may be summarized as follows: (1) evaluation of homogeneity of a seismic catalogue for the analyzed time interval, including magnitude calculations, taking into consideration changes in methods of determination of coordinates and depth of seismic events; (2) determination of a minimum magnitude of earthquakes being recorded without any omissions; (3) removal of groups and aftershocks in order to analyze so-called background seismicity; (4) quantitative evaluation of size and significance of an anomaly whereas significance shall be considered a statistically proven difference of an anomaly from random variations of the background seismicity; (5) quantitative determination of beginning of an anomaly; and (6) evaluation of sizes of an anomalous region. M. Wyss and R. Habermann determined that numerous events of seismic quiescence described in the literature are explained by changes of a minimum representative magnitude which were not taken into account owing to expansion of a seismic network or

development of information processing technique. However, there remain dozens of published events of seismic quiescence which may not be explained by any omissions of technical nature (artifacts).

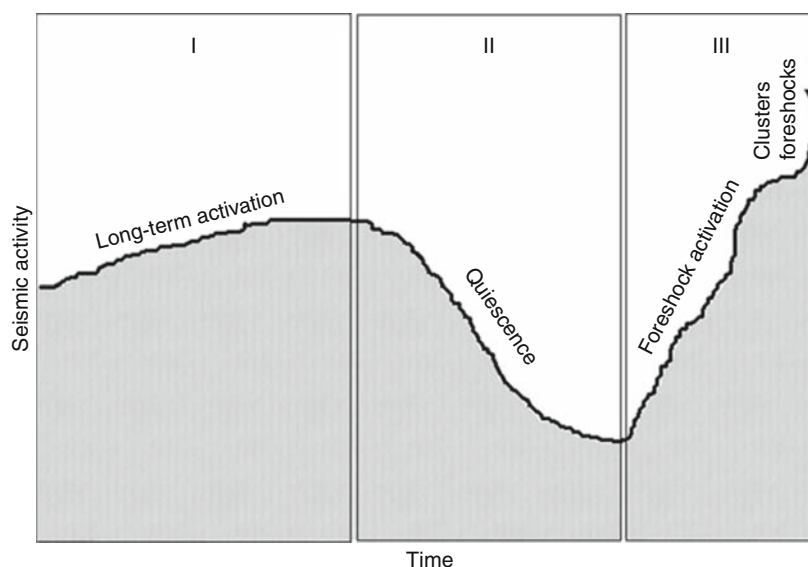
Description of certain events of seismic quiescence exceeds the limits of this entry. We will mention only as an example two widely discussed events. M. Ohtake et al. studied the seismicity along the Pacific Ocean coast of Mexico and found that after the middle of 1973 the seismicity rate in the region of 16.5°N and 96.5°W with the linear dimensions of about 100–200 km declined sharply what was interpreted as a seismic quiescence (Ohtake et al. 1977). On November 29, 1978, after 5.5 years, the Oaxaca earthquake occurred with $M = 7.6$ in this region. C. Kisslinger found a seismic quiescence before an Andreanof Islands earthquake of May 7, 1986, with $M = 8$ (Kisslinger 1988). Our task does not include any evaluation of authenticity and significance of these and other events, especially because it is not correctly to do it without the authors’ participation. We will note only that about 100 authors of scientific works published their investigations of seismic quiescence in regions with different geological and tectonic structure. It may be assumed that such phenomenon exists objectively in the nature and sometimes arises before a large earthquake. Since a seismic process to a great extent is self-similar in a wide range of magnitudes, a seismic quiescence most likely arises before a relative weak earthquake as well.

Main Phases of Seismic Quiescence and Activation

On the basis of thousands of observations of seismicity in various regions of the world there exists a generally accepted opinion among the seismologists that the large earthquake arises after a long-term (dozens of years) increase in seismicity in a relative region connected with a gradual growth of tectonic stresses. The latter, in their turn, arise at the joints of the earth’s crustal plates moving with a different rate. Let us name this period of increase in seismicity as a *phase of long-term seismic activation* (phase I at the Fig. 1). Spatial dimensions of a region of long-term activation before the large earthquake ($M = 7$ –8) may exceed 1,000 km.

On the background of such process, in some area of the region of long-term activation a mean seismicity rate decreases what, as a rule, results in decrease in rate of release of the seismic energy. A researcher has a task to detect such decrease in the seismic activation and to prove its significance at a quantitative level, that is, to detect the seismic quiescence. All the suggested methods are based somehow or other on an analysis of statistics of earthquakes which at the today’s level of development of seismological networks of observations includes events contained in catalogues, the magnitude of which is from three to four units less in comparison with the magnitude of the large earthquake. This statistics includes at

Seismic Quiescence and Activation, Fig. 1 Main phases of seismic quiescence and activation



best the first thousands of events within a period of some years what makes the task of quiescence detection difficult and not always solvable. Below we will describe in brief two methods which were analyzed to reveal a seismic quiescence before several earthquakes both *a retrospective one* (after the large earthquake) and *a prospective one* (before the large earthquake).

In the Z-value method there is calculated the standard deviate, Z, to estimate the significance of the rate change,

$$Z = (m_1 - m_2) / (S_1/n_1 + S_2/n_2)^{1/2},$$

where m is the mean rate, S the variance, and n is the number of events in the first and second period to be compared (Wyss and Habermann 1988).

For the purpose of significance evaluation there is applied the normal distribution law, but when having a rather large statistics of earthquakes the method gives plausible evaluations by other distribution law as well. The more the Z value is the more the difference between a rate of a seismic flow in the area of a supposed seismic quiescence and in comparison with a long-term rate. The calculations are carried out by the enumeration of nodes of the geographical grid and a changing number of earthquakes located near each node. It results the averaged Z values are presented in the form of maps in the researched seismic active region. The area of a seismic quiescence is determined according to high values of Z isolines, above the given rate of statistic significance. The experience showed that the more reliable results appeared in the task if there were removed swarms and aftershocks of previous earthquakes from the catalog.

The RTL-method uses three functions to measure the state of seismicity at a given location as a function of time (Sobolev 2001).

$$R(x, y, z, t) = \left[\sum \exp(-r_i/r_0) \right] - R_{ltr}$$

$$T(x, y, z, t) = \left[\sum \exp(-(t - t_i)/t_0) \right] - T_{ltr}$$

$$L(x, y, z, t) = \left[\sum (l_i/r_i)^p \right] - L_{ltr}$$

In these formulas, x, y, z, and t are the coordinates, the depth, and the analysis time, respectively. r_i is the epicentral distance from the location selected for analyses, t_i is the occurrence time of the past seismic events, and l_i is the length of rupture estimated by empirical relation with the magnitude. The R_{ltr} , T_{ltr} , L_{ltr} are the long-term averages of these functions. r_0 is a coefficient that characterizes the diminishing influence of more distant seismic events; t_0 is the coefficient characterizing the rate at which the preceding seismic events are “forgotten” as the time of analysis moves on; and p is the coefficient that characterizes the contribution of size of each preceding event. R, T, and L are dimensionless functions. They are normalized by their standard deviations. The product of the above three functions is calculated as the RTL-parameter, which describes the deviation from the background level of seismicity and is in units of the standard deviation, $\sigma_R \sigma_T \sigma_L$. The RTL decrease in relation to the background level means a seismic quiescence, and a structure of isolines of such parameter on the map indicates to the size of a quiescence area. In the RTL method the outgoing catalogue is cleaned of the aftershocks but the swarms are not removed.

An improvement of the method was proposed by Huang and Ding, which leads to a map of the quality of parameters. They select an area into the r_0 – t_0 space where at least 70% of the correlation coefficients over pairs exceeds a specified

threshold for r_0 and t_0 . Based on this map, they propose a method to find the optimal values of r_0 and t_0 (Huang and Ding 2012). Nagao et al. suggested a modified RTL algorithm, called the RTM algorithm, where magnitude parameter M_i was used instead of length of rupture L_i (Nagao et al. 2011).

Both methods (Z-value and RTL) require that the duration of period, which is included in calculations of the rate of background seismicity, must be by an order more than the supposed duration of the seismic quiescence. The experience of application of the described and other methods of revealing of the seismic quiescence in different tectonic conditions has shown (Sobolev 2011) that the duration of a quiescence period before the large earthquakes amounts to several years, that is, by an order less than the phase of the long-term activation (Fig. 1).

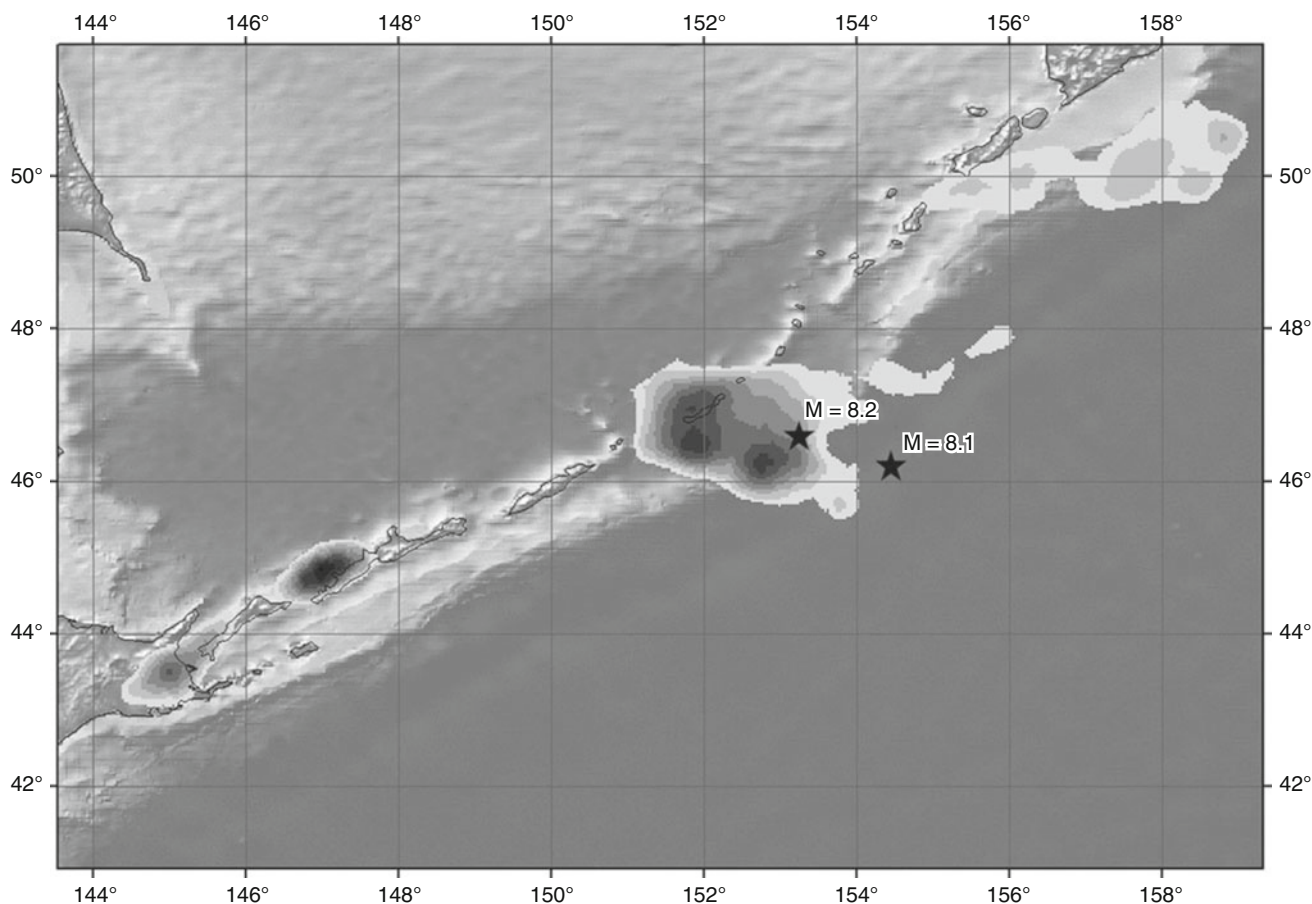
The area of significant values of the quiescence has linear dimensions about 300 km. In the most published works the seismic quiescence was determined by the application of only one method. The first exception, apparently, was a work by M. Wyss et al., in which a situation before the large earthquakes on the Sakhalin Island was analyzed by using of one and the same regional seismic catalogue but by the application of both methods mentioned above: Z-value and RTL (Wyss et al. 2004). The duration of quiescence anomalies before the Neftegorsk earthquake on May 27, 1995, $M = 7.1$, amounted according to results of both methods to 2.7 years. The size of a quiescence region was determined according to Z-value in the form of a rectangle with dimensions of 200×600 km, and according to the RTL data – in the form of a circle with the radius up to 400 km. The epicenter of the earthquake lied at the periphery of anomalies in both determinations. Before the Ulegorsk earthquake on August 4, 2000, $M = 6.8$, the duration of quiescence was according to Z-value 2.5 years and according to the RTL 3.0 years. The dimensions of the anomalous regions were determined in the form of a circle with the radius up to 165 and 200 km, respectively. The epicenter of the earthquake was also at the lateral part of quiescence area. The examination conducted with random catalogues has showed that the probability that the anomaly occurred by chance before the Neftegorsk earthquake did not exceed 1% and before the Ulegorsk earthquake 2% according to the results of both methods.

In spite of available facts of the seismic quiescence before some large earthquakes, the practical application of such anomaly for the forecast of earthquakes remains undecided. Authors of publications on the subject of seismic quiescence do not cite any data on a number of false alarms when no large earthquake occurred in the region of quiescence. According to our evaluations, the number of false alarms exceeds by several times the number of forecasts proven to be correct (even by a retrospective analysis). There is no statistic evaluation of a real *prospective* forecast with the use of such anomaly

indeed. Let us illustrate the above stated by a Fig. 2 where the areas of seismic quiescence in the region of the Kuriles arch are showed which were detected before the Simushir earthquake on November 15, 2006, $M = 8.2$. In this case the RTL method was applied, but it is not a matter of principal. The dark spots mean the areas of seismic quiescence. The main one of them with linear dimensions of about 200 km is located in the territory with coordinates of $46^\circ\text{--}48^\circ\text{N}$ – $151^\circ\text{--}154^\circ\text{E}$; the epicenter of the main shock is located at its periphery, and the strongest aftershock, $M = 8.1$, outside the anomaly. Another several anomalies of a lower size are showed in the figure: the territory with coordinates of 45°N – 147°E ; of 43.5°N – 145°E ; and of 50°N – 158°E . Any large earthquakes in these regions did not occur up to now.

It follows from the experience of investigations of the seismic quiescence that a large earthquake seldom occurs during the most significant rate of quiescence. It arises more often after the rate of seismic activity has increased again and even exceeded that which has been observed before the beginning of the quiescence. Let us name the time interval from minimal values of activity at the stage of quiescence to the moment of the large earthquake as a phase of foreshock activation (phase III on Fig. 1). Its duration has a wide range: from several days to the dozens of months. The area of the foreshock activation in general coincides with the area of aftershocks (see ► [“Earthquake, Aftershocks”](#)) and its size is estimated in hundreds of kilometers. These properties of the phase of foreshock activation bring to the understanding of the term “foreshock” in the broad sense. At the final step of development of such phase there may occur seismic events, the location of which coincide practically with the source of the following large earthquake. They are called traditionally foreshocks and should be understanding as foreshocks in the narrow sense (see ► [“Earthquake, Foreshocks”](#)). Usually they arise several hours before a large earthquake, reflecting apparently the nucleation of a rupture. This phenomenon is of great interest for a short-term forecast, but the practical application thereof has some difficulties because no reliable criteria are found till today in order to differentiate foreshocks from usual earthquakes. In exceptional events when a number of events occurring during several hours and at one and the same place amounts to dozens, they may be interpreted presumably as foreshocks of a future large earthquake and serve as a ground for a short-term forecast. The striking example is a prediction of the Highcheng earthquake, February 4, 1975, $M = 7.3$. The foreshocks determined after the large earthquake (retrospectively) are described in a great number of publications and the description thereof is beyond the scope of this entry.

Returning to the phase of foreshock activation (Fig. 1) we cite the example of its detection before the Simushir earthquake mentioned above as of November 15, 2006, $M = 8.2$. The RTL method developed for detection of a seismic



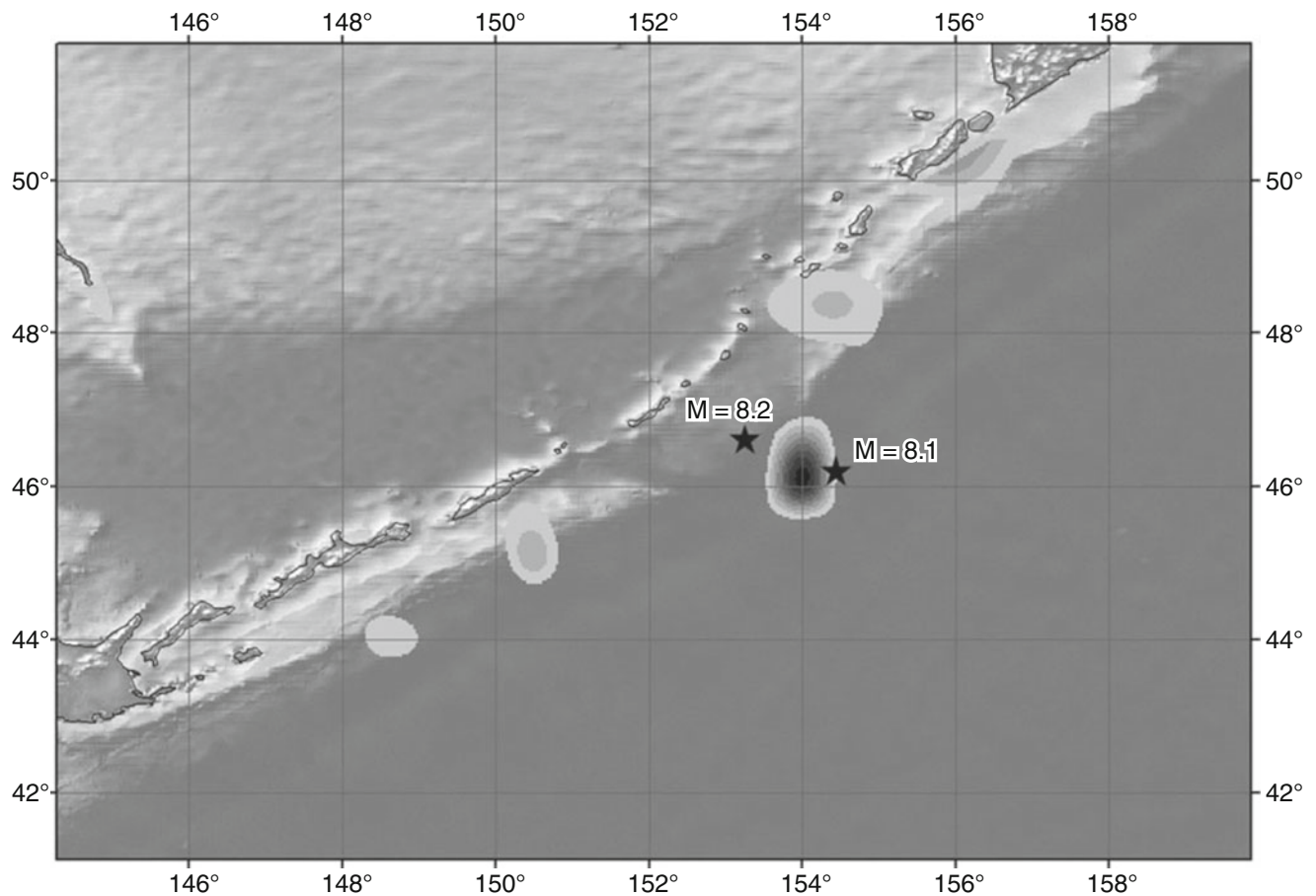
Seismic Quiescence and Activation, Fig. 2 Areas of seismic quiescence before the Simushir earthquake

quiescence allows also to identify the foreshock activation as a period of increase in the seismic activity after the quiescence. The situation showed in Fig. 3 is a continuation in time of that which appeared in this area of the Kurile Island arch during the seismic quiescence (Fig. 2). The development of the stage of seismic quiescence from the level of a long-term background to the lowest values of the seismic activity lasted in this event 1.5 years and the period of recovery of the stage of activity to the level of a long-term background (foreshock activation) lasted 1.5 years as well. The linear dimensions of foreshock activation amounted to about 100 km (Fig. 3). It was located in the area of aftershocks between the main shock with the $M = 8.2$ and the strongest aftershock, $M = 8.1$. Also, like at a seismic quiescence, within the region as showed on the map (Fig. 3) during the development of the main anomaly, there was selected a number of anomalies with a lower intensity: on the territory of coordinates 44°N – 149°E ; 45°N – 150.5°E ; 48.5°N – 154.5°E . In these places no large seismic events occurred up to now. Thus, the foreshock activation, like the seismic quiescence, is not of high reliability in the prognostic aspect because there exist anomalies qualified as false.

A detailed analysis of the redetermined earthquake catalog between 1996 and 2007, using a gridding technique (ZMAP), showed a pair of seismic quiescence and activation anomalies that start around the middle of 2005, and last about 30 months (Katsumata and Sakai 2013). The pair of quiescence and activation anomalies are located very close to each other. Nanjo et al. reported precursory seismic patterns prior to the 2016 Kumamoto earthquakes, as measured by four different methods: the b -value, the RTM-algorithm, the Z -value, and analysis of seismicity density in space and time. The 2016 Kumamoto earthquakes, including a magnitude $M7.3$ mainshock that struck on April 16, 2016, were preceded by periods of pronounced quiescence anomalies which lasted from 1.5 to >3 years (Nanjo et al. 2016).

Physical Mechanisms Leading to Seismic Quiescence and Activation

Let us consider possible physical mechanisms which bring to occurrences of the seismic quiescence and of the foreshock activation. There is no unified explanation at present and it is



Seismic Quiescence and Activation, Fig. 3 Areas of foreshock activation before the Simushir earthquake

expedient to discuss various hypotheses. We would like to note at first that except for the three phases discussed above and the occurrence of foreshocks in the narrow sense, there may appear sometimes also other anomalies. C. Scholz selected additional phases: doughnut pattern was observed in the period of development of the seismic quiescence around its external limits, and seismic silence was noticed just prior to the main event (Scholz 1990).

Let us consider in a more detailed manner hypotheses of nature of the stages of a seismic quiescence and of a foreshock activation. Ma et al. investigated the seismic patterns before nine large earthquakes in China (Ma et al. 1989). The authors did not select specially occurrences of a seismic quiescence or activation. They did describe mainly a migration of sources of seismic events for several years before large events in the regions adjacent to their epicenters with the dimensions of about first 100 km. Actual data cited by them are of interest as applied to a problem of physical nature of occurrences discussed in this entry. There was considered a migration of sources mainly with the magnitudes 3–6, while the large earthquakes had the magnitudes of above 8. The regions round the epicenters of large events were separated in three

parts: a hypocenter area A, which included a hypocenter of a large event; area of aftershocks B and external area C.

The following main characteristics of migration were selected. During 3–10 years till the moment of the large earthquake there was varied two phases in time. In the first one the events occurred mainly in the area C, while the areas A and B were characterized as relatively quiet. During the second phase which finished in the large earthquake the activity was detected in the areas A and B, while the seismic activity in the area C was reduced. As the second phase developed, the activity in the areas A and B continued to last up to several months or days in relation to the moments of the large earthquake; at the end of such series the activity either accumulated inside the hypocentral area A or continued to last in the area B, while the area A was calm. We should note that the authors did not cite any quantitative evaluations of the seismicity rate, thus conclusions may be made based in a qualitative manner only. Summarizing, there may be made a conclusion that the process of development of a large earthquake may occur according to various scripts, though in described events it related to continental events only.

The seismic quiescence may be a consequence of increase in the strength of rocks within the seismically active area. In the dilatancy-diffusion model the increase in the strength is explained by laboratory experiments, being a consequence of appearance of open microcracks and a relative drying of the rocks. However, it is not explained how such process can develop in the heterogeneous lithosphere in the spatial regions with the linear dimension of above 100 km.

The quiescence occurs also when the tectonic stress decreases. This can be a consequence of development of an unstable deformation in the source of a future large earthquake what constitutes one of the corner stones of the model of avalanche-like fracturing (Sobolev 2019). Expenditure of the accumulated potential energy brings to a decrease in stress both in the source and in the surrounding space. This explanation leads to expect an increase in seismic activity in the hypocentral area of a future large earthquake and a simultaneous occurrence of a seismic quiescence in the external area. This cannot always be observed. But the situation that the activation in the source occurs at the level of seismic events with small magnitudes, which are not registered by the available seismological network, cannot be excluded.

Decrease in the tectonic stress in some area of lithosphere can be a consequence of the motion of neighboring blocks with a different rate. If one of the blocks stops because of an strong asperity at its tectonic boundaries with any neighboring blocks, then the stress will be accumulated at points of a asperity, while in the middle of the stopped block it will relax gradually. The activation will start at boundaries of the stopped block and then it will migrate to central parts when the asperity is destroyed.

As a modification of such assumption is the situation when the stress increases and brings to the seismicity in the stronger blocks, and the block between them is relatively discharged. The quiescence occurring in the discharged block in such situation causes an effect of “false forerunners” as no large earthquake follows after them.

Kanamori suggested an explanation of different phases of seismicity based on a model of existence of a strong inclusion at a fault being heterogeneous in its strength (Kanamori 1981). When the stress increases gradually, the less strong inclusions disintegrate sequentially what brings to an accelerated growth of the stress at points being not disintegrated. There is observed a background seismicity at the low stress. As the stress increases, there occurs the doughnut pattern – a mass destruction of rocks around the strong inclusion. The latter has a seismic quiescence. When the stress approaches to a critical level, there occurs a destruction of subunits of the strong inclusion which characterizes the phase of foreshock activation. In this hypothesis, the facts of existence of the seismic quiescence at distances which significantly exceed the rupture sizes when a large earthquake occurs are not explained.

The reason of seismic quiescence can be a change of orientation of the tensor of the current stress due to development of a creep at the fault where a future earthquake will occur. The existing fractures, which get into new conditions, require the time for unstable development (kinetics of destruction). It is not proven that this, undoubtedly, existing phenomenon can cause a quiescence at distances which significantly exceed the length of a rupture of a future earthquake.

Another reason can be a transformation of the medium surrounding the source to the quasi-plastic state, for example, when the temperature or the hydrostatic compression increases. Then the process of fracturing will occur at a lower scale level which is beyond the registration of earthquakes by the seismic network. There remains a question what are the physical reasons of increase of the temperature or the compression in the lithosphere or the earth's crust within a relatively short time period (years).

It is necessary to mention another effect which influences on the fracturing in the geological medium. Laboratory and field observations prove that an increase or a decrease in the degree of water saturation of rocks results in a significant acceleration or retardation of occurrence of the seismic events (see ► [“Geomagnetically Induced Currents”](#)). Further, if the rate of the seismic activity increases due to that, then it results in a relative increase in the number of events with the relatively small magnitudes and in a decrease in the number of big events. The seismic quiescence may occur: (a) if a degree of water saturation decreases; (b) if under the influence of the increased water saturation the seismicity transforms to the level of small events being not registered by the seismic network. The activation is a direct consequence of the increased water saturation and may appear also in the form of swarms.

It follows from the most available experimental facts that the foreshock activation develops in the epicentral region. Its size is less than the area of seismic quiescence, and the centers of anomalous areas of these two phenomena do not coincide as a rule. There is a ground to suppose that the physics of foreshock activation is connected with the development of unstable deformation which is localized under the laws of mechanics in the zone of mainly two-dimensional extension. The complexity of fault systems, which display a fractal geometry, causes a successive occurrence of several zones of unstable deformation.

At the final period of foreshock activation there often occur clusters of seismic events, that is, groups of events, the distances between the hypocenters of which and the times between the successive events are less than mean values of the background seismicity. The occurrence of clusters at the phase of foreshock activation can be explained by two reasons. First, they occur by chance because of the increase in rate of formation of events in the narrow zone of unstable

deformation. Second, when the spatial density of accumulated active faults exceeds the critical level, there arise stress interactions among neighboring faults with the formation of the faults of a larger length. In the latter case such effect must appear with the increase of middle magnitudes of seismic events, that is, with the decrease of the *b*-value.

Braun and Peyrard investigated the origin of seismic quiescence with a generalized version of the Burridge–Knopoff model for earthquakes and found that it can be generated by a multi-peaked probability distribution of the thresholds at which contacts break. Such a distribution is not assumed a priori but naturally results from the aging of the contacts. It was shown that the model can exhibit quiescence as well as enhanced foreshock activity, depending on the value of some Parameters (Braun and Peyrard 2018).

Kawamura et al. applied two different approaches: the pattern informatics (PI) method and the ZMAP method, which is a gridding technique based on the standard deviate (*Z*-value) test and found that the epicenter of the 2013 ML6.2 Nantou earthquake was surrounded by three main seismic quiescence regions prior to its occurrence. The assumption that this is due to precursory slip (stress drop) on fault plane or its deeper extent of the ML6.2 Nantou earthquake was supported by previous researches based on seismicity data, geodetic data, and numerical simulations using rate- and state-dependent friction laws (Kawamura et al. 2014).

Yi-Ying Wen et al. investigated seismicity rate changes associated with the 2010 ML 6.4 Jiashan earthquake applying the RTL and pattern informatics algorithms (Wen et al. 2016). Both temporal and spatial results exhibited signatures of abnormal seismicity change related to the seismic activation and quiescence prior to main shock. Seismicity changes coincided with Coulomb stress change during the same period.

The 2004 Sumatra (Mw 9.1) earthquake was preceded by a seismic quiescence that began 13 years before the mainshock (Katsumata 2015). A detailed analysis of the earthquake catalog using a gridding technique (ZMAP) shows the quiescent area is located between 3° and 6° N, which covers the southeastern part of the focal area, including the rupture initiation point of the 2004 Sumatra earthquake. The observed spatial pattern of quiescence can be explained by stress perturbation due to a long-term slow slip located on the deeper edge of the mainshock fault.

Summary

It follows from the mentioned hypotheses that there exist various physical mechanisms of occurrence both of the seismic quiescence and of the foreshock activation. For the purpose of more fundamental understanding of such phenomena, the additional laboratory and field works are required. In our

opinion, special attention should be paid to the following directions:

To investigate on the same catalogues of earthquakes both the quiescence and the activation in a complex, paying attention to their distribution in space, time and to magnitudes
To compare, where possible, the seismic patterns with the field of deformations being estimated based on the data of satellite geodesy

To compare, where possible, the seismic patterns with the data of deep geoelectrical and hydrogeological investigations with the purpose of better understanding of the role of water

Cross-References

- [Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India](#)
- [Earthquake, Aftershocks](#)
- [Earthquake, Foreshocks](#)
- [Earthquake, Magnitude](#)
- [Earthquakes, Energy](#)
- [Earthquakes, Intensity](#)
- [Geomagnetically Induced Currents](#)

Bibliography

- Braun OM, Peyrard M (2018) Seismic quiescence in a frictional earthquake model. *Geophys J Int* 213:676–683
- Huang QH, Ding X (2012) Spatiotemporal variations of seismic quiescence prior to the 2011 M 9.0 Tohoku earthquake revealed by an improved Region-Time-Length algorithm. *Bull Seismol Soc Am* 102(4):1878–1883
- Kanamori H (1981) The nature of seismicity patterns before large earthquakes. In: *Earthquake prediction: an international review*. American Geophysical Union, Washington, DC, pp 1–19
- Katsumata K (2015) A long-term seismic quiescence before the 2004 Sumatra (Mw 9.1) earthquake. *Bull Seismol Soc Am* 105(1):167–176
- Katsumata K, Sakai S'i (2013) Seismic quiescence and activation anomalies from 2005 to 2008 beneath the Kanto district, central Honshu, Japan. *Earth Planets Space* 65:1463–1475
- Kawamura M, Chen C-c, Wu Y-M (2014) Seismicity change revealed by ETAS, PI, and *Z*-value methods: a case study of the 2013 Nantou, Taiwan earthquake. *Tectonophysics* 634:139–155
- Kisslinger C (1988) An experiment in earthquake prediction and the 7 May 1986 Andreanof Islands earthquake. *Bull Seismol Soc Am* 78:218–229
- Ma Z, Zhengxiang F, Zhang Y, Wang C, Zhang G, Liu D (1989) *Earthquake prediction: nine major earthquake in China*. Seismological Press/Springer, Berlin, p 332
- Mogi K (1979) Two kinds of seismic gap. *Pageoph* 117:1172–1186
- Nagao T, Takeuchi A, Nakamura K (2011) A new algorithm for the detection of seismic quiescence: introduction of the RTM algorithm, a modified RTL algorithm. *Earth Planets Space* 63:315–324
- Nanjo KZ, Izutsu J, Orihara Y, Furuse N, Togo S, Nitta H, Okada T, Tanaka R, Kamogawa M, Nagao T (2016) Seismicity prior to the 2016 Kumamoto earthquakes. *Earth Planets Space* 68:187

- Ohtake M, Matumoto T, Latham GV (1977) Seismicity gap near Oaxaca, Southern Mexico, as a probable precursor to a large earthquake. *Pageoph* 115:375–385
- Scholz CH (1990) *The mechanics of earthquakes and faulting*. Cambridge University Press, New York, p 439
- Sobolev G (2001) The examples of earthquake preparation in Kamchatka and Japan. *Tectonophysics* 338:269–279
- Sobolev GA (2011) Seismicity dynamics and earthquake predictability. *Nat Hazards Earth Syst Sci* 11:1–14
- Sobolev GA (2019) Avalanche unstable fracturing formation model. *Izv Phys Solid Earth* 55(1):1–14
- Wen Y-Y, Chen C-C, Wu Y-H, Chan C-H, Wang Y-J, Yeh Y-L (2016) Spatiotemporal investigation of seismicity and Coulomb stress variations prior to the 2010 ML 6.4 Jiashian, Taiwan earthquake. *Geophys Res Lett* 43:8451–8457
- Wyss M, Habermann RE (1988) Precursory seismic quiescence. *Pageoph* 126:319–332
- Wyss M, Sobolev G, Clippard JD (2004) Seismic quiescence precursors to two M7 earthquakes on Sakhalin Island, measured by two methods. *Earth Planets Space* 56:725–740

Seismic Ray Theory

Vlastislav Červený¹ and Ivan Pšenčík^{1,2}

¹Department of Geophysics, Faculty of Mathematics and Physics, Charles University, Praha, Czech Republic

²Institute of Geophysics, Academy of Sciences of Czech Republic, Praha, Czech Republic

Definition

Seismic ray theory High-frequency asymptotic method of study of seismic wavefields in complex inhomogeneous isotropic or anisotropic media with curved structural interfaces.

Introduction

The ray theory belongs to the methods most frequently used in seismology and seismic exploration for forward and inverse modelling of high-frequency seismic body waves. In smoothly varying media with smooth interfaces, it can provide useful approximate solutions of the elastodynamic equation of satisfactory accuracy. Starting from an intuitive description of the propagation of seismic waves along special trajectories – rays, it has developed into a highly sophisticated method, described briefly in this review paper.

The ray method has its advantages and limitations. The basic advantages are its applicability to complex, isotropic or anisotropic, laterally varying layered media and its numerical efficiency in such computations. It provides a physical insight into the wave propagation process by separating the wavefield into individual elementary waves and by allowing their

identification. In addition, it makes possible to track the paths (rays) in the medium along which energy of individual waves propagates, an aspect very important in tomography. The ray method also represents an important basis for other related, more sophisticated methods, such as the paraxial ray method, the Gaussian beam summation method, the Maslov method, the asymptotic diffraction theory, various combinations of the ray method with perturbation theory, etc. The ray method also has some limitations. It is approximate and applicable only to smooth media with smooth interfaces, in which the characteristic dimensions of inhomogeneities are considerably larger than the prevailing wavelength of the considered waves. The ray method can yield distorted results and may even fail in some special regions called *singular regions*.

The seismic ray method owes a lot to optics and radio-physics. Although the techniques used in different branches of physics are very similar, there are some substantial differences. The ray method in seismology is usually applied to more complicated structures than in optics or radiophysics. There are also different numbers and types of waves considered in different branches of physics.

The first seismological applications of ray concepts date back to the end of the nineteenth century. Then, only *kinematics*, specifically travel times, were used. Probably the first attempts to use also *dynamics* (amplitudes and waveforms) were made by Sir H. Jeffreys. The ray series solutions of elastodynamic equation with variable coefficients were first suggested by Babich (1956) and Karal and Keller (1959) for inhomogeneous isotropic media, and by Babich (1961) for inhomogeneous anisotropic media.

The Earth's interior is anisotropic or weakly anisotropic in some of its parts. Seismic anisotropy and its effects on wave propagation play an important role in contemporary seismology and seismic exploration. Consequently, it has also been necessary to develop the ray theory for elastic anisotropic media. It is important to emphasize that, for S waves, the ray theory for anisotropic media does not yield the ray theory for isotropic media in the zero anisotropy limit. For this reason, we describe systematically the ray theory for anisotropic media and also present corresponding formulae for isotropic media, and explain the differences between both of them.

S waves require generally a special attention. Well-understood phenomenon is propagation of two separate shear waves in anisotropic media. An underestimated phenomenon is shear-wave coupling, which occurs in weakly anisotropic media or in vicinities of shear-wave singularities. In such regions, standard ray theories for anisotropic as well as isotropic media do not work properly. Therefore, we also briefly describe the coupling ray theory for S waves, which fills the gap between ray theories for isotropic and anisotropic media.

We give here neither a detailed derivation of ray-theoretical expressions nor a relevant systematic

bibliography. This would extend the text considerably. We refer, however, to several textbooks, in which the ray theory is treated in a considerably greater detail (Červený et al. 1977; Kravtsov and Orlov 1990; Červený 2001; Chapman 2004; Babich and Kiselev 2018). The reader may also find useful information in several review papers devoted to seismic ray theory and its various aspects (Červený et al. 1988, 2007; Virieux 1996; Chapman 2002). Examples of computations based on the ray theory can be found, for example, in Červený et al. (1977) and Gjøystdal et al. (2002). Here we refer only to papers, in which the relevant methods and procedures were first proposed, and/or which give a useful more recent treatment of the subject.

We use the following notation. We denote Cartesian coordinates x_i and time t . The dots above letters denote partial derivatives with respect to time ($\ddot{u}_i = \partial^2 u_i / \partial t^2$) and the index following the comma in the subscript indicates the partial derivative with respect to the relevant Cartesian coordinate ($u_{i,j} = \partial u_i / \partial x_j$). We consider high-frequency time-harmonic seismic body waves, with the exponential factor $\exp(-i\omega t)$, where ω is fixed, positive, real-valued circular frequency. The lower-case Roman indices take the values 1, 2, 3, the upper-case indices 1, 2. Hats over bold symbols indicate 3×3 matrices, bold symbols without hats denote 2×2 matrices. The Einstein summation convention over repeating Roman indices is used, with exception of indices in parentheses.

Basic Equations of the Seismic Ray Method

For smoothly varying elastic media, the source-free *equation of motion* reads

$$\tau_{ij,j} - \rho \ddot{u}_i = 0. \quad (1)$$

Here $\tau_{ij}(x_n, t)$, and $u_i(x_n, t)$ are Cartesian components of stress tensor and displacement vector, respectively, and ρ is the density. In anisotropic media, the stress tensor τ_{ij} and the infinitesimal strain tensor $e_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i})$ are related by Hooke's law:

$$\tau_{ij} = c_{ijkl} e_{kl} = c_{ijkl} \partial u_k / \partial x_l. \quad (2)$$

$c_{ijkl}(x_n)$ is a tensor of elastic moduli (stiffness tensor), satisfying symmetry relations $c_{ijkl} = c_{jikl} = c_{ijlk} = c_{klij}$. There are, at the most, 21 independent elastic moduli. Inserting Eq. 2 into Eq. 1, we get the *elastodynamic equation*

$$(c_{ijkl} u_{k,l})_{,j} - \rho \ddot{u}_i = 0. \quad (3)$$

In the seismic ray method, high-frequency seismic body waves propagating in smoothly varying, isotropic or

anisotropic, media are studied. The formal ray series solution of the elastodynamic equation, see Eq. 3, for the displacement vector $\mathbf{u}(x_n, t)$ is sought in the form of an asymptotic series in inverse powers of circular frequency ω ,

$$\mathbf{u}(x_n, t) = \exp[-i\omega(t - T(x_n))] \left[\mathbf{U}^{(0)}(x_n) + \frac{\mathbf{U}^{(1)}(x_n)}{(-i\omega)} + \frac{\mathbf{U}^{(2)}(x_n)}{(-i\omega)^2} + \dots \right]. \quad (4)$$

Here $T(x_n)$ is the real-valued travel time, $\mathbf{U}^{(k)}$, $k = 0, 1, 2, \dots$ are complex-valued *vectorial amplitude coefficients*. Surfaces $T(x_i) = \text{const.}$ are called *wavefronts* (or phase fronts). In perfectly elastic media, functions $T(x_n)$, and $\mathbf{U}^{(k)}(x_n)$ are frequency independent.

Also other forms of the ray series have been used in the seismic ray method. For example, Chapman (2004) developed the seismic ray method using the ray series for particle velocity and traction. Such a formal ray series has certain advantages with respect to Eq. 4. Here, however, we consider systematically the traditional ray series for the displacement vector in the form of Eq. 4.

Inserting Eq. 4 into elastodynamic equation, see Eq. 3, we obtain a series in inverse powers of ω , which equals zero. Consequently, the coefficients of the individual powers of ω must also equal zero. This yields a system of equations called the basic recurrence system of equations of the ray method. This system can be used to determine the eikonal equation for travel times $T(x_n)$ and, successively, the equations for the amplitude coefficients $\mathbf{U}^{(0)}(x_n)$, $\mathbf{U}^{(1)}(x_n)$, $\mathbf{U}^{(2)}(x_n), \dots$. The equations for $\mathbf{U}^{(k)}(x_n)$ yield, among others, transport equations. For a detailed derivation of the basic system of equations of the ray method see Červený (2001, Sect. 5.7).

The vectorial amplitude coefficients $\mathbf{U}^{(k)}(x_n)$, $k = 1, 2, \dots$ can be expressed as a sum of the *principal component* and *additional component*. The zero-order amplitude coefficient $\mathbf{U}^{(0)}(x_n)$, which has no additional component, is an exception. Its principal component is determined from the transport equation discussed later. The principal component of $\mathbf{U}^{(k)}(x_n)$ for $k > 0$ is the projection of $\mathbf{U}^{(k)}(x_n)$ into the unit vector parallel to the zero-order amplitude coefficient $\mathbf{U}^{(0)}(x_n)$, the additional component of $\mathbf{U}^{(k)}(x_n)$ is the remaining part of $\mathbf{U}^{(k)}(x_n)$. The complexity of the equations for higher-order amplitude coefficients $\mathbf{U}^{(k)}$ increases rapidly with increasing k . Moreover, the higher-order amplitude coefficients are unstable and usually do not increase the accuracy notably. They are very sensitive to fine details of the medium. The instability of the amplitude coefficients increases with increasing k . For these reasons, only the zero-order coefficient $\mathbf{U}^{(0)}(x_n)$, with the additional component of $\mathbf{U}^{(1)}(x_n)$, at the most, has been used in seismological applications. In the following, we shall concentrate on the zero-order ray approximation $\mathbf{U}^{(0)}(x_n)$ only.

The zero-order approximation of the ray method reads:

$$\mathbf{u}(x_n, t) = \mathbf{U}(x_n) \exp[-i\omega(t - T(x_n))]. \quad (5)$$

In Eq. 5, we have dropped the superscript (0) of $\mathbf{U}(x_n)$. We call $\mathbf{U}(x_n)$ the complex-valued *vectorial amplitude*. In smooth, laterally varying media, containing smooth structural interfaces, the zero-order approximation of the ray method in Eq. 5 usually offers sufficiently accurate results, particularly for travel time $T(x_n)$. Its great advantage is that it allows one to work with frequency-independent travel time and amplitude. However, if the medium under consideration becomes more and more complex (less smooth) or if a singularity is close, vectorial amplitude $\mathbf{U}(x_n)$ becomes less accurate. In structures exceeding certain degree of complexity or close to singularities, the ray method may yield inaccurate results or even fail.

The first equation of the basic recurrence system of equations of the ray method reads:

$$(\Gamma_{ik} - \delta_{ik})U_k = 0, \quad i = 1, 2, 3. \quad (6)$$

Here Γ is the 3×3 *generalized Christoffel matrix* with elements given by the relation:

$$\Gamma_{ik} = a_{ijkl}p_j p_l. \quad (7)$$

In Eq. 7, p_i are the Cartesian components of the slowness vector $\mathbf{p}$,

$$p_i = \partial T / \partial x_i \quad (8)$$

and $a_{ijkl} = c_{ijkl}/\rho$ are density-normalized elastic moduli. Note that the classical Christoffel matrix, with elements $a_{ijkl}n_j n_l$, contains components of the real-valued unit vector $\mathbf{n}$ (perpendicular to the wavefront) instead of $\mathbf{p}$. For this reason, we call the matrix in Eq. 7 the “generalized” Christoffel matrix. The relation between p_i and n_i is $p_i = n_i/C$, where C is the phase velocity.

The generalized 3×3 Christoffel matrix in solid elastic media is symmetric ($\Gamma_{ik} = \Gamma_{ki}$), positive definite ($\Gamma_{ik}a_i a_k > 0$, where a_i are components of any non-vanishing real-valued vector), and homogeneous function of the second degree in p_i ($\Gamma_{ik}(x_n, ap_j) = a^2 \Gamma_{ik}(x_n, p_j)$ for any nonvanishing constant a). It has three real-valued positive eigenvalues $G_m(x_n, p_j)$, and three corresponding real-valued unit eigenvectors $\mathbf{g}^{(m)}(x_n, p_j)$, $m = 1, 2, 3$. G_m and $\mathbf{g}^{(m)}$ are solutions of the eigenvalue equations

$$(\Gamma_{ik} - \delta_{ik}G_m)g_k^{(m)} = 0, \quad i = 1, 2, 3. \quad (9)$$

Eigenvectors $\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \mathbf{g}^{(3)}$ are mutually perpendicular. Eigenvalue G_m and the relevant eigenvector $\mathbf{g}^{(m)}$ are mutually related as follows:

$$G_m = \Gamma_{ik}g_i^{(m)}g_k^{(m)} = a_{ijkl}p_j p_l g_i^{(m)}g_k^{(m)}. \quad (10)$$

For isotropic media, elastic moduli $c_{ijkl}(x_n)$ can be specified in terms of Lamé’s elastic moduli $\lambda(x_n)$ and $\mu(x_n)$, describing isotropic media, as follows:

$$c_{ijkl} = \lambda \delta_{ij} \delta_{kl} + \mu (\delta_{ik} \delta_{jl} + \delta_{il} \delta_{jk}). \quad (11)$$

Elements of the generalized Christoffel matrix are then given by the relation:

$$\Gamma_{ik} = \frac{\lambda + \mu}{\rho} p_i p_k + \frac{\mu}{\rho} \delta_{ik} p_n p_n. \quad (12)$$

In isotropic media, the expressions for eigenvalues and eigenvectors of the generalized Christoffel matrix can be determined analytically:

$$G_1 = G_2 = \beta^2 p_k p_k, \quad G_3 = \alpha^2 p_k p_k. \quad (13)$$

Here

$$\alpha^2 = (\lambda + 2\mu)/\rho, \quad \beta^2 = \mu/\rho. \quad (14)$$

The eigenvector relevant to the eigenvalue G_3 equals $\mathbf{n}$, the unit vector perpendicular to the wavefront. The eigenvectors relevant to coinciding eigenvalues G_1 and G_2 are mutually perpendicular unit vectors situated arbitrarily in the plane perpendicular to $\mathbf{n}$ and tangent to the wavefront.

Eikonal Equation. Polarization Vector

The comparison of Eq. 6 with the eigenvalue equation for the 3×3 generalized Christoffel matrix, see Eq. 9, shows that Eq. 6 is satisfied, if the eigenvalue G_m of the generalized Christoffel matrix satisfies the relation

$$G_m(x_i, p_j) = 1, \quad (15)$$

and if the complex-valued vectorial amplitude $\mathbf{U}$ of the wave under consideration is related to eigenvector $\mathbf{g}^{(m)}$ as follows:

$$\mathbf{U} = A \mathbf{g}^{(m)}. \quad (16)$$

Equation 15 is the important *eikonal equation*. It is a nonlinear, first-order partial differential equation for travel time $T(x_n)$. Equation 16 shows that displacement vector $\mathbf{U}$ is parallel to the appropriate eigenvector $\mathbf{g}^{(m)}$ of the generalized Christoffel matrix. For this reason, we call $\mathbf{g}^{(m)}$ the *polarization vector*. Symbol $A(x_n)$ denotes the complex-valued, frequency-independent, *scalar amplitude*.

Taking into account that G_m is a homogeneous function of the second degree in p_i , where $\mathbf{p} = C^{-1}\mathbf{n}$, we obtain $G_m(x_i, p_j) = C^{-2}G_m(x_i, n_j)$. This and Eqs. 10 and 15 yield

$$C^2(x_i, n_j) = G_m(x_i, n_j) = a_{ijkl}n_j n_l g_i^{(m)} g_k^{(m)}. \quad (17)$$

Phase velocity C is the velocity of the propagation of the wavefront in the direction $\mathbf{n}$. The phase-velocity vector $\mathbf{C} = C(x_i, n_j)\mathbf{n}$ has the direction of $\mathbf{n}$, i.e., it is perpendicular to the wavefront. It follows from Eq. 17 that the squares of phase velocity C are eigenvalues $G_m(x_i, n_j)$ of the classical Christoffel matrix with elements $a_{ijkl}n_j n_l$.

Generally, eigenvalues G_m , $m = 1, 2, 3$, of the generalized Christoffel matrix are mutually different. They correspond to three high-frequency body waves propagating in inhomogeneous anisotropic media. Smaller eigenvalues, G_1 and G_2 , belong usually to S1 and S2 waves, largest, G_3 , to P wave. Such situations are considered here. If the eigenvalues are mutually different, their polarization vectors can be determined uniquely.

If two eigenvalues coincide, we speak of the *degenerate case* of the eigenvalue problem. The corresponding eigenvectors can then be chosen as mutually perpendicular vectors situated arbitrarily in the plane perpendicular to the eigenvector related to the different eigenvalue. Eigenvalues may coincide *locally*, along certain lines or at certain points, which correspond to the so-called *S wave singular directions*, or may be close to one another *globally* in a vicinity of singular directions or in *weakly anisotropic media*. The approximate but unique determination of polarization vectors in the latter situations is possible using perturbation approach (Jech and Pšenčík 1989).

In isotropic media, the S wave eigenvalues G_1 and G_2 coincide *globally*, see Eq. 13. Consequently, in isotropic media, the S waves are controlled by a single eikonal equation and we have thus only two different eikonal equations corresponding to P and S waves. As the equations for the eigenvalues in isotropic media can be determined analytically, see Eq. 13, we can express the eikonal equations, see Eq. 15, for P and S waves explicitly:

$$\alpha^2 p_k p_k = 1 \quad \text{for P waves}, \quad (18)$$

$$\beta^2 p_k p_k = 1 \quad \text{for S waves}. \quad (19)$$

In isotropic media, the generally complex-valued amplitude vector $\mathbf{U}$ can be expressed in the simple form, see Eq. 16, only for P waves. In this case the polarization vector $\mathbf{g}^{(3)} = \mathbf{n}$, i.e., it is perpendicular to the wavefront. For S waves, $\mathbf{U}$ must be considered in the following form:

$$\mathbf{U} = B\mathbf{g}^{(1)} + C\mathbf{g}^{(2)}. \quad (20)$$

Here $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ are two mutually perpendicular unit vectors in the plane tangent to the wavefront, i.e., perpendicular to the vector $\mathbf{n}$. The determination of $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ along the

ray is explained later, see Eq. 38. Symbols $B(x_n)$ and $C(x_n)$ are the corresponding, generally complex-valued scalar amplitudes.

In the seismic ray method, it is common to express the eikonal equation, see Eq. 15, in Hamiltonian form. Hamiltonian $\mathcal{H}(x_i, p_j)$ may be introduced in various ways. We shall consider the Hamiltonian, which is a homogeneous function of the second degree in p_i . For inhomogeneous anisotropic media, we can introduce the Hamiltonian expressed in terms of $G_m(x_i, p_j)$, see Eq. 10:

$$\mathcal{H}(x_i, p_j) = \frac{1}{2}G_m(x_i, p_j) = \frac{1}{2}a_{ijkl}p_j p_l g_i^{(m)} g_k^{(m)}. \quad (21)$$

The eikonal equation, see Eq. 15, then reads:

$$\mathcal{H}(x_i, p_j) = \frac{1}{2}. \quad (22)$$

It holds for anisotropic as well as isotropic media.

From Eqs. 13 and 21, we get for isotropic inhomogeneous media:

$$\mathcal{H}(x_i, p_j) = \frac{1}{2}V^2(x_i)p_k p_k, \quad (23)$$

where $V = \alpha$ for P waves and $V = \beta$ for S waves.

Ray Tracing and Travel-Time Computation

The eikonal equation in Hamiltonian form as shown in Eq. 22, with $p_j = \partial T / \partial x_j$, is a nonlinear partial differential equation of the first order for travel time $T(x_i)$. It can be solved by the method of characteristics. The characteristics of eikonal equation, see Eq. 22, are spatial trajectories, along which Eq. 22 is satisfied, and along which travel time T can be computed by quadratures. The characteristics of the eikonal equation represent *rays*.

The characteristics of the eikonal equation expressed in general Hamiltonian form are described by a system of nonlinear, ordinary differential equations of the first order:

$$\frac{dx_i}{du} = \frac{\partial \mathcal{H}}{\partial p_i}, \quad \frac{dp_i}{du} = -\frac{\partial \mathcal{H}}{\partial x_i}, \quad \frac{dT}{du} = p_k \frac{\partial \mathcal{H}}{\partial p_k}. \quad (24)$$

Here u is a real-valued parameter along the ray. The relation between parameter u and the travel time along the ray depends on the form of the Hamiltonian used; see the last equation in Eq. 24. For Hamiltonians, which are homogeneous functions of the second degree in p_i , the Euler equation for homogeneous functions yields $p_k \partial \mathcal{H} / \partial p_k = 2\mathcal{H}$. If we consider Hamiltonian in the form of Eq. 22, we get $dT/du = 1$ from Eq. 24. For travel time T along the ray, denoted $\tau = T$, Eq. 24 simplifies to:

$$\frac{dx_i}{d\tau} = \frac{\partial \mathcal{H}}{\partial p_i}, \quad \frac{dp_i}{d\tau} = -\frac{\partial \mathcal{H}}{\partial x_i}. \quad (25)$$

This system of equations is usually called the *ray tracing system*. Solution of the ray tracing system, see Eq. 25, with appropriate initial conditions yields $x_i(\tau)$, the coordinates of points along the ray trajectory, and $p_i(\tau)$, the Cartesian components of the slowness vectors along the ray. The travel time T along the ray is obtained automatically, $T = \tau$.

Inserting Eq. 21 to Eq. 25, we obtain the ray tracing system for m -th wave in the inhomogeneous anisotropic medium:

$$\frac{dx_i}{d\tau} = a_{ijkl} p_l g_j^{(m)} g_k^{(m)}, \quad \frac{dp_i}{d\tau} = -\frac{1}{2} \frac{\partial a_{ijkl}}{\partial x_i} p_k p_n g_j^{(m)} g_l^{(m)}. \quad (26)$$

In the derivation of the first equation of Eq. 26 for $\partial \mathcal{H} / \partial p_i$, we took into account that $\Gamma_{ik} \partial (g_i^{(m)} g_k^{(m)}) / \partial p_n = 0$. An alternative version of ray-tracing equations, see Eq. 26, was derived by Červený (1972), in which the eigenvectors $\mathbf{g}^{(m)}$ were not used.

The initial conditions for the ray tracing system, see Eq. 26, are $x_i = x_{0i}$, $p_i = p_{0i}$, where x_{0i} and p_{0i} satisfy the eikonal equation, see Eq. 22, corresponding to the wave we wish to compute (P, S1 or S2). Components p_{0i} of the initial slowness vector $\mathbf{p}_0$ can be expressed as $p_{0i} = n_{0i} / C(x_{0j})$, where C is the relevant phase velocity. The eikonal equation, see Eq. 22, is then satisfied along the whole ray.

In inhomogeneous isotropic media, the ray tracing system, see Eq. 25 with Eq. 23, yields

$$\frac{dx_i}{d\tau} = V^2 p_i, \quad \frac{dp_i}{d\tau} = -\frac{\partial \ln V}{\partial x_i}. \quad (27)$$

The initial conditions for the ray tracing system, see Eq. 27, are again $x_i = x_{0i}$, $p_i = p_{0i}$, where $p_{0i} = n_{0i} / V(x_{0j})$. Here $V = \alpha$ for P waves, and $V = \beta$ for S waves.

As τ is the travel time along the ray, $dx_i / d\tau$ in Eq. 26 represents the Cartesian components $\mathcal{U}_i$ of the *ray-velocity vector* $\mathbf{U}$ of the m -th wave:

$$\mathcal{U}_i = a_{ijkl} p_l g_j^{(m)} g_k^{(m)}. \quad (28)$$

In elastic anisotropic media, the ray-velocity vector $\mathbf{U}$ is also called the group-velocity vector or the energy-velocity vector. As indicated by the name, the energy velocity vector $\mathbf{U}$ represents the velocity of the energy propagation.

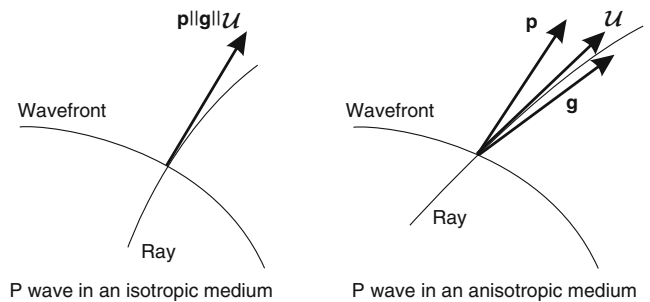
In anisotropic media, the ray-velocity vector $\mathbf{U}$ must be strictly distinguished from the phase-velocity vector $\mathbf{C}$. In inhomogeneous anisotropic media, the ray-velocity and phase-velocity vectors $\mathbf{U}$ and $\mathbf{C}$ are generally different, both in size and direction. Vector $\mathbf{U}$ is always greater than $\mathbf{C}$. The two vectors are equal (in size and direction) only in special directions, called *longitudinal directions*.

In inhomogeneous isotropic media, Eq. 28 for the ray-velocity vector yields $\mathbf{U} = V^2 \mathbf{p}$. For the phase-velocity vector, using Eq. 17, we get $\mathbf{C} = V^2 \mathbf{p}$, where $V = \alpha$ for P waves, and $V = \beta$ for S waves. Thus, the ray-velocity and phase-velocity vectors are identical in isotropic media.

Figure 1 shows mutual orientation of ray-velocity vector $\mathbf{U}$, phase-velocity vector $\mathbf{C}$ (parallel to slowness vector $\mathbf{p}$), and polarization vector $\mathbf{g}$ of a P wave propagating in an isotropic (left) and anisotropic (right) medium. While $\mathbf{U}$, $\mathbf{C}$, and $\mathbf{g}$ are parallel in isotropic media, they generally differ in anisotropic media. For S waves, the vectors $\mathbf{U}$ and $\mathbf{C}$ have similar orientation as in the case of P waves. The polarization vectors $\mathbf{g}$ are, however, perpendicular (isotropic medium) or nearly perpendicular (anisotropic medium) to the ray.

Ray tracing systems in Eqs. 26 and 27 can be simply solved if the initial values x_{0i} and p_{0i} are specified at some point S of the ray. We then speak of *initial-value ray tracing*. The standard numerical procedures of solving the system of ordinary differential equations of the first order with specified initial conditions can then be used (for example, Runge-Kutta). A very important role in seismology is played by *boundary-value ray tracing*, in which we seek the ray, satisfying prescribed boundary conditions. The typical boundary value problem is *two-point ray tracing*, in which we seek the ray connecting two specified points. Mostly, the controlled initial-value ray tracing (controlled shooting method) is used to solve this problem (Červený et al. 2007). Boundary-value ray tracing is considerably more complicated than initial-value ray tracing.

There are four important differences between initial value ray tracing in isotropic and anisotropic media. First: In anisotropic media, we deal with three waves, P, S1, and S2, in isotropic media with two waves, P and S, only. Second: In inhomogeneous anisotropic media, ray tracing system, see Eq. 26, is the same for all three waves. The wave under consideration is specified by the initial conditions, which must satisfy the eikonal equation of the considered wave. In isotropic inhomogeneous media, the ray tracing systems are different for P and S waves, see Eq. 27 with $V = \alpha$ and $V = \beta$,



Seismic Ray Theory, Fig. 1 Slowness vector $\mathbf{p}$ (perpendicular to the wavefront), ray-velocity vector $\mathbf{U}$ (tangent to the ray), and polarization vector $\mathbf{g}$ of a P wave propagating in an isotropic (left) and anisotropic (right) medium. For simplicity, the three vectors in the right-hand plot are shown in one plane. In general, this is not the case in anisotropic media

respectively. Third: In isotropic media, the initial direction of the slowness vector specifies directly the initial direction of the ray (as the tangents to the ray and the slowness vector have the same directions). In anisotropic media, the direction of the ray is, generally, different from the direction of the slowness vector. Nevertheless, we have to use p_{0i} as the initial values for the ray tracing system. The ray-velocity vector $\mathbf{U}$ can be simply calculated from slowness vector $\mathbf{p}$ at any point of the ray, including the initial point, however, not vice versa. Fourth: Ray tracing for P and S waves is regular everywhere in inhomogeneous isotropic media. In anisotropic media, problems arise with tracing S-wave rays in vicinities of singular directions, or if medium is nearly isotropic (quasi-isotropic).

The problem of ray tracing and travel-time computation in inhomogeneous media has been broadly discussed in the seismological literature, particularly for inhomogeneous isotropic media. Many ray tracing systems and many suitable numerical procedures for performing ray tracing have been proposed. For 1-D isotropic media (vertically inhomogeneous, radially symmetric), the ray tracing systems may be simplified so that they reduce to simple quadratures, well-known from classical seismological textbooks. For P-wave travel time T and angular distance Δ , they read:

$$T(p) = \int \left(1 - \frac{p^2 \alpha^2}{r^2} \right)^{-1/2} \frac{dr}{\alpha},$$

$$\Delta(p) = \int \frac{p\alpha}{r^2} \left(1 - \frac{p^2 \alpha^2}{r^2} \right)^{-1/2} dr. \quad (29)$$

Here, $\alpha = \alpha(r)$ is the P wave velocity as the function of the radius r , and p is the ray parameter, $p = (r \sin i)/\alpha$, where i is the angle made by the ray and the radius, see Eq. (9.21) of Aki and Richards (2002), where additional details can be found.

Standard programs for ray tracing and travel-time computations in laterally varying isotropic and anisotropic structures are available, see, for example, program packages SEIS (2D isotropic models), CRT, and ANRAY (3D isotropic/anisotropic models) available at <http://sw3d.cz/>. Programs for anisotropic media have, however, problems with S wave computations in weakly anisotropic media and in the vicinities of shear-wave singularities. In such cases, the standard ray theory should be replaced by the coupling ray theory.

Ray tracing may also serve as a basis for the so-called *wavefront construction method* (Gjøystdal et al. 2002). In this case, for a selected wave, wavefronts with travel times $T = T_0 + k\Delta T$ are computed successively from the previous wavefronts with travel times $T = T_0 + (k - 1)\Delta T$. The wavefront construction method has found broad applications in seismic exploration.

Let us consider a two-parametric system of rays, call it the *ray field*, and specify the individual rays in the ray field by *ray parameters* γ_1, γ_2 . Ray parameters γ_1, γ_2 may represent,

e.g., the take-off angles at a point source or the curvilinear Gaussian coordinates of initial ray points along the initial surface. The family of rays with ray parameters within the limit $[\gamma_1, \gamma_1 + d\gamma_1], [\gamma_2, \gamma_2 + d\gamma_2]$, is called the *elementary ray tube* or briefly the ray tube. We further introduce *ray coordinates* $\gamma_1, \gamma_2, \gamma_3$ in such a way that γ_1, γ_2 are ray parameters, and γ_3 is some monotonic parameter along a ray (arclength s , travel time τ , etc.). Here we consider $\gamma_3 = \tau$, but our results may be simply modified for any other monotonic parameter γ_3 . We further introduce the 3×3 transformation matrix $\hat{\mathbf{Q}}$ from ray to Cartesian coordinates with elements $Q_{ij} = \partial x_i / \partial \gamma_j$. The Jacobian of transformation from ray to Cartesian coordinates, $\det \hat{\mathbf{Q}}$, can be expressed as follows:

$$\det \hat{\mathbf{Q}}(\tau) = (\partial \mathbf{x}(\tau) / \partial \gamma_1 \times \partial \mathbf{x}(\tau) / \partial \gamma_2)^T \mathbf{U}(\tau). \quad (30)$$

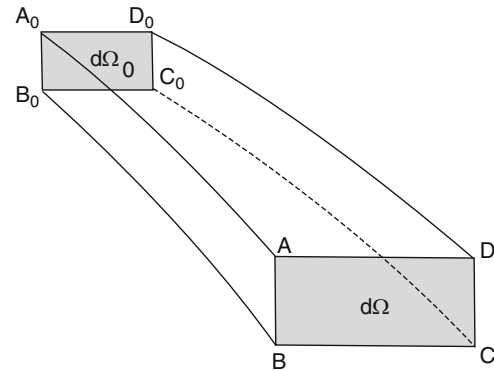
The vectorial product in Eq. 30 has the direction of the normal to the wavefront, specified by $\mathbf{n} = C \mathbf{p}$. As $\mathbf{p}(\tau) \cdot \mathbf{U}(\tau) = 1$, see Eqs. 10, 15, and 28, we also obtain

$$\det \hat{\mathbf{Q}}(\tau) = \pm C(\tau) |(\partial \mathbf{x}(\tau) / \partial \gamma_1 \times \partial \mathbf{x}(\tau) / \partial \gamma_2)|. \quad (31)$$

Thus the Jacobian $\det \hat{\mathbf{Q}}(\tau)$ equals $\pm C(\tau) d\Omega(\tau)$, where $d\Omega(\tau) = |(\partial \mathbf{x}(\tau) / \partial \gamma_1 \times \partial \mathbf{x}(\tau) / \partial \gamma_2)|$ is the scalar surface element cut out of the wavefront by the ray tube. It measures the expansion or contraction of the ray tube, see Fig. 2. For this reason, the 3×3 matrix $\hat{\mathbf{Q}}(\tau)$ is also often called the *geometrical spreading matrix* and various quantities related to $\det \hat{\mathbf{Q}}(\tau)$ are called *geometrical spreading*. It plays an important role in the computation of the ray-theory amplitudes.

Transport Equation. Computation of Ray-Theory Amplitudes

The second equation of the basic recurrence system of equations of the ray method yields the transport equation for the



Seismic Ray Theory, Fig. 2 Elementary ray tube. $d\Omega_0$ and $d\Omega$ are scalar surface elements cut out of the wavefront by the ray tube. This means that in isotropic media, the normals to $d\Omega_0$ and $d\Omega$ are parallel to rays. In anisotropic media, they are not

scalar ray-theory amplitude $A(x_i)$ specifying the zero-order vectorial amplitude $\mathbf{U}(x_n)$, see Eq. 16. The transport equation is a partial differential equation of the first order. It can be expressed in several forms. One of them, valid for isotropic and anisotropic media, and for any wave, reads

$$\nabla \cdot (\rho A^2 \mathbf{U}) = 0. \quad (32)$$

It is common to solve the transport equation along the ray. $\nabla \cdot \mathbf{U}$ can then be expressed as follows:

$$\nabla \cdot \mathbf{U} = d \left[\ln (\det \hat{\mathbf{Q}}) \right] / d\tau \quad (33)$$

(Červený 2001, Eq. 3.10.24). Inserting Eq. 33 into Eq. 32 yields the transport equation in the form of the first-order ordinary differential equation along the ray:

$$d \left(\rho(\tau) A^2(\tau) \det \hat{\mathbf{Q}}(\tau) \right) / d\tau = 0. \quad (34)$$

This yields a simple form of the continuation relation for $A(\tau)$ along the ray:

$$A(\tau) = \left[\frac{\rho(\tau_0) \det \hat{\mathbf{Q}}(\tau_0)}{\rho(\tau) \det \hat{\mathbf{Q}}(\tau)} \right]^{1/2} A(\tau_0). \quad (35)$$

We obtain another suitable continuation relation for amplitudes along the ray by introducing a special local Cartesian coordinate system y_1, y_2, y_3 , varying along the ray. We call it the *wavefront orthonormal coordinate system*. At any point of the ray specified by $y_3 = \tau$, the y_3 axis is parallel to slowness vector $\mathbf{p}$, and the y_1, y_2 axes are confined to the plane tangential to the wavefront at $y_3 = \tau$. Axes y_1 and y_2 are mutually perpendicular. If we denote the 3×3 transformation matrix from ray coordinates to wavefront orthonormal coordinates by $\hat{\mathbf{Q}}^{(y)}$, then

$$\det \hat{\mathbf{Q}}(\tau) = \det \hat{\mathbf{Q}}^{(y)}(\tau) = C(\tau) \det \mathbf{Q}^{(y)}(\tau). \quad (36)$$

Here $C(\tau)$ is the phase velocity, and $\mathbf{Q}^{(y)}(\tau)$ is the 2×2 upper-left submatrix of $\hat{\mathbf{Q}}^{(y)}(\tau)$. Using Eq. 36 in Eq. 35, we obtain the continuation relation in an alternative form:

$$A(\tau) = \left[\frac{\rho(\tau_0) C(\tau_0) \det \mathbf{Q}^{(y)}(\tau_0)}{\rho(\tau) C(\tau) \det \mathbf{Q}^{(y)}(\tau)} \right]^{1/2} A(\tau_0). \quad (37)$$

An important property of the continuation relation presented in Eq. 37 is that $\det \mathbf{Q}^{(y)}(\tau)$ is uniquely determined by coordinates y_1 and y_2 , confined to the plane tangential to the wavefront at τ . Thus, Eq. 37 remains valid for any coordinate systems q_i (even nonorthogonal), in which the coordinate axes q_1 and q_2 are orthogonal in the plane tangential to the

wavefront, but the axis q_3 is taken in a different way than y_3 , for example, along the ray. This is, e.g., the case of the well-known *ray-centered coordinate system* q_1, q_2, q_3 . We have $\det \mathbf{Q}^{(q)}(\tau) = \det \mathbf{Q}^{(y)}(\tau)$, where $\mathbf{Q}^{(q)}$ is the 2×2 transformation matrix from ray (y_1, y_2) to ray-centered (q_1, q_2) coordinates.

Transport equations for P and S waves in isotropic media may be also expressed in the form of Eq. 32. The expression is straightforward for P waves. For S waves, transport equations for scalar amplitudes B and C in Eq. 20 are generally coupled. They decouple only if the unit vectors $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ in Eq. 20 satisfy the following relation along the ray:

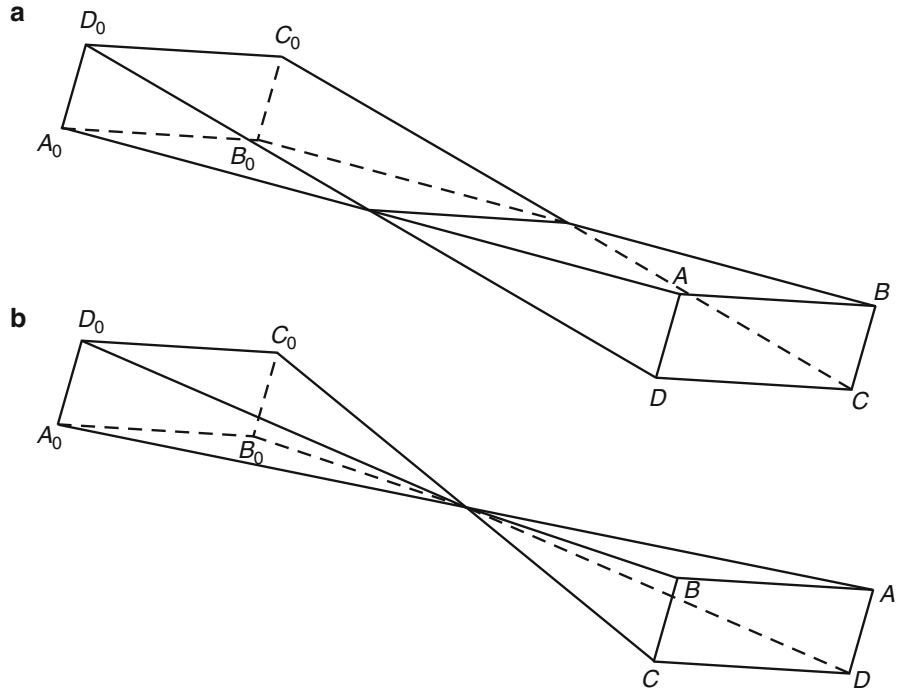
$$d\mathbf{g}^{(M)} / d\tau = (\mathbf{g}^{(M)} \cdot \nabla \beta) \mathbf{n}, \quad M = 1, 2. \quad (38)$$

In the terminology of the Riemannian geometry, vector $\mathbf{g}^{(M)}$ satisfying Eq. 38 is transported parallelly along the ray. If $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ are chosen as mutually perpendicular and perpendicular to $\mathbf{n}$ at one point of the ray, Eq. 38 guarantees that they have these properties at any point of the ray. Consequently, $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ are always perpendicular to the ray and do not rotate around it as the S wave progresses. As $\mathbf{g}^{(1)}$, $\mathbf{g}^{(2)}$, and $\mathbf{n}$ are always orthonormal, and $\mathbf{n}$ is known at any point of the ray, it is not necessary to use Eq. 38 to compute both vectors $\mathbf{g}^{(M)}$. One of them can be determined from the orthonormality condition, once the other has been computed using Eq. 38.

Quantity $\det \hat{\mathbf{Q}}(\tau)$ in Eq. 35 may become zero at some point $\tau = \tau^C$. This means that the cross-sectional area of the ray tube shrinks to zero at $\tau = \tau^C$. The relevant point $\tau = \tau^C$ of the ray is called the *caustic point*. At the caustic point, the ray solution is singular and yields an infinite amplitude there. In passing through the caustic point τ^C along the ray, the argument of $[\det \hat{\mathbf{Q}}(\tau)]^{1/2}$ may change by $\pm\pi/2$ or $\pm\pi$ (Kravtsov and Orlov 1999). The former case corresponds to the caustic point of the first order, see Fig. 3a, during which the ray tube shrinks to an elementary arc, the latter case corresponds to the caustic point of the second order, see Fig. 3b, during which the ray tube shrinks to a point. If the caustic point τ^C is situated between τ_0 and τ , it is common to introduce the *phase shift* $T^C(\tau, \tau_0)$ due to caustic using the relation:

$$\left[\frac{\det \hat{\mathbf{Q}}(\tau_0)}{\det \hat{\mathbf{Q}}(\tau)} \right]^{1/2} = \left| \frac{\det \hat{\mathbf{Q}}(\tau_0)}{\det \hat{\mathbf{Q}}(\tau)} \right|^{1/2} \exp [iT^C(\tau, \tau_0)]. \quad (39)$$

The phase shift due to the caustic is cumulative. If the ray passes through several caustic points along the ray between τ_0 and τ , the phase shift due to caustics is the sum of the individual phase shifts. For the choice of the exponential factor $\exp(-i\omega t)$ in Eq. 4, it has the form $T^C(\tau, \tau_0) = -\frac{1}{2}\pi k(\tau, \tau_0)$, where $k(\tau, \tau_0)$ is an integer, called the KMAH index (to acknowledge the work by Keller, Maslov, Arnold and Hörmander in this field). The continuation relation for ray-theory amplitudes given in Eq. 35 can then be modified to read:

Seismic Ray Theory,**Fig. 3** Caustic points of (a) the first order and (b) second order (Fig. 3.13 of Červený 2001)

$$A(\tau) = \left[\frac{\rho(\tau_0) |\det \hat{\mathbf{Q}}(\tau_0)|}{\rho(\tau) |\det \hat{\mathbf{Q}}(\tau)|} \right]^{1/2} \exp [iT^C(\tau, \tau_0)] A(\tau_0). \quad (40)$$

Equation 37 can be transformed to the form analogous to Eq. 40 as the zeros of $\det \hat{\mathbf{Q}}^{(v)}(\tau)$ are situated at the same points τ^C on the ray as the zeros of $\det \hat{\mathbf{Q}}(\tau)$.

The KMAH index can be calculated along the ray as a byproduct of dynamic ray tracing. For detailed derivations and discussion see Bakker (1998) and Klimeš (2010).

There are some differences between the KMAH indices along the rays in isotropic and anisotropic media. In isotropic media, the KMAH index always increases when the ray passes through a new caustic point, either by one or two. In anisotropic media, however, it may also decrease by one or two at some caustic points. This happens only for S waves as a consequence of the concave form of the slowness surface of the corresponding S wave.

Alternative expressions for ray-theory amplitudes and their mutual relations can be found in Klimeš (2019).

Dynamic Ray Tracing. Paraxial Approximations

As we can see in Eq. 35, the computation of the ray-theory amplitudes requires knowledge of $\det \hat{\mathbf{Q}}$, where $\hat{\mathbf{Q}}(\tau)$ characterizes the properties of the ray field in the vicinity of the ray under consideration. $\hat{\mathbf{Q}}(\tau)$ can be computed by the procedure called *dynamic* (or *paraxial*) *ray tracing*. In addition to $\hat{\mathbf{Q}}(\tau)$

with elements $Q_{ij}(\tau) = \partial x_i / \partial \gamma_j$, we also have to introduce a new 3×3 matrix $\hat{\mathbf{P}}(\tau)$ with elements $P_{ij}(\tau) = \partial p_i / \partial \gamma_j$. The equation for P_{ij} must be included to obtain the linear dynamic ray tracing system. Differentiating ray tracing system shown in Eq. 25 with respect to γ_j , we can easily obtain a system of linear ordinary differential equations of the first order for Q_{ij} and P_{ij} ,

$$\begin{aligned} \frac{dQ_{ij}}{d\tau} &= \frac{\partial^2 \mathcal{H}}{\partial p_i \partial x_k} Q_{kj} + \frac{\partial^2 \mathcal{H}}{\partial p_i \partial p_k} P_{kj}, \\ \frac{dP_{ij}}{d\tau} &= -\frac{\partial^2 \mathcal{H}}{\partial x_i \partial x_k} Q_{kj} - \frac{\partial^2 \mathcal{H}}{\partial x_i \partial p_k} P_{kj}, \end{aligned} \quad (41)$$

see Červený (1972). This system is usually called the *dynamic ray tracing system*, and the relevant procedure dynamic ray tracing. It can be solved along a given ray Ω , or together with ray tracing system, see Eq. 25.

The dynamic ray tracing system shown in Eq. 41 may be expressed in various forms. Instead of Cartesian coordinates x_i , we can use the wavefront orthonormal coordinates y_i , or the ray-centered coordinates q_i . Then, instead of the 3×3 matrices $\hat{\mathbf{Q}}$ and $\hat{\mathbf{P}}$, it is sufficient to seek the 2×2 matrices $\mathbf{Q}^{(v)}$, $\mathbf{P}^{(v)}$, or $\mathbf{Q}^{(q)}$, $\mathbf{P}^{(q)}$. This reduces the number of DRT equations, but complicates their right-hand sides (Červený 2001, Sect. 4.2).

As the dynamic ray tracing system presented in Eq. 41 is of the first order and linear, we can compute its fundamental matrix consisting of six linearly independent solutions. The 6×6 fundamental matrix of the dynamic ray tracing system in Eq. 41, specified by the 6×6 identity matrix at an arbitrary

point $\tau = \tau_0$ of the ray is called the *ray propagator matrix* and denoted by $\mathbf{\Pi}(\tau, \tau_0)$.

The 6×6 ray propagator matrix $\mathbf{\Pi}(\tau, \tau_0)$ is symplectic:

$$\mathbf{\Pi}^T(\tau, \tau_0) \mathbf{J} \mathbf{\Pi}(\tau, \tau_0) = \mathbf{J}, \quad \text{with} \quad \mathbf{J} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{pmatrix}. \quad (42)$$

If we know the matrices $\hat{\mathbf{Q}}(\tau_0)$, $\hat{\mathbf{P}}(\tau_0)$, we can compute $\hat{\mathbf{Q}}(\tau)$, $\hat{\mathbf{P}}(\tau)$ at any point τ of the ray by a simple matrix multiplication

$$\begin{pmatrix} \hat{\mathbf{Q}}(\tau) \\ \hat{\mathbf{P}}(\tau) \end{pmatrix} = \mathbf{\Pi}(\tau, \tau_0) \begin{pmatrix} \hat{\mathbf{Q}}(\tau_0) \\ \hat{\mathbf{P}}(\tau_0) \end{pmatrix}. \quad (43)$$

The ray propagator matrix $\mathbf{\Pi}(\tau, \tau_0)$ satisfies the chain rule, $\mathbf{\Pi}(\tau, \tau_0) = \mathbf{\Pi}(\tau, \tau_1) \mathbf{\Pi}(\tau_1, \tau_0)$, where point τ_1 is situated arbitrarily on the ray. It is simple to compute the inverse of $\mathbf{\Pi}(\tau, \tau_0)$: $\mathbf{\Pi}^{-1}(\tau, \tau_0) = \mathbf{\Pi}(\tau_0, \tau)$. We can express $\mathbf{\Pi}(\tau, \tau_0)$ in the following way:

$$\mathbf{\Pi}(\tau, \tau_0) = \begin{pmatrix} \hat{\mathbf{Q}}_1(\tau, \tau_0) & \hat{\mathbf{Q}}_2(\tau, \tau_0) \\ \hat{\mathbf{P}}_1(\tau, \tau_0) & \hat{\mathbf{P}}_2(\tau, \tau_0) \end{pmatrix}, \quad (44)$$

where $\hat{\mathbf{Q}}_1(\tau, \tau_0)$, $\hat{\mathbf{Q}}_2(\tau, \tau_0)$, $\hat{\mathbf{P}}_1(\tau, \tau_0)$, and $\hat{\mathbf{P}}_2(\tau, \tau_0)$ are 3×3 matrices.

Equation 43 can be used to obtain a very important quantity – the 3×3 matrix $\hat{\mathbf{M}}(\tau)$ of second derivatives of the travel-time field with respect to Cartesian coordinates, with elements $M_{ij} = \partial^2 T / \partial x_i \partial x_j$:

$$\hat{\mathbf{M}}(\tau) = \hat{\mathbf{P}}(\tau) (\hat{\mathbf{Q}}(\tau))^{-1}. \quad (45)$$

Matrix $\hat{\mathbf{M}}(\tau)$ plays an important role in the computation of travel time not only along the ray Ω but also in its “quadratic” paraxial vicinity:

$$T(\mathbf{x}) = T(\mathbf{x}^\Omega) + (\mathbf{x} - \mathbf{x}^\Omega)^T \mathbf{p}(\tau) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^\Omega)^T \hat{\mathbf{M}}(\tau) (\mathbf{x} - \mathbf{x}^\Omega). \quad (46)$$

In Eq. 46, $\mathbf{x}$ denotes an arbitrary point in the paraxial vicinity of the ray Ω , close to point $\mathbf{x}^\Omega = \mathbf{x}^\Omega(\tau)$ on the ray Ω ; slowness vector $\mathbf{p}(\tau)$ and the matrix $\hat{\mathbf{M}}(\tau)$ are given at $\mathbf{x}^\Omega$. The possibility of computing the travel time in the paraxial vicinity of the ray has many important applications.

The properties of the 6×6 ray propagator matrix $\mathbf{\Pi}(\tau, \tau_0)$ described above are also valid for the 4×4 ray propagator matrices $\mathbf{\Pi}^{(v)}(\tau, \tau_0)$ or $\mathbf{\Pi}^{(q)}(\tau, \tau_0)$ expressed in wavefront orthonormal coordinates y_i or ray-centered coordinates q_i . The ray propagator matrices $\mathbf{\Pi}^{(v)}(\tau, \tau_0)$ and $\mathbf{\Pi}^{(q)}(\tau, \tau_0)$

are identical, therefore, they can be expressed in terms of the same 2×2 matrices $\mathbf{Q}_1(\tau, \tau_0)$, $\mathbf{Q}_2(\tau, \tau_0)$, $\mathbf{P}_1(\tau, \tau_0)$, and $\mathbf{P}_2(\tau, \tau_0)$. Matrices $\mathbf{Q}_1(\tau, \tau_0)$, $\mathbf{P}_1(\tau, \tau_0)$ correspond to the plane-wavefront initial conditions at τ_0 , and matrices $\mathbf{Q}_2(\tau, \tau_0)$, $\mathbf{P}_2(\tau, \tau_0)$ to the point-source initial conditions at τ_0 , see Fig. 4. The 2×2 matrix $\mathbf{Q}_2(\tau, \tau_0)$ plays an important role in computing the ray-theory Green function. The quantity

$$\mathcal{L}(\tau, \tau_0) = |\det \mathbf{Q}_2(\tau, \tau_0)|^{1/2} \quad (47)$$

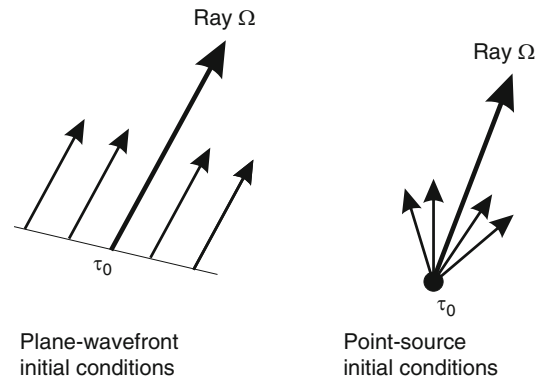
is called the *relative geometrical spreading*.

As in Eq. 45, we can define the 2×2 matrix of the second derivatives of the travel-time field with respect to y_1, y_2 or q_1, q_2 as follows:

$$\mathbf{M}(\tau) = \mathbf{P}(\tau) (\mathbf{Q}(\tau))^{-1}. \quad (48)$$

Let us briefly summarize several useful ray-theory quantities and applications, which rely fully or partly on dynamic ray tracing. For derivations and more detailed explanations, see Červený (2001, Chap. 4), where also many other applications and references can be found: (1) Paraxial travel times. (2) Paraxial slowness vectors. (3) Paraxial rays. (4) Curvature of the wavefront. (5) Matrix of geometrical spreading $\hat{\mathbf{Q}}$ and the relevant matrix $\hat{\mathbf{P}}$. (6) Continuation relations for ray-theory amplitudes along the ray. (7) Relative geometrical spreading. (8) Phase shift due to caustics. (9) Ray-theory elastodynamic Green function. (10) Higher-order spatial derivatives of the travel-time field. (11) Fresnel volumes and Fresnel zones. (12) Surface-to-surface propagator matrix. (13) Boundary-value problems in four-parametric system of paraxial rays, including two-point ray tracing. (14) Factorization of the geometrical spreading.

Dynamic ray tracing is also required in the computation of the two-point paraxial travel times in inhomogeneous, isotropic or anisotropic media (bin Waheed et al. 2013). This approach offers an efficient way of the determination of travel times between sources and receivers situated close to a known



Seismic Ray Theory, Fig. 4 Plane-wavefront and point-source initial conditions for dynamic ray tracing. In anisotropic media, rays are not perpendicular to the wavefront

reference ray and specified in Cartesian coordinates. Dynamic ray tracing is also needed in the investigation of chaotic rays and in computations of Lyapunov exponents (Červený et al. 2007), in ray perturbation methods and in modifications and extensions of the ray method such as coupling ray method, Maslov method, Gaussian beam and Gaussian packet summation methods, in Kirchhoff-Helmholtz method, and in various diffraction problems.

Coupling Ray Theory for S Waves in Anisotropic Media

In inhomogeneous weakly anisotropic media, the standard ray theory described above yields distorted results since it is unable to describe the coupling of S1 and S2 waves propagating with approximately equal phase velocities. This problem can be removed by using the coupling ray theory. In the coupling ray theory, the amplitudes of the two S waves can be computed along a trajectory called the *common ray* (Bakker 2002; Klimeš 2006). The closer the common ray approximates actual S wave rays, the more accurate results the coupling ray theory yields. The common rays can be constructed in a reference isotropic medium or in the actual anisotropic medium. A convenient option is to compute common rays using ray-tracing equations given in Eq. 25 with the Hamiltonian given as

$$\mathcal{H}(x_i, p_j) = \frac{1}{4} [G_1(x_i, p_j) + G_2(x_i, p_j)]. \quad (49)$$

In Eq. 49, G_1 and G_2 are eigenvalues of the generalized Christoffel matrix given in Eq. 7, corresponding to S1 and S2 waves.

The coupling ray theory solution is sought in the form (Coates and Chapman 1990; Bulant and Klimeš 2002):

$$\mathbf{u}(\tau, t) = A(\tau) \left[r_1(\tau) \mathbf{g}^{(1)}(\tau) \exp(i\omega\tau_1) + r_2(\tau) \mathbf{g}^{(2)}(\tau) \exp(i\omega\tau_2) \right] \exp(-i\omega t). \quad (50)$$

Here, $A(\tau)$ is the scalar amplitude given in Eqs. 35 or 37 calculated along the common ray. The symbols $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ denote the S wave eigenvectors of the generalized Christoffel matrix $\Gamma(x_i, p_j)$ calculated along the common ray. The travel times τ_1 and τ_2 are travel times corresponding to the above vectors $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$. They can be obtained by quadratures along the common ray:

$$d\tau_1/d\tau = \left[\Gamma_{ik} g_i^{(1)} g_k^{(1)} \right]^{-1/2}, \quad d\tau_2/d\tau = \left[\Gamma_{ik} g_i^{(2)} g_k^{(2)} \right]^{-1/2}. \quad (51)$$

The amplitude factors r_1 and r_2 are solutions of two coupled ordinary differential equations (Coates and Chapman 1990):

$$\begin{pmatrix} dr_1/d\tau \\ dr_2/d\tau \end{pmatrix} = \begin{pmatrix} 0 & \exp(i\omega[\tau_2(\tau) - \tau_1(\tau)]) \\ -\exp(i\omega[\tau_1(\tau) - \tau_2(\tau)]) & 0 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}, \quad (52)$$

where the angular velocity $d\varphi/d\tau$ of the rotation of the eigenvectors $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ is given by

$$\frac{d\varphi}{d\tau} = \mathbf{g}^{(2)} \frac{d\mathbf{g}^{(1)}}{d\tau} = -\mathbf{g}^{(1)} \frac{d\mathbf{g}^{(2)}}{d\tau}. \quad (53)$$

For detailed description of the algorithm, see Bulant and Klimeš (2002), for its generalization to the prevailing-frequency approximation of the coupling ray theory, which is effectively independent of frequency, see Klimeš and Bulant (2016). It allows to treat the coupling ray-theory results as standard ray-theory results because it avoids repeated calculations for varying frequencies.

There are many possible modifications and approximations of the coupling ray theory. In some of them, the amplitude vector $\mathbf{U}$ of coupled S waves is sought along the common ray in the form presented in Eq. 20, in which the amplitude factors B and C can be expressed as

$$B(\tau) = A(\tau) \mathcal{B}(\tau) \quad C(\tau) = A(\tau) \mathcal{C}(\tau). \quad (54)$$

In Eq. 54, $A(\tau)$ is again the scalar ray amplitude given in Eqs. 35 or 37 calculated along the common S-wave ray. There are many ways how to evaluate factors $\mathcal{B}$ and $\mathcal{C}$ (Kravtsov 1968; Červený et al. 2007). Here we present a combination of coupling ray theory and of the first-order ray tracing (Farra and Pšenčík 2010, see also section on “[Ray Perturbation Methods](#)”). In the approximation of Farra and Pšenčík (2010), the common ray is obtained as the first-order ray, see section on “[Ray Perturbation Methods](#)”. The vectors $\mathbf{g}^{(K)}$, appearing in Eq. 20, specify the first-order approximation of the S wave polarization plane. The factors $\mathcal{B}$ and $\mathcal{C}$ in Eq. 54 are then obtained as a solution of two coupled ordinary differential equations, which result from the corresponding two coupled transport equations:

$$\begin{pmatrix} d\mathcal{B}/d\tau \\ d\mathcal{C}/d\tau \end{pmatrix} = -\frac{i\omega}{2} \begin{pmatrix} \mathcal{M}_{11} - 1 & \mathcal{M}_{12} \\ \mathcal{M}_{12} & \mathcal{M}_{22} - 1 \end{pmatrix} \begin{pmatrix} \mathcal{B} \\ \mathcal{C} \end{pmatrix}. \quad (55)$$

Evaluation of the matrix $\mathcal{M}$ with elements $\mathcal{M}_{IJ}$ is simple, see Eqs. 7 and 20 of Farra and Pšenčík (2010).

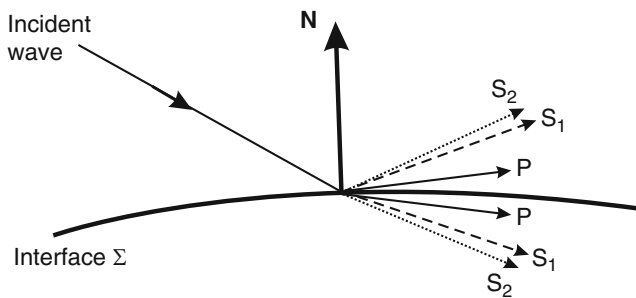
The resulting equations reduce to standard ray-theory equations in inhomogeneous isotropic media, they describe properly S wave coupling in inhomogeneous weakly anisotropic media and even yield separate S waves when anisotropy is stronger. Common S wave rays are regular

everywhere. They do not suffer from problems well known from tracing rays of individual S waves in anisotropic media and are suitable for investigating shear-wave splitting, see Pšenčík et al. (2012).

Effects of Structural Interfaces

Assume that the ray is incident on a curved structural interface. If we wish to continue the ray computations for the reflected, transmitted, monotypic, or converted waves, see Fig. 5, we have to use relevant transformation relations for the ray tracing system, dynamic ray tracing system, and for the ray theory amplitudes at the interface.

The transformation relations for ray tracing and dynamic ray tracing systems at interfaces are well known (Červený 2001). For the sake of brevity, we do not present them here. We shall, however, discuss the transformation of amplitudes. In the zero-order ray approximation, the transformation of ray-theory amplitudes across an interface is described by plane-wave reflection/transmission coefficients. In other words, amplitudes of generated waves do not depend on the curvature of the wavefront of the incident wave and the curvature of the interface at the point of incidence Q . Neither do they depend on the gradients of the density and on gradients of the density-normalized elastic moduli at Q , on both sides of the interface. They depend only on the local values of the density and density-normalized elastic moduli at Q (on both sides of the interface) and on the angle of incidence (the acute angle between the slowness vector of the incident wave and the normal $\mathbf{N}$ to the interface at the point of incidence Q).



Seismic Ray Theory, Fig. 5 Slowness vectors of P, S1, and S2 waves generated at the point of incidence Q of a curved interface separating two inhomogeneous anisotropic media. All slowness vectors at Q are situated in the plane of incidence specified by the slowness vector of the incident wave and the normal to the interface $\mathbf{N}$ at Q . Ray-velocity vectors (tangent to rays) of individual waves at Q are, in general, not confined to the plane of incidence. In isotropic media, instead of reflected and transmitted S1 and S2 waves, single reflected and transmitted S waves are generated. Ray-velocity vectors of individual waves at Q are situated in the plane of incidence

Various types of R/T coefficients may be used. The displacement R/T coefficients are used most frequently (Aki and Richards 2002; Červený et al. 1977 for isotropic media; Fedorov 1968 for anisotropic media). Very useful are the energy R/T coefficients, as they are reciprocal. The relation between the energy R/T coefficient $\mathcal{R}(Q)$ and the displacement R/T coefficient $R(Q)$ is as follows:

$$\mathcal{R}(Q) = R(Q) \left[\frac{\rho(\tilde{Q}) u_n(\tilde{Q})}{\rho(Q) u_n(Q)} \right]^{1/2} \quad (56)$$

(Červený 2001, Sect. 5.4.3). Here Q is the point of incidence, and $\tilde{Q}$ the relevant initial point of the R/T wave, both points being, of course, identical. u_n is the normal component (perpendicular to the interface) of the ray-velocity vector. We further introduce the complete energy R/T coefficients $\mathcal{R}^C$ along the ray using the relation

$$\mathcal{R}^C = \prod_{k=1}^N \mathcal{R}(Q_k). \quad (57)$$

The complete energy R/T coefficient $\mathcal{R}^C$ corresponds to the ray, which interacts N -times with interfaces (at points of incidence $Q_1, Q_2, \dots, Q_N$) between the initial and end point of the ray.

Generalization of the continuation relation presented in Eq. 37 for the ray-theory amplitudes along the ray situated in a laterally varying anisotropic medium containing curved interfaces then reads:

$$A(\tau) = \left[\frac{\rho(\tau_0) C(\tau_0) |\det \mathbf{Q}^{(y)}(\tau_0)|}{\rho(\tau) C(\tau) |\det \mathbf{Q}^{(y)}(\tau)|} \right]^{1/2} \times \mathcal{R}^C \exp [i\Gamma^C(\tau, \tau_0)] A(\tau_0). \quad (58)$$

In seismic prospecting, in the technique called *amplitude variation with offset* (AVO), it is common to work with the so-called *weak-contrast R/T coefficients*. They are linearized versions of exact R/T displacement coefficients. Linearization is mostly made with respect to the contrasts of the density and elastic moduli across the interface, often also under the assumption of weak anisotropy. There is a great variety of linearized formulae depending on the type of media surrounding the interface (isotropic, anisotropic), strength of anisotropy (weak, strong), etc. The coefficients yield reasonable approximation in the vicinity of normal incidence. For increasing incidence angles, their accuracy decreases. The advantage of the weak-contrast coefficients is their simplicity and the possibility of expressing them in explicit form. The effects of the individual medium parameters on the coefficients can then be easily evaluated.

Ray-Theory Elastodynamic Green Function

The elastodynamic Green function $G_{in}(R, t, S, t_0)$ represents the i -th Cartesian component of the displacement vector at location R and time t , due to a single-force point source situated at location S and oriented along the n -th Cartesian axis, with the time dependence $\delta(t - t_0)$. We introduce quite analogously the ray-theory elastodynamic Green function, with only two differences. The first difference is that ray-theory Green function is defined as a sum of *elementary ray-theory Green functions* computed along rays of selected elementary waves (direct, multiply reflected/transmitted, etc.). The second difference is that the elementary ray-theory Green functions are not exact, but only zero-order ray approximations.

In the frequency domain, the elementary ray-theory elastodynamic Green function $G_{in}(R, S, \omega)$ for $t_0 = 0$ reads:

$$G_{in}(R, S, \omega) = \frac{g_n(S)g_i(R) \exp [iT^C(R, S) + i\omega T(R, S)]}{4\pi[\rho(S)\rho(R)C(S)C(R)]^{1/2}\mathcal{L}(R, S)} \mathcal{R}^C. \quad (59)$$

Here $\mathcal{L}(R, S)$ is the relative geometrical spreading, given in Eq. 47, $g_i(R)$ and $g_n(S)$ are the eigenvectors of the generalized Christoffel matrix at R and S (polarization vectors corresponding to the considered elementary wave), $T(R, S)$ is the travel time along the ray from S to R , $\mathcal{R}^C$ is the complete energy R/T coefficient resulting from interactions of the wave under consideration with interfaces between S and R , and $T^C(R, S)$ is the complete phase shift due to caustics along the ray between S and R . The relevant KMAH index in anisotropic media may also include a contribution at a point source S (if the slowness surface of the considered wave is concave at S). In isotropic media, this contribution is always zero.

The complete energy R/T coefficient $\mathcal{R}^C$, the travel time $T(R, S)$, the relative geometrical spreading $\mathcal{L}(R, S)$, and the complete phase shift due to caustics are always reciprocal. Consequently, the elementary ray-theory elastodynamic Green function satisfies a very important property of reciprocity:

$$G_{in}(R, S, \omega) = G_{ni}(S, R, \omega). \quad (60)$$

This relation is valid for any elementary seismic body wave generated by a point source.

Chaotic Rays. Lyapunov Exponents

In homogeneous media, geometrical spreading increases linearly with increasing length of the ray. In heterogeneous

media, behavior of geometrical spreading is more complicated and depends considerably on the degree of heterogeneity of the medium. In models, in which the heterogeneity exceeds certain degree, average geometrical spreading increases exponentially with increasing length of the ray. Rays in such a medium often exhibit chaotic behavior, which is characterized by a strong sensitivity of rays to the initial values (for example, to ray parameters). The rays with only slightly differing specification of their initial values tend to diverge *exponentially* at large distances from the initial point. Consequently, the rays intersect many times and many rays pass through the same point. With such *chaotic rays*, two-point ray tracing is practically impossible, and the ray tubes are not narrow enough for travel time interpolation. The chaotic behavior of rays increases with increasing length of rays and prevents applicability of the ray theory.

The exponential divergence of chaotic rays in the phase space (the space formed by spatial coordinates x_i and slowness-vector components p_j) can be quantified by the so-called *Lyapunov exponents*. They may be introduced in several ways. It is common to express them in terms of characteristic values of the ray propagator matrix. The relevant expressions for the Lyapunov exponents and several numerical examples for 2D models without interfaces can be found in Klimeš (2002a). See also Červený et al. (2007), where other references can be found.

The estimate of the Lyapunov exponent of a single finite ray depends on its position and direction. The Lyapunov exponents associated with rays of different positions and directions can be used to calculate *average Lyapunov exponents* for the model. The average Lyapunov exponents play a very important role in smoothing the models to make them suitable for ray tracing (Červený et al. 2007).

Ray Perturbation Methods

Ray perturbation methods represent an important part of the ray theory. They can be used for approximate but fast and transparent solutions of forward problems in complicated models. They play even more important role in the inverse problems.

Ray perturbation methods are useful everywhere, where we wish to compute the wavefield or its constituents (travel times, amplitudes, and polarization) in complicated models, which deviate only little from *reference models*, for which computations are simpler. The solutions for complicated models are then sought as perturbations of simple solutions for the reference models. Examples are computations in weakly anisotropic media, which use an isotropic medium as a reference, or in weakly dissipative media, which use a perfectly elastic medium as a reference. Basic role in these approaches is played by *reference rays* traced in reference

media. Solutions in perturbed media can be given in the form of a power series in the deviations of the perturbed and reference models. Mostly, the *first-order approximation*, i.e., the first term of the power series, is used.

The most frequent application of ray perturbation methods is in travel-time computations. First-order travel-time perturbation formulae for isotropic media are known and have been used (mostly in tomography) for several decades. Well known and broadly applied are also first-order travel-time formulae for anisotropic media (Hanyga 1982; Červený 2001, Sect. 3.9). Travel-time perturbations are obtained by quadratures along reference rays. As integration parameters, the parameters along reference rays are used.

Another application of the perturbation method to the computation of travel times is the use of the so-called weak-anisotropy approximation to evaluate approximately reflection moveout. Reflection moveout is commonly calculated as a Taylor series expansion of the squared reflection travel time (for example, Tsvankin and Grechka 2011), which loses rapidly accuracy with increasing offset. Weak-anisotropy approximation of the moveout formulae is free of this limitation. The moveout formulae are available for unconverted or converted waves, for horizontal or dipping reflectors, for anisotropy of arbitrary symmetry. For example, formula for P wave reflection moveout in a homogeneous layer of tilted orthorhombic symmetry can be found in Farra and Pšenčík (2017).

Several procedures for computation of higher-order travel-time perturbations for weakly anisotropic media (note that anisotropy of the Earth is mostly weak) also exist. The procedure based on the so-called perturbation Hamiltonians (Klimeš 2002b; Červený et al. 2007) allows computation of highly accurate travel times along a fixed reference ray in a reference medium. Another procedure is based on the so-called first-order ray tracing described briefly below. In the latter method, second-order travel-time perturbations can be calculated along first-order rays.

Perturbation methods are also used in *first-order ray tracing* and *first-order dynamic ray tracing* for P and S waves in inhomogeneous, weakly anisotropic media (Pšenčík and Farra 2007; Farra and Pšenčík 2010). They allow to compute, approximately, not only rays and travel times but whole wavefields. To derive first-order ray tracing and dynamic ray tracing, the perturbation approach is used, in which deviations of anisotropy from isotropy are considered to be small. Then it is just sufficient to use Eqs. 25 and 41 with Eqs. 21 or 49, in which the exact eigenvalues G_m are replaced by their first-order approximations. The resulting ray tracing provides first-order rays, first-order travel times, and the first-order geometrical spreading. By simple quadratures along first-order rays, second-order travel-time corrections can be computed. This approach is applicable to P and S waves. In case of S waves, it can include the computation of coupling effects. First-order ray tracing and dynamic ray tracing are used in this

case for computing common S wave rays, first-order travel times and geometrical spreading along them. The wavefield of S waves is obtained by solving second-order coupling equations along the S wave common rays. The procedure yields standard ray-theory results for S waves propagating in isotropic media, and approximate results in anisotropic media when the S waves are coupled or even decoupled.

Ray Perturbation Method for Weakly Dissipative Media

Study of wave propagation in dissipative media can be often simplified by the use of the *correspondence principle* (Borchardt 2009; Carcione 2014). When applicable, it allows to obtain solutions for viscoelastic waves from solutions for elastic waves by replacing real-valued elastic parameters by their complex-valued, frequency-dependent counterparts. The density-normalized stiffness tensor $a_{ijkl}(x, \omega)$ then reads:

$$a_{ijkl} = a_{ijkl}^R - i a_{ijkl}^I. \quad (61)$$

If a_{ijkl}^I is small, the dissipative medium can be considered as a perturbation of an elastic medium (Gajewski and Pšenčík 1992; Červený 2001, Sect. 5.5.3), which has a form of the imaginary-valued term $-i a_{ijkl}^I$. Reference ray in the reference elastic medium and corresponding real-valued travel time T along the reference ray between points S and R can be obtained by standard ray tracing in the elastic medium. The travel-time perturbation ΔT (the perturbation due to the perturbation of the reference elastic medium $-i a_{ijkl}^I$), responsible for the exponential amplitude decay along the reference ray, can be obtained by quadratures along this ray:

$$\Delta T = \frac{i}{2} \int_S^R Q^{-1}(\tau) d\tau. \quad (62)$$

The quantity Q in Eq. 62 is a direction-dependent *quality factor* for anisotropic media, corresponding to the Hamiltonian given in Eq. 21:

$$Q^{-1} = a_{ijkl}^I p_j p_l g_i g_k. \quad (63)$$

For general Hamiltonians, the quality factor Q is given by the relation $Q^{-1} = -\text{Im} \mathcal{H}(x_i, p_j)$.

For causal dissipation, the stiffness tensor given in Eq. 61 is frequency dependent. The above-described perturbation approach is then equivalent to the perturbation scheme, in which $a_{ijkl}^I(x_m, \omega)$ is considered to be of the order of ω^{-1} for $\omega \rightarrow \infty$ (Kravtsov and Orlov 1990; Gajewski and Pšenčík 1992).

In an inhomogeneous isotropic, weakly dissipative medium, the expression shown in Eq. 63 reduces to the well-known formula

$$Q^{-1} = -\text{Im}V^2 / \text{Re} V^2 \doteq -2\text{Im}V / \text{Re} V, \quad (64)$$

in which V is the complex-valued velocity, $V = \alpha$ for P waves and $V = \beta$ for S waves. Complex-valued quantities α and β are generalizations (to the complex space) of real-valued α and β from Eq. 14.

The perturbation method described above uses the real ray tracing in a reference elastic medium. Vavryčuk (2012) shows that more accurate results are obtained by the procedure, which he calls the real viscoelastic ray tracing. Vavryčuk (2012) illustrates it on examples of isotropic dissipative media. Klimeš and Klimeš (2011) extend Vavryčuk's (2012) work to anisotropic dissipative media. They construct the reference Hamiltonian function using the Hamiltonian function corresponding to a given complex-valued eikonal equation. They show that the reference Hamiltonians can be constructed in various ways leading to travel time perturbations of varying accuracy.

Continuing attention is devoted to the study of effects of attenuation on reflection/transmission coefficients (Krebes and Daley 2007; Borchardt 2009; Ursin et al. 2017). There are still some open questions, mostly concerning the applicability of the correspondence principle to the reflection/transmission coefficients of overcritically reflected waves.

Concluding Remarks. Applications, Modifications, and Extensions of the Ray Method

Seismic ray method has found many applications both in global seismology and in seismic exploration. The advantages of the seismic ray method consist in its numerical efficiency, universality, conceptual clarity, and in its ability to treat various seismic body waves independently of other waves. Although its accuracy is only limited, the seismic ray method is the only method which is able to give an approximate answer to many problems of high-frequency seismic body wave propagation in laterally varying, isotropic or anisotropic, perfectly elastic or dissipative, layered and block structures.

In classical global seismology, the seismic ray method has been traditionally used to study the internal structure of the whole Earth, assuming that the Earth is radially symmetric. The standard Earth's model, obtained in this way, is expressed in terms of distribution of elastic parameters varying with depth.

At present, the applications of the seismic ray method are considerably broader. It is used to study the 3-D local laterally inhomogeneous structures, their anisotropy and attenuation,

and the form and physical properties of structural interfaces. In addition to forward modelling, the ray perturbation methods are also broadly used in inversions using observed travel times or whole waveforms. In lithospheric structural studies, particularly in crustal seismology, the ray-synthetic seismograms have been also often used for ultimate comparison with observed seismograms. The computation of ray-synthetic seismograms requires determination of not only travel times but also ray-theory amplitudes and polarizations of individual waves. Seismic ray method has also found applications in other branches of seismology. Important examples are the localization of seismic sources either in a known structure or simultaneously with structural inversion.

In most applications of the ray method in seismic exploration for oil, the use of local 3-D structures with structural curved interfaces is a necessity. Sophisticated algorithms have been developed and used to image the structures under consideration. At present, the most important role is played by migration algorithms. Seismic ray theory and its extensions have found important applications in these algorithms.

The ray method is not valid universally. We have briefly described three serious limitations of the ray method: (a) The ray method can be used only for high-frequency signals. (b) In models, in which heterogeneity of the medium exceeds certain degree, the ray field has chaotic character, particularly at large distances from the source. (c) The standard ray method cannot be used for computing S waves propagating in inhomogeneous, weakly anisotropic media. It must be replaced by the coupling ray theory. The coupling ray theory must be used in the vicinity of shear-wave singular directions.

The ray method fails, however, even in other singular situations. In smooth isotropic or anisotropic media, the most important types of singularity are *caustics*. Caustics may attain various forms. Various extensions of the ray method can be used to compute wavefields in caustic regions. These extensions are frequency dependent. See a detailed treatment of wavefields in caustic regions in Kravtsov and Orlov (1999), and also in Stamnes (1986). In models with smooth structural interfaces, other singularities often appear. For edge and vertex points, see Ayzenberg et al. (2007). For critical singular regions, at which head waves separate from reflected waves, see Červený and Ravindra (1971). For the waves, whose rays are tangential to interfaces, see Thomson (1989).

Specific methods, designed for different types of singularities may be used for computing wavefields in singular regions. Disadvantage of these methods is that they are different for different singularities. Moreover, singular regions often overlap, and the wavefield in the overlapping region requires again different treatment. It is desirable to have available a more general extension of the ray method, applicable uniformly in any of the mentioned singular regions, or, at least, in most of them. Such an extension

would simplify ray computations considerably and could even lead to more accurate results.

Several such extensions of the ray method have been proposed. We do not describe them here in detail. Instead, we merely present references, in which more details and further references can be found. Let us mention the Maslov asymptotic ray theory introduced to seismology by Chapman and Drummond (1982); see also Thomson and Chapman (1985) and Chapman (2004). Another extension of the ray method is based on the summation of Gaussian beams (Popov 1982; Červený et al. 1982; Klimeš 2018). For the relation of this method with the Maslov method see Klimeš (1984). The Gaussian beam summation method has found applications both in the forward modelling of seismic wavefields and in migrations in seismic exploration. It is closely related to the method of summation of Gaussian packets (Červený et al. 2007). Ray theory can be also used in the Born scattering theory (Chapman and Coates 1994; Chapman 2004). For waves reflected from a smooth structural interface separating two heterogeneous, isotropic or anisotropic media, the Kirchhoff surface integral method can be used. For details and many references see Chapman (2004, Sect. 10.4). Another useful extension of the ray method is the one-way wave equation approach (Thomson 1999).

Cross-References

- [Energy Partitioning of Seismic Waves](#)
- [Magnetic Anisotropy](#)
- [Seismic Diffraction](#)
- [Seismic Viscoelastic Attenuation](#)
- [Seismic Waves, Scattering](#)
- [Seismic, Migration](#)
- [Seismic, Waveform Modeling and Tomography](#)
- [Traveltime Tomography Using Controlled-Source Seismic Data](#)

Acknowledgments The authors are very grateful to Luděk Klimeš and Ravi Kumar for valuable comments and recommendations. The research was supported by the consortium project Seismic Waves in Complex 3-D Structures, by research projects 205/07/0032, 205/08/0332, and 16-05237S of the Grant Agency of the Czech Republic; and by research project MSM0021620860 of the Ministry of Education of the Czech Republic.

Bibliography

- Aki K, Richards P (2002) Quantitative seismology, 2nd edn. University Science Books, Sausalito
- Ayzenberg MA, Aizenberg AM, Helle HB, Klem-Musatov KD, Pajchel J, Ursin B (2007) 3D diffraction modeling of singly scattered acoustic wavefields based on the combination of surface integral propagators and transmission operators. *Geophysics* 72:SM19–SM34

- Babich VM (1956) Ray method of the computation of the intensity of wave fronts (in Russian). *Dokl Akad Nauk SSSR* 110:355–357
- Babich VM (1961) Ray method of the computation of the intensity of wave fronts in elastic inhomogeneous anisotropic medium. In: Petrashev GI (ed) *Problems of the dynamic theory of propagation of seismic waves* (in Russian), vol 5. Leningrad University Press, Leningrad, pp 36–46. (Translation to English: *Geophys J Int* 118:379–383, 1994)
- Babich VM, Kiselev AP (2018) *Elastic waves. High frequency theory*. CRC Press, London
- Bakker PM (1998) Phase shift at caustics along rays in anisotropic media. *Geophys J Int* 134:515–518
- Bakker PM (2002) Coupled anisotropic shear-wave ray tracing in situations where associated slowness sheets are almost tangent. *Pure Appl Geophys* 159:1403–1417
- bin Waheed U, Pšenčík I, Červený V, Iversen E, Alkhalifah T (2013) Two-point paraxial traveltimes formula for inhomogeneous isotropic and anisotropic media: tests of accuracy. *Geophysics* 78:WC65–WC80
- Borchardt RD (2009) *Viscoelastic waves in layered media*. Cambridge University Press, Cambridge
- Bulant P, Klimeš L (2002) Numerical algorithm of the coupling ray theory in weakly anisotropic media. *Pure Appl Geophys* 159:1419–1435
- Carcione JM (2014) *Wave fields in real media. Wave propagation in anisotropic, anelastic, porous and electromagnetic media*. Elsevier, Amsterdam
- Červený V (1972) Seismic rays and ray intensities in inhomogeneous anisotropic media. *Geophys J R Astron Soc* 29:1–13
- Červený V (2001) *Seismic ray theory*. Cambridge University Press, Cambridge
- Červený V, Ravindra R (1971) *Theory of seismic head waves*. Toronto University Press, Toronto
- Červený V, Molotkov IA, Pšenčík I (1977) *Ray method in seismology*. Univerzita Karlova, Praha
- Červený V, Popov MM, Pšenčík I (1982) Computation of wave fields in inhomogeneous media. Gaussian beam approach. *Geophys J R Astron Soc* 70:109–128
- Červený V, Klimeš L, Pšenčík I (1988) Complete seismic ray tracing in three-dimensional structures. In: Doornbos DJ (ed) *Seismological algorithms*. Academic Press, New York, pp 89–168
- Červený V, Klimeš L, Pšenčík I (2007) Seismic ray method: recent developments. *Adv Geophys* 48:1–126. <http://www.sciencedirect.com/science/bookseries/00652687>
- Chapman CH (2002) Seismic ray theory and finite frequency extensions. In: Lee WHK, Kanamori H, Jennings PC (eds) *International handbook of earthquake and engineering seismology, part A*. Academic Press, New York, pp 103–123
- Chapman CH (2004) *Fundamentals of seismic wave propagation*. Cambridge University Press, Cambridge
- Chapman CH, Coates RT (1994) Generalized Born scattering in anisotropic media. *Wave Motion* 19:309–341
- Chapman CH, Drummond R (1982) Body-wave seismograms in inhomogeneous media using Maslov asymptotic theory. *Bull Seismol Soc Am* 72:S277–S317
- Coates RT, Chapman CH (1990) Quasi-shear wave coupling in weakly anisotropic 3-D media. *Geophys J Int* 103:301–320
- Farra V, Pšenčík I (2010) Coupled S waves in inhomogeneous weakly anisotropic media using first-order ray tracing. *Geophys J Int* 180:405–417
- Farra V, Pšenčík I (2017) Weak-anisotropy moveout approximations for P waves in homogeneous TOR layers. *Geophysics* 82:WA23–WA32
- Fedorov FI (1968) *Theory of elastic waves in crystals*. Plenum, New York
- Gajewski D, Pšenčík I (1992) Vector wavefield for weakly attenuating anisotropic media by the ray method. *Geophysics* 57:27–38

- Gjøystdal H, Iversen E, Laurain R, Lecomte I, Vinje V, Åstebol K (2002) Review of ray theory applications in modelling and imaging of seismic data. *Stud Geophys Geod* 46:113–164
- Hanyga A (1982) The kinematic inverse problem for weakly laterally inhomogeneous anisotropic media. *Tectonophysics* 90:253–262
- Jech J, Pšenčík I (1989) First-order perturbation method for anisotropic media. *Geophys J Int* 99:369–376
- Karal FC, Keller JB (1959) Elastic wave propagation in homogeneous and inhomogeneous media. *J Acoust Soc Am* 31:694–705
- Klimeš L (1984) The relation between Gaussian beams and Maslov asymptotic theory. *Stud Geophys Geod* 28:237–247
- Klimeš L (2002a) Lyapunov exponents for 2-D ray tracing without interfaces. *Pure Appl Geophys* 159:1465–1485
- Klimeš L (2002b) Second-order and higher-order perturbations of travel time in isotropic and anisotropic media. *Stud Geophys Geod* 46:213–248
- Klimeš L (2006) Common-ray tracing and dynamic ray tracing for S waves in a smooth elastic anisotropic medium. *Stud Geophys Geod* 50:449–461
- Klimeš L (2010) Phase shift of the Green tensor due to caustics in anisotropic media. *Stud Geophys Geod* 54:269–289
- Klimeš L (2018) Superpositions of Gaussian beams and column Gaussian packets in heterogeneous anisotropic media. *Geophys J Int* 215:1739–1746
- Klimeš L (2019) Calculation of the amplitudes of elastic waves in anisotropic media in Cartesian or ray-centred coordinates. *Stud Geophys Geod* 63:229–246
- Klimeš L, Bulant P (2016) Prevailing-frequency approximation of the coupling ray theory for electromagnetic waves or elastic waves. *Stud Geophys Geod* 60:419–450
- Klimeš M, Klimeš L (2011) Perturbation expansions of complex-valued traveltime along real-valued reference rays. *Geophys J Int* 186:751–759
- Kravtsov YA (1968) “Quasiisotropic” approximation to geometrical optics. *Dokl Akad Nauk SSSR* 183(1):74–77. (in Russian)
- Kravtsov YA, Orlov YI (1990) Geometrical optics of inhomogeneous media. Springer, Heidelberg
- Kravtsov YA, Orlov YI (1999) Caustics, catastrophes and wave fields. Springer, Heidelberg
- Krebes ES, Daley PF (2007) Difficulties with computing anelastic plane-wave reflection and transmission coefficients. *Geophys J Int* 170:205–216
- Popov MM (1982) A new method of computation of wave fields using Gaussian beams. *Wave Motion* 4:85–97
- Pšenčík I, Farra V (2007) First-order P-wave ray synthetic seismograms in inhomogeneous weakly anisotropic media. *Geophys J Int* 170:1243–1252
- Pšenčík I, Farra V, Tessmer E (2012) Comparison of the FORT approximation of the coupling ray theory with the Fourier pseudospectral method. *Stud Geophys Geod* 56:35–64
- Stamnes JJ (1986) Waves in focal regions. Adam Hilger, Bristol
- Thomson CJ (1989) Corrections for grazing rays to 2-D seismic modelling. *Geophys J Int* 96:415–446
- Thomson CJ (1999) The “gap” between seismic ray theory and full wavefield extrapolation. *Geophys J Int* 137:364–380
- Thomson CJ, Chapman CH (1985) An introduction to Maslov’s asymptotic method. *Geophys J R Astron Soc* 83:143–168
- Tsvankin I, Grechka V (2011) Seismology of azimuthally anisotropic media and seismic fracture characterization. Society of Exploration Geophysicists, Tulsa
- Ursin B, Carcione JM, Gei D (2017) A physical solution for plane SH waves in anelastic media. *Geophys J Int* 209:661–671
- Vavryčuk V (2012) On numerically solving the complex eikonal equation using real ray-tracing methods: a comparison with the exact analytical solution. *Geophysics* 77:T109–T116
- Virieux J (1996) Seismic ray tracing. In: Boschi E, Ekström G, Morelli A (eds) *Seismic modelling of Earth structures*. Editrice Compositori, Bologna, pp 223–304

Seismic Reservoir Characterization

Dhananjay Kumar

Chevron, Houston, TX, USA

BP, Houston, TX, USA

Definition

Reservoir

A petroleum reservoir is a column of layers of sedimentary rocks in the subsurface containing oil and gas.

Reservoir properties

In petroleum exploration, some of the common reservoir properties of interest are thickness, rock type, porosity, permeability, fluid type, and saturation.

Reservoir

characterization

It is a method to describe the reservoir in terms of its structure and in situ rock and fluid properties in the subsurface.

Seismic

reservoir

characterization

It is the use of seismic data for 3D reservoir characterization away from the well locations.

Well tie

It is the comparison of a measured seismic response with a well log-based synthetic seismic response at a well location. Also, it is a comparison of a seismically derived property with a well log property at well locations.

Seismic

amplitude

calibration

It is a process to relate seismic amplitude to geology (often based on well log synthetics) for quantitative estimation of reservoir properties from seismic.

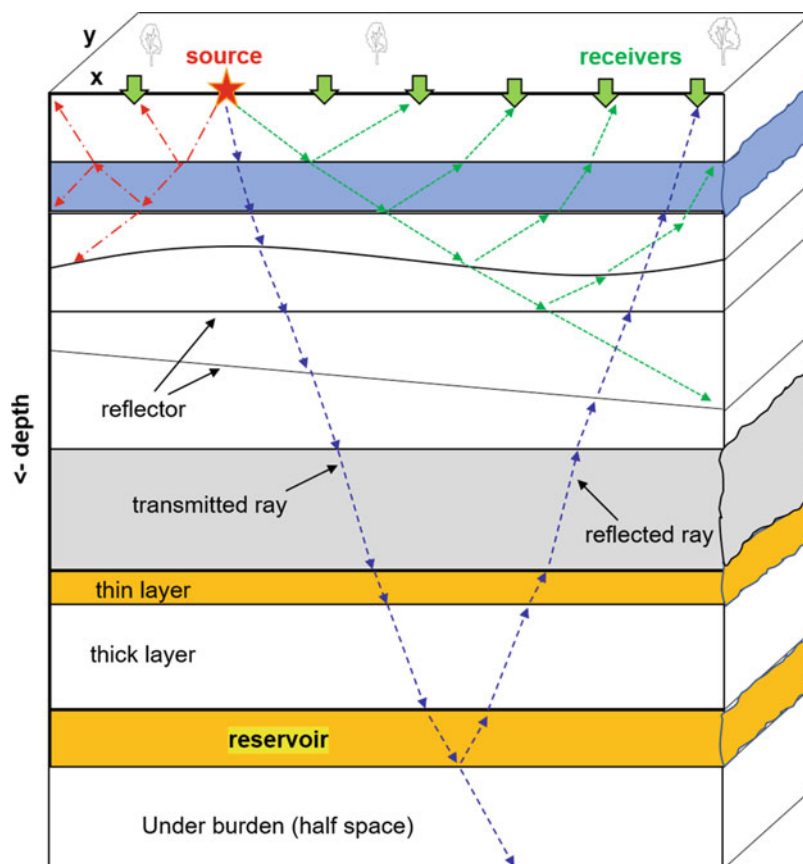
Introduction

In petroleum exploration, geoscientists want to locate and characterize reservoirs that may be a few kilometers deep in the subsurface. Seismic imaging is a noninvasive method that uses seismic data to produce three-dimensional (3D) images of the subsurface in terms of its structure, layered earth stratigraphy, and elastic properties of the rock formations – all the way from the top surface to deeper zones. However, a reservoir characterization study generally focusses on the hydrocarbon-bearing reservoir intervals of interest. To help understand how best to produce hydrocarbons from a reservoir, we want to understand the reservoir in terms of its size, shape, depth, thickness, rock type, rock strength, porosity, permeability, fluid type, saturation, pore pressure, and connectivity. Wells, including well log data, provide abundant information about the reservoir. However, wells usually must be drilled at sparse locations due to cost and safety

practicalities. Seismic data can be acquired on the surface, and the role of seismic reservoir characterization is to define the structure and interpret reservoir properties away from and between the well locations.

Seismic data is a time series representation of the earth's subsurface that contains information about the reservoir properties filtered with the source wavelet and wave propagation effects. The seismic response is a cumulative effect of wave propagation in the overburden and reservoir. Figure 1 shows a schematic plot of wave energy propagation in a subsurface model highlighting that the wave, represented by ray paths, spends most of the time in the overburden and only a short time in the reservoir. Data processing is carried out to attenuate unwanted seismic noise and enhance spatial and temporal resolution to detect finer details of the reservoir. Processed seismic volumes can be interpreted to map geological boundaries (horizons) and generate a variety of attributes for geological interpretation. For quantitative interpretation, the

seismic amplitudes are first calibrated to well log-based synthetic responses to establish relationships between the seismic waveform response and geology (reservoir properties) at well locations. While seismic reflectivity (amplitude) images have been used for decades in seismic interpretation, seismic inversion is now a standard technique used for quantitative reservoir characterization. Seismic inversion can be undertaken in two ways, either by inverting for elastic rock properties (P-wave velocity V_p , S-wave velocity V_s , impedance), from which reservoir properties (lithology, porosity, fluid saturation) can be estimated using rock physics relationships, or by inverting directly for reservoir properties. The properties estimated from both approaches are used to further estimate reservoir quality (net sand, porosity) and quantity (volumetric hydrocarbon in place), impact well-drilling decisions, and optimize hydrocarbon production and recovery. Figure 2 shows a seismic reservoir characterization workflow.



Seismic Reservoir Characterization, Fig. 1 Schematic subsurface model. An example of seismic wave propagation in the subsurface layers represented by three rays emitting from a source on the surface. Seismic receivers are placed on the top surface and record reflected waveforms in certain time-sampling intervals (i.e., 1 millisecond). Seismic wave propagates in all directions following physics (Snell's law). At each reflective interface (where contrasts in elastic properties and density are present), part of the energy is reflected, and part of the energy is transmitted, and

with each reflection and transmission, the wave may undergo a change in mode to other wave types. Here, only the primary P-P wave reflection and transmission are shown using three rays: (i) dash-dot red lines, (ii) dotted green lines, and (iii) dashed blue lines. Only the dashed blue color ray is reflecting from the base of the target reservoir, and note that this ray is spending most of the time in the overburden, and therefore its seismic response will be dominated by overburden geology and propagation effects

Seismic Data Processing

Seismic data are processed, including migration, to attenuate noise and remove 3D wave propagation effects (Yilmaz 2001). After processing, an ideal seismic response is the one responding to P-wave primary reflections only (without converted waves and multiple reflected waves) and without 3D propagation effects. Processing for the P-wave response is common, because our conventional models for interpretation are based on 1D earth models and P-P wave reflections. Advanced methods are being occasionally used (and studied) to consider the 3D and mode conversion effects. In practice, even after careful processing, seismic amplitudes can still be affected by various overburden geologic and wave propagation effects superimposed on those caused by reservoir layer properties (Sheriff 1975). All unwanted seismic responses are termed “seismic noise.” After processing, we use zero-phase seismic data and generate angle gathers, angle stacks, amplitude variation with angle (AVA) attributes, and other attributes to facilitate reservoir characterization (Chopra and Marfurt 2007).

Seismic-Well Tie

Calibration of a seismic data with synthetic data (generated using well logs) is referred to as the seismic-well tie process. Figure 3 shows an example. A seismic-well tie offers many benefits:

- (i) Provides a basis for interpreting seismic events in geologic terms
- (ii) Helps to establish a time-depth relationship between seismic data and the well depths, respectively
- (iii) Helps to estimate the phase of the data
- (iv) Helps to estimate the wavelet for use in synthetic seismogram computation and in seismic inversion
- (v) Enables general quality control of both seismic and well logs
- (vi) Enables the understanding of seismic resolution and tuning effects

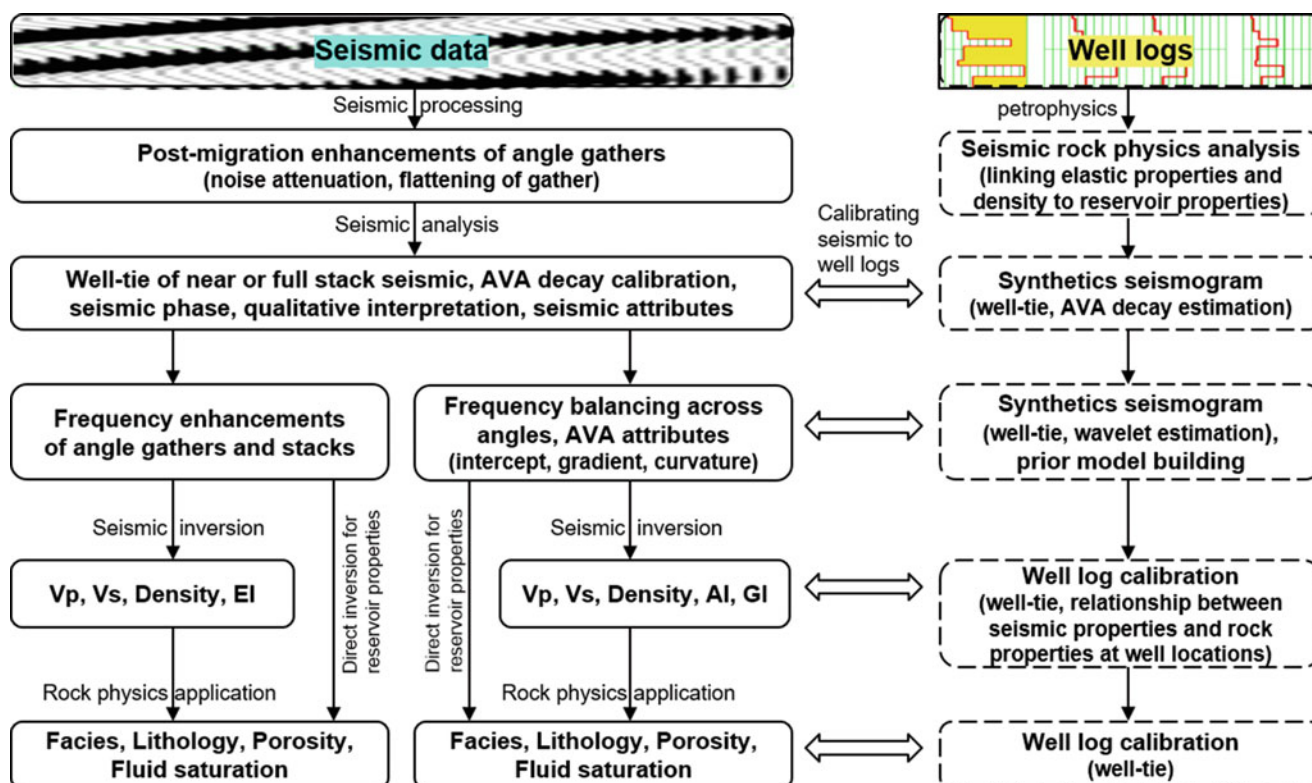
Seismic Rock Physics

Well logs are used to establish relationships between reservoir properties of interest and elastic properties that can be estimated from seismic data (Mavko et al. 1998). For example, relationships can be established to show how seismic impedance (product of velocity and density) can be used to predict reservoir porosity, or how multiple reservoir properties such as lithology and hydrocarbon saturation can be distinguished

using impedance and V_p/V_s ratio simultaneously. If reservoir properties and elastic properties are strongly related, then we have a better chance of successful reservoir characterization in the presence of good-quality seismic data. But if rock physics relationships are non-favorable, then even in the presence of good-quality seismic data, reservoir characterization will be ambiguous.

Seismic Inversion for Reservoir Properties

Seismic reflection data respond to changes in elastic properties and bulk density across interfaces between layered geologic formations. Therefore, the seismic reflectivity response must be inverted for layer properties, as they are better suited for reservoir characterization. Seismic inversion (Menke 1984; Tarantola 1986) is used either to estimate reservoir properties directly or first to estimate elastic properties followed by the application of rock physics models to estimate reservoir properties. Figure 2 shows both methods. The main objective in inversion is to estimate earth model parameters such that the difference between real seismic data and the synthetic seismic data (based on the estimated earth model, the forward modeling method, and the wavelet) is minimum. For the direct estimation of reservoir properties from seismic, rock physics models are included as part of the inversion algorithm and formulate inversion in terms of reservoir properties as model parameters. In this approach, rock physics models are used to compute elastic properties and density for the given reservoir properties (Mavko et al. 1998). Because seismic reflection data is bandlimited in temporal frequency (about 6–60 Hz), inversion using seismic data alone provides relative values of desired properties in seismic frequencies; one such method is colored inversion (Lancaster and Whitcombe 2000). Often, we want to estimate absolute values of reservoir properties, and for that, we use low-frequency (smooth) models based on available well logs and other geologic understanding including 3D seismic structural maps (Simm and Bacon 2014). Seismic models with relative values will have both negative and positive numbers, and models with absolute values will only have positive numbers. Figure 4 compares porosity predictions using two types of seismic inverted acoustic impedance (V_p times density): (i) relative impedance and (ii) absolute impedance. Various sources of noise in seismic data and our use of various models and approximations give rise to uncertainties in the estimated reservoir properties. Two main areas of research are being undertaken to help deal with uncertainties, as well as the increasing abundance of seismic data. One current focus for research is to estimate uncertainties in seismic predictions, for example, through a Bayesian framework, and consider these in the decision-making for petroleum exploration. Another focus area is machine learning methods.



Seismic Reservoir Characterization, Fig. 2 Seismic reservoir characterization workflow. Characterization starts with quality control of both seismic and well log data. Seismic data are conditioned, and desired angle gathers, stacks, or AVA attributes are generated. Well logs are used to generate similar synthetic gathers and stacks, followed by well tie analysis (Fig. 3) to calibrate seismic amplitude with synthetics generated using the estimated wavelet. The seismic reflectivity data can be interpreted for reservoir characterization. Also, we can perform seismic inversion to either directly invert for reservoir properties or first invert for

elastic properties and then estimate reservoir properties. These interpretations and inversions require information and calibration from geology (often using well logs). Common practices in terms of seismic inversion for elastic properties includes (i) pre-stack seismic inversion for V_p , V_s , and density; (ii) limited angle stack inversion for elastic impedance (EI); (iii) inversion of AVO intercept and gradient for acoustic impedance (AI) and gradient impedance (GI), respectively; and (iv) inversion of AVO attributes for V_p , V_s , and density. Refer to Whitcombe et al. (2002) for EI and GI

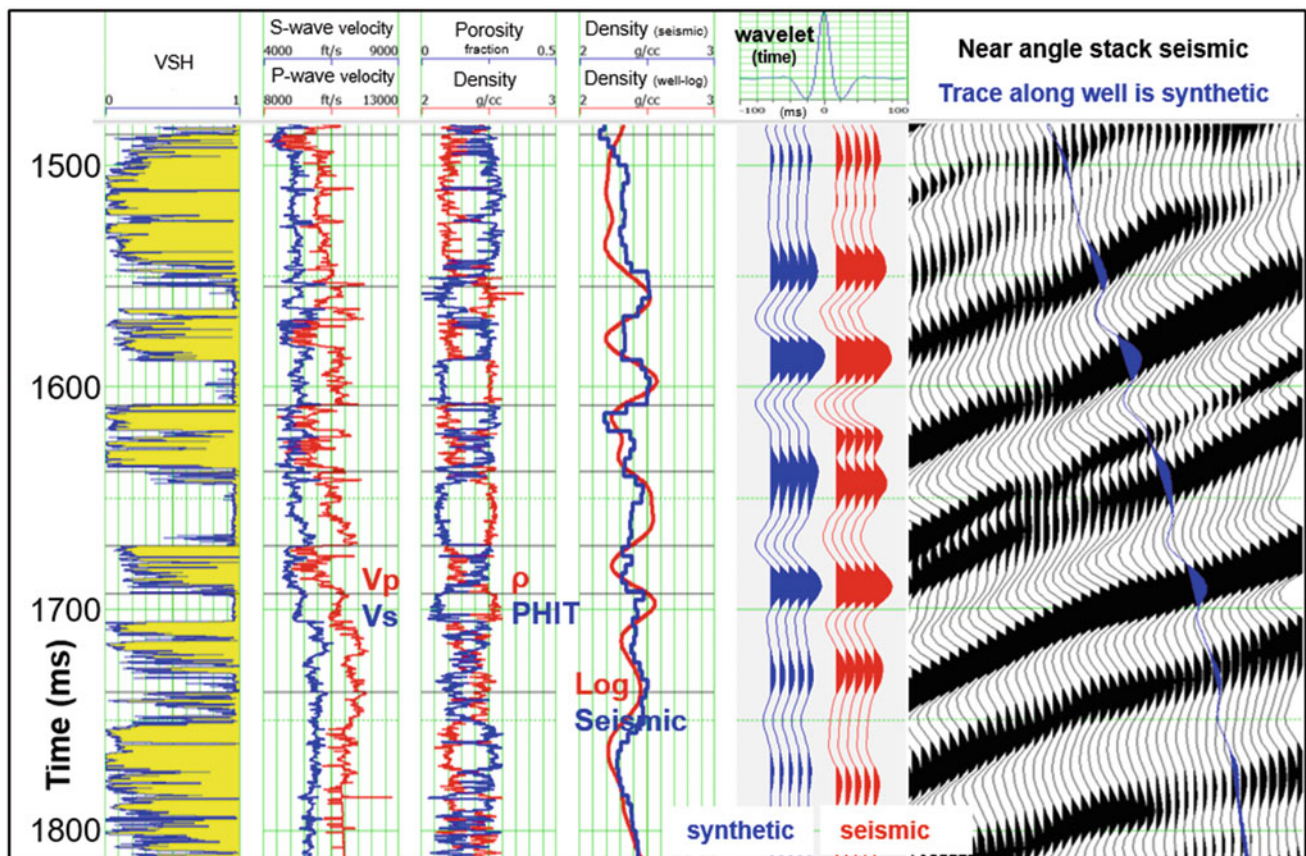
Applications of Machine Learning Methods in Reservoir Characterization

Machine learning (ML) methods take a numerical approach to estimate reservoir properties (labels in ML), recognizing the inherent imperfection between physics-based models and seismic amplitudes (data). This is possible because of the advancements in computer algorithms and speed and is done by first deriving abstract numerical relationships between data and desired properties using known data and label combinations, followed by predictions of labels for new datasets. In the absence of the availability of labeled data combination, synthetic models can be used to train the algorithms. This approach of ML is known as supervised learning. ML is becoming an alternative method for geophysical inversion in the presence of a large amount of data and a poor understanding of physics (Russell 2019). Structural (fault, horizon) mapping using ML is becoming a routine (Zheng et al. 2019), but reservoir properties prediction is still not fully developed and accepted. However, physics-guided ML takes the unsupervised approach, where a labeled data is not

required. It has been used for reservoir property prediction as one way to improve the validity and acceptable geological interpretations (Biswas et al. 2019b). Most ML algorithms are based on a type of statistical learning (Gareth et al. 2014). Note that conventional reservoir characterization is done using the processed seismic data to make sure seismic data contain only the P-wave primary reflection response that can be explained with known physics (e.g., Zoeppritz AVA equation-based convolutional synthetics). One should think about using the raw seismic data in ML methods for reservoir characterization. Although ML methods have potential to provide fast turnaround of a project and reduce human bias, human interactions will be needed to come up with physically interpretable reservoir properties.

Summary

Seismic data provides images of the subsurface in terms of the 3D structure, where there are changes in elastic properties and density. Various seismic reflectivity attributes are generated to

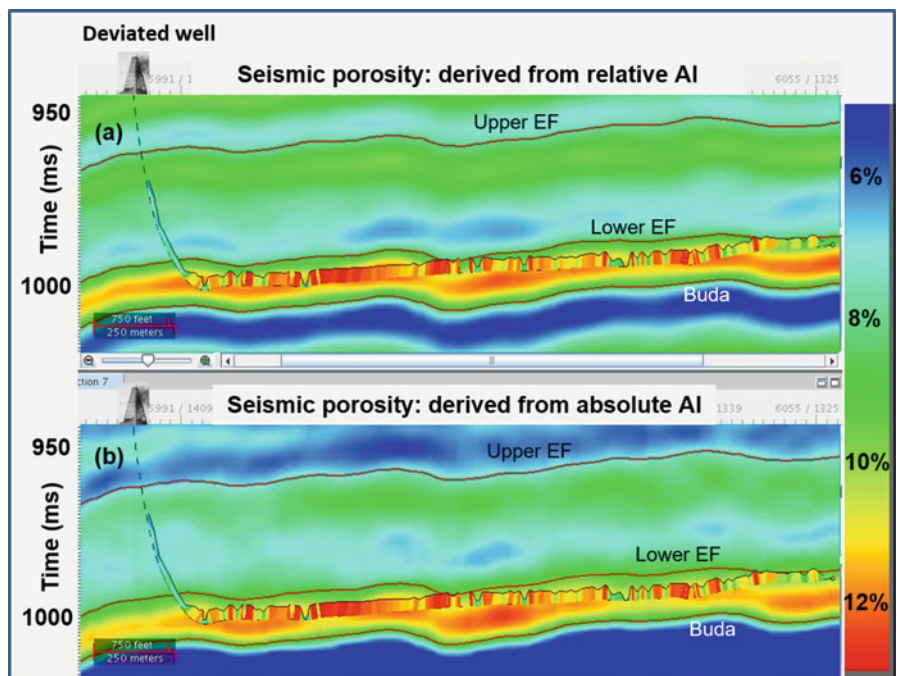


Seismic Reservoir Characterization, Fig. 3 Seismic-well tie (Biswas et al. 2019a). The first column shows VSH curve to highlight lithology, second column shows V_p (red) and V_s (blue), and third column shows bulk density (red) and porosity (blue). The fourth column evaluates seismic inverted density (blue) with smooth well log density (red), and

the two compare nicely here. The fifth column shows seismic-well tie as synthetic (blue) and seismic (red) reflectivity comparison, and the last column shows another view of seismic-well tie as seismic section along a well and synthetic (blue) displayed along well trajectory

Seismic Reservoir Characterization, Fig. 4

Seismic-derived porosity section views along a well (Kumar et al. 2014). Log porosity along the wellbore is compared with porosity predicted from two seismic acoustic impedance (AI) models: relative seismic AI (a) and absolute seismic AI (b). The two seismic porosities are comparable in the Eagle Ford (EF) shale, and they also match well with log porosity. The same color scale is used for both seismic and log porosities. Seismic formations (upper EF, lower EF, and Buda) are labelled



enhance seismic sensitivity to reservoir properties. Seismic inversion is utilized further to estimate layer properties for geological interpretation. The two critical ingredients for a successful seismic reservoir characterization are (i) favorable rock physics showing a strong sensitivity of seismic amplitude to reservoir properties and (ii) good-quality seismic data. Uncertainties in estimated reservoir properties exist due to low-quality seismic data, bandlimited resolution of seismic, poor calibration of seismic data to well logs, and approximate seismic physics used in the analysis. Future emphasis is on probabilistic seismic analysis and machine learning methods to estimate reservoir properties from seismic.

Cross-References

- [Earthquake Source Theory](#)
- [Seismic Data Acquisition and Processing](#)
- [Seismic Noise](#)
- [Uncertainties in Geomodelling Related to Geophysical Inversion](#)

Acknowledgments The author thanks Stan Davis, Chandan Kumar, Reetam Biswas, Margarita Corzo, John Etgen, Lisa Davies, and Kalachand Sain for reviews and BP for permission to publish.

Bibliography

- Biswas R, Kumar D, Sen MK, Paul A, Packer K (2019a) Density inversion from seismic using a trans-dimensional approach: a field dataset example. In: 89th Annual International Meeting, Society of Exploration Geophysicists, Expanded Abstracts, pp 539–543
- Biswas R, Sen MK, Das V, Mukerji T (2019b) Pre-stack and post-stack inversion using a physics-guided convolutional neural networks. Interpretation 1–76. <https://doi.org/10.1190/int-2018-0236.1>
- Chopra S, Marfurt KJ (2007) Seismic attributes for prospect identification and reservoir characterization. SEG geophysical development series no 11. Society of Exploration Geophysicists, Tulsa
- Gareth J, Witten D, Hastie T, Tibshirani R (2014) An introduction to statistical learning: with applications in R. Springer Publishing, New York. ISBN 978-1-4614-7137-0
- Kumar D, Sugianto H, Li S, Patel H, Land S (2014) Using relative seismic impedance to predict porosity in the Eagle Ford shale. In: 84th Annual International Meeting, Society of Exploration Geophysicists, Expanded Abstracts, pp 2688–2692
- Lancaster S, Whitcombe D (2000) Fast-track coloured inversion. In: 70th Annual International Meeting, Society of Exploration Geophysicists, Expanded Abstracts, pp 1572–1575
- Mavko G, Mukerji T, Dvorkin J (1998) The rock physics handbook: tools for seismic analysis in porous media. Cambridge University Press, Cambridge
- Menke W (1984) Geophysical data analysis: discrete inverse theory. Academic, Orlando
- Russell B (2019) Machine learning and geophysical inversion: a numerical study. The Leading Edge 38(7):512–519. <https://doi.org/10.1190/tle38070512.1>
- Sheriff RE (1975) Factors affecting seismic amplitudes. Geophys Prospect 23:125–138
- Simm R, Bacon M (2014) Seismic amplitude: an interpreter's handbook. Cambridge University Press, Cambridge
- Tarantola A (1986) A strategy for nonlinear elastic inversion of seismic reflection data. Geophysics 51:1893–1903
- Whitcombe DN, Connolly PA, Reagan RL, Redshaw TC (2002) Extended elastic impedance for fluid and lithology prediction. Geophysics 67:63–67
- Yilmaz O (2001) Seismic data analysis: processing, inversion, and interpretation of seismic data. Society of Exploration Geophysicists, Tulsa
- Zheng Y, Zhang Q, Yisifov A, Yunzhi S (2019) Applications of supervised deep learning for seismic interpretation and inversion. The Leading Edge 38(7):528–532. <https://doi.org/10.1190/tle38070526.1>

Seismic Seiches

Art McGarr

Earthquake Science Center, U.S. Geological Survey, Menlo Park, CA, USA

Definition

Seismic seiche is a term first used by Kvale (1955) to discuss oscillations of lake levels in Norway and England caused by the Assam earthquake of August 15, 1950. This definition has since been generalized to apply to standing waves set up in closed, or partially closed, bodies of water including rivers, shipping channels, lakes, swimming pools, and tanks due to the passage of seismic waves from an earthquake.

The first published mention of seismic seiches is thought to be reports of those observed throughout much of Europe due to the great earthquake at Lisbon, Portugal, in 1755 (Wilson 1953; Richter 1958). In addition to the Lisbon and Assam earthquakes, seismic seiches at teleseismic distances have been observed for many other large earthquakes including the 1964 Alaska (McGarr and Vorhis 1968) and the 2002 Denali, Alaska, an earthquake that caused damaging seiches in Lake Union, Seattle, Washington, at an epicentral distance of 2400 km (Barberopoulou et al. 2004). Most recently, Bondevik et al. reported observations of seiches in Norwegian fjords caused by S waves and surface waves from the 2011 M9 Tohoku, Japan, earthquake. Based on evidence from eyewitnesses and surveillance and cell phone cameras, the seiches began about 30 min after the origin time of the earthquake.

Kvale (1955) showed that seismic surface waves from the Assam earthquake were the most probable cause of the seiches observed in Norway and England at that time. Moreover, he concluded that the natural period of a basin must be matched by the periods of the passing seismic surface waves. Motivated by observations reported by Donn (1964) of a seiche generated in a channel near Freeport, Texas, at an epicentral distance of about 5040 km from the 1964 Alaska earthquake, McGarr (1965) developed a relation between the ground motion of seismic waves and the resulting seiche. The passing seismic

wave exerts a horizontal acceleration on a closed, or partially closed, body of water, which can be idealized as a long channel of uniform depth. This causes a seiche, composed of standing water waves whose periods depend on the dimensions of the channel. The amplitude of the seiche depends on channel depth, the amplitudes of the horizontal accelerations of the seismic waves, and the extent to which the periods of the seismic waves match those of standing water waves. The gravest mode of the standing waves has a period given by $T = 2L/\sqrt{gH}$, where T is the period in seconds, L is the channel width in meters, H is the channel depth in meters, and g is gravity. For instance, if $L = 100$ m and $H = 10$ m, then the period of the gravest seiche mode is 20 s, which tends to be the period where surface waves show maximum amplitudes at teleseismic distances. Any factor that enhances the amplitudes of seismic waves, such as basins containing low-velocity sediments, tends to result in greater production of observed seiches from a given earthquake (McGarr and Vorhis 1968; Barberopoulou et al. 2004).

Bibliography

- Barberopoulou A, Qamar A, Pratt TL, Creager K, Steele WP (2004) Local amplification of seismic waves from the Denali earthquake and damaging seiches in Lake Union, Seattle, Washington. *Geophys Res Lett* 31:L03607. <https://doi.org/10.1029/2003GL018569>
- Bondevik S, Gjevik B, Sorensen MB (2013) Norwegian seiches from the giant 2011 Tohoku earthquake. *Geophys Res Lett* 40:3374–3378. <https://doi.org/10.1002/grl.50639>
- Donn WL (1964) Alaska earthquake of 27 March, 1964: remote seiche stimulation. *Science* 146:261–262
- Kvale A (1955) Seismic seiches in Norway and England during the Assam earthquake of August 15, 1950. *Bull Seismol Soc Am* 45:93–113
- McGarr A (1965) Excitation of seiches in channels by seismic waves. *J Geophys Res* 70:847–854
- McGarr A, Vorhis RC (1968) Seismic seiches from the March 1964 Alaska earthquake. U.S. Geological Survey professional paper 544E. pp E1–E43
- Richter CF (1958) *Elementary seismology*. W. H. Freeman and Company, San Francisco, 768 pp
- Wilson BW (1953) Coastal seiches, pt. 1 of oscillations of the sea and the phenomenon of range. The Dock and Harbour Authority [London], June, pp 41–45

Seismic Signals in Well Water Observations

R. K. Chadha

CSIR:National Geophysical Research Institute, Hyderabad, India

Synonyms

Anomalous water level changes; Groundwater anomalies; Hydrological earthquake precursor

Definition

Seismic signals in well water observations – temporal rise or fall in well water level fluctuations induced by earthquakes.

Introduction

Seismic waves generated during an earthquake cause oscillations and permanent offsets in well water levels. Liu et al. (1989) explained the oscillations in terms of expansion and contraction of pore pressure changes induced by seismic waves. Systematic monitoring of well water levels has led to identification of anomalous pre-, co-, post-, and transient changes in well water levels related to earthquakes (Roeloffs 1996). These signals are believed to reflect pore pressure changes related to the redistribution of stress in the near and far fields of dislocation sources. The co-, post-, and transient seismic changes appear as a step or spikelike anomaly in well water oscillations and recover post-earthquake depending on the transmissivity of the medium. While the co- and post-seismic signals are attributed to local and regional earthquakes, the transient changes are in response to teleseismic (>1000 km) earthquakes. Several attempts have been made, in hindsight, to interpret preseismic well water levels changes as earthquake precursor.

Co- and Post-seismic Signals

The coseismic signals generally coincide with earthquake time and are observed mostly in wells which respond to earth tides. Earth tides are very small deformations in the earth's surface caused by the gravitational fields of the sun and moon as the earth rotates within their fields. The presence of tidal signals in well level recordings indicates that they are connected to confined, that is, hydraulically locked and fully water-saturated aquifers. This is favorable for the detection of weak pore pressure anomalies induced in crustal stress and hence can be considered as volume strainmeter of resolution of the order of 10 nanostrain (1 in 10^{-8}) or better. As an alternate mechanism, Rojstaczer et al. (1995) stressed the role of earthquake-induced enhancement of permeability in the shallow crust to explain the co- and post-hydrological changes observed in springs and groundwater flows after the 1989 Loma Prieta earthquake in California.

The 1999 Chi-Chi earthquake of $M_w 7.6$ in Taiwan is one of the well-documented events for water level changes in wells. Chia et al. (2001) reported the coseismic changes and compared them to the distances between the earthquake fault and the observations wells. Lee et al. (2002) reported coseismic hydrological changes associated with dislocation caused due to Chi-Chi earthquake, and Lai et al. (2004) compared these changes with the regional geological setting

and seismic ground motions. They showed that in the nearby Choshui River alluvial fan area, the groundwater levels rose coseismically, and amplitudes increased with increase in ground acceleration and hydraulic conductivity, whereas in the slope regions near the earthquake fault, the water levels dropped with increasing intensity as the ground accelerations increased. In a recent study, Anhua He and Singh (2019) reported coseismic step changes in 138 wells in China associated with the M_w 7.9 Wenchuan earthquake of 2008. Earlier, Zheng et al. (2012) analyzed the coseismic responses of the fluid well pattern system in Jiangsu Province to the 2008 Wenchuan and 2011 Tohoku earthquakes. Shi et al. (2013) studied coseismic response of groundwater level in the three gorges well network and its relationship to aquifer parameters.

Akita and Matsumoto (2004) showed coseismic changes induced by M 8.0 Tokachi-oki earthquake in 2003 in Japan. These coseismic changes were observed in 29 wells with increases in wells located in southeast part of Hokkaido and decreases in other areas. Matsumoto et al. (2003) studied 28 cases of coseismic changes during 1981 to 1997 at Haibara well in central Japan and obtained an empirical relation between earthquake magnitudes and distances. They attributed these changes to the strong ground motions due to passing of seismic waves. Koizumi et al. (1996) reported groundwater anomalies associated with the 1995 Hyogoken Nanbu earthquake, in Japan.

Chadha et al. (2003) reported four cases of coseismic changes in wells in the Koyna region, India. To comprehend the cause-and-effect relationship between local earthquakes and water level changes, they drilled 21 bore holes surrounding a well-constrained seismic source volume of $30 \times 15 \times 10 \text{ km}^3$ dimension. Analysis of data from these wells from 1997 to 2003 revealed four cases of coseismic step changes associated with local earthquakes of $5.3 \leq M \leq 4.2$ within a 50 km radius. Rudnicki et al. (1993) observed water levels at 17 wells in the Parkfield area along the San Andreas fault during January 1989 to 1990 as a part of the Parkfield earthquake prediction experiment. Using the recovery rates of water level changes due to fault creep or slow slip, they inferred the position of the fault zone. Later, Roeloffs (1998) analyzed persistent water level changes at a well near Parkfield caused by local and distant earthquakes.

Gavrinlenko et al. (2000) reported permanent water level drop associated with Spitak earthquake of 1988 at a well located 110 km from the epicenter. This drop occurred post-earthquake and might have been influenced by closeness of the fault zone. Following a sequence of nine M_w 5.0–5.9 earthquakes in central Italy, changes in water table and spring discharge were reported for several months (Petitta et al. 2018). Data from 22 measurement sites, located within 100 km of the epicentral zones, were analyzed. The intensity of the induced changes was correlated with earthquake magnitude and distance to epicenters.

Transient Signals

Transient changes in well water levels due to teleseismic earthquakes were reported for earthquakes of $M > 7.5$. Chadha et al. (2008) reported transient and persistent water level changes at few wells in Koyna, India, following M_w 9.3 Sumatra earthquakes in 2004 (Fig. 1). The transient changes were either a spike- or a steplike signal. They attributed these changes to the dynamic strain induced by the passage of seismic waves, most probably long period surface waves. Sil and Freymueller (2006) reported anomalous well water level changes in Fairbanks, Alaska, with the passing of seismic waves due to the Sumatra earthquake of 2004 in the Indian Ocean. Kitagawa et al. (2006) reported well water level changes in 38 observation wells due to Sumatra earthquake at a distance of more than 5000 km in Japan. Ishii-type borehole strainmeters installed in 10 of the observation wells also recorded changes in crustal strains.

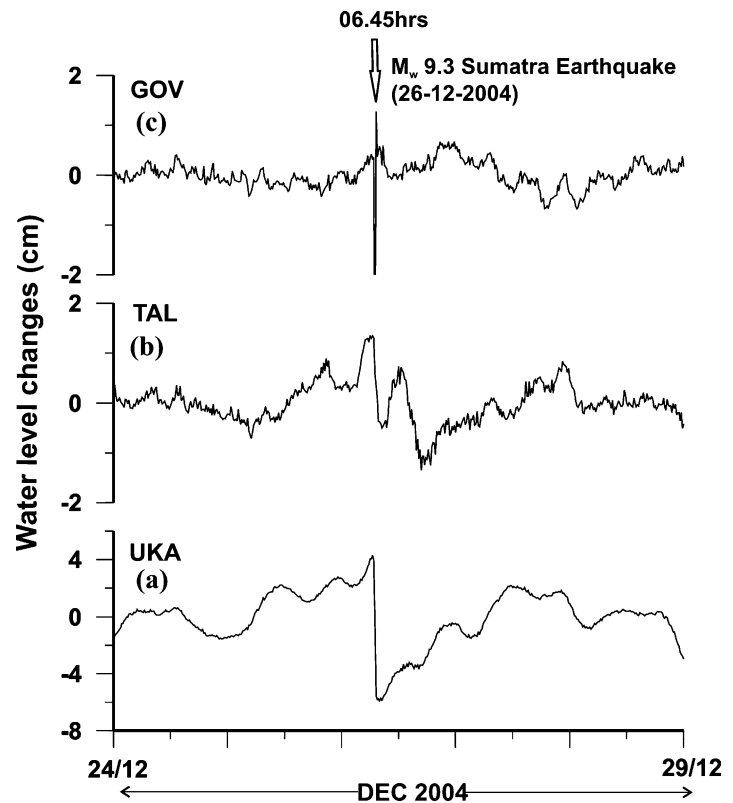
Sun et al. (2015) showed water level changes at Changping well in China related to M_w 9.0 Tohoku earthquake in Japan. They termed these changes as coseismic induced mostly by the surface waves of this distant earthquake. Similarly, He et al. (2017) reported coseismic response of water levels in the Jingle well in China associated with the Gorkha Nepal (M_w 7.8) earthquake in 2015 in the Himalaya. Several reports of earthquake-induced well water level changes at distant locations came out after the M_w 7.9 Denali earthquake in 2002 in Alaska (Cassidy and Rogers 2004). In the Korean Peninsula, Yun et al. (2019) analyzed groundwater level changes in a monitoring well close to Yangsan fault zone, induced by local and offshore earthquakes around the world.

Preseismic Signals

During 1970s there was a great optimism to study hydrological and geochemical signals for predicting earthquakes after the success of 1975 Haicheng earthquake in China. Wakita et al. (1988) presented preseismic changes in water level, temperature, radon, and strain associated with the 1978 Izu Oshima earthquake of M 7.0 in Japan. Later, Tsunogai and Wakita (1996) reported preseismic data on radon, chlorine, and strain prior to the 1995 Kobe earthquake of M 7.2 in Japan and showed the remarkable similarity in the precursory pattern 1 month prior to the earthquakes. Roeloffs (1996) reported preseismic changes at the observation stations located within 35 km from the epicenter of the 1985 M 6.1 Kettleman Hills earthquake along the San Andreas fault system. Two of the four water level sites and two of three volumetric strain sites showed covarying precursory signals beginning 3 days before the event. King et al. (1999) studied the water level drops observed at 16 wells in Tono mines in central Japan, prior to a M 5.8 local earthquake, and suggested

Seismic Signals in Well Water

Observations, Fig. 1 Transient changes in water levels at UKA, TAL, and GOV wells in the Koyna region, western India, induced by $M_w 9.3$ Sumatra earthquake 2000 km away in the Indian Ocean. Arrow shows the arrival of surface waves at the well location



that these changes were premonitory in nature. Chadha et al. (2003) reported preseismic water level drops prior to a coseismic step associated with a $M 4.4$ local earthquake in the Koyna region (Fig. 2). Igarashi et al. (1992) explained the preseismic water level drop as a “rebound anomaly” related to an increase of porosity and permeability due to fracturing, with subsequent recovery attributable either to influx of fluid or to compression.

Mechanisms

Three mechanisms have been put forward to explain the seismic signals in the wells.

Fault dislocation and static poroelastic strain mechanism states that coseismic water level changes can be explained by the coseismic static strain and the change in pore pressure predicted by the poroelastic theory (Wakita 1975; Ge and Stover 2000). Chadha et al. (2005) computed static volumetric strain change to explain the coseismic steps observed in four cases associated with $M \geq 4.2$ local earthquakes in the Koyna region, India. They showed a good correlation between steplike increases in wells located in compression and decrease in dilatational zones but reported a misfit in amplitudes of the anomalies. Based on this study, they concluded that coseismic strains may be a function of site effects controlled by local heterogeneity in geological structures. Thus, simple elastic models cannot wholly explain the

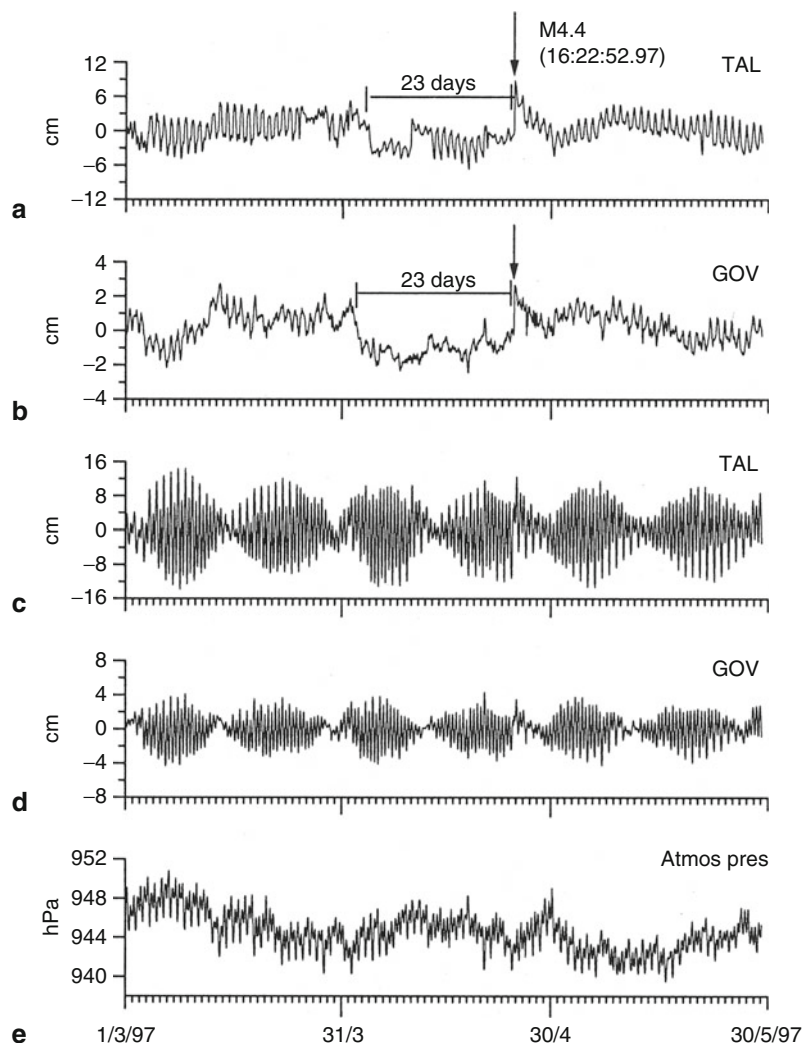
amplitudes of the hydrological anomalies. Huang et al. (1995) and Grecksch et al. (1999) have suggested a nonlinear response of water levels to coseismic strain changes due to local heterogeneities. King et al. (1999) computed strain from dislocation of seismic faults in Japan and showed a good quantitative agreement with the strain inferred from the water level changes around the epicenter. Quilty and Roeloffs (1997) showed good agreement between predicted strain and observed water level changes due to 1994 earthquake in Parkfield, California. Lee et al. (2002) found a similar pattern using the 178 auto-recording monitoring wells associated with the 1999 Chi-Chi earthquake.

Consolidation hypothesis proposes that seismic shaking causes sediments to consolidate or dilate under undrained conditions (Wang and Chia 2008). The water level will decrease or increase within the near field, and this change depends on the magnitude between the shear strain caused by seismic shaking and the critical threshold (10^{-4}). However, the water level will increase at regional distances when seismic wave energy exceeds the amount of dissipated energy required to cause undrained consolidation (Yoshimi and Oh-Oka 1975; Hsu and Vucetic 2004; Lai et al. 2004; Wang and Manga 2010).

Earthquake-enhanced permeability mechanism suggests enhancement in rock permeability either due to rock fracturing during earthquakes (Rojstaczer et al. 1995) or other processes like clogging and unclogging associated with aquifer deformation (Shi and Wang 2016) and consolidation of

Seismic Signals in Well Water

Observations, Fig. 2 Records of air pressure, and residual water levels at two wells in the Koyna region, western India, during March-May, 1997. A 23 days precursory drop is clearly seen



surficial deposits (Xue et al. 2016). Sometimes fractures in the earth are filled up with broken rock material during the normal course of groundwater flow and thus become points of local stress concentration increasing local pore pressure. These local stress points are broken when there is a rapid flow of water during an earthquake (Brodsky et al. 2003) redistributing the pore pressure as the permeability increases that can lead to changes in water levels in wells.

Summary

Earthquakes cause dislocations along faults in earth's crust which leads to redistribution of stresses resulting in alteration of subsurface medium properties and flow of groundwater. Anomalous changes were observed in well water levels after local, regional, and teleseismic earthquakes. These induced anomalous changes termed here as "seismic signals" in well water levels are categorized as pre-, co-, post-, and transient anomalies which manifest as preseismic water level drops, a

coseismic step- or a transient spikelike change, and sometimes a gradual post-seismic increase. These signals indicate the spatial distribution of seismic energy density that triggers the different responses. Three mechanisms have been put forward to explain these seismic signals in terms of a fault dislocation model and static poroelastic strain hypothesis where predicted strain is found to be in good agreement with the strain predicted by water level changes around the epicentral area. According to static strain hypothesis, the coseismic steplike anomalies can be explained in terms of coseismic static strain and changes in pore pressure predicted by poroelastic theory. The second is the consolidation hypothesis which states that the increase or decrease in well water levels depends on the magnitude of the shear strain caused by seismic shaking and the critical threshold of strain at local distances. However, the water level will increase at regional distance when seismic wave energy exceeds the amount of dissipated energy required to cause undrained consolidation. Finally, the earthquake-enhanced permeability mechanism suggests enhancement in rock permeability either due to

rock fracturing during earthquakes or other processes like clogging and unclogging associated with aquifer deformation and consolidation of surficial deposits.

Seismic signals in well water levels, while clearly are strain indicators, do not have a well-understood relation to strain. They exhibit “response heterogeneity” in the signals which are highly variable in space and time and thus have no straightforward explanation. The mechanisms put forward to explain these signals may not be mutually exclusive but may occur together and may even be causally related.

The preseismic changes in well water levels are still far from using them as short-term earthquake precursors. Very often, these precursory changes which represent the pre-earthquake phase are very small and therefore not obvious in the raw data which is also affected by atmospheric pressure, significant precipitation, and earth tides. Innovative filtering techniques have to be developed to decipher the earthquake-related anomalies from non-tectonic effects in the data. Definite patterns of earthquake-related water level anomalies have to be established, both site specific and global, so that predictive capabilities can be developed. We realize that earthquake prediction using short-term precursors is not easy, but sound observations which may lead to a physical model must continue. The earthquake-induced water level changes in groundwater have other implications, too. The supply of groundwater, both in terms of quantity and quality, contaminant transport, and underground waste repositories are some of the important issues which need attention. Plans made to manage these issues are based on pre-earthquake scenario which may drastically change, post event. Thus, there is a need to accelerate these studies by casting wider net-expanding observing programs to capture many more events and their precursors.

Bibliography

- Akita F, Matsumoto N (2004) Hydrological responses induced by the Tokachi–Oki earthquake in 2003 at hot spring wells in Hokkaido, Japan. *Geophys Res Lett* 31(16):371–375
- Brodsky EE, Roeloffs E, Woodcock D, Gall I, Manga M (2003) A mechanism for sustained groundwater pressure changes induced by distant earthquakes. *J Geophys Res* 108(B8):503–518
- Cassidy JF, Rogers GC (2004) The M_w 7.9 Denali fault earthquake of 3 November 2002: felt reports and unusual effects across western Canada. *Bull Seismol Soc Am* 94:53–57
- Chadha RK, Pandey AP, Kuempel HJ (2003) Search for earthquake precursors in well water levels in a localized seismically active area of reservoir triggered earthquakes in India. *Geophys Res Lett* 30(7):1416. <https://doi.org/10.1029/2002GL016694>
- Chadha RK, Srivastava K, Kuempel HJ (2005) Earthquake related changes in well water level and their relation to a static deformation model for the seismically active Koyna-Warna region. In: Rummel F (ed) *Rock mechanics with emphasis on stress*. Oxford & IBH, New Delhi, pp 135–150
- Chadha RK, Singh C, Shekar M (2008) Transient changes in well water level in bore wells in western India due to 2004 mw 9.3 Sumatra earthquake. *Bull Seismol Soc Am* 98(5):2553–2558
- Chia Y, Wang YS, Chiu JJ, Liu CW (2001) Changes of groundwater level due to the 1999 ChiChi earthquake in the Choshui River alluvial fan in Taiwan. *Bull Seismol Soc Am* 91(5):1062–1068
- Gavrinlenko P, Melikadze G, Chelidze T, Gilbert D, Kumsiashvili G (2000) Permanent water level drop associated with Spitak earthquake: observation at Lisi borehole (republic of Georgia) and modeling. *Geophys J Int* 143:83–98
- Ge S, Stover SC (2000) Hydrodynamic response to strike- and dip-slip faulting in a half-space. *J Geophys Res* 105(11):25513–25524
- Grecksch G, Roth F, Kuempel HJ (1999) Coseismic well level changes due to the 1992 Roermond earthquake compared to static deformation of half space solutions. *Geophys J Int* 138:470–478
- He A, Singh RP (2019) Groundwater level response to the Wenchuan earthquake of May 2008. *Geomat Nat Haz Risk* 10(1):336–352
- He A, Fan X, Zhao G, Liu Y, Singh RP, Hu Y (2017) Co-seismic response of water level in the jingle well (China) associated with the Gorkha Nepal (mw 7.8) earthquake. *Tectonophysics* 714/715:82–89
- Hsu CC, Vucetic M (2004) Volumetric threshold shear strain for cyclic settlement. *J Geotech Geoenviron Eng* 130(1):58–70
- Huang W, Rojstaczer S, Breau S (1995) Coseismic response of water level to earthquakes in the San Jacinto fault, southern California. *EOS Tran Am Geophys Un* 76:369
- Igarashi G, Wakita H, Sato T (1992) Precursory and coseismic anomalies in well water levels observed for the February 2, 1992 Tokyo Bay earthquake. *Geophys Res Lett* 19:1583–1586
- King CY, Azuma S, Igarashi G, Ohno M, Saito H, Wakita H (1999) Earthquake-related waterlevel changes at 16 closely clustered wells in Tono, Central Japan. *J Geophys Res* 104(B6):13073–13082
- Kitagawa Y, Koizumi N, Takahashi M, Matsumoto N, Sato T (2006) Changes in groundwater levels or pressures associated with the 2004 earthquake off the west coast of northern Sumatra (M9. 0). *Earth Planet Space* 58(2):173–179
- Koizumi N, Kano Y, Kitagawa Y, Sato T, Takahashi M, Nishimura S, Nishida R (1996) Groundwater anomalies associated with the 1995HyogokenNanbu earthquake. *J Phys Earth* 44:373–380
- Lai W-C, Koizumi N, Matsumoto N, Kitagawa Y, Lin C-W, Shieh C-L, Lee Y-P (2004) Effects of seismic ground motion and geological setting on the coseismic groundwater level changes caused by the 1999 Chi-Chi earthquake, Taiwan. *Earth Planet Space* 56(9):873–880
- Lee M, Liu TK, Ma KF, Chang YM (2002) Coseismic hydrological changes associated with dislocation of the September 21, 1999 Chi-chi earthquake, Taiwan. *Geophys Res Lett* 29(17):51–54
- Liu LB, Roeloffs E, Zheng XY (1989) Seismically induced water level fluctuations in the Wali Well, Beijing, China. *J Geophys Res* 94(B7):9453–9462
- Matsumoto N, Kitagawa G, Roeloffs EA (2003) Hydrological response to earthquakes in the Haibara well, Central Japan—I. Groundwater level changes revealed using state space decomposition of atmospheric pressure, rainfall and tidal responses. *Geophys J Int* 155(3):885–898
- Petitta M, Mastroiello L, Preziosi E et al (2018) Water-table and discharge changes associated with the 2016–2017 seismic sequence in Central Italy: hydrogeological data and a conceptual model for fractured carbonate aquifers. *Hydrogeol J* 26(4):1009–1026
- Quilty EG, Roeloffs EA (1997) Water-level changes in response to the 20 December 1994 earthquake near Parkfield, California. *Bull Seismol Soc Am* 87(2):310–317
- Roeloffs EA (1996) Poroelastic techniques in the study of earthquake-related hydrologic phenomena. *Adv Geophys* 37:135–195
- Roeloffs EA (1998) Persistent water level changes in a well near Parkfield, California, due to local and distant earthquakes. *J Geophys Res* 103(B1):869–889
- Rojstaczer S, Wolf S, Micheli R (1995) Permeability enhancement in the shallow crust as a cause of earthquake-induced hydrological changes. *Nature* 373:237–239

- Rudnicki JW, Yin J, Roeloffs EA (1993) Analysis of water level changes induced by fault creep at Parkfield, California. *J Geophys Res* 98:8143–8152
- Shi Z, Wang G (2016) Aquifers switched from confined to semiconfined by earthquakes. *Geophys Res Lett* 43(21):11166–11172
- Shi ZM, Wang GC, Liu CL, Mei JC, Wang JW, Fang HN (2013) Coseismic response of groundwater level in the three gorges well network and its relationship to aquifer parameters. *Chin Sci Bull* 58(25):3080–3087
- Sil S, Freymueller JT (2006) Well water level changes in Fairbanks, Alaska, due to the great Sumatra-Andaman earthquake. *Earth Planet Space* 58(2):181–184
- Sun X, Wang G, Yang X (2015) Coseismic response of water level in Changping well, China, to the mw 9.0 Tohoku earthquake. *J Hydrol* 531:1028–1039
- Tsunogai U, Wakita H (1996) Anomalous changes in groundwater chemistry. *J Phys Earth* 44(4):381–390
- Wakita H (1975) Water wells as possible indicators of tectonic strain. *Science* 189(4202):553–555
- Wakita H, Nakamura Y, Sano Y (1988) Short-term and intermediate-term geochemical precursors. *Pure Appl Geophys* 126:267
- Wang CY, Chia Y (2008) Mechanism of water level changes during earthquakes: near field versus intermediate field. *Geophys Res Lett* 35(12):109
- Wang CY, Manga M (2010) Earthquakes and water. Lecture notes in earth sciences, vol 114. Springer-Verlag, Berlin
- Xue L, Brodsky EE, Erskine J, Fulton PM, Carter R (2016) A permeability and compliance contrast measured hydrogeologically on the San Andreas fault. *Geochem Geophys Geosyst* 17(3):858–871
- Yoshimi Y, Oh-Oka H (1975) Influence of degree of shear stress reversal on the liquefaction potential of saturated sand. *Soil Found* 15(3):27–40
- Yun S, Hamm S, Cheong J, Lee C, Seo W, Woo N (2019) Analyzing groundwater level anomalies in a fault zone in Korea caused by local and offshore earthquakes. *Geosci J* 23(1):137–148
- Zheng J, Jiang H, He Z (2012) Analysis of the co-seismic responses of the fluid well pattern system in Jiangsu Province to the Wenchuan and Tohoku earthquakes. *Earthq Sci* 25(3):263–274

Seismic Stereotomography

Gilles Lambaré and Thibaut Allemand
CGG, Massy, France

Synonyms

Slope tomography

Definition

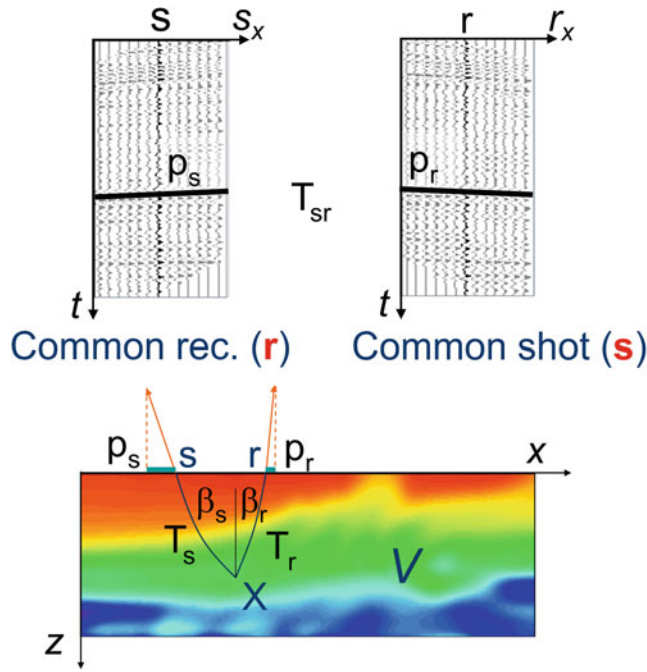
Stereotomography is a special form of reflection tomography proposed by Billette and Lambaré in 1998. It aims to reconstruct smooth, spatially varying properties of the subsurface (in particular the wave velocity) from seismic reflection data. Conventional reflection travel-time tomography relies on the

travel-time of continuous reflectors tracked in the prestack dataset over their full extent. On the other hand, stereotomography uses locally coherent reflected events characterized not only by the source and receiver positions and the reflection time, but also by their local slopes (the spatial derivatives of the reflection time with respect to receiver and source positions). Like travel-time tomography methods, stereotomography methods are based on seismic ray theory.

History

The roots of stereotomography go back to 1936 with the publication by F. Rieber of a paper in *Geophysics* describing a new reflection system with controlled directional sensitivity. Soviet scientists recognized the potential of this approach for velocity model building for seismic migration (see ► [“Seismic, Migration”](#)) and developed a method called CDR (controlled-directional-reception) in the 1950s and 1960s, principally by Professor Riabinkin (Billette and Lambaré 1998). A few years later, the approach gained interest from western geophysicists and in particular from Stanford’s Charles Sword, who spread it in his community. By 1998, Billette and Lambaré recast CDR and related approaches in a common theoretical framework under the generic term of slope tomography. These approaches have in common their aim to estimate the velocity model, $c(\mathbf{x})$ ($\mathbf{x}$ denoting the position in the subsurface), from a set of locally coherent reflected events (events that can be visually followed on a few adjacent traces in a given observation domain) characterized by their reflected travel-time, T_{sr} , source and receiver positions, $\mathbf{s}$ and $\mathbf{r}$, and slopes, that is, the spatial derivative of the travel-time, p_s and p_r (Fig. 1). These locally coherent events are interpreted within seismic ray theory (see ► [“Seismic ray theory”](#)) as pairs of ray segments connecting the reflecting point, $\mathbf{X}$, to the shot and receiver positions (Fig. 1). Compared to reflection traveltime tomography, the additional information brought by the slopes allows one to consider the reflection events locally: there is no need to track a reflection event over its full extent in the migrated or unmigrated prestack dataset. The picking (the extraction of the kinematic characteristics of the events, that is, their travel-time and slopes) becomes much easier and can therefore be performed more densely compared with conventional methods.

Within this theoretical frame, Billette and Lambaré also proposed a new approach, based on a generalization of the other slope tomography methods and with the expected benefit of improved robustness with respect to complex media involving multipath ray trajectories. They called it “stereotomography” and detailed its numerical implementation based on an efficient local optimization scheme involving paraxial ray tracing (see ► [“Seismic Ray Theory”](#) and ► [“Inverse Theory, Linear”](#)). Since this pioneering publication,



Seismic Stereotomography, Fig. 1 A locally coherent event in the prestack data cube, characterized by its central source and receiver positions $\mathbf{s}$ and $\mathbf{r}$, slopes p_s and p_r , and travel-time, T_{sr} (Fig. 3 from Lambaré (2008))

stereotomography and, more generally, slope tomography methods have gained in effectiveness and established themselves as powerful approaches for industrial velocity model building. One of their big advantages over the standard tomographic velocity model building methods is the non-linear update of velocity model (see Lambaré et al. (2014))

Data and Model in Slope Tomography

All slope tomography methods (and stereotomography in particular) have in common the description of the dataset. It consists of a set of N locally coherent reflected events characterized by their source and receiver positions, source and receiver slopes, and their travel-times,

$$\mathbf{d} = [(\mathbf{s}, \mathbf{r}, p_s, p_r, T_{sr})_i]_{i=1}^N. \quad (1)$$

Such a dataset needs to be explained by a model consisting of an associated set of pairs of ray segments and a velocity field (Fig. 1). Various cost functions and strategies have been proposed to estimate the optimum model, with important consequences for the allowed model space and optimization process. In all formulations, the model space contains ray parameters and a velocity model, which can be estimated through either joint or sequential optimization, that is, the

tracing of rays followed by a velocity update. In stereotomography, they are optimized jointly. The model ($\mathbf{m}$) consists of:

$$\mathbf{m} = [(\mathbf{X}, T_s, T_r, \beta_s, \beta_r)_i]_{i=1}^N, V, \quad (2)$$

where $\mathbf{X}$ denotes the common starting point for the two ray segments, β_s and β_r denote the starting angles of the rays towards the source and receiver, respectively, T_s and T_r represent the two one-way travel-times associated with the two ray segments, and V represents the set of parameters describing the velocity model (velocity and anisotropic parameter grids, etc.) (Fig. 1). The cost function to minimize reads:

$$C(\mathbf{m}) = \frac{1}{2} (d_{calc}(\mathbf{m}) - d_{obs})^T C_D^{-1} (d_{calc}(\mathbf{m}) - d_{obs}) + Regul(\mathbf{m}, \mathbf{d}). \quad (3)$$

where C_D is the a priori covariance matrix for data parameters (Billette and Lambaré 1998) and $Regul(\mathbf{m}, \mathbf{d})$ stands for regularization terms. When ray segments and velocity are optimized sequentially, ray segments can be determined by relaxing partly the data fitting or the reflection conditions (the two ray segments starting from a common point). Once the rays are traced, the velocity is estimated based on fitting the remaining condition (Billette and Lambaré 1998). Among the most popular methods of this kind, the slope tomographic approach initially proposed by Chauris et al. (2002) is widely used in the industry following Guillaume et al. (2008). In this approach, the ray segments are first recovered for a given velocity fitting source and receiver positions, midpoint slopes, and traveltimes (this is called kinematic migration). The velocity is then updated considering the misfit of the residual move-out slope in the offset direction in the common image gathers.

The numerical scheme used for solving the inverse problem aims at minimizing the cost function (3) through iterative local optimization, using typically a quasi-Newton scheme (see ► “Inverse Theory, Linear”). A local optimization offers an efficient solution for ensuring convergence of the optimization process. It requires the computation of the data and Fréchet derivatives (derivatives of the data with respect to the model parameters), which can be estimated by using ray tracing and paraxial ray tracing, respectively (see ► “Seismic Ray Theory”), provided the model exhibits proper smoothness. Cubic cardinal B-Splines are a good choice which can be used in gridded or multilayered velocity models (for references, see Lambaré et al. (2014)) not only because of their smoothness but also because they act as a model preconditioner to ease the inversion.

Velocity model building from surface reflection data being by nature an ill-posed inverse problem it requires

regularization methods, such as Tikhonov regularization. More advanced, dip-dependent regularizations have also been proposed, allowing the production of structurally-conformed velocity models. Another strategy requires the introduction of other types of data. For example, in dip-constrained tomography, the velocity is updated considering the misfit between the structural dips of the kinematically remigrated reflected events and an expected dip model (Lambaré et al. 2014). Another example is the incorporation of direct arrival travel times measured from surface sources with receivers located in a well, or first arrival travel times from surface-seismic measurements (see Allemand et al. (2017)). Including this kind of data is helpful, for example, to estimate anisotropy in order to reconcile the kinematics of reflected and diving waves.

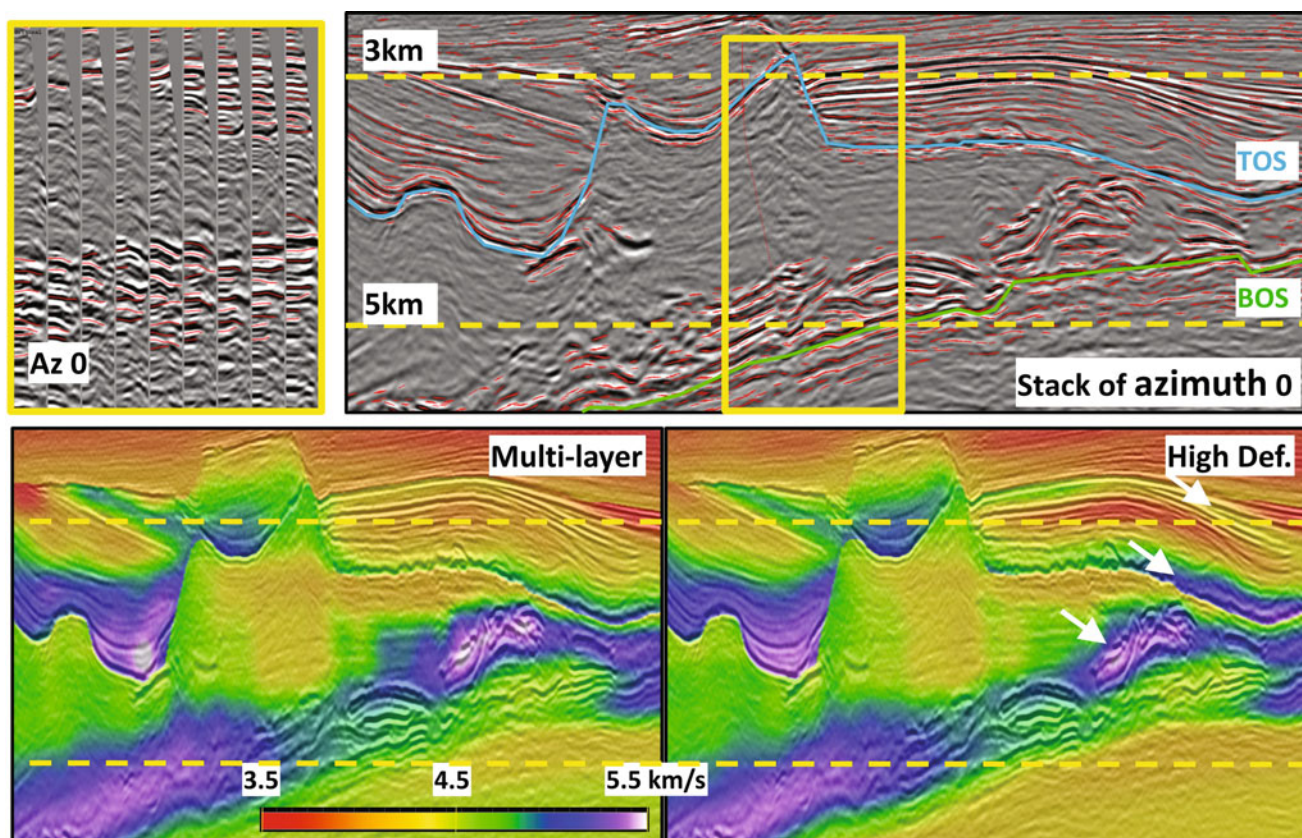
One of the first and most critical steps in velocity model building is the picking of locally coherent events (see Lambaré (2008) for a review). This can be performed either in the un-migrated data domain or in the (time- or depth-) migrated domain using automatic slope or dip picking tools. If done in the migrated domain, a kinematic demigration step is needed to

recover the so-called kinematic invariants, namely, the kinematic characteristics of the picked locally coherent events in the un-migrated domain (Guillaume et al. 2008). This versatility is a great strength in the approach, which has been extended from surface common-offset migration to any type of prestack depth migration, such as common-angle migration (Montel and Lambaré 2019). After picking, careful editing is needed to remove outliers. This is important as the quality of the picked data has a major influence on the inversion result.

Applications of Slope Tomography

Since it was first published in 1998, stereotomography and slope tomography have been implemented and applied in various contexts: marine streamer, ocean bottom seismic and land data, in 2D and 3D, isotropic and anisotropic contexts, using PP or PS waves (see Lambaré (2008) and Lambaré et al. (2014)).

We show in Fig. 2 an example of such an application (Hermant et al. 2014). The dataset is a high-density full-



Seismic Stereotomography, Fig. 2 Application of slope tomography to a land dataset. Top right: dip picks (in red) superimposed on the depth migrated stack for acquisition azimuth 0 (black to white). Blue curve indicates the location of top of salt (TOS), green curve the base of salt (BOS). Top left: residual move out curves (in red) superimposed on a

selection of common image gathers (black to white) for azimuth sector 0, located in the yellow box. Bottom: Velocity models obtained by multi-layer (left) and high-definition (right) tomography (Hermant et al. 2014). Courtesy of CGG, Petroleum Development Oman, and the Ministry of Oil and Gas of the Sultanate of Oman

azimuth seismic land dataset from South of the Sultanate of Oman. The area exhibits complex geological settings with an alternation of fast carbonate layers and slower shale or clastic layers overlying salt bodies. The associated strong lateral velocity variations are particularly challenging for velocity model building. The picking is done in the migrated domain to get the local dip of the events on the migrated stack and the corresponding local slope of the residual move-out curves in offset in the common image gathers (Fig. 2 (Top)). Figure 2 (Bottom left) shows the velocity model obtained by multi-layer tomography. This slope tomographic approach updates a layered velocity model with explicit velocity discontinuities iteratively repositioned during the updating process. Figure 2 (Bottom right) shows the velocity model after a further step of high definition tomography (an extension of slope tomography aimed at enhancing vertical resolution).

Challenges and Perspectives

Stereotomography and slope tomography methods in general are powerful velocity model building tools widely used in the seismic industry (Lambaré et al. 2014). The picking remains critical, with a major influence on the result. It should be accurate but also as dense as possible. The ability to handle a dense set of picks not limited to a few reflectors, combined with efficient nonlinear solvers and geologically constrained regularizations, allows high-resolution models to be obtained, far exceeding resolution of conventional travel-time tomography. On the other hand suboptimal picking will degrade the result, and a thorough editing should be applied prior to the inversion, for example using machine learning techniques.

In terms of tomographic velocity updates, Sambolian et al. (2019) recently investigated replacing ray tracing of individual pairs of ray segments by the use of travel-time maps (computed here by an eikonal solver). They promote the use of a matrix-free approach in order to invert for denser sets of picks.

An obvious limitation of these methods is the high frequency approximation, especially when dealing with complex media where waveforms have complicated behaviors with tuning effects. Hence, efforts should be made to extend the approach towards full-wave techniques. The connection between stereotomography and full-wave velocity model building, such as Differential Semblance Optimization, has been pointed out (“differential semblance is essentially data-weighted stereotomography, implicitly computed without event picking” Symes (2008)). Note, however, that the introduction of a picking step allows the extraction of the most significant kinematic information from the data, and the selection of events represents, to our minds, a serious advantage that would be worth preserving in full-wave approaches.

Conclusions

In the last 20 years, stereotomography and its variations (under the generic term of slope tomography) have earned their stripes in terms of industrial applications. This success is due to their versatility and efficiency at dealing with the critical velocity model building challenge (a non-linear and ill-constrained inverse problem). Even if full-wave methods are gaining popularity in complex areas, slope tomography remains and will remain a key tool for industrial velocity model building.

Cross-References

- [Inverse Theory, Linear](#)
- [Seismic Data Acquisition and Processing](#)
- [Seismic Imaging, Overview](#)
- [Seismic Ray Theory](#)
- [Seismic Tomography](#)
- [Seismic, Migration](#)
- [Seismic, Waveform Modeling and Tomography](#)
- [Traveltime Tomography Using Controlled-Source Seismic Data](#)

Bibliography

- Allemand T, Sedova A, Hermant O (2017) Flattening common image gathers after full-waveform inversion: the challenge of anisotropy estimation. In: 87th annual international meeting, SEG, expanded abstract, pp 1410–1415
- Billette F, Lambaré G (1998) Velocity macro-model estimation by stereotomography. *Geophys J Int* 135:671–680
- Chauris H, Noble M, Lambaré G, Podvin P (2002) Migration velocity analysis from locally coherent events in 2D laterally heterogeneous media, Part I: theoretical aspects. *Geophysics* 67:1202–1212
- Guillaume P, Lambaré G, Leblanc O, Mitouard P, Le Moigne J, Montel J-P, Prescott A, Siliqi R, Vidal N, Zhang X, Zimine S (2008) Kinematic invariants: an efficient and flexible approach for velocity model building. In: 78th annual international meeting, SEG, expanded abstract, pp 3687–3692
- Hermant O, Mansoor K, Baudon H, Al Amri A, Al Maamari A (2014) High-definition tomography: an application to intra-salt velocity update. In: EAGE/SPE workshop on subsalt imaging the challenges of subsalt exploration and imaging in deep water of the Middle East and North Africa, 16–19 February 2014, Limassol, Cyprus, SS21
- Lambaré G (2008) Stereotomography. *Geophysics* 73(5):VE25–VE34
- Lambaré G, Guillaume P, Montel JP (2014) Recent advances in ray-based tomography. In: 76th annual international conference and exhibition, EAGE, Extended Abstracts, We G103 01
- Montel JP, Lambaré G (2019) Kinematics of common-image gathers – Part I: theory. *Geophysics* 84(5):S437–S447
- Sambolian S, Operto S, Ribodetti A, Tavakoli FB, Virieux J (2019) Parsimonious slope tomography based on eikonal solvers and the adjoint-state method. *Geophys J Int* 218(1):456–478
- Symes W (2008) Migration velocity analysis and waveform inversion. *Geophys Prospect* 56:765–790

Seismic Structure at Mid-Ocean Ridges

Donald W. Forsyth

Department of Geological Sciences, Brown University,
Providence, RI, USA

Definition

Lithosphere	The cool, uppermost layer of the solid earth that moves as a unit and has some long-term elastic rigidity. It usually consists of both crust and the uppermost mantle. The mantle portion is typically characterized by high seismic velocities and low attenuation.
Asthenosphere	A more deformable, low-viscosity layer in the mantle underlying the lithosphere. Characterized by low seismic velocities and relatively high attenuation.
Anisotropy	Physical properties at any one point varying depending on direction; for seismic waves, velocity may depend on direction of propagation and direction of polarization of the wave.

Introduction

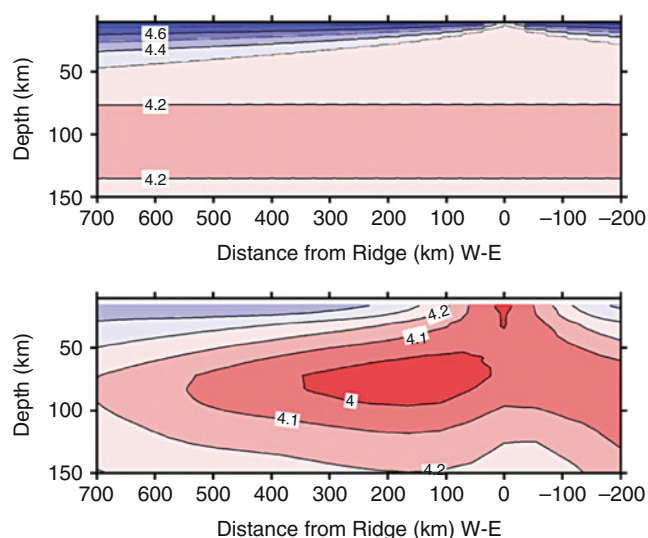
Mid-ocean ridges are spreading centers where two oceanic, lithospheric plates move apart. The separation of the plates induces upwelling in the underlying asthenospheric mantle. Because melting temperature decreases with decreasing pressure, as the mantle upwells, it undergoes pressure-release partial melting, producing magma that migrates upward to form new oceanic crust. As the plates move away from the spreading center, heat is lost to the surface conductively by diffusion and convectively by hydrothermal circulation. As the crust and mantle lose heat, the magma solidifies, and the plates thicken and cool with increasing age of the seafloor.

Although the general outline of the processes beneath mid-ocean ridges leading to the formation of new seafloor is well known, there are many questions remaining about the details of the mantle flow, melt generation, and melt migration that have not yet been fully answered. For example, along mid-ocean ridges, spreading centers are offset and segmented by transform faults and overlapping spreading centers. Are there distinct centers of mantle upwelling beneath each ridge segment or is the segmentation a shallow manifestation of stresses within the lithosphere with upwelling a more or less continuous phenomenon in the asthenosphere? How broad and deep is the melt production region?

The propagation of seismic waves through the crust and mantle provides one of the most direct ways of probing the structure beneath the ridges associated with plate separation and crust formation. The velocity, attenuation, and anisotropy of the waves are affected by temperature, composition, the presence of melt or cracks, and the crystal fabric. Tomographic images and maps of crustal and uppermost mantle structure are typically created in active source experiments where artificial sound sources, such as air guns, are recorded by ocean bottom seismometers (OBS) and/or long arrays of hydrophones towed behind ships. Deeper mantle structure is probed using signals generated by distant, teleseismic earthquakes recorded on arrays of ocean-bottom and land seismometers over periods of months to years. The logistical difficulty and expense of such experiments means that relatively few mid-ocean ridges have been studied in detail.

Mantle Structure

A cross section of the shear velocity structure beneath the East Pacific Rise spreading center is shown in Fig. 1 in comparison to the velocity variations that would be expected for simple conductive cooling of the plates if shear velocity were sensitive only to temperature and pressure. This tomographic image is based on the propagation of Rayleigh surface waves across two OBS arrays deployed for periods of 6 and 12 months. The expected thickening of the plate with increasing distance from the ridge axis as the plate cools is clearly observed in the form of increasing shear wave velocity near the



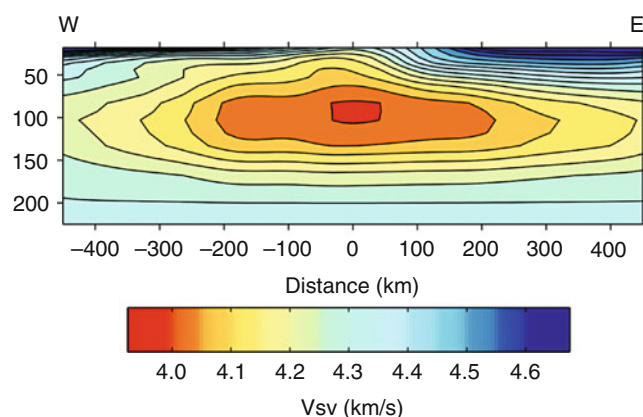
Seismic Structure at Mid-Ocean Ridges, Fig. 1 Tomographic cross section of the East Pacific Rise, comparing predicted shear velocity structure (*top*) to observed structure (*bottom*). Contours are labeled in km/s. Velocities are significantly lower than predicted for temperature effects alone, suggesting the presence of partial melt. (After Harmon et al. (2009))

surface. Velocity changes extend substantially deeper than is predicted and the shear velocity is lower than expected for the direct effects of temperature variations. Both of these departures from the predictions are indications that partial melt may be present, which could reduce the shear (S) wave velocity.

The lowest velocities are observed at depths of 60–80 km, where petrological models predict maximum melt production. There also are very low shear velocities at shallow depths immediately beneath the ridge axis, which may represent higher melt concentrations in the mantle that accumulated as melt migrates upward and inward to form the new oceanic crust at the ridge axis.

Another important feature of the shear velocity structure is the asymmetry across the ridge axis. To the east, beneath the Nazca plate, the high-velocity surface layer increases in thickness more rapidly and the very low-velocity region is absent. This asymmetry is also observed in the degree of shear wave splitting, an indicator of anisotropy, in delays of compressional (P) and S waves, in electrical conductivity, and in the rate of subsidence of the seafloor away from the ridge axis (MELT Seismic Team 1998; Evans et al. 1999). S wave tomography indicates that the asymmetry may extend to depths of 200 km or more (Hammond and Toomey 2003). The asymmetry is probably caused by large-scale mantle flow coming from the hotspot region to the west beneath the Pacific plate, coupled with migration of the spreading center to the west. The fast direction for seismic wave propagation, as indicated by shear wave splitting and Rayleigh wave anisotropy, is perpendicular to the East Pacific Rise, consistent with the alignment of olivine crystals in the mantle expected for plate formation and flow from the west.

The East Pacific Rise is one of the fastest spreading ridges, with full spreading rate of about 14 cm/year. Pronounced asymmetry is also observed across the Reykjanes Ridge south of Iceland (Fig. 2), which has a full spreading rate of



Seismic Structure at Mid-Ocean Ridges, Fig. 2 Tomographic cross section of shear wave velocity structure across Reykjanes Ridge south of Iceland. Note strong asymmetry in upper 100 km between North American plate (west) and European plate (east). (After Delorey et al. (2007))

only about 2 cm/year (Delorey et al. 2007). In this case, the tomographic study took advantage of the existence of arrays of stations on Iceland, which straddles the Reykjanes/Mid-Atlantic Ridge. Like the East Pacific Rise, very low shear velocities (~ 4.0 km/s) indicative of the presence of melt are found in a broad region beneath the ridge. For both the East Pacific Rise and the Reykjanes Ridge, there is too little attenuation of surface waves to attribute the very low velocities to the effect of high temperature alone. Velocities at depths shallower than 80 km are lower on the west side, beneath the North American plate, perhaps due to the westward migration of the ridge and upwelling in the mantle in the wake of the thicker North American lithosphere farther west. The anisotropy pattern is different than for typical mid-ocean ridges, perhaps indicating that there is along-axis flow away from the Iceland hotspot in the asthenosphere.

To date, there have been no experiments that provide good control of along-axis variations in mantle structure at depths of tens of kilometers or more where melt production is expected to occur. In the Gulf of California, a surface wave study showed that there are along-axis variations in shear velocity with minima spaced at intervals of about 250 km, perhaps indicative of discrete upwelling centers (Wang et al. 2009), but that spreading system is flanked by continental crust on both sides and dominated by long transform faults, so it may not be typical. At shallower depths just beneath the Moho (the transition from crust to mantle), there are clearly along-axis variations in P-wave velocity on the northern East Pacific Rise, the boundary between the Pacific and Cocos plates (Toomey et al. 2007). Using long offset arrivals refracting from the Moho (P_n) observed in an active experiment, velocity minima were found spaced about 25 km apart (Fig. 3). P-wave velocities in the minima are 7.4 km/s or less, suggesting the presence of 1–3% melt distributed in films or thin sheets (typical P-wave velocities at the Moho are 7.8–8.4 km/s). However, it is not clear whether these apparent centers of melt lie above centers of upwelling mantle or they represent a scale length associated with melt migration.

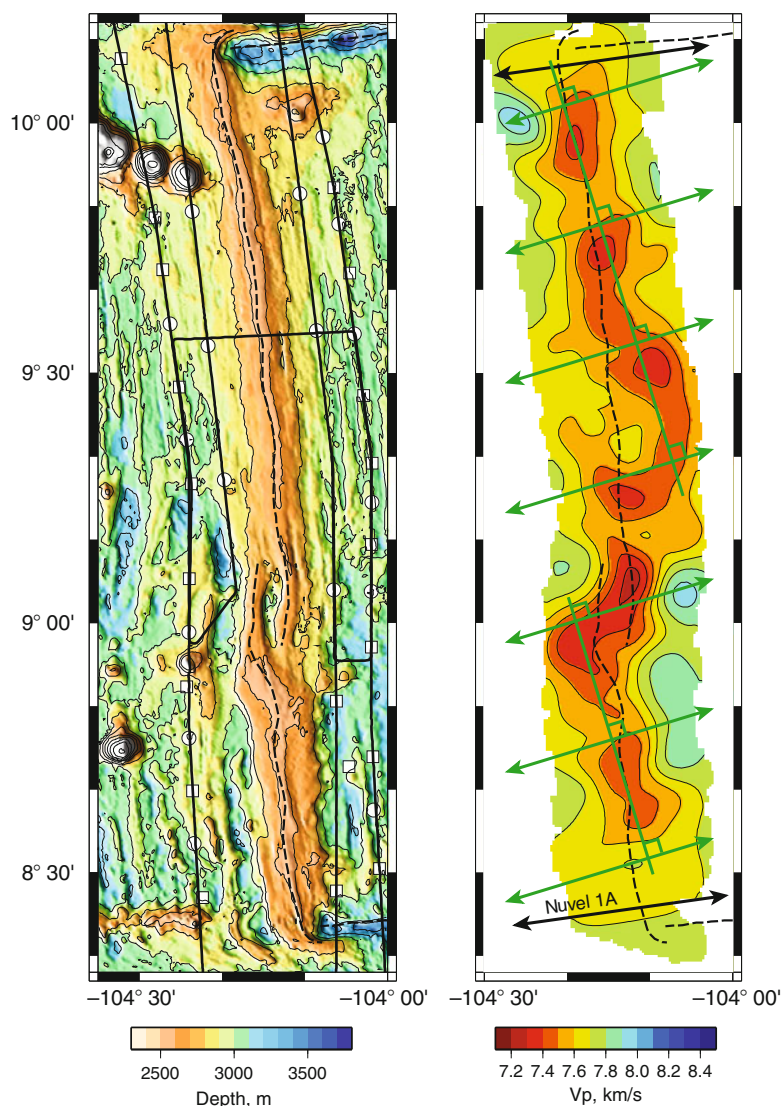
Most of the centers of melt concentration lie beneath or very close to the spreading center determined from detailed bathymetric surveys, but one at $9^{\circ}30'N$ is displaced several kilometers from the axis. Because the fast direction for P-wave propagation is skewed from the spreading direction and is not perpendicular to the strike of the ridge axis, Toomey et al., inferred that upwelling and mantle flow at depth may also be skewed, controlling the location of the velocity minima just below the Moho.

Crustal Structure

Basaltic melt migrates upward through the mantle and is focused at the ridge axis. The mechanisms for focusing are

Seismic Structure at Mid-Ocean Ridges,

Fig. 3 Bathymetry of the East Pacific Rise (*left*) and tomographic image of the mantle 9 km beneath the seafloor (*right*). Dashed lines show plate boundary. Solid lines show locations of air gun shots fired at 500-m intervals. Squares and circles are locations of ocean bottom receivers. Contour interval on tomographic image of P-wave velocity is 0.1 km/s. Green lines with double arrowheads indicate fast direction for anisotropic wave propagation. Black lines with arrows indicate direction of relative motion between the Pacific and Cocos plates. Note that some of the slowest region is displaced from the plate boundary. (After Toomey et al. (2007))



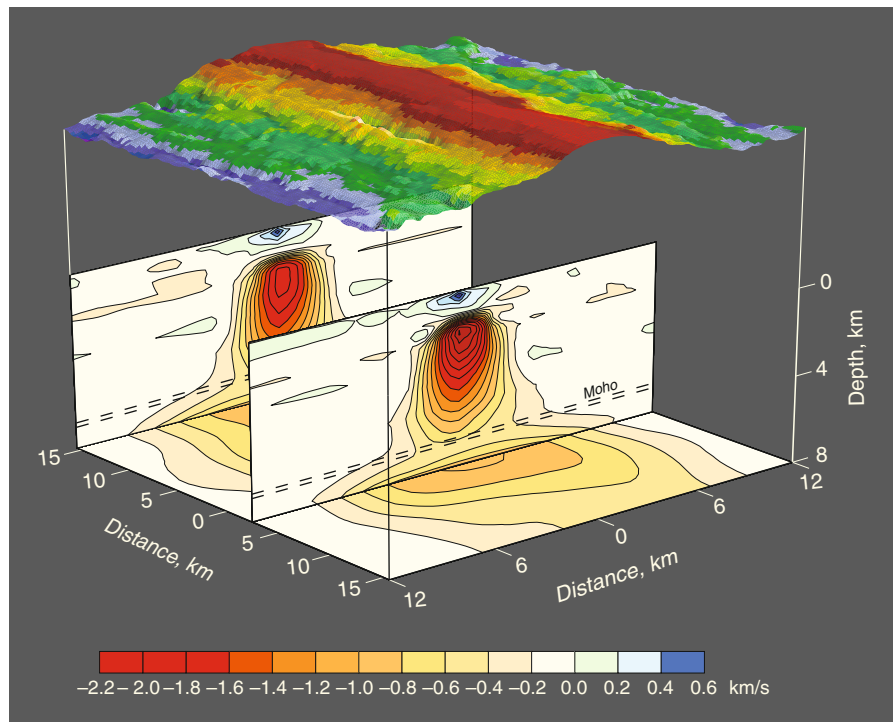
still not well known: there may be melt-rich channels at the base of the lithosphere that guide the melt upward and toward the ridge axis; there may be pressure gradients within the deforming mantle that help push the melt toward the axis; or there may be anisotropic cracks or dunite channels that form easy paths for melt migration. Once the melt reaches crustal levels, there may be redistribution along axis through dikes or a continuous magma chamber. It is clear from seismic studies, however, that there is very little magmatic addition to the crust outside the immediate vicinity of the ridge axis. The crust is essentially full thickness at the spreading center itself (Detrick et al. 1987).

The classic model of oceanic crust is a layer of extrusive basalts at the surface in the form of porous pillow basalts and sheet flows, underlain by a region of sheeted dikes that feed the extrusive layer from a magma chamber, and the lower crust consisting of gabbros that solidify from the magma chamber or underlying mush zone. The seismic structure

has also commonly been described in terms of layers that have often been equated to the lithological layering: layer 2A is a low-velocity layer at the surface several hundred meters thick (layer 1 is sediments that are deposited on top of the basaltic crust); layer 2B is a transition region in which the P velocity increases rapidly downward; and layer 3 has low vertical velocity gradients and high P-wave velocities of 6.5–7.0 km/s. A number of investigations, however, have demonstrated that the seismologic layering does not correspond exactly to the lithological layering and that seismic structure is primarily controlled by porosity. The middle to lower crust, layer 3, contains both dikes and gabbro bodies.

At the axis of fast-spreading ridges, there usually is a low-velocity region only a few kilometers wide at most that represents a zone of partial melting that extends throughout the crust (Fig. 4). At the top of this low-velocity region is a very low velocity layer, which is imaged as a prominent reflector in seismic reflection profiles at depths of 1.5–2.5 km on the

Seismic Structure at Mid-Ocean Ridges, Fig. 4 Three-dimensional view of the anomalous P-wave velocity structure of the East Pacific Rise at 9°30'N (a one-dimensional reference model has been subtracted). (From Dunn et al. (2000).) Overlying bathymetry is shown at top, with red shallow to blue, deep. The red regions below the ridge axis correspond to areas of partial melting. The axial magma chamber is at the top of these anomalously slow regions



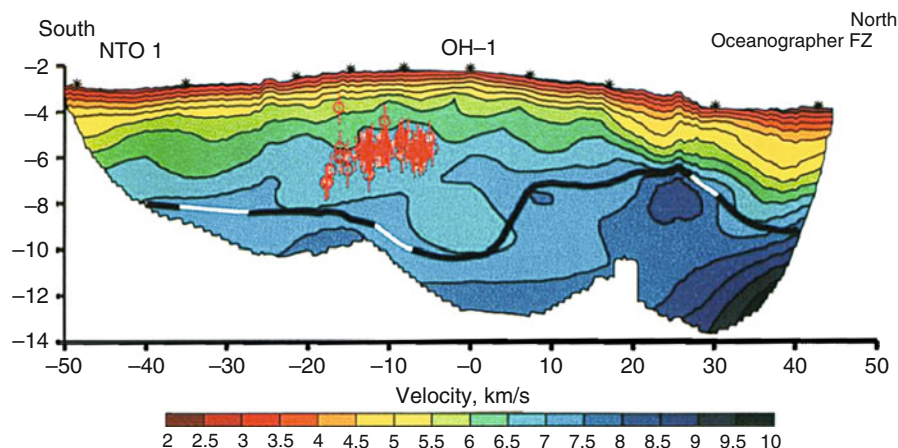
East Pacific Rise (Detrick et al. 1987). The depth to this reflector tends to increase with decreasing spreading rate or proximity to a fracture zone and it is typically absent at slow-spreading ridges (Singh et al. 2006). The width varies from about 250 m to several kilometers. It is interpreted as the top of an axial magma chamber (AMC) or melt sill, with typical thickness less than 100 m. Modeling of P to S conversions in wide-angle reflections indicates that melt content in the AMC varies from nearly 100% to less than 30% along the ridge (Singh et al. 1998; Canales et al. 2006). Estimates of melt content in the deeper crustal mush zone (consisting of a mixture of melt and crystals) are of the order of 2–8% in the lower crust and 3–12% near the Moho (Dunn et al. 2000). The fact that the zone in which melt is present is so narrow even at fast-spreading ridges means that hydrothermal circulation must be very efficient in removing heat, as purely conductive cooling of the crust would result in a much wider zone of partial melting.

Although there are variations in structure along-axis at fast-spreading ridges, the along-axis variations are much more pronounced at segmented, slow-spreading ridges. Typically, the crust thins approaching transform offsets. Within the fracture zone itself, the basaltic crust may be as thin as 1 km or less, but seismically there may be an altered, fractured layer that is a few kilometers thick and characterized by unusually low velocities, so that it looks like crust. The low-velocity region in the fracture zone, however, probably is mostly mantle that is altered by interaction with water

penetrating down cracks that are repeatedly opened by slip along the transform fault. Near the center of ridge segments between two transform offsets, the crust tends to be thicker and lower in velocity than elsewhere (Fig. 5), suggesting that melt is preferentially delivered to the crust from the mantle at that point. The upper crust at slow spreading ridges is anisotropic, with P-waves traveling faster along axis than perpendicular to it, indicating that faults and fissures are preferentially aligned parallel to the spreading center (Barclay and Toomey 2003).

Summary

The seismic velocity structure of mid-ocean ridges is controlled by crustal thickness, cracking or porosity, temperature, melt, and crystal orientation. Low-velocity regions in the mantle that are caused by high temperatures and partial melt indicate that melt is generated in a broad region beneath spreading centers. Asymmetries in the velocity structure show that upwelling and melting beneath ridges are strongly influenced by global mantle circulation and plate motions. Crustal low-velocity regions are concentrated very near the ridge axis, so melt must migrate both vertically and horizontally to the ridge axis from the broad melt production region. The narrowness of the low-velocity zone in the crust requires that hydrothermal circulation must rapidly cool the crust.



Seismic Structure at Mid-Ocean Ridges, Fig. 5 P-wave velocity structure along the Mid-Atlantic Ridge axis from a non-transform offset on the left to the Oceanographer fracture zone on the right. Crustal thickness from wide-angle reflections off the Moho is shown by thick solid line (open where there are gaps in coverage). The crust is thicker

and the lower crust has lower velocity near the center of the segment. (From Hooft et al. (2000)). The red circles are locations of microearthquakes with uncertainties indicated by error bars. Location of ocean-bottom hydrophones indicated by asterisks

Along-axis variations, particularly at slow-spreading ridges, suggest that melt is preferentially delivered to the center of ridge segments.

Cross-References

- [Crustal Reflectivity \(Oceanic\) and Magma Chamber](#)
- [Earth's Structure, Upper Mantle](#)
- [Lithosphere, Mechanical Properties](#)
- [Lithosphere, Oceanic](#)
- [Lithosphere, Oceanic: Thermal Structure](#)
- [Magnetic Anisotropy](#)
- [Ocean Bottom Seismics](#)
- [Seafloor Spreading](#)
- [Seismic Data Acquisition and Processing](#)
- [Seismic Tomography](#)
- [Seismic, Velocity, and Density Relationships](#)
- [Surface Waves](#)
- [Traveltime Tomography Using Controlled-Source Seismic Data](#)

Bibliography

- Barclay AH, Toomey DR (2003) Shear wave splitting and crustal anisotropy at the Mid-Atlantic Ridge, 35°N. *J Geophys Res* 108. <https://doi.org/10.1029/2001JB000918>
- Canales JP, Singh SC, Detrick RS, Carbotte SM, Harding A, Kent GM, Diebold JB, Babcock J, Nedimovic MR (2006) Seismic evidence for variations in axial magma chamber properties along the southern Juan de Fuca Ridge. *Earth Planet Sci Lett* 246:353–366
- Delorey AA, Dunn RA, Gaherty JB (2007) Surface wave tomography of the upper mantle beneath the Reykjanes ridge with implications for ridge-hot spot interaction. *J Geophys Res* 112:B08313. <https://doi.org/10.1029/2006JB004785>
- Detrick RS, Buhl P, Vera E (1987) Multi-channel seismic imaging of a crustal magma chamber along the East Pacific Rise. *Nature* 326:35–41
- Dunn RA, Toomey DR, Solomon SC (2000) Threedimensional seismic structure and physical properties of the crust and shallow mantle beneath the east pacific rise at 9 degrees 30'N. *J Geophys Res* 105:23537–23555
- Evans RL et al (1999) Asymmetric electrical structure in the mantle beneath the east pacific rise at 17°S. *Science* 286:756–759
- Hammond WC, Toomey DR (2003) Seismic velocity anisotropy and heterogeneity beneath the mantle electromagnetic and tomography experiment (MELT) region of the East Pacific Rise from analysis of P and S body waves. *J Geophys Res* 108. <https://doi.org/10.1029/2002JB001789>
- Harmon N, Forsyth DW, Weeraratne DS (2009) Thickening of young Pacific lithosphere from high-resolution Rayleigh wave tomography: a test of the conductive cooling model. *Earth Planet Sci Lett* 278:96–106
- Hooft EEE, Detrick RS, Toomey DR, Collins JA, Lin J (2000) Crustal thickness and structure along three contrasting spreading segments of the Mid-Atlantic Ridge, 33.5°–35°N. *J Geophys Res* 105:8205–8226
- MELT Seismic Team (1998) Imaging the deep seismic structure beneath a Mid-Ocean Ridge: the MELT experiment. *Science* 280:1215–1218
- Singh SC, Kent GM, Collier JS, Harding AJ, Orcutt JA (1998) Melt to mush variations in crustal magma properties along the ridge crest at the southern East Pacific Rise. *Nature* 394:874–878
- Singh SC, Crawford WC, Carton H, Seher T, Combier V, Cannat M, Canales JP, Dusanur D, Escartin J, Miranda JM (2006) Discovery of a magma chamber and faults beneath a Mid-Atlantic Ridge hydrothermal field. *Nature* 442:1029–1032
- Toomey DR, Jousset D, Dunn RA, Wilcock WSD, Detrick RS (2007) Skew of mantle upwelling beneath the East Pacific Rise governs segmentation. *Nature* 446:409–414
- Wang Y, Forsyth DW, Savage B (2009) Convective upwelling in the mantle beneath the Gulf of California. *Nature* 462:499–501. <https://doi.org/10.1038/nature08552>

Seismic Tomography

Guust Nolet

IRD Geoazur and Université Côte d'Azur, Sophia Antipolis, France

Definition

The term tomography derives from the Greek *τόμος*, or slice. “Seismic tomography” is used for a variety of methods that use transmitted seismic waves to estimate the spatial variations in properties (wave velocity, density, and attenuation) inside the Earth, which are often represented as images of two-dimensional cross-sections or “slices.” It is conceptually different from seismic migration, which uses reflected waves to image sharp discontinuities.

History

In 1971, P. Bois at the Institut Français de Pétrole was the first to suggest the tomographic method in order to locate the causes of delays in seismic waves between two boreholes. His paper predates many future developments but was initially written in French and remained largely unnoticed. In the mid-1970s, Keiti Aki from MIT applied a linear inversion to locate velocity heterogeneities beneath large nuclear monitoring arrays in Norway and Montana, and Harvard's Adam Dziewonski began interpreting the time residuals published by the International Seismological Center (ISC) in the UK in terms of global velocity anomalies.

In 1982, Guy Masters and his colleagues at the Scripps Institution of Oceanography discovered a strong degree-2 component in the geographical distribution of the slight shifts in the spectral peaks of the Earth's normal modes. Since then, the Earth's free oscillations have contributed to constrain the heterogeneity in the Earth at the longest wavelengths and as deep as the inner core. By 1984, John Woodhouse and Adam Dziewonski at Harvard published a first global model for shear velocity in the upper mantle based on long-period surface waves.

However, to image smaller scale anomalies, the shorter wavelengths of P and S-waves are indispensable. In particular, Steve Grand at the University of Texas, Rob van der Hilst and Wim Spakman at Utrecht University and Yoshio Fukao and colleagues at the University of Tokyo pioneered high-resolution body-wave tomography using iterative solvers for the huge systems of linearized equations and established in the early 1990s that some, but not all, subducting slabs are able to sink well into the depths of the lower mantle.

Thermal plumes in the lower mantle were for the first time reliably imaged in 2003 by Raffaella Montelli, using a new

technique of finite-frequency tomography developed by Tony Dahlen and Guust Nolet and their collaborators at Princeton University.

For references and a detailed account of the history of seismic tomography, see the reviews by Romanowicz (2003) and Rawlinson et al. (2010). Nolet (2008) provides a general introduction into the methods of seismic tomography, including the theoretical aspects that are here discussed only briefly.

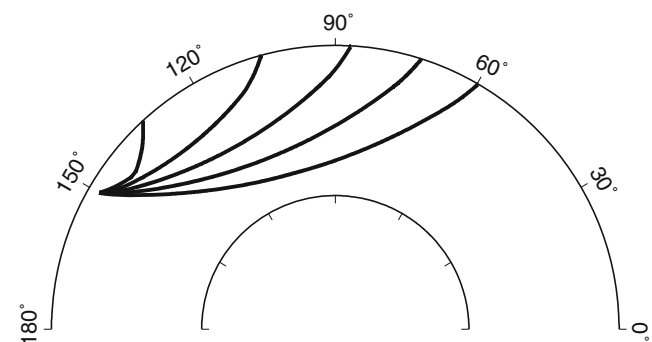
Onset Times

Much of seismic tomography is based on estimating the arrival time of a seismic body wave by picking the onset of a phase on the seismogram, and interpreting the travel time T using the infinite-frequency approximation of ray theory:

$$T = \int_{\text{raypath}} \sigma(\mathbf{r}) \, ds \quad (1)$$

where the raypath is determined by Snel's law and where σ is the inverse velocity v^{-1} of the wave at the location $\mathbf{r}$. The raypaths may be located between two boreholes, e.g., to monitor the exploitation of an oil or gas field, between several explosive sources and an array of seismographs at the surface (see article: Travel Time Tomography Using Controlled-source Seismic Data), between a seismogenic zone and a local array of seismographs, or between an earthquake and the global network of seismic stations.

The raypath is often approximated by the path calculated for a spherically symmetric Earth or a horizontally layered background model. Examples of global raypaths are given in Fig. 1. This approximation is permitted because ray trajectories render the travel time stationary, such that small deviations in the true ray path location only cause second order errors in the travel time calculated with Eq. 1, a property of rays known by the name of *Fermat's Principle*. In practice,



Seismic Tomography, Fig. 1 Examples of raypaths for a P wave in an equatorial cross-section from a hypothetical earthquake at a depth of 248 km located at 0°N, 150°W

one usually keeps the raypath fixed to its initial approximation. This keeps the inverse problem linear. One then inverts for the difference between the true Earth and a starting- or background model: $\delta\sigma(\mathbf{r}) = \sigma(\mathbf{r}) - \sigma_0(\mathbf{r})$, using the difference between observed time T , and its prediction T_0 from the background model as datum: $\delta T = T - T_0$:

$$\delta T = \int_{\text{raypath}} \delta\sigma(\mathbf{r}) d\mathbf{s} \quad (2)$$

For strongly heterogeneous regions such as subduction zones, one may need to use three-dimensional ray tracing and an iterative approach using Eq. 2 in which rays are recalculated after each modification of the model $\sigma(\mathbf{r})$.

Model Parameterization and Inversion

The change in the model $\delta\sigma(\mathbf{r})$ can be described by a finite number of parameters if it is developed in terms of a basis of M interpolating or “basis” functions $h_k(\mathbf{r})$:

$$\delta\sigma(\mathbf{r}) = \sum_{k=1}^M m_k h_k(\mathbf{r}) \quad (3)$$

The basis functions may be represented by homogeneous cells, linear interpolators in a tetrahedral or other mesh, spherical harmonics, or 3D wavelets. Substitution of Eq. 3 in Eq. 2 gives a formal system of linearized equations for N estimated travel times δT_i , $i = 1, \dots, N$, arranged in a vector $\delta \mathbf{T}$:

$$\delta \mathbf{T} = \mathbf{A} \mathbf{m} \quad (4)$$

where $A_{ik} = \int_{\text{raypath}} h_k(\mathbf{r}) d\mathbf{s}$. Since many raypaths may overlap, this system is usually overdetermined and needs to be solved by a least squares solver that minimizes χ^2 , the length of the misfit vector weighted by the data standard error e_i :

$$\chi^2(\mathbf{m}) = \sum_{i=1}^N \left(\frac{\sum_{k=1}^M A_{ik} m_k - \delta T_i}{e_i} \right)^2 \quad (5)$$

Mathematically, this is accomplished by first dividing the system of Eq. 4 by the standard error of the data (i.e., multiplying by the inverse square root of the covariance matrix $\mathbf{C}$ which is generally assumed to be diagonal), and back-projecting the system using the transpose $\mathbf{A}^T$ of $\mathbf{A}$:

$$\mathbf{A}^T \mathbf{C}^{-\frac{1}{2}} \mathbf{A} \mathbf{m} = \mathbf{A}^T \mathbf{C}^{-\frac{1}{2}} \delta \mathbf{T} \quad (6)$$

One often scales the equations a priori such that the data have unit error $e_i = 1$ and $\mathbf{C} = \mathbf{I}$, which has the same effect. In

the following we therefore ignore $\mathbf{C}$. The overdetermined part of the data allows one to obtain objective estimates for the standard errors in the data (Voronin et al. 2014).

Usually, Eq. 6 is at the same time overdetermined for some parameters and underdetermined for others (the determinant of $\mathbf{A}^T \mathbf{A}$ being zero): there are in that case infinitely many models that satisfy the data within the measurement uncertainty and one needs to regularize the solution. This can be done by choosing the solution that minimizes a weighted penalty of χ^2 , model norm and model roughness, by strongly truncating a development in spherical harmonics or by choosing the sparsest decomposition in wavelets that still satisfies the data. This generally involves a subjective choice between the degree of detail one allows in a model and the goodness of fit to the data. Generally, one aims for a χ^2 approximately equal to the number of data N – in other words, one attempts to fit the data at the level of about one standard deviation. Invariably, a trade-off exists between the detail allowed in the model and the statistical precision with which the model parameters can be determined: the sharper the detail, the more uncertain the m_k .

Regularization can also be done in a fundamentally different way by inverting for local averages in $\delta\sigma(\mathbf{r})$ with a prior specified variance that can be estimated linearly from the data (Backus-Gilbert theory). This can only be done at the expense of a significant increase in computing time. In this case, the tradeoff is between the size of the averaging volume and the model variance: averages over larger volumes are determined with smaller statistical uncertainty. Though computationally expensive, this is statistically superior to any of the other methods. A stable discretized version of Backus-Gilbert theory was developed by Nolet (1985) and for the first time applied to a large tomographic system by Zoroli (2016).

Whatever method is used, the system of Eq. 4 may be very large (e.g., more than 10^6 data for more than 10^5 unknown model parameters). Local parameterizations (cells, as opposed to spherical harmonics) render $\mathbf{A}$ sparse, and the system can be solved efficiently using iterative solvers that adapt to sparse matrices, such as LSQR (Paige and Saunders 1982).

Normal Modes and Surface Waves

The eigenfrequencies ${}_n\omega_\ell^m$ of the Earth are characterized by three quantum numbers ℓ , m , and n , related to the number of nodal surfaces in the displacement field of the Earth with latitude, longitude, and depth, respectively. For a non-rotating, isotropic, spherically symmetric Earth, the spectrum is degenerate in the sense that the frequency is independent of the azimuthal order m . For the real Earth, a weak dependence on m splits each eigenfrequency into $2\ell + 1$ separate frequencies that are too closely spaced to be resolvable except for the

very lowest angular order ℓ . Instead, a composite spectral line or “multiplet” is observed with a peak that depends on the location of the seismic station – a direct consequence of the fact that the $2\ell + 1$ single peaks have amplitudes that depend on geographical location by virtue of their spherical harmonic dependence on latitude and longitude.

Two major strategies exist to exploit the small fluctuations in the spectrum. Decomposing the free oscillation into surface waves traveling in opposite directions, ray theory may be used to establish a linear relationship between the heterogeneity along the great circle between source and receiver and the observed peak shift in the spectral line. At higher frequency, we separate single passages of the surface wave and exploit the linear relationship between fluctuations in the observed phase velocity and the Earth’s heterogeneity. In both cases, the relationship between the Earth’s heterogeneity and the observed datum is a two-dimensional integral along the great circle of the form:

$$\delta\omega = \int_0^a \int_{\text{gc}} K(\mathbf{r}) \delta\sigma(\mathbf{r}) d\mathbf{s} d\mathbf{r}, \quad (7)$$

where the kernel $K(\mathbf{r})$ is computed using first order perturbation theory of the differential equations governing the Earth’s free oscillations.

Alternatively, we may exploit the known distribution of amplitudes of single peaks over the surface of the Earth to invert for the location of these peaks (“mode splitting”). The small frequency shifts $\delta\omega^m$ are themselves the eigenvalues of a splitting matrix $\mathbf{H}$ of which the elements are linearly related to the variation of density and elastic parameters in the Earth. We can estimate $\mathbf{H}$ from the seismic data using autoregressive filtering techniques. This way we avoid any ray-theoretical approximations and obtain a three-dimensional integral constraint that can be used to solve for the large-scale variations in the Earth’s density and elastic properties.

Finite-Frequency Tomography

Modern, broadband digital instrumentation allows for a robust estimation of the delay of an observed seismic wave $s_1(t)$ with respect to a theoretically predicted (“synthetic”) waveform $s_2(t)$ or with respect to the same phase observed elsewhere, by locating the maximum in the cross-correlation $C(t)$ between the two signals:

$$\delta T = \arg \max_{t \in [\tau_1, \tau_2]} C(t), C(t) = \int_{\tau_1}^{\tau_2} s_1(\tau) s_2(\tau - t) d\tau, \quad (8)$$

where the integration interval (τ_1, τ_2) usually (but not necessarily) extends over the window containing a distinct arrivals such as P or S. The cross-correlation delay is influenced by the

distribution of energy within this window and thus fundamentally different from the delay measured by picking the onset of $s_1(t)$. Instead of a “shortest time” satisfying Snell’s law, it represents an integral measure over a time window that is at least as long as the dominant period of $s_1(t)$. While Eq. 1 is correct for onset times, it needs to be modified for delays measured by cross-correlation. Heterogeneities inside the Earth may influence the integral by scattering waves in the direction of the recording station that arrive within the time window. The most serious consequence is that energy may diffract around a small heterogeneity, thus masking or dominating the slower or faster arrival that has crossed the heterogeneity. This phenomenon of “wavefront healing” biases observed delay times to zero, especially if the anomaly is located at some distance from the source and receiver. Nolet and Dahlen (2000) give a theoretical analysis of wavefront healing.

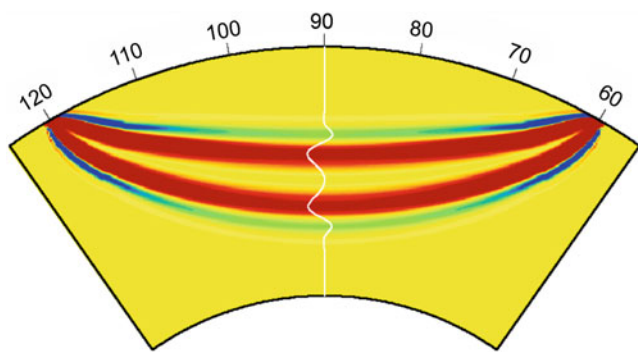
Taking wave diffraction into account leads to a three-dimensional integral constraint of the form:

$$\delta T = \int_V K(\mathbf{r}) \delta\sigma(\mathbf{r}) d^3\mathbf{r}, \quad (9)$$

where the kernel $K(\mathbf{r})$ can be efficiently calculated using ray theory for scattered waves. The effective volume of integration V is roughly equivalent to that of the Fresnel zone of the ray, where the sensitivity is largest. Since a positive change in rock velocity gives a negative change in the travel time, the kernel values within the Fresnel zone are negative. There are, however, sidelobes with a positive sign, where a positive velocity anomaly would give a positive change in the observed travel time. This may seem counter-intuitive, but it is possible because cross-correlations involve a finite time window, and it is the energy distribution in this window, rather than the onset of the wave, that defines the delay measured in this way.

An example of a finite-frequency kernel is shown in Fig. 2. Remarkably, the cross-correlation travel time is insensitive to perturbations in the Earth’s properties at the center of the kernel, which is at location of the ray path. Because of this hole in the sensitivity and their general shape, finite-frequency kernels are often referred to as “banana-doughnut” kernels. Numerical tests have shown that the finite-frequency kernels accurately model the loss of signal caused by wavefront healing, the gradual reduction of a delay caused by waves diffracting around a small heterogeneity. By estimating delays in different frequency bands, one gains information about the size of the heterogeneity in the Earth.

Though finite-frequency kernels were initially proposed to provide a better theoretical modeling of long-period signals affected by wavefront healing, they offer the advantage of being able to model the healing as a function of frequency, and thus obtain independent information on the size of the



Seismic Tomography, Fig. 2 The sensitivity of the cross-correlation time to the earth structure, for a long-period P wave from a surface source at 0° N, 120° W, recorded at a distance of 60°. The kernel is plotted in an equatorial cross-section of the mantle, the color scale runs from large negative values (red) to positive (blue). The sensitivity is zero (yellow) at the location of the geometrical ray-path, forming a “doughnut hole” in the banana-shaped kernel. The white line denoted the kernel values at 90°

heterogeneity. Such multiple-frequency tomography may significantly increase resolution (Sigloch et al. 2008). Finite-frequency sensitivity can also be formulated for the amplitude perturbations caused by attenuation and by focusing/defocusing (focusing cannot be handled with ray-theory because ray theory does not model amplitudes correctly at low frequency and is very nonlinear at high frequency).

Instead of inverting for a cross-correlation delay or an amplitude perturbation, one can formulate a 3D sensitivity directly for the observed time series, though this has the disadvantage that the observed and predicted waveforms have to be close enough in phase that the phase difference can be adequately modeled by adding a small perturbation to the waveform itself (in fact approximating $e^{i\phi} \approx 1 + i\phi$). This limits the inversion to waveforms with small time mismatches, in contrast to the delays estimated through Eq. 8, which remain linear for large anomalies.

All types of data – delays, amplitudes, and waveform mismatches – can be iteratively inverted using an adjoint approach, in which best fitting model is sought along the gradient $\nabla_m \chi^2$ (Tromp et al. 2005), e.g., for delay times scaled to unit variance:

$$\nabla_m \chi^2 = \mathbf{A}^T (\mathbf{A} \mathbf{m} - \delta \mathbf{T}) \quad (10)$$

For waveforms, the matrix $\mathbf{A}$ is in this case the matrix representation of a finite-difference or spectral-element algorithm that produces the predicted seismograms and $\mathbf{A}^T$ is the adjoint operator that projects the seismograms back in time. Since this allows one to backproject all residual seismograms for one earthquake with only one calculation, and since only the product of the adjoint matrix and the residual data vector is needed, this allows us to avoid computing the partial

derivatives separately for each datum δT_i with respect to each model parameter m_k . The adjoint approach has some advantages over a direct matrix inversion for large-scale problems, in particular those with few sources and many stations, because it does not require the storage of a large matrix, which is especially important for waveform inversions. Since the gradient is recalculated for every iteration, the method lends itself naturally to tomographic inversions that are strongly nonlinear, such as one expects for imaging of the crust or the D'' region just above the Earth's core. However, Mercier and Nolet (2013) showed that the inversion of delay times measured by cross-correlation are highly linear in most depth regions of the Earth. For a technical description of the adjoint method (also referred to as “residual back-propagation”) see the article: Seismic, Waveform Modeling and Tomography.

Summary

The equations of seismic tomography are integral equations; the observations are weighted averages of the properties of the Earth (Eq. 9). Though often treated as line integrals, assuming ray theory is valid, a delay of a 1 Hz P wave measured by cross-correlation senses a volume or Fresnel zone inside the Earth that is several hundred km wide. The inversion of integral equations demands care to avoid that noise (errors in the data) are interpreted as small differences in observed delays, which may require large variations in the Earth's properties. This imposes a fundamental limitation to the resolving power of seismic waves that even finite-frequency theory or waveform modeling cannot completely avoid. The theoretically best available horizontal resolution, using the highest observable seismic frequencies, is of the order of hundred km in the lower mantle. In practice, this lower limit is not yet reached because of the limited coverage of the Earth's surface with seismic stations. On land, the coverage can be improved markedly using temporary deployments of seismic arrays, but this is much more difficult and expensive to do in the oceanic domain. Major improvements on a global scale are therefore to be expected only if we solve the problem of oceanic seismometry. It is expected that floating seismographs (see article “Mermaids”), and possibly also strain signals extracted from fiber optic cables on the ocean floor, will soon make this possible.

Cross-References

- [Earth, Density Distribution](#)
- [Earth's Structure, Core](#)
- [Earth's Structure, Global](#)
- [Earth's Structure, Lower Mantle](#)

- Earth's Structure, Upper Mantle
- Floating Seismographs (MERMAIDS)
- Gravity, Global Models
- Inverse Theory, Linear
- Mantle Plumes
- Plate-Driving Forces
- Propagation of Elastic Waves: Fundamentals
- Seismic Imaging, Overview
- Seismic Ray Theory
- Seismic, Waveform Modeling and Tomography
- Subduction Zones
- Surface Waves
- Traveltimes Tomography Using Controlled-Source Seismic Data

Bibliography

- Aki K, Lee HK (1976) Determination of three-dimensional anomalies under a seismic array using first P arrival times from local earthquakes: 1. A homogeneous initial model. *J Geophys Res* 81:4381–4399
- Bois P, la Porte M, Lavergne M, Thomas G (1971) Essai de détermination automatique des vitesses sismiques par mesures entre puits. *Geophys Prospect* 19:42–81
- Dziewonski AM, Hager BH, O'Connell RJ (1977) Large-scale heterogeneities in the lower mantle. *J Geophys Res* 82:239–255
- Masters G, Jordan TH, Silver PG, Gilbert JF (1982) Aspherical earth structure from spheroidal mode oscillations. *Nature* 298:609–613
- Mercerat ED, Nolet G (2013) On the linearity of cross-correlation delay times in finite-frequency tomography. *Geophys J Int* 192:681–687
- Montelli R, Nolet G, Dahlen FA, Masters G, Engdahl ER, Hung SH (2004) Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* 303:338–343
- Nolet G (1985) Solving or resolving inadequate and noisy tomographic systems. *J Comput Phys* 61:463–482
- Nolet G (2008) A breviary of seismic tomography. Cambridge University Press, Cambridge, UK
- Nolet G, Dahlen FA (2000) Wave front healing and the evolution of seismic delay times. *J Geophys Res* 105:19043–19054
- Paige CC, Saunders M (1982) LSQR: an algorithm for sparse, linear equations and sparse least squares. *J ACM Trans Math Softw* 8:43–71
- Rawlinson N, Pozgay S, Fishwick S (2010) Seismic tomography: a window into deep earth. *Phys Earth Planet Inter* 178:101–135
- Romanowicz B (2003) Global mantle tomography: progress status in the past 10 years. *Annu Rev Earth Planet Sci* 31:303–328
- Sigloch K, McQuarrie N, Nolet G (2008) Two-stage subduction history under North America inferred from multiple-frequency tomography. *Nat Geosci* 1:458–462
- Tromp J, Tape C, Liu Q (2005) Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophys J Int* 160:195–216
- Voronin S, Mikesell D, Slezak I, Nolet G (2014) Solving large tomographic linear systems: size reduction and error estimation. *Geophys J Int* 199:276–285
- Woodhouse JH, Dziewonski AM (1984) Mapping the upper mantle: three dimensional modelling of the earth structure by inversion of seismic waveforms. *J Geophys Res* 89:5953–5986
- Zaroli C (2016) Global seismic tomography using Backus-Gilbert inversion. *Geophys J Int* 207:876–888

Seismic Velocity and Temperature Relationships

Kalachand Sain

Wadia Institute of Himalayan Geology, Dehradun, India

Definition

Seismic velocity is defined as the speed with which a seismic wave travels through a medium and is measured in km/s, whereas temperature is a physical property that quantitatively expresses hot and cold. It is the manifestation of thermal energy, present in all matter, and is expressed in °C. Seismic velocities of subsurface rocks decrease with temperature but experience opposite effects due to increase in pressure (depth).

Velocity-temperature equation: On an average, the pressure increases at a rate of 30 MPa/km and the temperature raises at a rate of 25 °C/km from surface to deeper parts of the Earth. Since seismic velocity (V) decreases with temperature (T) and increases with pressure (P), we need to know the combined effects of P and T for interpreting seismic velocities at different depths. The generalized relation for the variation of velocity with depth Z can be expressed as

$$\frac{dV}{dZ} = \left(\frac{dV}{dP}\right)_T \frac{dP}{dZ} + \left(\frac{dV}{dT}\right)_P \frac{dT}{dZ} \quad (1)$$

where $\left(\frac{dV}{dP}\right)_T$ denotes the change in velocity with pressure at constant temperature (isotherm), and $\left(\frac{dV}{dT}\right)_P$ is the change in velocity with temperature at constant pressure (isobar). $\frac{dP}{dZ}$ and $\frac{dT}{dZ}$ are the vertical pressure and temperature gradients respectively.

To make petrological inferences, the crustal and lithospheric seismic velocities are to be corrected to the experimental reference values (i.e., constant pressure of 100 MPa and room temperature of 20 °C). Rybach and Buntebarth (1984) have defined the correction factor (f) to the P-wave velocity (V_P) at a given P and T as

$$\begin{aligned} V_P(20^\circ\text{C}, 100\text{MPa}) &= V_P(P, T)f \\ &= V_P \left[1 + \Delta T \frac{\partial V_P}{\partial T} - \Delta P \frac{\partial V_P}{\partial P} \right] \end{aligned} \quad (2)$$

The correction factors for granites, gneisses, amphibolites, gabbros, and ultrabasic rocks of both Precambrian and Phanerozoic period are given in Table 2 of Rybach and Buntebarth (1984). As per eq. (2), the corrections to the field velocities at depths from 1 to 50 km are of the order of ± 0.1 km/s. As a

matter of fact, the increase in velocity with pressure is nearly compensated by the decrease in velocity due to temperature increasing with depth.

Velocity variation with temperature for some reservoir rocks: The change in seismic velocities with temperature also depends on the saturation of rock. The results of Wang and Nur (1990) in heavy oil sands and hydrocarbon-saturated rocks show that both V_P and V_S decrease with increasing temperature. Since V_S is not affected by fluids, the decrease in V_S is due to changes in rock frame and in rock fluid interactions. The V_P in heavy oil sands (Tosaya et al. 1987) shows a dramatic decrease. As temperature increases from 25 °C to 125 °C, V_P can drop by 35% to almost 90%. Heavy oils are highly viscous and a strong interfacial force exists between oil and rock grains. The viscosity of oil and interfacial force decrease due to rise in temperature, which decreases the rigidity and bulk modulus leading to reduction in seismic velocities. For temperature up to about 150 °C, changes in pore fluid properties play dominant role in changing the velocity. Though there is a negative correlation between seismic velocity and temperature, Ryan and Shalev (2014), from drilling results in Montserrat (West Indies), showed that above ~220 °C there is a positive correlation into a geothermal reservoir. They hypothesized that anomalous variation is controlled by the first-order hydrothermal mineral assemblage.

Perry et al. (2006) compared temperature and velocity of Moho refracted phase, Pn, for various mineralogical models of the upper mantle from seismic refraction and heat flow studies and demonstrated that the theoretical velocity-temperature conversion used for interpretation of seismic tomographic models is in good agreement with the observations. Schutt et al. (2018) used the velocity-temperature relationship in modeling the deep geotherm within few km of the uppermost mantle in western USA using the Pn velocity.

The temperature is a fundamental factor in determining the lithospheric strength, thickness, buoyancy, and mass flux for geoscientific investigation of the subsurface. For example, if temperature of the lower crust is found sufficiently warm, the lower crust may be weak that can decouple the upper crustal and mantle strain fields, affecting the style of deformation. Based on deep seismic studies and gravity modeling in a huge territory of north Eurasia, Yegorova and Pavlenkova (2015) showed that seismic velocities in the upper mantle mainly reflect the temperature regime with no direct velocity-density relationship, as characterized by high-velocity, low-density and low heat flow.

Acknowledgments The author is grateful to the Director of Wadia Institute of Himalayan Geology, Dehradun, for his consent to publish this work. This has a Wadia contribution WIHG/0037.

Bibliography

- Perry HKC, Jaupart C, Mareschal JC, Shapiro NM (2006) Upper mantle velocity-temperature conversion and composition determined from seismic refraction and heat flow. *J Geophys Res* 111(B07301):1–14
- Ryan GA, Shalev E (2014) Seismic velocity/temperature correlations and a possible new geothermometer: insights from exploration of a high-temperature geothermal system on Montserrat, West Indies. *Energies* 7(10):6689–6720
- Rybach L, Buntebarth G (1984) The variation of heat generation, density and seismic velocity with rock type in the continental lithosphere. *Tectonophysics* 103:335–344
- Schutt DL, Lowry AR, Buehler JS (2018) Moho temperature and mobility of lower crust in the western United States. *Geology* 46:1–4
- Tosaya C, Nur A, Vo-Thanh D, Da Prat G (1987) Laboratory seismic method for remote monitoring of thermal EOR. *SPE Reserv Eng* 2:238–242
- Wang Z, Nur A (1990) Wave velocities in hydrocarbon saturated rocks: experimental results. *Geophysics* 55:723–733
- Yegorova TP, Pavlenkova GA (2015) Velocity-density models of the Earth's crust and upper mantle from the quartz, craton, and kimberlite superlong seismic profiles. *Phys Solid Earth* 51:250–267

Seismic Viscoelastic Attenuation

Vernon F. Cormier

Physics Department, University of Connecticut,
Storrs, CT, USA

Synonyms

Seismic intrinsic attenuation

Definition

Linear viscoelastic attenuation. The fractional loss of seismic energy in a material in which elastic deformation (strain) induced by one cycle of a seismic wave or mode lags in time the applied stress associated with the wave or mode.

Apparent seismic attenuation. The loss of energy in a propagating seismic wave or standing mode due to viscoelasticity combined with the loss of scattered energy redistributed in time and space by heterogeneity.

Introduction

The amplitude of seismic waves decreases with increasing distance from earthquake, explosion, and impact sources. How this amplitude decrease occurs and how it depends on frequency of the seismic waves are fundamentally important to the efforts to describe Earth structure and seismic sources.

The decay of amplitude of seismic waves with increasing distance of propagation through Earth is known as seismic wave attenuation. The attenuation occurring under high-temperature rheological conditions in the Earth's interior can be called seismic viscoelastic attenuation.

Seismic attenuation and its variation with location within Earth are useful for determining the anelastic properties of Earth as a function of depth. Seismic attenuation also shows large lateral variations that can be related to lateral variations in geological and geophysical properties not as easily detected by measurement of seismic velocities. In addition to providing information on a physical property, research in seismic attenuation has also been strongly motivated by more practical problems. One problem has been the prediction of ground motion due to probable earthquakes in different regions. The frequency content and decay with distance of this strong ground motion is an important input to the design of earthquake-resistant structures and to disaster forecasting (see ► [“Earthquakes, Strong-Ground Motion”](#)). Another problem has been to estimate the size and detectability of underground nuclear tests (see ► [“Seismic Monitoring of Nuclear Explosions”](#)).

How Do Seismic Waves Attenuate?

The attenuation of seismic waves is due to three effects: geometric spreading, intrinsic attenuation, and scattering.

Geometric Spreading

Geometric spreading leads to an energy density decrease that occurs as an elastic wave front expands with increasing distance from its source. In a homogeneous Earth of constant velocity and density, the geometric spreading of a seismic body wave is proportional to the reciprocal of the distance between source and receiver. In the real Earth, velocity and density vary strongly with depth and less so laterally. Given a model of this variation, however, the geometric spreading of a body wave can be easily calculated (see ► [“Seismic Ray Theory”](#)).

Intrinsic Viscoelastic Attenuation

Intrinsic (viscoelastic) attenuation occurs at high temperatures due to internal friction during the passage of an elastic wave. It is controlled by the thermal and defect properties of the medium in which the wave is propagating. It can result in a phase lag between strain and stress giving rise to strain energy dissipation and associated frequency dependence (dispersion) of the relevant modulus or speed of the propagating elastic wave. The microscopic mechanisms of intrinsic attenuation have been described in several different ways, including the resistive and viscous properties of oscillator models of the atoms in crystalline lattices, the movement of interstitial fluids

between grain boundaries and cracks (O'Connell and Budiansky 1977), and the frictional sliding of cracks. Jackson (1993, 2007) reviews laboratory experiments that investigate microscopic mechanisms of intrinsic attenuation. This entry concentrates on the measurement of intrinsic attenuation from recordings of seismic waves at great distance.

Scattering Attenuation

Scattering attenuation occurs when elastic energy is scattered and redistributed into directions away from the receiver or into waves arriving in later time windows at the receiver (see ► [“Seismic Waves, Scattering”](#)). Scattering takes place by reflection, refraction, and mode conversion of elastic energy by wavelength-scale irregularities in the medium. These irregularities are discontinuous or rapid variations in the velocity and/or density of the medium. In the crust and uppermost mantle, variations in velocity and density can be particularly strong in the lateral as well as the vertical direction.

Linear Viscoelasticity

Rheology

A stress is a vector force per unit area applied to a solid. A strain is nondimensional measure of the deformation of the solid due to the applied stress, such as the change in a length element divided by the original length. The equation that relates stress and strain is sometimes termed the *rheology* or the *constitutive* relation (see ► [“Mantle Viscosity”](#)). A linear viscoelastic rheology can be described by a linear differential equation:

$$L_1 \sigma(t) = L_2 \varepsilon(t) \quad (1)$$

where L_1 and L_2 are any linear combinations of operators of the time, $\frac{d^n}{dt^n}$ or $\int dt^n$. This type of equation can describe both the elastic strain of a material over a short time interval of applied stress as well as its viscous behavior and flow over a longer time interval (Gross 1953; Nowick and Berry 1972; Jackson et al. 2005; Kohlstedt 2007).

Anelastic Hysteresis

Seismic oscillations at distances beyond several fault lengths from an earthquake excite small strains less than 10^{-6} . These strains are recoverable during a cycle of seismic oscillation and lag the applied stress of the oscillation in time. Because of the time lag, a cycle of increasing and decreasing stress does not produce a perfectly proportional increase and decrease in strain. Instead a hysteresis loop occurs (Fig. 1). The area enclosed by the hysteresis loop is a measure of the energy lost due to heat and internal friction. During the stress cycle associated with the passage of a seismic wave, the



Seismic Viscoelastic Attenuation, Fig. 1 Stress-strain hysteresis curve showing the behavior of strain during a cycle of applied stress induced by a propagating seismic wave

energy lost to this internal friction is not available to deform the adjacent regions of the solid ahead of the wave front, and therefore the amplitude of the wave decreases.

From the hysteresis curve, one can see that the stress-strain relation cannot be described by a simple constant of proportionality in the time domain. A more complicated relation involving an integral over time is required to describe strain at any instant of time as a function of the prior time history of the applied stress. By Fourier transforming the rheologic equation, however, and keeping only terms describing the short-term anelastic behavior, the stress-strain relation can be simply expressed either by means of a complex elastic modulus $\hat{G}(\omega)$ or by its reciprocal, $\hat{J}(\omega)$ the complex elastic compliance:

$$\hat{\sigma}(\omega) = \hat{G}(\omega) \hat{\varepsilon}(\omega) \quad (2a)$$

$$\hat{\varepsilon}(\omega) = \hat{J}(\omega) \hat{\sigma}(\omega) \quad (2b)$$

The elastic modulus $\hat{G}$ and compliance $\hat{J}$ must be complex numbers to describe the phase lag of strain. They must also be frequency dependent because the phase lag of strain depends on the time history of stress, the shape of the hysteresis curve changing with different load histories. All the usual measures of anelasticity, including the frequency-dependent quality factor $Q(\omega)$ and the dispersion of the complex phase velocity $\hat{v}(\omega)$, can be expressed in terms of the complex modulus or complex compliance (Jackson and Anderson 1970). The trend of the frequency dependencies can be inferred from the time lag of strain from applied stress.

A feature of the complex modulus is that its real part will be smaller at zero or very low frequency and larger at infinite or very high frequency. That is, there will be an instantaneous response of strain to the applied stress, which is smaller than the eventual equilibrium response after longer time. The difference between the modulus at infinite frequency $G(\infty)$, representing the instantaneous or *unrelaxed* response, and the

low-frequency limit of the modulus $G(0)$, for the equilibrium or *relaxed* response, is called the *modulus defect* ΔG , with

$$\Delta G = G(\infty) - G(0) \quad (3)$$

The relaxed and unrelaxed moduli are pure real numbers that can be determined by observing a sequence of hysteresis curves for increasing frequencies of monochromatic loads. The frequency dependence of the real part of the modulus G at frequencies between 0 and ∞ implies that the propagation of a stress pulse will be dispersive, with higher frequencies traveling faster than lower frequencies.

Q and Complex Velocity

Since simple mechanical systems, composed of springs and dashpots and simple electric circuits, also obey linear equations of the form of Eqs. 2a and 2b, there are analogies between the quantities describing these systems and quantities in the stress-strain relation. For example, strain behaves like voltage, stress like current, and the complex compliance $\hat{J}(\omega)$ like the complex impedance of an electric circuit. Similar to the resonance phenomenon in circuits and mechanical systems, a Q can be defined by the average energy W per cycle divided by the energy lost or work done per cycle, ΔW :

$$Q = \frac{W}{\Delta W} \quad (4)$$

Large Q 's imply small energy loss; small Q 's imply large loss. Q is a measure of the area contained in the hysteresis loop of a stress-strain cycle. The inverse of Eq. 4, Q^{-1} , is sometimes simply termed the *attenuation or internal friction* (Knopoff 1964).

Plane waves of frequency ω and propagating in the + or - x direction can be defined by the phasor $\exp[i(\omega t \mp kx)]$ where $\hat{k}$ is a complex wave number $\frac{\omega}{\hat{c}}$ and $\hat{c}$ is a complex velocity defined from the local density ρ and complex modulus $\hat{G}$, with

$$\hat{c} = \sqrt{\frac{\hat{G}}{\rho}} \quad (5)$$

From the average energy density and loss per cycle of a complex plane wave, it can be shown that $Q = \frac{\text{Re}(\hat{G})}{\text{Im}(\hat{G})}$.

It is often less confusing to report the reciprocal parameter Q^{-1} , which represents the usually small perturbations to perfect elasticity.

The $Q^{-1}(\omega)$ Relaxation Spectrum

Since $\hat{G}$ depends on frequency, Q also depends on frequency. Zener (1960) described the frequency-dependent effects on an elastic modulus of a solid having a single characteristic time τ

for the relaxation of stress. A distribution of relaxation times can be constructed to give a Q^{-1} having a general dependence on frequency. The function $Q^{-1}(\omega)$ is called the *relaxation spectrum*. In the Earth and in many solid materials, the relaxation spectrum is observed to be slowly varying and nearly constant over a broadband of frequencies (Lekić et al. 2009). A theoretical requirement is that the attenuation Q^{-1} cannot increase faster than ω^1 or decrease faster than ω^{-1} . Figure 2 shows how a continuous distribution of relaxations can produce a Q^{-1} that is nearly constant with a frequency over a broadband. Once the limits of an absorption band are specified, however, it is not possible to have an arbitrarily high Q^{-1} (low viscoelastic Q) over an arbitrarily broad frequency band without making an unrealistically large modulus defect ΔG . Measured modulus defects in shear are typically less than 25%.

Velocity Dispersion

Although the dispersion in elastic moduli had long been known and predicted from the theories of viscoelasticity, it only began to be widely recognized in seismology when velocity models determined in the low-frequency band from the normal modes of the Earth (0.0001–0.01 Hz) were compared with velocity models determined in a high-frequency band (0.1–10 Hz) of body waves (Dziewonski and Anderson 1981). The models were found to differ, and the difference was found to agree with the amount of dispersion predicted from average Q models of the Earth. For example, since the Preliminary Reference Earth Model (PREM) was derived from observations of both the travel times of body waves and the eigenfrequencies of free oscillations, it reports velocities referenced at both 0.001 Hz and 1 Hz.

Another more subtle effect of this velocity dispersion can be seen in the propagation of pulses as body waves. A stress

disturbance that propagates from its point of initiation as a symmetric narrow Gaussian or triangle-shaped function in time gradually evolves into an asymmetric pulse (Fig. 3). High frequencies traveling faster than low frequencies are preferentially loaded into the front of the pulse (Futterman 1962; Carpenter 1967). Common theories for the physical mechanism of earthquakes as either frictional slip on a plane or a propagating crack triggered by tectonic stress often predict a far-field displacement pulse that has either a different or opposite form of asymmetry than that predicted for the effect of viscoelastic attenuation. These differences can assist in separating the effects of the source-time history from the effects of viscoelastic attenuation.

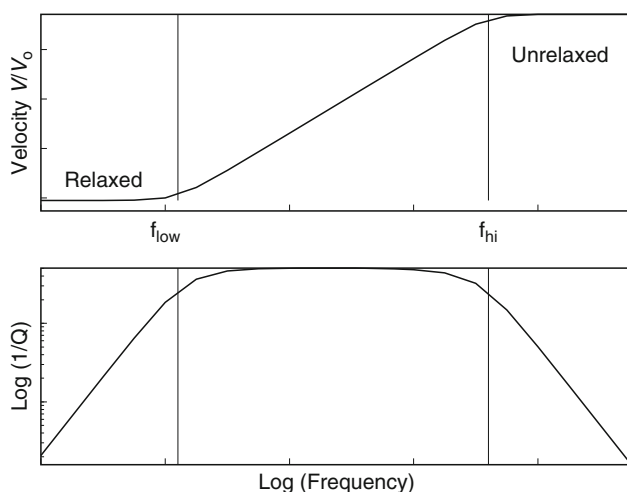
Effects of Scattering

Equivalent Medium

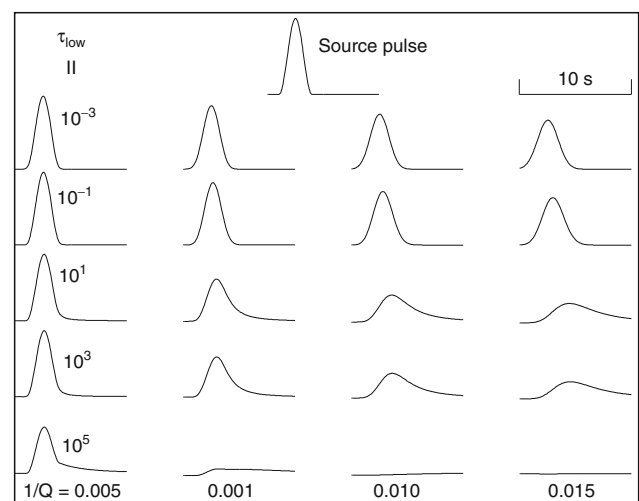
At frequencies that are so low that wavelengths are much larger than the characteristic scales of heterogeneity, the attenuative effects of scattering can usually be neglected. At sufficiently low frequency, little energy is lost to scattering, and the medium behaves like an equivalent medium, having properties that are an average of small-scale heterogeneities.

Stochastic Dispersion

The most complicated domain to calculate the effects of scattering is where the wavelength is of the order of the scale length of the heterogeneity (Fig. 4). In this domain, the presence of heterogeneities can profoundly alter the propagation of the wavefield, affecting both the initial cycle of a body wave pulse and the motion immediately following the initial cycle or *coda*. At distances less than 2000 km and frequencies



Seismic Viscoelastic Attenuation, Fig. 2 Viscoelastic dispersion of seismic velocity (*top*) and attenuation (*bottom*) showing a relaxation spectrum constant with frequency between two corner frequencies

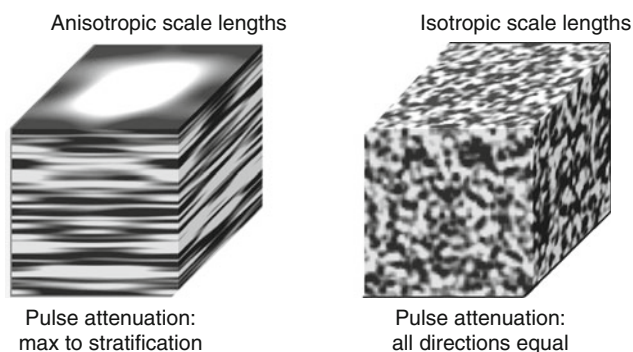


Seismic Viscoelastic Attenuation, Fig. 3 Pulse distortion showing the effects of viscoelastic dispersion for variable low-frequency corner and peak attenuation

higher than 1 Hz, scattered coda waves dominate the signature of S waves propagating in Earth's heterogeneous crust. Sato et al. (2012) review the theory and measurement of the apparent attenuation Q_C^{-1} of these coda waves. They describe strategies to separate the contribution of intrinsic from scattering attenuation to Q_C^{-1} based on the shape and frequency content of the energy envelopes of coda waves.

For longer range, some of the important effects of scattering can be illustrated by propagating body waves in a one-dimensional medium consisting of thin planar layers in which the velocity in each layer is assigned randomly (O'Doherty and Anstey 1971; Richards and Menke 1983). In fully three-dimensional structure, many of the effects of multiple scattering can be efficiently calculated by a radiative transport modeling of the energy of coda wave envelopes (Margerin 2013). These calculations predict body waves to exhibit a stochastic dispersion in which high-frequency energy is transferred into the coda. Neglect of stochastic dispersion can bias some estimates of intrinsic attenuation. Estimates of the spectral amplitude over a narrow time window can differ depending on the length of window analyzed, with less attenuation of higher frequencies estimated from longer time windows.

Pulse measurements such as width and rise time may also be biased because higher-frequency energy has been transferred out of the pulse into the later coda. This behavior is opposite to the effects of intrinsic attenuation on a propagating pulse, in which higher frequencies arrive at the beginning of the pulse. A symmetrically shaped displacement source pulse loses less of its symmetry as it propagates through the heterogeneous medium (Fig. 5). Anisotropy of the scale lengths of heterogeneity can also be important factor (Hong and Wu 2005), attenuation being strongest for paths for which the wavelength is of the order of the characteristic scale length in the medium in that direction.



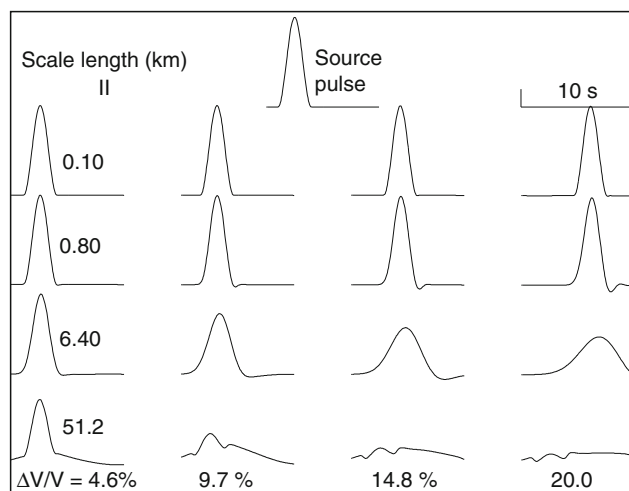
Seismic Viscoelastic Attenuation, Fig. 4 Example heterogeneity in the Earth and the directional dependence of attenuation of a body wave pulse for wavelengths that are approximately equal to either the dominant scale length (isotropic heterogeneity) or the dominant scale length in the direction of propagation (anisotropic heterogeneity)

Effects of Anisotropy

The existence of general anisotropy in the real part of the elastic modulus has the potential to bias some estimates of anelastic attenuation from either shear wave pulses or surface waves. In a medium having general anisotropy, the decompositions of shear wave motion into SH and SV motion will each contain the interference of two orthogonal shear wave polarizations that are neither SH nor SV (see ► [“Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting”](#)). The broadening of the SH component due to the interference of two quasi-S waves arriving close in time can be mistaken for the broadening due to anelastic attenuation. The regions of the deep Earth characterized by the strongest elastic anisotropy are the upper 400 km of the mantle (Silver 1996) and the lowermost 400 km of the mantle near the core-mantle boundary (Panning and Romanowicz, Panning and Romanowicz 2006). The effects of elastic anisotropy must be removed by combined analysis of SV and SH components of motion, resolving the polarizations of two quasi-S waves, before viscoelastic attenuation can be properly measured.

Measurement and Modeling Attenuation

Measurements of amplitude of seismic waves may be taken directly from seismograms or from their frequency spectra. To measure the attenuation, we must predict its effects from a model and vary the parameters of the model to fit the observed amplitude, amplitude ratio, or waveform. The effects of intrinsic attenuation in any modeling algorithm operating in



Seismic Viscoelastic Attenuation, Fig. 5 Pulse distortion showing the effects of scattering attenuation for variable scale lengths and velocity perturbation calculated by Cormier and Li (2002) using the Dynamic Composite Elastic Modulus (DYCEM) theory of Kaelin and Johnson (1998)

the frequency domain can be simply obtained by allowing elastic moduli and propagation velocities to become complex. Elastic boundary conditions, reflection and transmission at boundaries, travel times, and amplitudes are calculated exactly as in a non-attenuating solid but with elastic moduli and associated velocities analytically continued to complex values. This step of analytic continuation of real moduli to complex moduli is the same whether one wishes to predict the waveform of a body wave or surface wave or spectrum of free oscillations. The size of the imaginary part of the elastic moduli, parameterized by the value of Q^{-1} as a function of depth and frequency, is chosen to match an observed waveform, spectrum, amplitude ratio, or spectral ratio.

The Attenuation Operator for Body Waves

As an example of these procedures, consider an experiment with body waves. The effects on a body wave of source radiation, geometric spreading, reflection-transmission, and intrinsic attenuation are most conveniently expressed in the frequency domain by a product of complex functions. The complex $\hat{\Omega}(\vec{x}, \omega)$ spectrum of a body wave propagation from a point $\vec{x}_o$ to a receiver at $\vec{x}$ is

$$\hat{\Omega}(\vec{x}, \omega) = \bar{B}(\vec{x}_o, \vec{x}, \omega) \hat{S}(\omega) \hat{A}(\omega) \quad (6)$$

The function $\hat{S}(\omega)$ is the Fourier transform of the source-time function. $\bar{B}(\vec{x}_o, \vec{x}, \omega)$ incorporates a product of reflection-transmission coefficients, reverberations at source and receiver, geometric spreading, and source radiation pattern. In the frequency domain, an attenuation operator $\hat{A}(\omega)$ is defined by

$$\hat{A}(\omega) = \exp \left[i\omega \hat{T}(\omega) \right] \quad (7)$$

where $\hat{T}(\omega)$ is the complex travel time obtained by integrating the reciprocal of complex velocity along a ray or normal to the wave front of the body wave:

$$\hat{T}(\omega) = \int_{\text{ray}} \frac{ds}{\hat{c}(\omega)} \quad (8)$$

The effects of attenuation on reflection-transmission coefficients and geometric spreading through the complex velocity are much smaller and can be neglected unless the attenuation is very large (Q is very small). For $Q \gg 1$, $\hat{A}(\omega)$ can be rewritten as

$$\hat{A}(\omega) = \exp \left[\frac{-\omega t^*(\omega)}{2} \right] \exp \left\{ i\omega \left[\text{Re} \hat{T}(\infty) - \frac{H[t^*(\omega)]}{2} \right] \right\} \quad (9)$$

where

$$t^*(\omega) = \int_{\text{ray}} \frac{Q^{-1}}{\hat{c}(\omega)} ds \quad (10)$$

In Eq. 9, the attenuation effect is contained in the factor $\exp \left[\frac{-\omega t^*(\omega)}{2} \right]$, and the dispersive effect is in the factor $\exp \left\{ i\omega \left[\text{Re} \hat{T}(\infty) - \frac{H[t^*(\omega)]}{2} \right] \right\}$. The operator H is a Hilbert transform. In a band of frequencies in which Q and t^* are nearly constant

$$H[t^*(\omega)]/2 = \frac{\ln(\omega/\omega_0)}{\pi} t^* \quad (11)$$

where ω_0 is a reference frequency contained in the band (Liu et al. 1976). The value of $T(\infty)$ need not be known and can be replaced by some reference time or predicted from an Earth model for the phase being analyzed. The Hilbert transform relation in Eq. 11 for the dispersive phase of $\hat{A}(\omega)$ says that $\hat{A}(\omega)$ must be a minimum phase filter in the frequency domain. In general, the Fourier transform of the source-time function, $\hat{S}(\omega)$, is not a minimum phase filter, which can help in the separation and discrimination of the source spectrum from the effects of $\hat{A}(\omega)$ in the total expression for the far-field spectrum $\hat{\Omega}(\vec{x}, \omega)$.

The phase given by Eq. 11 will be accurate only between and far from the low- and high-frequency corners of the relaxation spectrum. Accurate representations of $\hat{A}(\omega)$ across a broad frequency band can be obtained for general relaxation spectra by substituting expressions for complex velocity $\hat{c}(\omega)$ in Eq. 8 obtained by superposing multiple Zener relaxations centered on single relaxation times whose strength is varied to achieve a desired shape for the relaxation spectrum. A useful expression for $\hat{c}(\omega)$ that is accurate for all frequencies across a relaxation spectrum, which is flat between two corner frequencies, can be derived from formulae for complex modulus given by Minster (1978) and is

$$\hat{c}(\omega) = c_{\text{ref}}(\omega_0) \frac{\sqrt{1 + 2\pi Q^{-1} \ln[\psi(\omega)]}}{\text{Re} \sqrt{1 + 2\pi Q^{-1} \ln[\psi(\omega_0)]}} \quad (12a)$$

where

$$\psi(\omega) = \frac{i\omega + 1/\tau_1}{i\omega + 1/\tau_2} \quad (12b)$$

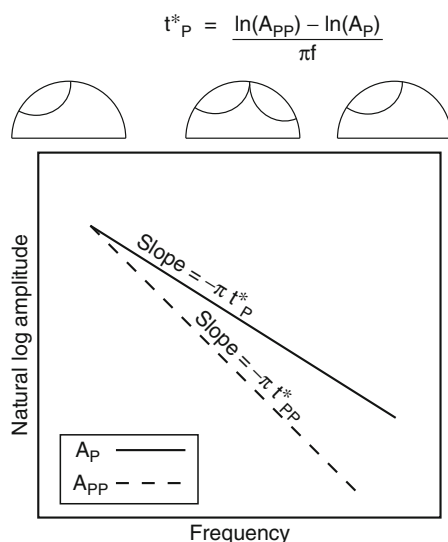
with τ_1 and τ_2 the relaxation times corresponding to the low- and high-frequency corners, respectively. $c_{\text{ref}}(\omega_0)$ is a real velocity at the reference frequency ω_0 .

Most measurements of attenuation attempt to measure only the amplitude effect of attenuation through the term $\exp[-\omega t^*(\omega)/2]$ from the spectral shape of body waves. There are basically two types of experiments commonly reported: (1) matching of spectral decay rates and (2) spectral ratios. In experiment (1), a shape for the displacement source spectrum $\hat{S}(\omega)$ is assumed usually to be a flat level followed by a decay of ω^{-2} above a corner frequency. The additional decay observed at high frequencies in data spectra is taken as a measure of t^* in $\exp[-\omega t^*(\omega)/2]$. In experiment (2), a ratio of two different seismic phases from the same source is observed in which the source spectrum is assumed to approximately cancel and the factor related to ratios of geometric spreading and near source and receiver crustal reverberations can be assumed to contribute a simple constant scalar factor. If the phases analyzed are recorded at the same receiver and are incident at nearly the same angles, then crustal reverberations at the source and receiver will approximately cancel. Both types of experiments usually apply some type of smoothing to the spectra to remove biasing effects of spectral holes caused by interfering crustal multiples, source complexities, scattering, and multipathing that are not included in the simple propagation model. Figure 6 illustrates an attenuation experiment of this type.

Since t^* measures only the path-integrated effect of attenuation, many such experiments for different ray paths, bottoming at a range of different depths, are needed to construct a model of Q as a function of depth. Serious

breakdowns in this approach, however, exist for cases in which the factorization of the observed spectrum into a product of a geometric spreading, source spectrum, and crustal effects is no longer accurate. One such case is when the body waves in question experience frequency-dependent effects of diffraction near caustics or grazing incidence to discontinuities. The spectral ratios of PKnKP waves, for example, are dominated by the effects of frequency-dependent reflection and transmission coefficients at grazing incidence to the core-mantle boundary. Instead of decreasing linearly with increasing frequency, an observed spectral ratio increases with frequency and exhibits a curvature in a log-log plot, which is consistent with a Q near infinity ($Q^{-1} = 0$) in the other core (Cormier and Richards 1976).

It is becoming more common to model and invert for viscoelastic attenuation parameters in the time domain, including not only the magnitude of the viscoelastic attenuation parameter Q^{-1} but also its frequency dependence. Examples of such a study are the inversions for Q^{-1} in the inner core assuming either a viscoelastic (Li and Cormier 2002) or a scattering origin of attenuation (Cormier and Li 2002). In these studies, the combined effects of mantle attenuation and source-time function were first modeled by fitting P waves observed in the great circle range 30–90°. Attenuation in the liquid outer core was assumed to be zero. Parameters defining a viscoelastic relaxation spectrum in the inner core were then varied to match the observed PKIKP waveforms. Care must be taken to examine a broad range of attenuation parameters because waveform inversions of this type are very nonlinear. Equally good fits between observed and predicted waveforms can sometimes be achieved with quite different domains of parameters.



Seismic Viscoelastic Attenuation, Fig. 6 The steps (top) to measure the path-integrated attenuation t^* of P waves in the mantle from a log-log plot (bottom) of stacked PP and P spectra (A_{PP} and A_P). The distances of observed P and PP spectra are chosen such that each turning ray path of PP is identical in shape and length to that of the single turning ray path of P in the mantle. (Adapted from figures in Warren and Shearer 2000)

Free Oscillations and Surface Waves

Measurements of attenuation in the low-frequency band of the free oscillations of the Earth are conducted in the frequency domain by observing the width of the individual resonance peaks associated with each mode. These measurements face special problems associated with the broadening produced by lateral heterogeneity of elastic Earth structure. This heterogeneity splits the degenerate modes of a radially symmetric Earth, making a set of modes that would have the same frequency have slightly different frequencies. The slightly different frequencies of the split modes may not be easily resolved in the data spectra and can be confused with the broadening of a single resonance peak of a mode caused by attenuation.

Lateral heterogeneity also complicates the measurement of viscoelastic attenuation of surface waves. Heterogeneity introduces focusing, defocusing, and multipathing, all of which must be accurately modeled to understand the separate attenuative effects of viscoelasticity.

The frequency band of free oscillation and surface waves (0.001–0.1 Hz), however, offers the best hope of obtaining radially symmetric whole-Earth models of viscoelastic attenuation in this frequency band. This is because lateral variations in attenuation structure are averaged by the gravest modes of oscillation and surface waves that make multiple circuits around the Earth. Computational advances have made the division between free oscillation and surface wave studies fuzzier, with common approaches now amounting to time-domain modeling of complete low-frequency (<0.1 Hz) seismograms for combined three-dimensional models of viscoelasticity and heterogeneity.

Numerical Modeling

Fully numerical modeling of the seismic wavefield allows the combined effects of heterogeneity and viscoelasticity in three dimensions to be predicted. If the numerical technique is formulated in the frequency domain, substituting a complex velocity for an assumed relaxation spectrum can incorporate viscoelastic attenuation.

If the technique is formulated in the time domain by a finite difference approach, it is neither simple nor efficient to incorporate attenuation by convolution of the wavefield calculated in a non-attenuating medium with an attenuation operator $A(t)$, where $A(t)$ is the inverse Fourier transform of $\hat{A}(\omega)$ defined in Eq. 9. Instead, time-domain *memory functions* can be defined to describe a viscoelastic relaxation (Robertsson et al. 1994; Blanch et al. 1995) that can be integrated over time simultaneously with the equations describing particle velocity or displacement and stress. In practice, only three-memory functions, distributed evenly over the logarithm of their characteristic times, are required to simulate a broad frequency band in which Q^{-1} varies slowly.

Interpretation of Attenuation Measurements in the Earth

Shear Versus Bulk Attenuation

In the most general theory of viscoelasticity, it is possible to have energy loss occur during both a cycle of volumetric strain and shear strain. Since the velocity of a P wave depends on both the bulk and shear moduli, the attenuation Q_P^{-1} of a P wave can be written as a linear combination of the attenuations Q_P^{-1} and Q_S^{-1} defined from complex shear and bulk moduli:

$$Q_P^{-1} = L Q_S^{-1} + (1 - L) Q_K^{-1} \quad (13)$$

where $L = (4/3) (V_S/V_P)^2$ and V_P and V_S are the compressional and shear velocities, respectively (Anderson 1989). Although

plausible mechanisms for defects in bulk moduli have been found in both laboratory measurements and analytic models of specific attenuation mechanisms, measurements on real data find that bulk dissipation in the Earth is small and, in most cases, can be neglected. One exception may occur when the pressure and temperature state in a narrow depth regions of the Earth are close to those near a phase transition, either solid-liquid (Stevenson 1983) or solid-solid (Ricard and Chambat 2009). Except for these regions, intrinsic attenuation occurs almost entirely in shear, associated with lateral movement of lattice defects, grain boundaries, and/or fluids rather than with changes in material volume. Hence, assuming $Q_K^{-1} = 0$ and a Poisson solid, for which $V_P = \sqrt{3} V_S$,

$$Q_P^{-1} = \frac{4}{9} Q_S^{-1} \quad (14)$$

and the parameter for path-integrated attenuation of S waves t_S^* approximately equals $4t_P^*$. Most experiments confirm these values. Jackson (2007) reviews several situations in which measureable bulk attenuation exists. In the majority of these cases, Q_K^{-1} is no more than several percent of Q_S^{-1} . Morozov (2015), however, has demonstrated that assuming zero bulk attenuation leads to an auxetic medium, in which Poisson's ratio becomes negative. He proposes a starting assumption for bulk attenuation in which $Q_K^{-1} = 0.4 Q_S^{-1}$. This instead predicts $Q_P^{-1} = \frac{2}{3} Q_S^{-1}$ and $t_S^* = 2.6 t_P^*$. In support are observations at frequencies higher than 1 Hz that typically find $t_S^* < 4 t_P^*$. These observations may alternatively be an effect of scattering rather than bulk attenuation. With scattering, the apparent Q_P^{-1} can approach the apparent Q_S^{-1} , especially when it is estimated from pulse widths or spectra taken in a frequency band in which wavelengths are on the order of richest heterogeneity scale lengths of the medium. In this case, the assumption of viscoelastic attenuation occurring mainly in shear can aid in separating the effects of scattering from intrinsic attenuation.

Frequency Dependence

When the results of attenuation measurements determined from free oscillations and body waves in the 0.0001–0.1 Hz band began to be compared with observations of body wave spectra in the 1–10 Hz band, it became apparent that even under the assumption of a white source spectrum, an increase in Q with frequency was necessary to explain the amplitude of spectra in the 1–10 Hz band.

Thermal Activation

Frequency dependence of viscoelastic attenuation has been interpreted in terms of physical mechanisms of attenuation that are thermally activated. In these mechanisms, the low-frequency corner of an absorption band f_L is tied to a

relaxation time τ_L , where $f_L = 1/(2\pi\tau_L)$. The time τ_L depends on temperature T and pressure P as follows:

$$\tau_L = \tau_0 \exp\left(\frac{E^* + PV^*}{RT}\right) \quad (15)$$

where E^* and V^* are the activation energy and volume, respectively. Both the low- and high-frequency corners (f_L , f_H) of an absorption band are assumed to be similarly affected, temperature and pressure acting to slide the absorption band through a band of frequencies. A typical width to expect for the relaxation spectrum of the mantle is about five orders of magnitude in frequency, $\tau_L/\tau_H = 10^5$ (Minster and Anderson 1981; Anderson and Given 1982). A simplified model of an absorption band with depth in the Earth's mantle is shown in Fig. 7. The movement of the absorption band toward lower frequencies (longer periods) in the mantle below 400 km depth is consistent with the type of behavior shown in Fig. 8 for the t^*_S measured from shear waves of an earthquake. The difference in the location of the absorption band with respect to the band of seismic frequencies is consistent with models of the temperature and pressure profiles of the Earth's mantle for specific values of E^* and V^* . Relaxation times are also affected by the grain size of minerals, which may increase from millimeter to centimeter in the upper 400 km of the mantle (Faul and Jackson 2005). A rapid increase in temperature with depth can rapidly change the location of the absorption band with respect to the seismic band. Given a specific temperature profile, estimated values of the activation energy and volume, and grain size, a combined shear velocity and Q_S profile can be predicted and modified to fit an observed shear velocity profile.

Seismic Viscoelastic

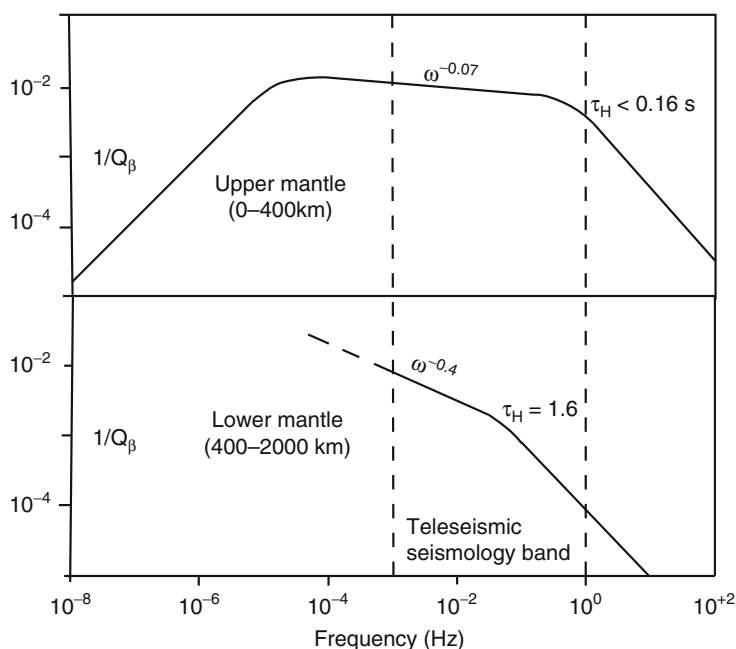
Attenuation, Fig. 7

A frequency- and depth-dependent model of shear attenuation in the Earth's mantle derived from modeling broadband shear waves. This model has been used by the National Earthquake Information Center (NEIC) to correct for viscoelastic attenuation in the reported radiated elastic energy from earthquakes (Boatwright and Choy 1986; Choy and Boatwright 1995)

Regional Variations

Romanowicz and Mitchell (2015) review and interpret both global and many regional variations in intrinsic attenuation, including correlations with velocity perturbations. Tomographic images of perturbations to seismic velocities and attenuations in the mantle can qualitatively be interpreted as images of lateral temperature variations, leaving open the possibility of additional contributions to the observed heterogeneity from chemical variations. In the upper mantle, tectonically active regions overlain by radiogenically younger crust are more attenuating than the mantle underlying inactive regions such as continental shields (Fig. 9). The shape of the frequency dependence across the seismic band seems to remain similar in different regions, although the Q at a given frequency is lower for a tectonically young region than for an older shield region. Although motion of interstitial water and partial melt can produce high attenuation, grain boundary and defect mechanisms in a dry mantle have been shown to be equally effective in explaining regions of high attenuation in the upper mantle, except possibly behind island arcs and directly beneath active spreading ridges (Karato and Jung 1998; Faul and Jackson 2005).

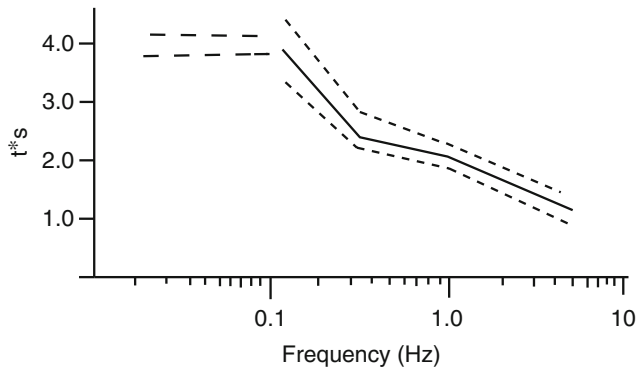
Generally, perturbations in attenuation Q^{-1} inversely correlate with those in shear wave velocity (Roth et al. 2000). The correlations between shear velocity and shear attenuation appear to be consistent with thermal activation, in which the dispersive effect of attenuation acts jointly with variations in the high-frequency corner of the mantle relaxation spectrum to produce the observed variations in travel time and frequency content. Deep chemical differences between the upper mantle beneath shields and that beneath young continents and oceans and in the deep mantle, however, have



been suggested by comparing anomalies in shear velocity versus bulk velocity V_K , where $V_K = \sqrt{V_P^2 - (4/3)V_S^2}$. Milder lateral temperature differences in the mid- and lower mantle tend to make the relaxation spectrum more laterally stable in height, width, and location within a frequency band, reducing the observed lateral heterogeneity in velocity and attenuation in these regions.

Global models of attenuation of the type shown in Fig. 9 often do not have the resolving power to detect spatially

concentrated regions of high-attenuation and sharp spatial gradients found in regional studies beneath and near island arcs, mid-ocean spreading ridges, and hot mantle plumes. Long-established lateral variations, such as high-attenuation west of the Rocky Mountains in North America and low-attenuation east (Der et al. 1982), are not very apparent. Denser path coverage from either body or surface waves, however, is becoming increasingly available from sensor networks such as the USArray, making it possible to resolve smaller spatial scales (Fig. 10).



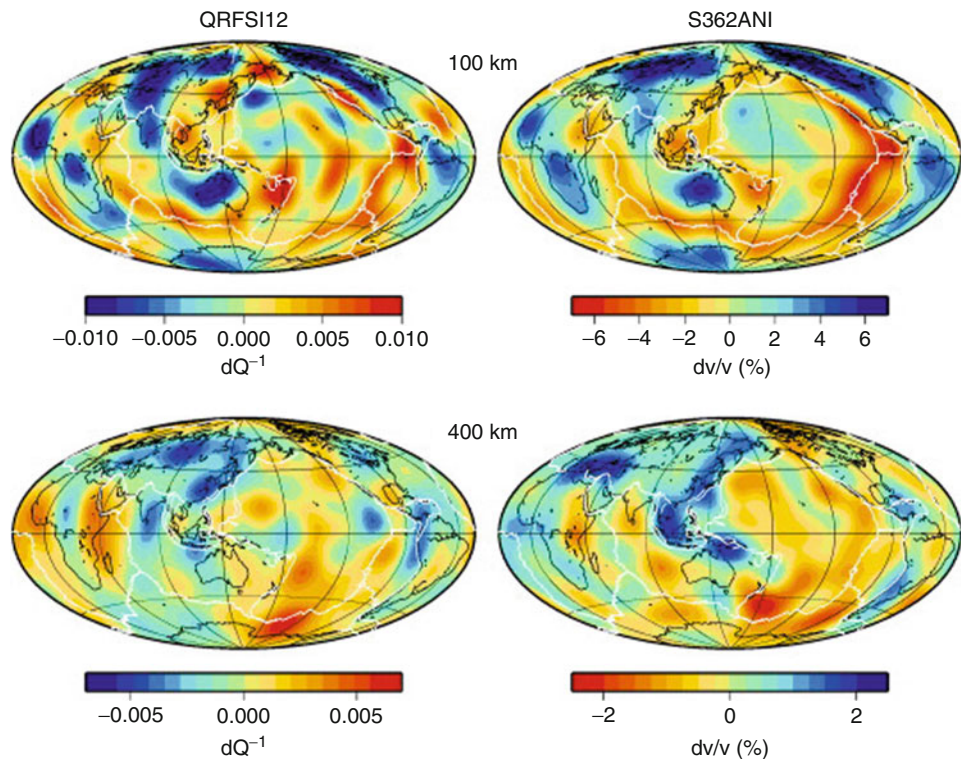
Seismic Viscoelastic Attenuation, Fig. 8 Path-integrated attenuation t^* of S waves in the mantle as a function of frequency determined from modeling broadband shear waves predicted from the frequency- and depth-dependent attenuation model shown in Fig. 7 (Choy and Cormier 1986)

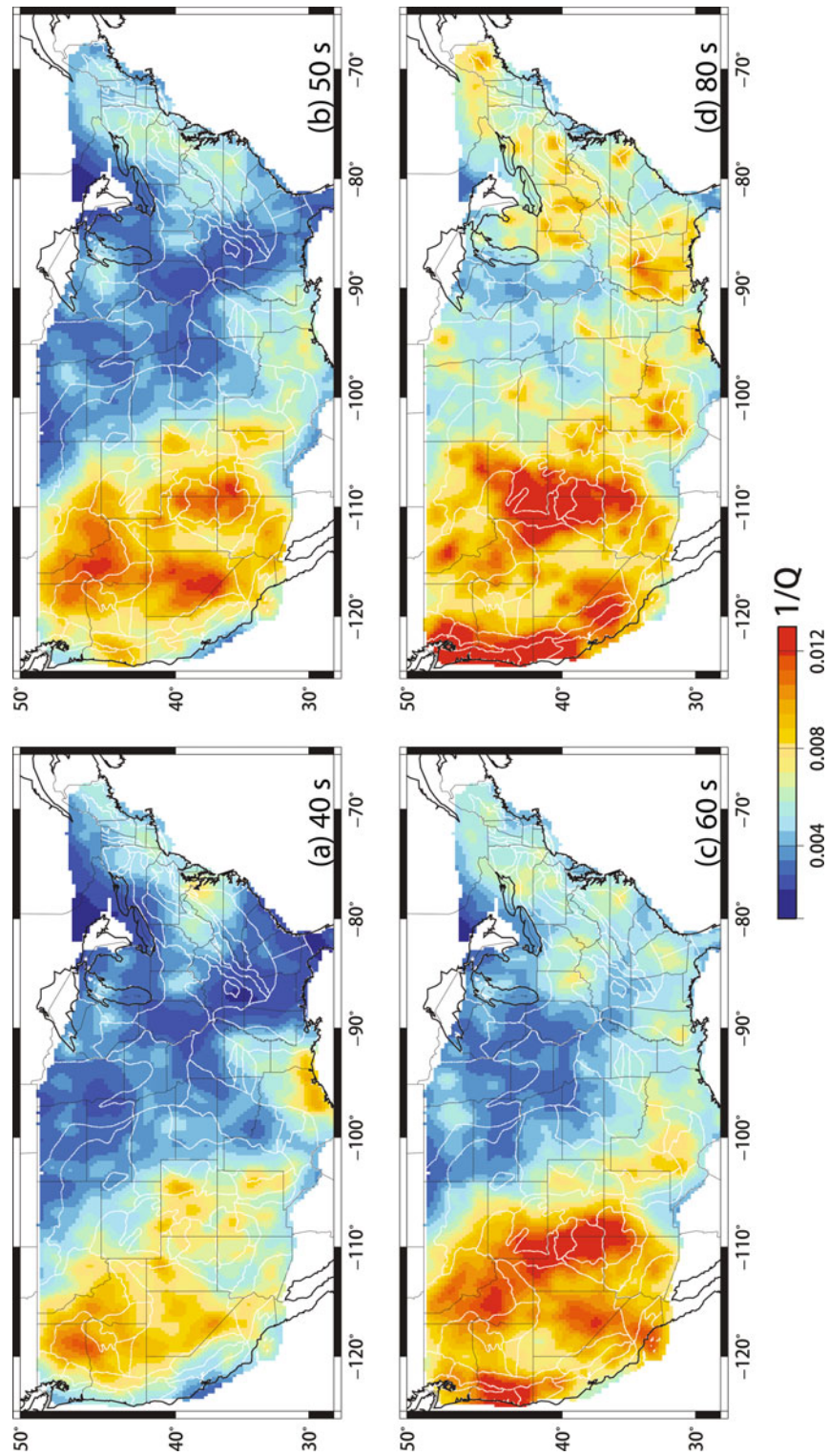
Strain Dependence

Laboratory measurements of Q in rocks find dependence in strain beginning at strains of about 10^{-6} . The strain dependence decreases with increasing confining pressure. The Q is also strongly dependent on moisture and interstitial fluids between cracks in rocks and grain boundaries and soils. These observations are consistent with a physical mechanism of frictional sliding across cracks. Unlike viscoelastic relaxations, which are representative of all linear mechanisms, frictional sliding is an inherently nonlinear mechanism, depending on strain amplitude.

Estimates of the onset of nonlinearity may be made by calculating the strain associated with the seismic wave being analyzed. A rough estimate can be obtained by assuming that the wave front is a plane wave and dividing the particle velocity by the propagation velocity. For example, the particle

Seismic Viscoelastic Attenuation, Fig. 9 Inverted upper mantle shear velocity perturbations and shear attenuation. (From Dalton et al. 2009)





Seismic Viscoelastic Attenuation, Fig. 10 Regionally averaged Rayleigh wave attenuation at four periods. (From Bao et al. 2016). Longer periods sample deeper into the upper mantle

velocities of body waves observed in strong ground motion recordings from 0 km to 10 km from the hypocenter of a magnitude 6 earthquake are typically 0.01 m/s. If the body wave propagates at 3 km/s, the strain observed at the strong ground motion site is roughly 0.01 m/s divided by 3×10^3 m/s or $\text{strain} = 3.3 \times 10^{-5}$. This value is likely to be in the nonlinear regime of surficial rocks having open cracks or pores. In this strain regime, it becomes important to solve the elastic equation of motion with nonlinear terms in its rheology. Bataille and Calisto (2008) describe a perturbative approach to nonlinear rheology, demonstrating that strain dependence behaves in much the same way as elastic scattering does in linear elasticity, transferring energy to the later coda of a body wave. Delépine et al. (2009) describe a finite element approach to solve the problem of a more general 3-D, nonlinear, viscoelastic medium.

Summary

The intrinsic attenuation of seismic waves in the Earth has been found to be consistent with loss mechanisms that are thermally activated. The observed regional and frequency dependences of seismic Q agree with the expected lateral variations in a geotherm having a rapid temperature increase in the upper 400 km of the mantle, followed by a slower vertical and lateral variation in the mid- and lower mantle. High velocities correlate with regions of low attenuation; low seismic velocities correlate with regions of high attenuation. Measurements are consistent with losses primarily in shear rather than bulk deformations.

The existence of lateral heterogeneity in the elastic properties of the Earth complicates the measurement of viscoelastic properties. The longer scale lengths of heterogeneity can split modes of free oscillation and focus and defocus body waves and surface waves. Shorter scale lengths scatter seismic energy, broaden the waveforms of body waves, and redistribute energy into different time and angular windows. Observations that are useful for discriminating between the effects of scattering attenuation versus viscoelastic attenuation include the ratio of apparent P wave attenuation to apparent S wave attenuation, the rate of velocity dispersion within a frequency band, and the apparent viscoelastic modulus defect. The intensity of heterogeneity in percent fluctuation of velocities and densities is higher at shorter scale lengths at shallower depths in the Earth's crust and upper mantle. There is still a need for experiments that determine finer details of how the distribution of heterogeneity changes with depth and lateral location in the Earth and its anisotropy of scale lengths. Many, if not most, experiments have not completely removed the effects of heterogeneity on the apparent attenuation, making their results an upper bound on the viscoelastic Q^{-1} .

Laboratory experiments find a transition from linear to nonlinear rheology at strains of the order of 10^{-6} . The observed strain dependence of Q and its dependence on pressure in the shallow crust agree with a mechanism of frictional sliding of cracks. It is still unknown how and at what strain levels linear superposition begins to break down close to a seismic source.

Although a consensus has been reached on the major features and thermal activation of intrinsic attenuation in most of the Earth's upper mantle, this is less true of other deep regions of the Earth. Definitive experiments are still needed for the determination of Q^{-1} in the lowermost 400 km of the mantle, where increased lateral heterogeneity exists across a broad spatial spectrum, complicating the separation of its effects from those of viscoelasticity. In the inner core, there is still a need to understand the linkage between observed lateral variations in velocity, elastic anisotropy, and scattering (Calvet and Margerin 2008) as well as an apparent anisotropy of attenuation (Makinen et al. 2014).

Cross-References

- [Body Waves](#)
- [Earth's Structure, Core](#)
- [Earth's Structure, Lower Mantle](#)
- [Earth's Structure, Upper Mantle](#)
- [Earthquakes, Early and Strong Motion Warning](#)
- [Earthquakes, Strong-Ground Motion](#)
- [Energy Partitioning of Seismic Waves](#)
- [Free Oscillations of the Earth](#)
- [Lithosphere, Mechanical Properties](#)
- [Magnetic Anisotropy](#)
- [Mantle D'' Layer](#)
- [Mantle Viscosity](#)
- [Numerical Methods, Finite Difference](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Diffraction](#)
- [Seismic Monitoring of Nuclear Explosions](#)
- [Seismic Phase Nomenclature: The IASPEI Standard](#)
- [Seismic Properties of Rocks](#)
- [Seismic Tomography](#)
- [Seismic Velocity and Temperature Relationships](#)
- [Seismic Wave Propagation in Real Media: Numerical Modeling Approaches](#)
- [Seismic Waves, Scattering](#)
- [Seismic Ray Theory](#)
- [Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting](#)
- [Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat](#)

Bibliography

- Anderson DL (1989) *Theory of the Earth*. Blackwell Scientific Publications, Boston
- Anderson DL, Given JW (1982) The absorption band Q model for the Earth. *J Geophys Res* 87:3893–3904
- Bao X, Dalton CA, Jin G, Gaherty JB, Shen Y (2016) Imaging Rayleigh wave attenuation with USArray. *Geophys J Int* 206:241–259
- Bataille K, Calisto I (2008) Seismic coda due to non-linear elasticity. *Geophys J Int* 172:572–580
- Blanch JO, Robertsson JOA, Symes WW (1995) Optimally efficient constant Q modeling. *Geophysics* 60:176–184
- Boatwright J, Choy G (1986) Teleseismic estimates of the energy radiated by shallow earthquakes. *J Geophys Res* 91:2095–2112
- Calvet M, Margerin L (2008) Constraints on grain size and stable iron phases in the uppermost inner core from multiple scattering modeling of seismic velocity and attenuation. *Earth Planet Sci Lett* 267:200–212
- Carpenter EW (1967) Teleseismic signal calculated for underground, underwater, and atmospheric explosions. *Geophysics* 32:17–32
- Choy GL, Boatwright JL (1995) Global patterns of radiated seismic energy and apparent stress. *J Geophys Res* 100:18205–18228
- Choy GL, Cormier VF (1986) Direct measurement of the mantle attenuation operator from broadband P and S waves. *J Geophys Res* 91:7326–7342
- Cormier VF, Li X (2002) Frequency dependent attenuation in the inner core: Part II. A scattering and fabric interpretation. *J Geophys Res* 107(B12). <https://doi.org/10.1029/2002JB1796>
- Cormier VF, Richards PG (1976) Comments on “The damping of core waves” by Anthony Qamar and Alfredo Eisenberg. *J Geophys Res* 81:3066–3068
- Dalton CA, Ekstrom G, Dziewonski AM (2009) Global seismological shear velocity and attenuation: a comparison with experimental observations. *Earth Planet Sci Lett* 284:65–75
- Delépine N, Lenti L, Bonnet G, Semblat J-F (2009) Nonlinear viscoelastic wave propagation: an extension of Nearly Constant Attenuation models. *J Eng Mech (ASCE)* 135(11):1305–1314
- Der ZA, McElfresh TW, O'Dannell A (1982) An investigation of regional variations and frequency dependence of anelastic attenuation in the United States in the 0.5–4 Hz band. *Geophys J R Astron Soc* 69:67–100
- Dziewonski AM, Anderson DL (1981) Preliminary reference Earth model. *Phys Earth Planet Inter* 24:297–356
- Faul UH, Jackson I (2005) The seismological signature of temperature and grain size variations in the upper mantle. *Earth Planet Sci Lett* 234:119–134
- Futterman WI (1962) Dispersive body waves. *J Geophys Res* 67:5279–5291
- Gross B (1953) *Mathematical structure of the theories of viscoelasticity*. Hermann, Paris
- Hong T-K, Wu R-S (2005) Scattering of elastic waves in geometrically anisotropic random media and its implication to sounding of heterogeneity in the Earth's deep interior. *Geophys J Int* 163:324–338
- Jackson I (1993) Progress in the experimental study of seismic attenuation. *Annu Rev Earth Planet Sci* 21:375–406
- Jackson I (2007) Properties of rocks and minerals – physical origin of anelasticity and attenuation in rocks. In: Schubert G (ed) *Treatise on geophysics*, vol 2. Elsevier, Amsterdam, pp 493–525
- Jackson DD, Anderson DL (1970) Physical mechanisms of seismic wave attenuation. *Rev Geophys Space Phys* 8:1–63
- Jackson I, Webb S, Weston L, Boness D (2005) Frequency dependence of elastic wave speeds at high temperature: a direct experimental demonstration. *Phys Earth Planet Inter* 148:85–96
- Kaelin B, Johnson LR (1998) Dynamic composite elastic medium theory. Part II. Three-dimensional media. *J Appl Phys* 84:5458–5468
- Karato S-I, Jung H (1998) Water partial melting and the origin of the seismic low velocity zone in the upper mantle. *Earth Planet Sci Lett* 157:193–207
- Knopoff L (1964) *Q*. *Rev Geophys* 2(4):625–660
- Kohlstedt DL (2007) Properties of rocks and minerals – constitutive equations, rheological behavior, and viscosity of rocks. In: Schubert G (ed) *Treatise on geophysics*, vol 2. Elsevier, Amsterdam, pp 390–417
- Lekić V, Matas J, Panning MP, Romanowicz B (2009) Measurement and implications of frequency dependence of attenuation. *Earth Planet Sci Lett* 282:295–303
- Li X, Cormier VF (2002) Frequency dependent attenuation in the inner core: Part I. A viscoelastic interpretation. *J Geophys Res* 107(B12). <https://doi.org/10.1029/2002JB001795>
- Liu H-P, Anderson DL, Kanamori H (1976) Velocity dispersion due to anelasticity: implications for seismology and mantle composition. *Geophys J R Astron Soc* 47:41–58
- Makinen A, Deuss A, Redfern SAT (2014) Anisotropy of Earth's inner core intrinsic attenuation from seismic normal mode models. *Earth Planet Sci Lett* 404:354–364
- Margerin L (2013) Introduction to radiative transfer of seismic waves. In: Levander A, Nolet G (eds) *Seismic Earth: array analysis of broadband seismograms*. <https://doi.org/10.1029/157GM14>
- Minster JB (1978) Transient and impulse responses of a one-dimensional linearly attenuating medium—I. Analytical results. *Geophys J R Astron Soc* 52:479–501
- Minster B, Anderson DL (1981) A model of dislocation-controlled rheology for the mantle. *Philos Trans R Soc Lond* 299:319–356
- Morozov IG (2015) On the relation between bulk and shear seismic dissipation. *Bull Seismol Soc Am* 105:3180–3188
- Nowick AS, Berry BS (1972) *Anelastic relaxation in crystalline solids*. Academic, New York, p 677
- O'Connell RJ, Budiansky B (1977) Viscoelastic properties of fluid-saturated cracked solids. *J Geophys Res* 82:5719–5735
- O'Doherty RF, Anstey NA (1971) Reflections on amplitudes. *Geophys Prospect* 19:430–458
- Panning MP, Romanowicz BA (2006) A three dimensional radially anisotropic model of shear velocity in the whole mantle. *Geophys J Int* 167:361–379
- Ricard Y, Chambat F (2009) Seismic attenuation in a phase change coexistence loop. *Phys Earth Planet Inter* 176:124–131
- Richards PG, Menke W (1983) The apparent attenuation of a scattering medium. *Bull Seismol Soc Am* 73:1005–1021
- Robertsson JOA, Blanch JO, Symes WW (1994) Viscoelastic finite-difference modeling. *Geophysics* 59:1444–1456
- Romanowicz B, Mitchell B (2015) Deep earth structure: Q of the earth from crust to core. In: Schubert G (ed) *Treatise on Geophysics*, vol 1. Elsevier, Amsterdam, pp 789–827
- Roth EG, Wiens DA, Zhao D (2000) An empirical relationship between seismic attenuation and velocity anomalies in the upper mantle. *Geophys Res Lett* 27:601–604
- Sato H, Fehler MC, Maeda T (2012) *Seismic wave propagation and scattering in the heterogeneous Earth*, 2nd edn. Springer, New York
- Shearer PM, Earle PS (2008) Observing and modeling elastic scattering in the deep Earth. *Advances in geophysics*, vol 50: Earth heterogeneity and scattering effects on seismic waves. 50 (Sato H, Fehler MC, Eds.), Elsevier Academic, San Diego, pp 167–193. [https://doi.org/10.1016/s0065-2687\(08\)00006-x](https://doi.org/10.1016/s0065-2687(08)00006-x)
- Silver PG (1996) Seismic anisotropy beneath the continents: probing the depths of geology. *Annu Rev Earth Planet Sci* 24:385–432
- Stevenson DJ (1983) Anomalous bulk viscosity of two-phase fluids and implications for planetary interiors. *J Geophys Res* 88:2445–2455
- Warren LM, Shearer PM (2000) Investigating the frequency dependence of mantle Q by stacking P and PP spectra. *J Geophys Res* 105(B11):25391–25402
- Zener C (1960) *Elasticity and anelasticity of metals*. The University of Chicago Press, Chicago

Seismic Wave Propagation in Real Media: Numerical Modeling Approaches

Ursula Iturrarán-Viveros¹ and Francisco J. Sánchez-Sesma²

¹Facultad de Ciencias, Departamento de Matemáticas, Universidad Nacional Autónoma de México, Ciudad Universitaria, México City, Mexico

²Instituto de Ingeniería, Universidad Nacional Autónoma de México, Ciudad Universitaria, México City, Mexico

Synonyms

Elastic waves in homogeneous and inhomogeneous media

Definition

The Earth is a complex medium containing heterogeneities over many scales. Its mechanical properties have great variability as well. The influence of heterogeneities on seismic wave propagation is therefore studied intensively by numerical modeling to explain observations. Modeling of seismic wave propagation is an essential ingredient in many aspects of exploration seismology, ranging from survey design methods to imaging and inversion algorithms. The increase in available computational power enabled the development of powerful numerical techniques for the accurate calculation of synthetic seismograms in three-dimensional (3D) models of the Earth. This chapter presents a glimpse of a variety of these numerical methods.

Introduction

In 1969, Keiiti Aki first focused attention on the appearance of continuous wave trains (called coda), in the tail portion of individual seismograms of local earthquakes as a direct evidence of the random heterogeneity of the lithosphere. Models for seismic wave propagation through inhomogeneous elastic media have been developed using deterministic approaches such as model theory for layered media. In 1954, Norman A. Haskell made a great contribution to theoretical geophysics with his famous paper in which he showed how seismic surface waves could be computed for an Earth modeled by an arbitrary number of plane parallel layers, each one with arbitrarily prescribed physical properties. His work was based on Thomson's work, and it is now known as the Thomson-Haskell method. In his pioneering work, Claerbout (1968) recognized the solution to the inverse problem of determining the medium from the seismogram. Numerical modeling of

seismic wave propagation in an inhomogeneous media is becoming more important in seismic exploration. A wide variety of recent numerical strategies implies treating Earth models with successively increasing complexity. The realistic computational models can include anisotropic media, non-planar interfaces between layers and blocks, velocity/density/quality factor gradients inside layers, and often with free surface topography. In particular, the rheology of the medium should allow for realistic attenuation. We can divide the different approaches to numerically solve the wave equation in complex media into three main classes: *direct methods*, *integral equation methods*, and *ray tracing methods* (see Table 1 and Carcione et al. (2002)). In order to solve the wave equation using *direct methods*, the geological model needs to be approximated by a numerical mesh. There are no restrictions about material variability and different rheologies can be implemented. These methods give solution to the wave equation in the time domain. They can be very accurate provided a sufficiently fine grid, but they are computationally more expensive than analytical or ray tracing methods. *Integral equation methods* are inspired in Huygens' principle. These methods formulate the solution of a problem in terms of values at the domain's boundary. Integral equation methods have an interesting conceptual advantage over direct methods, which is the reduction of one space dimension. Moreover, integral equation methods (in the frequency domain) do not need absorbing boundaries. They match easily the boundary conditions and do not suffer from grid dispersion. However, these methods require the knowledge of Green's function which is limited to a few cases (see Kausel 2006). *Ray tracing* or *asymptotic methods* are approximate because they do not take into account the complete wavefield. The wavefield is considered as an ensemble of certain events of relatively high frequency, each arriving at certain traveltimes and having certain amplitude.

Direct Methods

A convenient way to classify numerical methods to solve wave equation is to use a variational description. If the equation of motion and the boundary conditions are prescribed in their differential form and the solution approach deals directly with them, it corresponds to the *strong formulation*. Such are the cases of finite differences (FE) and pseudospectral (PS) methods. On the other hand, the methods that seek to reduce the errors use an integral form of the equation of motion weighting the error by means of trial functions, corresponding to the *weak formulation*. They are also known as variational methods. Such is the case of finite element (FE) method. The free or natural boundary conditions are automatically satisfied in the weak formulation.

Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Table 1 Classification of numerical methods to model seismic wave propagation in real media

Direct methods	Finite differences (FD), pseudospectral (PS), finite element (FE), spectral element method (SEM), finite volume (FV), discontinuous Galerkin (DG), discrete element method (DEM)
Integral equation methods	Boundary element method (BEM), indirect boundary element method (IBEM), fast multipole method (FMM), domain integral methods (DIM)
Asymptotic or ray tracing methods	Gaussian beams

Finite Difference Method

The basic idea of finite difference methods is to compute the partial derivatives by an approximation based on Taylor series expansions of functions near the point of interest. Forward and backward time derivatives lead to implicit or explicit schemes. In general, an implicit method is computationally more expensive than an explicit method, due to the requirement of solving large matrix equations. When applying finite differences in the implicit scheme only for 1D problems, the matrix to be inverted is tridiagonal, which is very convenient in terms of computational cost. However this is not the case for 2D problems. Nevertheless, using the method of dimensional splitting, one can replace the multidimensional problem by a succession of much simpler one-dimensional problems, and as a result it is possible to build tridiagonal matrices and combine their results to compute the desired time derivative. This ingenious approach is described in Bayliss (1986), Mufti (1985), and Emerman et al. (1982). The Taylor series expansion method is usually used to calculate finite difference coefficients, and the calculation error increases with the frequency of data. To further suppress numerical dispersion, numerous methods have been proposed based on the modification of finite difference coefficients. For example, Liu and Sen (2009) have obtained implicit finite difference formulae derived from fractional expansion of derivatives which form a system of linear equations. Obviously, the explicit schemes are computationally simpler. Later, Liu (2013, 2014) optimized the finite difference coefficients by least squares method in a wide frequency range, which completed the finite difference forward modeling results. Chu and Stoffa (2012) proposed a finite difference method, implicit both in space and in time, for the scalar wave equation. The authors used a three-level implicit splitting time integration method for the temporal derivative and implicit finite difference operators of arbitrary order for the spatial derivatives. Both the implicit splitting time integration method and the implicit spatial finite difference operators require solving systems of linear equations.

Therefore a vast majority of earthquake ground motion modeling and exploration seismology studies use explicit FD schemes. Virieux in a couple of papers (1984) and (1986) for SH and P-SV cases respectively was the first to introduced explicit time domain staggered FD in seismology. This method is easy to implement, to introduce different kinds of sources, and to consider viscoelasticity. In addition, the local nature of finite difference operators makes the method suitable for parallelization. In seismic applications, the following velocity-stress formulation is widely used:

$$\begin{aligned}\frac{\partial \sigma_{ij}}{\partial x_j} + f_i &= \rho \frac{\partial v_i}{\partial t} \\ \frac{\partial \sigma_{ij}}{\partial t} &= c_{ijkl} \frac{\partial e_{kl}}{\partial t}\end{aligned}\quad (1)$$

where the first time derivative of the displacement u is the velocity v , σ_{ij} is the stress, $\partial e_{kl} = (v_{k,l} + v_{l,k})/2$, being e_{kl} the strain and c_{ijkl} is the fourth-order stiffness tensor with elastic coefficients for linear elastic solids, f_i is the i th component of the body force vector $\mathbf{f}$ that can be written in terms of the moment tensor as follows:

$$\mathbf{f} = -\mathbf{M} \cdot \nabla \delta(\mathbf{x} - \mathbf{x}_s) S(t) \quad (2)$$

where $\mathbf{x}_s$ denotes the source position, δ is the Dirac delta, and $S(t)$ denotes the source time function. Evaluating the second time derivative of Eq. (1), i.e., $\ddot{u} = \partial_t v$ (double dot denotes twice time derivative), at times $(n+1)\Delta t$ and $(n-1)\Delta t$ by a Taylor expansion yields

$$\frac{\partial^2 u^n}{\partial t^2} = \frac{1}{(\Delta t)^2} \left[u^{n+1} - 2u^n + u^{n-1} - 2 \sum_{l=2}^L \frac{(\Delta t)^{2l}}{(2l)!} \frac{\partial^{2l} u^n}{\partial t^{2l}} \right] \quad (3)$$

When applied (3) to numerically approximate spatial and time derivatives in the 1D scalar wave equation

$$\frac{1}{\beta^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \quad (4)$$

we obtain the Lax-Wendroff scheme:

$$\frac{1}{\beta^2} \frac{u_i^{n+1} - 2u_i^n + u_i^{n-1}}{(\Delta t)^2} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{(\Delta x)^2}. \quad (5)$$

This equation represents the discretization of (4) in the x - t space. Now, consider that we have a Cartesian system of coordinates (x, y, z) with a fourth-order computational domain taking into account the time, i.e., (x, y, z, t) . A convenient spatial-time grid is given by a set of discrete points (x_i, y_j, z_k, t_n) expressed as $x_i = x_0 + i\Delta x$, $y_j = y_0 + j\Delta y$,

$z_k = z_0 + k\Delta z$, and $t_n = t_0 + n\Delta t$; $i, j, k, n = 1, 2, \dots$. The spatial increments Δx , Δy , and Δz are usually referred to as grid spacings, and (x_0, y_0, z_0, t_0) are initial values for these variables, while Δt is the time step. The value of a function u at a grid position (x_i, y_j, z_k, t_n) is given by $u_{i,j,k}^n = u(x_i, y_j, z_k, t_n)$. It is possible to have nonuniform grids, and the other important consideration is whether all functions are approximated at the same grid positions. The conventional, partly staggered, and staggered grids are illustrated in Fig. 1. An important observation is pointed out for finite difference solutions of hyperbolic equations. Accuracy increases with increasing Δt , for central differences, up to a limit defined by the Courant-Friederickson-Lewy stability condition (assuming $\Delta x = \Delta z$, in 2D):

$$\frac{V_{\max} \Delta t}{\Delta x} \leq \frac{1}{\sqrt{2}} \quad (6)$$

where $V_{\max}$ is the maximum velocity considered in the model.

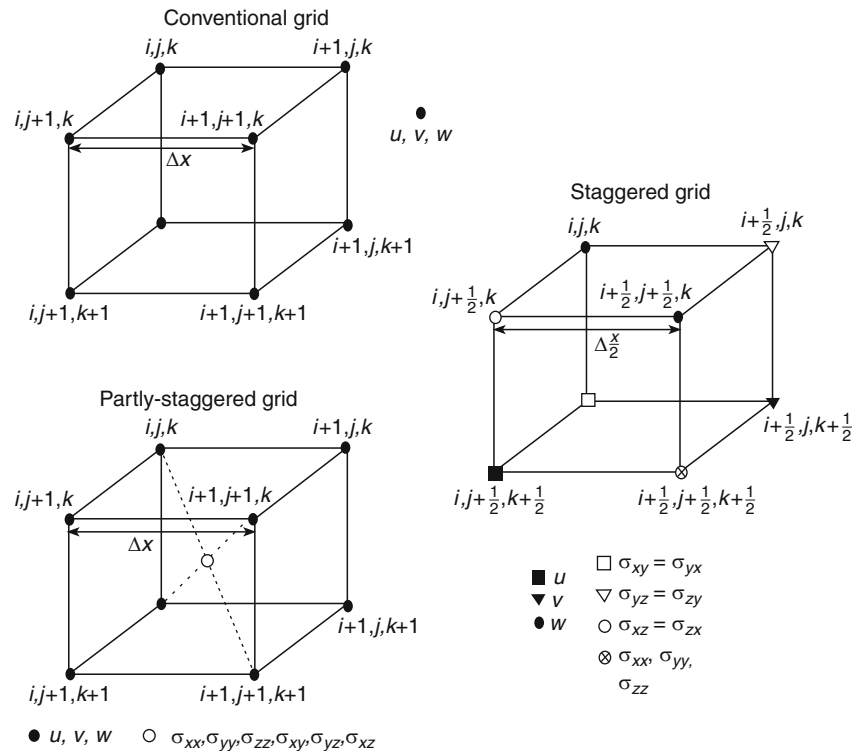
The choice among the different schemes is made to improve accuracy, reduce memory requirements, and deal with anisotropy media. Recently, the use of the partly staggered grid was upgraded by Saenger et al. (2000) and Saenger and Bohlen (2004). The authors called the grid rotated staggered grid since they obtained the spatial FD operator by the rotation of the standard staggered grid operator. They used this scheme to successfully simulate media with heterogeneities (cracks, pores, free surface, and anisotropy). Rotated

staggered grids have been also used to model planar and non-planar dynamic rupture models (see Cruz-Atienza et al. (2007)). The main limitations of FD are the extensive consumption of computational resources in terms of both core memory and CPU time. However, Furumura and Chen (2005) performed large-scale parallel FD simulations of seismic wave propagation to model strong ground motions during damaging earthquakes. Parallel FD simulation allowed Saito and Furumura (2009) to successfully simulate tsunami generation. A complete review of FD methods can be found in Moczo et al. (2007). Maeda et al. (2017) developed an open-source software package, called Open-source Seismic Wave Propagation Code (OpenSWPC), for parallel numerical simulations of seismic wave propagation in 3D and 2D (P-SV and SH) viscoelastic media based on the finite difference method in local-to-regional scales. The authors include a frequency-independent attenuation model based on the generalized Zener body and an efficient perfectly matched layer for absorbing boundary condition.

Irregular interfaces are difficult to deal with finite differences; however some successful efforts have been made to address this issue. One example is given in Rao and Wang (2018) where the authors propose a summation-by-parts finite difference method that can be used for models with non-rectangular meshes across fluctuating interfaces, and it can achieve desirable simulation accuracy. Almuhaideb and Tokzö (2015) developed a 2D finite difference solver for the elastic wave equation that combines a fourth-order

Seismic Wave Propagation in Real Media: Numerical Modeling Approaches,

Fig. 1 Spatial grid cells in the conventional, partly staggered, and staggered grids (see Moczo et al. 2007)



ADER scheme (Arbitrary high-order accuracy using DERivatives). Liu et al. (2018) modeled seismic wave propagation in fractured media using an integral method with tetrahedral grid cells. This approach is derived from finite element and finite difference methods in heterogeneous media. It is flexible in modeling irregular interfaces and surface topography. Sun et al. (2017) simulated seismic wave propagation across a 2D topographic fluid and solid interface at the sea bottom by the finite difference method. In this method, seismic waves in seawater are governed by acoustic wave equations, whereas seismic waves in solid earth are governed by elastic wave equations. Li et al. (2019) proposed a seismic forward modeling method for fluid-solid coupled media with a transition layer. The second-order acoustic wave pressure equation and the second-order elastic wave displacement equation in heterogeneous media are used in the fluid and solid media, respectively.

Anisotropy

Anisotropy is caused by structural alignment of elements that are smaller than the wavelength employed for seismic data acquisition. Examples of structures that cause anisotropic effects are fine layering, tectonic grain, orientation and intensity of cracking. One way to include these effects in the model is to consider that elastic properties at a given point might vary with direction. Therefore by including anisotropy in the model, alignments of small-scale heterogeneities are captured. The alignment effects might lead to important

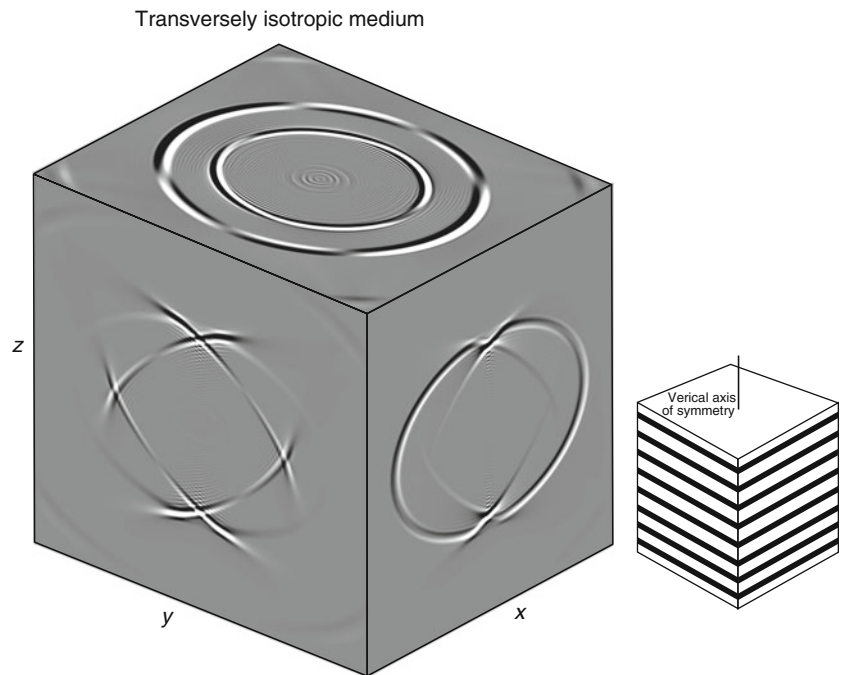
petrophysical information. To consider anisotropy in a FD simulation, one needs to define c_{ijkl} the fourth-order stiffness tensor in Eq. (1). For example, to model the effect of layering (see Fig. 2) one can consider the following stiffness tensor

$$C = \begin{bmatrix} c_{11} = 66.6 & c_{12} = 19.7 & c_{13} = 39.4 & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{13} & 0 & 0 & 0 \\ c_{13} & c_{13} & c_{33} = 39.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} = 10.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{66} \end{bmatrix} \quad (7)$$

where $c_{66} = (c_{11} - c_{12})/2$, the density given by $\rho = 2590 \text{ Kg/m}^3$, and the units for the elastic constant are in GPa. The example represents the Mesaverde clay shale (see Thomsen 1986 and Carcione et al. 1992). This kind of stiffness tensor with these five elastic constants characterizes a medium of class IV according to the classification given by Payton (1983), and it is called transversely isotropic medium (since it has the vertical z -axis as a symmetrical axis). Notice that due to the anisotropy, in Fig. 2, the different plane cuts xy -plane, xz -plane, and yz -plane the wave fronts do not propagate at the same speed. In Gao and Huang (2017), an improved rotated staggered grid method with fourth-order temporal accuracy is applied to simulate elastic wave propagation in anisotropic media. Inspired by the idea of rotated staggered grid finite difference method, Zou and Cheng (2017) proposed a modified pseudospectral method for wave propagation in arbitrary 3D anisotropic media.

Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Fig. 2

(Left) Illustration of a transversely isotropic medium. The example represents the Mesaverde clay shale. The six elastic constants (for the stiffness tensor in Eq. (1) that generate this medium are given by $c_{11} = 66.6 \text{ GPa}$, $c_{12} = 19.7 \text{ GPa}$, $c_{13} = 39.4 \text{ GPa}$, $c_{33} = 39.9 \text{ GPa}$, $c_{44} = 10.9 \text{ GPa}$, and $c_{66} = (c_{11} - c_{12})/2$ and the density $\rho = 2590 \text{ Kg/m}^3$. (Right) A layered medium that represents the anisotropic effect



Pseudospectral Method

The pseudospectral method (PS) is based mainly on Fourier and Chebyshev differential operators. The spatial derivative is computed in the frequency-wavenumber domain as follows:

$$\frac{\partial u(x, t)}{\partial x} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} u(\xi, t) e^{-ik\xi} d\xi \right) ike^{ikx} dk, \quad (8)$$

(see Kreiss and Oliger (1972), Kosloff and Baysal (1982), Furumura et al. (1998)). In Eq. (8) the interior integral corresponds to the spatial Fourier transform of $u(x, t)$. The product by ik within the integrand of the exterior integral makes the x -derivative when performing the pass to space domain. The PS with Fourier has the disadvantage of periodic properties, called wraparound. Both Fourier and Chebyshev are accurate up to the maximum wavenumber of the mesh that corresponds to a spatial wavelength of two grid points (at maximum grid spacing for the Chebyshev operator). Like FD methods, they are also unable to model surface waves with the same accuracy as body waves because of the one-way treatment that needs to be performed in order to implement the free surface condition (e.g., Carcione 1994).

Finite Element and Spectral Element Methods

The finite element (FE) method is more efficient than FD or PS when dealing with complex geometries (irregular interfaces) and heterogeneous media and handling boundary conditions. This method is also applicable with inelastic constitutive models. Bielak et al. (2005) have developed an efficient FE-based computational platform for large-scale ground motions. Low approximation orders may lead to a large numerical dispersion (see Marfurt (1984)). As a result, mesh refinement is required to reduce numerical dispersion but may lead to a large numerical cost even if parallelization is possible. High-order FE methods became more efficient when the spectral element method (SEM) appeared using Chebyshev polynomials as interpolating functions (see Seriani et al. (1992), Priolo et al. (1994), Komatitsch and Villotte (1998), Komatitsch and Tromp (1999)). This method introduced the use of the Lagrange polynomials as interpolating functions which produces a diagonal mass matrix when using the Gauss-Lobatto-Legendre quadrature.

Discontinuous Galerkin Method

The Discontinuous Galerkin method (DG) is a technique that uses discontinuous basis functions to formulate a Galerkin approximation. Given a mesh of the analysis domain, the DG method approximates the solution within each element by a

function from a low-dimensional vector space of functions, e.g., as a linear combination of basic functions like polynomials. For a pair of adjacent mesh elements, the approximate solution computed in the interior of the elements does not have to agree on their common boundary. Some of the advantages of this method are:

- It can sharply capture solution discontinuities relative to a computational mesh.
- It simplifies adaptation since inter-element continuity is neither required for mesh refinement and coarsening.
- It conserves the appropriate physical quantities (e.g., mass, momentum, and energy) on an element-by-element basis.
- It can handle problems in complex geometries to high order.
- Regardless of order, it has a simple communication pattern to elements sharing a common face that simplifies parallel computation. With a discontinuous basis, the DG method produces more unknowns for a given order of accuracy than traditional finite element or finite volume methods, which may lead to some inefficiency. The DG method is harder when applied to unstructured meshes; in particular, it is harder to formulate limiting strategies to reduce spurious oscillations when high-order methods are used. Nevertheless, Käser and Dumbster (2006) have successfully implemented the method for elastic wave propagation on unstructured meshes.

For an introduction to the key ideas, basic analysis, and efficient implementation of discontinuous Galerkin methods and a review of the state of the art, we refer the reader to the book by Hesthaven and Warburton (2008). Fernandez et al. (2018) shows a hybridizable and embedded discontinuous Galerkin method for wave propagation problems in fluids, solids, and electromagnetism. In Shukla et al. (2019) a nodal discontinuous Galerkin method with a Lax-Friedrichs flux is used to model the wave propagation in transversely isotropic and poroelastic media. Furthermore, Shukla et al. (2019b) applied a high-order weight-adjusted discontinuous Galerkin scheme for the numerical solution of three-dimensional (3D) wave propagation problems in anisotropic porous media including for the first time a sub-cell level to deal with cases where heterogeneities are present.

Integral Equation Methods

Boundary Element and Indirect Boundary Element Methods (BEM and IBEM).

Among Integral equation methods it is the boundary element method. The formulation of direct boundary integral equation method (BIE) in elastodynamics is back to the pioneering work of Somigliana (1885). Considering a volume

of elastic material V bounded by a surface S , the displacement field u_i at a point $\mathbf{x}$ and time t can be expressed as a function of the values of the displacement and traction t_j along the boundary through Somigliana's representation theorem:

$$cu_i(\xi, t) = \int_0^t d\tau \int_S \left[t_j(\xi, \tau) G_{ij}(\xi, t, \mathbf{x}, \tau) - u_j(\xi, \tau) \sum_{jik}(\xi, t, \mathbf{x}, \tau) n_k(\xi) \right] dS(\xi) \quad (9)$$

where $G_{ji}(\xi, t, \mathbf{x}, \tau)$ and $\sum_{jik}(\xi, t, \mathbf{x}, \tau)$ are the responses in terms of displacement and stress, time t , and point ξ of an infinite homogeneous medium to a unit force impulse, applied at time τ at point $\mathbf{x}$ in the direction i and $n_k(\xi)$ is the normal boundary pointing outside V . The volumetric body sources are assumed to be null, but if present, their contribution can be easily added as follows:

$$cu_i(\xi, t) = \int_0^t d\tau \int_S \left[t_j(\xi, \tau) G_{ij}(\xi, t, \mathbf{x}, \tau) - u_j(\xi, \tau) \sum_{jik}(\xi, t, \mathbf{x}, \tau) n_k(\xi) \right] dS(\xi) + \int_0^t d\tau \int_V G_{ij}(\xi, t, \mathbf{x}, \tau) f_i(\mathbf{x}, \tau) dV \quad (10)$$

where

$$c = \begin{cases} 1, & \xi \in V \\ 1/2, & \xi \in S \\ 0, & \mathbf{x} \notin V \end{cases} \quad (11)$$

assuming that S has smooth boundaries. The values of c come from the volume integration of the product $\delta(\mathbf{x} - \xi, t - \tau) \mathbf{u}_j(\xi, \tau)$, and the value accounts for the integration of the Dirac's delta and its position. Green's function for the homogeneous, isotropic, elastic media can be found in Aki and Richards (1980).

The indirect formulation of the elastodynamic problem expresses the wavefield as an integral over the boundary of elementary source radiations as follows:

$$u_i(\mathbf{x}, t) = \int_0^t d\tau \int_S \phi_j(\xi, \tau) G_{ij}(\mathbf{x}, t, \xi, \tau) dS(\xi) \quad (12)$$

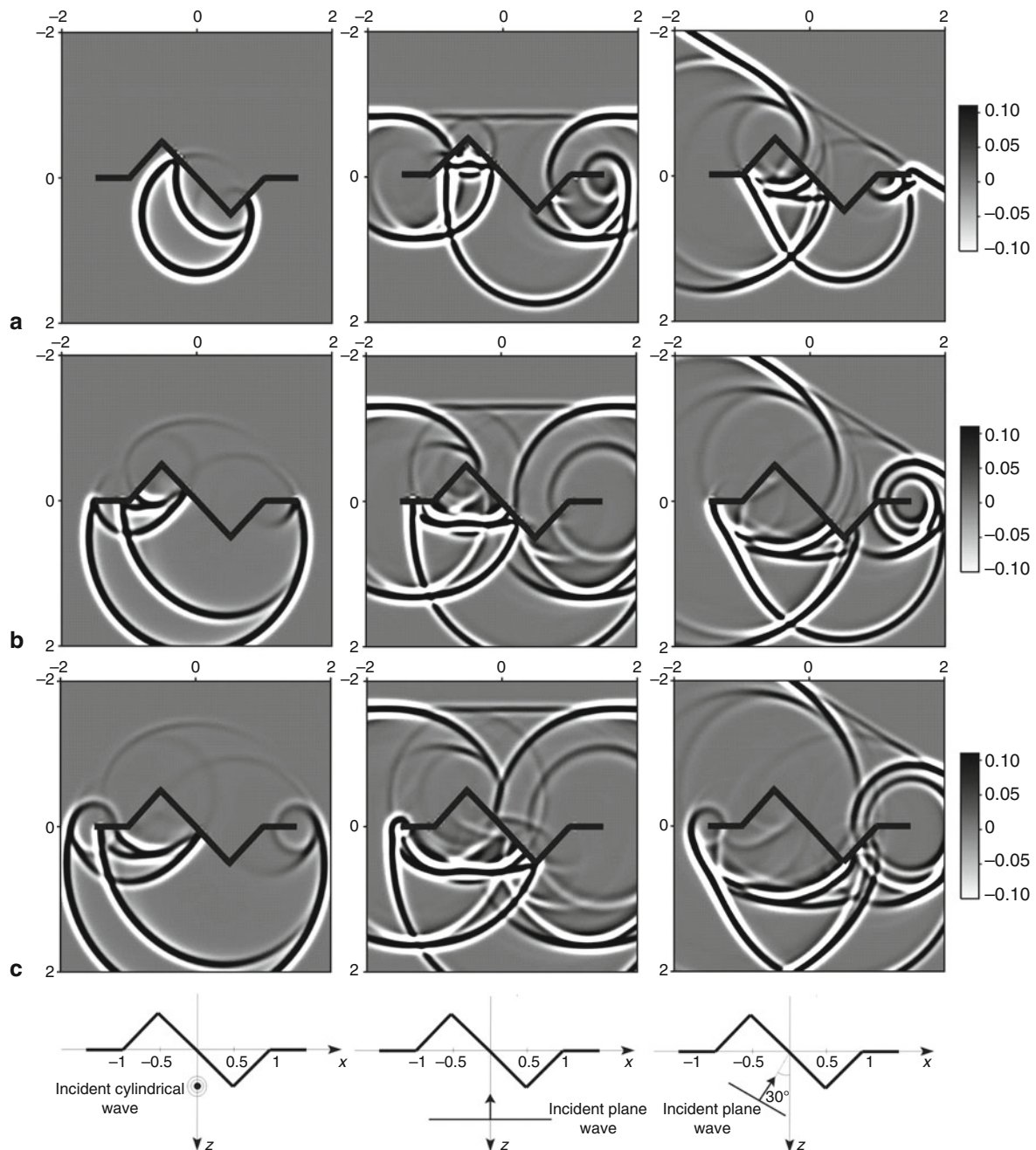
where $\phi_j(\xi, \tau)$ denotes τ the force density distribution applied at time at point ξ of the surface S . In this equation we have assumed that the volumic forces are null. If present, their contribution needs to be added. Equation (10) can be seen as the mathematical transcription of Huygens' principle, which states that each point of diffracting boundary acts as a secondary source of wave radiation. This principle applies to

any wave front as well. Equation (12) is also the mathematical expression of the exploding reflector concept widely used in seismic exploration (see Claerbout 1985). The fictitious distribution ϕ_j is an intermediate variable which needs to be solved in the boundary integral equations before the wavefield can be computed. This is the reason why this formulation is called indirect.

The discretization of the boundaries leads to the discretization of the BIE, and the application of boundary conditions transforms BIE into a system of linear equations which in general is not symmetric. The resolution of this system is usually done implicitly and in the frequency domain. Comprehensive introduction of integral equation methods can be found in the books by Bonnet (1999), Dominguez (1993), and Manolis and Beskos (1988). In seismology, the BEM is used to study the effect of irregular topography on earthquake ground motion, both in 2D and 3D problems. This method has been also used to simulate elastic wave propagation in media with cracks; for example, in Fig. 3 we have an irregular crack strike by a cylindrical SH incident wave (first column) and the same crack strike by a plane SH incident wave with 0° and 30° , measured with respect to the z -axis (second and third columns, respectively) (see Iturrarán-Viveros et al. 2005; Pointer et al. 1998). A complete review on the use of BEM for seismic problems can be found in Bouchon and Sánchez-Sesma (2007). In Liu et al. (2016) the IBEM has also been applied to model elastic wave propagation in a 3D fluid-saturated poroelastic half-space. Perton et al. (2016) shows a new implementation of the IBEM that allows to simulate elastic wave propagation in complex configurations made of embedded regions that are homogeneous with irregular boundaries or with flat layers.

Fast Multipole Method in Elastodynamics

Two types of fast multipole method (FMM) are available for elastodynamics in the frequency domain. The first one is the low-frequency FMM. As for static cases, the complexity of this algorithm is $O(N)$. This complexity is due to the fact that the wavelength is much larger than the domain size. On the other hand, if the wavelength is shorter than the geometrical feature, the complexity of low-frequency FMM increases to $O(N^2)$, and the method is not efficient any longer. For this reason, computational efficiency of fast BEMs in the mid-frequency regime is enhanced by using the so-called diagonal form for the Helmholtz Green's function, proposed by Rokhlin (1990), with a complexity of $O(N \log N)$. The upper limit stems from the fact that the size N becomes intractable at high frequencies, but the diagonal form also breaks down at very low frequencies and must be replaced with other types of expansions. The method for low frequencies was developed by Fujiwara (1998). The first 3D implementation was proposed by Fujiwara (2000) using a multilevel and diagonal form. The author presents some low-frequency seismic-oriented examples. Chaillat et al.



Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Fig. 3 Snapshots of displacements for (left) a cylindrical SH wave striking an irregular crack, (middle) a plane SH wave with

incident angle 0° striking an irregular crack, and (right) a plane SH wave with incident angle 30° striking an irregular crack. The set of snapshots are taken at (a) $t = 1.25$ s (b) $t = 1.71$ s, and (c) $t = 2.03$ s

(2009) also addressed problems in 3D. Yoshida (2001) proposed a low-frequency FMM for crack problems in 3D.

Analytic Solutions for Nonhomogeneous Media

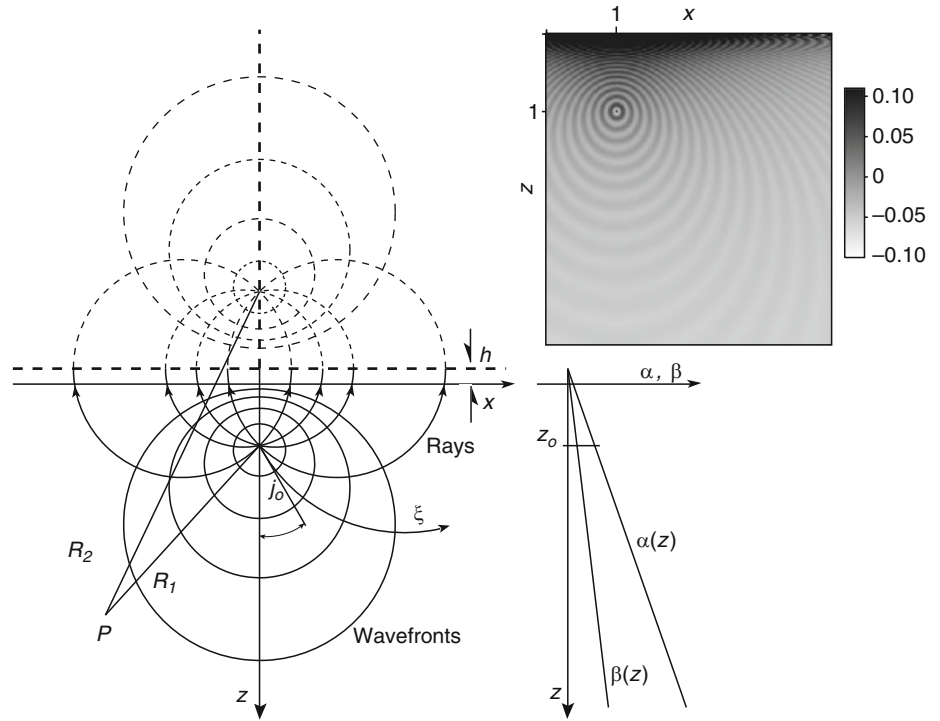
Since analytic solutions help to assess numerical methods, some attempts to develop analytic solutions for particular cases of mildly anisotropy media are briefly described. In

Sánchez-Sesma et al. (2001), authors developed approximate analytical formulae for elastic 2D Green's functions for a constant-gradient propagation velocity medium. These solutions correspond to unit line forces per unit length: the anti-plane SH line source and the in-plane P-SV line sources, respectively. They are based on the asymptotic ray theory and account for both near-source effects and low frequencies. The orthogonal bipolar coordinates are depicted in Fig. 4.

The approximate solution for the 2D scalar case is given by

Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Fig. 4

(Left) Schematic illustration of the bipolar coordinate system. Both rays and wave fronts are circular. Rays depend upon takeoff angle j_0 . $t = \xi t_0$, where $t_0 = h/\beta(0)$ and ξ is one of the bipolar coordinates (the other is j_0). The distances R_1 and R_2 from the poles to point P are displayed. The plane for which $z = -h$ corresponds to null propagation velocities. (Top right) This solution is used to compute Green's function for the scalar case in the time domain with a source located at (1,1), $\beta(0) = 500$ m/s and $\beta(z_{\max}) = 3000$ m/s. Wave fronts depend on traveltime



$$G_{22}(\omega, \tau) \approx \Lambda \frac{1}{4\mu_0} H_0^{(1)}(\omega\tau) \quad (13)$$

where Λ is defined as follows

$$\Lambda = \left(\frac{1 + \gamma z_0}{1 + \gamma z} \right)^{\frac{1+\eta}{2}} \left[2 \ln \left(\frac{R_2 + R_1}{R_2 - R_1} \right) \frac{(z_0 + h)(z + h)}{R_1 R_2} \right]^{1/2} \quad (14)$$

where z_0 = depth of the source, β_0 = shear wave velocity at the source level, $\beta(0)$ = shear wave velocity at $z = 0$, ρ_0 = mass density at source level, $\gamma = 1/h$, h = distance from depth $z = 0$ to the level where the propagation velocity is null ($z = -h$), $j_0 = \sin^{-1}[2x(z_0 + h)/R_1 R_2]$ = take off angle, $R_1 = (x^2 + (z - z_0)^2)^{1/2}$, and $R_2 = (x^2 + (z + z_0 + 2h)^2)^{1/2}$. The travel time t is given by

$$\tau = t_0 \ln \left(\frac{R_2 + R_1}{R_2 - R_1} \right) = 2t_0 \tanh^{-1} \left(\frac{R_1}{R_2} \right) = t_0 \xi \quad (15)$$

In Watanabe and Payton (2006) and (2002), Green's functions are obtained exactly for radially symmetric waves and for transient and time-harmonic SH waves in inhomogeneous anisotropic elastic solids, respectively. Since time-harmonic conditions have been assumed, viscoelastic material behavior can be captured through the introduction of complex-valued wave speeds.

Equivalent Media Theories

When one wishes to model highly heterogeneous media, one possibility is to make the continuum hypothesis (see Christensen 1979; Hudson 1991). A real material with its atomic and molecular structure is replaced by an equivalent continuum model, which remains a continuum no matter how much it is subdivided. The macroscopic mechanical properties of the material remain unchanged. This hypothesis works as long as the scale of observation is much larger than the scale length of the molecular structure (Rayleigh scattering regime). Microfractures, pores, and other heterogeneities with uniform statistical distributions can be replaced by an equivalent or effective medium provided that the scale of the observation is much larger than the scale of the heterogeneities. This is an attractive approach in exploration seismology since the scale of the observation, or seismic wavelengths, is much larger than the scale length of the heterogeneities present in the subsurface. The concept of equivalent media clearly implies that if the stress, strain, or displacements are measured on a large enough scale, the values obtained and the relationship between them will be that of homogeneous continuum. This implies some kind of spatial averaging process. Nevertheless, scales much smaller than the minimum wavelength (used in the numerical model) are present in the earth model. Therefore effective media theories which allow to average small scales of the original medium without losing

the accuracy of the wavefield computation are needed. In Capdeville and Marigo (2007), authors applied two-scale homogenization technique to the wave equation in layered media. The order 0 homogenization provides the same result as the one given by Backus (1962). The order 0 is not enough to obtain an accurate solution especially for surface waves. Higher-order homogenization terms (up to 2) allowed them to obtain accurate surface waves. Lately Capdeville and Marigo (2010) proposed an original nonperiodic homogenization technique in 1D. The extensions to 2D and 3D are foreseen.

Asymptotic Ray Tracing Methods

An accurate estimation of travel time is needed to map heterogeneities in the Earth's subsurface. Traveltime computation schemes fall within one of these two categories: first-arrival traveltime and ray theory-based traveltime. The first-arrival traveltime corresponds to the first arrival of the complete wavefield at a specified receiver position, and it is not a function of the type of wave (e.g., head wave, direct wave, etc.). Most of the schemes, which estimate first-arrival traveltimes, are based on the solution of eikonal equation. Vidale (1988, 1990) proposed different versions of finite difference solution of the eikonal equation along expanding square (in 2D) and along an expanding cube (in 3D) to compute first-arrival traveltimes in isotropic media. Schneider et al. (1992) proposed a method based on Fermat's principle for traveltime computation in isotropic media that uses a local ray-trace solution of the eikonal equation. Ray tracing can be classified into two categories: (1) initial-value ray tracing and (2) boundary value ray tracing. In initial-value ray tracing, a fan of rays is shot from the source in the model, and some kind of extrapolation scheme is used to estimate the traveltime at a particular point in the model. Paraxial extrapolation (Červený, 1985) is one of the most popular methods of traveltime extrapolation. An excellent review on ray theory can be found in Červený (2001).

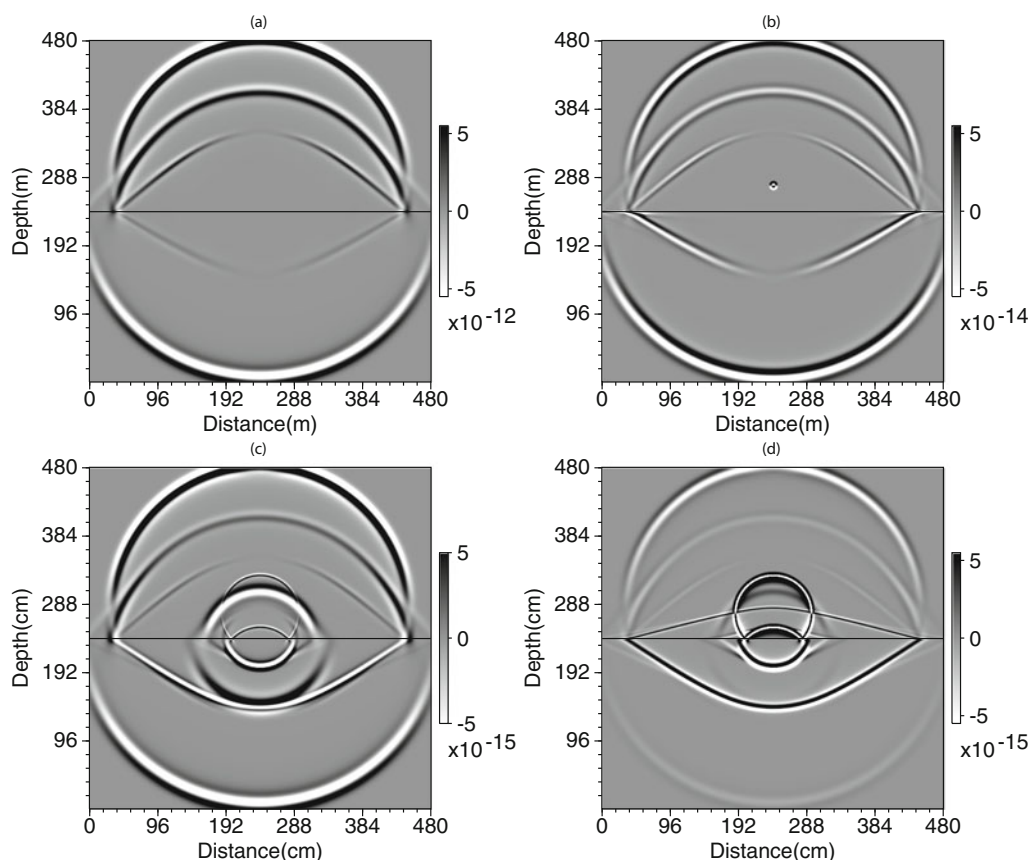
Poroelastic and Thermoelastic Media

In addition to complex geometries, different rheologies, and anisotropy, current numerical techniques are concentrating efforts to include porosity and fluid saturation in their models. Reservoir rocks can be modeled as a porous solid matrix where the pore space is fully saturated with a pore fluid. The analysis of wave propagation in fluid-saturated porous media is different in nature than that of elastic wave propagation. Biot's model is generally used to describe this effect (Biot 1956, 1962). Two approaches exist to derive the poroelastic equations of motion and the constitutive laws. One method is to use homogenization methods. The laws to describe the

medium are first derived on the microscopic scale of porosity. Then they are passed to macroscopic scale by homogenization. Pride et al. (1992) used an averaging procedure to derive the poroelastic equations similar to Biot (1956) and Biot (1962). O'Brine (2010) used rotated staggered grid to model wave propagation in 3D poroelastic media. Figure 5 shows snapshots of the vertical component of the solid velocity and the relative fluid velocity (with respect to the solid velocity) at times $t = 0.225$ ms and $t = 0.255$ ms, respectively. At the low-frequency range (Figs. 5a and b), the slow P wave is highly diffusive and vanishes that we can only see it in its static mode at the source location. However, at the high-frequency range, the slow P wave propagates through the medium (Figs. 5c and d) according to Biot's theory. This numerical simulations were produced using an optimal implicit finite difference method (Itzá et al. 2016) and are in agreement with the results given in the paper by Sheen et al. (2006). For a comprehensive account on this subject, the reader may be referred to the book by Carcione (2014). Moreover, a numerical algorithm for simulation of wave propagation in linear thermoelastic media, based on a generalized Fourier law of heat transport, computed by using a grid method based on the Fourier differential operator is presented in Carcione et al. (2019).

Boundaries

When modeling the propagation of elastic waves in any medium, the waves reaching the neighborhood of the boundaries will create reflected waves. This is a nonphysical phenomenon that we want to avoid. On top of the model, we would like to impose a free surface. Therefore for the top boundary, we need a reflected boundary, and for the other three boundaries (in 2D), we need nonreflecting absorbing boundaries. In most of the numerical schemes described here, there are boundary conditions to be imposed, or it is necessary to implement absorbing boundary conditions to eliminate spurious reflections due to the finite computational domain. The free surface boundary condition is implemented by adding a fictitious line of grid nodes next to the top boundary and using the one-side differences to approximate normal derivatives and centered difference to approximate tangential derivatives (see Kelly et al. (1976) and Graves (1996)). The simplest and more effective absorbing boundary conditions consist of a sponge along the artificially truncated edges of the domain. Clayton and Enquist (1977) proposed a kind of absorbing boundary condition based on replacing the wave equation in the boundary region by one-way wave equation that does not allow energy to propagate from the boundaries to the numerical domain. An alternative scheme suggested by Cerjan et al. (1985) is based on gradual reduction of the amplitudes in a strip of nodes along the boundaries for the mesh. Béranger (1994, 1996) first introduced perfectly



Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Fig. 5 Snapshots of vertical component from a gas-water contact model. (a, c) Snapshots of solid velocity. (b, d) Snapshots of relative fluid velocity (to solid velocity)

matched layers (PMLs) for electromagnetism. The main idea is to define a selective attenuation of the fields propagating in one prescribed direction. PMLs for elastic wave propagation were developed by Basu and Chopra (2003, 2004) and Komatitsch and Martin (2007). The velocity-stress formulation of PMLs for elastic wave equations has been introduced by Collino and Tsogka (2001). It was applied to viscoelastic (Martin and Komatitsch 2009) or poroelastic (Zeng et al. 2001; Martin et al. 2008) media. Meza-Fajardo and Papageorgiou (2008) demonstrated that for any anisotropic media, the classic PML exhibits instabilities. In addition, they presented a generalization of the classical perfectly matched layer PML (called multi-axial perfectly matched layer (M-PML)) to a medium in which damping profiles are specified in more than one direction.

Hybrid absorbing boundary conditions can significantly suppress artificial boundary reflections by inserting a transition area between the boundary and the inner area. The wavefield in the transition area is computed by linearly weighting the wavefield from one-way wave equation and that from two-way wave equation. This leads to a smooth variation from the computational area to the boundary via the transition area and thus greatly reduces boundary reflections.

In Liu and Sen (2018), the authors proposed two techniques to further enhance the absorbing effect. Firstly, they widen the boundary from one point to several points. Secondly, they adopt nonlinear weighting coefficients instead of linear ones.

Summary

Modeling the Earth is a challenging task. With the increase of computational power, direct modeling techniques start to open the door to inversion problems. All numerical methods share at the same time their own disadvantages and weaknesses (high computational costs and storage requirements, poor accuracy, nonstructured grids, hard implementations, system matrix full) with the agreeable flavor of equivalent advantages (reduction of one dimension of the problem, system matrix diagonal or sparse, no need for absorbing boundaries, easy implementation, easy to model topographies or complex geometries). The nature of the problem gives us the appropriate guide to choose one method over others. In this entry, we briefly cover some numerical techniques to model elastic waves in heterogeneous media. However, the literature on all available methods is very extensive, and we include only some references.

Cross-References

- [Body Waves](#)
- [Earthquake Rupture: The Inverse Problem](#)
- [Poroelectricity](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Signals in Well Water Observations](#)
- [Seismogram Interpretation](#)
- [Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting](#)
- [Statistical Seismology](#)
- [Surface Waves](#)
- [Tsunami](#)

Acknowledgements This work was partially supported by CONACYT, México, under PROINNOVA projects 251910 and, 241763 and DGAPA-UNAM Projects IN100917, IN107720 and IN121709, Mexico.

Bibliography

- Advances in Wave Propagation in Heterogeneous Earth. Editors: Ru-Shan Wu, Valerie Maupin. Series Editor: Renata Dmowska. Academic Press 2007, pp 626, Hardcover ISBN: 9780120188505
- Aki K, Richards PG (1980) Quantitative seismology. W. H. Freeman & Co, New York
- Almuhaidib AM, Toksöz MN (2015) Finite difference elastic wave modeling with an irregular free surface using ADER scheme. *J Geophys Eng* 12(3):435–447
- Backus G (1962) Long-wave elastic anisotropy produced by horizontal layering. *J Geophys Res* 67(11):4427–4440
- Basu U, Chopra AK (2003) Perfectly matched layers for time-harmonic elastodynamics of unbounded domains: theory and finite-element implementation. *Comp Meth Appl Mech Eng* 192:1337–1375
- Basu U, Chopra AK (2004) Perfectly matched layers for transient elastodynamics of unbounded domains. *Int J Numer Meth Eng* 59:1039–1074
- Bayliss A, Jordan KE, LeMesurier BJ, Turkel E (1986) A fourth-order accurate finite-difference scheme for the computation of elastic waves. *Bull Seism Soc Am* 76:1115–1132
- Béranger JP (1994) A perfectly matched layer for the absorption of electromagnetic waves. *J Comp Phys* 114:185–200
- Béranger JP (1996) Three-dimensional perfectly matched layer for the absorption of electromagnetic waves. *J Comp Phys* 127:363–379
- Bielak J, Ghattas O, Kim EJ (2005) Parallel octree-based finite element method for large-scale earthquake ground motion simulation. *Comput Model Eng Sci* 10:99–112
- Biot M (1956) Theory of propagation of elastic waves in a fluid saturated porous solid. I-low frequency range. *J Acoust Soc Am* 28:168–178
- Biot M (1962) Mechanics of deformation and acoustic propagation in porous solids. *J Appl Phys* 33:1482–1498
- Bonnet M (1999) Boundary integral equation method for solids and fluids. Wiley, New York
- Bouchon M, Sánchez-Sesma FJ (2007) Ch. 3, Boundary integral equations and boundary element methods in elastodynamics. In: Wu R-S, Maupin V, Dmowska R (eds) *Advances in wave propagation in heterogeneous earth*. Advances in geophysics, vol 48. Elsevier-Academic Press, New York, Boston, pp 157–190
- Capdeville Y, Marigo J-J (2007) Second order homogenization of the elastic wave equation for non-periodic layered media. *Geophys J Int* 170:823–838
- Capdeville Y, Marigo JJ (2010) 1-D non-periodic homogenization for the seismic wave equation. *Geophys J Int* (In press)
- Carcione JM (1994) The wave equation in generalized coordinates. *Geophysics* 59:1911–1919
- Carcione JM (2014) Wave field in real media. Wave propagation in anisotropic, anelastic, porous and electromagnetic media. In: *Handbook of geophysical exploration*, vol 38, 3rd edn. Elsevier, Amsterdam
- Carcione JM, Kosloff D, Behle A, Seriani G (1992) A spectral scheme for wave propagation simulation in 3-D elastic-anisotropic media. *Geophysics* 57(12):1593–1607
- Carcione JM, Herman GC, ten Kroode APE (2002) Seismic modeling. *Geophysics* 67(4):1304–1325
- Carcione JM, Wang Z-W, Ling W, Salusti E, Ba J, Li-Yun F (2019) Simulation of wave propagation in linear thermoelastic media. *Geophysics* 84(1):T1–T11
- Cerjan C, Kosloff D, Kosloff R, Reshef M (1985) A nonreflecting boundary condition for discrete acoustic and elastic wave equations. *Geophysics* 50:705–708
- Cervený V (1985) The application of ray tracing to the numerical modeling of seismic wavefields in complex structures: handbook of geophysical. Exploration. Geophysical Press, London
- Cervený V (2001) Seismic ray theory. Cambridge University Press, Cambridge, UK
- Chaillat S, Bonnet M, Semblat JF (2009) A new fast multi-domain BEM to model seismic wave propagation and amplification in 3D geological structures. *Geophys J Int* 177:509–531
- Christensen RM (1979) Mechanics of composite materials. Wiley, New York
- Chu C, Stoffa PL (2012) Implicit finite-difference simulations of seismic wave propagation. *Geophysics* 77(2):T57–T67
- Claerbout JF (1968) Synthesis of a layered medium from its acoustic transmission response. *Geophysics* 33(2):264–269
- Claerbout JF (1985) *Imaging the Earth's Interior*. Blackwell Scientific Publications, Ltd. 108 Cowley Rd. Oxford OX4 1JF United Kingdom ISBN:978-0-86542-304-6
- Clayton R, Enquist B (1977) Absorbing boundary conditions for acoustic and elastic wave equations, *Bulletin of the Seismological Society of America* 67:1529–1540
- Cockburn B, Karniadakis G, Shu C (2000) *Discontinuous Galerkin methods: theory, computation and applications*, vol 11 of Lecture notes in computational science and engineering. Springer-Verlag Berlin Heidelberg
- Collino F, Tsogka C (2001) Application of the perfectly matched absorbing layer model to the linear elastodynamic problem in anisotropic heterogeneous media. *Geophysics* 66:294–307
- Cruz-Atienza VM, Virieux J, Aochi H (2007) 3D finite-difference dynamic-rupture modelling along non-planar faults. *Geophysics* 72: SM123
- Dimitri Komatitsch; Jean-Pierre Vilotte (1998) The spectral element method: An efficient tool to simulate the seismic response of 2D and 3D geological structures. *Bulletin of the Seismological Society of America* No. 88(2):368–392
- Dominguez J (1993) *Boundary elements in dynamics*. Elsevier Applied Science Publishers, London
- Emmerman HS, Schmidt W, Stephen FLA (1982) An implicit finite-difference formulation of the elastic wave equation. *Geophysics* 47(11):1521–1526
- Fernandez P, Christophe A, Terrana S, Nguyen NC, Peraire J (2018) Hybridized discontinuous Galerkin methods for wave propagation. *J Sci Comput* 77(3):1566–1604
- Frankel A, Leith W (1992) Evaluation of topographic effects on P and S waves of explosions at the northern Novaya Zemlya test site using 3-D numerical simulations. *Geophys Res Lett* 19:1887–1890
- Frankel A, Vidale J (1992) A three-dimensional simulation of seismic waves in the Santa Clara Valley, California, from a Loma Prieta aftershock. *Bull Seismol Soc Am* 82:2045–2074

- Fujiwara H (1998) The fast multipole method for the integral equations of seismic scattering problems. *Geophys J Int* 133:773–782
- Fujiwara H (2000) The fast multipole method for solving integral equations of three-dimensional topography and basin problems. *Geophys J Int* 140:198–210
- Furumura T, Chen L (2005) Parallel simulation of strong ground motions during recent and historical damaging earthquakes in Tokyo, Japan. *Parallel Comput* 31(2):149–165
- Furumura T, Kennett BLN (1997) On the nature of regional phases – II On the effect of crustal barriers. *Geophys J Int* 129:221–234
- Furumura T, Kennett BLN, Takenaka H (1998) Parallel 3-D pseudo-spectral simulation of seismic wave propagation. *Geophysics* 63:279–288
- Gao K, Huang L (2017) An improved rotated staggered-grid finite-difference method with fourth-order temporal accuracy for elastic-wave modeling in anisotropic media. *J Comput Phys* 350:361–386
- Graves RW (1996) Simulating seismic wave propagation in 3D elastic media using staggered-grid finite differences. *Bull Seismol Soc Am* 86:1091–1106
- Hesthaven JS, Warburton T (2008) *Nodal Discontinuous Galerkin methods. Texts in applied mathematics*. Springer, New York
- Higdon RL (1991) Absorbing boundary-conditions for elastic-waves. *Geophysics* 56:231–241
- Hudson JA (1991) Overall properties of heterogeneous material. *Geophys J Int* 107:505–511
- Iturrarán-Viveros U, Vai R, Sánchez-Sesma FJ (2005) Scattering of elastic waves by a 2-D crack using the Indirect Boundary Element Method (IBEM). *Geophys J Int* 162:927–934
- Itzá R, Iturrarán-Viveros U, Parra JO (2016) Optimal implicit 2-D finite differences to model wave propagation in poroelastic media. *Geophys J Int* 206(2):1111–1125
- Käser M, Dumbster M (2006) An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes – I. The two-dimensional isotropic case with external source terms. *Geophys J Int* 166:855–877
- Kausel E (2006) *Fundamental solutions in elastodynamics. A compendium*. Cambridge University Press, New York
- Kelly KR, Ward R, Treitel W, Alford RM (1976) Synthetic seismograms: a finite difference approach. *Geophysics* 41:2–27
- Komatitsch D, Martin R (2007) An unsplit convolutional perfectly matched layer improved at grazing incidence for the seismic wave equation. *Geophysics* 72:SM155–SM167
- Komatitsch D, Tromp J (1999) Introduction to the spectral-element method for 3-D seismic wave propagation. *Geophys J Int* 139:806–822
- Komatitsch D, Vilotte J-P (1998) The spectral element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures. *Bull Seismol Soc Am* 88:368–392
- Kosloff D, Baysal E (1982) Forward modeling by the Fourier method. *Geophysics* 47:1402–1412
- Kreiss H-O, Oliger J (1972) Comparison of accurate methods for the integration of hyperbolic equations. *Tellus* 24:199–215
- Li Q, Wu G, Wu J, Duan P (2019) Finite difference seismic forward modeling method for fluid–solid coupled media with irregular seabed interface. *J Geophys Eng* 16(1):198–214
- Liu Y (2013) Globally optimal finite-difference schemes based on least squares. *Geophysics* 78:T113–T132
- Liu Y (2014) Optimal staggered-grid finite-difference schemes based on least-squares for wave equation modelling. *Geophys J Int* 197:1033–1047
- Liu Y, Sen MK (2009) A practical implicit finite-difference method: examples from seismic modelling. *J Geophys Eng* 6:31–249
- Liu Y, Sen MK (2018) An improved hybrid absorbing boundary condition for wave equation modeling. *J Geophys Eng* 15(6):2602–2613
- Liu Z, Liu L, Liang J, Zhou Y (2016) An indirect boundary element method to model the 3-D scattering of elastic waves in a fluid-saturated poroelastic half-space. *Eng Anal Bound Elem* 66:91–108
- Liu H, Sen MK, Spikes KT (2018) 3D simulation of seismic-wave propagation in fractured media using an integral method accommodating irregular geometries. *Geophysics* 83(1):1JF–1Z3
- Maeda T, Takemura S, Furumura T (2017) OpenSWPC: an open-source integrated parallel simulation code for modeling seismic wave propagation in 3D heterogeneous viscoelastic media. *Earth Planets Space* 69:102
- Manolis GD, Beskos DE (1988) *Boundary element methods in elastodynamics*. Unwin Hyman, London
- Manolis GD, Shaw RP (1996) Green's function for the vector wave equation in mildly heterogeneous continuum. *Wave Motion* 24:59–83
- Marfurt KJ (1984) Accuracy of finite-difference and finite element modeling of the scalar and elastic wave equations. *Geophysics* 49:533–549
- Martin R, Komatitsch D (2009) An unsplit convolutional perfectly matched layer technique improved at grazing incidence for the viscoelastic wave equation. *Geophys J Int* 179(1):333–344
- Martin R, Komatitsch D, Ezziani A (2008) An unsplit convolutional perfectly matched layer improved at grazing incidence for seismic wave equation in poroelastic media. *Geophysics* 73(4):T51–T61
- Meza-Fajardo K, Papageorgiou AS (2008) A non convolutional, split-field, perfectly matched layer for wave propagation in isotropic and anisotropic elastic media: stability analysis. *Bull Seismol Soc Am* 98(4):1811–1836
- Moczo P, Robertsson JOA, Eisner L (2007) The finite-difference time-domain method for modeling of seismic wave propagation. *Adv Geophys* 48:421–516
- Mufti IR (1985) Seismic modeling in the implicit mode. *Geophys. Prosp* 33:619–656
- O'Brine G (2010) 3D rotated and standard staggered finite-difference solutions to Biot's poroelastic wave equations: stability condition and dispersion analysis. *Geophysics* 75:T111–T119
- Olsen KB, Archuleta RJ, Matarrese JR (1995) Magnitude 7.75 earthquake on the San Andreas fault: three-dimensional ground motion in Los Angeles. *Science* 270:1628–1632
- Payton RG (1983) *Elastic wave propagation in transversely isotropic media*. Martinus Nijhoff Publishers, The Hague
- Perton M, Contreras-Zazueta MA, Sánchez-Sesma FJ (2016) Indirect boundary element method to simulate elastic wave propagation in piecewise irregular and flat regions. *Geophys J Int* 205(3):1832–1842
- Pointer D, Liu E, Hudson J (1998) Numerical modeling of seismic waves generated by hydrofractures: application of the indirect boundary element method. *Geophys J Int* 135:289–303
- Pride S, Gangi A, Morgan F (1992) Deriving the equations of motion for porous isotropic media. *J Acoust Soc Am* 92(6):3278–3290
- Priolo E, Carcione JM, Seriani G (1994) Numerical simulation of interface waves by high-order spectral modeling techniques. *J Acoust Soc Am* 95:681–693
- Rao Y, Wang Y (2018) Seismic waveform simulation for models with fluctuating interfaces. *Sci Rep* 8(1):3098
- Richtmyer RD, Morton KW (1967) *Difference methods for initial-value problems*, 2nd edn. Wiley, New York
- Rokhlin V (1990) Rapid solution of integral equations of scattering theory in two dimensions. *J Comp Phys* 86:414–439
- Saenger EH, Bohlen T (2004) Finite-difference modeling of viscoelastic and anisotropic wave propagation using the rotated staggered grid. *Geophysics* 69:583–591
- Saenger EH, Gold N, Shapiro SA (2000) Modeling the propagation of elastic waves using a modified finite-difference grid. *Wave Motion* 31:77–92
- Saito T, Furumura T (2009) Three-dimensional tsunami generation simulation due to sea-bottom deformation and its interpretation based on the linear theory. *Geophys J Int* 178:877–888
- Sánchez-Sesma FJ, Madariaga R, Irikura K (2001) An approximate elastic two-dimensional Green's function for a constant-gradient medium. *Geophys J Int* 146:237–248

- Schneider WA, Ranzinger KA, Balch AH, Kruse C (1992) A dynamic programming approach to first arrival 765 traveltimes computation in media with arbitrary distributed velocities. *Geophysics* 57:39–50
- Sheen D-H, Tuncaya K, Baagb C-E, Ortoleva PJ (2006) Parallel implementation of a velocity-stress staggered-grid finite-difference method for 2-D poroelastic wave propagation. *Comput Geosci* 32:1182–1191
- Shukla K, Hesthaven JS, Carcione JM, Ye R, de la Puente J, Jaiswal P (2019) A nodal discontinuous Galerkin finite element method for the poroelastic wave equation. *Comput Geosci* 23(3):595–615
- Shukla K, Chan J, de Hoop MV, Jaiswal P (2019b) A weight-adjusted discontinuous Galerkin method for the poroelastic wave equation: penalty fluxes and micro-heterogeneities. Submitted to *J Comput Phys*
- Seriani G, Priolo E, Carcione JM, Padovani E (1992) High-order spectral element method for elastic wave modelling. In: SEG 62 annual meeting, expanded abstracts, pp 1285–1288
- Somigliana C (1885) Sopra l'equilibrio di un corpo elastico isotropo. *Nuovo Cim* 17:140–148. <https://doi.org/10.1007/BF02817783>
- Sun YC, Zhang W, Xu J-K, Chen X (2017) Numerical simulation of 2-D seismic wave propagation in the presence of a topographic fluid–solid interface at the sea bottom by the curvilinear grid finite-difference method. *Geophys J Int* 210:1721–1738
- Thomsen L (1986) Weak elastic anisotropy. *Geophysics* 51:1954–1966
- Vidale JE (1988) Finite-difference calculation of travel times. *Bull Seismol Soc Am* 78:2062–2076
- Vidale JE (1990) Finite-difference calculation of traveltimes in three dimensions. *Geophysics* 55:521–526
- Virieux J (1984) SH-wave propagation in heterogeneous media: velocity-stress finite-difference method. *Geophysics* 49:1933–1957
- Virieux J (1986) P-SV wave propagation in heterogeneous media: velocity-stress finite-difference method. *Geophysics* 51:889–901
- Watanabe K, Payton R (2002) Green's function for SH-waves in a cylindrical monoclinic material. *J Mech Phys Solids* 50:2425–2439
- Watanabe K, Payton R (2006) Green's function for radially symmetric waves in inhomogeneous anisotropic solids. *Z Angew Math Mech* 86:642–648
- Yoshida KI (2001) Application of fast multipole method to boundary integral equation method. Ph.D thesis, Kyoto University
- Zeng YQ, He JQ, Liu QH (2001) The application of the perfectly matched layer in numerical modeling of wave propagation in poroelastic media. *Geophysics* 66(4):1258–1266
- Zou P, Cheng J (2017) Pseudo-spectral method using rotated staggered grid for elastic wave propagation in 3D arbitrary anisotropic media. *Geophys Prospect* 66(1):47–61

Seismic Waves, Scattering

Ludovic Margerin
Institut de Recherche en Astrophysique et Planétologie,
Observatoire Midi-Pyrénées/C.N.R.S., Toulouse, France

Definition

Scattering of seismic waves is best defined by reference to a laterally homogeneous or slowly varying medium where the wave fronts can be perfectly tracked and the propagation of waves can be successfully described by geometrical methods such as ray theory. In the presence of obstacles, or lateral

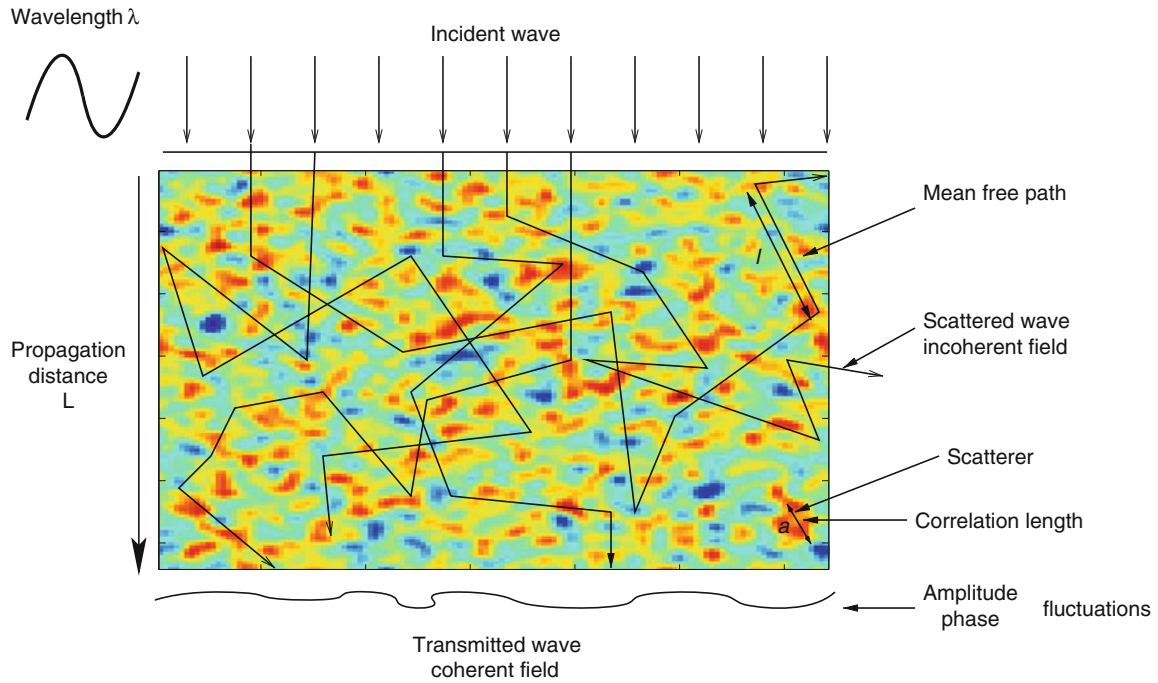
variations of elastic parameters, wave fronts are distorted and seismic energy can be deflected in all possible directions: a phenomenon known as wave scattering. Specific to seismic waves are the possible mode conversions between various polarizations: compressional, shear, and Rayleigh waves.

Introduction

The scattering of seismic waves is intimately related to the heterogeneous nature of Earth materials on a variety of spatial scales. Well-logs data or geological maps offer direct but limited access to the nature of heterogeneities in the Earth. As elastic waves can propagate all the way through the Earth, they constitute the primary tool to sound deep heterogeneities. While the pioneering studies of seismic scattering were largely focused on the lithosphere, the presence of scatterers in the mantle and solid core has been confirmed in a number of recent studies (see Shearer and Earle 2008, for a review). Hence, the field of seismic scattering has become of interest in global seismology too. For a thorough review of pre-1990s works in seismic scattering, we refer the reader to Wu (1989). A comprehensive reference on the topic is the book by Sato and Fehler (1998). Reviews of recent methodological and observational developments are given in Sato et al. (2008).

Statistical Description of the Earth

To introduce the various scale lengths that enter in the modeling of seismic scattering, it is convenient to consider a thought-experiment such as the one shown in Fig. 1. We imagine an incident plane wave propagating through a heterogeneous medium. As evidenced by the analysis of well-logs (see for instance Wu et al. 1994), the density and the elastic constants fluctuate in the Earth in a very complicated way. In such a situation, it is convenient to think of the Earth as a particular realization of an ensemble of random media sharing similar statistical properties. The goal of the stochastic imaging methods is to infer some information on the nature of the randomness from the statistical fluctuations of the wavefield. This is a difficult task because the statistical properties of the Earth are nonhomogeneous: they are expected to depend on the geological setting and on the depth. In addition, in the case of the Earth, the random fluctuations are superimposed on a deterministic background, usually represented by a 1-D reference model such as PREM (Dziewonski and Anderson 1981). It is customary to assume that the spatial variations of the background velocities are slow compared to the typical scale length of the random medium. In this limit of separation of spatial scales, it is possible to consider the scattering medium as locally homogeneous. We may therefore hope to characterize large-scale variations of small-scale heterogeneities.



Seismic Waves, Scattering, Fig. 1 Basic concepts of seismic scattering. We consider a thought-experiment where an incoming plane wave propagates through a chunk of random medium with typical dimension

L . As the plane wave progresses through the heterogeneous medium, scattering operates and deflects energy in all space directions. The transmitted wavefield is distorted and attenuated

If we think of the medium as random and continuous, the fluctuations of a parameter such as the density ρ are characterized to lowest order by a correlation function (Rytov et al. 1989b):

$$\psi(\mathbf{x}) = \langle \rho(\mathbf{r} - \mathbf{x}/2) \rho(\mathbf{r} + \mathbf{x}/2) \rangle, \quad (1)$$

where the brackets denote an average over the ensemble of random media. In Eq. 1, we have used an assumption of statistical homogeneity. In the case of a locally homogeneous medium, the correlation function may be allowed to vary slowly with the variable $\mathbf{r}$. $\psi(\mathbf{x})$ is maximum for $x = 0$, where it is equal to the total variance of the fluctuations denoted by ϵ^2 and goes to zero over a typical spatial scale a which is known as the correlation length. Common choices for ψ are Gaussian and exponential functions but a wide range of other possibilities exist (Klimeš 2002). An equivalent description of the medium fluctuations is provided by the power spectrum of heterogeneities $F(k)$ defined as:

$$\psi(\mathbf{x}) = \frac{1}{(2\pi)^3} \iint \int_{-\infty}^{+\infty} \Phi(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{x}} d^3k \quad (2)$$

$\Phi(\mathbf{k})$ quantifies the distribution of the total variance of the fluctuations over all possible length scales. A very useful tool to represent the heterogeneity of fluid or solid geophysical media is the Von-Karman correlation function, whose power spectrum can be written as (Sato and Fehler 1998):

$$\Phi(\mathbf{k}) = \frac{8\pi^{3/2} \epsilon^2 a^3 \Gamma(k + \frac{3}{2})}{\Gamma(k) (1 + k^2 a^2)^{k+3/2}}. \quad (3)$$

In Eq. 3, ϵ denotes the RMS fluctuations, a is the correlation length and $\kappa > 0$ is the Hurst exponent which acts as a cutoff of the small-scale features in the medium. The power spectrum (Eq. 3) shows a plateau at small wave numbers up to a corner value $k_c \sim 1/a$. For k much larger than k_c , the power spectrum decreases algebraically as $k^{2\kappa+3}$. Note, however, that this description does not apply to Gaussian random media. This is not a severe restriction, since Gaussian media are often regarded as too smooth to represent faithfully the heterogeneities in the Earth. This point of view is confirmed by the analysis of well-logs (Shiomi et al. 1997), which further shows that only the parameter κ can usually be recovered from observations.

The Coherent and Incoherent Fields

Following the statistical approach, it is natural to characterize the propagation in the random medium by calculating the moments of the wavefield u . This yields a simple classification of the various components of the random field. The first moment $\langle u \rangle$ defines the coherent wave (Rytov et al. 1989a; Sheng 1995) and corresponds to the part of the field which survives after an average over the statistical ensemble has been performed. In a typical seismological experiment, the measured field is the sum of two terms: the ensemble average

$\langle u \rangle$ and a fluctuating part u_f . The coherent field mostly contributes in a time window around the seismic primaries, which may also be termed the ballistic waves. Although the coherent field is an important component of the ballistic field, it has to be carefully distinguished from it (Derode et al. 2001). Statistical wave theory shows that the mean field decays like $e^{-L/2l}$, where L is the propagation distance and l is known as the mean free path (Sheng 1995). The mean free path gives the typical spatial scale over which a significant amount of energy has been transferred from the coherent to the incoherent or fluctuating part of the wavefield. The incoherent part of the wavefield contributes to all portions of the seismogram and forms the coda of the earthquake records. It is important to realize that the term “incoherent” just indicates the generation of waves that propagate in a direction which differs from that of the incident wave. It does not imply that the phase information has been lost. In section “[Interferometry with Scattered Waves](#)”, we will demonstrate that the so-called incoherent field is in fact highly “coherent” in the sense that it gives rise to very useful interference effects.

The Propagation Regimes

The distinction between the fluctuating and ensemble average part of the wavefield is useful to characterize the propagation regime. For propagation distances typically less than the mean free path, the measured field is dominated by the coherent wave: this is the weak fluctuation regime (Rytov et al. 1989a). As shown in Fig. 1, after propagating through several inhomogeneities, the most important observation is the distortion of the incident plane wave front. This gives rise to fluctuations of both phase and amplitude, from which information on the statistical properties of the medium can be extracted. For propagation distances much larger than the mean free path, the coherent wave completely vanishes and one is left with the fluctuating part only: this is the strong fluctuation regime (Rytov et al. 1989a). Since by definition the ensemble average of the fluctuating part u_f is zero, it is only through the consideration of the second moment of the field that one can model the propagation of scattering when fluctuations are strong. Specific approaches will be developed in sections “[Envelope Modeling: Markov Approximation](#)” and “[Envelope Modeling: Radiative Transfer](#)”.

At this stage, it is natural to raise the following question: How can we measure the mean free path in the Earth? First, it is important to note that one has to consider at least two different mean free paths since both longitudinal and transverse waves propagate through an elastic medium. In the laboratory, there are well-calibrated techniques to measure the mean free path based on ensemble averaging. Unfortunately, these methods cannot be transposed to field experiments. Instead, the common seismological practice consists in correcting the seismograms for the travel-time fluctuations

caused by long wavelength variations of the velocity before stacking or averaging the records. The field obtained after such a processing is called the traveltime-corrected mean field. Sato (1982, 1984) and Wu (1982) have demonstrated that this field decays exponentially on a scale which is much larger than the mean free path, because the traveltime correction in fact removes the long wavelength components of the medium fluctuations. A theoretical discussion of various averaging procedures and their application to seismic data is given by Shapiro and Kneib (1993).

Analysis of Transmission Fluctuations

As announced in introduction, we now review the various approaches that have been developed to infer the statistical properties of the Earth from an analysis of seismic wave scattering. We will first introduce a method which is well adapted to the analysis of fluctuations of amplitude and phase of teleseismic P waves recorded on a dense array. Such an approach has been pioneered in the 1970s by Aki (1973) and has been subsequently developed by Wu and coworkers (Wu and Flatté 1990; Zheng et al. 2007; Zheng and Wu 2008). In this approach, one considers a small time window around the direct P -wave at teleseismic distance. It is usually assumed that at sufficiently large distance from the source, the wave front incident from the mantle can be approximated locally by a plane. By using sufficiently deep earthquakes, it is possible to avoid the effects of strong heterogeneities in the vicinity of the source. As the plane wave progresses through the lithosphere, the wave front is distorted and the amplitude and phase of the ballistic waves measured on the array fluctuate. Zheng and Wu (2005) have very carefully discussed the correct measurement practice. The main pulse recorded on the vertical component of the seismometer is first isolated from the rest of the signal. After Fourier transformation, the amplitude A and the phase ϕ of the pulse is obtained at a given frequency. This operation is repeated at all stations of the array, and the phase field is subsequently unwrapped spatially, i.e., possible phase discontinuities are removed. It is not correct to measure the phase of the field by picking the first arrival since, as a consequence of scattering, the medium is dispersive and the first arrivals propagate with the group velocity.

Basic observables are the fluctuations of the phase and of the logarithm of the amplitude. The reason why one considers the logarithm of the amplitude is that this quantity enters in the Rytov representation of the wavefield (Rytov et al. 1989a):

$$u = u_0 e^{\psi}, \quad (4)$$

where u_0 is the reference plane wavefield in the background medium and ψ is a complex phase:

$$\operatorname{Re} \psi = \ln \frac{A}{A_0} \quad \operatorname{Im} \psi = \phi - \phi_0 \quad (5)$$

Scattering theory based on the Rytov representation is more accurate than the standard Born approximation in the transmission geometry. In Eq. 5, A (resp. A_0) and ϕ (resp. ϕ_0) denote the amplitude and unwrapped phase of the field (resp. reference field). From the array measurement, one can estimate the transverse coherence function of the phase field which is defined as:

$$C_\phi(\mathbf{r}) = \langle \phi(\mathbf{x}) \phi(\mathbf{r} + \mathbf{x}) \rangle, \quad (6)$$

where the points $\mathbf{x}$ and $\mathbf{x} + \mathbf{r}$ lie on the surface of the Earth. More generally, one can consider the transverse angular coherence function of the two phase fields corresponding to two incoming plane waves with different incident wavevectors $\mathbf{k}$ and $\mathbf{k}'$ (Wu and Flatté 1990; Chen and Aki 1991).

The interpretation of the phase coherence function is based on two major theoretical approaches, with different domains of validity. These approaches rely on the basic physical assumption that the medium fluctuations have a spatial scale, a , which is much larger than the probing wavelength, λ . In this regime, the scattering occurs preferentially in a small angular cone around the forward direction. As a consequence, one can usually neglect the backscattered waves and consider one-way propagation only. The crucial nondimensional parameter that governs the properties of the phase is known as the wave parameter, D , which is defined as (Rytov et al. 1989a):

$$D = \frac{\lambda L}{a^2} \quad (7)$$

In the numerator of Eq. 7, the length $\sqrt{\lambda L}$ provides the typical size of the first Fresnel zone which can be interpreted as the typical volume sampled by the waves around a ray of length L . The region $D \ll 1$ can be conveniently treated with the methods of geometrical optics. In this regime, there are some well-established results for the phase coherence function of a single plane wave with vertical incidence: (1) The variance of the phase is proportional to the total variance of the medium fluctuations and increases linearly with the propagation distance L . (2) The typical correlation length of the phase is proportional to the correlation length of the random medium. This result makes sense physically, since parts of the incident wave front which are located a distance a apart – a , the correlation length – will visit different inhomogeneities. This provides a very useful method to map the typical scale of heterogeneities. The geometrical optics approach is rather flexible and allows the treatment of more complex

situations. Kravtsov et al. (2003) have adapted the method to the reflexion geometry usually encountered in exploration geophysics.

The method of geometrical optics usually breaks down for $D > 1$ as caustics may develop for sufficiently large propagation distances. In order to account properly for diffraction effects, it is important to use a wave-based approach. As in the problem of transmission fluctuations in which the propagation is essentially one-way, it is convenient to consider the depth variable z as an evolution variable and to neglect the backscattered waves. In this case, the usual Helmholtz equation can be replaced by the following parabolic approximation for the wavefield:

$$2ik_0 \frac{\partial u}{\partial z} + \Delta_\perp u - 2k_0^2 \frac{\delta c}{c_0} u = 0, \quad (8)$$

where $\Delta_\perp = \partial_x^2 + \partial_y^2$ represents the Laplace operator in the plane perpendicular to the main propagation direction. The parabolic approximation is accurate for low-angle scattering only, which is consistent with the scaling $\lambda/a \ll 1$. In this regime, a very powerful approach couples scattering theory and the Rytov representation of the wavefield to obtain expressions for the coherence function that extend the results of geometrical optics to the region $D > 1$. We refer to Wu and Flatté (1990) for further details on the derivation.

The first application to data was made by Aki who estimated the correlation length and total variance of the velocity fluctuations from the cross-coherence of amplitude and phase under the LASA array (Aki 1973). In this pioneering work, he further approximated the correlation function of the medium with a Gaussian function. It was later realized that this choice was not always the most appropriate since many studies revealed that the more general Von-Karman correlation function could be used to describe a broad class of random media. Flatté and Wu (1988) developed a two-layer model of heterogeneity of the upper mantle based on observed transverse and angular correlation functions at NORSAR. By superimposing a layer with Hurst exponent 1/2 extending from 15 to 250 km depth upon a layer with flat power spectrum extending from 0 to 200 km depth, they were able to explain satisfactorily the joint measurements of phase and amplitude fluctuations. The most recent developments by Zheng et al. (2007) and Zheng and Wu (2008) allow the treatment of velocity gradient in the reference medium. Their general formalism offers the possibility to perform a nonparametric inversion, i.e., the full power spectrum of the fluctuations at depth can in principle be mapped from the surface measurements of angular correlation functions. For a complete description of the method as well as numerous references to the literature, we refer to the review by Zheng and Wu (2008).

Envelope Modeling: Markov Approximation

We now consider a regime of propagation where the fluctuating part of the field is larger than the coherent part. In this regime, an interesting phenomenon termed “envelope broadening” can be used to retrieve information on the statistical fluctuations of velocities in the medium. Enveloped broadening is illustrated in Fig. 2, where small earthquakes have been recorded at epicentral distances of the order of 150 km in eastern Japan. On these data, we can observe a delay between the onset of the direct shear wave and the maximum of the amplitude. Such a delay is characteristic of the strong forward-scattering regime where the waves follow “snake-like” paths around the ballistic trajectory between source and station. As clearly shown in Fig. 2, envelope broadening depends on propagation distance and frequency band. It is also noticeable that very heterogeneous regions such as volcano areas give rise to much stronger envelope broadening than the usual crust. The clear path dependence suggests the possibility to use regional variations of the onset to peak delay time to map heterogeneities in the crust, in order to develop a stochastic tomography in the strong fluctuation regime.

Such an approach has been pioneered by Sato in the late 1980s. A recent review of the topic in seismology is given by Sato and Korn (2009). Based on the parabolic approximation, Sato (1989) derived an equation for the two-point, two-frequency coherence functions of the wavefield for incident plane waves. This technique which was originally introduced in astrophysics has been termed “Markov approximation.” In the case of regional propagation, it is preferable however to consider the radiation from a point source. Following Saito et al. (2002), we write the coherence function of the field as follows:

$$\Gamma(\mathbf{r}_{\perp 1}, \mathbf{r}_{\perp 2}, r, \omega_d, \omega_c) = \langle U(\mathbf{r}_{\perp 1}, r, \omega_1) \langle U(\mathbf{r}_{\perp 2}, r, \omega_2)^* \rangle, \quad (9)$$

where ω_c and ω_d denote the central $(\omega_1 + \omega_2)/2$ and modulation $\omega_1 - \omega_2$ frequency, respectively. The introduction of two frequencies is demanded by the nonstationarity of the envelope records. The source station distance is denoted by r and the points $\mathbf{r}_{\perp 1}$ and $\mathbf{r}_{\perp 2}$ lie in a plane perpendicular to the wave front. The time dependence of the observed coda envelopes around the frequency ω_c is obtained by setting $\mathbf{r}_{\perp 1} = \mathbf{r}_{\perp 2}$ and taking a Fourier transform over ω_d .

The basic idea of the so-called Markov approximation is to consider one-way propagation and to neglect backscattering. The different slices of the medium perpendicular to the main propagation direction independently perturb the local ray direction in a random fashion. As a consequence, the distribution of ray directions at a distance r_0 from the source depends solely on the heterogeneities that have been met for

$r < r_0$ (Sato 1989; Sato and Korn 2009). Such an approximation is valid for strong forward-scattering, i.e., when the relation $\lambda > a$ applies. This does not mean that the Markov approximation cannot be used if the medium contains heterogeneities with scale lengths shorter than the wavelength. Indeed, the large-scale fluctuations contribute predominantly to the field measured in the transmission geometry. Scattering at large angles caused by small-scale fluctuations mainly plays the role of an apparent energy loss which adds to the intrinsic losses (Sato 1989). In order to quantify the typical validity region of the Markov approximation, it is useful to introduce a quantity known as the transport mean free path (Sheng 1995) and defined as:

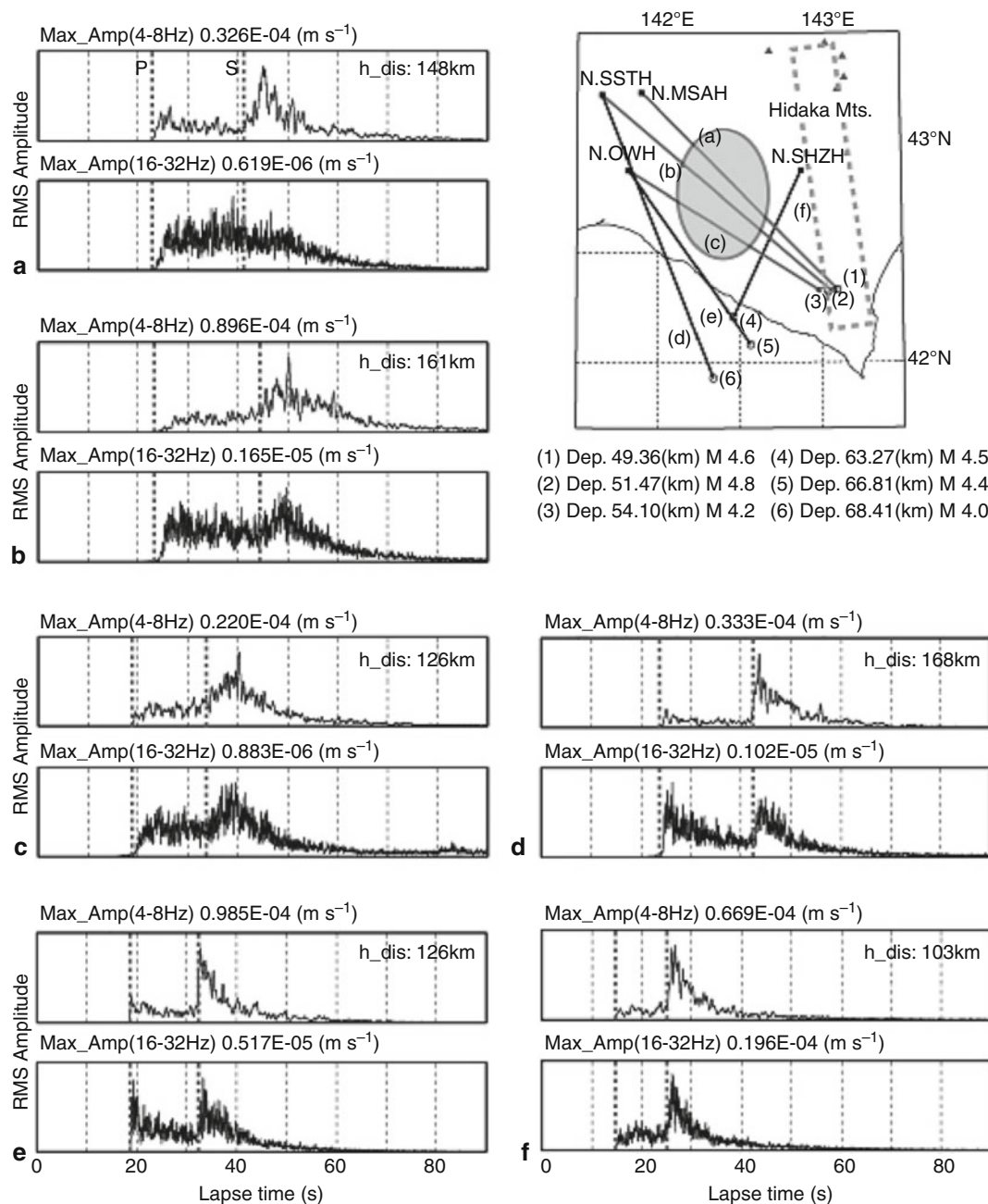
$$l^* = \frac{l}{1 - \langle \cos \theta \rangle}, \quad (10)$$

where $\langle \cos \theta \rangle$ is the mean cosine of the scattering angle. The transport mean free path can be interpreted as the typical length beyond which the waves have lost memory of their initial propagation direction. In the case of large-scale fluctuations, the transport mean free path l^* is typically one or two orders of magnitude larger than the scattering mean free path l . For propagation distances of the order of the transport mean free path, backscattering can no longer be ignored and one must appeal to other methods to be described in section “Envelope Modeling: Radiative Transfer”.

As carefully explained in the paper of Sato (1989), there are two basic phenomena that contribute to the observed envelope shape. The first one is termed “the wandering effect” and is purely statistical in nature. It corresponds to the small arrival time fluctuations in the line-of-sight propagation direction. It leads to a Gaussian spread of an initially delta-like pulse. This reshaping can only be observed upon ensemble averaging and is therefore not of primary importance in seismology. The second phenomenon is the increase of the typical duration of the seismogram with propagation distance which has been termed “pulse” – or may be more appropriately – “envelope broadening.” It reflects the angular spreading of the wave propagation directions as they interact with the medium heterogeneities. For a general Von-Karman correlation function, a time scale t_m characterizes the typical duration of the seismogram (Saito et al. 2002):

$$t_m = C \left(\varepsilon^{1/(p-1)} a^{-1} \right)^{2-2/p} \omega^{-2+4/p} r_0^{1+2/p}, \quad (11)$$

where ε denotes the root mean square fluctuations, ω is the central frequency of the signal, r_0 is the hypocentral distance and C is a numerical pre-factor which depends on the wavespeed, and on the Hurst exponent κ . The parameter p is an increasing function of κ and varies from 1.2 to 2 for



Seismic Waves, Scattering, Fig. 2 Example of earthquake records showing the phenomenon of envelope broadening. Paths a, b, c propagate through a very heterogeneous volcanic region. A large delay between the onset of S waves and the maximum amplitude is clearly

$0 < \kappa \leq 1$. Formula (11) reveals interesting characteristics of the envelope broadening phenomenon. For a smooth medium, i.e., a medium poor in small-scale features, the parameter p is close to 2 and, as a consequence, the typical envelop duration is frequency-independent. Such a result also holds for the Gaussian correlation function and is in sharp contradiction with observations. As a consequence, the frequency dependence of envelope duration offers access to the degree of roughness of Earth's heterogeneities. Accurate measurements of the exponent κ requires that absorption

observed. This phenomenon further depends on the frequency band considered. Paths d, e, f do not cross the heterogeneous region and show more impulsive envelopes with small onset to peak delay time. (Reproduced from Takahashi et al. 2007)

and – to some extent – scattering losses caused by small-scale heterogeneities be incorporated in envelope modeling (Saito et al. 2002).

In practical applications to data, a characteristic time, t_q , is measured as a function of frequency and hypocentral distance: $t_q = t_p + t_{1/2}$, where t_p is the duration from onset to peak and $t_{1/2}$ is the decay time of the envelope from the peak to RMS half-peak value. Note that the observations are usually focused on the S wave train for which pulse broadening is usually much clearer than for P waves. The observations are

then compared to those numerically obtained by the Markov approximation for a wide range of correlation lengths, RMS perturbations and intrinsic quality factor. In general it is not possible to estimate independently ϵ and a since various combinations of these two parameters may yield the same characteristic time t_m . However, by analyzing separately the distance and frequency dependence of t_q , the intrinsic quality factor Q_s , the parameter p , and in turn, the Hurst exponent κ can be estimated. For applications to seismic observations, we refer to, e.g., Scherbaum and Sato (1991), Obara and Sato (1995), and Saito et al. (2005).

As illustrated in Fig. 2, for comparable hypocentral distances and frequency bands, the envelope shapes may strongly depend on the direct ray path. Such an observation calls for the introduction of large-scale variations of small-scale heterogeneities. Such an approach has been taken by Takahashi et al. (2007) who mapped strongly heterogeneous regions beneath quaternary volcanoes in northeastern Japan from the observed path dependence of the characteristic broadening time t_q . A complete tomographic approach to estimate the power spectrum of small-scale heterogeneities from envelope broadening data was further developed and applied to northeastern Japan by Takahashi et al. (2009).

Gusev and Abubakirov (1999a, b) developed a tomographic method to map the transport mean free path in the Earth based on body wave envelope broadening. Their approach is simpler since they consider the inversion of a single parameter. From the knowledge of the frequency dependence of the transport mean free path, it is possible to put constraints on the parameters ϵ , a , and κ . The Markov method has recently been extended to the case of vector waves by Sato (2006, 2007, 2008). The theoretical approach relies on the introduction of scalar and vector potentials for P and S waves, respectively. The employed Helmholtz decomposition neglects the conversions between P and S waves, which is justified in the high-frequency regime $\lambda \gg a$, up to a characteristic coupling time τ^{PS} between P and S waves. Extensions of the Markov method to the vectorial case allow the modeling of the observed energy envelopes of the transverse components of the wavefield after the onset of the direct P -wave. The observation of such transverse energy is another manifestation of the randomization of the ray directions in a scattering medium. Alternative explanations for this phenomenon are mode conversions or anisotropy. In the case of a statistically homogeneous Gaussian medium with thickness h , Sato (2006) showed that the ratio between the peak total intensity and the peak intensity on the transverse component of the P pulse scales like $\epsilon^2 h/a$. Interestingly, this relation is free from absorption, since it cancels out in the intensity ratio. Based on this approach Kubanza et al. (2007) made a global study of energy envelopes of transverse components in the P wave train observed at teleseismic distances and found good correlation between the strength of heterogeneity and the tectonic setting.

Envelope Modeling: Radiative Transfer

When the propagation time and distances become very large compared to the ballistic time, or when the acquisition geometry is in the reflection mode, it becomes crucial to take into account scattering at large angle. A powerful approach to model seismogram envelopes in scattering media in such cases is provided by radiative transfer theory. In seismology, radiative transfer was introduced in the mid-1980s by Wu (1985) with a first application to data by Wu and Aki (1988). Physically, the equation of transfer is a local statement of energy conservation in a random medium and incorporates an arbitrarily high number of scattering events. The radiative transfer equation also takes into account the anisotropy of the scattering process, in particular large-angle scattering. Therefore, it enables the complete modeling of seismogram envelopes including coda waves, i.e., the late arrivals of the seismogram. A review of the radiative transfer approach in seismology is given by Margerin (2005).

The study of coda waves has been pioneered by Aki (1969) and Aki and Chouet (1975). In this last reference, the authors developed two famous models of scattering: the single-scattering approximation and the diffusion approximation. The single-scattering model applies for propagation distances of the order of, or less than the mean free path. This model has been very popular until studies by Gusev and Abubakirov (1987) and Hoshihara (1991) pointed out the importance of multiple scattering. Recently, the diffusion model has known a revived interest in connection with seismic experiments on volcanoes (Wegler and Lühr 2001; Wegler 2004). The single-scattering and diffusion models are respectively short-time and large-time asymptotics of the radiative transfer equation. Mathematically, radiative transfer takes the form of an integro-differential equation for an angularly resolved energy flux known as the specific intensity $I(\mathbf{r}, \hat{\mathbf{k}}, t, \omega)$. The specific intensity quantifies the amount of energy flowing around the direction $\hat{\mathbf{k}}$ in a small frequency band $[\omega, \omega + \delta\omega]$ during a time interval $[t, t + dt]$ through a small surface element located at $\mathbf{r}$. Although the concept of specific intensity was first introduced on a phenomenological basis, it is now well understood that it is in fact connected to the Wigner-Ville distribution of the wavefield (Ryzhik et al. 1996):

$$I(\mathbf{r}, \hat{\mathbf{k}}, t, \omega) = C \iint \langle u(t - \tau/2, \mathbf{r} + \mathbf{x}/2) u(t + \tau/2, \mathbf{r} - \mathbf{x}/2)^* \rangle e^{-i\mathbf{k} \cdot \mathbf{x} + i\omega\tau} d^3x d\tau, \quad (12)$$

where the brackets denote an ensemble average. The key to understand this definition is the notion of separation of scales. In Eq. 12, one must think of the intensity as a slowly varying function of the time t , as compared to the fast oscillations of

the wavefield described by the frequency ω . Similarly, the average intensity varies slowly in space as compared to the wavelength. Such a separation of scales is well satisfied by high-frequency seismograms and guarantees that the Wigner distribution can be interpreted as a local power spectrum of the wavefield. It is important to note that interference effects are neglected in the transfer approach. This can be justified on the basis of the small phase shifts that occur upon scattering and which will be different for waves visiting different heterogeneities.

An important step in the development of radiative transfer was the introduction of the Monte Carlo method, which is a flexible numerical technique to solve the transfer equation. It has been introduced in seismology by Gusev and Abubakirov (1987). The Monte Carlo method simulates the random walk of particles, which change direction of propagation at the end of each step. The step length has an exponential distribution with parameter the scattering mean free path, and the change of direction upon scattering is provided by a probabilistic interpretation of the differential scattering cross-section. The latter quantity can be obtained from the Born approximation at frequencies below the geometrical optics limit. The validity of this approximation has been discussed by Wegler et al. (2006a) and Przybilla and Korn (2008). The Monte Carlo method can be shown to yield an exact solution of the transfer equation (Papanicolaou et al. 2000). The variance of the result typically decreases like the square root of the number of simulated random walks.

An important application of the radiative transfer approach was developed by Fehler et al. (1992) and Hoshiya (1993) who proposed a method to infer the mean free path and intrinsic quality factor in the crust from the space time dependence of the energy radiated by small earthquakes. The method is based on the estimation of the energy in three successive time windows starting with the direct S wave arrivals. Such multiple lapse-time window analysis has been applied in various regions of the world to give estimates of the level of heterogeneity in the crust. Typically, the mean free path of the crust around 1 Hz ranges from a few tens to a few hundred kilometers. The method has some limitations since it assumes isotropic scattering and a uniform distribution of scatterers in a half-space.

It was pointed out by Gusev and Abubakirov (1987), Abubakirov and Gusev (1990) and Hoshiya (1995) that the observed shape of envelopes of body waves is incompatible with isotropic scattering, since no broadening occurs in this case. Based on the comparison between modeled and observed S wave seismogram envelopes, Gusev and Abubakirov (1996) were able to infer a Hurst exponent ranging between $1/4$ and $1/2$ in the crust. In the case of crustal propagation at regional distances, the field is dominated by guided waves. To take into account the role of stratification and scattering in the Earth, Hoshiya (1997), Margerin et al.

(1998), and Yoshimoto (2000) have shown how to incorporate depth-dependent velocity and mean free path in the Monte Carlo method. Margerin et al. (1999) numerically solved the radiative transfer equation in the crustal geometry and confirmed the importance of the leakage of energy at the Moho to explain the decay of the coda, as first proposed by Korn (1990).

Other applications of radiative transfer pertain to the study of the seismic source. Nakahara et al. (1998) proposed a method based on radiative transfer theory to infer the distribution of high-frequency radiation on the fault plane of large earthquakes. They considered the different part of the fault to be independent and modeled the propagation using a multiple isotropic scattering model. Sens-Schönfelder and Wegler (2006a) devised a method to infer the seismic moment of small crustal earthquakes recorded at regional distance. Another important application concerns the modeling of wave propagation in very heterogeneous structures such as volcanoes, where the mean free path can become as small as a few hundred meters. A review of recent development in the field is provided by Del Pezzo (2008). It is to be noted that in cases where the typical propagation distance is much larger than the mean free path, the radiative transfer equation can be simplified to a simple scalar diffusion equation for the total energy density of P and S waves. The diffusion approach to the modeling of seismic energy propagation in volcanoes has been advocated by Wegler (2004).

Radiative transfer theory is still in an active stage of development. A radiative transfer equation for elastic waves has been introduced in seismology by Zeng (1993), Sato (1994), and Ryzhik et al. (1996). This equation takes into account in a rigorous way the coupling between P and S waves. A Monte Carlo method to solve this elastic transfer equation was developed by Margerin et al. (2000). A comparison of finite-difference and Monte Carlo simulations by Przybilla and Korn (2008) demonstrated the accuracy of the radiative transfer approach to simulate the coupling and multiple scattering of P and S waves. An extension of the radiative transfer equation to incorporate the coupling between surface and body waves is a major challenge. An important step in this direction was made by Maeda et al. (2007) who incorporated the coupling between Rayleigh and body waves in the single-scattering approximation. The impact of anisotropic scale lengths has been studied by Margerin (2006) who derived a radiative transfer equation with an angular-dependent scattering mean free path.

Global-Scale Scattering

The subject of seismic scattering at the global scale was first developed in the early 1970s in connection with short-period precursors to phases such as PP (King et al. 1975) and PKP

(Cleary and Haddon 1972). Accompanying the improvement of the global seismic network and the availability of high-quality data, evidences for the importance of scattering in the deep Earth have accumulated in recent years. The study of precursors to deep phases such as *PKP* (Hedlin et al. 1997) and *PKKP* (Earle and Shearer 1997) has been used to put constraints on the heterogeneities in the deep mantle and core. Observation and modeling of anomalous propagation of seismic waves in the subducting lithosphere have put forward the role of scattering by elongated heterogeneities in slabs (Furumura and Kennett 2005). The presence of strong scatterers in Earth's inner core has been revealed by a number of studies (Vidale and Earle 2000; Koper et al. 2004; Poupinet and Kennett 2004). Vidale and Earle (2000) used data from nuclear explosions recorded on the LASA array to demonstrate that the long-lasting signal following the ballistic *PKiKP* phase was not caused by the response of the crust but were genuine scattered waves radiated from the inner core. They showed examples of slowly-emergent and long-lasting *PKiKP* signals recorded at an epicentral distance of about 60° . This last point is particularly important, since it corresponds to incidence angles of *P* waves at the inner core boundary such that the reflection coefficient vanishes. The coda of *PKiKP* thus provides a direct proof of the presence of scatterers in the inner core.

Since the deep Earth interior is usually thought to be much less heterogeneous than the lithosphere, much of the modeling work has been focused on the single scattering approximation (Hedlin et al. 1997; Vidale and Earle 2000; Leyton and Koper 2007). The radiative transfer approach was introduced in global seismology by Margerin and Nolet (2003a, b) to improve the modeling of the precursors to the *PKP* wave. In particular, they demonstrated the ability of radiative transfer to incorporate complicated ray geometry such as the one imposed by the severe drop of velocity at the core–mantle boundary. The Monte Carlo method was further developed by Shearer and Earle (2004) who proposed a scheme to simulate the multiple scattering of elastic waves at the global scale, including the coupling between *P* and *S* waves. Their method was applied to the study of the coda envelopes of teleseismic *P* waves. To explain the observations, they proposed a 1-D model of Earth small-scale heterogeneity composed of a heterogeneous 200 km top-layer (4% RMS perturbations) overlying a less heterogeneous 400-km layer (3% RMS perturbations). At the base of the mantle, a 600 km thick layer with 0.5% RMS perturbations is required to achieve a good fit to the data. Shearer and Earle (2004) concluded that attenuation is dominated by scattering in the upper-mantle and by absorption in the lower mantle.

In addition to global studies of scattering of short-period waves, an interest in the scattering of long-period surface waves has recently arisen. Maeda et al. (2006) performed an (f, k) analysis of the vertical components of three large

earthquakes recorded on the Japanese Hi-Net in the 90–180 s period band. They find that the signals recorded between the first multi-orbiting Rayleigh wave arrivals are composed of waves coming from all directions with phase velocities of the order of the Rayleigh wave fundamental mode. This suggests that in the period band of interest, scattered Rayleigh waves are present in the data at short lapse time. After typically 30,000 s, the multi-orbiting Rayleigh waves have strongly decayed. The (f, k) analysis demonstrates that the signal is dominated by waves with very large phase velocities, thereby implying the dominance of higher-mode Rayleigh waves. Such a phenomenon is to be expected on the ground that higher modes sample the deeper parts of the Earth with high intrinsic quality factor. To model the coda of multi-orbiting Rayleigh waves, Sato and Nishino (2002) developed a radiative transfer equation on the sphere assuming the multiple isotropic scattering of the fundamental mode Rayleigh wave. In their approach, the coupling with higher modes is neglected. They observe a systematic discrepancy between observed and modeled seismogram envelopes which also suggests that higher modes dominate the data at lapse times larger than 30,000 s.

Interferometry with Scattered Waves

Green Function Retrieval

As explained in section “Introduction”, the incoherent field refers to the waves that average out when a mean over an ensemble of random media is performed. These waves are incoherent in the sense that, at a given point of the medium, they propagate in space directions that differ from that of the mean field. But “incoherence” does not imply that the phase information has been lost (Campillo 2006). Campillo and Paul (2003) were the first to demonstrate the existence of correlations between the coda waves recorded at two distant points in the heterogeneous crust. Such correlations emerge after sufficient temporal and source averaging. Following Lobkis and Weaver (2001), Campillo and Paul (2003) proposed that the correlation tensor of the coda wavefields recorded at two points *A* and *B* is in fact proportional to the Green's tensor G_{ij} between these two points. In the frequency domain, such a relation can be mathematically formulated as follows:

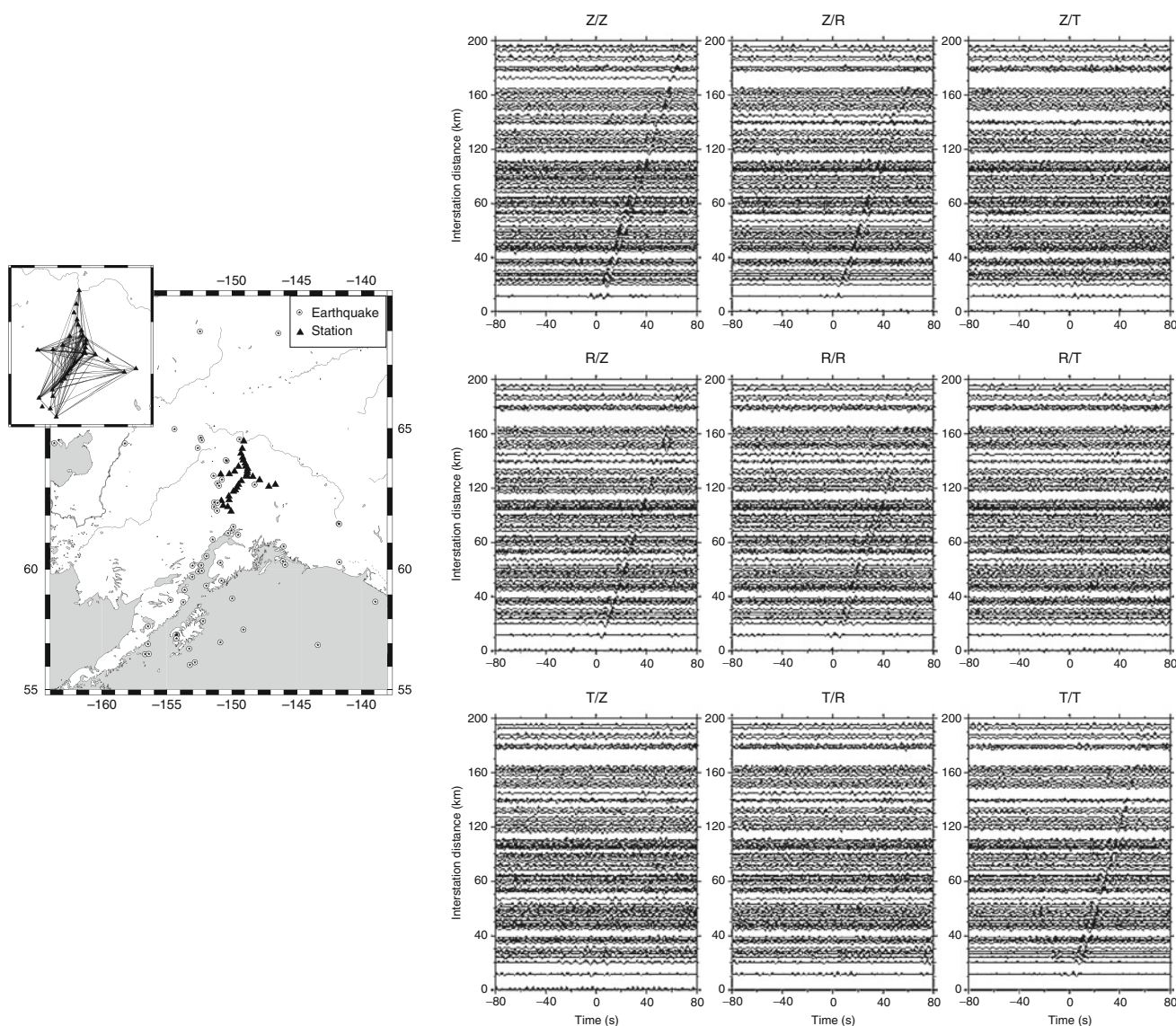
$$\langle u_i(B, \omega) u_j(A, \omega)^* \rangle \propto \text{Im} G_{ij}(B, A, \omega). \quad (13)$$

The proportionality factor depends on the details of the source spectrum. The occurrence of the imaginary part of the Green's function on the right-hand side of Eq. 13 physically means that the correlation of two wavefields is proportional to a combination of the retarded and advanced Green's functions. The appearance of the advanced Green's function is

most easily understood by noting the equivalence between correlations and time-reversal experiments. We will not discuss this analogy, but the interested reader can refer to Derode et al. (2003).

In seismology, the brackets in Eq. 13 usually denote an average over time windows in the coda and over earthquake sources. An example of emergence of the retarded and advanced Green's function in the correlation of coda waves recorded on a temporary network deployed in Alaska is illustrated in Fig. 3. The maximum of the correlation is always shown for positive times, but it is nevertheless possible to

distinguish some coherent arrivals at negative times. The reconstruction of the Green's function from coda waves is intimately related to the concept of equipartition, which is a fundamental consequence of multiple scattering. Loosely stated, equipartition stipulates that plane P , S , Rayleigh, and Love waves coming from all possible directions compose the wavefield. That the Green's function of the medium is recovered if the equipartition state is reached has been shown for an infinite elastic medium by Sánchez-Sesma and Campillo (2006). Based on the single-scattering approximation for scalar waves, Sato (2009) demonstrated that the correlation



Seismic Waves, Scattering, Fig. 3 Example of Green's function reconstruction from coda waves recorded in Alaska. A location map of the experiment is shown on the left. The inset shows all the station pairs analyzed in this study. On the right, all possible terms of the correlation tensor are represented as a function of the correlation time on the

horizontal axis and the epicentral distance between the two stations on the vertical axis. For each station pair, the three components of the seismometers have been rotated onto the radial (R), transverse (T), and vertical (Z) reference frame. (Reproduced from Paul et al. 2005)

of coda waves recorded at two stations enables the reconstruction of the ballistic Green's function between the two stations. Snieder (2004b) derived the reconstruction of the Green's function from coda waves based on the stationary phase approximation. An experimental verification of the equipartition principle has been performed by Hennino et al. (2001). In Fig. 3, we can observe a clear time asymmetry in the reconstructed Green's function. This seems to contradict the equipartition principle. Such a time asymmetry has been studied by Paul et al. (2005). These authors demonstrate that a flux of energy directed from the earthquake source to the station can persist even at very large lapse time in the coda. Such an energy flux breaks the temporal symmetry of the correlations unless the distribution of earthquakes around the station covers all azimuths.

Monitoring Temporal Variations

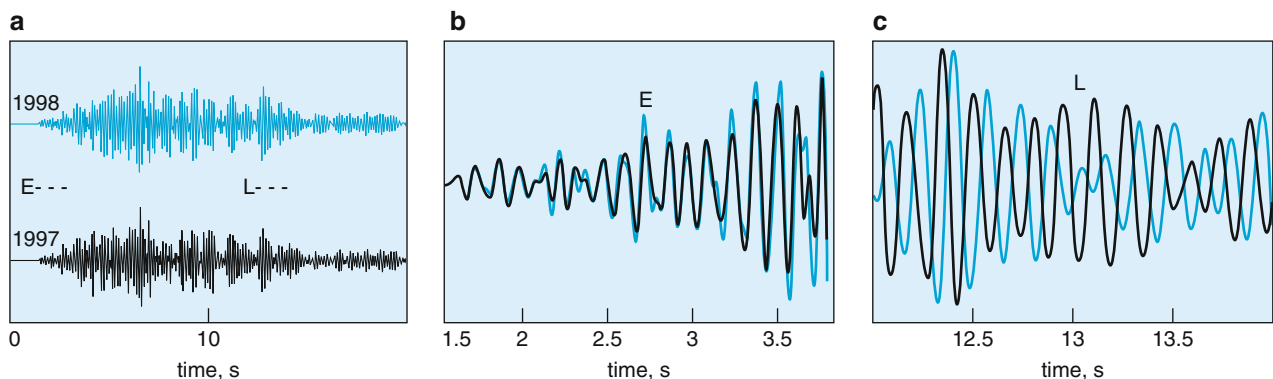
An interesting application of wave multiple scattering pertains to the detection of weak changes in dynamic media, i.e., media evolving with time. The methods which make use of scattered waves to probe changes in complex media have been termed “diffusing acoustic wave spectroscopy” in acoustics and “coda wave interferometry” (CWI) in seismology. In this article, we focus on applications to seismic waves. A theoretical approach of the method is provided by Snieder (2006). For a very accessible review of this broad topic in physics, we refer to Snieder and Page (2007). A comprehensive review of monitoring applications in seismology is given by Poupinet et al. (2008).

The key idea of CWI is to take advantage of the long propagation paths of scattered waves, which accumulate small phase differences on their way from source to station. An example of monitoring the medium changes at the Merapi volcano is shown in Fig. 4. In this experiment, a repeatable source is used to generate seismic waves that propagate

through the volcano. The fact that the source is highly reproducible is important, as any small difference between seismograms recorded at different time instants can be interpreted as a change of the propagating medium. Natural repeating sources such as earthquake doublets can also be exploited to probe medium changes. The doublets correspond to small earthquakes breaking the same fault patch at different times of occurrence (Poupinet et al. 1984). As shown in Fig. 4, the changes are virtually undetectable on the first arrivals. In the coda, the medium changes are manifested by a small time delay between two very similar waveforms. In the case of the Merapi data shown in Fig. 4, an analysis of the delay times as a function of the time lag in the coda reveals that the two seismograms are nearly stretched versions of one another. This means that the two waveforms can be related by a similarity transformation $t \rightarrow t/(1 + \epsilon)$ of the time variable. Such an observation can be interpreted as a very small change of the wave velocity δv in the propagating medium. The relation between the stretching parameter ϵ and the velocity variation reads (Sens-Schönfelder and Wegler 2006b):

$$\epsilon = -\frac{\delta v}{v} \quad (14)$$

In practice, the velocity change can be estimated by measuring the delay time which maximizes the correlation between the two signals. Such a procedure can be applied to moving time windows in the coda, of typical duration, a few central periods. The relation (14) is valid for acoustic waves only. A formula valid for elastic coda waves in the equipartition regime has been derived by Snieder (2002). Ratdomopurbo and Poupinet (1995), and Wegler et al. (2006b) found that the shear velocity increases prior to eruptions at the Merapi volcano, an observation interpreted as a



Seismic Waves, Scattering, Fig. 4 Illustration of the principle of coda wave interferometry. A controlled and repeatable source is used to probe the Merapi volcano with seismic waves. The signals recorded at a time interval of 1 year are shown by blue and black lines. The insets show the

comparisons of waveforms in the early part of the signal (E) and in the late part of the signal (L). In the later portion, a clear time shift between the two traces can be observed. (Reproduced from Snieder 2004a)

pressurization effect due to the magma ascent. Typically, relative velocity changes of the order of 10^{-5} – 10^{-4} can be detected (Poupinet et al. 2008). This makes coda wave interferometry the ideal tool to study the rheology of materials (Snieder et al. 2002).

As a final example of the use of scattered waves, we consider a method which has been termed “passive image interferometry” (Sens-Schönfelder and Wegler 2006b) which combines coda wave interferometry with the principle of Green’s function reconstruction from coda waves. Passive interferometry uses ambient noise – instead of coda waves – to reconstruct the Green’s function. The coda part of the reconstructed Green’s function is subsequently used to monitor temporal variations in the medium. A great advantage of this method as compared to the usual coda wave interferometry lies in the fact that noise records allow a continuous reconstruction of the Green’s function over very long time periods. Brenguier et al. (2008) demonstrated the possibility to detect very small medium changes at Le Piton de la Fournaise volcano in la Réunion Island. Applications to the study of the recovery phenomena in sediments affected by strong ground motion are developed in Sawazaki et al. (2009).

Summary

This article presents a nontechnical introduction to the various phenomena and methods pertaining to the scattering of seismic waves. The basic concepts of seismic scattering are introduced with the aid of a thought-experiment. The topics covered include the interpretation of transmission fluctuations, the broadening of envelope at regional distances, the interpretation of coda waves at local and global scales, and the use of scattered waves in seismic interferometry. The methods developed in seismic scattering offer powerful tools for the stochastic imaging of Earth’s structure at spatial scales that cannot be resolved by traditional tomographic methods. In addition, scattered seismic waves are extremely sensitive to slight temporal changes of the medium. Using the interferometric approach, multiply-scattered coda waves can be used to monitor the evolution of the dynamic Earth.

Bibliography

- Abubakirov I, Gusev A (1990) Estimation of scattering properties of lithosphere of Kamchatka based on Monte-Carlo simulation of record envelope of a near earthquake. *Phys Earth Planet Inter* 64:52–67
- Aki K (1969) Analysis of the seismic coda of local earthquakes as scattered waves. *J Geophys Res* 74:615–618
- Aki K (1973) Scattering of P waves under the Montana LASA. *J Geophys Res* 78:1334–1346
- Aki K, Chouet B (1975) Origin of coda waves, sources and attenuation. *J Geophys Res* 80:3322–3342
- Brenguier F, Shapiro NM, Campillo M, Ferrazzini V, Duputel Z, Coutant O, Nercessian A (2008) Toward forecasting volcanic eruptions using seismic noise. *Nat Geosci* 1:126–130
- Campillo M (2006) Phase and correlation in ‘random’ seismic fields and the reconstruction of the green function. *Pure Appl Geophys* 163:475–502
- Campillo M, Paul A (2003) Long range correlations in the diffuse seismic coda. *Science* 299:547–549
- Chen X, Aki K (1991) General coherence functions for amplitude and phase fluctuations in a randomly heterogeneous medium. *Geophys J Int* 105:155–162
- Cleary J, Haddon R (1972) Seismic wave scattering near the Core-mantle boundary: a new interpretation of precursors to PKP. *Nature* 240:549–551
- Del Pezzo E (2008) Seismic wave scattering in volcanoes. In: Sato H, Fehler M, Dmowska R (eds) *Earth heterogeneity and scattering effect on seismic waves. Advances in geophysics*, vol 50. Academic, New York, pp 353–369
- Derode A, Tourin A, Fink M (2001) Random multiple scattering of ultrasound. I. Coherent and ballistic waves. *Phys Rev E* 64:36–605
- Derode A, Larose E, Tanter M, de Rosny J, Tourin A, Campillo M, Fink M (2003) Recovering the Green’s function from field-field correlations in an open scattering medium. *J Acoust Soc Am* 113:2973–2976
- Dziewonski A, Anderson D (1981) Preliminary reference earth model. *Phys Earth Planet Inter* 25:297–356
- Earle P, Shearer P (1997) Observations of PKKP precursors used to estimate small-scale topography on the core-mantle boundary. *Science* 277:667–670
- Fehler M, Hoshiba M, Sato H, Obara K (1992) Separation of scattering and intrinsic attenuation for the Kanto-Tokai region, Japan, using measurements of S-wave energy versus hypocentral distance. *Geophys J Int* 108:787–800
- Flatté S, Wu R (1988) Small-scale structure in the lithosphere and asthenosphere deduced from arrival time and amplitude fluctuations at NOR-SAR. *J Geophys Res* 93:6601–6614
- Furumura T, Kennett B (2005) Subduction zone guided waves and the heterogeneity structure of the subducted plate: intensity anomalies in northern Japan. *J Geophys Res* 110:B10302
- Gusev A, Abubakirov I (1987) Monte-Carlo simulation of record envelope of a near earthquake. *Phys Earth Planet Inter* 49:30–36
- Gusev A, Abubakirov I (1996) Simulated envelopes of non-isotropically scattered body waves as compared to observed ones: another manifestation of fractal heterogeneity. *Geophys J Int* 127:49–60
- Gusev A, Abubakirov I (1999a) Vertical profile of effective turbidity reconstructed from broadening of incoherent body-wave pulses. I – General approach and the inversion procedure. *Geophys J Int* 136:295–308
- Gusev A, Abubakirov I (1999b) Vertical profile of effective turbidity reconstructed from broadening of incoherent body-wave pulses. II – Application to Kamchatka data. *Geophys J Int* 136:309–323
- Hedlin M, Shearer P, Earle P (1997) Seismic evidence for small-scale heterogeneity throughout the Earth’s mantle. *Nature* 387:145–150
- Hennino R, Trégourès N, Shapiro NM, Margerin L, Campillo M, van Tiggelen BA, Weaver RL (2001) Observation of equipartition of seismic waves. *Phys Rev Lett* 86:3447–3450
- Hoshiba M (1991) Simulation of multiple scattered coda wave excitation based on the energy conservation law. *Phys Earth Planet Inter* 67:123–136

- Hoshiba M (1993) Separation of scattering attenuation and intrinsic absorption in Japan using the multiple lapse time window analysis of full seismogram envelope. *J Geophys Res* 98: 15809–15824
- Hoshiba M (1995) Estimation of nonisotropic scattering in western Japan using coda wave envelopes: application of a multiple non-isotropic scattering model. *J Geophys Res* 100:645–657
- Hoshiba M (1997) Seismic coda wave envelope in depth-dependent S wave velocity structure. *Phys Earth Planet Inter* 104:15–22
- King D, Haddon R, Husebye E (1975) Precursors to PP. *Phys Earth Planet Inter* 10:103–127
- Klimeš L (2002) Correlation functions of random media. *Pure Appl Geophys* 159:1811–1831
- Koper K, Franks J, Dombrovskaya M (2004) Evidence for small-scale heterogeneity in Earth's inner core from a global study of PKiKP coda waves. *Earth Planet Sci Lett* 228:227–241
- Korn M (1990) A modified energy flux model for lithospheric scattering of teleseismic body waves. *Geophys J Int* 102:165–175
- Kravtsov Y, Müller T, Shapiro S, Buske S (2003) Statistical properties of reflection traveltimes in 3-D randomly inhomogeneous and anisotropic media. *Geophys J Int* 154:841–851
- Kubanza M, Nishimura T, Sato H (2007) Evaluation of strength of heterogeneity in the lithosphere from peak amplitude analyses of teleseismic short-period vector P waves. *Geophys J Int* 171: 390–398
- Leyton F, Koper K (2007) Using PKiKP coda to determine inner core structure: 2. Determination of QC. *J Geophys Res* 112:B05317
- Lobkis OI, Weaver RL (2001) On the emergence of the Green's function in the correlations of a diffuse field. *J Acoust Soc Am* 110:3011–3017
- Maeda T, Sato H, Ohtake M (2006) Constituents of vertical-component coda waves at long periods. *Pure Appl Geophys* 163:549–566
- Maeda T, Sato H, Nishimura T (2007) Synthesis of coda wave envelopes in randomly inhomogeneous elastic media in a half-space: single scattering model including Rayleigh waves. *Geophys J Int* 172:130–154
- Margerin L (2005) Introduction to radiative transfer of seismic waves. In: Nolet G, Levander A (eds) *Seismic earth: analysis of broadband seismograms*. Geophysical monograph, vol 157. American Geophysical Union, Washington, pp 229–252
- Margerin L (2006) Attenuation, transport and diffusion of scalar waves in textured random media. *Tectonophysics* 416:229–244
- Margerin L, Nolet G (2003a) Multiple scattering of high-frequency seismic waves in the deep earth: modeling and numerical examples. *J Geophys Res* 108:2234
- Margerin L, Nolet G (2003b) Multiple scattering of high-frequency seismic waves in the deep earth: PKP precursor analysis and inversion for mantle granularity. *J Geophys Res* 108:2514
- Margerin L, Campillo M, van Tiggelen BA (1998) Radiative transfer and diffusion of waves in a layered medium: new insight into coda Q. *Geophys J Int* 134:596–612
- Margerin L, Campillo M, Shapiro N, van Tiggelen BA (1999) Residence time of diffuse waves in the crust as a physical interpretation of coda Q: application to seismograms recorded in Mexico. *Geophys J Int* 138:343–352
- Margerin L, Campillo M, van Tiggelen BA (2000) Monte Carlo simulation of multiple scattering of elastic waves. *J Geophys Res* 105:7873–7892
- Nakahara H, Nishimura T, Sato H, Ohtake M (1998) Seismogram envelope inversion for the spatial distribution of high-frequency energy radiation from the earthquake fault: application to the 1994 far east off Sanriku earthquake, Japan. *J Geophys Res* 103:855–867
- Obara K, Sato H (1995) Regional differences of random inhomogeneities around the volcanic front in the Kanto-Tokai area, Japan, revealed from the broadening of S wave seismogram envelopes. *J Geophys Res* 100:2103–2121
- Papanicolaou G, Bal G, Ryzhik L (2000) Probabilistic theory of transport processes with polarization. *SIAM J Appl Math* 60:1639–1666
- Paul A, Campillo M, Margerin L, Larose E, Derode A (2005) Empirical synthesis of time-asymmetrical green functions from the correlation of coda waves. *J Geophys Res* 110:B08302
- Poupinet G, Kennett B (2004) On the observation of high frequency PKiKP and its coda in Australia. *Phys Earth Planet Inter* 146:497–511
- Poupinet G, Ellsworth W, Frechet J (1984) Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras fault, California. *J Geophys Res* 89:5719–5731
- Poupinet G, Got J-L, Brenguier F (2008) Monitoring temporal variations of physical properties in the crust by cross-correlating the waveforms of seismic doublets. In: Sato H, Fehler M, Dmowska R (eds) *Advances in geophysics: earth heterogeneity and scattering effect on seismic waves*, vol 50. Academic, New York, pp 374–399
- Przybilla J, Korn M (2008) Monte Carlo simulation of radiative energy transfer in continuous elastic random media-three-component envelopes and numerical validation. *Geophys J Int* 173:566–576
- Ratdomopurbo A, Poupinet G (1995) Monitoring a temporal change of seismic velocity in a volcano: application to the 1992 eruption of Mt. Merapi (Indonesia). *Geophys Res Lett* 22:775–778
- Rytov SM, Kravtsov YA, Tatarskii VI (1989a) Principles of statistical radiophysics. Wave propagation through random media, vol 4. Springer, New York
- Rytov SM, Kravtsov YA, Tatarskii VI (1989b) Principles of statistical radiophysics. Elements of random fields, vol 1. Springer, New York
- Ryzhik L, Papanicolaou G, Keller JB (1996) Transport equation for elastic and other waves in random media. *Wave Motion* 24:327–370
- Saito T, Sato H, Ohtake M (2002) Envelope broadening of spherically outgoing waves in three-dimensional random media having power law spectra. *J Geophys Res* 107:2089
- Saito T, Sato H, Ohtake M, Obara K (2005) Unified explanation of envelope broadening and maximum-amplitude decay of high-frequency seismograms based on the envelope simulation using the Markov approximation: forearc side of the volcanic front in north-eastern Honshu, Japan. *J Geophys Res* 110:B01304
- Sánchez-Sesma FJ, Campillo M (2006) Retrieval of the green function from cross correlation: the canonical elastic problem. *Bull Seismol Soc Am* 96:1182–1191
- Sato H (1982) Amplitude attenuation of impulsive waves in random media based on travel time corrected mean wave formalism. *J Acoust Soc Am* 71:559–564
- Sato H (1984) Attenuation and envelope formation of three-component seismograms of small local earthquakes in randomly inhomogeneous lithosphere. *J Geophys Res* 89:1221–1241
- Sato H (1989) Broadening of seismogram envelopes in the randomly inhomogeneous lithosphere based on the parabolic approximation: southeastern Honshu, Japan. *J Geophys Res* 94:17735–17747
- Sato H (1994) Multiple isotropic scattering model including P-S conversion for the seismogram envelope formation. *Geophys J Int* 117:487–494
- Sato H (2006) Synthesis of vector wave envelopes in three-dimensional random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: plane wave case. *J Geophys Res* 111:B06306
- Sato H (2007) Synthesis of vector wave envelopes in three-dimensional random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: spherical wave case. *J Geophys Res* 112:B01301
- Sato H (2008) Synthesis of vector-wave envelopes in 3-D random media characterized by a nonisotropic Gaussian ACF based on the Markov approximation. *J Geophys Res* 113:B08304

- Sato H (2009) Green's function retrieval from the CCF of coda waves in a scattering medium. *Geophys J Int* 179:1580–1583
- Sato H, Fehler M (1998) *Seismic wave propagation and scattering in the heterogeneous earth*. Springer, New York
- Sato H, Korn M (2009) Synthesis of vector-wave envelopes in random elastic media on the basis of the markov approximation. In: Sato H, Fehler M, Dmowska R (eds) *Earth heterogeneity and scattering effect on seismic waves*. *Advances in geophysics*, vol 50. Academic, New York, pp 43–93
- Sato H, Nishino M (2002) Multiple isotropic-scattering model on the spherical earth for the synthesis of Rayleigh-wave envelopes. *J Geophys Res* 107:2343
- Sato H, Fehler M, Dmowska R (eds) (2008) *Earth heterogeneity and scattering effects on seismic waves*. *Advances in geophysics*, vol 50. Academic, Amsterdam
- Sawazaki K, Sato H, Nakahara H, Nishimura T (2009) Time-lapse changes of seismic velocity in the shallow ground caused by strong ground motion shock of the 2000 Western-Tottori earthquake, Japan, as revealed from coda deconvolution analysis. *Bull Seismol Soc Am* 99:352–366
- Scherbaum F, Sato H (1991) Inversion of full seismogram envelopes based on the parabolic approximation: estimation of randomness and attenuation in Southeast Honshu, Japan. *J Geophys Res* 96:2223–2232
- Sens-Schönfelder C, Wegler U (2006a) Radiative transfer theory for estimation of the seismic moment. *Geophys J Int* 167:1363–1372
- Sens-Schönfelder C, Wegler UC (2006b) Passive image interferometry and seasonal variations of seismic velocities at merapi volcano, Indonesia. *Geophys Res Lett* 33:L21302
- Shapiro SA, Kneib G (1993) Seismic attenuation by scattering: theory and numerical results. *Geophys J Int* 114:373–391
- Shearer P, Earle P (2004) The global short-period wavefield modelled with a Monte Carlo seismic phonon method. *Geophys J Int* 158:1103–1117
- Shearer P, Earle P (2008) Observing and modeling elastic scattering in the deep earth. In: Sato H, Fehler M, Dmowska R (eds) *Earth heterogeneity and scattering effect on seismic waves*. *Advances in geophysics*, vol 50. Academic, New York, pp 167–193
- Sheng P (1995) *Introduction to wave scattering, localization and mesoscopic phenomena*. Academic, San Diego
- Shiomi K, Sato H, Ohtake M (1997) Broad-band power-law spectra of well-log data in Japan. *Geophys J Int* 130:57–64
- Snieder R (2002) Coda wave interferometry and the equilibration of energy in elastic media. *Phys Rev E* 66:46615
- Snieder R (2004a) Coda wave interferometry. In: McGraw-Hill yearbook of science & technology. McGraw-Hill, New York, pp 54–56
- Snieder R (2004b) Extracting the greens function from the correlation of coda waves: a derivation based on stationary phase. *Phys Rev E* 69:46610
- Snieder R (2006) The theory of coda wave interferometry. *Pure Appl Geophys* 163:455–473
- Snieder R, Page J (2007) Multiple scattering in evolving media. *Phys Today* 60:49–55
- Snieder R, Grêt A, Douma H, Scales J (2002) Coda wave interferometry for estimating nonlinear behavior in seismic velocity. *Science* 295:2253–2255
- Takahashi T, Sato H, Nishimura T, Obara K (2007) Strong inhomogeneity beneath quaternary volcanoes revealed from the peak delay analysis of S-wave seismograms of microearthquakes in northeastern Japan. *Geophys J Int* 168:90–99
- Takahashi T, Sato H, Nishimura T, Obara K (2009) Tomographic inversion of the peak delay times to reveal random velocity fluctuations in the lithosphere: method and application to northeastern Japan. *Geophys J Int* 178:1437–1455
- Vidale J, Earle P (2000) Fine-scale heterogeneity in the Earth's inner core. *Nature* 404:273–275
- Wegler U (2004) Diffusion of seismic waves in a thick layer: theory and application to Vesuvius volcano. *J Geophys Res* 109:07303
- Wegler U, Lühr B (2001) Scattering behaviour at Merapi volcano (Java) revealed from an active seismic experiment. *Geophys J Int* 145:579–592
- Wegler U, Korn M, Przybilla J (2006a) Modeling full seismogram envelopes using radiative transfer theory with born scattering coefficients. *Pure Appl Geophys* 163:503–531
- Wegler U, Lühr B, Snieder R, Ratdomopurbo A (2006b) Increase of shear wave velocity before the 1998 eruption of Merapi volcano (Indonesia). *Geophys Res Lett* 33:L09303
- Wu RS (1982) Attenuation of short period seismic waves due to scattering. *Geophys Res Lett* 9:9–12
- Wu RS (1985) Multiple scattering and energy transfer of seismic waves-separation of scattering effect from intrinsic attenuation I. Theoretical modelling. *Geophys J Int* 82:57–80
- Wu RS (1989) Seismic wave scattering. In: James DE (ed) *The encyclopedia of solid earth geophysics*. Van Nostrand Reinhold, New York, pp 1166–1187
- Wu RS, Aki K (1988) Multiple scattering and energy transfer of seismic waves-separation of scattering effect from intrinsic attenuation II. Application of the theory to Hindu Kush region. *Pure Appl Geophys* 128:49–80
- Wu R, Flatté S (1990) Transmission fluctuations across an array and heterogeneities in the crust and upper mantle. *Pure Appl Geophys* 132:175–196
- Wu R, Xu Z, Li X (1994) Heterogeneity spectrum and scale-anisotropy in the upper crust revealed by the German continental deep-drilling (KTB) holes. *Geophys Res Lett* 21:911–914
- Yoshimoto K (2000) Monte Carlo simulation of seismogram envelopes in scattering media. *J Geophys Res* 105:6153–6161
- Zeng Y (1993) Theory of scattered P-and S-wave energy in a random isotropic scattering medium. *Bull Seismol Soc Am* 83:1264–1276
- Zheng Y, Wu R (2005) Measurement of phase fluctuations for transmitted waves in random media. *Geophys Res Lett* 32:L14314
- Zheng Y, Wu R (2008) Theory of transmission fluctuations in random media with a depth-dependent background velocity structure. In: Sato H, Fehler M, Dmowska R (eds) *Earth heterogeneity and scattering effect on seismic waves*. *Advances in geophysics*, vol 50. Academic, New York, pp 21–41
- Zheng Y, Wu R, Lay T (2007) Inverting the power spectrum for a heterogeneous medium. *Geophys J Int* 168:1005–1010

Seismic Zonation

Yanxiang Yu, Mengtan Gao and Guangyin Xu
Institute of Geophysics, China Earthquake Administration,
Beijing, China

Definition

Seismic zonation. The process of subdividing the territory into regions with respect to the level of seismic hazard. The result of seismic zonation is usually presented as a map, which is based on seismic hazard map.

Introduction

Seismic zonation is useful for hazard reduction such as earthquake-resistant design of structures, risk analysis, land-use planning, etc. Many earthquake-prone countries developed seismic zonation maps. Seismic zonation map is usually revised or updated periodically with the progress in methodology and accumulation of new data. Seismic intensity (see ► [“Earthquakes, Intensity”](#)) or ground motion parameters such as peak ground acceleration (PGA), peak ground velocity (PGV), and spectral acceleration at specific natural period are mostly adopted in seismic zonation map. In the early stage, most seismic zonation maps were in terms of intensity, but since the 1980s, ground motion parameters have become popular. More commonly, such maps take the exceeding probability of 10% within 50 years (return period 475 years) as standard.

Methodology

The basic method to develop seismic zonation map is the seismic hazard analysis (see ► [“Seismic Hazard”](#)) approach. Both deterministic and probabilistic approaches are adopted. Although some new deterministic approaches based on the computation of synthetic seismograms are used in recent years in developing seismic zonation maps (Parvez et al. 2003), nevertheless probabilistic approaches are more popular. The seismic zonation maps, based on seismic hazard maps (see ► [“Seismology, Global Earthquake Model”](#)), of many countries are developed by the use of probabilistic approach. The well-known USGS National Seismic Hazard Maps (Algermissen and Perkins 1976; Algermissen et al. 1990; Frankel et al. 1996, 2002; Petersen et al. 2008) and the Global Seismic Hazard Assessment Programme (GSHAP) are the typical (see ► [“Seismic Hazard”](#)) ones. In this article, only probabilistic method is briefly introduced.

The probabilistic method (usually referred to as PSHA – probabilistic seismic hazard analysis) was first introduced by Cornell (1968) and since then widely adopted and modified (McGuire 1978; Bender and Perkins 1982). There are four basic steps for assessment of PSHA (Fig. 1):

- Step 1: Definition of seismic sources. Sources may range from small faults to large seismotectonic provinces with uniform seismicity. The type of seismic source can be both line or area sources.
- Step 2: Definition of seismicity recurrence characteristic for the sources, where each source is described by an earthquake probability distribution, or recurrence relationship. A recurrence relationship indicates the chance of an earthquake of a given magnitude to occur anywhere inside the

source during a specified period of time. An upper bound earthquake is chosen for each source, which represents the maximum event to be considered.

- Step 3: Development of ground motion attenuation relationships. This is usually done empirically from strong motion records (see ► [“Earthquakes, Strong-Ground Motion”](#)).

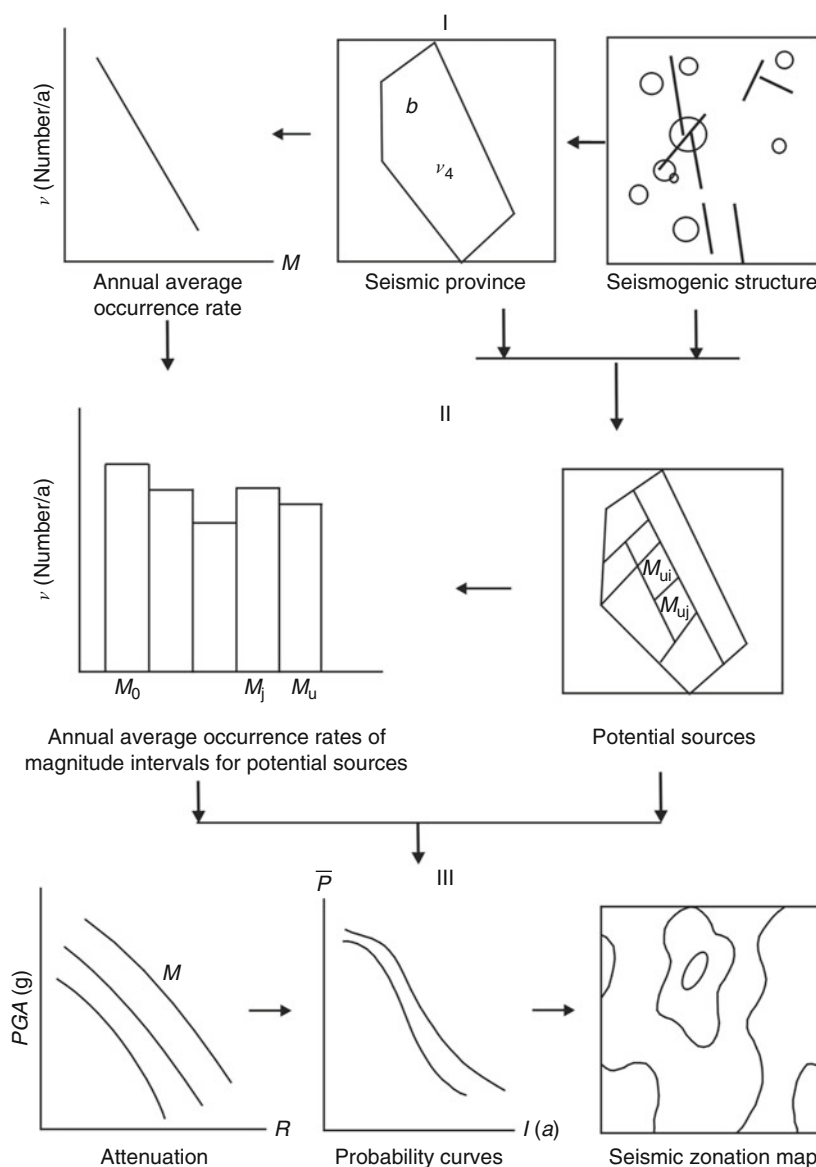
- Step 4: Determination of the hazard at each site. In this case, the effects of all the earthquakes of different sizes occurring at different locations in different earthquake sources at different probabilities of occurrence are integrated into one curve that shows the probability of exceeding different levels of ground motion level (such as PGA) at the site during a specified period of time.

There are two basic assumptions in the seismic hazard analysis method at present: (a) Seismicity in the region around the site in the past indicates that in future, the recurrence rate of given site is the same as that of historic period. (b) Seismicity of the region can be expressed by tectonic earthquakes in the region, i.e., the seismic activities distribute homogeneously in a certain tectonic area or an active fault.

The two assumptions accord with the two principles of historic earthquake repeatness and tectonic extrapolation. The difference is that the PSHA adds a new concept of magnitude interval recurrent rate and the hazard is evaluated with probabilistic analysis method. The method can provide the exceedence probability of different ground motion extent (intensity, acceleration, etc.) at the site in specific time intervals, so that the earthquake resistant parameters can be selected with different exposure period, risk level, and various engineering structures.

Example: Seismic Zonation Map of China (2001)

The first seismic zonation map of China was compiled by Wong (1921) after the Haiyuan earthquake, which occurred in 1920. After that, three versions of seismic zonation map were developed in 1957, 1977, and 1990 (Shi et al. 1992). The 1957 version demonstrated the maximum affected intensity of China. The 1977 version was provided by using the methodology of long-term and middle-term earthquake prediction. This version of seismic zonation map demonstrated the maximum encountered intensity in the forthcoming 100 years. This map was adopted by the building code. The 1990 version used probabilistic method. The seismic intensity with exceeding probability of 10% within 50 years was given in the map. It was used in the building code and other regulations related with seismic design. In 2001, a new seismic zonation map of China was issued. This map also used probabilistic method.

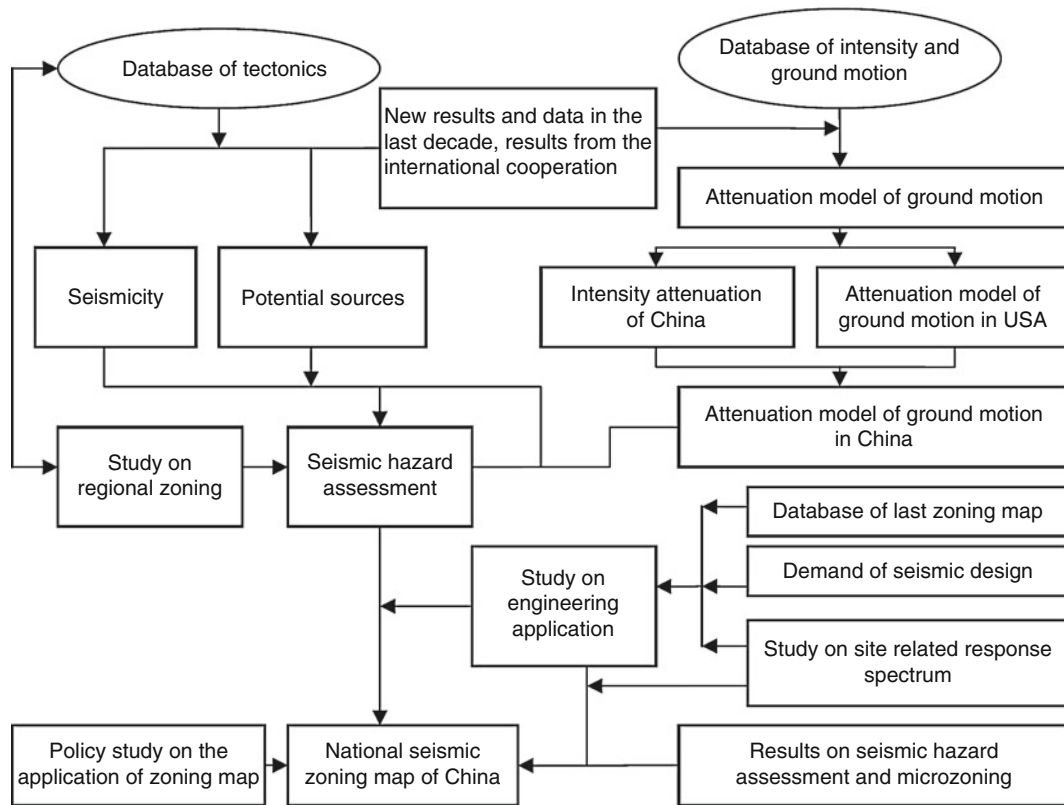
Seismic Zonation,**Fig. 1** Approaches of probabilistic seismic hazard analysis

The PGA and characteristic period of response spectrum T_g with exceeding probability of 10% within 50 years were given in the map. The probabilistic method used in developing seismic zonation of China (2001) was a little bit different from PSHA.

Most earthquakes in China are intraplate earthquakes (see ► [“Seismicity, Intraplate”](#)). Their non-homogeneity in space, non-stationarity in time were shown in the historical earthquake catalogs. The PSHA method applied in China can reveal non-homogeneity in space, non-stationarity in time. The approaches applied differ from PSHA in two aspects:

1. Evaluating seismicity and determining magnitude distribution, total annual occurrence rate of the province are on the basis of the seismic tendency estimation and the seismic characteristics analysis.
2. Annual rates of sources in all magnitude intervals are determined by spatial distribution functions, which describe relative risk among sources in the province.

The technical approach for developing the national seismic zonation map of China (2001) is shown in Fig. 2 (Gao 2003).



Seismic Zonation, Fig. 2 Technical approach of developing national seismic zonation map of China (2001)

The key scientific problems in compiling this map are: (a) how to treat the uncertainties in the evaluation of seismicity parameters and the delineation of potential sources as well as the attenuation relationship; (b) how to select the suitable parameters in the zonation map to fit the need of seismic design; and (c) how to use the domestic data and the data from the world to get the attenuation of the ground motion parameters.

The logic tree method is adopted to treat the uncertainties in seismic hazard analysis. The multi-set of potential source delineations was used. Four groups of scientists participated in the work of potential sources delineations based on the independent background and database. There are four sets of potential sources adopted in the logic tree analysis. Figure 3 is one of the potential source set.

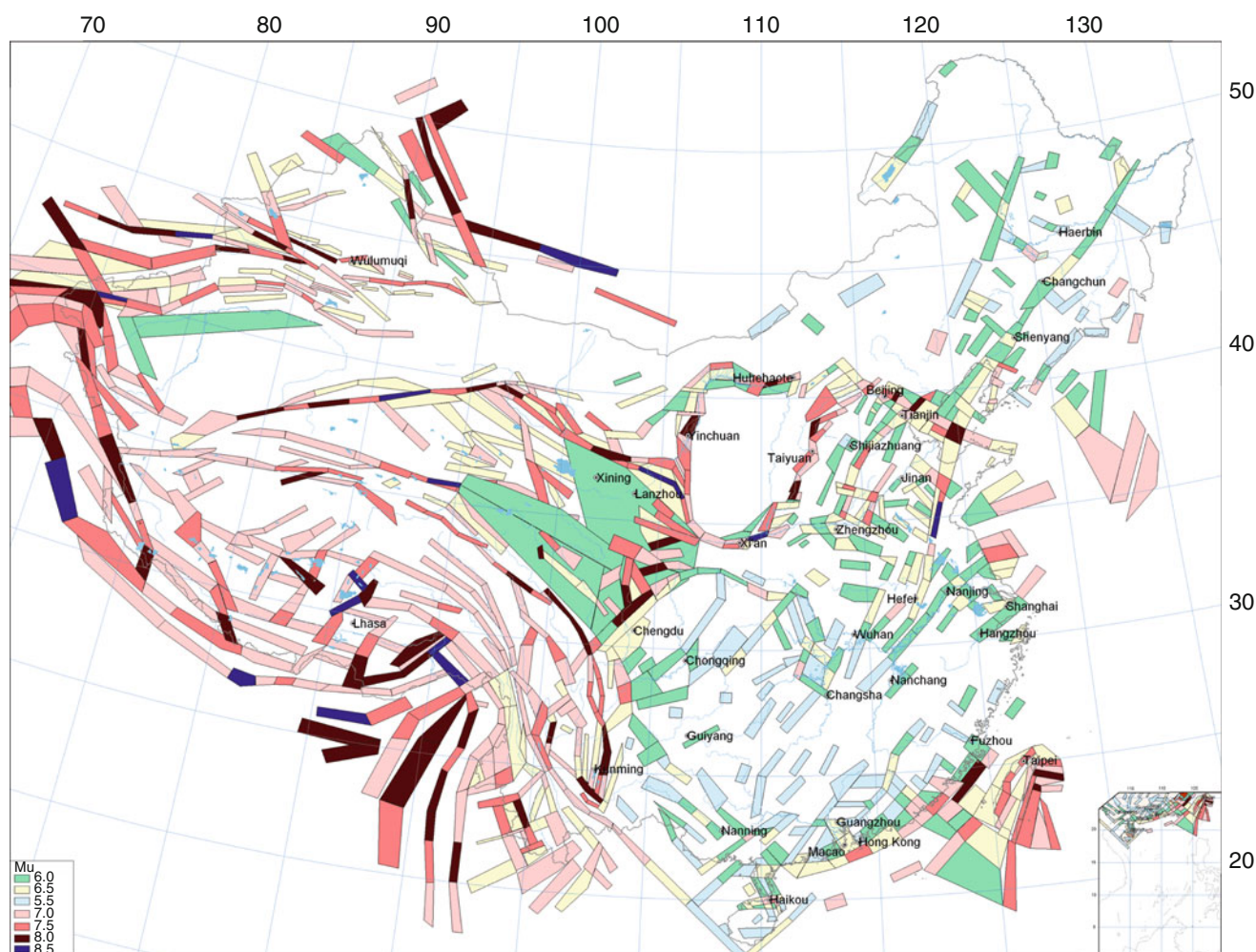
Site-related seismic response spectrum is the base of seismic design for ordinary industry and civil construction. The site-related response spectrum is not only closely related to the earthquake environment, but also related to the soil condition. It is very difficult to decide site-related response spectrum by a single parameter such as intensity or peak

acceleration. According to the results from seismic hazard assessment and the building code, the basic parameters used in the seismic zoning map to provide site-related response spectrum are PGA and the characteristic period of the response spectrum.

The attenuation relationships for the acceleration response spectrum platform value S_a and the velocity response spectrum platform value S_v were developed by modifying corresponding attenuation relations in the western United States according to the differences of intensity attenuation relations. Then the PGA and the characteristic period of response spectrum of acceleration T_g were defined as:

$$\begin{aligned} PGA &= \frac{S_a}{2.5} \\ T_g &= 2\pi \frac{S_v}{S_a} \end{aligned} \quad (1)$$

The country was divided into 40,000 grids. The probabilistic seismic hazard analysis for every grid was performed. The ground motion parameters with exceeding probability



Seismic Zonation, Fig. 3 One of the seismic potential source set used in developing national seismic zonation of China (2001)

10% within 50 years (return period 475 years) were determined.

The new zoning map includes two specific maps. One is PGA in gravity unit g (Fig. 4) and the other is T_g in second (Fig. 5). The scale of the maps are 1:4,000,000. The soil condition is medium hard soil. From these two parameters, the design response spectrum can be determined easily by (Eq. 2):

$$Sa(T) = 2.5a \begin{cases} 6T + 0.4 & 0 \leq T < 0.1 \\ 1 & 0.1 \leq T < T_g \\ T_g/T & T \geq T_g \end{cases} \quad (2)$$

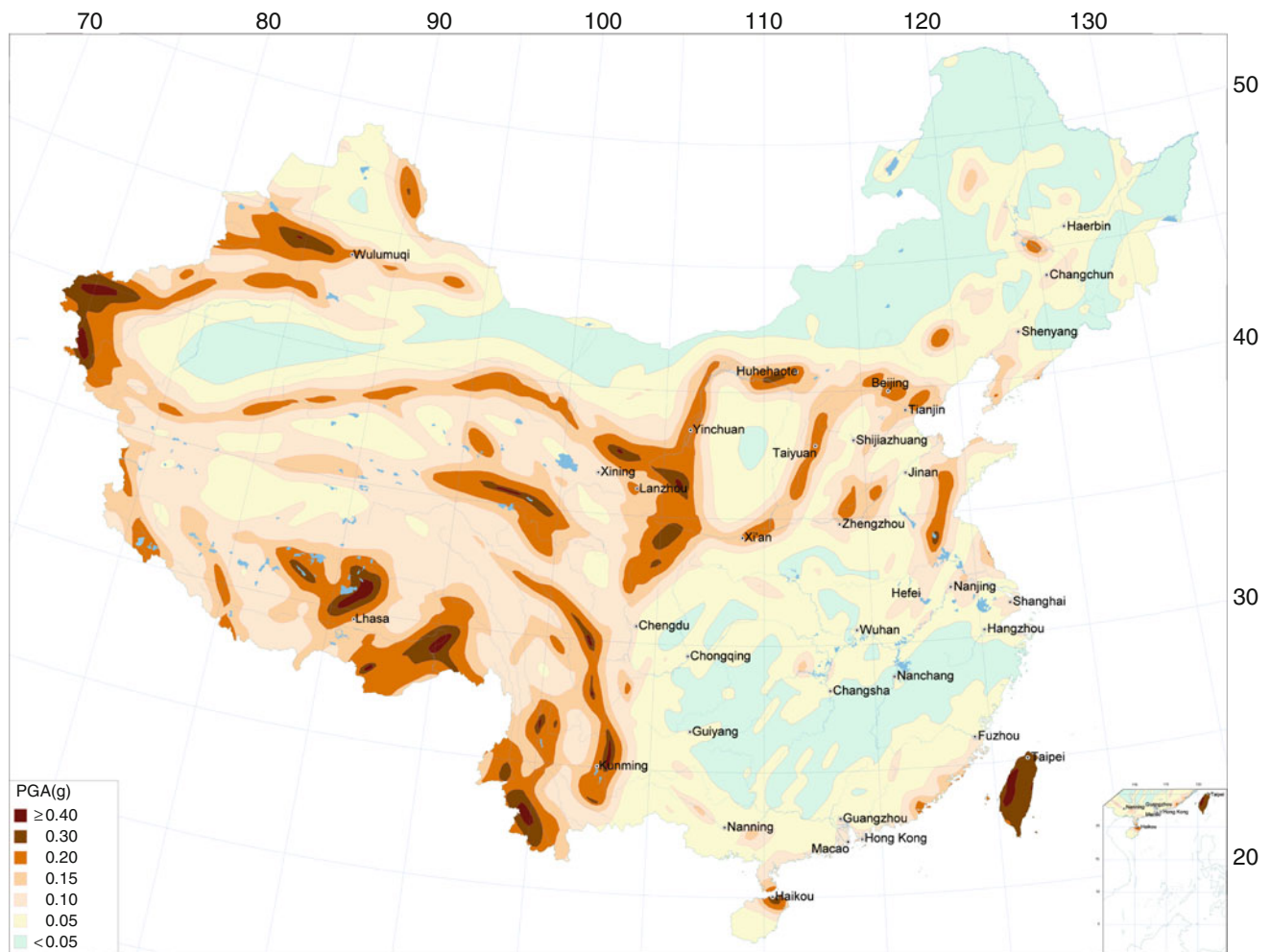
where a is PGA in g , T is the natural period in second.

In the acceleration map (Fig. 4), the territory is divided into seven zones. The acceleration for the seven zones are 0.05, 0.05, 0.10, 0.15, 0.20, 0.30, and ≥ 0.40 g , respectively.

In the characteristic period zoning map (Fig. 5), the territory of China is divided into three zones. Zone 1 is the region with $T_g = 0.35$ s, zone 2 with $T_g = 0.40$ s, and zone 3 with $T_g = 0.45$ s.

For different soil conditions, the T_g value should be adjusted as Table 1, but the PGA value does not change with soil conditions.

The new national seismic zonation map of China serves as the obligatory state standard, which took effect in August 1, 2001. For the ordinary newly constructed buildings, the standard must be followed.



Seismic Zonation, Fig. 4 Illustration of acceleration zonation map of China

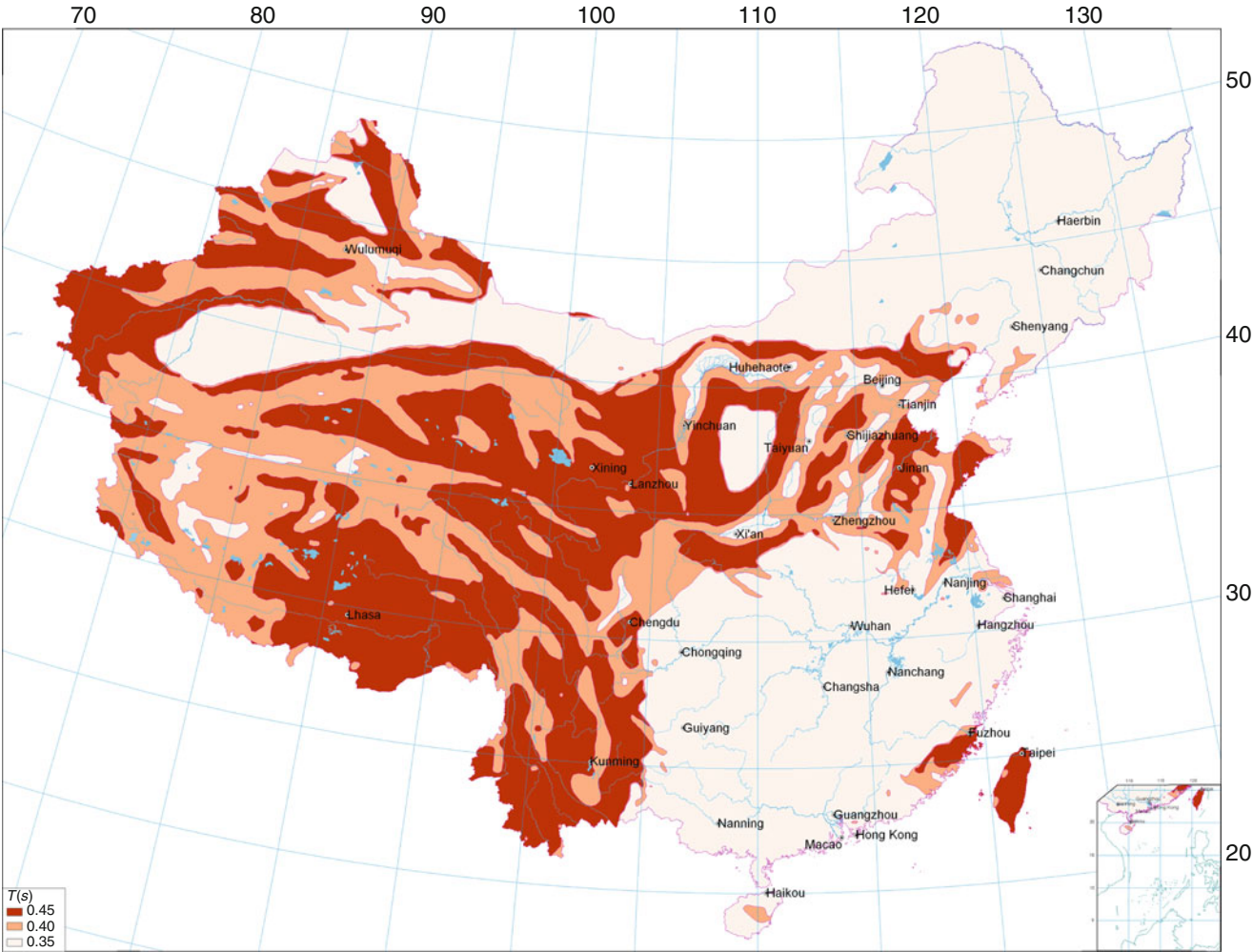
Summary

Seismic zonation is a process of estimation of the seismic hazard in terms of parameters of ground motion for a certain area. Assessment results in seismic zonation map compilation, which reflects territorial distribution of the seismic hazard (see ► [“Seismic Hazard”](#)). Seismic zonation map is useful for hazard reduction such as earthquake-resistant design of structures, risk analysis, land-use planning, etc. Many countries apply seismic hazard maps in anti-seismic codes. Recently, the probabilistic seismic hazard analysis method is more commonly used in compiling seismic zonation map. The seismic zonation maps take the exceeding probability of 10% within 50 years (return period 475 years)

as standard. Seismic zonation will develop with the development of seismic hazard assessment methodology and anti-seismic policy.

Cross-References

- [Earthquakes, Intensity](#)
- [Earthquakes, Strong-Ground Motion](#)
- [Seismic Hazard](#)
- [Seismic Microzonation](#)
- [Seismicity, Intraplate](#)
- [Seismology, Global Earthquake Model](#)



Seismic Zonation, Fig. 5 Illustration of characteristic period of seismic response spectrum zonation map of China

Seismic Zonation, Table 1 T_g value for different soil conditions (unit: second)

Zone	Soil conditions			
	Hard	Medium hard	Medium soft	Soft
1	0.25	0.35	0.45	0.65
2	0.30	0.40	0.55	0.75
3	0.35	0.45	0.65	0.90

Bibliography

Algermissen ST, Perkins DM (1976) A probabilistic estimate of the maximum acceleration in rock in the contiguous United States: U.S. Geological Survey open-file report 76-416, 45 pp, 2 pls., scale 1:7,500,000

Algermissen ST, Perkins DM, Thenhaus PC, Hanson SL, Bender BL (1990) Probabilistic earthquake acceleration and velocity maps for the United States and Puerto Rico: U.S. Geological Survey Miscellaneous Field Studies Map MF-2120, 2 sheets, scale 1:7,500,000

Bender BK, Perkins DM (1982) SEISRISK II: a computer program for seismic hazard estimation. U.S. Geological Survey open-file report, 82-293

Cornell CA (1968) Engineering seismic risk analysis. *Bull Seismol Soc Am* 58:1583-1606

Frankel AD, Mueller C, Barnhard T, Perkins D, Leyendecker E, Dickman N, Hanson S, Hopper M (1996) National Seismic Hazard Maps – Documentation June 1996: U.S. Geological Survey open-file report, 96-532

Frankel AD, Petersen MD, Mueller CS, Haller KM, Wheeler RL, Leyendecker EV, Wesson RL, Harmsen SC, Cramer CH, Perkins DM, Rukstales KS (2002) Documentation for the 2002 update of the National Seismic Hazard Maps: U.S. Geological Survey open-file report 2002-420

Gao M (2003) New national seismic zoning map of China. *Acta Seismol Sin* 16:639-645

McGuire RK (1978) FRISK: computer program for seismic risk analysis using faults as earthquake sources. U.S. Geological Survey open file-report 78-1007

Parvez IA, Vaccari F, Panza GF (2003) A deterministic seismic hazard map of India and adjacent areas. *Geophys J Int* 155:489-508

Petersen MD, Cramer CH, Reichle MS, Frankel AD, Hanks TC (2008) Discrepancy between earthquake rates implied by historic earthquakes and a consensus geologic source model for California. *Bull Seismol Soc Am* 90:1117-1132

Shi Z, Yan J, Gao M (1992) Research on the principle and methodology of seismic zonation – results of the trials in North China. *Acta Seismol Sin* 5:305-314

Wong WH (1921) Zhuizhi collections. The Commercial Press, Beijing. (in Chinese)

Seismic, Ambient Noise Correlation

Michel Campillo¹, Philippe Roux¹ and Nikolai M. Shapiro²

¹Observatoire de Grenoble, Université Joseph Fourier and CNRS, Grenoble, France

²Institut de Physique du Globe de Paris, Paris, France

Definition

Seismic noise: permanent motion of the Earth surface that is not related to earthquakes or specific controlled sources.

Introduction

Traditional observational methods in seismology are based on earthquake records. It results in two main shortcomings. First, most techniques are based on waves emitted by earthquakes that occurred only in geologically active areas, mainly plate boundaries. This results in a limited resolution in all other areas where earthquakes are not present. Second, the repetition of earthquakes is rare, preventing the study of continuous changes within active structures such as volcanoes or faults.

Also at smaller scales in the context of geophysics prospecting, the resolution is limited by the number and power of sources, making it difficult to image large areas and/or deep structures. Similarly, reproducible sources are necessary for time-lapse monitoring leading to long-duration surveys that are difficult to achieve.

Nowadays, the seismic networks are producing continuous recordings of the ground motion. These huge amounts of data consist mostly of so called seismic noise, a permanent vibration of the Earth due to natural or industrial sources. Passive seismic tomography is based on the extraction of the coherent contribution to the seismic field from the cross-correlation of seismic noise between station pairs.

As described in many studies where noise has been used to obtain the Green's function between receivers, coherent waves are extracted from noise signals even if, at first sight, this coherent signal appears deeply buried in the local incoherent seismic noise. Recent studies on passive seismic processing have focused on two applications, the noise-extracted Green's functions associated to surface waves leads to sub-surface imaging on scales ranging from thousands of kilometers to very short distances; on the other hand, even when the Green's function is not satisfactorily reconstructed from seismic ambient noise, it has been shown that seismic monitoring is feasible using the scattered waves of the noise-correlation function.

Theoretical Basis for the Interpretation of Noise Records at Two Stations

Passive seismology is an alternative way of probing the Earth's interior using noise records only. The main idea is to consider seismic noise as a wave field produced by randomly and homogeneously distributed sources when averaged over long time series. In this particular case, cross-correlation between two stations yields the Green's function between these two points. In the case of a uniform spatial distribution of noise sources, the crosscorrelation of noise records converges to the complete Green's function of the medium, including all reflection, scattering, and propagation modes. However, in the case of the Earth, most of ambient seismic noise is generated by atmospheric and oceanic forcing at the surface. Therefore, the surface wave part of the Green's function is most easily extracted from the noise cross-correlations. Note that the surface waves are the largest contribution of the Earth response between two points at the surface.

Historically speaking, helioseismology was the first field where ambient-noise cross-correlation performed from recordings of the Sun's surface random motion was used to retrieve time-distance information on the solar surface. More recently, a seminal paper was published by Weaver and Lobkis (2001) that showed how, at the laboratory scale, diffuse thermal noise recorded and cross-correlated at two transducers fastened to one face of an aluminum sample provided the complete Green's function between these two points. This result was generalized to the case where randomization is not produced by the distribution of sources, but is provided by multiple scattering that takes place in heterogeneous media.

By summing the contributions of all sources to the correlation, it has been shown numerically that the correlation contains the causal and acausal Green's function of the medium. Cases of non-reciprocal (e.g., in the presence of a flow) or inelastic media have also been theoretically investigated. Derode et al. (2003) proposed to interpret the Green's function reconstruction in terms of a time-reversal analogy that makes it clear that the convergence of the noise-correlation function towards the Green's function is bonded to the stationary phase theorem. For the more general problem of elastic waves, one could summarize that the Green's function reconstruction depends on the equipartition condition of the different components of the elastic field. In other words, the emergence of the Green's function is effective after a sufficient self-averaging process that is provided by random spatial distribution of the noise sources when considering long-time series as well as scattering (e.g., Gouédard et al. 2008 and references herein).

Applications in Seismology

For the first time, Shapiro and Campillo (2004) reconstructed the surface wave part of the Earth response

by correlating seismic noise at stations separated by distances of hundreds to thousands of kilometers, and measured their dispersion curves at periods ranging from 5 s to about 150 s. Then, a first application of passive seismic imaging in California (e.g., Shapiro et al. 2005; Sabra et al. 2005) appeared to provide a much greater spatial accuracy than for usual active techniques. More recently, the feasibility of using the noise cross-correlations to monitor continuous changes within volcanoes and active faults was demonstrated (e.g., Brenguier 2008a, b). These results demonstrated a great potential of using seismic noise to study the Earth interior at different scales in space and time. At the same time, the feasibility of both noise-based seismic imaging and monitoring in every particular case depends on spatio-temporal properties of the available noise wavefield. Therefore, a logical initial step for most of noise-based studies is to characterize the distribution of noise sources. Also, in many cases, knowledge of the distribution of the noise sources can bring very important information about the coupling between the Solid Earth with the Ocean and the Atmosphere. So far, we can identify three main types of existing seismological applications related to noise correlations: (1) studies of spatio-temporal distribution of seismic noise sources, (2) noise-based seismic imaging, and (3) noise-based seismic monitoring.

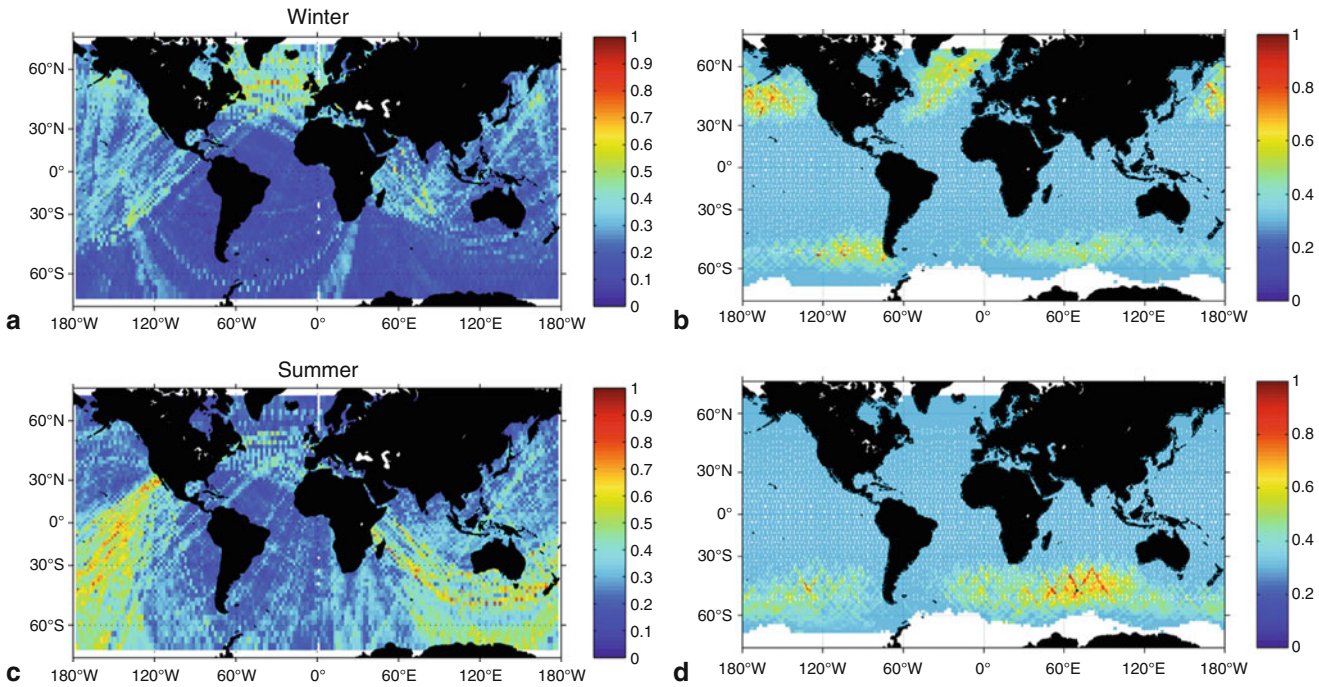
Noise Source Origin and Distribution

Distribution of noise sources strongly depends on the spectral range under consideration. At high frequencies (>1 Hz), the noise is strongly dominated by local sources that may have very different origins and are often anthropogenic. At these scales, the properties of the noise wavefield should be studied separately for every particular case and no reasonable generalization can be done. At longer periods, noise is dominated by natural sources. In particular, it is well established that two main peaks in the seismic noise spectra in so-called microseismic band (1–20 s) are related to forcing from oceanic gravity waves. It has been also argued that at periods longer than 20 s, the oceanic gravity and infragravity waves play a major role in the seismic noise excitation. The interaction between these oceanic waves and the solid Earth is governed by a complex non-linear mechanism (Longuet-Higgins 1950) and, as a result, the noise excitation depends on many factors such as the intensity of the oceanic waves but also the intensity of their interferences as well as the seafloor topography (e.g., Kedar et al. 2008). Overall, the generation of seismic noise is expected to be strongly modulated by strong oceanic storms and, therefore, to have a clear seasonal and non-random pattern.

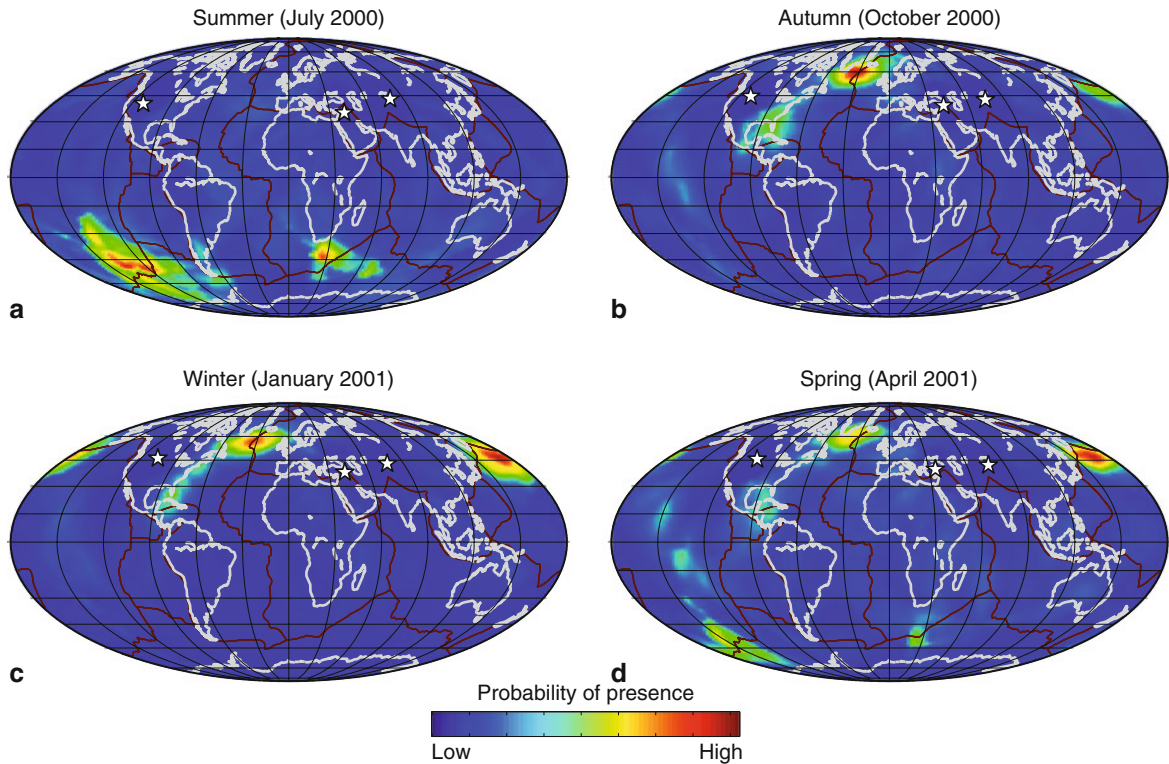
Seismic noise in the microseismic spectral band is dominated by fundamental mode surface waves. It is currently debated whether the surface wave component of microseisms is generated primarily along coastlines or if it is also generated in deep-sea areas. Inhomogeneous distribution and seasonality of microseismic noise sources is clearly revealed by the amplitude of the Rayleigh wave reconstructed in noise cross-correlations (e.g., Stehly et al. 2006) as shown in Fig. 1. At the same time, body waves were detected in the secondary microseismic band and can be sometimes associated with specific storms. Figure 2 shows that sources of microseismic P waves are located in specific areas in deep ocean and exhibit strong seasonality as determined from the analysis of records by dense seismic networks (Landes et al. 2010).

Noise-based Seismic imaging

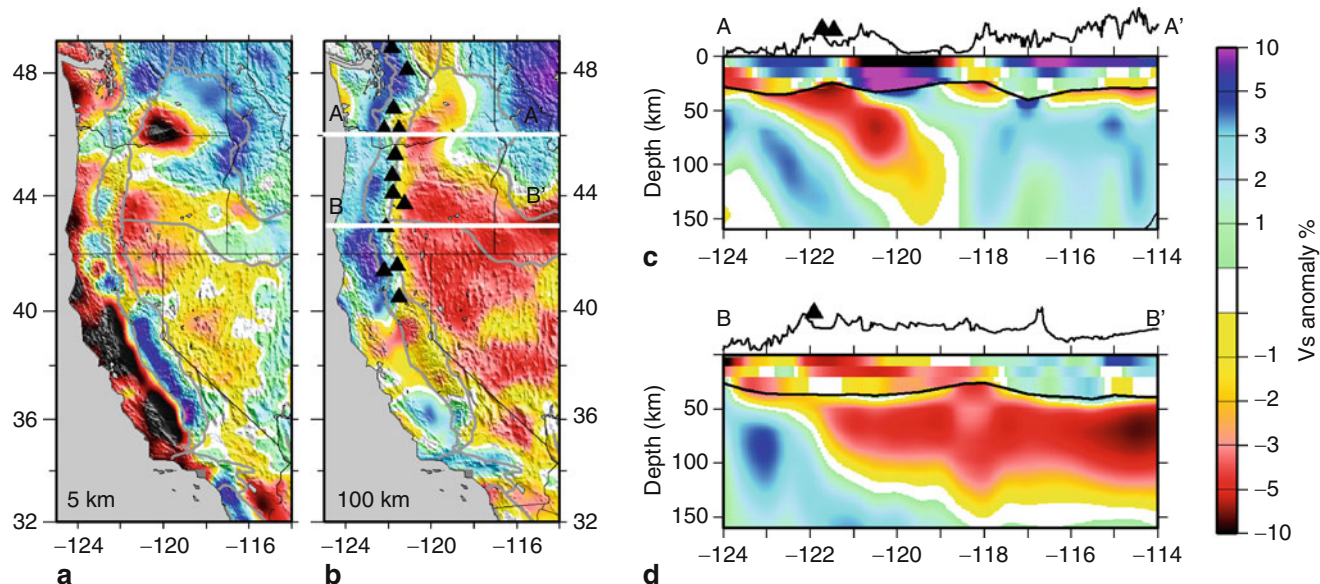
Numerous studies has demonstrated that, when considered over sufficiently long times, the noise sources become sufficiently well distributed over the Earth's surface and that dispersion curves of fundamental mode surface waves can be reliably measured from correlations of seismic noise at periods between 5 and 50 s for most of interstation directions. This led to the fast development during recent years of the ambient-noise surface wave tomography. It consists of computing cross-correlations between vertical and horizontal components for all available station pairs followed by measuring group and phase velocity dispersion curves of Rayleigh and Love waves (e.g., Bensen et al. 2007). This dispersion curves are then regionalized (e.g., Lin et al. 2009) and inverted to obtain three-dimensional distribution of shear velocities in the crust and the uppermost mantle. After first results obtained in southern California (Shapiro et al. 2005; Sabra et al. 2005), this method has been applied with many regional seismological networks (e.g., Yao et al. 2006; Lin et al. 2007; Yang et al. 2008a). At smaller scales, it can be used to study shallow parts of volcanic complexes (e.g., Brenguier et al. 2007). The ambient-noise surface wave tomography is especially advantageous in context of dense continent-scale broadband seismic networks such as available in USA (e.g., Moschetti et al. 2007; Yang et al. 2008b) and Europe (e.g., Stehly et al. 2009). At these scales, noise-based imaging can be used to obtain high-resolution information about the crustal and the upper mantle structure including seismic anisotropy (e.g., Moschetti et al. 2010) and can be easily combined with earthquake-based measurements to extend the resolution to larger depths (e.g., Yang et al. 2008b). An example of results obtained from combined noise and earthquakes based surface wave tomography in western USA is shown in Fig. 3.



Seismic, Ambient Noise Correlation, Fig. 1 Comparison between seasonal variations of the location of seismic noise sources and significant wave height. (a) and (c) Geographical distribution of the apparent source of the Rayleigh waves detected in the 10–20 s noise cross correlations during the winter and the summer, respectively. (b) and (d) Global distribution of the square of wave height measured by TOPEX/Poseidon during the winter and the summer, respectively. (From Stehly et al. 2006)



Seismic, Ambient Noise Correlation, Fig. 2 Seasonal variation of the location of P-wave seismic noise sources in the secondary microseismic band (0.1–0.3 Hz) determined from the analysis of records at the three seismic networks indicated with white stars. (From Landes et al. 2010)



Seismic, Ambient Noise Correlation, Fig. 3 Shear-velocity structure of the crust and the upper mantle obtained from the inversion of the USArray data. (a) and (b) Horizontal cross-sections at depths of 5 and 100 km. (c) and (d) Vertical cross-sections along profiles delineated by

the white lines in (b). Black lines outline the Moho. Topography is superimposed above individual cross sections. The black triangles represent active volcanoes in the Cascade Range. (From Yang et al. 2008b)

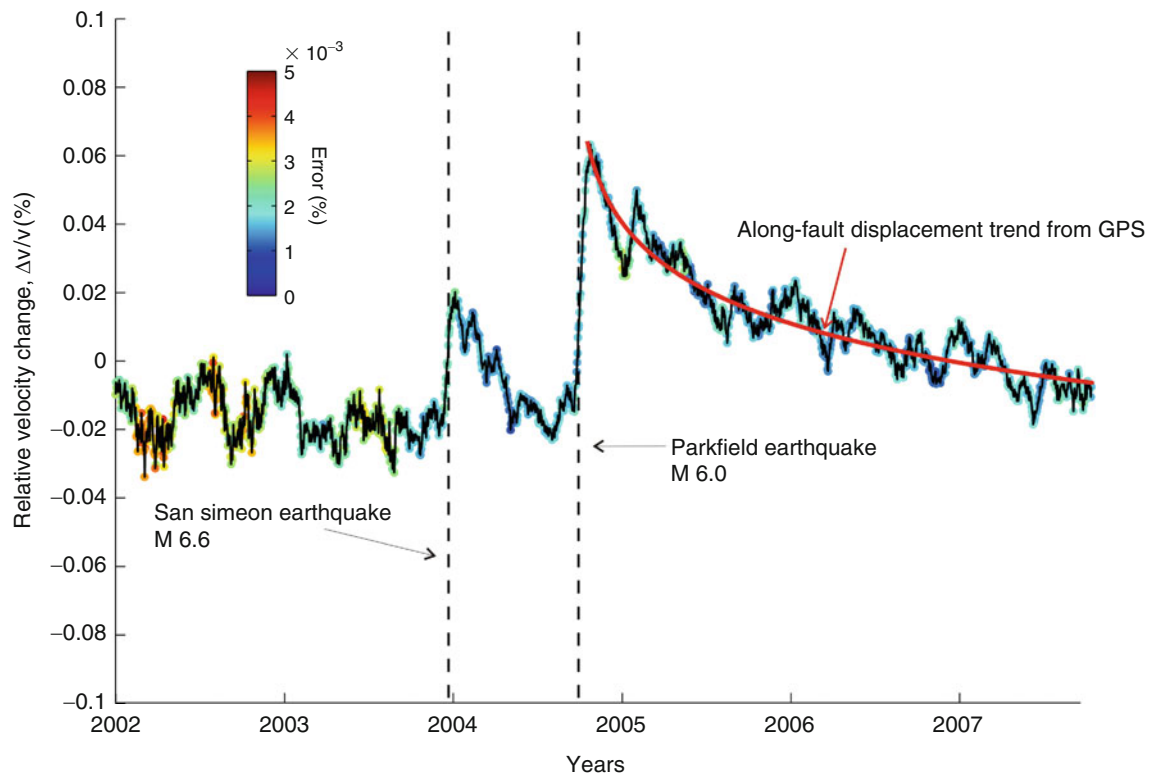
Noise-based Monitoring

One of the advantages of using continuous noise records to characterize the earth materials is that a measurement can easily be repeated. This led recently to the idea of a continuous monitoring of the crust based on the measurements of wave speed variations. The principle is to apply a differential measurement to correlation functions, considered as virtual seismograms. The technique developed for repeated earthquakes (doublets), proposed by Poupinet et al. 1984, can be used with correlation functions. In a seismogram, or a correlation function, the delay accumulates linearly with the lapse time when the medium undergoes a homogeneous wave speed change, and a slight change can be detected more easily when considering late arrivals. It was therefore reasonable, and often necessary, to use coda waves for the measurements of temporal changes. Noise-based monitoring relies on the auto-correlation or cross-correlation of seismic noise records (Sens-Schönfelder and Wegler 2006; Brenguier et al. 2008a, b). When data from a network are available, using cross-correlation take advantage of the number of pairs with respect to the number of stations. It is worth noting that the use of the coda of the correlation functions is also justified by the fact that its sensitivity to changes in the origin of the seismic noise is much smaller than the sensitivity of the direct waves. Several authors noted that an anisotropic distribution of sources leads to small errors in the arrival time of the direct waves, which can be evaluated quantitatively (e.g., Weaver et al. 2009). While in most of the cases, they are acceptable for imaging, they can be larger than the level of precision required

when investigating temporal changes. The issue of the nature of the tail (coda) of the cross-correlation function is therefore fundamental and was analyzed by Stehly et al. (2008). These authors showed that it contains at least partially the coda of the Green function, i.e., physical arrivals which kinematics is controlled by the wave speeds of the medium. It can therefore be used for monitoring temporal changes. As an illustration of the capability of this approach, we present in Fig. 4 a measure of the average wave speed change during a period of 6 years in the region of Parkfield, California. Two main events occurred in this region during the period of study: the 2003 San Simeon and 2004 Parkfield earthquakes. In both cases, noise-based monitoring indicates a co-seismic speed drop. The measured relative variations of velocity before de San Simeon earthquake are as small as 10^{-4} . The changes of velocity associated with earthquakes are associated with at least two different physical mechanisms: (1) the damage induced by the strong ground motions in shallow layers and fault zone, as illustrated by the co-seismic effect of the distant San Simeon event, and (2) co-seismic bulk stress change followed by the post-seismic relaxation, as shown with the long-term evolution after the local Parkfield event, similar in shape to the deformation measured with GPS.

Summary

Continuous recordings of the Earth surface motion by modern seismological networks contain a wealth of information on the structure of the planet and on its temporal evolution.



Seismic, Ambient Noise Correlation, Fig. 4 Relative seismic velocity change during 6 years measured from continuous noise correlations in Parkfield. The dashed lines indicated two major earthquakes: the San

Simeon event that occurred 80 km from Parkfield and the local Parkfield event. (Modified from Brenguier et al. 2008b)

Recent developments shown here make it possible to image the lithosphere with noise only and to detect temporal changes related to inner deformations.

Cross-References

- Body Waves
- Earthquakes and Crustal Deformation
- Seismic Noise
- Seismic Tomography
- Seismic Waves, Scattering
- Surface Waves

Bibliography

- Bensen GD, Ritzwoller MH, Barmin MP, Levshin AL, Lin F, Moschetti MP, Shapiro NM, Yang Y (2007) Processing seismic ambient noise data to obtain reliable broadband surface wave dispersion measurements. *Geophys J Int* 169:1239–1260. <https://doi.org/10.1111/j.1365-246X.2007.03374.x>, 2007.
- Brenguier F, Shapiro NM, Campillo M, Nercessian A, Ferrazzini V (2007) 3-D surface wave tomography of the Piton de la Fournaise volcano using seismic noise correlations. *Geophys Res Lett* 34: L02305. <https://doi.org/10.1029/2006GL028586>
- Brenguier F, Shapiro N, Campillo M, Ferrazzini V, Duputel Z, Coutant O, Nercessian A (2008a) Toward forecasting volcanic eruptions using seismic noise. *Nat Geosci* 1(2):126–130
- Brenguier F, Campillo M, Hadziioannou C, Shapiro NM, Nadeau RM, Larose E (2008b) Postseismic relaxation along the San Andreas fault in the Parkfield area investigated with continuous seismological observations. *Science* 321(5895):1478–1481
- Derode A, Larose E, Tanter M, de Rosny J, Tourin A, Campillo M, Fink M (2003) Recovering the Green's function from field-field correlations in an open scattering medium. *J Acoust Soc Am* 113:2973–2976
- Gouédard P, Stehly L, Brenguier F, Campillo M, de Verdière Colin Y, Larose E, Margerin L, Roux P, Sanchez-Sesma FJ, Shapiro NM, Weaver RL (2008) Cross-correlation of random fields: mathematical approach and applications. *Geophys Prospect* 56:375–393
- Kedar S, Longuet-Higgins M, Webb F, Graham N, Clayton R, Jones C (2008) The origin of deep ocean microseisms in the North Atlantic Ocean. *Proc Royal Soc London A* 464:777–793. <https://doi.org/10.1098/rspa.2007.0277>
- Landes M, Hubans F, Shapiro NM, Paul A, Campillo M (2010) Origin of deep ocean microseisms by using teleseismic body waves. *J Geophys Res*. <https://doi.org/10.1029/2009JB006918>
- Lin F, Ritzwoller MH, Townend J, Savage M, Bannister S (2007) Ambient noise Rayleigh wave tomography of New Zealand. *Geophys J Int*. <https://doi.org/10.1111/j.1365-246X.2007.03414.x>
- Lin F-C, Ritzwoller MH, Snieder R (2009) Eikonal tomography: surface wave tomography by phase-front tracking across a regional broadband seismic array. *Geophys J Int* 177(3):1091–1110
- Longuet-Higgins MS (1950) A theory of the origin of microseisms. *Philos Trans R Soc Lond A* 243:1–35

- Moschetti MP, Ritzwoller MH, Shapiro NM (2007) Surface wave tomography of the western United States from ambient seismic noise: Rayleigh wave group velocity maps. *Geochim Geophys Geosyst* 8:Q08010. <https://doi.org/10.1029/2007GC001655>.
- Moschetti MP, Ritzwoller MH, Lin FC (2010) Seismic evidence for widespread crustal deformation caused by extension in the western USA. *Nature* 464:885–889. <https://doi.org/10.1038/nature08951>.
- Poupinet G, Ellsworth WL, Frechet J (1984) Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras Fault, California. *J Geophys Res* 89:5719–5731
- Sabra KG, Gerstoft P, Roux P, Kuperman WA, Fehler MC (2005) Extracting time domain Green's function estimates from ambient seismic noise. *Geophys Res Lett* 32:L03310
- Sens-Schönfelder C, Wegler U (2006) Passive image interferometry and seasonal variations of seismic velocities at Merapi Volcano, Indonesia. *Geophys Res Lett* 33:L21302
- Shapiro NM, Campillo M (2004) Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise. *Geophys Res Lett* 31:L07614. <https://doi.org/10.1029/2004GL019491>.
- Shapiro NM, Campillo M, Stehly L, Ritzwoller M (2005) High resolution surface wave tomography from ambient seismic noise. *Science* 307:1615–1618
- Stehly L, Campillo M, Shapiro N (2006) A Study of the seismic noise from its long range correlation properties. *J Geophys Res* 111: B10306
- Stehly L, Campillo M, Froment B, Weaver RL (2008) Reconstructing Green's function by correlation of the coda of the correlation (C3) of ambient seismic noise. *J Geophys Res* 113:B11306
- Stehly L, Fry B, Campillo M, Shapiro NM, Guilbert J, Boschi L, Giardini D (2009) Tomography of the Alpine region from observations of seismic ambient noise. *Geophys J Int* 178:338–350
- Weaver RL, Lobkis OI (2001) Ultrasonics without a source: thermal fluctuation correlations at MHz frequencies. *Phys Rev Lett* 87(13):134301. <https://doi.org/10.1103/PhysRevLett.87.134301>
- Weaver RL, Froment B, Campillo M (2009) On the correlation of non-isotropically distributed ballistic scalar diffuse waves. *J Acoust Soc Am* 126:1817–1826
- Yang Y, Li A, Ritzwoller MH (2008a) Crustal and uppermost mantle structure in southern Africa revealed from ambient noise and teleseismic tomography. *Geophys J Int*. <https://doi.org/10.1111/j.1365-246X.2008.03779.x>
- Yang Y, Ritzwoller MH, Lin F-C, Moschetti MP, Shapiro NM (2008b) The structure of the crust and uppermost mantle beneath the western US revealed by ambient noise and earthquake tomography. *J Geophys Res* 113:B12310. <https://doi.org/10.1029/2008JB005833>.
- Yao H, van der Hilst RD, de Hoop MV (2006) Surface-wave array tomography in SE Tibet from ambient seismic noise and two-station analysis – I. Phase velocity maps. *Geophys J Int* 166:732–744

Seismic, Artificial Intelligence to Neural Intelligence for Advanced Interpretation

Kalachand Sain and Priyadarshi Chinmoy Kumar
Wadia Institute of Himalayan Geology, Dehradun, India

Definition

The human brain is an expert system for analyzing patterns. In other words, human or neural intelligence is quick in

demonstrating its proficiency for making qualitative analysis. When machines (or machine intelligence) are trained to work under the guidance of the human brain (or intelligence), optimum solutions to large and complex problems can be achieved with much ease. This motivates to amalgamate machine (or artificial) intelligence with human (or neural) intelligence to solve complicated nonlinear problems. Such practices have been applied to many scientific areas such as face recognition, image processing, seismic data analysis, and interpretation.

Introduction: A Motivation for Automating Interpretation

Seismic reflection data are rigorously processed for producing an appreciable image of the subsurface, which can be interpreted for geologically meaningful features related to hydrocarbon exploration, mineralized prospects, geotectonics, etc. The image can be enhanced by using advanced tools to aid interpreters for better understanding of the subsurface. The interpretation varies from a young interpreter to an experienced interpreter. Strictly speaking, an anomaly in the seismic data for a young interpreter could be a noise, whereas it may be a geologic feature to an experienced interpreter. Hence, interpreters continue to improve their skills, such that a realistic and geologically meaningful interpretation of the subsurface can be achieved.

Advanced interpretation techniques, such as seismic attribute analysis, have brought out revolutionary changes for the interpretation of subsurface. The attribute analysis workflow is one of the major steps of the subsurface interpretational sequence adopted by interpreters. Seismic attributes help in capturing every single anomalous response of the targets from the data and provide guidance to interpret the responses in terms of geological features such as faults, fractures, and folds; stratigraphic features such as pinch outs, debris, slumps, and channels; intrusive features such as sills, dykes, submarine volcanoes, and reefs. Seismic attributes have been recently used as major inputs for generating three dimensional bodies of subsurface, such that the entire tectonic architecture and geometry of these features can be imaged clearly from seismic data. The major and most important problem which agitates an interpreter is the fact that an attribute hardly ever responds to a particular geologic object of interest (Kumar and Sain 2018, 2020; Kumar et al. 2019a, b, c). Hence, the output is a mixed set of geological information overlapped with noisy or unwanted features. Thus, the final interpretation becomes erroneous and perturbs the interpreters.

A plausible way to alleviate such challenges is amalgamating artificial intelligence with neural intelligence. This amalgamation is carried out by designing meta-attributes or hybrid attributes, which are robust computer models (Kumar

and Sain 2018) that are obtained through unification of several attributes trained over interpreter's acquaintances. The meta-attribute is much more efficient than single attribute in interpreting the subsurface geological features.

Automating interpretation of seismic data has been an active area of research. However, most of these automation approaches have concentrated on improving interpretation of discontinuous structures such as the faults and gas migrations from seismic data. Aminzadeh et al. (2013), Connolly and de Groot (2016), and Singh et al. (2016) demonstrated the use of chimney technology (an automated approach) for the interpretation of gas seepages from reflections of seismic data. Kluesner and Brothers (2016) and Kumar and Mandal (2017) highlighted the use of conditioned seismic data in extracting different structural attributes and simultaneously combined them to capture the fault discontinuities from seismic data. Di et al. (2017) further advanced the multiattribute approach using the support vector machine (SVM) technique for interpreting the geologic discontinuities from 3D seismic volume. Kumar and Sain (2018) introduced a thinned fault cube meta-attribute to bring out sharp thinned fault images from seismic data. Wu et al. (2019) adopted deep neural learning approaches for improving the fault images from seismic data. Many advances in this process of interpretation include the works of Kumar et al. (2019a, b, c), where the combination of artificial intelligence and machine intelligence is incorporated for interpreting magmatic bodies, fluid migrations, volcanic systems, and submarine landslides from seismic data.

Here we present examples of a set of meta-attributes that have been designed by the unification of artificial and neural intelligences to produce geologically-meaningful subsurface images of sedimentary basins through an automated approach.

Examples of Meta-attributes

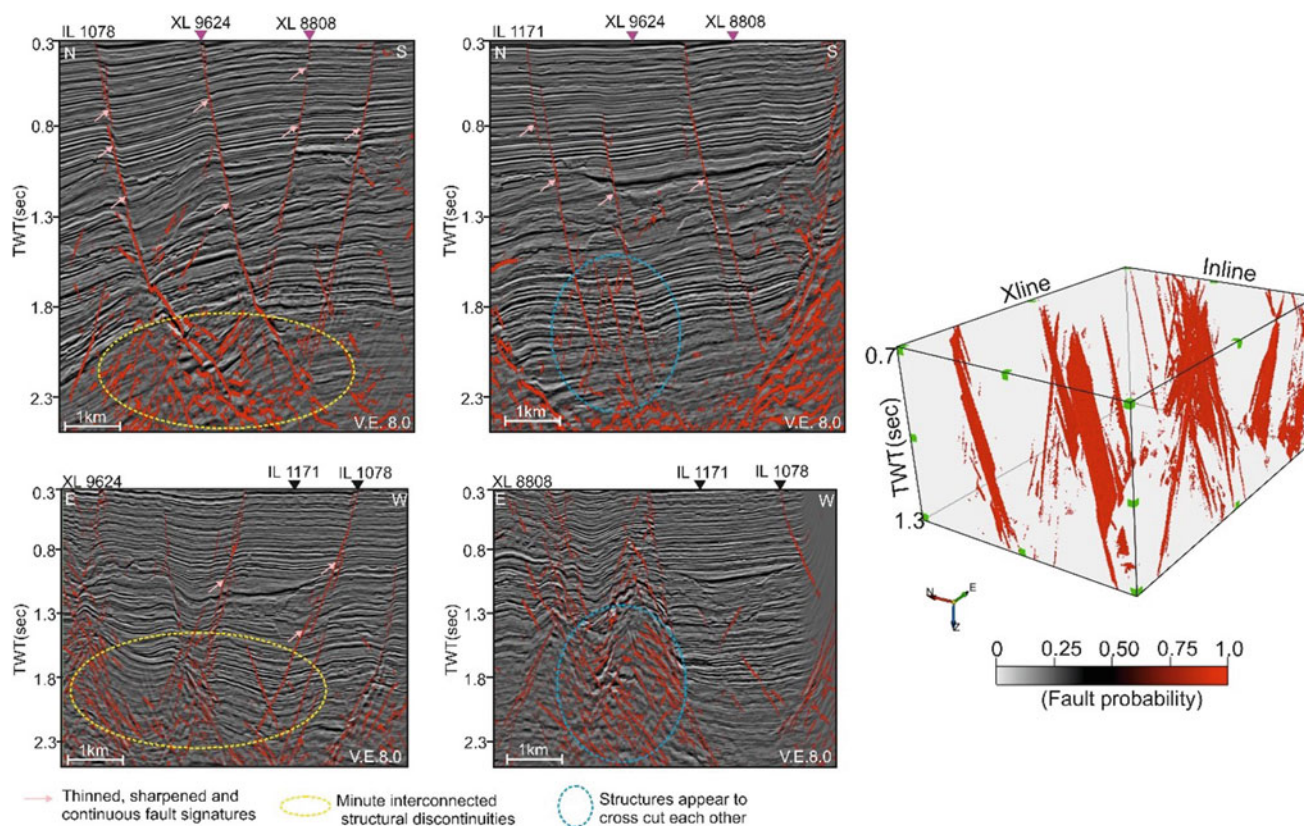
Example 1: Thinned Fault Cube (TFC) Meta-attribute

The first example is a case study performed over the Waitara 3D prospect, off New Zealand, where our objective is to generate a Thinned Fault Cube (TFC) meta-attribute from 3D time-migrated seismic data. Geologic faults are discontinuous structures, which are associated with reflector terminations, vertical disturbances, and discontinuities in seismic reflectors (Tingdahl 2003; Kumar and Sain 2018). These features are characterized with abrupt changes in the dips of seismic reflectors and are generally oriented with different geometric shapes (e.g., curve, straight, or bend) (Tingdahl 2003). The fault zones are also characterized with the loss of energy and frequency content. Seismic attributes, e.g., similarity, coherency, curvature, dip angle variance, dip magnitude, azimuth, and energy attributes, when amalgamated

and trained under interpreter's knowledge, generate a hybrid attribute called as the Thinned Fault Cube (TFC) meta-attribute. Interpreter's past acquaintances play a crucial role as he/she decides all possible locations associated with the faults and all nonfault locations devoid of structural discontinuities within the data. The TFC meta-attribute generates thinned, linear, prominent, distinct, and razor-shaped fault images from the seismic data (Fig. 1).

Example 2: Intrusion Cube (IC) Meta-attribute

The second example is a case study performed over the Kora 3D prospect, off New Zealand, where our objective is to generate an Intrusion Cube (IC) meta-attribute from the 3D time-migrated seismic data. This meta-attribute is very much suitable for interpreting submarine buried volcanic system. Buried volcanic systems are accompanied with igneous and sedimentary processes that may have both positive and negative impacts over the petroleum system, e.g., maturity of the source rocks, migration pathways for hydrocarbons, sealing and trapping mechanism, geothermal history of the basin, etc. (Kumar et al. 2019a). This type of complexities sometimes changes the porosity and permeability of the subsurface reservoirs and poses high risks for exploration and exploitation. Thus, such a complex geological system should be examined critically and investigated thoroughly for prospect evaluation. Submarine buried volcanoes have a tendency to pierce vertically through the host sedimentary succession. Internally they are associated with distorted and chaotic reflections and dissimilar seismic events. Over the crater and along the flanks, they are associated with high amplitudes due to larger impedance contrast with the surrounding sedimentary rocks (Kumar et al. 2019a). While piercing through the host geologic formations, these features tend to possess variable dips. Seismic energy, when passed through these features, gets diminished. The flanks and the crater of the volcano are linked with high coherent energy. Sills and dykes generally act as conducive pathways for the movement of magma into the overlying sedimentary units. Sills are concordant intrusive bodies, commonly subhorizontal with a gentle inclination cross-cutting the subsurface stratigraphy, and exhibit major impact on basin history and petroleum system (Thomson and Hutton 2004). On seismic section, these appear as saucer shaped, gently inclined, distinct high amplitude and discontinuous in nature. However, dykes vertically intrude into the overlying younger strata and exhibit steep dips. Thus, attributes, e.g., reflection strength, dip angle variance, energy gradient, azimuth, and similarity, when amalgamated and trained over interpreter's knowledge to capture the extension and distribution of such a complex plumbing system from seismic data, generate a meta-attribute called Intrusion Cube (IC) meta-attribute. Kumar et al. (2019a) make a first documentation of such a meta-attribute that has brought out a realistic image of the plumbing system



Seismic, Artificial Intelligence to Neural Intelligence for Advanced Interpretation, Fig. 1 The TFC meta-attribute corendered over amplitude data, in the Waitara prospect off New Zealand, shows thin and razor-shaped images (cream arrows) of subsurface faults from seismic data

(consisting of buried volcano, sill networks, dyke swarms, and magmatic ascent) from seismic data (Fig. 2).

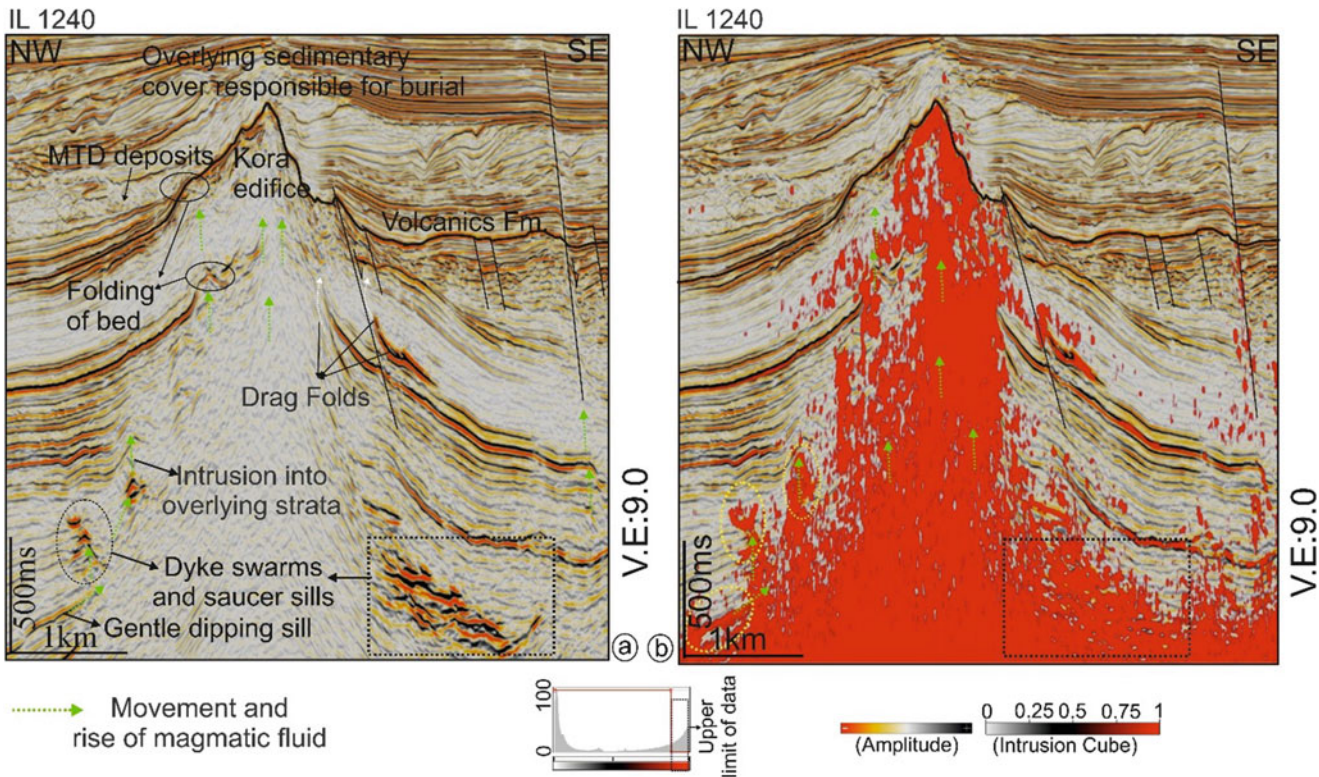
Example 3: Sill Cube (SC) Meta-attribute

The third example is a case study performed over the Utgard prospect, Vøring Basin, off Norway, where our objective is to generate a Sill Cube (SC) meta-attribute from 3D time-migrated seismic data. This meta-attribute is very much suitable for automatic interpretation of magmatic sill complexes from seismic data. Conventionally, magmatic sills are interpreted as tabular intrusive rocks with concordant surfaces showing concave upward cross-sectional geometries and discordant limbs (Hansen et al. 2004). These are concordant intrusive bodies, commonly subhorizontal with a gentle inclination cross-cutting the subsurface stratigraphy, and exhibit major impact on basin history and petroleum system (Thomson and Hutton 2004). These geologic features help in transporting magmatic fluids into the overlying stratigraphic units. Magmatic sills exhibit high amplitude and entropy character due to larger impedance contrast with the surrounding sedimentary rocks. Moreover, they possess an appreciable contrast in comparison to the host sedimentary succession (Chopra and Marfurt 2007). Thus, these seismic characteristics make suitable attributes, e.g., reflection strength, texture contrast, and texture entropy

for amalgamation and training them under interpreter's guidance to generate a hybrid attribute, called Sill Cube meta-attribute (Kumar et al. 2019b). Figure 3 shows the meta-attribute corendered over a conventional seismic section.

Example 4: Fluid Cube (FC) Meta-attribute

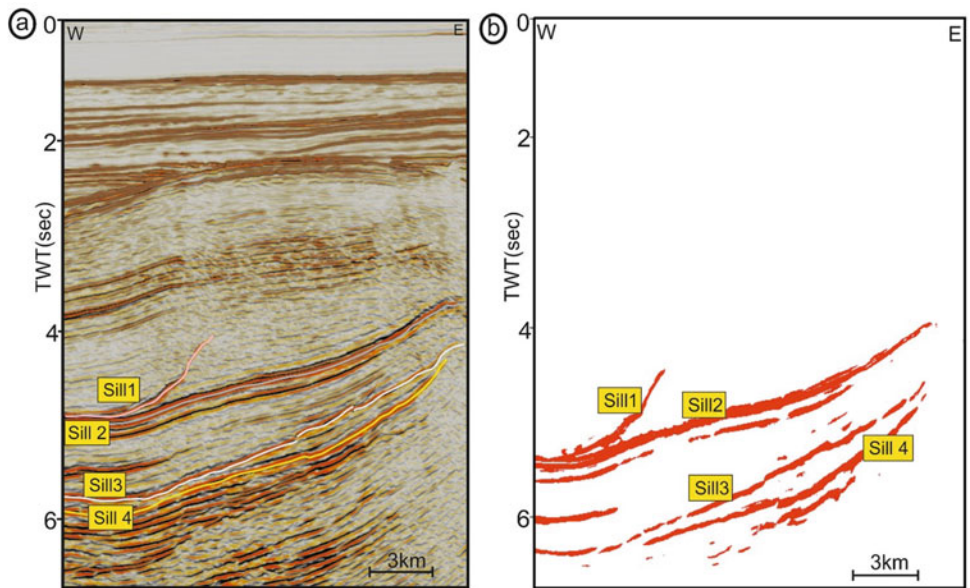
The fourth example is a case study performed over the Parihaka 3D prospect, off New Zealand, where our objective is to generate a Fluid Cube (FC) meta-attribute from the 3D time-migrated seismic data. This meta-attribute is very much suitable for interpreting fluid migration (e.g., hydrocarbon fluids and magmatic fluids) through subsurface geological formations. On seismic section, these migration zones are associated with disordered reflection patterns, low similarity among seismic traces, loss in energy, and frequency and appear to have vent-like structure. Suitable attributes such as similarity, energy, and frequency variance, when amalgamated and trained over interpreter's knowledge, generate a meta-attribute called Fluid Cube (FC) meta-attribute. Kumar et al. (2019c) have brought out a realistic interpretation of such migration zones using the FC meta-attribute from seismic data. Fluids are observed to get migrated from the subsurface geological formations bounded by a system of hard-linked normal faults (Fig. 4a, b).



Seismic, Artificial Intelligence to Neural Intelligence for Advanced Interpretation, Fig. 2 (a) Interpreted Inline 1240 demonstrating the buried volcanic system with different geological elements

(black arrows and dotted ovals); (b) The IC meta-attribute corendered over the amplitude data, exhibiting the architecture of buried volcanic system in Kora prospect off New Zealand

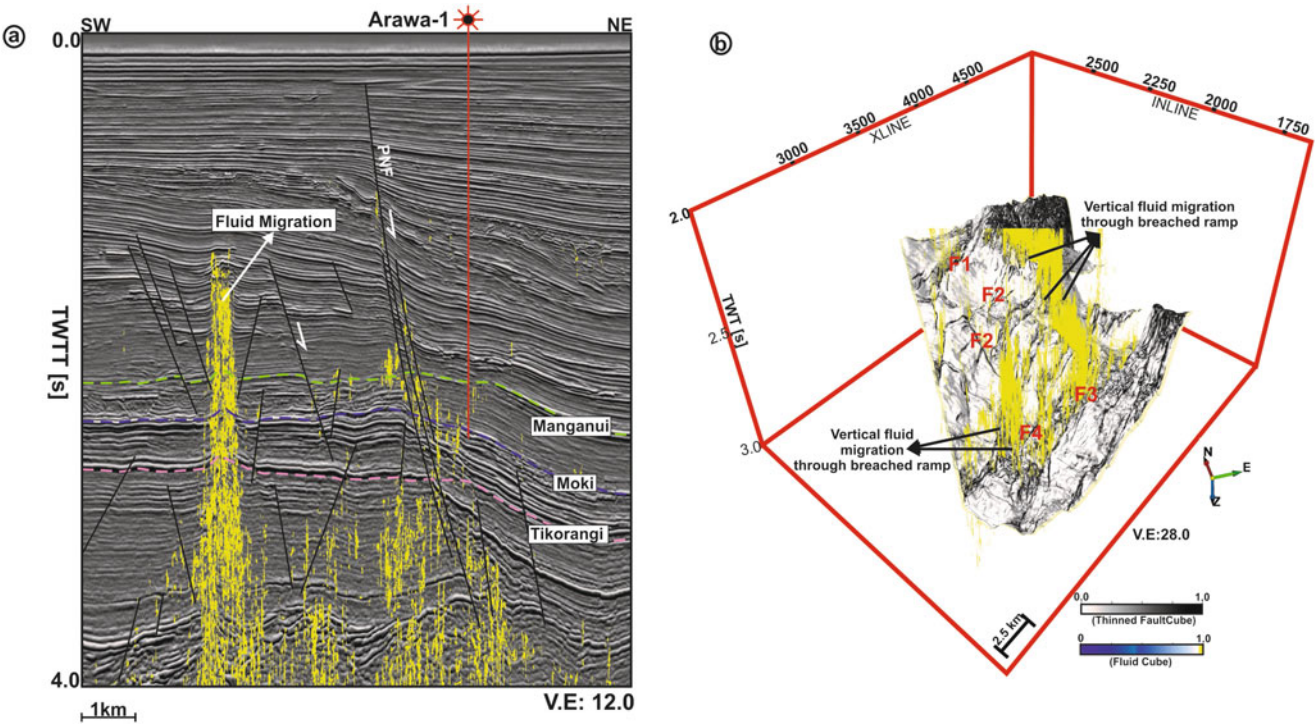
Seismic, Artificial Intelligence to Neural Intelligence for Advanced Interpretation, Fig. 3 (a) Amplitude section showing manually identified sills from seismic data; (b) automatic interpretation of magmatic sills using SC meta-attribute



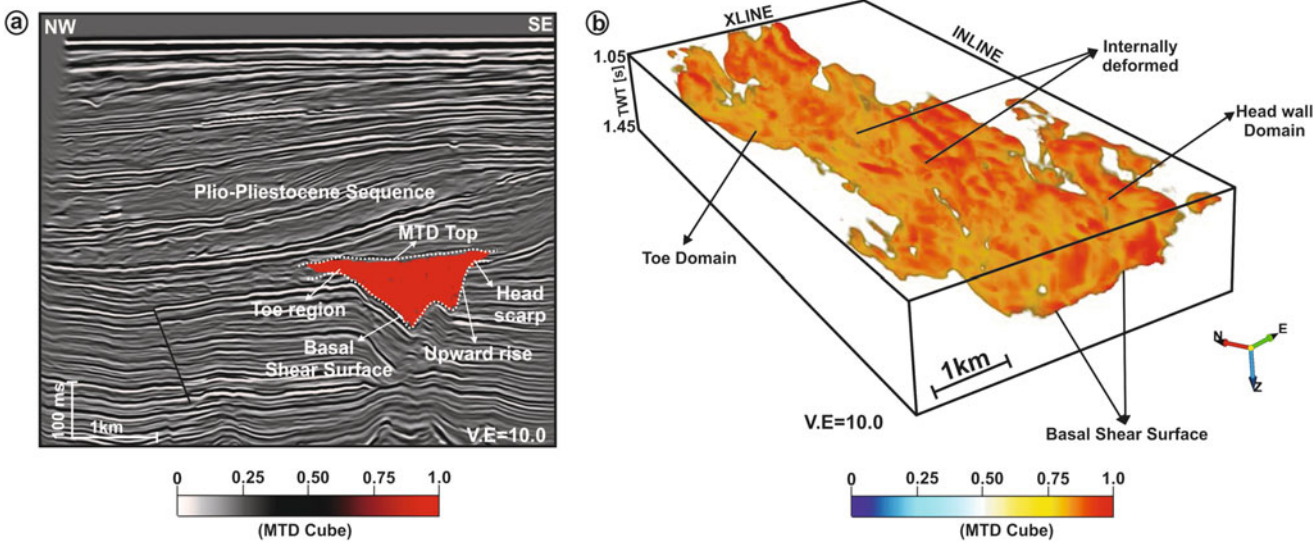
Example 5: Mass Transport Deposit Cube (MTDC) Meta-attribute

The fifth example is a case study performed over the Karewa 3D prospect, off New Zealand, where the objective is to generate a Mass Transport Deposit Cube (MTDC) meta-

attribute from the 3D time-migrated seismic data. Mass Transport Deposits (MTDs), occurring in different tectonic and depositional settings, are generally defined as gravity induced slope failure deposits that include creeps, slides, slumps, or debris flows. These deposits are internally deformed and



Seismic, Artificial Intelligence to Neural Intelligence for Advanced Interpretation, Fig. 4 (a) The FC meta-attribute corendered over amplitude data in the Parihaka prospect revealing the migrated fluids from the subsurface formations; Top of different formations (Tikorangi, Moki and Manganui) are marked by dashed lines. Faults are shown by black solid lines. (b) 3D view of the TFC and FC meta-attributes bringing out realistic interpretation of subsurface geological scenario from 3D seismic data



Seismic, Artificial Intelligence to Neural Intelligence for Advanced Interpretation, Fig. 5 (a) The MTDC meta-attribute corendered over the amplitude section of the Karewa prospect; (b) 3D view of the MTDC meta-attribute revealing the structural geometry of the MTD in the Karewa prospect, off New Zealand

associated with variable shape and size. During slope failure, masses tend to flow downslope over a shearing surface called as the basal shear surface (BSS) that forms the base of the MTDs. Moreover, the BSS preserves the record of all erosional and deformational activities experienced by these

deposits or masses during their transportation. Their interpretation is crucial, as such deposit during transportation over the instable slope may lead to several catastrophic submarine landslides, tsunamis, and avalanches and is a precursor for subsea installation. When suitable attributes such as

similarity, dip angle variance, energy, and average frequency variance are amalgamated and trained over interpreter's knowledge, generates a meta-attribute called Mass Transport Deposit Cube (MTDC) meta-attribute. This meta-attribute reveals the detailed structural geometry of the MTD, which is very crucial for its interpretation (Fig. 5).

Conclusion

When seismic attributes (that act as corner stones) are combined suitably and trained over an interpreter's acquaintances or past knowledge through artificial intelligence, a meta-attribute is generated that augments the image of subsurface and hence its interpretation. Meta-attribute is robust and automatically derived that has brought revolutionary changes in the interpretation of 3D seismic data. The most exciting part of the meta-attribute concept is the amalgamation of artificial intelligence of neural networks and human or neural intelligence to improve the interpretation task. This concept has been applied to 3D seismic data in different geological provinces demonstrating how the seismic images can be enhanced to aid the interpretation of several subsurface features. To have more detailed information on the meta-attributes, the readers may explore the following articles, e.g., Singh et al. 2016; Kumar and Mandal 2017; Kumar and Sain 2018, 2020; Kumar et al. 2019a, b, c. This approach can be extended for the interpretation of other geophysical data also.

Bibliography

- Aminzadeh F, Berge T, Connolly D (2013) Hydrocarbon seepage: from source to surface. SEG/AAPG, Tulsa, USA
- Chopra S, Marfurt KJ (2007) Seismic attributes for prospect identification and reservoir characterization. SEG, Tulsa
- Connolly D, De Groot P (2016) Chimney atlas to quantify top seal and charge risk: case study from Maari oil field, Taranaki Basin, New Zealand. In: EAGE 78th EAGE Conference & Exhibition, Vienna
- Di H, Shafiq MA, AlRegib G (2017) Seismic-fault detection based on multi-attribute support vector machine analysis. In: SEG technical program expanded abstracts, pp 2039–2042
- Hansen DM, Cartwright JA, Thomas D (2004) 3D seismic analysis of the geometry of igneous sills and sill junction relationships. *Geol Soc Lond Mem* 29:199–208
- Kluesner JW, Brothers DS (2016) Seismic attribute detection of faults and fluid pathways within an active strike-slip shear zone: new insights from high-resolution 3D P-Cable™ seismic data along the Hosgri Fault, offshore California. *Interpretation* 4:SB131–SB148
- Kumar PC, Mandal A (2017) Enhancement of fault interpretation using multi-attribute analysis and artificial neural network (ANN) approach: a case study from Taranaki Basin, New Zealand. *Explor Geophys CSIRO* 49:409–424
- Kumar PC, Sain K (2018) Attribute amalgamation-aiding interpretation of faults from seismic data: an example from Waitara 3D prospect, offshore Taranaki basin, New Zealand. *J Appl Geophys* 159:52–68
- Kumar PC, Omosanya KO, Sain K (2019a) Sill cube: a novel automated approach for interpreting magmatic sills from reflection seismic data. *Mar Pet Geol* 100:60–84
- Kumar PC, Sain K, Mandal A (2019b) Delineation of buried volcanic system in Kora prospect off New Zealand using artificial neural networks and its implications. *J Appl Geophys* 161:56–75
- Kumar PC, Omosanya KO, Alves T, Sain K (2019c) A neural network approach elucidating fluid leakage along hard linked normal faults. *J Mar Petrol Geol* 110:518–538
- Kumar PC, Sain K (2020) Interpretation of magma transport through saucer sills in shallow sedimentary strata: An automated machine learning approach. *Tectonophysics* 789(228541):1–16
- Singh D, Kumar PC, Sain K (2016) Interpretation of gas chimney from seismic data using artificial neural network: a study from Maari 3D prospect in the Taranaki basin, New Zealand. *J Nat Gas Sci Eng* 36:339–357
- Thomson K, Hutton D (2004) Geometry and growth of sill complexes: insights using 3D seismic from the North Rockall Trough. *Bull Volcanol* 66:364–375
- Tingdahl KM (2003) Improving seismic chimney detection using direction attributes. In: Nikarvesh M, Aminzadeh F, Zadeh LA (eds) *Soft computing and intelligent data analysis in oil exploration, developments in petroleum science*, vol 51, Elsevier, Amsterdam, The Netherlands, pp 157–173
- Wu X, Liang L, Shi Y, Fomel S (2019) FaultSeg3D: using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation. *Geophysics* 84(3):1–36

Seismic, Migration

Samuel H. Gray¹ and Nittala Satyavani²

¹CGGVeritas, Calgary, AB, Canada

²Seismic Group, National Geophysical Research Institute, Hyderabad, India

Synonyms

Seismic imaging

Definition

Seismic migration. A set of techniques for transforming recorded (elastic-wave) seismic reflection data into an image of reflecting boundaries in the earth's interior. In simplest form, these are intended to correct certain distortions present in recorded wavefields. The distortions can be caused by: diffractors inside the earth, which scatter the incident energy to a range of receiver locations; geometric effects caused by dipping reflectors; and velocity effects, which cause the seismic waves to change direction as they propagate from source to reflector to receiver. Migration is also used to estimate seismic velocity and to provide amplitudes for rock property analysis.

The migration heuristic assumes that subsurface reflectors are made up of point diffractors (Trorey 1970). The response of a reflector to input seismic energy is the sum of responses to

all the diffractors, and migration transforms this response to the actual reflector shape by collapsing each diffractor response to a point (diffraction stack). Migration can be applied both in prestack and poststack (zero-offset data) domains, and in general, the quality of migration depends on the data quality (S/N), velocity field used in migration, migration algorithm, and the type of migration (2D/3D; depth/time; prestack/poststack).

History

Mapping Using Pencil and Paper: Mechanical Migration

The first migrations were performed in the 1920s as a manual operation. A single-fold reflection profile was acquired with sources and receivers on the surface of the earth, and two-way travel times for events from a shallow dipping reflector were picked on the recorded traces. The extent of the lateral offset between source and receiver locations was neglected, which is equivalent to assuming that sources and receivers were colocated, or at zero offset. Then circular arc segments were drawn, centered at the source-receiver midpoints and with radii equal to one-half the product of the picked travel times and the velocity of propagation in the overburden. The envelope of these mapped arcs gave a rough picture of the reflecting interface (Fig. 1).

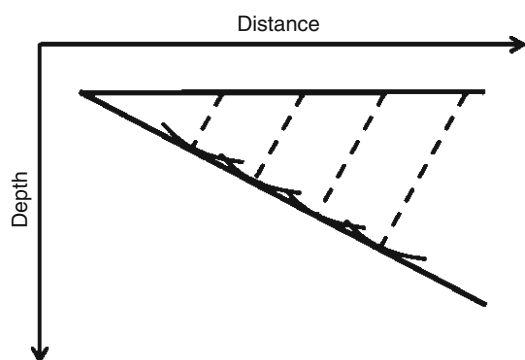
This technique reveals principles and problems of migration. First, reflection events move sideways, or *migrate*, from their picked time locations on the recording surface (the source-receiver midpoints) to subsurface locations at the reflection interface. When reflectors in the earth are not flat, the lateral position of a subsurface reflector is different from that of the recording location of the same event. Second, the assumption of colocated sources and receivers is problematic. As the use of seismic data increased, reflection profiles were typically acquired using higher fold, or

multiplicity: for each source, a spread of many receivers recorded reflection data in order to provide a greater multiplicity of observations. Some of the receivers were at considerable distance from the source locations, violating the zero-offset assumption. Later, the common-midpoint (CMP) stack (Mayne 1962) performed approximate corrections intended to shift each event on a nonzero-offset recorded trace to its zero-offset time, once again allowing the use of the zero-offset assumption in migration (poststack migration). In areas of complicated geologic structure, however, the time corrections used in the CMP stacking process are not accurate, requiring that each recorded trace be migrated using its actual source and receiver location (prestack migration). Third, if more than one reflector exists inside the earth, the various migrated events can interfere with one another as the circular arcs that build up their envelopes overlap. This problem is solved by the more recent application of wave theory to migration. This allows the rigorous migration of waveforms, with energy buildup near reflector locations, and energy cancellation away from reflector locations. Fourth, if the velocity used in the migration process is incorrect, the image of a subsurface reflector will be located incorrectly. This fundamental velocity problem persists today.

The process of swinging reflection events through arcs to construct approximate reflector segments spawned a number of mechanical devices that migrated reflection records. They allowed the migration of picked events from nonzero-offset traces. However, they were all based on the same principle of swinging events over a restricted range of angles. This process was formalized by Hagedoorn (1954).

Digital Migration

Seismic recording, processing, and migration became digital in the 1960s. Digital processing emphasized the discrete sampling of signals, in both time and space. It also allowed reflection records to be treated as wavefields, and wavefield algorithms could be applied to them. Wave-equation migration methods arose, based on differential equations, integral solutions to the wave equation, or mapping in the Fourier domain (Berkhout 1982; Yilmaz 1987). These methods are in use today. They have different realms of validity and different efficiencies. One of the methods, integral (Kirchhoff) migration, deserves mention. This technique is the direct descendant of pencil-and-paper migration. In Kirchhoff migration, samples on the recorded input traces are swung out over a range of angles. The earlier techniques could consider only isolated events, because constructing the envelopes corresponding to all reflection events of a record would produce an uninterpretable mess. However, including the wave equation as an essential part of the process allowed all samples to be swung out. Constructive and destructive interference, a natural consequence of treating the samples as components of a wavefield, caused reflection events to appear



Seismic, Migration, Fig. 1 Schematic illustration of early pencil-and-paper migration. Assuming constant velocity and zero-offset source and receiver, an event on a seismic trace is swung out along an arc of a circle. Repeating this for the corresponding event on all traces produces an image of a reflector, which lies along the envelope of all the arcs

on the image, not as envelopes, but as waveforms moved to reflector locations.

Poststack Migration

The first digital migrations were performed on computers with limited memory and processing capability. Migrating all traces of a two-dimensional (2-D) reflection survey involving hundreds of source locations, each with tens of receiver locations, would have taken months on early computers. So it became necessary to make use of the intermediate process of CMP stack to reduce the number of traces to be migrated. As mentioned above, this process performs time (“normal moveout,” or NMO) corrections to traces with nonzero source-receiver offset. These corrections are time shifts, different for each sample, intended to produce equivalent zero-offset traces to be summed, or stacked, together. The quality of the normal-moveout process depends on the stacking velocity, which determines the time shift to be applied to each sample of each nonzero-offset trace. In order to stack together traces in a domain where the velocity varies as little as possible, it is desirable to sort the input traces from their original recorded common-shot configuration to a set of records where each record has a common source-receiver midpoint – thus the term common-midpoint stack. The stacking velocity for a common-midpoint gather is taken to be the velocity at the midpoint location. This velocity can vary with time down the zero-offset trace of the record. For each value of time, it is a type of average velocity. When all the traces in a CMP gather have been NMO-corrected and summed (stacked) together, the CMP gather has been replaced with a single, zero-offset, trace. The NMO/stack process is illustrated on a single reflection event in Fig. 2.

The ensemble of zero-offset traces, from all CMP gathers indexed by their midpoint locations, is a stacked record ready to be (poststack) migrated.

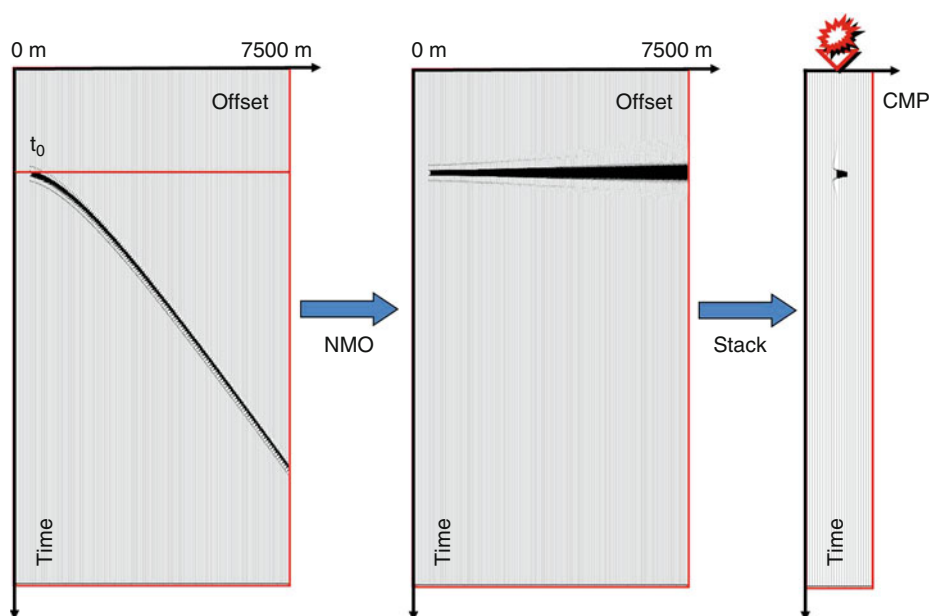
The CMP stacking process makes several approximations that are incompatible with the physics of wave propagation inside the earth. For example, sorting into CMP records and assuming that the seismic velocity does not change for all traces within a record ignores the lateral variability of the actual velocity, and it assumes that the reflectors inside the earth are locally flat below each midpoint location. These approximations were not usually damaging in earlier times, when drilling prospects were in shallow, relatively simple geology. As shallow targets were drilled, leaving deeper prospects under complex overburden with complex velocity behavior, the approximations broke down. Since about 2000, migration performed without CMP stacking (prestack migration) has become commonplace.

Prestack Migration

Dropping the NMO/stack process from the migration flow eliminated some problems, but it made the velocity estimation problem more complicated. NMO velocities were useful for migration as well as stacking. These velocities were not necessarily accurate, but they provided a guide for the velocities to be used in migration. Using a rough estimate of velocity, poststack migration generates plausible migrated images, even if they lack the resolution required for interpretation of subtle targets. Accurate prestack migration, on the other hand, requires the velocity to be estimated without a prior stacking velocity analysis. New tools, consistent with wave propagation inside a geologically complicated earth, were needed for this task. These tools have been under

Seismic, Migration,

Fig. 2 NMO and CMP stack illustrated on a single reflection event



development since the 1980s, but are not yet completely satisfactory. Unlike NMO/stack, they allow arbitrary velocity structure inside the earth, and there is no “locally flat” assumption.

2-D, 3-D, Wide-Azimuth 3-D Acquisition, and Migration

Early seismic reflection profiles were acquired along 2-D lines, both land and marine. On land, a line of receivers was laid out, and shots were fired into the earth at locations along the line. As the shot locations moved along the line, receivers were picked up from the back end of the spread and placed at the front. Marine data were acquired using a vessel towing a linear array of receivers (a streamer), firing at regular locations along the line. Two-dimensional processing assumes that the subsurface of the earth is invariant in the direction perpendicular to the line direction. When this assumption is violated, “out-of-plane” reflectors (not vertically below the acquisition line) contribute energy to the reflection records; 2-D migration cannot map this energy to its correct locations. For this reason, 3-D acquisition became commonplace in the 1970s and 1980s. On land, the receivers were laid out in a 3-D pattern (patch) that could be repeated as in 2-D. Gradually, the receiver patch became larger (thousands of meters on a side), with a wider range of azimuths, or compass directions of vectors joining the source and receiver locations. Today, the recording patch often contains several thousand receivers. Wide-azimuth marine acquisition is more difficult than land because it is physically impossible for a single vessel to tow the large number of streamers required. However, adding a second vessel as a source vessel alongside the first vessel at a large lateral distance produced data over a wide range of azimuths as for land.

Moving from 2-D to 3-D enabled the accurate migration of geologically complex structures such as salt bodies and overthrust structures, and it altered the economics of migration. In the initial days of 3-D seismic data, 2-D prestack migration was available but not commonly applied because it was still too computationally expensive. A fortiori, 3-D prestack migration was not yet feasible. In 3-D, the numbers of input and output locations are increased by an order of magnitude over 2-D, increasing migration cost by two orders of magnitude. Three-dimensional migration was earlier accomplished by a two-pass migration algorithm, which migrates the data first in inline direction and then in crossline direction, using the standard 2D migration algorithms and the choice of migration algorithms was open. With the availability of greater computational facilities, the 3D migration now works in the true sense of 3D by treating the hyperboloids in the inline and crossline directions simultaneously.

Deep Crustal Imaging

Most migration projects are performed by the oil and gas industry for hydrocarbon exploration, whose target depths

rarely exceed 10 km. However, solid-earth geophysicists acquire and migrate seismic data to map crustal structures tens of kilometers deep. Imaging very deep targets with cumulative overburden effects such as wavefield scattering and absorption presents added challenges, and migration cannot be expected to be successful as often as it is for hydrocarbon exploration (Warner 1987).

Time Migration, Depth Migration

Seismic traces are recorded in time, with a particular event occurring at different times on traces with different lateral positions. The CMP stack produces a composite section, with a trace at each position. Although migration is used to image reflectors inside the earth, the CMP stack itself provides a crude image. With its collapse of many seismic records into one, and its use of redundancy (fold) to suppress noise, the CMP stack produces images that can sometimes be interpreted without migration. If the CMP stack is migrated with the depth axis on the migrated image replaced by time, events on the unmigrated stack and the migrated stack can be compared directly, helping with interpretation. Such a migration, called “time migration,” is opposed to the more natural “depth migration.”

This fact has caused a diversity of approaches to migration. Geophysical interpreters often prefer time migration, but structural geologists and engineers usually prefer depth migration. Structural geology deals with space (depth) and how space is filled with material; geologists and engineers are concerned with the physical dimensions of a potential hydrocarbon trap. Even if a migrated image is interpreted in time, in principle, it is preferable to perform the migration and display the final image in depth. In practice, however, there are valid arguments against this statement.

First, time migration is sometimes more robust than depth migration; that is, it is often easier to produce an interpretable image with time migration than with depth migration, even if the image is known not to be accurate. This is because time and depth migration treat velocity differently. Time migration, following procedures developed for NMO/stack, uses what is best called an imaging velocity field, i.e., one that best focuses the migrated image. At each image location, this velocity is an average of the seismic velocity values in the overburden above the image depth. The imaging velocity is free to vary spatially. In essence, time migration performs a constant-velocity migration at each image location, where the constant can change from point to point. The imaging velocity field that produces the best image need not be consistent with any possible field of actual velocities inside the earth, and any attempt to convert the imaging velocities to a field of geologically possible earth velocities can fail. For example, assuming that the imaging velocities are root-mean-square

velocities and using the Dix (1955) equation to invert these to interval velocities often produces physically impossible values. This inconsistent handling of velocity allows time migration to be robust. Second, the precision implied by depth migration is not always justified. A migrated image purports to show locations of reflectors and diffractors inside the earth, but often the locations are incorrect. It is easy to see that this can happen for time migration, which is not precise because of its inconsistent handling of velocity. For depth migration, the implied accurate positioning of reflectors assumes a correct migration velocity, which is an interval velocity field, i.e., a field of local velocities. Using accurate interval velocities allows depth migration to image more accurately than time migration can, but using inaccurate interval velocities can degrade image coherence that the less precise velocities of time migration often preserve.

Purposes of Migration

Structural Imaging

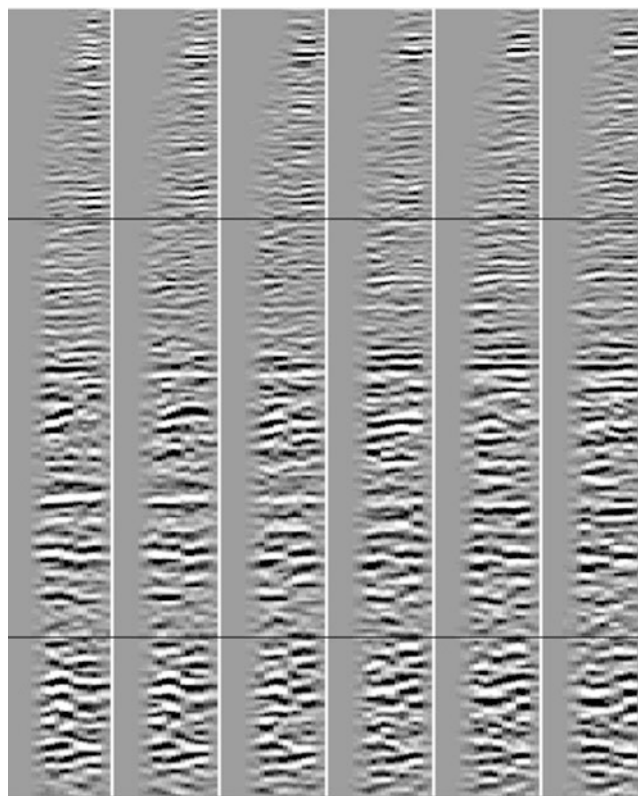
The primary goal of seismic migration is to correct distortions present in recorded seismic records, due to uncollapsed diffraction energy, geometric effects, or velocity effects. Even in areas of fairly flat geology, diffracted energy from abrupt fault terminations can obscure deeper reflection events, making their interpretation difficult. Therefore, collapsing unwanted diffraction energy to scatterer locations is important. Geometric and velocity effects are caused by complicated geologic structures. Familiar examples are “bowtie” signatures on stacked records converted by migration into synclines, and broad anticlinal structures converted by migration into narrower, steeper structures. Other examples include complicated water bottoms, complicated salt geometry, thrust structures, heavily faulted areas, and gas-charged sediments. In these, reflector geometries or lateral velocity variations cause wavefield distortions that must be corrected before the image can be interpreted.

Velocity Estimation

Poststack migration is performed using a velocity field that is specified beforehand. The velocity can be derived from NMO velocity analysis or from an assumed geologic model. It is difficult to analyze a poststack migrated image for velocity errors. By contrast, we can analyze prestack migration for velocity errors. To do this, one performs prestack migration separately on all the input records, with individual images that overlap considerably. In the overlap areas, the different images can be checked for consistency by sorting the migrated traces into common-image-point gathers (CIGs), which are the migration analog of NMO-corrected unmigrated CMP gathers. Within a CIG, each trace refers to a single surface location, imaged from a different input

record. As an example, Fig. 3 shows six CIG's from a land survey, referring to images below six different locations on the earth's surface. Each CIG displays migrated traces from different source-receiver offsets. If all the individual migrations have imaged one particular event at the same depth, the images are consistent at that location. That is the case in this example: except at the greatest depths, most of the events on each CIG appear flat. If the different migrations image a particular event at different depths, the images will be inconsistent, and the event will not appear flat on a CIG. In order to make the images consistent for such an event, the migration velocity will need to be adjusted in the overburden above the event location.

For prestack depth migration, the velocity estimation problem is to build an interval velocity field that makes the migrated events appear flat on all the CIGs. This is usually posed as an algebraic inverse problem, called seismic tomography, involving the interval velocity field, the reflector locations, and the raypaths from the reflectors to the source and receiver locations. This problem is solved iteratively for velocity model refinements. In order to avoid converging to an unrealistic model, tomography usually needs to be constrained, using partial knowledge of the seismic velocities



Seismic, Migration, Fig. 3 Prestack depth migration common-image-point gathers from six different locations on a survey from Western Canada. Each of the gathers contains traces from migrations of data with offsets ranging from 0 m to 4500 m. The events on the gathers are flat except at the greatest depths, indicating that the velocity field is largely correct

and reflecting horizon locations. This biases the solution, and it can result in incomplete event flattening on the CIGs. Tomography is widely used, but it usually produces velocity models with limited spatial resolution.

An attempt to circumvent the issue of velocity dependency of prestack time migration is made and the concept of common migration reflector element stacking (Dell et al. 2012) is introduced. In this scheme, a stack of common scatter point (CSP) gathers is generated, while including an automatic update of time-migration velocities. The CSP gathers are generated by a direct mapping of summed amplitude into the common offset apex of the double square root (DSR) migration operator. The DSR operator is calculated as follows:

$$t_D = \sqrt{\frac{t_0^2}{4} + \frac{(m-h)^2}{V^2(t_0)}} + \sqrt{\frac{t_0^2}{4} + \frac{(m+h)^2}{V^2(t_0)}}$$

where h is the half source-receiver offset, m is the midpoint displacement with respect to the considered CMP position, m_0, t_0 correspond to the vertical ZO two-way traveltime, and $V(t_0)$ is the time-migration velocity belonging to t_0 . The CSP gathers are generated exactly at the locations of the input traces as that of a CMP gather and so, the method does not suffer from generation of extra traces, and is more suitable for application in 3D case.

More recently, the approach of automatic migration velocity estimation (Dong et al. 2019) is becoming popular, wherein the velocity field is constructed from the common reflection point (CRP) gathers. The scheme is similar to the conventional velocity analysis, except that it is based on residual moveout and not the normal move out. The event picking is done in the common reflection point gathers, and the scheme is implemented by semblance scanning, as in conventional velocity analysis. The velocities are then picked at discrete grids using the least squares method and the spatial interpolation is carried out to create a velocity model. The principal advantage is that the velocity model can be obtained in few iterations and fitting curves of the picked events in a range of offsets prevents the scheme from entering into the local minima issue as is the case with curve fitting to a normal moveout.

Increasingly, seismic velocities are recognized as anisotropic: at each location, seismic velocity varies with direction. Including anisotropy makes velocity estimation more difficult, but anisotropy is usually needed to produce accurate images. The most common assumption for anisotropy is transverse isotropy, often with a tilted axis of symmetry (TTI), which is generally the case in presence of fractures. The issue of fracture induced anisotropy is routinely addressed using the azimuth sector migration (Peng et al. 2013) aimed at analyzing the azimuthal anisotropy and for

detecting fractures. However, these are applied to a limited offset bin sizes in the prestack data and hence suffer from low S/N. Subsequently, the aspect of considering the full-azimuth offset into consideration came into use so as to carry out anisotropic migration on basis of common offset migration (Wang et al. 2013). However, these schemes are based on assumption that the ray path can be invariantly parameterized by source-receiver offset and azimuth, which is not always the case because the ray path is continuously changing during the course of energy propagation. The anisotropic migration carried out in local angle domain was introduced by Sun and Sun (2015) and proved to be an efficient way of extracting the azimuth preserved common image gathers (CIGs), so as to retain the fracture induced amplitude responses. In case of steep dips, the spatial aliasing is quite significant and needs to be handled by increasing the spatial sampling.

Migrated Amplitude Analysis for Rock Property Determination

In the 1980s, researchers observed that careful handling of amplitude terms within migration algorithms produced migrated amplitudes that are proportional to reflection coefficient values at rock boundaries. This observation is the basis of true-amplitude migration (Bleistein 1987). Migrated amplitudes are now often analyzed in terms of angle-dependent reflection coefficients at rock boundaries. The reflection coefficients can be used to estimate the types of rock above and below each interface, and whether the rocks can bear hydrocarbons. This endeavor has had some success in areas of moderate velocity and structural complexity, but less success below areas of high complexity, such as below salt.

Imaging Conditions

The general assumptions are that the data is multiple free; the seismic velocity is known at every data sample and that there are no out of plane reflections.

Isochron Imaging

Pencil-and-paper migrations were performed by swinging reflection events to their migrated locations, which are the envelopes of all possible swing locations. Kirchhoff migration can be performed the same way, with the wave equation canceling the energy away from the envelopes and leaving only images of reflectors. The time of a sample on an input trace is the sum of time from the source to a reflector plus time from the reflector to the receiver; Kirchhoff migration acts by placing the sample at all subsurface locations (isochrons) that share the same two-way travel time.

Isochron imaging, useful in describing Kirchhoff migration, is plausible and correct, but it is not a wavefield concept,

and cannot be applied to all migrations based on the wave equation. These require more general treatment, which is different for poststack and prestack migration.

Poststack Migration: Exploding Reflector Model

The CMP stack simulates a zero-offset record, where energy has been excited and recorded at the same location. It can also be considered as a single wavefield from a far-fetched experiment. Suppose all the reflectors and diffractors inside the earth exploded at the same instant (time zero), sending energy in all directions; some of the energy would emerge at the earth's surface at later times to be recorded at the receiver locations. Suppose also that the propagation velocity of the wavefield from the exploding reflectors is exactly one-half the actual propagation velocity inside the earth. Then the recording time for an event from an exploding reflector is the same as the two-way zero-offset reflection time, with the actual velocity, from the same reflector. This thought experiment (Loewenthal et al. 1976) provides the basis for a poststack migration imaging principle, breaking migration into two parts: downward continuation and imaging. The first part expresses the wavefield at one depth in terms of the wavefield at a shallower depth. As the wavefield moves down, time decreases towards zero (the time of the explosion). The second part picks off the downward-continued wavefield at each location when the time reaches zero. The result of the process is a snapshot of the reflectors in the act of explosion.

Reflected Wavefield Amplitude Normalized by Source Wavefield Amplitude

For prestack migration, downward continuation is useful, but exploding reflectors and zero-time imaging are not. As in poststack migration, the wavefield from the many receiver locations is downward continued into the earth and backwards in time. Likewise, the wavefield from the source location is downward continued, but forward in time so that it can interact with the receiver wavefield at actual reflection times. If the two downward-continued wavefields intersect at a location at a particular time, the process has captured the instant of reflection at that point. At that location and time, the receiver wavefield amplitude equals the source wavefield amplitude times the value of the reflection coefficient at the location (possibly zero, if the location is not on an actual reflector). That is, the migrated image at that location is the receiver wavefield divided by the source wavefield.

Deconvolution and Cross-Correlation Imaging Conditions

Some migration techniques are applied in the time domain, others in the frequency (ω) domain. For frequency-domain methods, a precise statement of the prestack migration imaging principle is that reflectors exist inside the earth where downward continued source and receiver wavefields are in

phase, and the reflection strength is the ratio of the two wavefields. For a single frequency, there is complete uncertainty about the location of a reflector; only by adding the results from many frequencies can reflector positions be precisely determined. Thus, the prestack-migration imaging principle can be written as:

$$R(\mathbf{x}) = \int \frac{U(\mathbf{x}; \omega)}{D(\mathbf{x}; \omega)} d\omega, \quad (1)$$

where $R(\mathbf{x})$ is reflection strength at location $\mathbf{x}$, U is the downward-continued receiver wavefield, and D is the downward-continued source wavefield. The quotient can cause stability problems when the source wavefield D becomes weak. To overcome instability, note that if both numerator and denominator are multiplied by the complex conjugate of D , denoted D^* , the phase of the denominator will disappear (the denominator becomes a positive real number). Instability, caused by the nearly vanishing denominator DD^* , can be eliminated by replacing the denominator with a constant value, leaving the modified imaging condition:

$$R(\mathbf{x}) = \int U(\mathbf{x}; \omega) D^*(\mathbf{x}; \omega) d\omega. \quad (2)$$

Equations 1 and 2 migrate events to the same locations, but migrated amplitudes obtained using Eq. 2 are different from those obtained using Eq. 1. Imaging conditions (1) (*deconvolution*) and (2) (*cross-correlation*) were introduced by Claerbout (1971).

Migration Techniques

Integral (Kirchhoff) Migration

Kirchhoff migration is the digital embodiment of the earliest migration heuristics, and it has a basis in the downward continuation of wavefields: the wavefield at depth is an integral of the recorded wavefield (Schneider 1978). The mathematical physics of this operation requires amplitude and phase manipulations of the recorded wavefield, yielding various expressions for poststack and prestack migration with different recording configurations. For example, shot-record Kirchhoff migration is expressed as:

$$I(\mathbf{x}; \mathbf{x}_s) = \int d\mathbf{x}_r \int dt W \frac{\partial U(\mathbf{x}_r; \mathbf{x}_s; t)}{\partial t} \delta[t - (t_s + t_r)], \quad (3)$$

where $I(\mathbf{x}, \mathbf{x}_s)$ is the migrated image at location $\mathbf{x}$ due to a source at $\mathbf{x}_s$, $\mathbf{x}_r$ are the receiver locations, W is a weight function, U is the recorded wavefield, δ is the Dirac delta function, and t_s and t_r are travel times from $\mathbf{x}_s$ and $\mathbf{x}_r$ to $\mathbf{x}$.

The integral accumulates data samples from the receiver locations into the image at $\mathbf{x}$. Alternatively, Eq. 3 allows an input sample at time $t_s + t_r$ from receiver location $\mathbf{x}_r$ to be smeared out into the isochron surface.

Kirchhoff migration is flexible, allowing the accumulation of trace values from any subvolume of the recorded traces into any subvolume of the image. This has made Kirchhoff migration very popular. However, complicated velocity models used in many depth migration projects have shown its accuracy limits, which are due mostly to a reliance on asymptotic ray theory to provide travel times and amplitudes. Kirchhoff migration can be implemented in 2D and 3D data in both pre- and poststack domains.

Practical Considerations in Kirchhoff's Migration

Migration aperture: Migration aperture can be defined as the summation span of the diffraction hyperbola. In practical terms, it is defined as the number of traces involved in the computation. The shape of the hyperbola is governed by the velocity at apex and for a vertical increase in velocity, the slower velocity will have smaller aperture and the higher velocity will have wider aperture. In other words, wider apertures are needed to collapse deeper diffraction hyperbolas. Also, wider aperture means higher velocity and thereby more migration and the deeper events are migrated more than the shallow events, provided the dip is constant. The effectiveness of the migration scheme is proportional to the aperture width, i.e., wider the aperture, better the lateral resolution. On the flip side, the runtime is also directionally proportional to the aperture, meaning wider apertures will require more time for migration. Therefore, an optimum migration aperture must be used for effective migration. In general, for a specific reflection event with a constant dip θ and depth h , the aperture width d can be computed as:

$$d = h / \tan \theta$$

Too small apertures can sometimes cause total loss of steep dips, e.g., flanks of diapirs, and cause spurious subhorizontal events at deeper parts, called as “smearing.” These are especially significant at later times, where signal strength is weaker than the random noise.

Dip Limits The maximum dip can be defined in the migration scheme and is used to suppress the steeply dipping coherent noise. The flanks of theoretical hyperbolas used for summation are truncated after the allowed dip. Dip parameter is again related to the migration aperture to some extent. The narrower the migration aperture, the smaller maximum allowed dip. The effect of dip is especially significant in areas of steep dips and dip tests need to be carried out for areas of steep dip.

Velocity Field Effectiveness of the migration undisputedly depends on the velocity field used, irrespective of the migration algorithm. A slower velocity results in under migration and “frowns” occur in the final image. The inclined events are not relocated to their accurate lateral positions and bow-tie effects remain unresolved. Inversely, higher velocity causes over migration and “smiles” occur in the final image. The inclined events are migrated too much in an updip direction.

Beam Migration

In its simplest form, beam migration is a Kirchhoff migration with the integral over the recording surface broken into two stages: first, the domain of integration is divided into a number of overlapping regions; second, a directional decomposition of the data within each region is performed. This decomposition produces a set of local plane waves, over a range of angles, emerging within each region. Each of these local-plane wave data components is mapped back into the earth using raypaths sent in the directions of the emerging plane waves. A more complete description (Hill 2001) uses Green's identity to downward continue the wavefields, with Gaussian beams as Green's functions. Gray et al. (2009) provide an elementary physical explanation of this complicated method, which is generally more accurate than Kirchhoff migration because it naturally allows several arrivals from each wavefield to accumulate at any image location (which is difficult for Kirchhoff migration). It has many of the flexibilities of Kirchhoff migration, but, like Kirchhoff migration, relies on asymptotic ray theory.

One-Way Wave-Equation Migration (OWEM)

Strictly speaking, applying the term “wave-equation migration” solely to migration methods that downward-continue wavefields recursively one depth step at a time is an abuse of terminology. This usage excludes Kirchhoff and beam migrations, implying that they do not derive from the wave equation, which is not true. This usage has become entrenched.

Downward continuing a wavefield from one depth to the next requires that the wavefield is propagating either up or down, but not both at once. (The wavefield can propagate laterally, but up or down is a distinguished direction.) Further, it requires an appropriate (upgoing or downgoing) wave equation. In the simple constant-velocity case, the one-way solution can be obtained easily. The two-way wave equation for wavefield P (acoustic, ignoring elastic effects such as mode conversion) is:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) P = 0. \quad (4)$$

When this equation is expressed in the frequency/horizontal wave number domain, it factors into separate equations for upgoing and downgoing wavefields. Analytical solutions of

these equations provide phase shifts that perform downward continuation. Once the downward-continued wavefields have been inverse Fourier transformed from wave number back to space, they can be combined and summed over frequency using Eq. 1 or 2 to produce the migrated image (Gazdag 1978). This method, phase-shift migration, can accurately image dips up to 90° . Its only limitation is the lateral invariance of the velocity field. It has proven successful in imaging salt flanks in marine sedimentary basins. Beneath the salt flanks, where significant velocity contrast has occurred, the phase-shift image is inaccurate.

Factoring the wave equation exactly into one-way equations works only for constant velocity. When velocity varies laterally, the factoring, again with z as a distinguished direction, is approximate, incurring kinematic (time) and dynamic (amplitude) errors. A number of methods downward continue wavefields using approximate one-way wave equations. The earliest of these used finite differences to discretize the equations, using low-angle approximations (e.g., Claerbout 1970). Later methods generalized phase-shift migration to handle lateral velocity variations, and other methods combine aspects of finite-difference and phase-shift approaches. Development of these methods continues today. Because of the errors inherent in even the best one-way wave equations, none of the methods can be completely accurate.

Finite-Difference Migration

As the name indicates, the finite-difference migration is based on the numerical solution of the scalar wave equation by the finite-difference method. Its theoretical basis consists in downward continuation of the recorded wavefield, by moving the receivers downward in finite depth steps and by shortening the spread length with each step. This process will shorten the diffraction hyperbolas, and when the receiver arrives at the apex, the hyperbola is completely collapsed to a single point. An undermigration occurs if the downward continuation is completed before the source depth is reached and the hyperbola cannot be collapsed perfectly. The same situation also occurs when a slower velocity is used for the migration. Inversely, if the downward continuation is continued beyond the source depth, the mirror image of diffraction hyperbola is formed. This also happens if a faster velocity is used for migration.

Depth Step The minimum depth step size can be the same as the sampling rate of the input seismic data. Higher depth step results in undermigration and lower depth steps result in overmigration. Undermigration can sometimes lead to specific type of noise (Yilmaz 2001) called “dispersion,” along the flanks of the diffraction hyperbola. Large depth steps are incapable of collapsing the diffraction events and the dipping events, thereby are not moved to their correct lateral

position, and this becomes more prevalent if there are updipping event. Smaller depth steps are usually preferred in steeply dipping events as they are capable of producing better results. In practice, the value for depth step is between T_s and $0.5 T_s$, where T_s is the dominant period of the input dataset.

Frequency–Wave Number (f-k) Migration

In this scheme, the migration is implemented using the Fourier transform in f-k domain and also using the phase shift method in Fourier domain. The former is called Stolt migration, and the later is called Gazdag migration, in the name of their proposers. Of these, the Stolt migration does not handle the velocity variations in the horizontal and vertical directions and instead requires a constant velocity medium. Therefore, a modification is made to the method where a constant called “Stolt stretch factor” is defined so that the algorithm can be applicable to areas with complex geology, wherein the velocity varies both laterally and vertically. Before migration, the vertical axis in the f-k domain is the frequency that is transformed into wave number. Stolt migration is by far the fastest method and is mostly used in QC purposes.

Gazdag migration also works on the principle of downward continuation of the recorded wave field, like finite-difference method, but corresponds to a phase shift in the f-k domain.

Stolt Stretch Factor The crux of applying a stretch factor lies in the fact that it transforms the input data into a constant velocity section, inducing a stretch to the seismic data along the temporal axis. Subsequently, an inverse stretch is applied and final migrated section is obtained after inverse Fourier transform. The stretch factor also partially controls the aperture of migration and changes between 0 and 2. The amount of stretch depends on the ratio between the V_{rms} of the time sample and the smallest value of RMS velocity for the entire dataset. For constant velocity media, the stretch factor equals 1 and if it is greater than 1, then it indicates overmigration and less than 1 indicates undermigration.

A special case of this class of migration is the zero-order WKB approximation, which can be used in the datasets that involve both upgoing and downgoing wavefields, e.g., ocean bottom seismometer data, vertical seismic profiling data, etc. The set of equations governing the (f-k) migration are transformed to give new set of equations that govern the interaction between the up- and downgoing wave fields (Ursin 1984). The zero-order WKB approximation (consisting of neglecting all interaction between the down- and upgoing wave fields) is used to downward extrapolate both the downgoing wavefield from the source and the recorded upgoing wavefield. The ratio between the up- and the downgoing wave fields is used to compute the depth-migrated seismic trace.

Two-Way Wave-Equation Migration (Reverse-Time Migration)

Reverse-time migration (RTM) uses the full two-way wave Eq. 4. It does not operate by downward continuation of downgoing or upgoing wavefields. Instead, it takes the source and receiver wavefields at the recording surface and injects them into the earth by treating them as boundary values for the wave equation (McMechan 1983), computing full wavefields inside the earth at each time step. The source wavefield is computed forwards in time, so successive wavefield snapshots show an expanding wavefront moving into the earth and interacting with reflectors and diffractors. The receiver wavefield is computed backwards in time (beginning with the final time sample on each trace of the recorded data), so successive snapshots show reflection events moving generally downward. Imaging is performed using a temporal version of the imaging condition (1) or (2). In principle, RTM is the most accurate migration method: it does not rely on asymptotic ray theory, and it suffers from neither the steep-dip nor the kinematic and dynamic errors of OWEM. The source and receiver wavefield are reconstructed forward and backward separately at finite time steps. At each time step, the migrated image is achieved by cross-correlating the two extrapolated wave fields. The caveat is that the time steps selected must be small, so as to ensure numerical stability. Until recently, its use has been limited because it tends to be slower than other migration methods, and numerical approximations to the wave equation (e.g., low-order finite-difference approximations) produced errors (e.g., numerical dispersion) that other methods do not incur.

To overcome this drawback, RTM is implemented in time-wave number (t - k) domain, where it is relatively faster, but accounts only for vertical velocity variations. However, modern computational hardware has both the speed and the memory availability to allow very good approximations to the wave equation, and high-quality RTM images are now the norm. RTM is especially effective in areas of complex geology with steep dips, such as salt flanks, that are not sufficiently illuminated by primary reflections.

RTM can be effectively reformulated (Zhan et al. 2014) so that it can be used as a tool to filter the coherent noise. In this scheme, the migration image at a point is computed by summing the trace amplitudes over all the scattered events which in turn leads to filters that can be applied to mitigate the coherent noise and associated artifacts, like crosstalk and aliasing. However, the main drawback is that it is more memory intensive and input dependent than the conventional RTM.

Dispersion in seismic waves caused by attenuation generally degrades the resolution of the migrated images. The method of compensating for the attenuation and then applying RTM is called Q-RTM, and it effectively mitigates the amplitude/phase dispersion effects. The dispersion and attenuation

compensated operators are constructed by extrapolating the time-domain visco-acoustic wave equation (Zhu et al. 2014). The results obtained showed marked improvement in the amplitude and phase of the seismic events compared to the conventional RTM.

In a nutshell, the various types of migration can be grouped into three categories, each solving for scalar wave equation using different approaches:

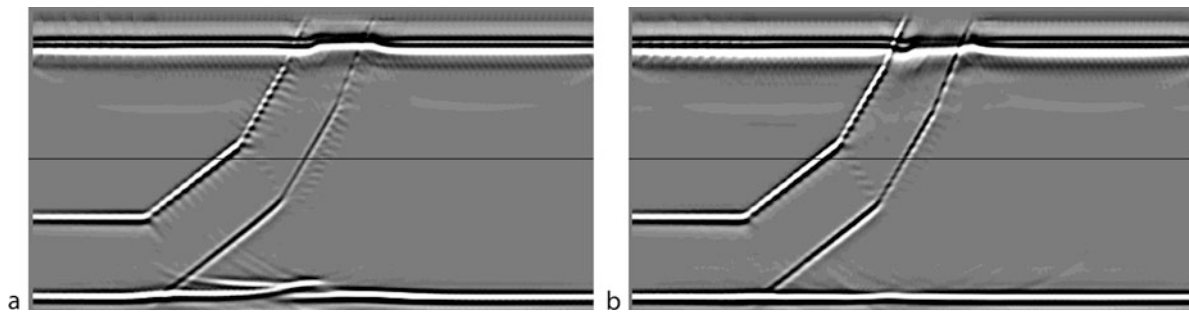
- Integral or diffraction summation methods (Schneider 1978) – can handle structural dips up to ~ 90 degrees; migration cannot handle lateral velocity variations.
- Finite-difference methods (Claerbout and Doherty 1972) – can handle structural dips up to ~ 35 degrees only; migration is not affected by lateral velocity variations.
- Fourier transform methods (Stolt 1978; Gazdag 1978) – can handle dips up to structural 90 degrees; migration is not affected by horizontal and vertical velocity variations.
- Reverse-time migration (Baysal et al. 1983) – very effective in imaging the geologically complex areas.

Migration Examples

There are many reasons to migrate seismic data; we will need to be content with a few illustrations.

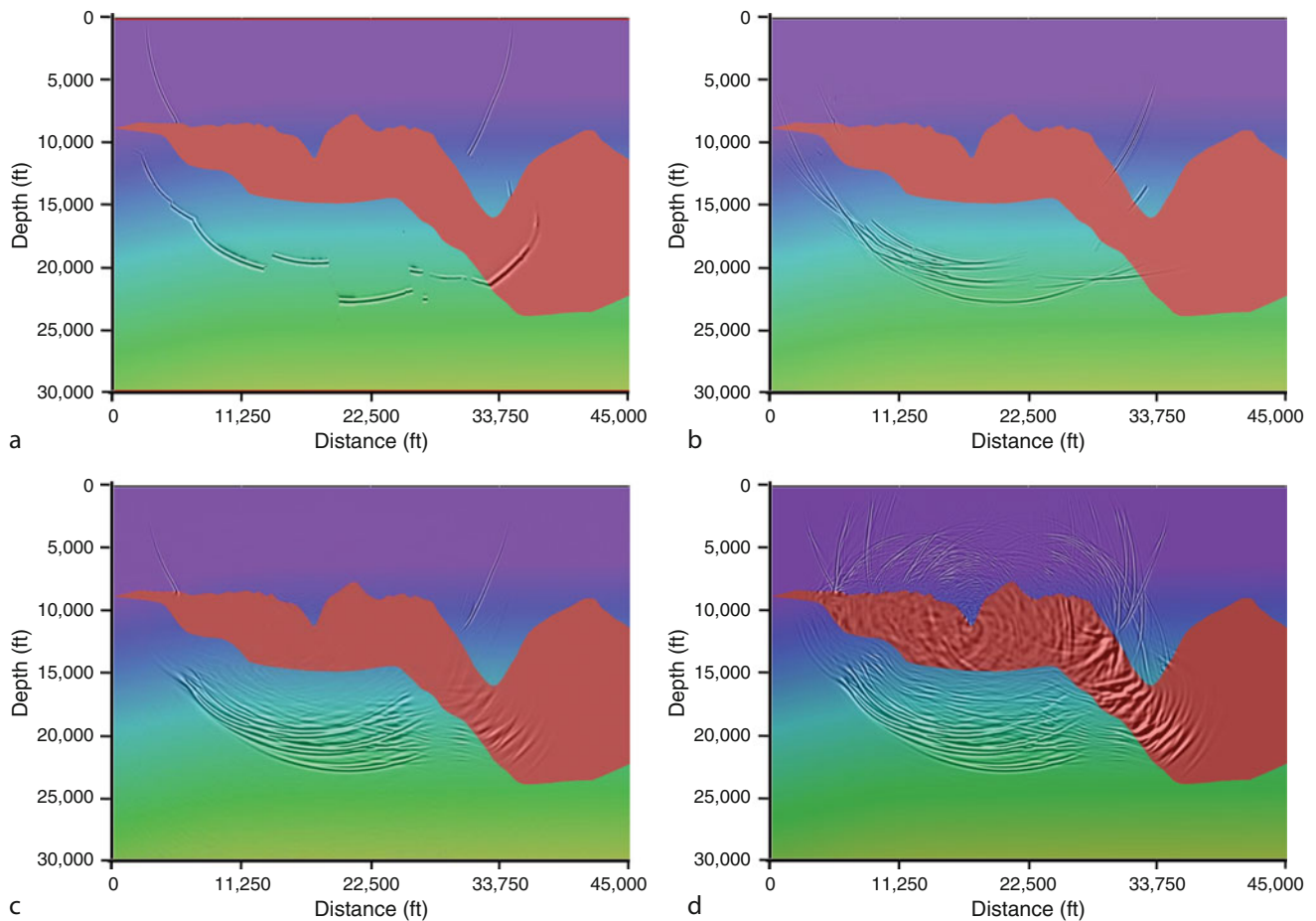
The first illustrates structural imaging in the presence of anisotropy. Fig. 4 shows two images of a cross section designed to mimic the structural geology of a dipping thrust sheet, simplified to isolate the effect of anisotropy on image positioning and quality. The velocity is generally simple, with a constant velocity of 2740 m/s outside the thrust sheet. The velocity complication lies in the thrust sheet, which is anisotropic (TTI), with a symmetry axis parallel to the boundaries of the sheet. In the thrust sheet, the velocity is 2925 m/s in the direction of the symmetry axis, and higher in all other directions. Both images in Fig. 4 are from prestack beam migration of the same input data set, which was generated by finite-difference elastic modeling (courtesy BP). In Fig. 4a, anisotropy was ignored, and in Fig. 4b, anisotropy was taken into account. The reflector below the dipping thrust sheet should be flat, as it is in Fig. 4b. In Fig. 4a, the flat reflector is pulled up beneath the thrust sheet and is poorly imaged. The high-amplitude flat event cutting across the entire section at shallow depths is a shear-wave artifact that has been mis-migrated by the P-wave migration program.

The second example illustrates the action of various migration methods: Kirchhoff, beam, OWEM, and RTM. Figure 5 shows four images overlaid on a geologic cross section, each obtained by migrating a single event on a single zero-offset trace through the complicated geology/velocity model shown (called Sigsbee2a). A high-velocity salt body of complex



Seismic, Migration, Fig. 4 Prestack depth migrated images from a synthetic thrust model data set. In (a), the anisotropy in the dipping thrust sheet is ignored by the migration, leading to inaccurate imaging below

the thrust sheet. In (b), the migration has accounted for the anisotropy, leading to correct structural imaging



Seismic, Migration, Fig. 5 The action of migration, illustrated on four different migration methods: (a) Kirchhoff; (b) beam; (c) one-way wave equation; (d) reverse time. In each, a single event from a single zero-

offset trace was migrated. The migration methods have interacted differently with the velocity model, producing different wavefronts. (From Etgen et al. 2009)

shape has been inserted into a set of lower-velocity sedimentary layers, and we note the effect of the geometry and velocity on the migration algorithms. The migrated images are, in effect, wavefronts modeled by the migration methods; full migrated images are formed by adding together such partial images from all samples on all traces. The Kirchhoff migrated image (Fig. 5a) shows the effect of the assumption,

made by most Kirchhoff migration programs, that at most one travel path for seismic energy connects a given upper-surface location with each subsurface location. The wavefield discontinuities in the subsalt area are physically impossible and will cause inaccurate, noisy images in areas of structural complexity. The remaining images show wavefields of greater continuity and complexity in the subsalt area. The most

complicated wavefield, produced by RTM, should produce the most accurate image. This happens in practice when the velocity field is accurate, and the input traces are free from unwanted noise such as elastic-wave mode-converted events.

The Evolving Role of Migration in Seismic Data Processing

Before seismic processing became digital, migration was a standalone process. Pencil-and-paper or mechanical migration provided a cross section that showed where to drill. Later, poststack migration was the final seismic processing step; all other processing was intended to prepare a stacked data set for migration. With prestack migration, migration has moved to a central location in the processing flow. In tomography, it is common to process the signals on a migrated stack or in a set of CIG's. In postmigration amplitude analysis, it is common to analyze and alter the migrated waveforms. In the future, when the wave equation used for migration can model earth effects more completely than at present, seismic data will undergo less processing before migration, and the migrated data will predict rock properties directly.

Selection of Migration Algorithm

Ideally, the migration algorithm should be able to handle full range of dips; preserve amplitude and phase information without generating artifacts and should be computationally efficient and fast. Poststack time migration is the simplest one and can be applied to the regions of less geological complexity, velocity variation and gentle dips. In case of greater

geological complexity, prestack time of depth migration should be applied. In both the types of migrations, the velocity plays a vital role. In areas typical of salt tectonics, the depth imaging has to be preferred. In such cases, RTM, although costly, provides good results. The most common algorithm used by industry is the Kirchhoff migration, because the shortcomings, the advantages, and restrictions of it are well understood. A comparative statement of the various migration algorithm, their applicability and computational simplicity are presented in Table 1.

Some Aspects of Migration

- Input data should be properly conditioned before migration. Missing traces/irregular fold distribution/spiky data induces artifacts in migration. So also, the input velocity should be properly interpolated and regularized before migration.
- Spatial aliasing is of two types, i.e., data aliasing and operator aliasing. Data aliasing is caused due to coarse trace spacing in the input data. Operator aliasing is caused due to the coarse sampling of hyperboloid (3D) or hyperbola (2D). Data aliasing is more pronounced in dipping events and the aliased energy is located at the negative panel in the f-k spectrum and migration treat them as reverse dipping events, causing aliased amplitudes. The distance between CDP inputs directly effects the migration output. Finer intervals give higher resolution and sharper migrated images. Data interpolation is generally required in 3D data, especially in crossline direction so as to have properly migrated images. Data aliasing can be solved either by data interpolation or by using the anti-aliasing

Seismic, Migration, Table 1 Comparative statement of migration algorithm, applicability, and computational simplicity

Type of migration	Applicability	Simplicity level (1–4, 4 is most complex)
Kirchhoff (prestack or poststack)	• Vertical and horizontal velocity variations • Steep dips ($< 80^\circ$) • Widely used • Might produce horizontal events at depth	3 (especially in depth)
Finite difference (prestack or poststack)	• Dips up to 35° • Large depth step might cause kinks, dispersion, and undermigration • Less sensitive to velocity errors	2
Stolt (only in poststack)	• Computationally efficient • Constant velocity medium • Steep dips up to 90° • Good for QC	1
Phase shift (only in poststack)	• Valid only for vertical velocity variations • Low noise in output • Steep dips up to 90°	1
RTM (prestack or poststack)	• Promising algorithm for less illuminated zones • Complex geology • Valid for vertical and horizontal velocity variations	4

filter. The operator aliasing can be avoided by removing high-frequency components.

- Borders of the input seismic line must be treated carefully, since there is no place for updip replacement of dipping events located close to the borderline. The section should be expanded in both directions by incorporating the zero-amplitude traces at the end of the data traces. This process is called zero padding.
- Tapering also needs to be carefully selected before migration in order to avoid the edge effects due to the migration aperture. Conical taper zones (narrow at time zero and wider at maximum time) can be considered since the diffraction hyperbolas (and aperture) get wider with an increase in depth. However, unnecessarily wide taper zones might induce noise in migration output.
- Mute zones have zero amplitude and cause a characteristic noise called as swings. These effects are mostly seen in the beginning of the line and also in zones which have been muted as part of trace kill, surgical mute, irregular data acquisition, etc., and hence it is important to reapply mute after migration to avoid swings.
- Migration should be carried out only in seismic lines that are long enough. Migration on short lines could be disastrous as the entire data would be consumed by edge effects.
- Spikes or the high amplitude transient noise bursts present along the seismic line at regular intervals also lead to the migration output that is completely degraded. Every single spike causes a distinct smile and in such case, it is desired to go back to preprocessing stage and remove the spike.
- Migration algorithms, although very robust, still depend a lot on the interval velocity model. Poststack migrations are less sensitive to velocity errors, but the prestack migrations, especially depth migrations are highly degraded by improper velocity input. To reduce the effect of localized velocity errors, it is preferred to smooth the RMS velocity field both in space and time before migration.
- Near surface structures may result in velocity issues in the deeper parts of the data, e.g., mass transport deposits in the shallow subsurface/near surface channel filled with low/high velocity sediment results in vertical disturbances in the deeper parts, such as velocity pull-up or push-down in deeper section. In such cases, tomography velocity model maybe used instead of the RMS velocity model.

Summary

Seismic migration has been used in prospecting for oil and natural gas since the 1920s. Since the 1970s, migration has been applied as a wave-equation process, explicitly recognizing that reflections recorded at the earth's surface are the result of wavefields propagating and reflecting inside the earth. Wavefield migration has taken several forms (time-domain,

frequency-domain; finite-difference, integral) because of the variety of ways for expressing the propagation of wavefields. All these different forms of migration are still in use because each of them has features that the others lack: flexibility, steep-dip capability, etc. When computer power was limited, migration was usually performed after NMO/stack, allowing a reduction of the size of the data volume input to migration. Nowadays most migrations are performed before stack in 3-D. Another shift has been from time to depth, bringing migration closer to the problem of estimating seismic velocities inside the earth. A separate development has been the capability of migration to preserve amplitudes for estimating rock properties near reflector locations. With migration velocity estimation and amplitude analysis, migration has moved from the final step of the seismic processing flow to a more central role. This means that an increasing amount of processing and analysis is performed on migrated gathers before they are stacked to form the final structural image.

Cross-References

- [Seismic Data Acquisition and Processing](#)
- [Seismic Diffraction](#)
- [Seismic Imaging, Overview](#)
- [Seismic Properties of Rocks](#)
- [Seismic Ray Theory](#)
- [Seismic Waves, Scattering](#)
- [Seismic, Reflectivity Method](#)
- [Seismic, Waveform Modeling and Tomography](#)
- [Traveltime Tomography Using Controlled-Source Seismic Data](#)

Bibliography

- Baysal E, Kosloff DD, Sherwood JWC (1983) Reverse-time migration. *Geophysics* 48:1514–1524
- Berkhout AJ (1982) *Seismic migration – imaging of acoustic energy by wavefield extrapolation*. Elsevier, Amsterdam
- Bleistein N (1987) On the imaging of reflectors in the earth. *Geophysics* 52:931–942
- Claerbout J (1970) Coarse-grid calculations of waves in inhomogeneous media with application to delineation of complicated seismic structure. *Geophysics* 35:407–418
- Claerbout J (1971) Toward a unified theory of reflector mapping. *Geophysics* 36:467–481
- Claerbout JF, Doherty SM (1972) Downward continuation of moveout-corrected seismograms. *Geophysics* 37:741–768
- Dell S, Gajewski D, Vanelle C (2012) Prestack time migration by common-migrated-reflector-element stacking. *Geophysics* 77(3): S73–S82
- Dix CH (1955) Seismic velocities from surface measurements. *Geophysics* 20:68–86
- Dondurur D (2018) *Seismic migration in acquisition and processing of marine seismic data*. Elsevier Publications

- Dong C, Wang S, Zhang J, Ma J, Zhang H (2019) Automatic migration velocity estimation for pre stack time migration. *Geophysics* 84(3): U1–U11
- Etgen J, Gray SH, Zhang Y (2009) An overview of depth migration in exploration geophysics. *Geophysics* 74:WCA5–WCA18
- Gazdag J (1978) Wave equation migration with the phase-shift method. *Geophysics* 43:1342–1351
- Gray SH, Xie Y, Nottfors C, Zhu T, Wang D, Ting C-O (2009) Taking apart beam migration. *Lead Edge* 28:1098–1109
- Hagedoorn JG (1954) A process of seismic reflection interpretation. *Geophys Prospect* 6:449–453
- Hill NR (2001) Prestack Gaussian-beam depth migration. *Geophysics* 66:1240–1250
- Loewenthal D, Roberson R, Sherwood J, Lu L (1976) The wave equation applied to migration. *Geophys Prospect* 24:380–399
- Mayne WH (1962) Common reflection point horizontal data stacking techniques. *Geophysics* 27:927–938
- McMechan GA (1983) Migration by extrapolation of time-dependent boundary values. *Geophys Prospect* 31:413–420
- Peng LL, Sun SZ, Zhang YY, Yang HJ (2013) Fracture prediction and its fluid identification of complex carbonate reservoir: A case study in Ha7 area, Tarim Basin. In: 83rd annual international meeting, SEG, Expanded abstracts, pp 3247–3251
- Schneider WA (1978) Integral formulation for migration in two and three dimensions. *Geophysics* 43:49–76
- Stolt RH (1978) Migration by Fourier transform. *Geophysics* 43:23–48
- Sun X, Sun ZS (2015) Full azimuth anisotropic prestack time migration in local angle domain and its application on fracture detection. *Geophysics* 80(2):C37–C47
- Trorey AW (1970) A simple theory of seismic diffractions. *Geophysics* 35:762–784
- Ursin (1984) Seismic migration using the WKB approximation. *Geophys J Int* 79(1):339–352
- Wang D, Sun SZ, Zhou J, Wang H, Yue, Liu Y (2013) 3D amplitude-preserved full azimuth anisotropic imaging: offering more reliable azimuthal AVO for fracture detection. In: 75 annual international conference and exhibition, EAGE incorporating SPE EUROPEC 2013, Extended abstracts, P1305
- Warner M (1987) Migration – why doesn't it work for deep continental data. *Geophys J R Astron Soc* 89:21–26
- Yilmaz O (1987) Seismic data processing. SEG, Tulsa
- Yilmaz O (2001) Seismic data analysis, processing, inversion and interpretation of seismic data. In: *Investigations in geophysics*, 2nd edn. Society of Exploration Geophysicists, Vol. I, Tulsa
- Zhan G, Wei D, Zhou M, Schuster GT (2014) Generalized diffraction stack migration and filtering of coherent noise. *Geophys Prospect* 62(3):1–23
- Zhu T, Harris JM, Biondi B (2014) Q-compensated reverse time migration. *Geophysics* 79(3):1MJ–Z84

Seismic, Receiver Function Technique

Rainer Kind^{1,2} and Xiaohui Yuan¹

¹Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany

²Freie Universität, Fachrichtung Geophysik, Berlin, Germany

Synonyms

Receiver functions; Scattered teleseismic waves

Definition

Receiver function. Response of the Earth's structure below seismic stations to incident teleseismic waves. Seismic discontinuities in the Earth's crust or upper mantle cause waves from distant earthquakes to be converted from P to S waves or vice versa or to be multiply reflected between discontinuities before arriving the stations. Such scattered waves carry information about the seismic structure of the lithosphere and upper mantle. Scattered waves are weak signals, and summation of many records is usually required to visualize the useful information.

Receiver Function Technique

If a seismic wave hits a discontinuity between two solid materials with different physical parameters, a part of the wave will be reflected, and another part will be transmitted. There is also mode conversion between compressional (P) waves and shear (S) waves (see ► [“Energy Partitioning of Seismic Waves”](#)). Therefore, a P wave crossing a discontinuity will generate an S wave (called Ps), which will follow the P wave with a slower speed; and an S wave will produce a P wave (called Sp), which will run ahead of the S wave with a faster speed. If these wave types are recorded at the surface and their incidence angles and the material velocities are known, the depth of the discontinuity, where they are generated, can be determined from the differential times of the two seismic phases, mother phase and daughter phase. The waveforms of both phases carry information about the structure of the discontinuity, which can be isolated by the receiver function technique. The ray path of such waves is shown in Fig. 1. Figure 2 compares incident P and S phases for a simple model with two discontinuities (Moho and lithosphere–asthenosphere boundary, LAB; see ► [“Earth's Structure, Core”](#); ► [“Lithosphere, Oceanic: Thermal Structure”](#)), which cause converted and multiply reflected phases that lead finally to the receiver functions. There are a number of processing steps required in the receiver function technique, which will be considered in the following.

Separation of P and S Waves

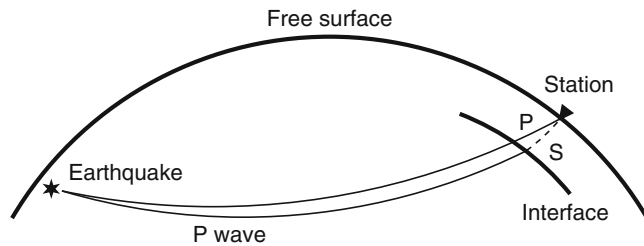
Since the converted daughter phase is a different wave type with different particle motion, it can be separated from the mother phase. The usual vertical, north–south, and east–west (ZNE) components of a seismic record should be rotated, for better identification of signals, into a coordinate system where P, SV, and SH wave types are on different components. Due to heterogeneities close to the station, such a coordinate system (frequently called LQT or P-SV-SH system; see Fig. 3)

depends on the local angle of incidence and back azimuth. In many cases, only the horizontal components N and E are rotated into the radial (R) and transverse (T) components. Frequently the rotation is done according to the theoretical back azimuth between receiver and source and the theoretical angle of incidence according to epicentral distance. A local ray coordinate system may be determined from the diagonalization of the covariance matrix of the ZNE recordings within a certain time window (Montalbetti and Kanasevich 1970;

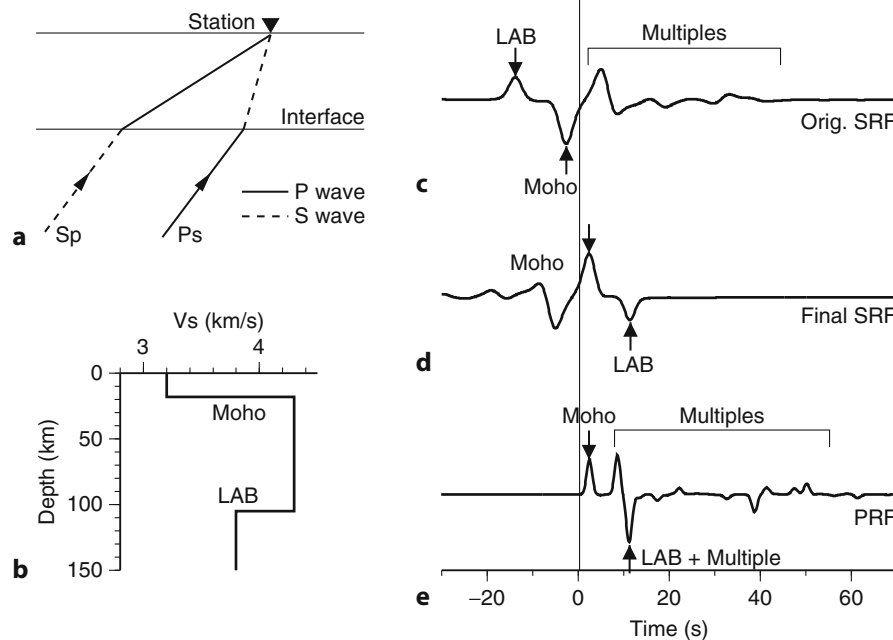
Kind and Vinnik 1988). The length of the time window used plays an important role for determination of rotation angles with this technique. The influence of the free surface of the Earth is small and in most cases not considered. In Fig. 4, theoretical seismograms in the ZR and LQ coordinate system are shown, along with ray diagrams for direct Ps phase and multiple phases PpPs and PpSs.

Deconvolution

After separation of P and SV wave types on separate components, it is much easier to identify and interpret the small converted signals. In earlier times when only analog data records were available, only a few records could be searched for converted phases. With the large amount of digital data now available (see ► “Seismological Networks”), weak signals can be enlarged relatively easily by summation of many records. A number of problems need to be considered before seismic records can be summed. Different waveforms and amplitudes generated by different earthquakes are another problem, which needs to be taken into account before summation of many traces can be applied. After rotation, records of one source and many distributed stations may be summed because all records are caused by a single event with the same source

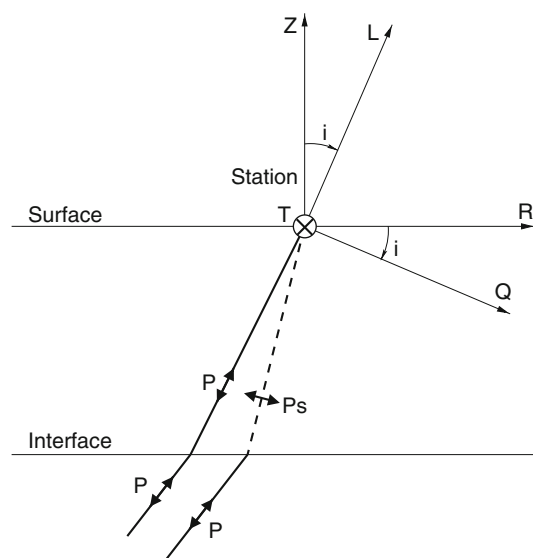


Seismic, Receiver Function Technique, Fig. 1 P receiver function geometry. The target zone below the three-component receiver is illuminated from below. The source is an earthquake at a large epicentral distance. When a plane P wave encounters interfaces below the receiver, Ps waves are generated. The incidence angle of the P wave is always bigger than that of the Ps waves. The travel path of S receiver functions (SRFs) is very similar, except that the incidence angle of the S wave is smaller than that of the Sp wave



Seismic, Receiver Function Technique, Fig. 2 (a) Sketch showing ray paths of teleseismic Ps and Sp converted waves below a station. Note that the piercing point of Sp is offset farther away from the station than that of Ps. (b) Velocity model for the calculation of P and S receiver functions. (c) Theoretical S receiver function (SRF), primary conversions arrive prior to S (which arrives at zero time and is omitted here), and crustal multiples arrive after S. (d) S receiver function (trace c) after

reversal of timescale and amplitude for better comparison with P receiver functions (PRFs). (e) P receiver functions (PRFs) of model in (b). The main difference of P and S receiver functions is that S receiver functions are free of crustal multiples at the time of the lithosphere–asthenosphere boundary (LAB) arrivals. Note that in the case of a homogeneous half-space, no signal would be contained in the above traces (Yuan et al. 2007)



Seismic, Receiver Function Technique, Fig. 3 Rotation of the radial (R) and vertical (Z) components around the angle of incidence into the L and Q system (Vinnik 1977). The P wave is in this case only on the L component, and the SV wave is only on the Q component. The SH wave remains on the T component

time function. This is the well-known delay and sum technique. However the results of such a summation will be the response function at the source region (including, e.g., depth phases) and not the “receiver function.” If we are interested in the structure beneath a seismic receiver, records from many earthquakes at this receiver with very different source time functions and magnitudes must be summed. Only in this case we will obtain as the result of the summation the response of the structure beneath a receiver. Influences of the receiver structure are common to all records and will be enhanced by summation. In contrast, influences of the source or the source structure (depth phases, conversions near the source) will be reduced. Obtaining the structure beneath a receiver is especially interesting because in contrast to sources, we can move the receivers where we need them and record teleseismic events to study the structure at nearly any location (with many problems at ocean bottom stations). The procedure used for source equalization is usually deconvolution. In the frequency domain, deconvolution of P receiver functions (PRFs) means division of the SV component by the P component, and in S receiver functions (SRFs), it means division of the P component by the SV component (Langston 1979; Gurrola et al. 1995). In time domain, a Wiener filtering approach (Kind et al. 1995) can be applied. For P receiver functions, an inverse filter can be generated by minimizing the least-square difference between the observed P waveform on the vertical component and the desired delta-like spike function. The inverse filter is then convolved with other components to obtain receiver functions. Ligorria and Ammon (1999) proposed an alternative approach to iteratively remove the source and

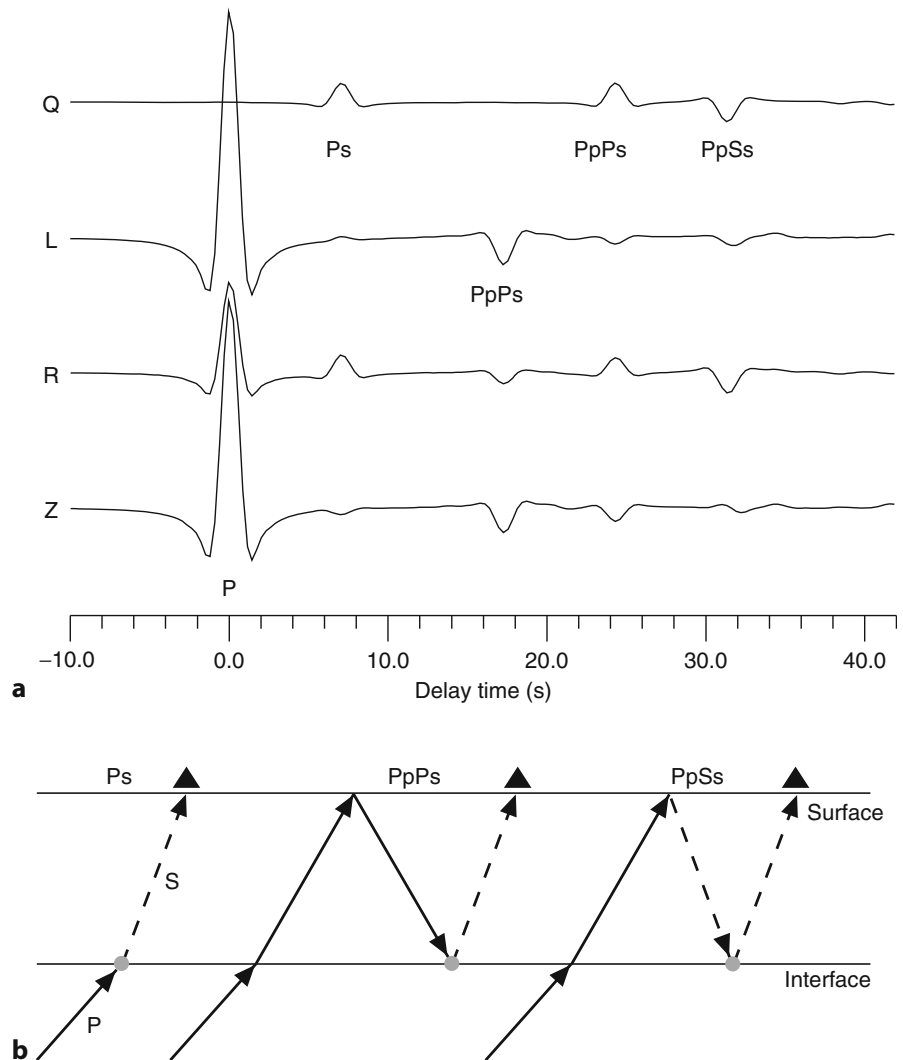
propagation effects. The question which type of deconvolution is carried out does not seem to be very significant. After deconvolution, amplitude normalization is applied, resulting in amplitude values of the converted phase in percent of the incident phase. Deconvolution, however, has one principal disadvantage. One entire component is considered as the source time function. For example, the entire P component over duration of perhaps 100 s is deconvolved from the SV component. This means all signals within this window on the P component are considered source. For example, P multiples within the crust at the receiver site will also be eliminated by this kind of deconvolution. In this sense, deconvolution prevents computation of the complete Green’s function (impulse response) at the receiver site (Kumar et al. 2010; Baig et al. 2005; Bostock 2004; Langston and Hammer 2001). The modification of seismic records caused by deconvolution is demonstrated in Fig. 5. Kumar et al. (2010) have shown that plain summation of many records of a seismic station results in SV component traces, which are nearly identical with deconvolved traces and P component traces are preserved. They only applied amplitude and sign equalization and summed all traces aligned along the maximum of the P signal (see Kumar et al. 2010). The only disadvantage of the new technique is that more traces are needed to obtain the same signal-to-noise ratio.

Moveout Correction and Summation

An important problem is that seismic phases travel with different velocities. This means that the differential times of seismic phases in one record depend on epicentral distances. Therefore, only records from similar epicentral distances can be summed. Traces from different distances may also be summed if the conversions arrive very closely in time to their mother phase (perhaps still possible for crustal conversions). For converted phases arriving later from greater depths, it is necessary to consider effects of their different slowness. Vinnik (1977) has solved this problem by using a delay and sum technique known from array processing and applying it to the upper mantle discontinuities. Results are displayed in a slowness-time plot (or conversion-depth-time plot; see Fig. 6). Another solution is that the distance moveout correction procedure could be applied before summation, which is very well known in applied seismics (see ▶ “[Deep Seismic Reflection and Refraction Profiling](#)”). In moveout correction, the timescale is stretched or compressed in order to parallelize the same type of travel time curves. A fixed reference slowness of 6.4 s/degree (or 67° epicentral distance) is frequently used at which the timescale remains unchanged. At smaller distances the timescale is compressed and expanded at larger distances, in order to transform the travel time curve of the converted phase under investigation into one parallel to the mother signal. After moveout correction, a

Seismic, Receiver Function

Technique, Fig. 4 (a) Effects of rotation of theoretical Z and R coordinates into the LQ system. Phases arriving as S wave (Ps, PpPs, PpSs) at the station remain on the Q component only. Phases that arrive as P waves remain on the L component (P, PpPp). (b) Ray paths of crustal multiples

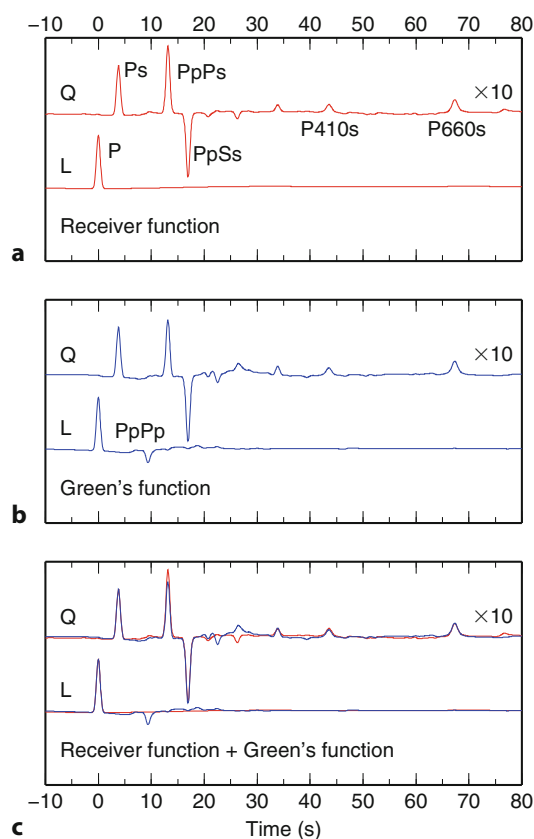


number of traces may be lined up and summed into a single trace. A group of traces used for summation could be all records from either one station or records with piercing points in a certain geographical region (box) at a certain depth (common conversion point technique, Dueker and Sheehan 1997; see Fig. 7). The advantage of the moveout correction and summation technique is that single traces representing groups of traces can be easily compared. After moveout correction for direct conversions, the summation traces contain only such phases. Surface multiples are destructively superimposed because of their different slowness. Moveout corrections cannot only be applied to Ps or Sp conversions but also to other phases like multiples. In this case, the summation trace contains only a certain type of multiples (see Fig. 6). Moveout correction and summation also solve a special problem in S receiver functions. Not every precursor of S on the P component is an S-to-P conversion. Surface multiples of the

P phase can also arrive on the vertical component as precursors of S (Bock 1994). These phases have, fortunately, a different slowness and are suppressed in the delay and sum technique or the moveout correction and sum technique.

CCP Stack and Migration

Stacking by single stations is meaningful if piercing points from different stations are not overlapping in the depth range studied. If several stations are closely spaced and piercing points are overlapping at the study depth, summation by common regions of piercing points at that depth may be more meaningful (common conversion point stacking, CCP). In this case, piercing point locations for conversions at a defined depth are computed, and all traces with piercing points inside a defined area are summed (see Fig. 7).



Seismic, Receiver Function Technique, Fig. 5 (a) Deconvolved, (b) original theoretical seismograms (Kind 1985), and (c) the overlay of both for a simple crust–mantle model. All phases except P are eliminated on the deconvolved P component. The deconvolved SV (Q) component differs significantly from the theoretical SV component. The $P_M P$ phase has disappeared on the deconvolved SV component, as expected. However, the PpPs Moho multiple reflection has a significantly distorted amplitude (Kumar et al. 2010)

Stacked time domain receiver functions are not yet the goal of the processing, even if clear converted phases can be recognized. We need to know where in the three-dimensional Earth the conversion occurs. Therefore, the final goal is the migration from the time series into a depth model (see ► “Seismic, Migration”). Depth migration with a known velocity depth model is necessary for this purpose. Migration is a very well-known technique in controlled source seismics. In earthquake seismology with relatively sparse station distribution, more simple versions of migration with one-dimensional models are often used. The amplitudes of the receiver function traces are distributed along the ray path of a known velocity depth model. The Earth model is divided in boxes, and amplitudes of all rays in one box are summed. The result is an approximate two-dimensional (or three-dimensional) distribution of seismic energy in space where hopefully seismic discontinuities can be identified (see Fig. 8). In the case of P receiver functions, such images are blurred with multiples from the surface, which produce

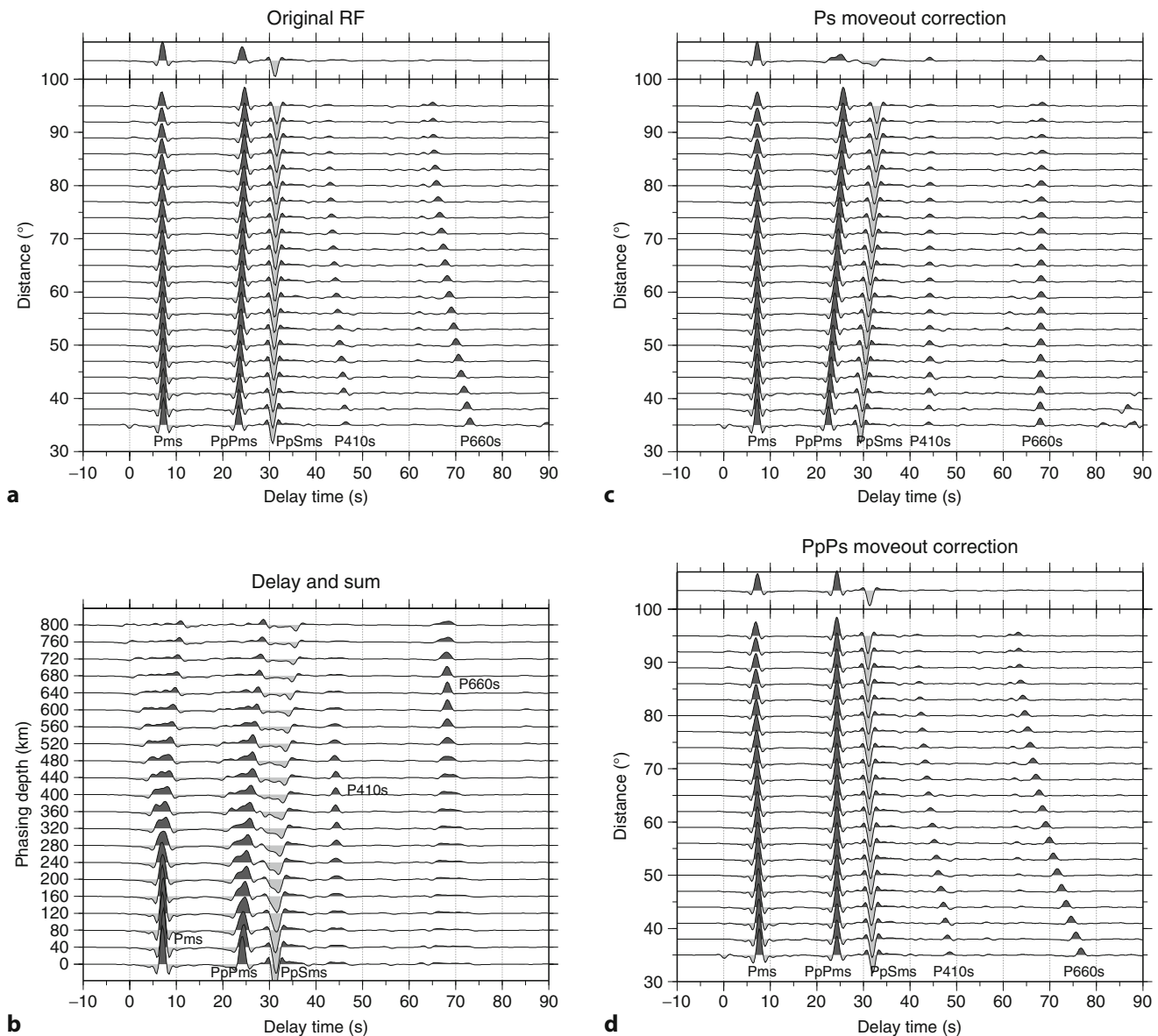
apparent discontinuities. One of the first examples of receiver function depth migration is given by Kosarev et al. (1999). A more sophisticated development is given, e.g., by Bostock (2002).

The results of the processing steps described so far indicate the great similarity with steep angle-controlled source techniques. The topography of seismic discontinuities in the interior of the Earth is displayed either in the time or space domain. Distance moveout corrections and depth migration are the most important common steps. The main differences are the frequency content (which determines the resolution) and depth penetration. Steep angle seismics uses frequencies of many Hertz, whereas the teleseismic signals have periods from one to many seconds. Steep angle seismics does not always reach the Moho, whereas receiver functions have practically unlimited depth penetration since the signals are incident from below. Ryberg and Weber (2000) conclude from computations of theoretical seismograms of two-dimensional models that a station spacing of a few kilometers is required if structures with a few kilometer scale length need to be resolved (see Fig. 9). A station density of a few kilometers has so far been achieved only in very few deployments (e.g., Hansen and Dueker 2009).

Waveform Modeling

The above described processing steps for receiver functions lead to the detection and location of discontinuities in the interior of the Earth and their topography, ranging from the base of the sedimentary layers to the upper mantle transition zone between 410- and 660-km depth. A frequently applied additional step in receiver functions is the inversion of the waveforms over the entire length of the receiver function at a station into a one-dimensional velocity depth model (e.g., Ammon et al. 1990; Kind et al. 1995). Such an inversion technique has the capability of fitting the data nearly perfectly. However, problems involved are the nonuniqueness of the inversion and the assumption that seismic phases, which have been generated by a three-dimensional Earth, can be transformed into a one-dimensional model (Ammon et al. 1990). Although the nonuniqueness problem has been improved recently by using long-period signals (Jacobsen and Sverningsen 2008), even a perfect fit of the observed waveforms with theoretical waveforms encounters the problem of fitting three-dimensional effects and possible noise into a one-dimensional model.

Somewhat closer to reality seems to be the inversion of isolated waveforms of conversions from individual discontinuities in P or S receiver functions (e.g., Moho or LAB). An example is given by Li et al. (2007) (see Fig. 10). They summed about 8000 S-to-P converted signals from the LAB (S receiver functions) in the western USA and inverted the



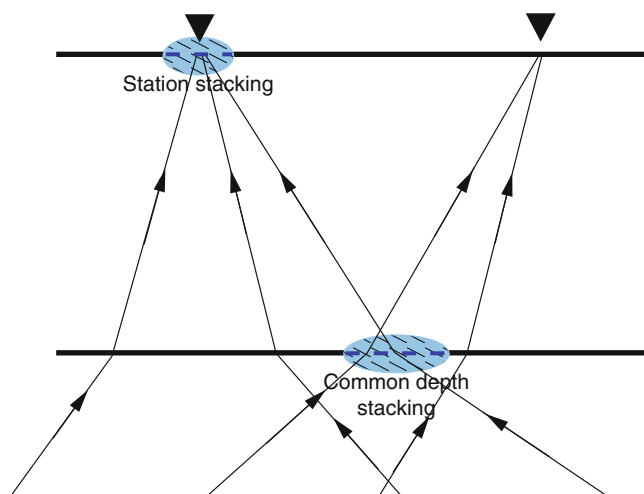
Seismic, Receiver Function Technique, Fig. 6 Theoretical P receiver functions in original form and moveout corrected. A velocity model is needed to apply the moveout correction. (Bottom

left) Delay and sum technique developed by Vinnik (1977). Delays are computed for a number of assumed depths of conversion

width of the summed signal in terms of thickness of the LAB gradient. They obtained a transition zone of not more than 20 km. However, determinations of the size of the velocity jump and of the thickness of the LAB from summation traces may depend on focusing and defocusing effects of the topography within the region of piercing points and erroneous transform topography into the vertical velocity structure. Therefore, this 20-km gradient must be considered as an upper limit in this case. Another example is given by Mohsen et al. (2005) in Fig. 11. They show how each one of the observed seismic phases may be modeled with good accuracy with a relatively simple model.

Two Data Examples

In Fig. 12 are shown P, S, and SKS receiver functions in time domain from a seismically quiet station in northern Canada. The interesting point in these data is that in S and SKS receiver function (SKSRF), a clear negative phase is observed at 12–13 s, which corresponds to a depth of about 110 km. A negative signal indicates a low velocity zone. Such a zone in that depth may be caused by the LAB. However, from surface wave studies, a much larger depth of the LAB is expected in this region. These contradicting results pose very interesting questions about the structure of the upper



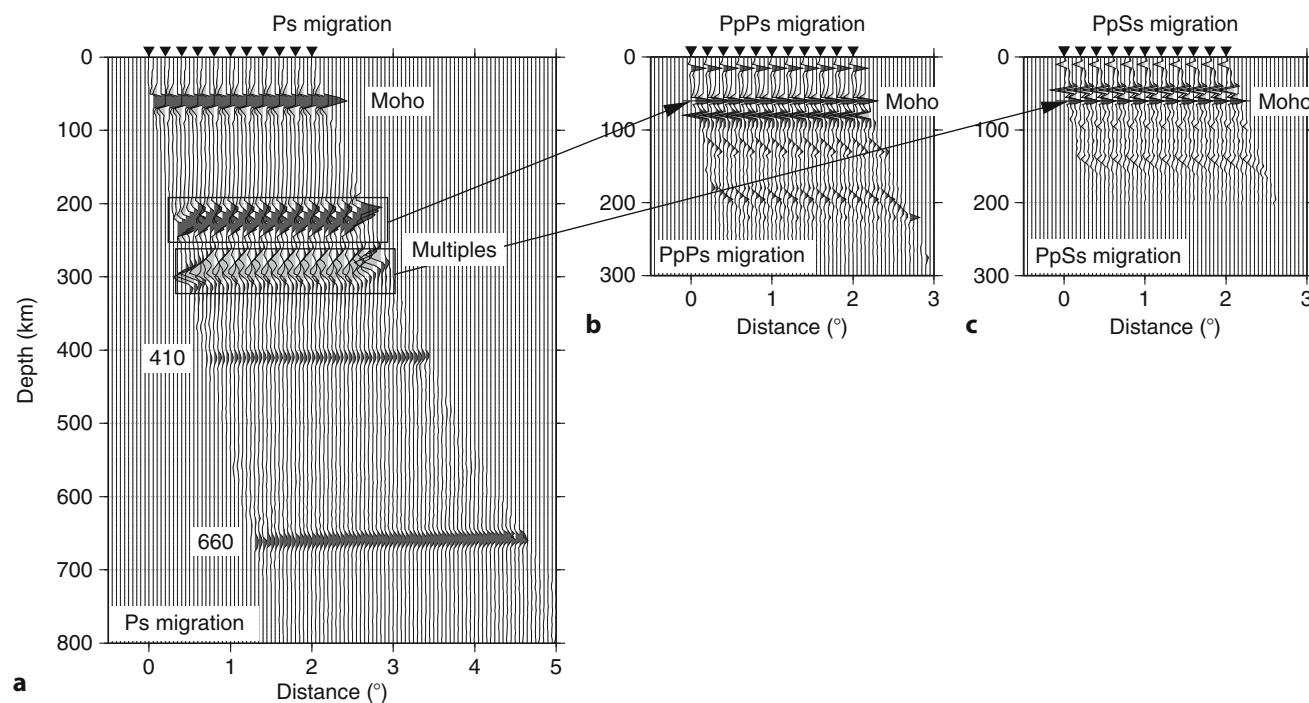
Seismic, Receiver Function Technique, Fig. 7 Cartoon showing receiver function rays in a simple Earth model. Single-station summation is useful for shallower studies or for averages over larger regions in greater depths. In case of dense deployments of several stations, common depth summation (CCP) may be used

mantle beneath old cratons. In Fig. 13 is shown an excellent example of depth migrated P receiver functions beneath a subduction zone (Kawakatsu et al. 2009). The subducting oceanic lithosphere is completely visible, including upper and lower boundary of the oceanic crust and the lower boundary of the oceanic lithosphere (see ► [“Lithosphere, Oceanic”](#)).

New Technical Development of the S-Receiver Function Method

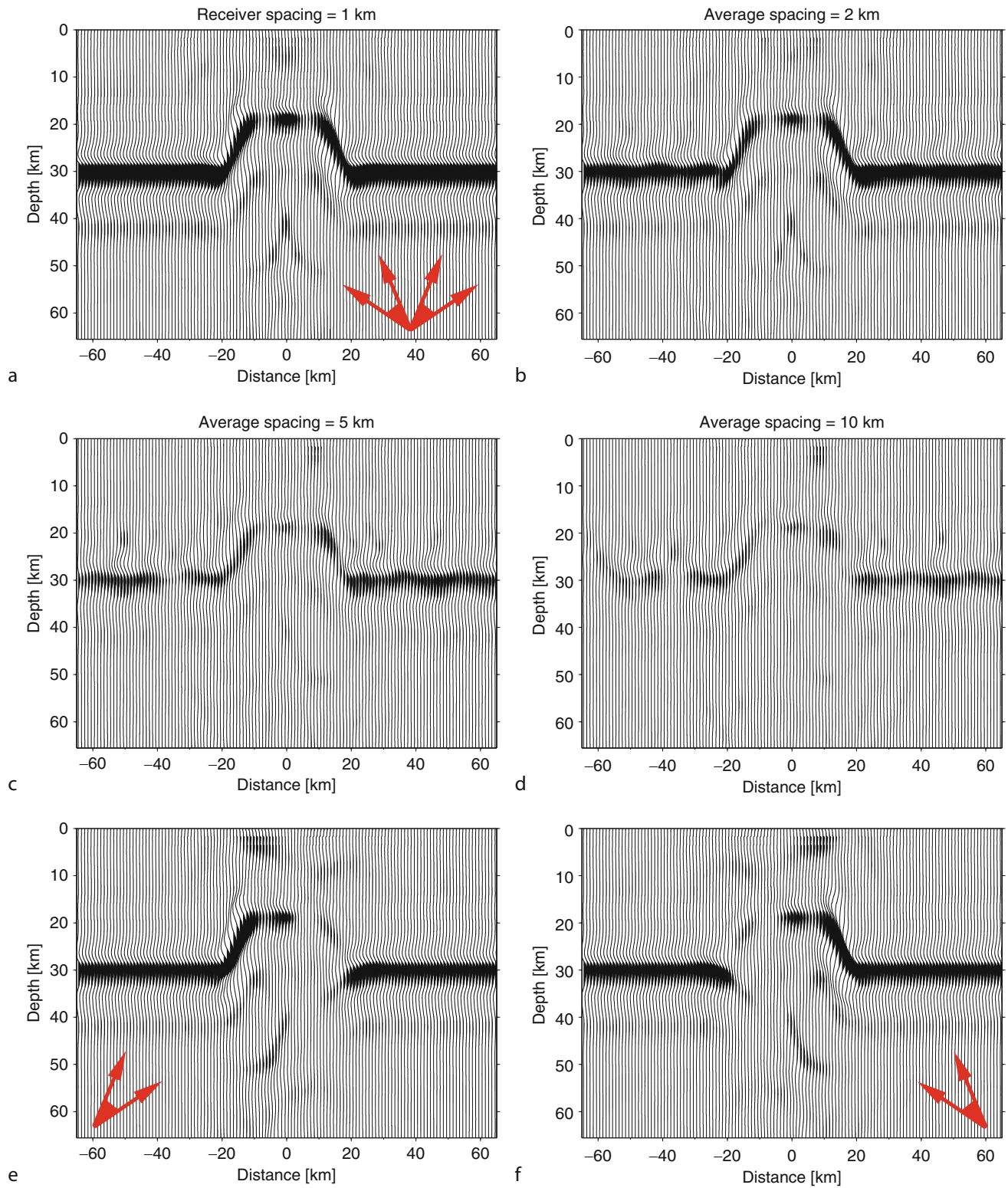
The question of sidelobes in the receiver function technique caused by the data processing has frequently been asked but not satisfyingly been answered so far. A new method was suggested which avoids sidelobes completely and leads in some geological regions to surprising new results.

The traditional S-receiver function (SRF) method uses several acausal processing steps. The first of such steps is the deconvolution which may cause significant precursors of



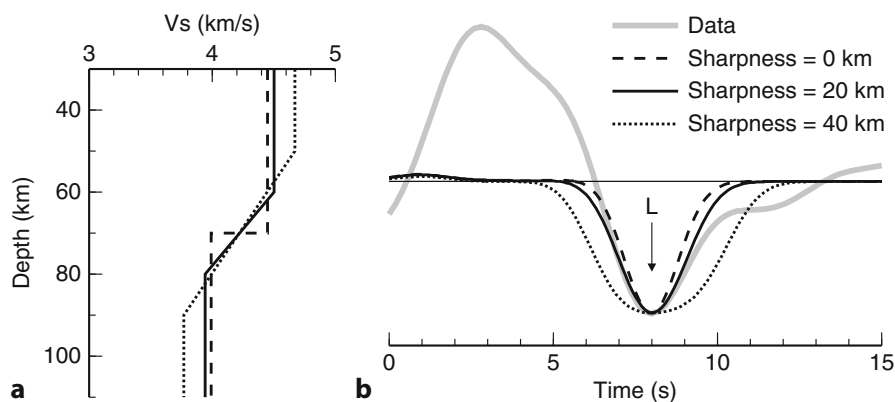
Seismic, Receiver Function Technique, Fig. 8 Principle of depth migration of seismic records (theoretical seismograms in this case). Each ray from each station is back projected into an assumed model of the Earth. The amplitudes of the seismogram are distributed along this ray path. In the left figure it is assumed that all energy in the seismogram

originates from direct conversions; in the other figures, it is assumed that all energy is caused by multiples. Migration can, like moveout correction, only be done for one wave type at a time. The energy of the other phase type is migrated to the wrong depth and not focused



Seismic, Receiver Function Technique, Fig. 9 (a-d) Migrated theoretical 1-Hz receiver functions of a dome structure with different receiver spacing. A spacing of 5 or less kilometers is required to obtain

a good image of the structure. (e-f) Effects of illuminating the structure from limited back azimuths (see *arrows*, Ryberg and Weber 2000). Station spacing is 1 km



Seismic, Receiver Function Technique, Fig. 10 Modeling of the LAB summation phase in the western USA (Li et al. 2007, Fig. 8). (a) Models with different sharpness and velocity contrast of the LAB. Dashed line: LAB as a first-order discontinuity; solid line: LAB as a 20-km thick gradient transition zone and dotted line: LAB as a 40-km thick gradient transition zone. (b) Waveforms of data and synthetic S receiver functions. The thick gray line is the summation of all the

S receiver function stacks for each station after an alignment of the LAB phase. A 3–20-s bandpass filter is applied prior to alignment and summation for a better phase correlation. Synthetic S receiver functions correspond to the models shown in (a). The large positive swing in the data at less than 5 s is the averaged Moho signal, which is not computed in the synthetic seismograms

the SV signal. The second step is the definition of the reference time. The reference time is usually chosen at the time of the maximum of the deconvolved SV signal. Summation of different traces along these reference times causes always significant sidelobes. Applying acausal filters contributes also to sidelobes.

In the new method, deconvolution is avoided and the causal onset time of the SV signal is chosen as reference time. We correct for the sign of the SV onset and normalize the amplitudes by the maximum of the SV wave group. Also unfiltered broadband records are used. The new causal method is described by Kind et al. (2019). A comparison of a profile across the northern part of the United States computed with the traditional acausal method and the new causal method is shown in Fig. 14. The same data from the USARRAY are used in both techniques. The main difference is that the mid-Lithospheric discontinuity (MLD) in the traditional acausal deconvolution method (marked MLD in Fig. 14b) has completely disappeared in the causal method (Fig. 14a). The signal marked LAB (Lithosphere-Asthenosphere Boundary), in Fig. 14a, b, is identical in both methods. This indicates that this signal is physically existing and may be interpreted as LAB of the tectonic active western USA.

A new signal marked X appears in the causal method (Fig. 14a), which is overwhelmed by the sidelobes in the acausal method. This signal was interpreted by Kind et al. (2019) as indication of low-angle fossil subduction below nearly the entire Superior Craton. More geologically significant new findings of the new method in the USA are shown by Kind et al. (2019) and in China by Shen et al. (2019).

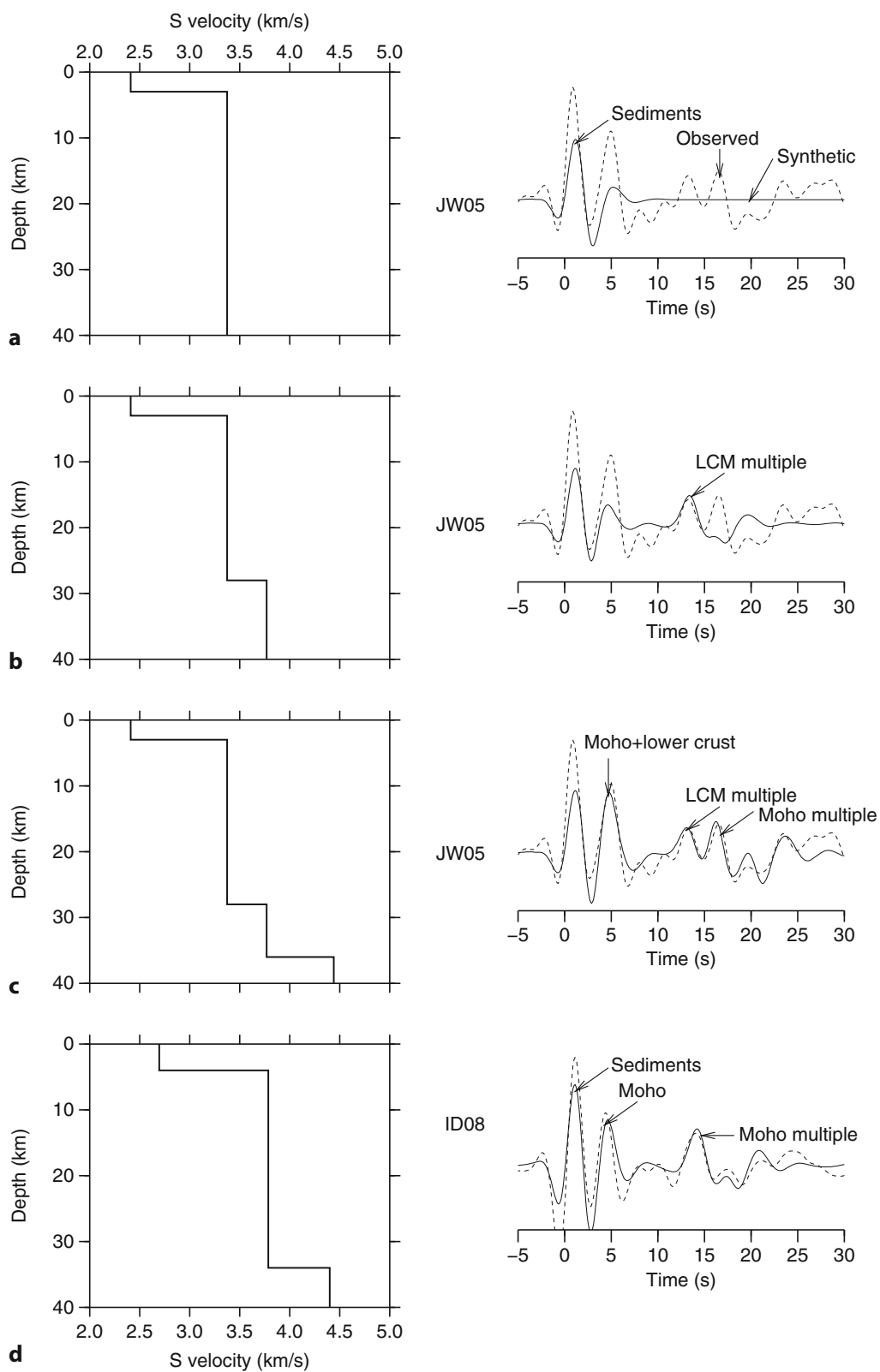
Summary

Receiver functions or converted waves belong to a class of small secondary scattered waves generated by a relatively strong main phase like P or S. Similar scattered phases are precursors of other main phases like PP, SS, or P'P', which have the great advantage to cover areas that are difficult to access, like oceans (see ► [“Body Waves”](#); ► [“Seismic Phase Nomenclature: The IASPEI Standard”](#)). These techniques are since about 30 years the main tools to study discontinuities of material properties in the Earth. The other main seismic tool is tomography, which is sensitive to smooth changes of material properties. The success of the scattered wave techniques was made possible by the great extension of high-quality seismic networks in the last decades. Permanent and mobile networks supplement each other. Especially studies of the lithospheric plates and the mantle transition zone have gained much from the new techniques. Many more detailed results can be expected when much denser networks will be used in future.

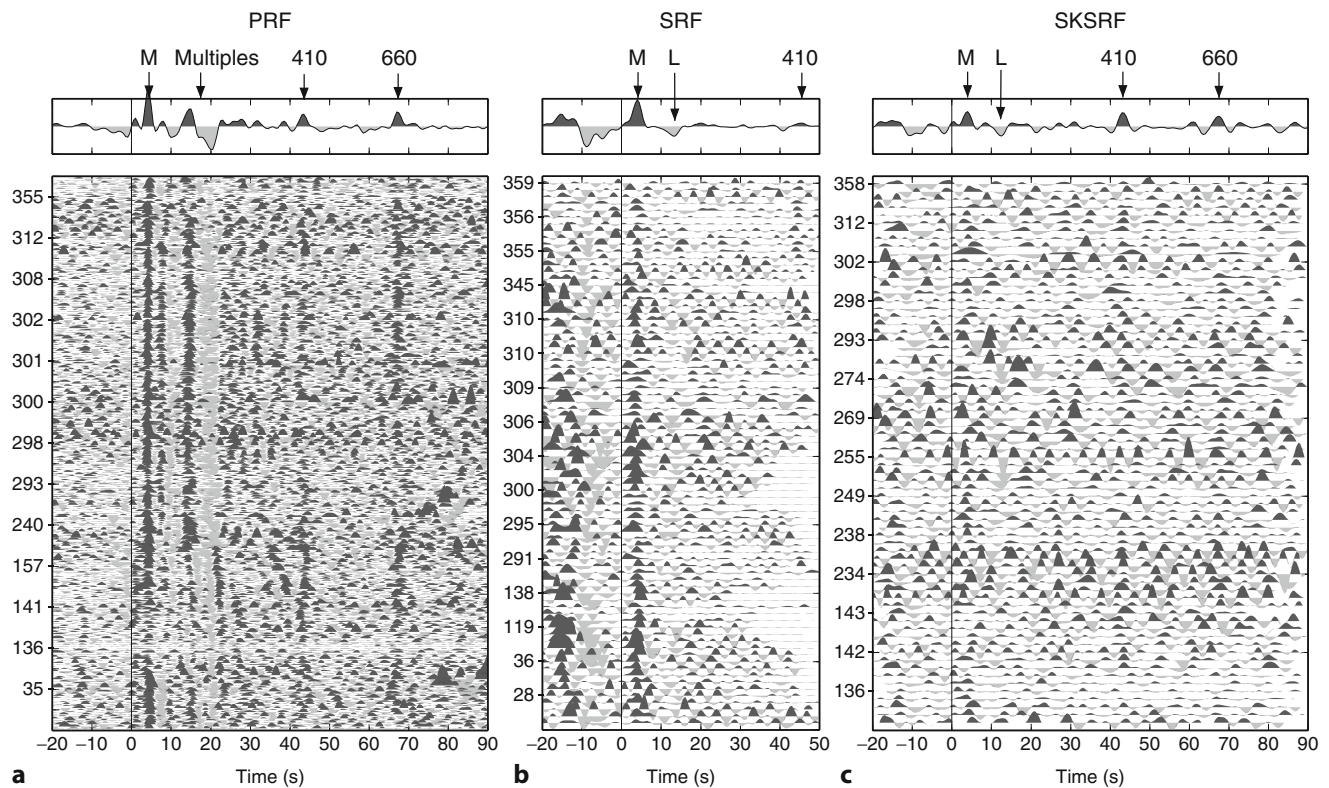
A recent review of the receiver function technique has been given by Rondenay (2009). A number of additional processing steps are the use of multiples to determine average crustal properties (e.g., Zhu and Kanamori 2000), the role of anisotropy (e.g., Levin and Park 1997; Schulte-Pelkum et al. 2005), or three-dimensional inversion (e.g., Bostock 2002).

Cross-References

- [Body Waves](#)
- [Continental Crustal Structure](#)



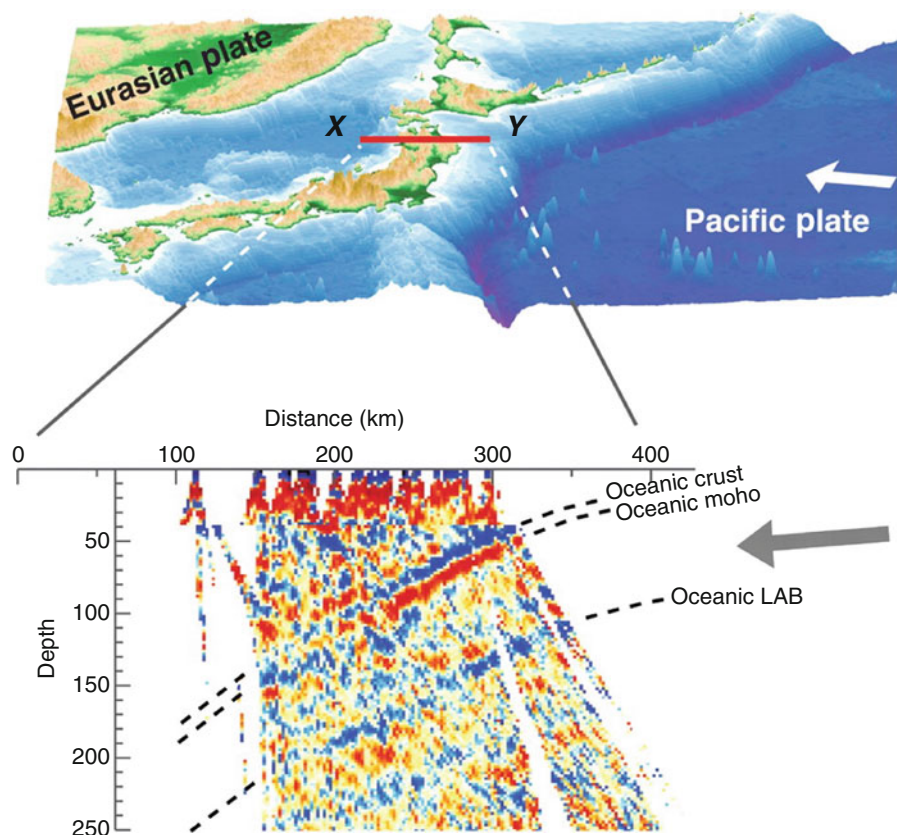
Seismic, Receiver Function Technique, Fig. 11 Forward waveform modeling of P receiver functions with relatively simple models (Mohsen et al. 2005)

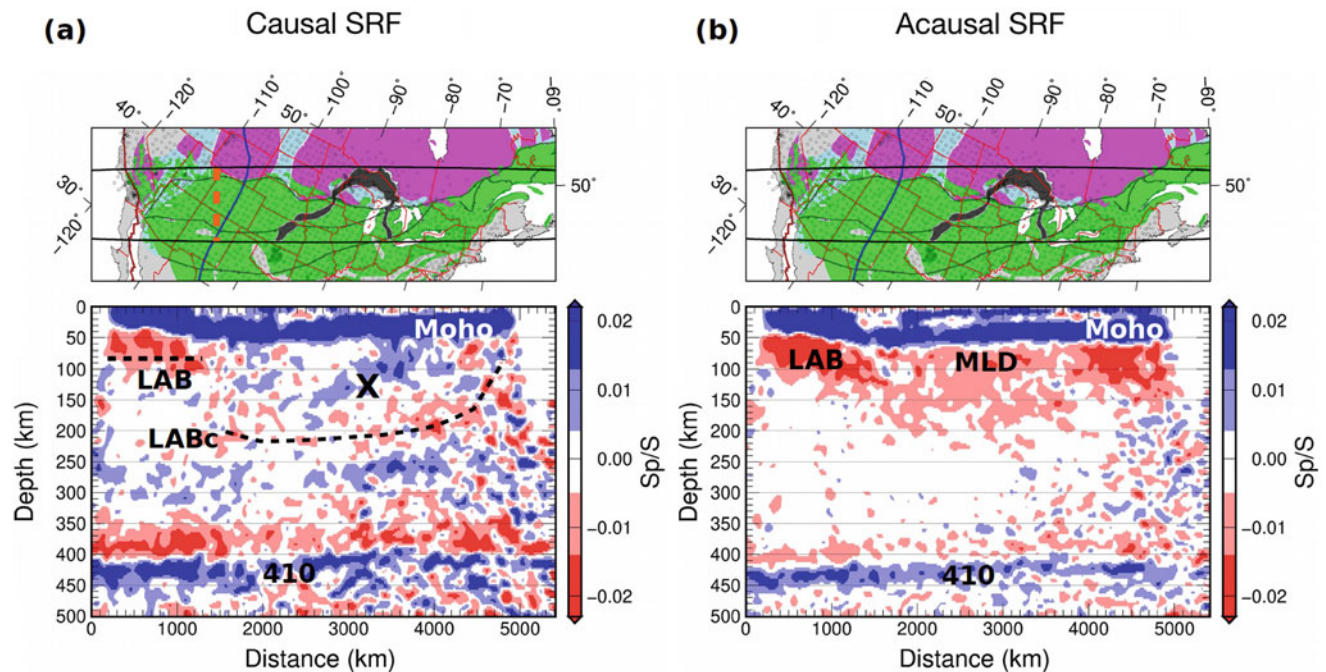


Seismic, Receiver Function Technique, Fig. 12 PRFs, SRFs, and SKSRFs of station YKW3 (Yellow Knife, Canada). The receiver functions are corrected for moveout with a reference slowness of 6.4 s/degree and a reference model of the IASP91 (Kennett and Engdahl 1991). The number of the PRFs is more than 200, and the numbers of the SRFs and

SKSRFs are ~80 each. A three-pole low-pass filter with corner frequency of 3 S is applied to all the data. Back azimuth values for some sample traces are labeled to the left of each trace. The summation trace is plotted above each receiver function section. Main conversion phases are labeled on the summation traces

Seismic, Receiver Function Technique, Fig. 13 P receiver functions for ocean bottom station WP2 and receiver function image of subducting Pacific plate beneath northeast Japan. In the topographic map, the inverted red triangle indicates the location of WP2, which is deployed at a depth of 460 m below the seafloor with a water column of 5566-m thick. (a) P receiver function image along the profile XY using dense land seismic data of Hi-net from Japan. Red and blue colors indicate velocity increase (from shallow to deep) and decrease at the point, respectively. The top surface of the slab and the oceanic Moho are clearly imaged, as well as the bottom surface of the slab (i.e., subducting LAB). (b) P receiver function and S receiver function for WP2. Negative phases associated with a shear wave velocity drop marked as LAB appear to correspond to the strong negative (blue) signature of the subducting slab below Japan. (After Kawakatsu et al. 2009)





Seismic, Receiver Function Technique, Fig. 14 Comparison of an SRF profile computed with two methods. The MLD has disappeared in the causal method and a new structure X appears which might be interpreted as fossil subduction below the Superior Craton. (From Kind et al. 2019)

- Deep Seismic Reflection and Refraction Profiling
- Energy Partitioning of Seismic Waves
- Lithosphere, Continental
- Lithosphere, Oceanic
- Seismic Phase Nomenclature: The IASPEI Standard
- Seismic, Migration
- Seismological Networks

Bibliography

- Ammon CJ, Randall GE, Zandt G (1990) On the nonuniqueness of receiver function inversions. *J Geophys Res* 95:15303–15318
- Baig AM, Bostock MG, Mercier J-P (2005) Spectral reconstruction of teleseismic Green's functions. *J Geophys Res* 110:B08306
- Bock G (1994) Multiples as precursors of S, SKS and ScS. *Geophys J Int* 119(2):421–427
- Bostock MG (2002) Kirchhoff-approximate inversion of teleseismic wave fields. *Geophys J Int* 149:787–795
- Bostock MG (2004) Green's functions, source signatures, and the normalization of teleseismic wave fields. *J Geophys Res* 109
- Dueker KG, Sheehan AF (1997) Mantle discontinuity structure from midpoint stacks of converted P to S waves across the Yellowstone hotspot track. *J Geophys Res* 102:8313–8327
- Gurrola H, Baker GE, Minster JB (1995) Simultaneous time-domain deconvolution with application to the computation of receiver functions. *Geophys J Int* 120:537–543
- Hansen S, Dueker K (2009) P- and S-wave receiver function images of crustal imbrication beneath the cheyenne belt in Southeast Wyoming. *Bull Seismol Soc Am* 99:1953–1961. <https://doi.org/10.1785/0120080168>
- Jacobsen BH, Sverningesen L (2008) Enhanced uniqueness and linearity of receiver function inversion. *Bull Seismol Soc Am* 98:1756–1767
- Kawakatsu H, Kumar P, Takei Y, Shinohara M, Kanazawa T, Araki E, Suyehiro K (2009) Seismic evidence for sharp lithosphere-aesthenosphere boundaries of oceanic plates. *Science* 324:499–502
- Kennett BLN, Engdahl ER (1991) Travel times for global earthquake location and phase identification. *Geophys J Int* 105:429–465
- Kind R (1985) The reflectivity method for different source and receiver structures and comparison with GRF data. *J Geophys* 58:146–152
- Kind R, Vinnik LP (1988) The upper-mantle discontinuities underneath the GRF array from P-to-S converted phases. *J Geophys* 62:138–147
- Kind R, Kosarev GL, Petersen NV (1995) Receiver functions at the stations of the German Regional Seismic Network (GRSN). *Geophys J Int* 121:191–202
- Kind R, Mooney WD, Yuan X (2019) New insights into structural elements of the upper mantle beneath the contiguous United States from S-to-p converted waves. *J Geophys Res Solid Earth*. (in review)
- Kosarev G, Kind R, Sobolev SV, Yuan X, Hanka W, Oreshin S (1999) Seismic evidence for a detached Indian Lithospheric mantle beneath Tibet. *Science* 283:1306–1309
- Kumar P, Kind R, Yuan XH (2010) Receiver function summation without deconvolution. *Geophys J Int* 180:1223–1230
- Langston CA (1979) Structure under Mount Rainer, Washington, inferred from teleseismic body waves. *J Geophys Res* 84:4749–4762
- Langston CHA, Hammer JK (2001) The vertical component P-wave receiver function. *Bull Seismol Soc Am* 91(6):1805–1819
- Levin V, Park J (1997) P-SH conversions in a flat-layered medium with anisotropy of arbitrary orientation. *Geophys J Int* 131:253–266
- Li XQ, Yuan XH, Kind R (2007) The lithosphere-aesthenosphere boundary beneath the western United States. *Geophys J Int* 170(2):700–710
- Ligorria JP, Ammon CJ (1999) Iterative deconvolution and receiver function estimation. *Bull Seismol Soc Am* 89:1395–1400

- Mohsen A, Hofstetter R, Bock G, Kind R, Weber M, Wylegalla K, Rumpker G, DESERT Group (2005) A receiver function study across the Dead Sea Transform. *Geophys J Int* 160:948–960
- Montalbetti JF, Kanasewich ER (1970) Enhancement of teleseismic body phases with a polarization filter. *Geophys J Int* 21(2):119–129
- Rondenay S (2009) Upper mantle imaging with array recordings of converted and scattered teleseismic waves. *Surv Geophys* 30:377–405
- Ryberg T, Weber M (2000) Receiver function array: a reflection seismic approach. *Geophys J Int* 41:1–11
- Schulte-Pelkum V, Monsalve G, Sheehan A, Pandey MR, Sapkota S, Bilham R, Wu F (2005) Imaging the Indian subcontinent beneath the Himalaya. *Nature* 435:1222–1225. <https://doi.org/10.1038/nature03678>
- Shen X, Kind R, Huang Z, Yuan X, Liu M (2019) Imaging the mantle lithosphere below the China cratons using S-top converted waves. *Tectonophysics* 754:73–79
- Vinnik LP (1977) Detection of waves converted from P to SV in the mantle. *Phys Earth Planet Inter* 15:39–45
- Yuan X, Kind R, Li X, Wang R (2006) S receiver functions: synthetics and data example. *Geophys J Int* 175(2):555–564
- Yuan X, Li X, Wölbern I, Kind R (2007) Tracing the Hawaiian mantle plume by converted seismic waves. In: Ritter JRR, Christensen UR (eds) *Mantle plumes – a multidisciplinary approach*. Springer, Berlin, pp 49–69
- Zhu LP, Kanamori H (2000) Moho depth variation in southern California from teleseismic receiver functions. *J Geophys Res* 105:2969–2980

Seismic, Reflectivity Method

Mrinal K. Sen

Department of Geological Sciences, UT Institute for Geophysics, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX, USA

Synonyms

Reflection matrix method

Definition

Reflectivity method is a semi-analytic method for computing synthetic seismograms in layered elastic media. The method has now been extended to anisotropic and laterally varying media.

The Reflectivity Method

Introduction

Synthetic seismograms are now used routinely for interpretation of field seismic data and evaluating the performance of seismic processing algorithms. Generation of synthetic seismograms essentially involves solution of the partial differential equation (PDE) for seismic wave propagation.

Different forms of the PDE are available for different earth models (e.g., acoustic, isotropic, elastic, anisotropic, homogeneous, laterally inhomogeneous, etc.). Brute force numerical methods such as finite differences and finite elements can, in principle, handle models of any complexity. However, they are limited by the speed of computation and computer memory. Even with our latest computer technology, generation of realistic 3D synthetic seismograms is not possible within reasonable clock time.

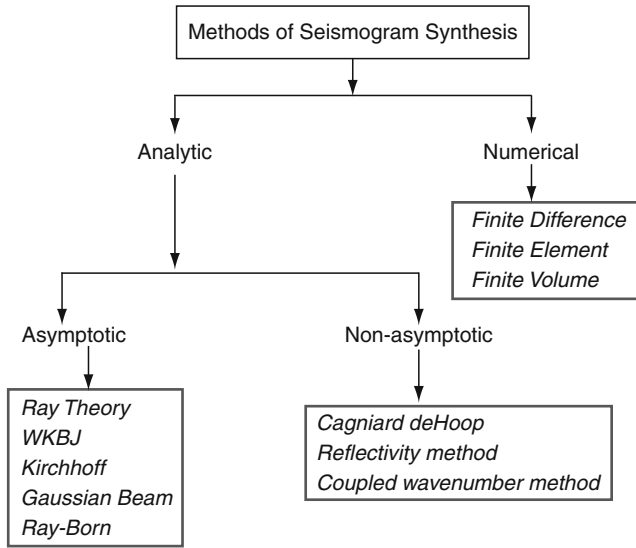
Much effort has been invested over the years in developing methods that are fast and accurate. We are left with three choices:

- Approximate the earth model and generate analytic or semi-analytic solutions
- Derive approximate (asymptotic) solutions for realistic heterogeneous models
- Derive pure numerical solutions for general heterogeneous models

Figure 1 provides a fairly complete list of the methods currently available for seismic modeling. Pure numerical methods based on finite difference (Virieux 1984; Liu and Sen 2009) and finite element (Marfurt 1984; Komatitsch and Tromp 1999, DeBasabe and Sen 2007; Vamaraju et al. 2020) approaches generate complete solutions but these methods become prohibitively expensive at high frequencies. The ray-based methods generate asymptotic solutions at infinite frequency and can be applied to models of general complexity as long as rays can be traced through the medium (Cerveny 2001). Although the original WKB method is valid for layered media, its extension to laterally varying media is the Maslov method (Chapman and Drummond 1982).

The methods such as the reflectivity (Fuchs and Mueller 1971) or Cagniard-deHoop method (e.g., Aki and Richards 2000) are included in the category of non-asymptotic analytic method. They were originally proposed for layered (1D) isotropic media and are still commonly used in such media although some extensions of these methods to anisotropy and lateral heterogeneity have indeed been reported. Although I included these methods in the category of “analytic method,” their final evaluation is done numerically. Thus, strictly speaking, they should be included in the category of “semi-analytic” methods. The speed of computation of these methods can be improved by using an asymptotic approximation at the final stage of evaluation of an oscillatory integral.

The primary goal of this entry is to provide an overview of the reflectivity method – perhaps the most popular method of seismogram synthesis. In the following sections, I will summarize the theory and numerical methods pertaining to reflectivity and provide some numerical examples. Finally, I will discuss an extension of the reflectivity method to laterally inhomogeneous media.



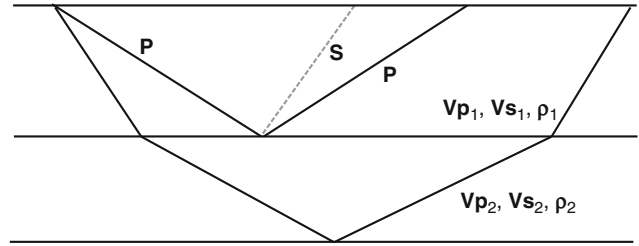
Seismic, Reflectivity Method, Fig. 1 A table showing different methods for computing synthetic seismograms: the reflectivity method is a semi-analytic method that is generally valid for layered earth models

Plane Waves

The plane waves are fundamental to understanding wave propagation. It has been demonstrated that a point source response can be generated by a weighted sum of individual plane wave responses (e.g., Aki and Richards 2000). In a homogeneous medium, plane waves propagate without spreading. In a stratified medium, plane waves remain planar during their propagation. However, as they propagate through a stack of flat layers (Fig. 2), they undergo changes in amplitude and phase.

- The amplitude changes can be accounted for by imposing boundary conditions of continuity of displacement and stress resulting in expressions for the so-called plane wave reflection and transmission coefficients.
- Travel time changes can be computed using individual layer velocities for locally linear ray paths.
- Changes in propagation angle can be computed by imposing Snell's law of preservation of horizontal slowness.
- In a stratified medium, an incident plane wave generates converted waves and internal multiples with appropriate amplitudes given by reflection and transmission coefficients.

Figure 2 displays a stack of layers in which each layer is characterized by its elastic moduli (P-wave and S-wave velocities in an elastic isotropic medium), density, and thickness. An incident plane P-wave characterized by a ray (normal to the plane) given by the angle it makes with the vertical axis is shown in the top layer. At the first layer boundary, it generates a reflected P-wave and a reflected SV-wave (converted). It



Seismic, Reflectivity Method, Fig. 2 A stack of layers in which each layer is characterized by its elastic moduli, density, and layer thickness: As an incident plane wave propagates through the stack of layer, it undergoes mode conversion and changes in amplitude due to reflection and transmission

also generates several reverberations with the top free surface (not shown in the figure). As the plane wave enters the second layer, it changes its propagation direction and similarly generates several converted waves and multiples. The primary task in generating synthetic seismograms is to keep track of all of these paths and account for all these amplitude and phase changes. The reflectivity method does exactly that. However, before we look into the details of the reflectivity method, let us first examine some simple cases with a few rays.

Simple Plane Wave Synthetics

It is fairly straightforward to generate plane wave synthetic seismograms for a layered medium for a few rays or plane waves. Two fundamental parameters involving plane wave seismograms are the ray parameter or horizontal slowness denoted by p and vertical delay time denoted by τ . They are defined as follows

$$p = \frac{\sin \theta}{v}, \quad q = \frac{\cos \theta}{v} \quad (1)$$

where θ is the vertical angle of the ray, q is called the vertical slowness, and

$$\tau = 2hq, \quad (2)$$

where h is the layer thickness.

Now referring to Fig. 2, the PP reflection delay time for layer 1 is

$$\tau_1^{PP} = 2h_1 q_1^P, \quad (3)$$

and the PS reflection delay time is

$$\tau_1^{PS} = h_1 (q_1^P + q_1^S). \quad (4)$$

In the above equations, the subscript represents the layer and the superscript represents the wavemode under consideration. The plane wave response of the primary PP mode from

the interface between layer 1 and layer 2 can be written down in the frequency-slowness domain as

$$R(\omega, p) = R_1^{PP} \exp(i\omega 2h_1 q_1^{PP}), \quad (5)$$

where R_{1PP} is the PP plane wave reflection coefficient for interface 1. Similar expression can be written down for the PS mode primary reflection.

$$R(\omega, p) = R_1^{PS} \exp(i\omega h_1 (q_1^P + q_1^S)). \quad (6)$$

To include the reflection from the second layer, we simply need to add a term appropriate for the second layer (similar to Eq. 3) resulting in the following expression

$$R(\omega, p) = T_{1D}^{PP} R_2^{PP} T_{1U}^{PP} \exp(i\omega (2h_1 q_1^{PP} + 2h_2 q_2^{PP})), \quad (7)$$

where we include the effects of transmission using downgoing (with subscript D) and upgoing (with subscript U) transmission coefficients. Expressions similar to Eq. (7) can also be written down for mode-converted waves.

Examples of plane wave and point source synthetic seismograms are shown in Fig. 3; these are generated for a vertical point force and only the traces of vertical component of displacement are shown. The upper panel shows the plane

wave seismograms in τ - p domain while the lower panel shows the corresponding x - t domain synthetic seismograms. The left panel shows seismograms for primary PP reflections only. Note two distinct reflections from the two-layer boundaries. The right panel shows seismograms that include primary PP , PS , and internal multiples from the second layer.

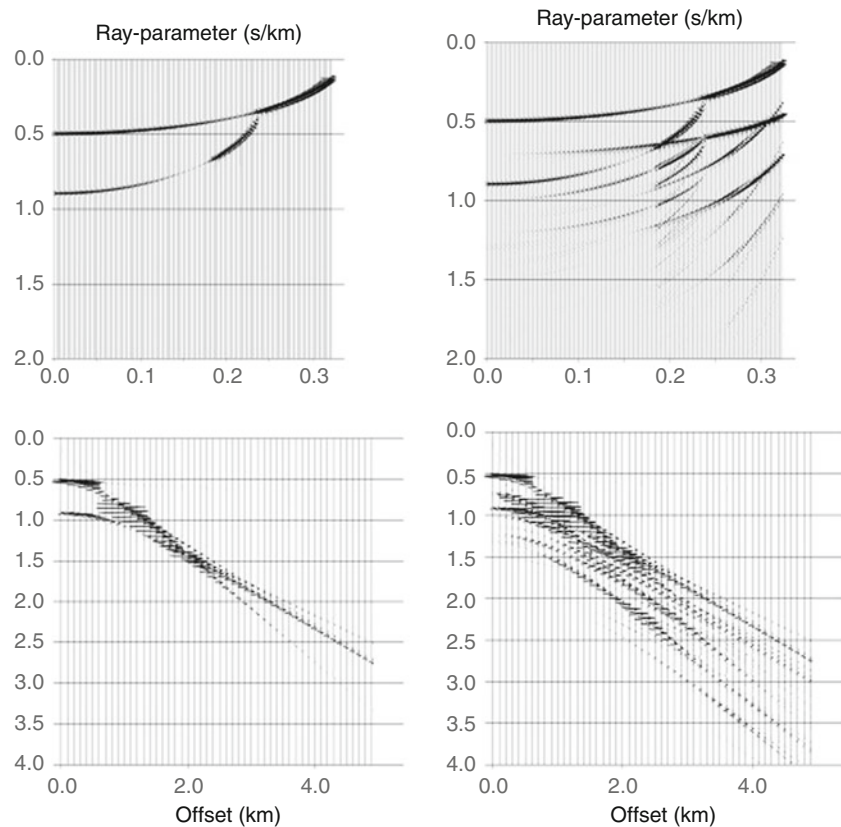
It is fairly straightforward to compute seismograms for a few selected phases for a few layers. In reality, we have many layers and an infinite number of rays paths – these are not easy to count. The reflectivity method offers an elegant algorithm for computing full response without having to count rays. Of course, we may choose to compute a few selected phases under reflectivity formulation as well, if desired.

Theory

Detailed developments of the theory of the reflectivity method for isotropic media can be found in Fuchs and Mueller (1971) and Kennett (1983). Derivations for an anisotropic layered medium can be found in Booth and Crampin (1983) and Fryer and Frazer (1984). The method has been used in a wide variety of applications in global and near surface seismology (e.g., Piera and Zanzi 1993). We start with the two fundamental equations in seismology, namely, the constitutive relation and the equation of motion given by

Seismic, Reflectivity Method,

Fig. 3 Examples of plane wave and point source synthetic seismograms for a model comprising two elastic layers over a half space: These were generated for a vertical point force and only the traces of vertical component of displacement are shown. The upper panel shows the plane wave seismograms in τ - p domain while the lower panel shows the corresponding x - t domain synthetic seismograms. The left panel shows seismograms for primary PP reflections only. The right panel shows the seismograms that include primary PP , PS , and internal multiples from the second layer



$$\tau = \mathbf{C} : \nabla \mathbf{u}, \quad (8)$$

$$\rho \omega^2 \mathbf{u} = \nabla \cdot \tau + \mathbf{f}, \quad (9)$$

where τ is a second rank stress tensor, $\mathbf{C}$ is the fourth rank elastic stiffness tensor, $\mathbf{u}$ is the displacement vector, ω is the angular frequency, and $\mathbf{f}$ is a body force term. Note that the above equations are in frequency domain.

The fundamental assumption in reflectivity is that the medium is layered and therefore, we can apply transformations over x and y coordinates to horizontal wavenumbers K_x and K_y or horizontal slowness p_x and p_y . For isotropic and layered transversely isotropic media, we can make use of cylindrical symmetry and transform to radial ray parameter p and azimuth ϕ . For a general anisotropic medium, we simply apply Fourier transforms over x and y coordinates to transform Eqs. (8) and (9) into horizontal wave numbers or ray parameter domain. Once that is achieved, all the analytic developments are done in the frequency ray-parameter domain.

The equation of motion and the constitutive relation are thus transformed into the following system of ordinary differential equations in depth z , by applying a Fourier transform in x and y

$$\mathbf{b} = i\omega \mathbf{A} \mathbf{b} + \mathbf{f} \quad (10)$$

where $\mathbf{b} = [u_x \ u_z \ u_y \ \tau_{xz} \ \tau_{zz} \ \tau_{yz}]^T = \mathbf{b}(\omega, \mathbf{p})$ is the stress-displacement vector which is a function of frequency ω and horizontal slowness $\mathbf{p}$, $\mathbf{A}(\omega, \mathbf{p})$ is the system matrix that is a function of elastic coefficients, and $\mathbf{f}$ is a body force term. The stress-displacement vector for isotropic and transversely isotropic media, the Eq. (10) decouples into two systems, namely, a P-SV (4X4) system and an SH (2X2) system. The solution of the ODE (Eq. 10) can be carried out by a propagator matrix method (Gilbert and Backus 1966). It is well known that the propagator matrix is generally unstable due to

growing exponentials (e.g., Jensen et al. 1993). Stable solutions can be obtained by one of the three methods:

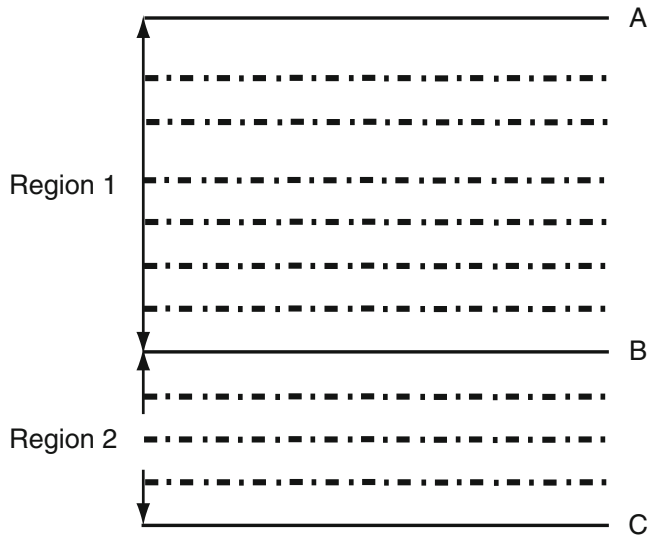
- A global matrix approach (Schmidt and Tango 1986).
- Compound matrix approach where we define a new system of ODE in which the elements of the new system matrix are the minors of the original system matrix. The original 4X4 P-SV system maps into a 6X6 system (Dunkin 1965, Phinney et al. 1987).
- An invariant imbedding or a reflection matrix approach (Kennett 1983).

Of the three methods listed above, the compound matrix and the reflectivity methods have been widely used in seismology. Table 1 summarizes the two methods. The reflectivity method has been very popular because of its ray-interpretation and easy generalization to azimuthally anisotropic media. In the unconditionally stable reflection matrix approach (Kennett 1983), the propagation uses the eigenvalues and eigenvectors of the system matrix $\mathbf{A}$; the eigenvalues are the vertical phase functions. These eigenvalues and eigenvectors are used to define four upgoing and downgoing reflection and transmission coefficient matrices, $\mathbf{R}_D$, $\mathbf{T}_D$, $\mathbf{R}_U$, and $\mathbf{T}_U$. These are propagated through the stack of layers to obtain a composite reflection matrix that includes the effects of reflection, transmission, mode-conversion, and internal multiples. Kennett (1983) derived the following iteration equation (Kennett 1983, p. 127), which can be used to compute the R/T matrices of a zone AC when those of zones AB and BC are known (Fig. 4):

$$\begin{aligned} \mathbf{R}_D^{AC} &= \mathbf{R}_D^{AB} + \mathbf{T}_U^{AB} \mathbf{R}_D^{BC} [\mathbf{I} - \mathbf{R}_U^{AB} \mathbf{R}_D^{BC}]^{-1} \mathbf{T}_D^{AB} \\ &\quad \mathbf{T}_D^{AC} + \mathbf{T}_D^{BC} [\mathbf{I} - \mathbf{R}_U^{AB} \mathbf{R}_D^{BC}]^{-1} \mathbf{T}_D^{AB} \\ \mathbf{R}_U^{AC} &= \mathbf{R}_U^{BC} + \mathbf{T}_D^{BC} \mathbf{R}_U^{AB} [\mathbf{I} - \mathbf{R}_D^{BC} \mathbf{R}_U^{AB}]^{-1} \mathbf{T}_U^{BC} \\ &\quad \mathbf{T}_U^{AC} + \mathbf{T}_U^{AB} [\mathbf{I} - \mathbf{R}_D^{BC} \mathbf{R}_U^{AB}]^{-1} \mathbf{T}_U^{BC} \end{aligned} \quad (11)$$

Seismic, Reflectivity Method, Table 1 Reflection matrices or compound matrices?

Reflectivity approach	Compound matrix approach
Unconditionally stable	Unconditionally stable
For the P-SV case, the algorithm evaluates four 2X2 complex reflection/transmission matrices and iteration equations for each frequency and ray-parameter.	For the P-SV case, the algorithm fewer floating-point operations than the reflectivity method. The method is computationally faster than the reflectivity method.
Derivatives or differential seismograms can be computed analytically for isotropic and VTI media	Derivatives or differential seismograms can be computed analytically for isotropic and VTI media
Derivatives can be computed by semi-analytic approach for general anisotropic media	—
Requires VERY CAREFUL coding	Coding is very EASY
Most efficient for general anisotropic media	Very slow for general anisotropy
Ray-interpretation is possible. Selected rays and selected number of multiples can be included. The algorithm is VERY FLEXIBLE	Complete solution is derived



Seismic, Reflectivity Method, Fig. 4 Zones AB and BC from a stack of layers; if the upgoing and downgoing reflection/transmission matrices for these two zones are known, those for the entire stack AC can be generated using the iteration equations

Equation (11) is the most fundamental development of the reflectivity formulation. Given the R/T matrices (which include layer propagation and interaction terms) of two consecutive regions AB and AC, we can compute the response of region AC using Eq. (11). Thus, we can propagate through the stack of layers resulting in an unconditionally stable algorithm.

Let us examine the downward reflection matrix in Eq. (11) in slightly more detail. Using the matrix identity

$$(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{I} + \mathbf{A} + \mathbf{A}\mathbf{A} + \mathbf{A}\mathbf{A}\mathbf{A} + \dots, \quad (12)$$

we can write

$$\begin{aligned} \mathbf{R}_D^{AC} &= \mathbf{R}_D^{AB} + \mathbf{T}_U^{AB} \mathbf{R}_D^{BC} [\mathbf{I} - \mathbf{R}_U^{AB} \mathbf{R}_D^{BC}]^{-1} \mathbf{T}_D^{AB} \\ &= \mathbf{R}_D^{AB} + \mathbf{T}_U^{AB} \mathbf{R}_D^{BC} \mathbf{T}_D^{AB} + \mathbf{T}_U^{AB} \mathbf{R}_D^{BC} \mathbf{R}_U^{AB} \mathbf{R}_D^{BC} \mathbf{T}_D^{AB} \\ &\quad + \dots \end{aligned} \quad (13)$$

Figure 5 clearly shows that Eq. (13) includes all the internal multiples and mode converted waves. We may also choose to retain a few terms in the expansion and thus compute exact response of a few rays.

While computing the up/downgoing reflection/transmission matrices through the stack of layers, we need to save the appropriate matrices through the source, receiver layers and the free surface. Finally, one can compute the entire stress/displacement vector comprising the components of displacement and traction at any receiver location. Combining these R/T matrices through one can obtain the required displacement or traction components.

Note that we obtain our solution in the frequency-ray parameter domain. An inverse temporal Fourier transforms results in $(\tau-p)$ seismograms. Synthetics in the offset-time domain can be obtained by plane wave synthesis of the (ω, p) or $(\tau-p)$ seismograms. A general flow chart for computing synthetic seismograms by the reflectivity method is shown in Fig. 6. Note that the ray-parameter, layer and frequency loops can be interchanged depending on the application. As stated earlier, intermediate results in $(\tau-p)$ domain are very useful for understanding wave propagation and amplitude effects since they are devoid of spherical spreading loss (Fryer 1980).

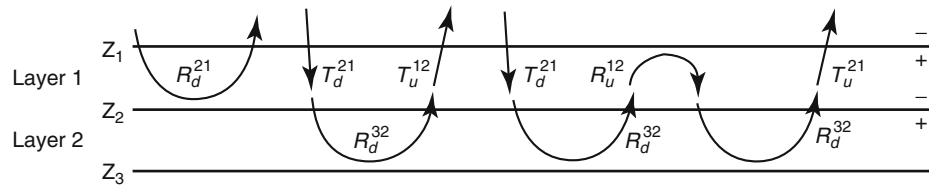
Computational Issues

It has been demonstrated that the reflectivity algorithm reviewed in this entry is unconditionally stable (Kennett 1983). However, one must incorporate some practical strategies in computing noise-free synthetic seismograms (e.g., Mallick and Frazer 1987).

- **Spatial aliasing:** Generation of offset synthetics from plane wave seismograms requires evaluation of an oscillatory integral. The integrand becomes more and more oscillatory with increase in offset and frequency. Thus, a practical strategy would be to use a frequency-dependent sampling interval in ray-parameters. Alternatively, one can use the minimum ray-parameter increment appropriate for the highest frequency and the largest offset of interest. We also need to use smooth tapers at the high and low ends of the ray-parameter window to avoid truncation phases.
- **Temporal aliasing:** Temporal aliasing can be avoided either by using a very large time window or adding a small imaginary part in the frequency following a method outlined in Phinney (1965)
- **Parallelization:** For high frequency and large offset calculation, the reflectivity calculation may be computationally intensive especially for azimuthally anisotropic media where reflectivity calculation needs to be performed for a large number of ray parameters. Note that (Fig. 6) almost the entire reflectivity computation can be done in parallel resulting in an algorithm that may be termed “embarrassingly parallel.” One of the simplest ways to parallelize is to distribute the computation of ray-parameter traces equally to the available nodes of a computer (e.g., Roy et al. 2005). The master node collects all the ray parameter traces (note that there is no communication in between the nodes required for the calculation of ray parameter trace). This results in an almost linear speedup as a function of the number of nodes. MPI can be used for parallelization; the reflectivity code can also be parallelized using simple compiler directives, if coded carefully.
- **Attenuation:** In order to include attenuation, we need to specify Q values for P- and S- waves for each layer. For

Seismic, Reflectivity Method,

Fig. 5 A demonstration of how the iteration equation includes all the reflection and transmission effects

**General Flow chart**

- Loop over ray-parameter
 - Loop over layer
 - Compute R/T (up going and downgoing) for all the interfaces
 - loop over frequency
 - evaluate iteration equation
 - end loop over frequency
 - end loop over layer
 - inverse FFT for (tau-p) seismograms
- End loop over ray-parameter
- Sum the plane wave responses to compute point source synthetics

Seismic, Reflectivity Method, Fig. 6 A general flow chart for developing a reflectivity code

frequency-dependent attenuation, the velocities are complex and dependent on frequency. In such cases, the layer loop (Fig. 6) becomes the deepest loop.

Applications

Exploration scale Modeling: Numerous applications of reflectivity synthetic seismograms for exploration can be found in literature. This algorithm is often used to generate synthetic seismograms from well logs and then used in calibrating a seismic gather at the well location and well tie (e.g., Sen 2006). Figure 7a shows V_p , V_s , density, and VTI anisotropy parameters ϵ and δ derived from a well log as a function of two-way time. The plane wave synthetics for an isotropic and VTI model and their differences are shown in Fig. 7b. Note that for isotropic calculations, ϵ and δ values were set equal to zero; for VTI calculations, V_p and V_s logs were treated as vertical P and S velocities. Offset synthetics for isotropic and VTI models and their differences are shown in Fig. 7c. As expected, the effect of anisotropy is more pronounced at large offsets or ray-parameters.

Modeling anisotropy in the core-mantle transition zone (D''): For modeling at regional and teleseismic distances, we need to apply earth flattening transformation to the models prior to computing synthetic seismograms using a reflectivity code. An example of such modeling is given in Pulliam and Sen (1998) in which attempts were made to model shear wave splitting observations of S-wave phases that propagate nearly horizontally through the core-mantle transition zone. Figure 8

shows the data and the corresponding synthetic seismograms for a station at Hockley from earthquakes from Tonga-Fiji. The model used for reflectivity calculation includes transversely isotropic layers with a horizontal axis of symmetry within the D'' zone.

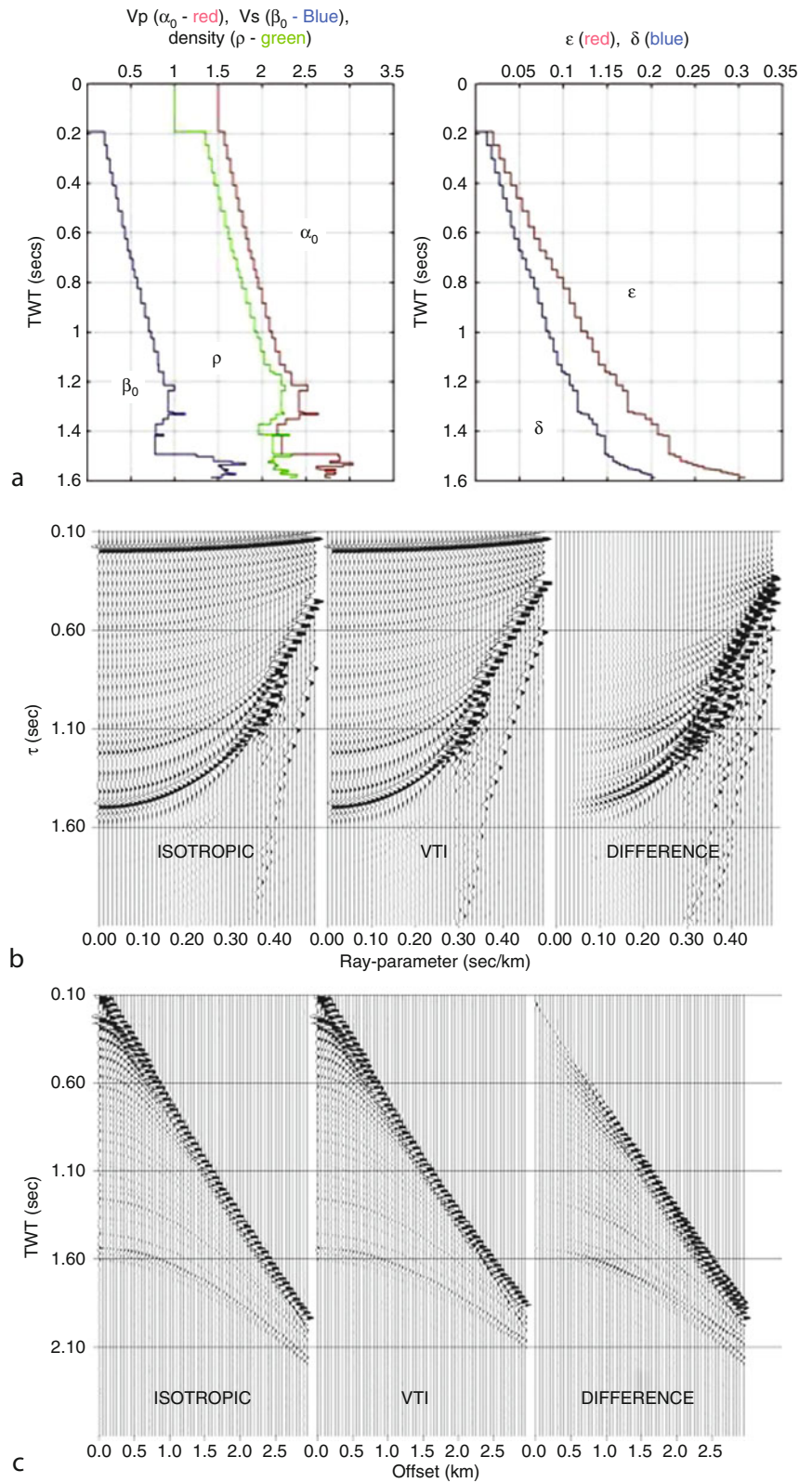
Modeling in azimuthally anisotropic media (comparison with FD): An example of comparison of azimuthally anisotropic reflectivity synthetic seismograms with those computed by a 3D azimuthally anisotropic finite difference code is shown in Fig. 9 (Bansal and Sen 2008). The synthetics computed by the two methods are in excellent agreement.

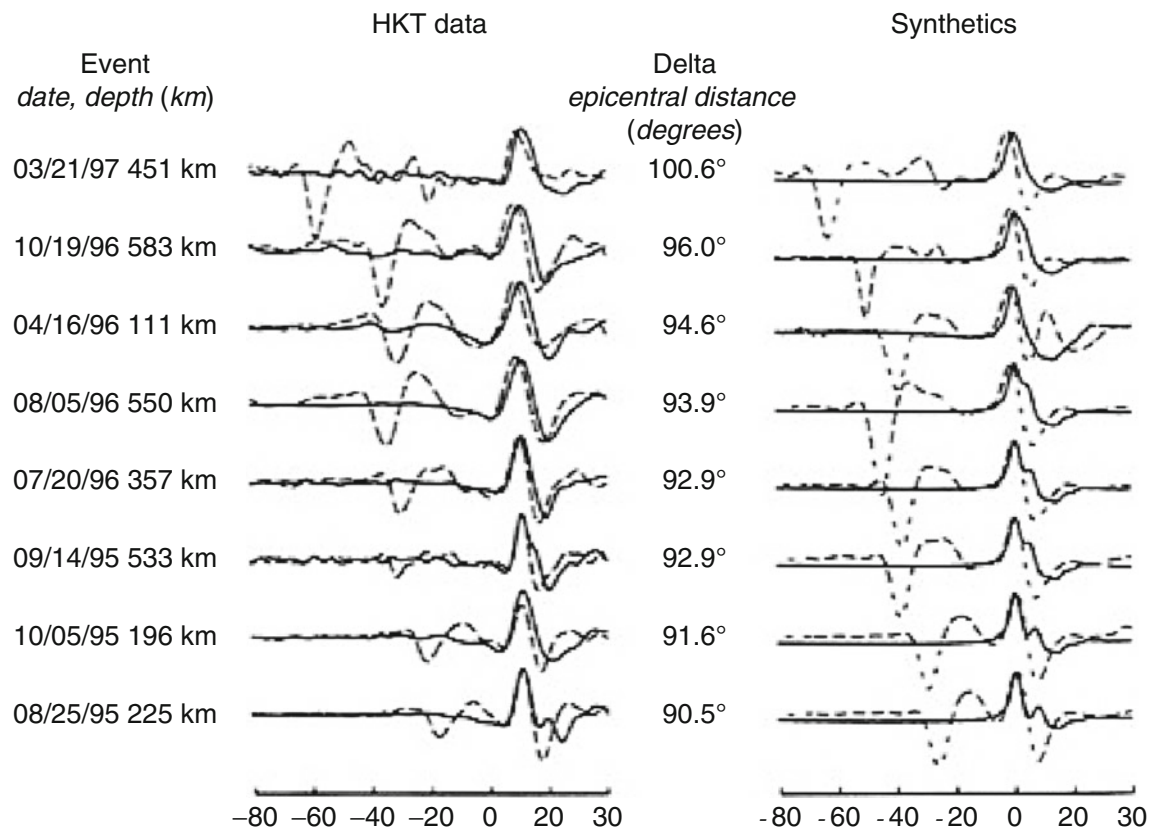
Extension to Laterally Heterogeneous Media

Extension of the reflectivity method to two dimensions, where the velocities and densities are allowed to vary laterally as well, is nontrivial. Koketsu et al. (1991) and Sen and Pal (2009) developed extended reflectivity methods for the case of homogeneous layers separated by irregular or curved interfaces which do not cross each other. Lateral media variations within a layer cannot be easily taken into account in a reflectivity-type formalism because the concept of an interface is inherent in it except using the concepts of pseudo-differential operators and Fourier Integral operators (McCoy et al. 1986).

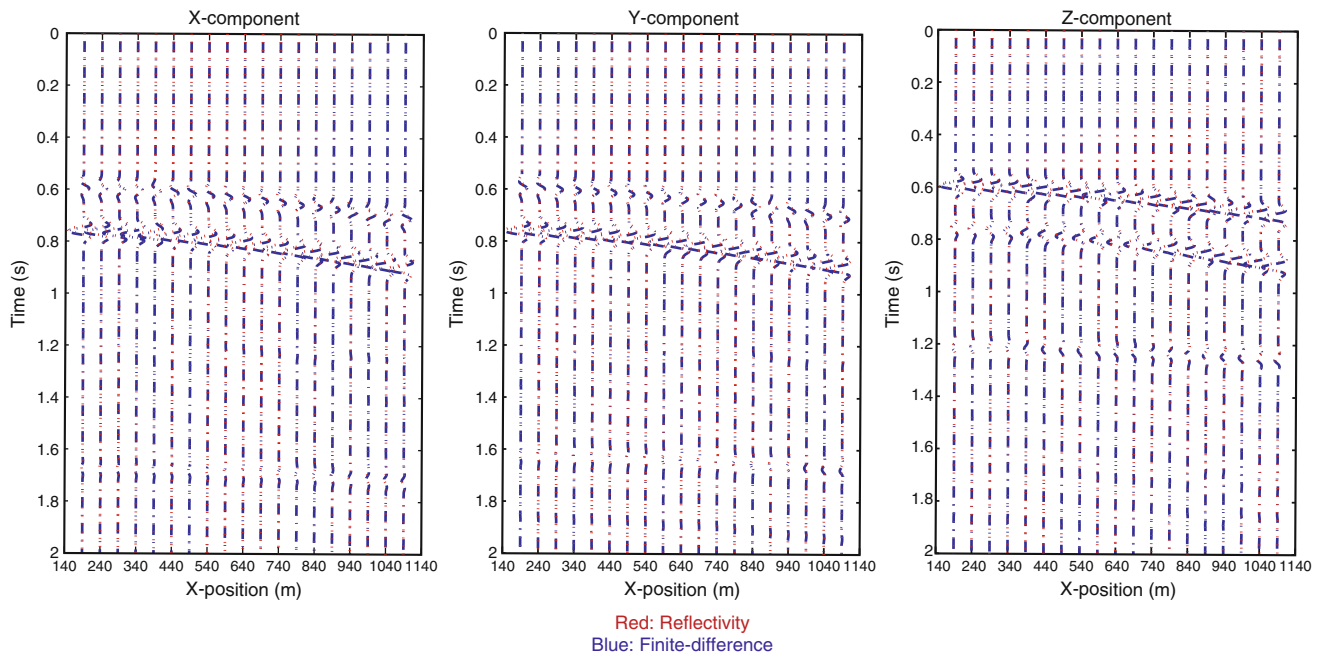
Across a flat surface, the horizontal slowness p is conserved according to Snell's law. However, across an irregular interface, an incident ray with a particular p is scattered into different plane waves with slownesses p' . This is mathematically equivalent to Fourier transforming the dependence of z on x to the scattered set of p' or k' (wavenumber). Now the coefficients $\mathbf{R}_D(p)$, $\mathbf{R}_U(p)$, $\mathbf{T}_D(p)$, $\mathbf{T}_U(p)$ become functions of both p and p' . The 2D extension of the method as proposed by Koketsu et al. (1991) involves explicit evaluation of boundary conditions that results in a matrix formulation involving several matrix inversions in the coupled ray-parameter domain; numerous numerical artifacts are caused by such matrix operations. An alternate asymptotic approach was proposed by Sen and Pal (2009) by examining a tangent-plane or Kirchhoff formulation in the plane wave domain to replace the exact boundary condition evaluation (Sen and Frazer 1991). All other reflectivity operations such as the invariant imbedding or iterative computation of reflection, transmission, multiple and mode conversions can be readily applied even under this approximation. This new algorithm computes noise-free

Seismic, Reflectivity Method,
Fig. 7 (a) V_p , V_s , density, ϵ , and δ for a well log model; (b) plane wave synthetic seismograms for an isotropic model (left), VTI model (middle), and the difference (right); (c) offset synthetic seismograms for an isotropic model (left), VTI model (middle), and the difference (right)

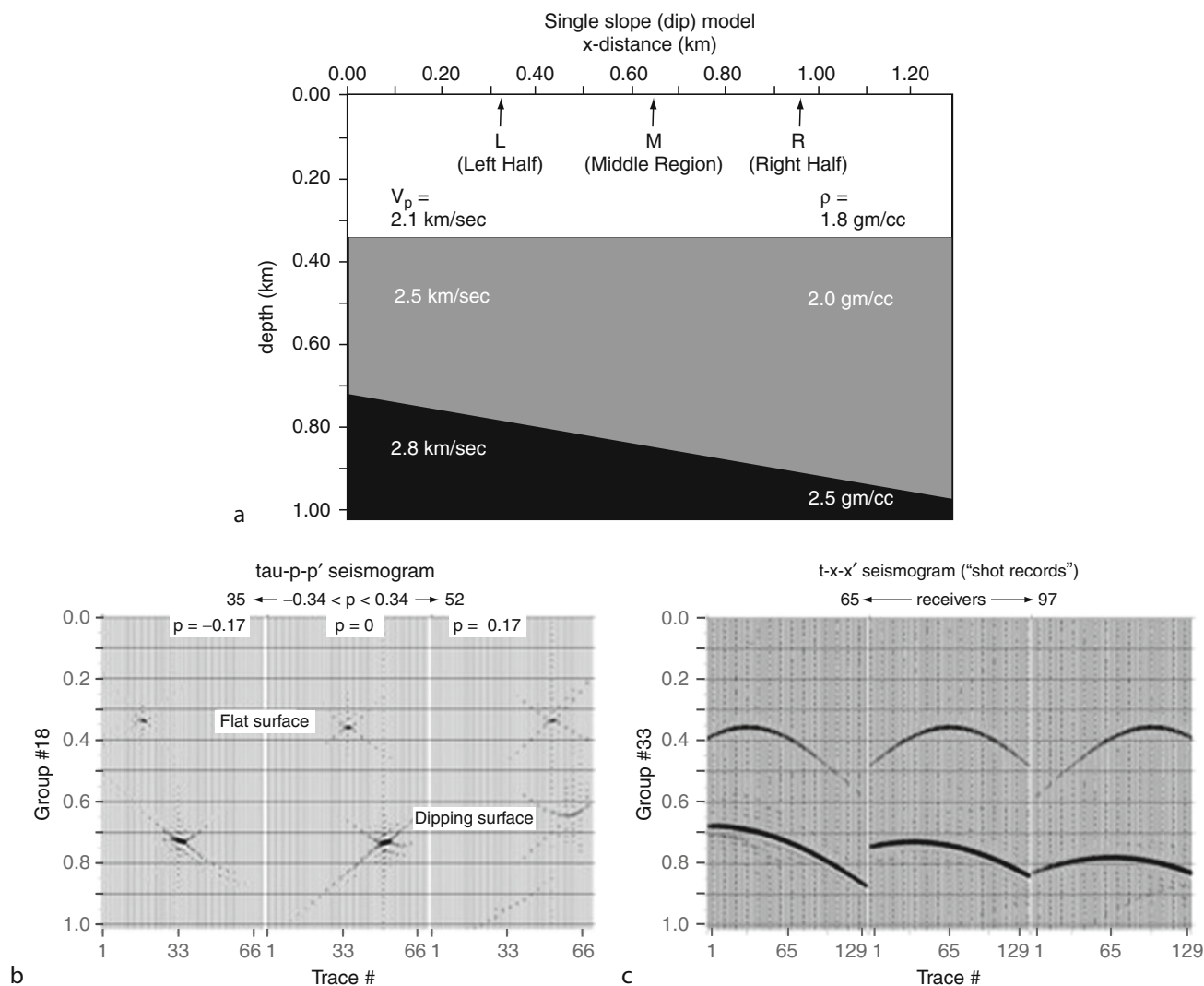




Seismic, Reflectivity Method, Fig. 8 Comparison of S-wave arrivals from Tonga-Fiji events recorded at Hockley station. Synthetic seismograms (right panel) were computed using a reflectivity code for a model containing transversely isotropic layers with a horizontal axis of symmetry within the core mantle transition zone



Seismic, Reflectivity Method, Fig. 9 A comparison between reflectivity and 3D finite difference synthetic seismograms for a two-layer azimuthally anisotropic earth model



Seismic, Reflectivity Method, Fig. 10 (a) A model with two homogeneous layers over a half space; (b) τ - p - p' synthetic seismograms for the model shown in (a): note the shift in p value for the reflection from

the dipping interface; (c) shot gathers for the model shown in (a) – asymmetry in the reflection pattern from the second interface is modeled very well

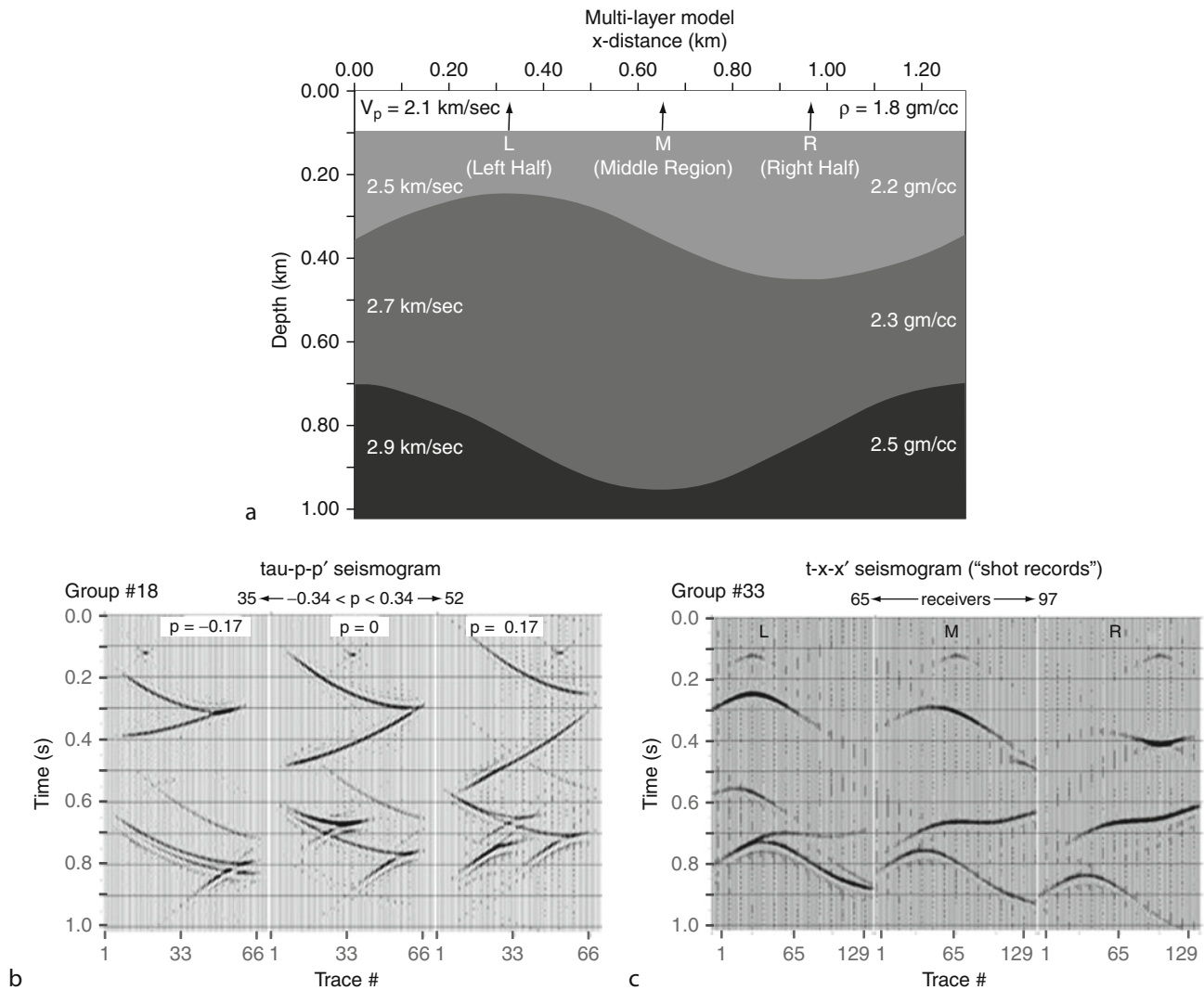
seismograms even with coarse sampling of the interfaces and ray-parameters. Approximate calculation of reflection/transmission coefficients, however, does not include multiple interaction of a plane wave with an interface.

Sen and pal (2009) performed computations in the coupled slowness domain and thus, multiple shot-receiver data can be synthesized rapidly. Intermediate results in the coupled slowness space provide important insight into understanding wave propagation in heterogeneous media. Figure 10 shows synthetic seismograms for a model with two homogeneous layers over a half space (Fig. 10a) such that the top interface is flat and the second interface is dipping with a constant dip. Figure 10b shows synthetic seismograms in (τ - p - p') domain; note that for the flat layer the scattered p is the same as the incident ray-parameter and the same for the dipping reflector is different from the incident ray-parameter. Synthetic shot

gathers for three shot locations shown in Fig. 10a are shown in Fig. 10c. These were all computed using a 2D reflectivity code in which internal multiples were not included. Similarly, in Fig. 11, I show synthetic seismograms for a two layer over half space model where both the interfaces are curved. Note the triplicated arrivals in the seismograms in both τ - p - p' domain (Fig. 11b) and shot gathers (Fig. 11c). Note the triplicated arrivals in the seismograms due to curved interfaces.

Summary

Seismic modeling is crucial to understanding seismic wave propagation and interpreting field seismic records. Several methods for seismic modeling are available; they vary in



Seismic, Reflectivity Method, Fig. 11 (a) A model with two homogeneous layers over a half space; (b) τ - p - p' synthetic seismograms for the model shown in (a); note the shift in p values and triplications for the

reflection from the curved interfaces; (c) shot gathers for the mode shown in (a) – triplications in the reflections modeled very well

accuracy, model assumptions, and computational speed. Of these, the reflectivity method is perhaps the most popular. The original reflectivity formulation is valid for 1D elastic media. It is capable of computing a complete response of a layered earth model. If needed, this can also be designed to compute responses of a few selected phases. The cost of computation grows with the number of layers, increase in frequency, and offset. Most isotropic models can now be computed fairly rapidly on standard desktop workstations of personal computers. Azimuthally anisotropic models require computations in two wave numbers and therefore can be computationally demanding. Parallel algorithms can be easily developed for rapid computation of reflectivity synthetic seismograms for large models. For 1D full waveform inversion, the reflectivity forward modeling is invoked a large number of times where a parallel reflectivity algorithm is also essential. The reflectivity

algorithm has been used in a wide variety of applications including seismic modeling in exploration and whole earth scales. Full waveform inversion methods based on reflectivity forward modeling are becoming increasingly popular (e.g., Sen and Roy 2003). The method has also been extended to modeling in laterally heterogeneous media and in electromagnetic modeling in layered media (Sena et al. 2008).

Cross-References

- [Numerical Methods, Finite Difference](#)
- [Numerical Methods, Finite Element](#)
- [Seismic Ray Theory](#)
- [Seismic Waves, Scattering](#)
- [Seismic, Waveform Modeling and Tomography](#)

Bibliography

- Aki K, Richards PG (2000) Quantitative seismology. W. H. Freeman and Company, San Francisco
- Bansal R, Sen MK (2008) Finite difference of S-wave splitting in anisotropic media. *Geophys Prospect* 56:293–312
- Booth DC, Crampin S (1983) The anisotropic reflectivity technique: theory. *Geophys J R Astr Soc* 72(3):755–766
- Cerveny V (2001) Seismic ray theory. Cambridge University Press, New York
- Chapman CH, Drummond R (1982) Body-wave seismograms in inhomogeneous media using Maslov asymptotic theory. *Bull Seism Soc Am* 72:S277–S317
- DeBasabe J, Sen MK (2007) Grid dispersion and stability criteria of some finite element methods for seismic wave propagation modeling. *Geophysics* 72(6):T81–T95
- Dunkin JW (1965) Computation of modal solutions in layered elastic media at high frequencies. *Bull Seis Soc Am* 55:335–338
- Fryer GJ (1980) A slowness approach to the reflectivity method of seismogram synthesis. *Geophys J R Astr Soc* 63(3):747–758
- Fryer GJ, Frazer LN (1984) Seismic waves in stratified anisotropic media. *Geophys J Int* 78(3):697–710
- Fuchs K, Mueller G (1971) Computation of synthetic seismograms with the reflectivity method and comparison with observations. *Geophys J R Astr Soc* 23:417–433
- Gilbert F, Backus GE (1966) Propagator matrices in elastic wave and vibration problems. *Geophysics* 31:326–332
- Jensen FB, Kuperman WA, Porter MB, Schmidt H (1993) Computational ocean acoustics. AIP Press, New York
- Kennett BLN (1983) Seismic wave propagation in stratified media. Cambridge University Press, Cambridge
- Koketsu K, Kennett BLN, Takenaka H (1991) 2-D reflectivity method and synthetic seismograms for irregularly layered structures – II. Invariant embedding approach. *Geophys J Int* 105:119–130
- Komatitsch D, Tromp J (1999) Introduction to the spectral element for three dimensional seismic wave propagation. *Geophys J Int* 139:806–822
- Liu Y, Sen MK (2009) Advanced finite-difference methods for seismic modeling. *Geohorizons* 14(2):5–16
- Mallick S, Frazer LN (1987) Practical aspects of reflectivity modeling. *Geophysics* 52:1355–1364
- Marfurt KJ (1984) Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations. *Geophysics* 49:533–549
- McCoy JJ, Fishman L, Frazer LN (1986) Reflection and transmission at an interface separating transversely inhomogeneous acoustic half spaces. *J R Astr Soc* 85:543–562
- Phinney RA (1965) Theoretical calculation of spectra of first arrivals in layered elastic mediums. *J Geo Res* 70:5107–5123
- Phinney RA, Odom RI, Fryer GJ (1987) Rapid generation of synthetic seismograms in layered media by vectorization of the algorithm. *Bull Seis Soc Am* 77:2218–2226
- Piera F, Zanzi L (1993) The reflectivity method as a tool for evaluating the seismic response of layered structures. *J Appl Geophys* 1-2:35–41
- Pulliam J, Sen MK (1998) Anisotropy in the core-mantle transition zone may indicate chemical heterogeneity. *Geophys J Int* 135:113–128
- Roy IG, Sen MK, Torres-verdin C (2005) Full waveform inversion using a distributed system of computers. *Concurr Comput Pract Exp* 17:1635–1685
- Schmidt H, Tango G (1986) Efficient global matrix approach to the computation synthetic seismograms. *Geophys J R Astr Soc* 84:331–359
- Sen MK (2006) Seismic inversion. Society of Petroleum Engineers, Richardson
- Sen MK, Frazer LN (1991) Multifold phase space path integral synthetic seismograms. *Geophys J Int* 104:479–487
- Sen MK, Pal A (2009) A reflectivity method for laterally varying media. *Geophys J Int* 178(2):792–812
- Sen MK, Roy IG (2003) Computation of differential seismograms and iteration adaptive regularization in pre-stack seismic inversion. *Geophysics* 89(6):2026–2039
- Sena A, Sen MK, Stoffa PL (2008) Modeling of ground penetrating radar data in stratified media using the reflectivity technique. *J Geophys Eng* 5:129–146
- Vamaraju J, Sen MK, DeBasabe J, Wheeler MF (2020) A hybrid Galerkin finite element for seismic wave propagation in fractured media. *Geophys J Int* 221(2):857–878
- Virieux J (1984) SH-wave propagation in heterogeneous media: velocity-stress finite-difference method. *Geophysics* 49:1933–1957

Seismic, Super-Virtual Refraction Interferometry

Sherif M. Hanafy

College of Petroleum Engineering and Geosciences (CPG),
King Fahd University of Petroleum & Minerals (KFUPM),
Dhahran, Saudi Arabia

Definition

Super-virtual refraction interferometry (SVI) is a seismic data processing method that enhances the signal-to-noise ratio (SNR) of head-waves generated from different sources and recorded by many receivers. The enhancement of the SNR can be equal to the square root of N , where N is the number of postcritical sources.

Introduction

Seismic interferometry estimates the detailed properties of the Earth by analyzing the interference patterns of seismic waves. These patterns are constructed by correlating or convolving traces, and then summing pairs of seismic traces to robustly image the Earth's elastic properties (Schuster 2009).

The reciprocity equation of correlation type for a two-state system (Schuster 2009) is given as

$$G(y|x) - G_o(x|y)^* = \int_{S_o} \left[G_o(x|s)^* \frac{\partial G(y|s)}{\partial n_s} - G(y|s) \frac{\partial G_o(x|s)}{\partial n_s} \right] d^2s, \quad (1)$$

where x and y represent receiver positions, s represents the location of the physical shot, $G(b|a)$ is the Green's function for the actual Earth model in the frequency domain for a source located at a and a receiver located at b , and S_o is the

integration boundary. Here, the receiver line should be slightly below the source line, while the integration along the free surface vanishes because both Green's functions are zero. $G_o(x|y)$ is the acausal Green's function, and n is an outward pointing unit vector normal to the boundary.

The far-field approximation to eq. (1) is given as

$$G(y|x) - G_o(x|y)^* \approx 2ik \int_{S_o} G(y|s) G_o(x|s)^* dx^2. \quad (2)$$

An important property of the correlation is that the phases in the spectral product are subtractive, so that subtracting the traveltimes leads to shorter traveltime and earlier events with shorter raypaths. On the contrary, temporal convolution leads to the spectral product with additive phases. This means that the resulting trace has a longer traveltime and, hence, the new event has a longer raypath. For more information on seismic interferometry, see Schuster (2009) and Wapenaar et al. (2010).

The integration boundary should completely surround the receivers at x and y , however, for practical reasons, this is not possible; hence, artifacts may occur due to incomplete cancellation of nonphysical events. However, eq. 2 is considered acceptable for seismic interferometry since the approximation does not affect the phase (Wapenaar et al. 2010).

One of the major benefits of seismic interferometry is that velocity model or statics are not required for the redatuming because the data act as natural wavefield extrapolators. Actually, the scatterers in the medium have the effect of virtually enlarging the aperture of the recording array, which leads to what is known as super-resolution of the image (Leroy et al. 2007).

Interferometry helps in the surface-wave inversion of earthquake data (Gerstoft et al. 2006), surface-wave prediction and elimination (Halliday et al. 2007), trace interpolation (Hanafy and Schuster 2014) and extrapolation, refraction interferometry (Dong et al. 2006), passive seismic imaging (Draganov et al. 2004), and super-illumination (He et al. 2006) of the subsurface. In this entry, we will discuss the enhancement of the noisy signals through the use of natural Green's functions, which is known as super-virtual refraction interferometry (SVI).

Virtual Refraction Interferometry

In a refraction acquisition experiment where the sources and receivers are located on the ground surface, assume that the subsurface seismic velocities are homogenous, then the first arrival traveltime from a source located at s to a receiver located at x (left panel of Fig. 1a) is a head wave and can be approximated by

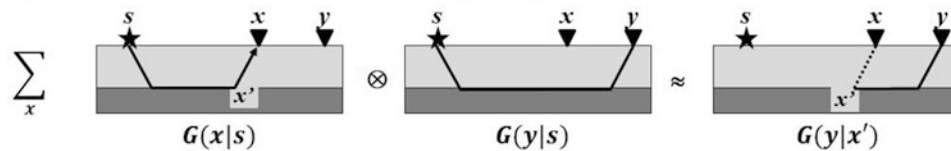
$$u(s, x) = A(s, x) e^{i\omega(\tau_{ss'} + \tau_{s'x})} \quad (3)$$

The amplitude term $A(s, x)$ in equation (3) accounts for both the geometrical spreading and the source wavelet, $\tau_{ss'}$ is the traveltime of the seismic wave that propagates from the surface source (s) to a point x' along the refractor with angular frequency ω , and $\tau_{s'x}$ is the traveltime from x' to the surface receiver located at point x .

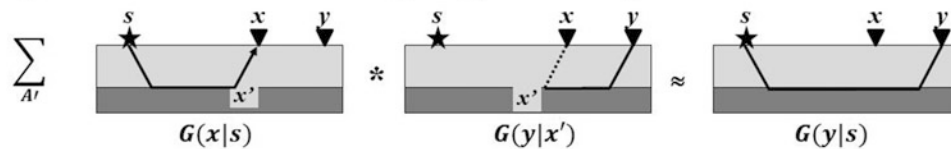
Dong et al. (2006) redatum the surface sources (s) to the refractor by cross-correlating two nearby traces; here, the source is redatumed to the refractor at point x' (eq. 4 and Fig. 1a),

Super-virtual Refraction Interferometry

a) Crosscorrelation and stacking to generate virtual refractions



b) Convolution and stacking to generate super-virtual refractions



Seismic, Super-Virtual Refraction Interferometry, Fig. 1 (a) The two traces $u(s, x)$ and $u(s, y)$ recorded at the receivers in the left and middle velocity models are cross-correlated to generate the virtual trace $D_s^v(x, y)$ (see right velocity model). To enhance the SNR we stack over different source positions to generate the virtual trace $d^v(x, y)$. (b)

The virtual trace $d^v(x, y)$, recorded at y in the middle model, is now convolved with the recorded trace $u(s, x)$ in the left model, to generate the super-virtual refraction trace $D_s^{sv}(s, y)$ recorded at y in the right model. To enhance the SVI trace, we stack the traces for all locations (x) which are between s and y

$$\begin{aligned}
D_s^v(x, y) &= u(s, y)u(s, x)^*, \\
&= |A(s, x)|^2 e^{i\omega(\tau_{sx'} + \tau_{x'y} - \tau_{sx} - \tau_{x'y})}, \\
&= |A(s, x)|^2 e^{i\omega(\tau_{x'y} - \tau_{x'x})}, \tag{4}
\end{aligned}$$

where $D_s^v(x, y)$ is a virtual trace recorded at the physical receiver y and excited at a virtual source with unknown location x' ; the unknown excitation time is equal to the negative of the time for waves to propagate from point x' to point x . Here, we assume that the amplitude terms are roughly the same for each receiver position. The $\tau_{x'y}$ and $\tau_{x'x}$ terms in equation 4 represent the traveltime difference between a direct wave measured at x and a refraction arrival at y for a buried source located on the refractor at x' (Fig. 1a). The traveltime calculated by equation (4) is the same for any source position s located at the surface that generates head waves at receivers x and y from the same refractor. Therefore, these correlated traces for different source positions can be stacked together to give an N -fold stacked trace to enhance the signal-to-noise ratio (SNR), i.e.,

$$\begin{aligned}
d^v(x, y) &= \sum_{s=1}^N D_s^v(x, y), \\
&= e^{i\omega(\tau_{x'y} - \tau_{x'x})} \sum_{s=1}^N |A(s, x)|^2, \tag{5}
\end{aligned}$$

where $d^v(x, y)$ is the stacked virtual interferometry trace for a virtual source located at x' and arrivals recorded by a receiver located at y (Fig. 1a). Equation 5 shows an enhancement of the SNR equal to $\sqrt{N}$, where N is the number of stacked traces.

To avoid the artifacts due to a limited recording aperture and discrete sampling, Dong et al. (2006) suggested windowing about the first arrival traveltimes so that only the head-wave arrivals are correlated with one another.

Advantages. The advantage of the virtual refraction interferometry is that the correlated traces show improvement in the SNR equal to $O(\sqrt{N})$.

Limitations. The virtual refraction interferometry method suffers from the following limitations: (a) the source location is redatumed to the unknown position in the subsurface (Fig. 1a, right panel), (b) the excitation time of the correlated trace is unknown, the new excitation time after the redatuming step is equal to the negative traveltime from point x to point x' (the dashed line shown on Fig. 1a, right panel), and (c) the length of the raypath of the correlated trace is shorter than the original lengths of the input traces, which decreases the depth of penetration. These problems can be remedied by selecting the time of a visible refraction arrival

and adjusting the times of the SVI traces so that the virtual sources are automatically redatumed to the surface.

Super-Virtual Refraction Interferometry (SVI)

Bharadwaj and Schuster (2010) and Mallinson et al. (2011) suggested an alternative solution to the problem associated with the unknown excitation time and unknown source location, where they convolved the virtual traces $d^v(x, y)$ in equation 5 with the real trace $u(s, x)$ recorded at x excited by a source located at s (Fig. 1b),

$$\begin{aligned}
D_x^{sv}(s, y) &= u(s, x)d^v(x, y), \\
&= |A(s, x)| e^{i\omega(\tau_{sx'} + \tau_{x'y})} e^{i\omega(\tau_{x'y} - \tau_{x'x})} \sum_{s=1}^N |A(s, x)|^2, \\
&= |A(s, x)| \sum_{s=1}^N |A(s, x)|^2 e^{i\omega(\tau_{sx'} + \tau_{x'y})}, \tag{6}
\end{aligned}$$

The SVI head-wave arrivals, $D_x^{sv}(s, y)$ given by equation 6, are kinematically identical to those seen in traces with a high SNR and recorded by a receiver at y with a source at s . This process can be repeated for all M postcritical positions of x which are located between s and y ; hence, a total of M traces for the same source and receiver positions are generated. Then, similar to the virtual refraction interferometry, we stack these traces to enhance the output trace which has a known surface source position, and an enhanced SNR compared to the recorded trace (Bharadwaj et al. 2012, Fig. 1b):

$$\begin{aligned}
d^{sv}(s, y) &= \sum_{x=1}^M D_x^{sv}(s, y), \\
&= \sum_{x=1}^M |A(s, x)| \sum_{s=1}^N |A(s, x)|^2 e^{i\omega(\tau_{sx'} + \tau_{x'y})}, \tag{7}
\end{aligned}$$

The output of equation 7 is the “*super-virtual*” head-wave refraction arrival at y for a source located at s , as shown in Fig. 1b. Here, the location of the source point is known and located at point s , and the source excitation time is now equal to zero (Hanafy et al. 2011; Alshuhail et al. 2012).

Advantages. The SVI approach is an alternative solution to the limitations of the virtual refraction interferometry: (a) The source location is now redatumed back to the ground surface at point s , (b) the excitation time of the SVI is equal to zero, and (c) the lengths of the raypaths associated with the SVI traces are the same as the original raypath lengths of the input traces. In addition, the SNR of the SVI traces is

improved by an additional $\sqrt{M}$, where M is the number of stacked traces in the convolution step.

Limitations. The limitations of the SVI include: (a) we need to mute around the expected first-arrival traveltimes because the smaller the window the better the SVI result and (b) SVI is valid only in the case of head waves. However, it can also work if the velocity gradients only generate diving that does not dip too steeply.

SVI in Practice

The SVI method can be summarized (as shown in Fig. 2a–c) in the following steps:

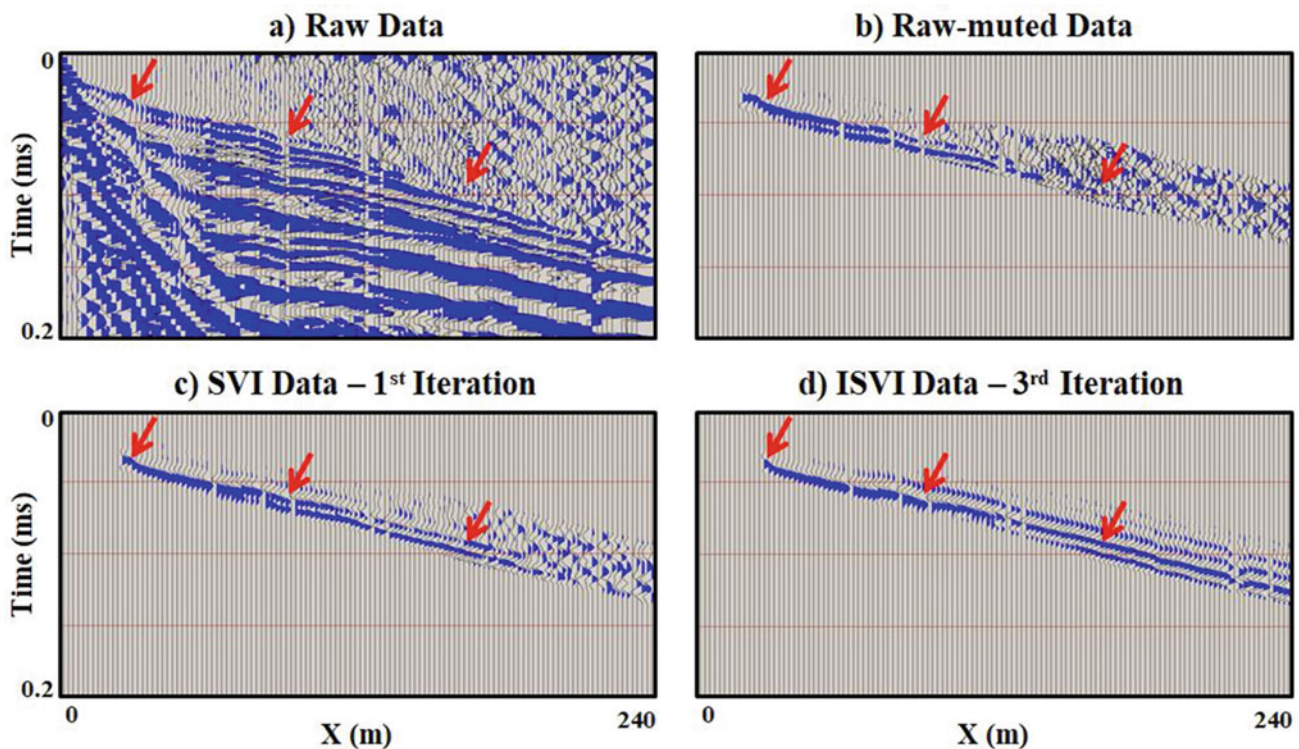
1. On the raw data (Fig. 2a), a time window is identified that includes the head waves at both the near- and far-offset traces.
2. Mute the arrivals outside the window selected in step 1 (Fig. 2b), so that only the head wave arrivals are correlated with one another. This step is essential in order to suppress unwanted artifacts related to the correlations between reflections, direct arrivals, and refractions. The length of the window should be 2–3 periods long.
3. Generate the virtual interferometry data set by cross-correlating the muted traces together (eq. 4); the two

correlated traces must be from the same shot point (s), and the head waves must propagate along the same refractor (Fig. 1a).

4. Stack the virtual traces that share the same virtual source (x') and same receiver point (y) to enhance the virtual refractions (eq. 5).
5. Convolve the stacked virtual interferometry data set with the recorded traces to produce the super-virtual refraction interferometry (SVI) data set (eq. 6). The convolved traces must share the same virtual source point (x'), and the associated head waves must refract along the same refractor (Fig. 1b).
6. Stack the SVI traces that share the same source (s) and same receiver point (y) to enhance the SNR of the SVI arrivals (Fig. 2c and eq. 7).

For refractions propagating along from more than one refractor, the data are divided into subsections, and each subsection contains traces refracted from only one refractor. Then the SVI method is applied to each subsection separately, and the SVI data sets of all sections are blended together to get the final SVI data set.

In general, the SVI traces are more ringy than the raw traces because the amplitude spectrum of the wavelet (eq. 7) is raised to the third power. This might lead to ambiguous



Seismic, Super-Virtual Refraction Interferometry, Fig. 2 A field example from the Qademah data set recorded 20 km north of KAUST, Saudi Arabia: (a) raw CSG; (b) same as (a) after muting around the first

arrival traveltimes; (c) the result of SVI, 1st iteration; and (d) the result of the ISVI after three iterations. The red arrows point to the location of the first arrival traveltimes. (Modified from Fig. 7, Al-Hagan et al. 2014)

identifications of the first arrivals (Fig. 2c). To solve this problem, the raw traces with high SNR are used as guidance for better identification of the first arrival traveltimes in the SVI traces (Note red arrows in Fig. 2b, c).

The SVI method has been tested using synthetic and field 3D data set (Lu et al. 2014, Edigbue and Al-Shuhail 2016). There are several differences between the 2D and 3D SVI approaches, where the stationary positions of the source-receiver pairs could be any place along the x-y plane, also the SNR enhancing effect could be limited due to the limited number and irregularity of the survey lines and azimuthal range of arrivals. Lu et al. (2014) suggested integrating along the source or receiver lines, which enhances the SNR without the requirement of knowing the stationary locations. In addition, the amplitudes SVI traces are largely strengthened by integration, which is helpful to further enhance the SNR.

Iterative Super-Virtual Refraction Interferometry

There are some cases where the SNR enhancement from SVI is insufficient for accurate picking of noisy far-offset arrivals. This noise might result from weak sources shooting into distant receivers, and the area is polluted with environmental noise, etc. In such cases, Al-Hagan et al. (2014) suggested that the SVI traces can be used as an input to another round of SVI processing (eq. 8), and the output will be traces with an increase in SNR (Fig. 2d). They called this procedure iterative super-virtual refraction interferometry (ISVI),

$$d^{(k+1)} = L[d^{(k)}] \quad (8)$$

where $L[]$ represents the normalized super-virtual interferometry operations of summed correlation and convolution in equations 5 and 7, where $d^{(0)}$ at $k = 0$ represents the raw windowed traces and the output ($d^{(1)}$) is the super-virtual data obtained by the sequential application of equations 5 and 7. Then, the input to the 2nd SVI iteration at $k = 1$ is the SVI output $d^{(1)}$ to produce $d^{(2)}$, and this procedure is repeated until acceptable convergence.

Al-Hagan et al. (2014) recommended that the number of iterations should not exceed 3 ($k \leq 3$). For large number of iterations, the ISVI data become overly smoothed.

Advantages. ISVI shares the same advantages as the SVI, except it produces a higher SNR.

Limitations. ISVI also shares the same limitations as the SVI, and it also shows more ringing in the ISVI traces by increasing the number of iterations (note red arrows in Fig. 2d). To partially solve this problem, we can use a deconvolution filter (Snieder and Safak 2006, Vasconcelos and Snieder 2008; Al-Hagan et al. 2014), which is applied to the virtual traces after the correlation step to reduce the ringing in the first arrivals.

Very Noisy Data

In case of very noisy traces, it could be very difficult to identify a small (2–3 periods) window around the first arrival traveltimes. In these cases, Hanafy (2019) suggested the use of a 2-stage ISVI procedure.

1. In the first stage a very large window is identified to make sure that the first arrival traveltime is included. Then the ISVI with 2-iterations is applied on the windowed data. The result of the first stage shows ISVI data with slightly improved SNR, however, the SNR is not good enough for the first arrival traveltime picking.
2. In the second stage we use the ISVI data set of the first stage to identify a smaller window around the first arrival traveltimes. The updated window is now used to mute around the first arrival traveltime in the raw data, and then the ISVI is applied again to the muted-raw data using the small window. In this case, we use 2 or 3 ISVI iterations.

Summary

Super-virtual refraction interferometry (SVI) is used to enhance the SNR of noisy head waves recorded at the far offsets. The SVI starts with correlating nearby traces to generate the virtual refraction data traces, which are then convolved with the recorded data to generate the SVI data set. The SVI enhancement can be as large as $O(\sqrt{M})$, where M is the number of stacked traces in the convolution step. SVI procedure can be applied iteratively to enhance very low SNR data, however, a maximum of three iterations is recommended in this case. SVI is valid in case of head waves, and it can be applied to data recorded at any velocity model; however, it fails if the first arrival traveltimes mainly consist of diving waves with steep arrival angles. Traces from different refractors need to be handled separately to decrease the artifacts in the final SVI data set.

Bibliography

- Al-Hagan O, Hanafy SM, Schuster GT (2014) Iterative supervirtual refraction interferometry. *Geophysics* 79(3):Q21–Q30
- Alshuhail A, Aldawood A, Hanafy S (2012) Application of super-virtual seismic refraction interferometry to enhance first arrivals: a case study from Saudi Arabia. *TLE* 31:34–39. <https://doi.org/10.1190/1.3679326>
- Bharadwaj P, Schuster GT (2010) Extending the aperture and enhancing the S/N ratio of refraction imaging by super-virtual interferometry. In: Abstract S23C-04 presented at 2010 Fall Meeting, AGU
- Bharadwaj P, Schuster G, Mallinson I, Dai W (2012) Theory of super-virtual refraction interferometry. *Geophys J Int* 188:263–273. <https://doi.org/10.1111/j.1365-246X.2011.05253.x>

- Dong S, Sheng J, Schuster GT (2006) Theory and practice of refraction interferometry. In: 76th SEG annual meeting, Expanded Abstracts, pp 3021–3025
- Draganov D, Wapenaar K, Artman B, Biondi B (2004) Migration methods for passive seismic data. In: 74th Annual international meeting, SEG expanded abstracts, pp 1123–1126
- Edigbue P, Al-Shuhail A (2016) Three-dimensional supervirtual seismic refraction interferometry. *Arab J Geosci* 9:537–549
- Gerstoft P, Sabra K, Roux P, Kuperman WA, Fehler M (2006) Green's functions extraction and surface-wave tomography from microseisms in southern California. *Geophysics* 71:SI23–SI32
- Halliday D, Curtis A, Robertsson J, van Manen D (2007) Interferometric surface-wave isolation and removal. *Geophysics* 72:A69–A73
- Hanafy SM (2019) Iterative super-virtual refraction interferometry and Traveltime tomography of seismic data: field example at Gulf of Aqaba. EAGE meeting, London, UK, 3–6 June 2019
- Hanafy SM, Schuster G (2014) Interferometric interpolation of sparse marine data. *Geophys Prospect* 62:1–16. <https://doi.org/10.1111/1365-2478.12066>
- Hanafy SM, AlHagan O, Al-Tawash F (2011) Super-virtual refraction interferometry: field data example over a Colluvial Wedge. In: 81st Annual international meeting, SEG expanded abstracts, pp 3814–3818
- He R, Hornby B, Schuster GT (2006) 3D wave-equation interferometric migration of VSP multiples. In: 76th annual international meeting, SEG expanded abstracts, pp 3442–3445
- Lerosey G, de Rosny J, Tourin A, Fink M (2007) Focusing beyond the diffraction limit with far-field time reversal. *Science* 315:1120–1122
- Lu K, AlTheyab A, Schuster GT (2014) 3D super-virtual refraction interferometry. In: 84th annual international meeting, SEG expanded abstracts. <https://doi.org/10.1190/segam2014-0822.1>
- Mallinson I, Bharadwaj P, Schuster GT, Jakubowicz H (2011) Enhanced refractor imaging by super-virtual interferometry. *Lead Edge* 30:546–550. <https://doi.org/10.1190/1.3589113>
- Schuster GT (2009) *Seismic interferometry*. Cambridge University Press, Cambridge
- Snieder R, Safak E (2006) Extracting the building response using seismic interferometry: theory and application to the Millikan Library in Pasadena, California. *Bull Seismol Soc Am* 96:586–598. <https://doi.org/10.1785/0120050109>
- Vasconcelos I, Snieder R (2008) Interferometry by deconvolution, Part 1 – theory for acoustic waves and numerical examples. *Geophysics* 73(3):S115–S128. <https://doi.org/10.1190/1.2904554>
- Wapenaar K, Slob E, Snieder R, Curtis A (2010) Tutorial on seismic interferometry: Part 2–Underlying theory and new advances. *Geophysics* 75(5):75A211–75A227

Seismic, Velocity, and Density Relationships

Kalachand Sain
Wadia Institute of Himalayan Geology, Dehradun, India

Definition

Seismic velocity is defined as the speed with which a seismic wave travels through a medium and is measured in km/s. The density is the mass per unit volume and is expressed in g/cc.

The relationship between seismic velocity and density is very much required for subsurface geological interpretation

and reservoir characterization. As a rock property, density plays an important role for the identification of lithologies, delineation of pore fluids, estimation of porosity and overburden stress, and prediction of well planning. It is possible to derive velocity-depth function from seismic data, but the density variation along a well path can be hardly obtained from gravity data until a well is drilled. Hence, density is calculated from seismic velocities using a set of relationships (Birch 1961; Gardner et al. 1974; Lindseth 1979), if the well information is not available.

Types of seismic waves: The *body wave* and the *surface wave* are the main seismic waves. The body waves are the longitudinal (P) and the shear (S) waves in which the particles vibrate parallel and perpendicular to the direction of wave propagation, respectively. The velocities of P-wave (V_P) and S-wave (V_S) are related to the density, ρ as

$$V_P = \sqrt{\frac{\kappa + \frac{4}{3}\mu}{\rho}} \quad (1)$$

$$V_S = \sqrt{\frac{\mu}{\rho}} \quad (2)$$

Where κ and μ are the bulk and rigidity moduli, respectively. It appears that the seismic velocity and density are inversely related, but the denser rocks are characterized by higher velocities compared to lighter rocks, as the density also depends on κ and μ . The halite with low density of 1.8 g/cc and high V_P of 4.5 km/s shows an exception. Generally, velocities and densities increase with depth. The inversion of this trend is also observed due to pore pressure, fluids, or sediments below volcanic rocks.

General Relation Between Velocity and Density

- (a) Birch's law – To fit measurements from crustal and mantle rocks, Birch (1961) established a linear relationship between V_P and ρ as

$$V_P = a + b\rho \quad (3)$$

Where a and b are empirical parameters. This relation was tested at high pressure by Campbell and Heinz (1992).

- (b) Gardner relation – Gardner et al. (1974) conducted field and laboratory measurements on saturated sedimentary rocks from a wide-range of basins, ages, and depths and established the velocity-density relation as

$$\rho = aV_P^b \quad (4)$$

Where $a = 0.31$ and $b = 0.25$ are Gardner coefficients. Though many sedimentary rocks (Shale, Sandstone,

Limestone, and Dolomite) fall within a narrow corridor around the Gardner curve, Coals, Anhydrite, Salts, etc. exhibit large deviations from this trend. The deviation increases with porosity and becomes higher for lower densities.

- (c) Based on data of Gardner et al. (1974), Lindseth (1979) proposed an empirical relation between V_p and acoustic impedance to extract the density from the reflectivity as

$$V_p = 0.308 \rho V_p + 3460$$

Krasovsky (1981) showed that the Gardner equation is non-unique for many sedimentary basins in the world, and suggested for local calibration to improve the accuracy of rock property prediction. As a way forward, Ojha and Sain (2014) and Nwozor et al. (2017) used the well logs data in the Kerala-Konkan (India) and Niger Delta (Egypt) basins, respectively, to derive the lithology-specific Gardner coefficients for geological analyses accurately. The sands usually trend lower densities than shales of corresponding velocity due to their different mineral composition, and shales are more responsive to compaction than sands. Though the Gardner coefficients are calibrated from sonic and density well log information, but in the absence of these, Gardner coefficients are still used in petroleum exploration.

Estimation of seismic velocity and density from surface measurements provides important inputs in understanding the structure and tectonics of a geotectonic province. Since gravity is a potential field, velocity structure is often used to constrain the gravity modeling (Barton 1986; Ravat et al. 1999; Behera et al. 2004; Boore 2016; Emudianughe and Ahmad 2019) by converting the velocity into density based on a relationship. Use of Ludwig (1970) formula to relate V_p with ρ is also very common in crustal studies.

Cross-References

- [Body Waves](#)
- [Surface Waves](#)

Acknowledgement The author is grateful to the Director of Wadia Institute of Himalayan Geology, Dehradun, for his consent to publish this work. This has a Wadia contribution WIHG/0035.

Bibliography

- Barton PJ (1986) The relationship between seismic velocity and density in the continental crust – a useful constraint? *Geophys J Int* 87:195–208
- Behera L, Sain K, Reddy PR (2004) Evidence of underplating from seismic and gravity studies in the Mahanadi delta and its tectonic significance. *J Geophys Res* 109:1–25
- Birch F (1961) The velocity of compressional waves in rocks to 10 kilobars. *J Geophys Res* 66:2199–2224
- Boore D (2016) Determining generic velocity and density models for crustal amplification calculations, with an update of the Boore and Joyner (1997) Generic site amplification for graphic site amplification. *Bull Seismol Soc Am* 106(1):306–320
- Campbell AZ, Heinz BJ (1992) A high pressure testing of Birch's law. *Science* 257:66–68
- Emudianughe JE, Ahmad MM (2019) Comparative analysis of Gardner relation in lithology discrimination for density modeling in the Niger Delta. *J Sci Eng Res* 6(6):22–27
- Gardner GHF, Gardner LW, Gregory AR (1974) Formation velocity and density – the diagnostic basics for stratigraphic traps. *Geophysics* 39:770–780
- Krasovsky SS (1981) Reflection of continental-type crustal dynamics in the gravity field. *Naukova Dumka, Kiev*
- Lindseth RO (1979) Synthetic sonic logs – a process for stratigraphic interpretation. *Geophysics* 44:3–26
- Ludwig WJ, Nafe JE, Drake CL (1970) Seismic refraction. In: Maxwell (ed) *The sea*, vol 4(1). Wiley-Intersciences, New York, pp 53–84
- Nwozor KK, Onuorah LO, Onyekuru SO, Egbuachor CJ (2017) Calibration of Gardner coefficient for density-velocity relationships of tertiary sediments in Niger Delta basin. *J Pet Explor Prod Technol* 7:627–635
- Ojha M, Sain K (2014) Velocity-porosity and velocity-density relationship for shallow sediments in the Kerala-Konkan basin of western Indian margin. *J Geol Soc India* 84:187–191
- Ravat D, Lu Z, Braile LW (1999) Velocity-density relationships and modeling the lithospheric density variations of the Kenya Rift. *Tectonophysics* 302:225–240

Seismic, Waveform Modeling and Tomography

Yanghua Wang¹ and Ying Rao²

¹Resource Geophysics Academy, Department of Earth Science and Engineering, Imperial College London, London, UK

²College of Geophysics, China University of Petroleum (Beijing), Beijing, China

Synonyms

Waveform inversion

Definition

Tomography. A seismic inversion method to produce slicing image of the internal structures of an object, by recording seismic wavefield propagating through and scattered-refracted-reflected back from the object and observing the difference in the effects on the wave energy impinging on those structures.

Introduction

Seismic waveform tomography is an imaging technology using seismic data to reconstruct the Earth subsurface physical properties, represented by seismic velocity, attenuation coefficient, etc. It is usually formulated as an inverse problem (Wang 2016). Many seismic tomography methods use the arrival time or amplitude information extracted from the waveform recordings as the input data to an inversion process. These methods are generically referred to as travel time or amplitude tomography (Wang 2003). Comparing to either travel time or amplitude tomography, waveform tomography has at least two major features. First, the input data consist of the seismic waveforms themselves, as opposed to travel times, amplitudes, or some other secondary attributes of the recorded data. This makes waveform tomography a better resolution than travel time or amplitude tomography. Second, the underlying numerical method is based on the full-wave equation, as opposed to a ray approximation or a Born approximation. This makes waveform tomography more accurate than the travel time or amplitude counterpart. These two features also make the inverse problem of waveform tomography more difficult to solve.

However, as field seismic data usually lack low-frequency information, travel time or amplitude tomography may provide a reliable starting model for the iterative waveform inversion. If waveform tomography is implemented in the frequency domain, the inversion process proceeds from low- to high-frequency components. In the time domain, a band-pass filter can be applied to the seismic data, and inversion uses band-pass filtered data with low frequencies first and proceeds to higher frequencies (Rao and Wang 2017). Even for the frequency-domain implementation, in order to suppress the low signal-to-noise ratio effect of frequency data and to make the inversion procedure robust, waveform tomography usually uses a group of frequencies simultaneously in an iterative inversion and proceeds from low- to high-frequency groups, generating a high-resolution image of the subsurface model (Wang and Rao 2006).

Waveform tomography has been used successfully on transmission data, such as crosshole seismic data, and to wide-angle reflection/refraction seismic data with a certain degree of success. It has also been used for regional-scale studies on the crustal and upper mantle velocity structure, using scattering waves, surface waves, and SH-waves from either exploration seismic data or broadband teleseismograms. In addition, there is a good progress on its application to seismic reflection data with limited source-receiver offsets (Wang and Rao 2009). The seismic reflection method is a routine practice in the hydrocarbon exploration, and the data dominated by the precritical reflection energy, reflected back from subsurface contrasts in physical parameters, are well suited for seismic migration for the structural image. Applying the waveform tomography

technique to these reflection data can quantitatively extract the geophysical parameters, for identifying different lithologies and different fracture characteristics and even for indicating the hydrocarbon distribution directly.

Waveform Modeling

Waveform modeling is the key engine in seismic waveform tomography. Seismic wave propagation satisfies both Newton's second law and Hook's law and can be expressed by the following five simultaneous first-order differential equations for a two-dimensional (2D) case:

$$\begin{aligned}\rho \frac{\partial u}{\partial t} &= \frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xz}}{\partial z}, \\ \rho \frac{\partial w}{\partial t} &= \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{zz}}{\partial z}, \\ \frac{\partial \tau_{xx}}{\partial t} &= c_{11} \frac{\partial u}{\partial x} + c_{13} \frac{\partial w}{\partial z} + c_{15} \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right), \\ \frac{\partial \tau_{zz}}{\partial t} &= c_{31} \frac{\partial u}{\partial x} + c_{33} \frac{\partial w}{\partial z} + c_{35} \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right), \\ \frac{\partial \tau_{xz}}{\partial t} &= c_{51} \frac{\partial u}{\partial x} + c_{53} \frac{\partial w}{\partial z} + c_{55} \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right),\end{aligned}\quad (1)$$

where u and w are the particle velocity components in the horizontal and vertical direction, respectively; τ_{xx} , τ_{xz} , and τ_{zz} are stress components; ρ is density; and

$$\begin{bmatrix} c_{11} & c_{13} & c_{15} \\ c_{31} & c_{33} & c_{35} \\ c_{51} & c_{53} & c_{55} \end{bmatrix}\quad (2)$$

are the elastic constants in Hook's law relating stress to strain in the 2D case. The first two equations are derived from Newton's second law and the last three from Hooke's law for an elastic medium. Eliminating stress components, these five simultaneous equations can be coupled into two equations:

$$\begin{aligned}-\rho \frac{\partial^2 u}{\partial t^2} + \frac{\partial}{\partial x} \left(c_{11} \frac{\partial u}{\partial x} + c_{15} \frac{\partial w}{\partial x} \right) + \frac{\partial}{\partial x} \left(c_{15} \frac{\partial u}{\partial z} + c_{13} \frac{\partial w}{\partial z} \right) \\ + \frac{\partial}{\partial z} \left(c_{51} \frac{\partial u}{\partial x} + c_{55} \frac{\partial w}{\partial x} \right) + \frac{\partial}{\partial z} \left(c_{55} \frac{\partial u}{\partial z} + c_{53} \frac{\partial w}{\partial z} \right) = 0, \\ -\rho \frac{\partial^2 w}{\partial t^2} + \frac{\partial}{\partial x} \left(c_{51} \frac{\partial u}{\partial x} + c_{55} \frac{\partial w}{\partial x} \right) + \frac{\partial}{\partial x} \left(c_{55} \frac{\partial u}{\partial z} + c_{53} \frac{\partial w}{\partial z} \right) \\ + \frac{\partial}{\partial z} \left(c_{31} \frac{\partial u}{\partial x} + c_{35} \frac{\partial w}{\partial x} \right) + \frac{\partial}{\partial z} \left(c_{35} \frac{\partial u}{\partial z} + c_{33} \frac{\partial w}{\partial z} \right) = 0,\end{aligned}\quad (3)$$

which are defined in terms of horizontal wave component u and vertical wave component w . These equations are

valid for the Earth media with arbitrary anisotropy and heterogeneity.

In the forward modeling of seismic wave propagation based upon either five or two simultaneous equations, there has been a steady development in the finite-difference implementations. As the time derivatives are approximated by $\partial u/\partial t \approx -(u_{i+1} - u_i)/\Delta t$ and $\partial^2 u/\partial t^2 \approx (u_{i+1} - 2u_i + u_{i-1})/\Delta t^2$, where i is the time index, $t_i = i\Delta t$, and Δt is the step size in time; the wavefield u_{i+1} at the current time can be calculated based on the previous wavefield: $u_{i+1} = f(u_i)$ or $u_{i+1} = f(u_i, u_{i-1})$. This is an explicit finite-difference scheme (Alford et al. 1974; Kelly et al. 1976). For the spatial derivatives, a second-order staggered grid scheme is widely used for numerical stability (Virieux 1986) and can also be extended to a fourth-order scheme for greater accuracy (Levander 1988).

In the frequency domain, the two simultaneous equations become

$$\begin{aligned} \omega^2 \rho u + \frac{\partial}{\partial x} \left(c_{11} \frac{\partial u}{\partial x} + c_{15} \frac{\partial w}{\partial x} \right) + \frac{\partial}{\partial x} \left(c_{15} \frac{\partial u}{\partial z} + c_{13} \frac{\partial w}{\partial z} \right) \\ + \frac{\partial}{\partial z} \left(c_{51} \frac{\partial u}{\partial x} + c_{55} \frac{\partial w}{\partial x} \right) + \frac{\partial}{\partial z} \left(c_{55} \frac{\partial u}{\partial z} + c_{53} \frac{\partial w}{\partial z} \right) = 0, \\ \omega^2 \rho w + \frac{\partial}{\partial x} \left(c_{51} \frac{\partial u}{\partial x} + c_{55} \frac{\partial w}{\partial x} \right) + \frac{\partial}{\partial x} \left(c_{55} \frac{\partial u}{\partial z} + c_{53} \frac{\partial w}{\partial z} \right) \\ + \frac{\partial}{\partial z} \left(c_{31} \frac{\partial u}{\partial x} + c_{35} \frac{\partial w}{\partial x} \right) + \frac{\partial}{\partial z} \left(c_{35} \frac{\partial u}{\partial z} + c_{33} \frac{\partial w}{\partial z} \right) = 0, \end{aligned} \quad (4)$$

where ω is the angular frequency. This system of equations can be solved in parallel for individual frequencies, using also a finite-difference method. Once all frequency components of the wavefield at any spatial position are obtained, performing inverse Fourier transform, with respect to the frequency variable, will produce a time-domain seismic trace at this spatial position.

However, the frequency-domain implementation is much more time-consuming, compared to its counterpart in the time domain. This is because the frequency-domain calculation involves the solution of linear algebraic equations as the follows. Using a finite-difference method, Equation (4) may be presented in a vector-matrix form as

$$\mathbf{M}\mathbf{u} = \mathbf{s}, \quad (5)$$

where $\mathbf{M}$ is a matrix approximating the partial differential operators, $\mathbf{u}$ is a vector representing the two components of the wavefield at all grids, and $\mathbf{s}$ is a vector representing the source term, which is zero everywhere except at the location of the source (added to the right-hand side of the wave equations). To obtain wavefield $\mathbf{u}$, one needs to solve the inverse matrix $\mathbf{M}^{-1}$. This is therefore an implicit finite-difference scheme.

The matrix $\mathbf{M}$ represents a significant storage requirement, which is largely determined by the numerical bandwidth of $\mathbf{M}$ and by the manner in which the structural sparsity of the matrix is maintained in any solution method. The bandwidth of the differencing matrix is determined by the number of nodes needed for a spatial derivative. A minimal, rotated computational star is a scheme that can minimize the bandwidth of the matrix and meanwhile improve the accuracy for finite-difference modeling. It does not require any new grid points as used in a second-order finite differencing but can produce an equivalent of fourth-order accuracy in finite differencing (Štekl and Pratt 1998).

While it is feasible to calculate the inverse matrix $\mathbf{M}^{-1}$ for a small, 2D model, in many practical problems such as 3D seismic modeling, the system $\mathbf{M}\mathbf{u} = \mathbf{s}$ with a large, sparse matrix is solved iteratively rather than directly. An approximate solution $\tilde{\mathbf{u}}$ is first obtained by solving using a related, preconditioning matrix. This preconditioner is chosen such that it is related to the true matrix but is much faster to invert. The approximate solution $\tilde{\mathbf{u}}$ from the preconditioner is then substituted into the true matrix equation to discover the effective source, $\tilde{\mathbf{s}} = \mathbf{M}\tilde{\mathbf{u}}$. The difference $\mathbf{s} - \tilde{\mathbf{s}}$ between this effective source and the true source can then be treated as an actual source for the next iteration. The final solution $\mathbf{u}$ to the original matrix equation is found if the difference is sufficiently small.

Figure 1 shows the snapshot of wave propagation through an anisotropic medium. The elastic coefficients are summarized as $c_{11} = 29.26$, $c_{33} = 23.98$, $c_{55} = 4.29$, $c_{13} = c_{31} = 17.27$, $c_{15} = c_{51} = 1.98$, and $c_{35} = c_{53} = 2.53$ (Gpa), and the density of the medium is a constant of 2.5 g/cm^3 . The source is a Ricker wavelet with dominant frequency of 10 Hz. The snapshot shows clearly qP -wave front propagating in difference directions with different velocities. S-wave front can also be observed in these figures.

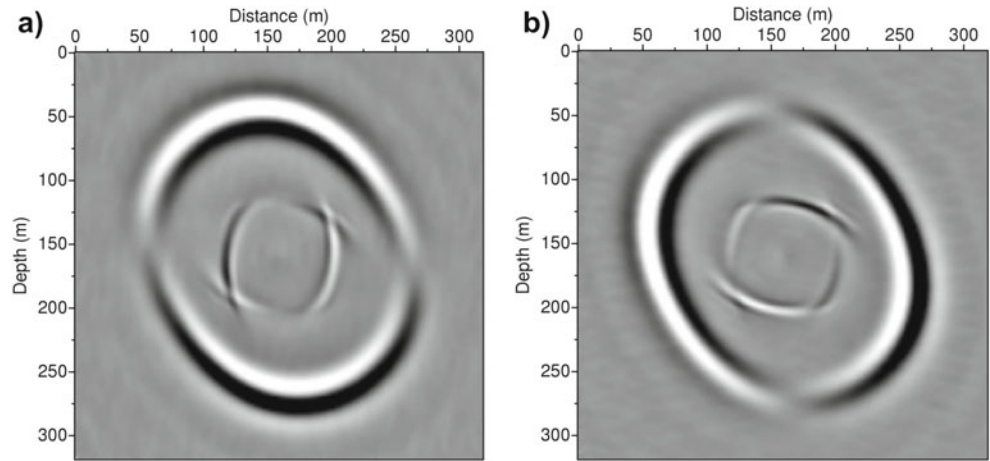
Waveform modeling uses an elastic wave equation, for the generality, and produces a wavefield presented as particle velocity components. In a homogeneous, isotropic media, $c_{11} = c_{33} = \lambda + 2\mu$, $c_{13} = c_{31} = \lambda$, and $c_{55} = \mu$, where λ and μ are Lamé parameters, and $c_{15} = c_{51} = c_{35} = c_{53} = 0$. In an acoustic assumption, $c_{11} = c_{33} = c_{13} = \lambda$ and $c_{55} = c_{15} = c_{35} = 0$. With the acoustic assumption, most field seismic records are pressure data. According to Newton's second law, the relationship between particle velocity (u, w) and pressure P is

$$\rho \frac{\partial u}{\partial t} = -\frac{\partial P}{\partial x}, \quad \rho \frac{\partial w}{\partial t} = -\frac{\partial P}{\partial z}. \quad (6)$$

Structural complexities such as the fluctuations of subsurface interfaces significantly influence seismic wavefields. For instance, fluctuating interfaces have both focusing and defocusing effects on seismic amplitudes. These interface-

Seismic, Waveform Modeling and Tomography,

Fig. 1 Snapshot of wave propagation in an anisotropic medium: (a) horizontal component and (b) vertical component



related amplitude effects can be exploited to reconstruct the interface geometry by inverting the seismic amplitudes (Wang and Houseman 1994, 1995; Wang and Pratt 2000; Wang et al. 2000) and are also included implicitly in the waveform inversion of seismic reflection data. Seismic reflection data, which are dominated by precritical reflected energy from subsurface contrasts in physical parameters, are suitable for structural imaging following seismic migration. However, these data can also be utilized in waveform inversions to quantitatively reconstruct velocity models for the subsurface of the Earth.

To conduct an accurate waveform modeling for seismic waveform tomography, it is necessary to have an accurate representation of the subsurface interfaces within the numerical model, as the contrast in elastic properties across an interface has a dominant impact on the seismic wavefield. A straightforward and cost-effective waveform modeling method is the finite-difference method with quadrilateral grids. A body-fitted gridding scheme can be adopted to partition each horizontal layer that is confined by two fluctuating interfaces in the model. The grids should be coincided with fluctuations of the irregular surface and interfaces while satisfying the pseudo-orthogonality condition. This pseudo-orthogonality condition is necessary to effectively avoid any scattering caused by artificial stair-type grids used within conventional rectangular gridding (Rao and Wang 2013; Zhang et al. 2014).

However, for a fluctuating interface, it is difficult to simultaneously adjust grids on both sides of the interface and to ensure that grids across the interface satisfy the pseudo-orthogonality condition. Then, a summation-by-part (SBP) discretization approach can be effective for the finite-difference implementation of the wave equation (Strand 1994; Mattsson and Nordström 2004; Svärd et al. 2005; Mattsson 2012). The SBP discretization satisfies the SBP

identity in Abel's lemma that guarantees the stability of the finite-difference scheme. This SBP finite-difference method can cope with non-rectangular meshes in the vicinity of a fluctuating interface and consequently achieve a desirable accuracy for the waveform modeling (Rao and Wang 2018).

Inverse Method

Seismic waveform tomography is an inverse problem (Wang 2016). Its objective can be defined as

$$\phi(\mathbf{m}) = [\mathbf{P}(\mathbf{m}) - \mathbf{P}_{\text{obs}}]^H \mathbf{C}_D^{-1} [\mathbf{P}(\mathbf{m}) - \mathbf{P}_{\text{obs}}] + \mu [\mathbf{m} - \mathbf{m}_0]^H \mathbf{C}_M^{-1} [\mathbf{m} - \mathbf{m}_0], \quad (7)$$

where $\mathbf{m}_0$ is a reference model; $\mathbf{m}$ is the model to invert for; $\mathbf{P}(\mathbf{m})$ is a modeled data set based on model $\mathbf{m}$; $\mathbf{P}_{\text{obs}}$ is an observed data set; $\mathbf{C}_D$ is the data covariance matrix with units of $(data)^2$, defining the uncertainties in the data set; $\mathbf{C}_M$ is the model covariance matrix with units of $(model\ parameter)^2$; and μ is a scalar that controls the relative weights of the data fitting term and the model constraint in the objective function. In Equation (7), the superscript H denotes the complex conjugate transpose.

For minimizing the objective function, a gradient method can be used. It starts with the differentiation of the objective function with respect to the model parameters:

$$\frac{\partial \phi}{\partial \mathbf{m}} = 2(\mathbf{F}^H \mathbf{C}_D^{-1} \delta \mathbf{P} + \mu \mathbf{C}_M^{-1} \delta \mathbf{m}), \quad (8)$$

where $\delta \mathbf{m} = \mathbf{m} - \mathbf{m}_0$ is the model perturbation, $\delta \mathbf{P} = \mathbf{P}(\mathbf{m}) - \mathbf{P}_{\text{obs}}$ is the data residual, and $\mathbf{F}$ is a matrix of the Fréchet derivative of $\mathbf{P}(\mathbf{m})$ with respect to the model $\mathbf{m}$. The first term in Equation (8) is the gradient direction of the data misfit:

$$\hat{\mathbf{g}} = \mathbf{F}^H \mathbf{C}_D^{-1} \delta \mathbf{P} = \mathbf{F}^H \delta \hat{\mathbf{P}}, \quad (9)$$

where $\delta \hat{\mathbf{P}} = \mathbf{C}_D^{-1} \delta \mathbf{P}$ is a weighted data residual. Setting $\partial \phi / \partial \mathbf{m} = 0$ in Equation (8), one obtains the following equation:

$$\delta \mathbf{m} = -\alpha \mathbf{C}_M \hat{\mathbf{g}}, \quad (10)$$

where α is a update step length that needs to be determined.

In order to evaluate the gradient $\hat{\mathbf{g}}$ using Equation (9), one needs to know the Fréchet matrix $\mathbf{F}$, which is obtained from the following linear formula:

$$\delta \mathbf{P} = \mathbf{F} \delta \mathbf{m}. \quad (11)$$

This is the first term in a Taylor's series for $\delta \mathbf{P}$ and relates the data perturbation $\delta \mathbf{P}$ to the model perturbation $\delta \mathbf{m}$. However, a direct computation of $[\mathbf{F}]_{ij} = \partial P_i / \partial m_j$ is a formidable task when P_i are seismic waveforms. The action of matrix $\mathbf{F}^H$ on the weighted data residual vector $\delta \hat{\mathbf{P}}$ can be computed by a series of forward modeling steps, summarized as follows (Tarantola 1984, 1987).

The frequency-domain acoustic wave equation for a constant density medium with velocity $c_0(\mathbf{x})$ is

$$\left(\nabla^2 + \frac{\omega^2}{c_0^2(\mathbf{x})} \right) P_0(\mathbf{x}, \omega) = -S(\omega) \delta(\mathbf{x} - \mathbf{x}_0), \quad (12)$$

where $\mathbf{x}$ is the position vector, $\mathbf{x}_0$ locates the source position, $S(\omega)$ is the source signature of frequency ω , and $P_0(\mathbf{x}, \omega)$ is the (pressure) wavefield of this frequency, and we drop off the frequency reference in the following discussion. If the velocity is perturbed by a small amount, $\delta c(\mathbf{x}) \ll c_0(\mathbf{x})$, that is, $c_0(\mathbf{x}) \rightarrow c(\mathbf{x}) = c_0(\mathbf{x}) + \delta c(\mathbf{x})$, then the total wavefield is correspondingly perturbed to $P_0(\mathbf{x}) \rightarrow P(\mathbf{x}) = P_0(\mathbf{x}) + \delta P(\mathbf{x})$. Following Equation (12), δP approximately satisfies

$$\left(\nabla^2 + \frac{\omega^2}{c_0^2(\mathbf{x})} \right) \delta P(\mathbf{x}) = 2\omega^2 P_0(\mathbf{x}) \frac{\delta c(\mathbf{x})}{c_0^3(\mathbf{x})}. \quad (13)$$

Considering $2\omega^2 P_0(\mathbf{x}) \delta c(\mathbf{x}) / c_0^3(\mathbf{x})$, the term on the right-hand side, as a series of "virtual sources" over $\mathbf{x}$ the integral solution for $\delta P(\mathbf{x})$ can be expressed as

$$\delta P(\mathbf{x}) = - \int_{\Omega} \delta c(\mathbf{x}') \frac{2\omega^2}{c_0^3(\mathbf{x}')} P_0(\mathbf{x}') G(\mathbf{x}, \mathbf{x}') d\mathbf{x}', \quad (14)$$

where $G(\mathbf{x}, \mathbf{x}')$ is the Green's function for the response at $\mathbf{x}$ to a point source at $\mathbf{x}'$ for the original velocity field. Note that in the acoustic case where one assumes density to be constant and defines the model by the velocity field only, $\mathbf{m} \equiv \mathbf{c}$. Then,

comparing Equation (14) against the matrix-vector form of Equation (11), one can see that the Fréchet matrix is defined with element, $F(\mathbf{x}, \mathbf{x}') = -[2\omega^2 / c_0^3(\mathbf{x}')] P_0(\mathbf{x}') G(\mathbf{x}, \mathbf{x}')$. Substituting this Fréchet kernel into Equation (9), one obtains

$$\hat{\mathbf{g}}(\mathbf{x}) = \left(\frac{2\omega^2}{c_0^3(\mathbf{x})} \right)^* \int_D P_0^*(\mathbf{x}') G^*(\mathbf{x}, \mathbf{x}') \delta \hat{P}(\mathbf{x}'). \quad (15)$$

Replacing the integral over the data space with a summation over source and receiver pairs, denoted by s and g , respectively, as the source and receiver position are inherently discrete and finite in number, one can obtain

$$\begin{aligned} \hat{\mathbf{g}}(\mathbf{x}) &= \left(\frac{2\omega^2}{c_0^3(\mathbf{x})} \right)^* \sum_s \left(P_0^*(\mathbf{x}; \mathbf{x}_s) \sum_r G^*(\mathbf{x}, \mathbf{x}_r) \delta \hat{P}(\mathbf{x}_r; \mathbf{x}_s) \right) \\ &= \left(\frac{2\omega^2}{c_0^3(\mathbf{x})} \right)^* \sum_s (P_0^*(\mathbf{x}; \mathbf{x}_s) P_b^*(\mathbf{x}; \mathbf{x}_s)), \end{aligned} \quad (16)$$

where

$$P_b(\mathbf{x}; \mathbf{x}_s) = \sum_r G(\mathbf{x}, \mathbf{x}_r) \delta \hat{P}^*(\mathbf{x}_r; \mathbf{x}_s) \quad (17)$$

representing the wavefield generated by a series of virtual sources $\delta \hat{P}^*(\mathbf{x}_r)$, corresponding to a single source $\mathbf{x}_s$. Note that wavefield $P_b(\mathbf{x}; \mathbf{x}_s)$ is not calculated directly from Equation (17) but is computed using the same forward modeling scheme as used for Equation (12) with the virtual sources $\delta \hat{P}^*(\mathbf{x}_r)$. This procedure is referred to as data residual backpropagation.

In summary, waveform tomography is performed iteratively. For each iteration, the inversion procedure may be divided into four steps:

1. For a given model estimate, calculating the synthetic wavefield $P_0(\mathbf{x}; \mathbf{x}_s)$ at space position $\mathbf{x}$ corresponding to a source point at $\mathbf{x}_s$
2. Using the weighted data residual $\delta \hat{\mathbf{P}} = \mathbf{C}_D^{-1} \delta \mathbf{P}$ as virtual sources to generate a so-called back propagation wavefield $P_b(\mathbf{x}; \mathbf{x}_s)$
3. Cross-correlation of the original wavefield $P_0(\mathbf{x}; \mathbf{x}_s)$ and the backpropagation wavefield $P_b(\mathbf{x}; \mathbf{x}_s)$ to get the gradient direction $\mathbf{g} = \mathbf{C}_M \hat{\mathbf{g}}$, where $\mathbf{C}_M$ is the model covariance matrix with units of $(\text{model parameter})^2$
4. Estimating the model update $\delta \mathbf{m} = -\alpha \mathbf{g}$, where α is the optimal step length that can be found by using the linear approximation or simply line search for a minimum of the objective function

In a time-domain implementation of waveform inversion, the gradient direction is calculated by

$$\hat{\mathbf{g}}(\mathbf{x}) = \frac{2}{c_0^3(\mathbf{x})} \sum_s \int \left(\frac{\partial}{\partial t} p_0(\mathbf{x}; \mathbf{x}_s) \frac{\partial}{\partial t} p_b(\mathbf{x}; \mathbf{x}_s) \right) dt, \quad (18)$$

where $\partial p_0(\mathbf{x}; \mathbf{x}_s)/\partial t$ denotes the time derivative of the time-domain forward-propagated wavefields and $\partial p_b(\mathbf{x}; \mathbf{x}_s)/\partial t$ is the time derivative of the backprojected residual waveforms in the time domain. Zero-leg correlation of these two wavefields provides us the gradient direction for model updating.

Shot-Encoded Waveform Tomography

Seismic waveform tomography often demonstrates high resolution and high accuracy. However, the computational cost obstructs the wide application of this inversion technology, especially for the inversion of seismic reflection data, which is routinely acquired during exploration geophysics. As described in the previous section, the subsurface model is updated iteratively following the negative direction of the gradient. This gradient vector is the first-order derivative of the objective function with respect to the unknown model parameters and is computed for each shot separately. The gradient vector is formed by cross-correlation between the incident wavefield and the adjoint wavefield, and these two wavefields are simulated for each seismic shot individually. Thus, this process requires a very high operational speed, particularly for practical 3D seismic data with thousands of shots.

In seismology, a shot gather refers to a group of seismic traces recorded by a series of receivers for which the recorded wavefield has originated from a single source. To reduce the number of wavefield simulations in the inversion process, the shot gathers are multiplied by either +1 or -1 randomly and are then stacked into a super-shot gather. Equivalently, the source signatures are multiplied by random codes and are subsequently excited simultaneously to simulate forward wave propagation. For the calculation of the gradient, the residual of the waveform between the shot-encoded synthetic wavefield and the shot-encoded observed wavefield is backward propagated and cross-correlated with the forward wavefield. Note that the cross-correlation will exhibit a cross talk effect between different shots, which can directly deteriorate the quality of tomographic imagery.

For typical shot-encoded waveform tomography, seismic shots are assigned into different super shots. The shots contained within a super shot should be randomly encoded rather than simply superimposed so that the cross talk effect can be offset by the randomness of encoding. The objective

function for the shot-encoded waveform tomography is defined in terms of the weighted wavefields according to the weights in the encoding procedure. Then, the gradient of the objective function consists of both the ordinary gradient (without shot encoding) and an extra term representing the cross talk between different shots. Fortunately, this cross talk effect can be effectively suppressed by increasing the randomness of the shot encoding (Rao and Wang 2017). The quality improvement occurs because the weights w_i used in the encoding procedure are randomly generated values of +1 and -1 and the sum of a large number of $w_i w_j$ (i.e., the multiplication between different encoded shots) in the cross talk term has an expectation of zero.

The inversion algorithm used in the shot-encoded waveform tomography must be stochastic. Because the model update procedure is stochastic, both gradient-based and Newton-type methods used in conventional deterministic inversions must be properly adjusted to fit within a shot-encoded framework. The restarted L-BFGS algorithm described in the next section exploits the features of shot encoding and of the L-BFGS algorithm and uses these features to enhance the computational efficiency and the image quality of waveform tomography by mitigating the cross talk effect.

The Restarted L-BFGS Algorithm

Shot encoding in waveform tomography will induce instability in the inversion regardless of whether it uses a robust limited-memory BFGS (L-BFGS) algorithm. Rao and Wang (2017) proposed a restarted L-BFGS algorithm which is both stable and efficient. This breakthrough ensures the applicability of waveform tomography methods to 3D seismic field data. In a standard L-BFGS algorithm, if the shot encoding remains unchanged, it will generate a cross talk effect between different shots. This cross talk effect can only be suppressed by employing sufficient randomness in the shot encoding. Therefore, the implementation of the L-BFGS algorithm is restarted at every segment. Each segment consists of a number of iterations; the first few iterations use an invariant encoding, while the remainder use random re-coding. This restarted L-BFGS algorithm balances the computational efficiency of shot encoding, the convergence stability of the L-BFGS algorithm, and the inversion quality characteristic of random encoding in waveform tomography.

In waveform tomography, the subsurface model is updated iteratively, and the model update is defined by the gradient according to a precondition of the inverse Hessian matrix in the form of

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \alpha_k \mathbf{B}_k \hat{\mathbf{g}}_k, \quad (19)$$

where k is the iteration number, α_k is the step length, and $\mathbf{B}_k$ is the inverse Hessian matrix. Since it exploits the inverse Hessian matrix, in contrast to Equation (10), this is a quasi-Newton method.

In a BFGS algorithm, the inverse Hessian matrix $\mathbf{B}_k$ is derived recursively from $\mathbf{B}_{k-1}$ using the following relation:

$$\mathbf{B}_k = \left(\mathbf{I} - \frac{\mathbf{z}_{k-1} \mathbf{y}_{k-1}^T}{\mathbf{y}_{k-1}^T \mathbf{z}_{k-1}} \right) \mathbf{B}_{k-1} \left(\mathbf{I} - \frac{\mathbf{y}_{k-1} \mathbf{z}_{k-1}^T}{\mathbf{y}_{k-1}^T \mathbf{z}_{k-1}} \right) + \frac{\mathbf{z}_{k-1} \mathbf{z}_{k-1}^T}{\mathbf{y}_{k-1}^T \mathbf{z}_{k-1}}, \quad (20)$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{y}_{k-1}$ is the difference in the gradient between two consecutive iterations, and $\mathbf{z}_{k-1}$ is the model difference. Because of its recursive nature, $\mathbf{B}_{k-1}$ does not need to be stored, and $\mathbf{B}_k$ can be calculated quickly and adaptively based on the gradient differences $\{\mathbf{y}_{k-1}, \mathbf{y}_{k-2}, \mathbf{y}_{k-3}, \dots\}$ and the model differences $\{\mathbf{z}_{k-1}, \mathbf{z}_{k-2}, \mathbf{z}_{k-3}, \dots\}$ from the previous iterations. Therefore, Equation (20) represents the limited-memory version of the BFGS algorithm (i.e., L-BFGS).

If $\mathbf{y}_{k-1}^T \mathbf{z}_{k-1} > 0$, the inverse Hessian matrix $\mathbf{B}_k$ is positive definite. This is because in the following formula

$$\mathbf{v}^T \mathbf{B}_k \mathbf{v} = \left(\mathbf{v} - \frac{\mathbf{z}_{k-1}^T \mathbf{v}}{\mathbf{z}_{k-1}^T \mathbf{y}_{k-1}} \mathbf{y}_{k-1} \right)^T \mathbf{B}_{k-1} \left(\mathbf{v} - \frac{\mathbf{z}_{k-1}^T \mathbf{v}}{\mathbf{z}_{k-1}^T \mathbf{y}_{k-1}} \mathbf{y}_{k-1} \right) + \frac{(\mathbf{z}_{k-1}^T \mathbf{v})^2}{\mathbf{y}_{k-1}^T \mathbf{z}_{k-1}}, \quad (21)$$

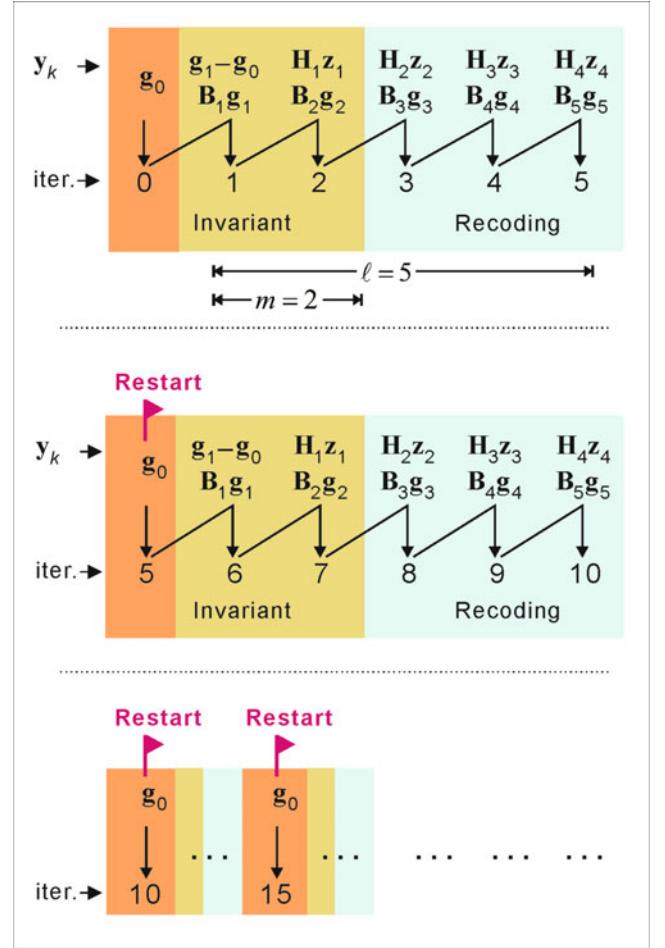
if $\mathbf{B}_{k-1} > 0$, both terms on the right-hand side are nonnegative, while the first term is zero only if $\mathbf{v} = 0$ and the second term is zero only if $\mathbf{z}_{k-1}^T \mathbf{v} = 0$. This positive definite property ensures that the update $-\mathbf{B}_k \hat{\mathbf{g}}_k$ is in a descent direction and consequently the model update during the inversion will converge. If shots are re-coded at each iteration, two gradients $\hat{\mathbf{g}}_k(\mathbf{m}_k, \mathbf{w}_k)$ and $\hat{\mathbf{g}}_{k-1}(\mathbf{m}_{k-1}, \mathbf{w}_{k-1})$ are calculated from the objective functions with different sets of encoding, which are $\mathbf{w}_{k-1}$ and $\mathbf{w}_k$. For the gradient difference

$$\mathbf{y}_{k-1} = \hat{\mathbf{g}}_k - \hat{\mathbf{g}}_{k-1}, \quad (22)$$

even if the step length α_{k-1} is sufficiently small, the condition $\mathbf{y}_{k-1}^T \mathbf{z}_{k-1} > 0$ is difficult to satisfy. Therefore, shot encoding will induce instability and potentially non-convergence during waveform tomography.

The restarted L-BFGS algorithm (Fig. 2) contain the following three distinguishing features:

1. Structurally, the L-BFGS calculation is restarted after every ℓ iterations. The total number of L-BFGS iterations



Seismic, Waveform Modeling and Tomography, Fig. 2 The restarted L-BFGS algorithm. The implementation of the L-BFGS algorithm is restarted after a number of iterations (ℓ). Within the segment between two starts, only the coding of the first few iterations (m) is kept unchanged (yellow) and is the same as the previous one (orange). The shots for the rest of the iterations (gray) are re-coded with a sufficient randomness to suppress cross talk effects. Within a restarted segment, the gradient differential is given by the secant equation $\mathbf{y}_{k-1} = \mathbf{H}_{k-1} \mathbf{z}_{k-1}$ except for the first iteration where $\mathbf{y}_0 = \mathbf{g}_1 - \mathbf{g}_0$. Both $\mathbf{H}_{k-1} \mathbf{z}_{k-1}$ and $\mathbf{B}_k \mathbf{g}_k$ are calculated using the limited-memory formulas

is divided into a series of segments, and each segment consists of ℓ iterations. The recursive L-BFGS calculation is only based on the gradient differences $\mathbf{y}_k$ and model differences $\mathbf{z}_k$ within the individual segment. Therefore, the segment interval ℓ is also referred to as the storage length in the L-BFGS algorithm (i.e., for storing the vectors $\mathbf{y}_k$ and $\mathbf{z}_k$).

2. Within a segment, shot encoding of the first few iterations is kept invariant (yellow) and is the same as the previous iteration (orange). A new reassignment of the weights (either +1 or -1) is triggered for the remainder of the iterations (gray). We denote m as the number of iterations that employ invariant coding. Generally, $2 \leq m < \ell$, which

maximizes the overall randomness of shot encoding to effectively suppress the cross talk effect. Meanwhile, maintaining the weights of coding as invariant for a few of iterations, rather than re-coding at every iteration, will significantly improve the convergence speed and stability during the inversion.

3. For the first iteration, the gradient difference is calculated by $\mathbf{y}_0 = \mathbf{g}_1 - \mathbf{g}_0$. For the rest of the iterations, the gradient difference is calculated based on the following secant equation:

$$\mathbf{y}_{k-1} = \mathbf{H}_{k-1} \mathbf{z}_{k-1}, \quad (23)$$

where $\mathbf{H}_{k-1} = \nabla^2 \phi(\mathbf{m}_{k-1})$ is the Hessian matrix in the $k-1$ -th iteration. Equation (23) represents the gradient differential rather than the gradient difference. Using the gradient differential will improve the accuracy and, in turn, the stability of convergence of waveform tomography.

Note that the secant equation in Equation (23) differs slightly from the conventional secant equation used in the BFGS algorithm. The latter imposes a condition wherein $\mathbf{H}_k$ is close to the current matrix $\mathbf{H}_{k-1}$ and presents the following approximation as the secant equation:

$$\mathbf{y}_{k-1} = \mathbf{H}_k \mathbf{z}_{k-1}. \quad (24)$$

This approximation can be justified because

$$\mathbf{y}_{k-1} \equiv \mathbf{g}_k - \mathbf{g}_{k-1} = \frac{\mathbf{g}_k - \mathbf{g}_{k-1}}{\mathbf{m}_k - \mathbf{m}_{k-1}} (\mathbf{m}_k - \mathbf{m}_{k-1}), \quad (25)$$

wherein the fraction can be treated as either a forward difference or a backward difference, i.e.

$$\frac{\mathbf{g}_k - \mathbf{g}_{k-1}}{\mathbf{m}_k - \mathbf{m}_{k-1}} \approx \mathbf{H}_k \approx \mathbf{H}_{k-1}. \quad (26)$$

However, the approximated secant equation in Equation (24) cannot be used to directly calculate $\mathbf{y}_{k-1}$, as it depends on $\mathbf{H}_k$ and not $\mathbf{H}_{k-1}$. Rao and Wang (2017) suggested use Equation (23) to calculate $\mathbf{y}_{k-1}$.

To calculate the Hessian matrix, we use the DFP (Davidon-Fletcher-Powell) method, which represents the Hessian matrix based on the gradient difference and the model difference as follows:

$$\mathbf{H}_k = \left(\mathbf{I} - \frac{\mathbf{y}_{k-1} \mathbf{z}_{k-1}^T}{\mathbf{y}_{k-1}^T \mathbf{z}_{k-1}} \right) \mathbf{H}_{k-1} \left(\mathbf{I} - \frac{\mathbf{z}_{k-1} \mathbf{y}_{k-1}^T}{\mathbf{y}_{k-1}^T \mathbf{z}_{k-1}} \right) + \frac{\mathbf{y}_{k-1} \mathbf{y}_{k-1}^T}{\mathbf{y}_{k-1}^T \mathbf{z}_{k-1}}. \quad (27)$$

Exploiting its recursive characteristic, we can derive the following formula:

$$\begin{aligned} \mathbf{H}_k &= \mathbf{V}_{k-1}^T \mathbf{V}_{k-2}^T \cdots \mathbf{V}_1^T \mathbf{H}_0 \mathbf{V}_1 \cdots \mathbf{V}_{k-2} \mathbf{V}_{k-1} \\ &+ \rho_1 \mathbf{V}_{k-1}^T \mathbf{V}_{k-2}^T \cdots \mathbf{V}_2^T \mathbf{y}_1 \mathbf{y}_1^T \mathbf{V}_2 \cdots \mathbf{V}_{k-2} \mathbf{V}_{k-1} \\ &\vdots \\ &+ \rho_{k-2} \mathbf{V}_{k-1}^T \mathbf{y}_{k-2} \mathbf{y}_{k-2}^T \mathbf{V}_{k-1} \\ &+ \rho_{k-1} \mathbf{y}_{k-1} \mathbf{y}_{k-1}^T, \end{aligned} \quad (28)$$

where $\rho_k = 1/\mathbf{y}_k^T \mathbf{z}_k$ and $\mathbf{V}_k = (\mathbf{I} - \rho_k \mathbf{z}_k \mathbf{y}_k^T)$. According to the secant equation, where $\mathbf{H}_k \mathbf{z}_k = \mathbf{y}_k$, we define the initial Hessian matrix $\mathbf{H}_0$ as

$$\mathbf{H}_0 = \mathbf{y}_0 \mathbf{y}_0^T [\mathbf{z}_0 \mathbf{y}_0^T]^{-1} \approx \frac{\mathbf{y}_0 \mathbf{y}_0^T}{\mathbf{z}_0^T \mathbf{y}_0} \mathbf{I}. \quad (29)$$

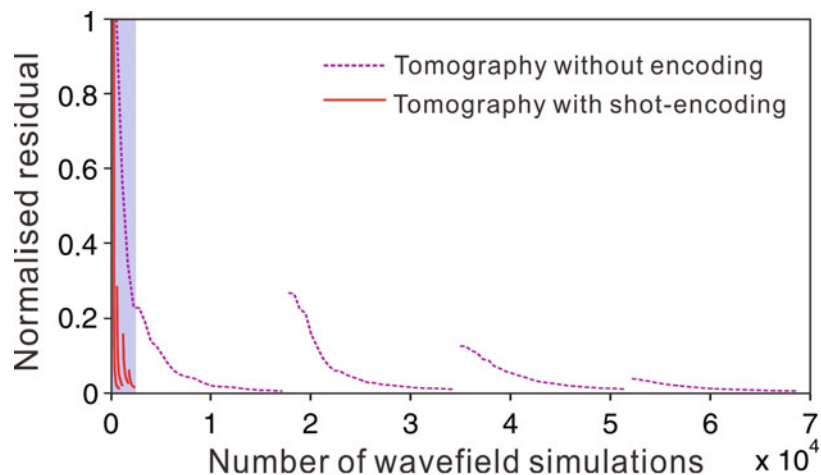
Equation (28) does not need to store matrices and is a limited-memory version of the DFP formula. This equation is similar to the L-BFGS formula, which uses the same recursive form as Equation (28) for the inverse Hessian matrices $\mathbf{B}_k$, for $k = 0, 1, \dots$, and does not need to store these matrices as well. Therefore, in shot-encoded waveform tomography, we use the L-BFGS and DFP algorithms and calculate $\mathbf{B}_k \mathbf{g}_k$ and $\mathbf{H}_k \mathbf{z}_k$ based only on the vectors $\mathbf{y}_k$ and $\mathbf{z}_k$, which are stored within the segment of ℓ iterations.

The computational efficiency of waveform tomography is determined by the actual number of wavefield simulations required. These wavefield simulations include wave back-propagation and step-length calculations for every iteration. Figure 3 shows the number of wavefield simulations required to achieve convergence in a synthetic test, in which a multi-scale inversion method is adopted and seismic data have been divided into four frequency bands: 2–4, 4–6, 6–8, and 8–10 Hz. Therefore, the inversion procedure consists of four stages. The total number of simulations in the restarted L-BFGS algorithm ($<0.25 \times 10^4$, shaded region in Fig. 3) is substantially less than that in the standard L-BFGS algorithm (7×10^4). Therefore, the restarted L-BFGS algorithm fully reflects an accelerated shot-encoding inversion procedure. Compared with waveform tomography without shot encoding, the restarted L-BFGS algorithm significantly reduces computational costs.

As the restarted L-BFGS algorithm can improve image quality by suppressing the cross talk effect and can simultaneously reduce computational costs and storage requirements, shot-encoded waveform tomography is shown to be applicable to cases involving 3D data sets, as demonstrated in Fig. 4. The size of the actual 3D overthrust SEG/EAGE model is $4.65 \times 20 \times 20 \text{ km}^3$. The model is partitioned into $93 \times 401 \times 401$ grids with a cell size of $50 \times 50 \times 50 \text{ m}^3$. The source signature is a Ricker wavelet with a domain frequency of 7 Hz. There are 400 shots spread out over the surface at an interval of $1 \times 1 \text{ km}^2$. Each shot gather consists of 32,761 traces. In the inversion, every 100 shots are assembled into a super shot.

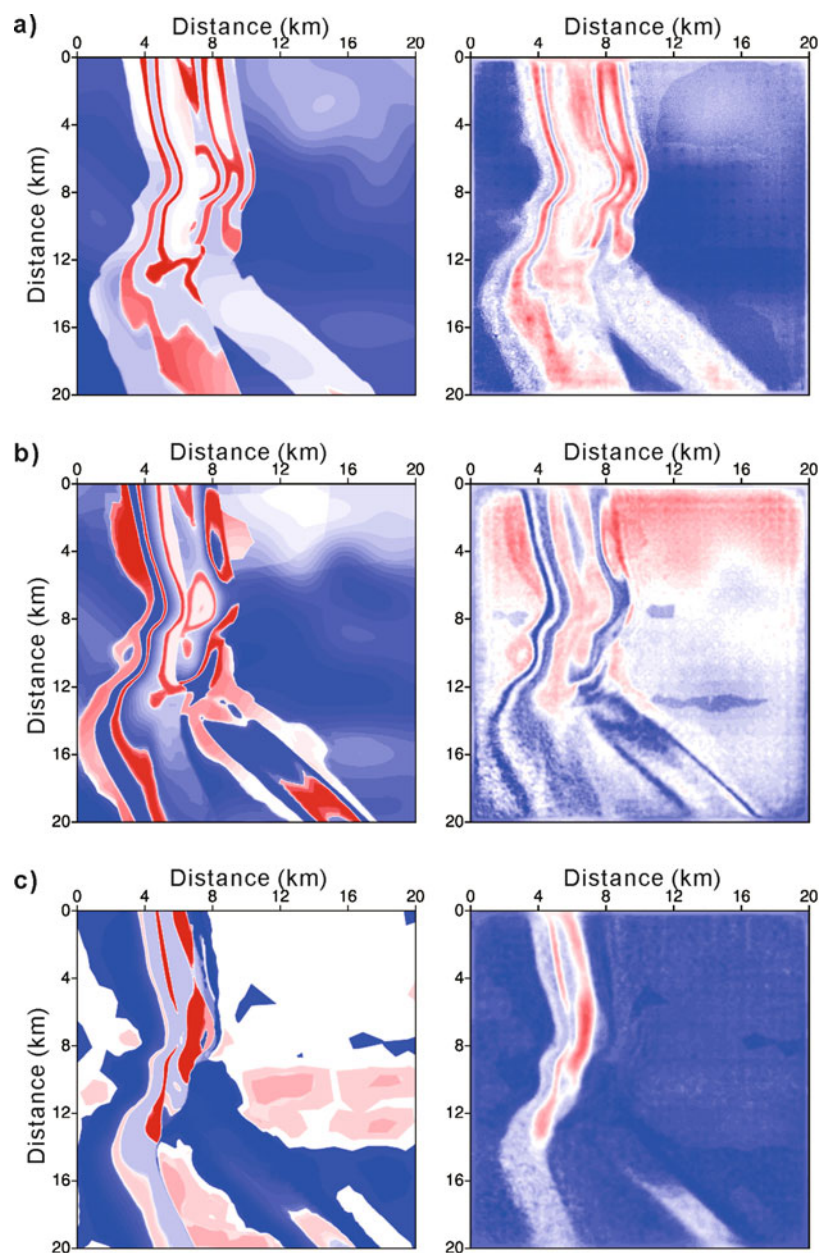
Seismic, Waveform Modeling and Tomography,

Fig. 3 Convergence rates of waveform tomography without shot encoding (purple dashed lines) and of shot-encoded waveform tomography using a restarted L-BFGS algorithm (red solid lines within the shadowed zone). The horizontal axis denotes the number of wavefield simulations, and the vertical axis is the normalized residual. Each line segment corresponds to an inversion of a data set for a particular frequency band



Seismic, Waveform Modeling and Tomography, Fig. 4

(a) Velocity slices at a depth of 1 km. (b) Velocity slices at a depth of 1.5 km. (c) Velocity slices at a depth of 2 km. The left column is the true velocity model, and the right column is the velocity model reconstructed using shot-encoded waveform tomography



The initial model is a Gaussian smoothed version of the true velocity model. For the inversion of each of four frequency bands, 100 iterations are executed, and the restarted L-BFGS algorithm is used with $\ell = 5$ and $m = 2$. Figure 4 displays the resulting velocity slices at different depths to compare the true velocity model and the reconstructed velocity model. At depths of 1.0 and 1.5 km, the morphology of an ancient channel is clearly reconstructed. At a depth of 2 km, the main body of the channel can also be inverted, although there is limited wave coverage.

Preprocessing for Field Data Waveform Tomography

Let us demonstrate the application of waveform tomography to a real, marine seismic data set (Fig. 5). This example profile shows the structure along the section. However, the input to waveform tomography is shot records which should be subject to some necessary processing, such as multiple attenuation and transformation of the original point sources to equivalent line sources, before they are used in tomography.

Figure 6a is a sample shot record, consisting of 120 traces with a minimum source-receiver offset of 337.5 m and a maximum source-receiver offset of 1825 m. We investigate the feasibility of reflection seismic tomography within such a narrow source-receiver offset range.

Waveform tomography usually does not include free-surface multiples. Otherwise including seismic multiples bouncing back and forwards within the water layer in the tomographic inversion will increase the nonlinearity of the problem. As the number of multiples increases, the errors in model (and in turn in synthetics) will also increase. We use a narrow-offset shot record in multiple attenuation, also to

avoid the wide-angle refraction of the water bottom and their multiples, as the current methodology for free-surface multiple prediction cannot properly model the refraction multiples. As marked in Fig. 6a, the most difficult part of multiple attenuation is where the refraction wave just starts appearing. Figure 6b displays a shot record and the one after free-surface multiple attenuation, using a multiple prediction through inversion (MPI) method (Wang 2004, 2007). The real shot record is generated by a point source, but Fig. 6c is an equivalent line-source shot gather, after partial compensation.

Before input to waveform tomography, a shot record of real seismic data needs to be partially compensated, to become a gather generated from a line source. The 2D Green's function is

$$G_{2D}(\mathbf{r}, \mathbf{r}_s, \omega) = \left(\frac{c}{i 8 \pi \omega R} \right)^{1/2} \exp \left(-i \omega \frac{R}{c} \right), \quad (30)$$

where $R = |\mathbf{r} - \mathbf{r}_s|$ is the distance from the source and c is the acoustic velocity of the media. For a 3D case, Green's function is

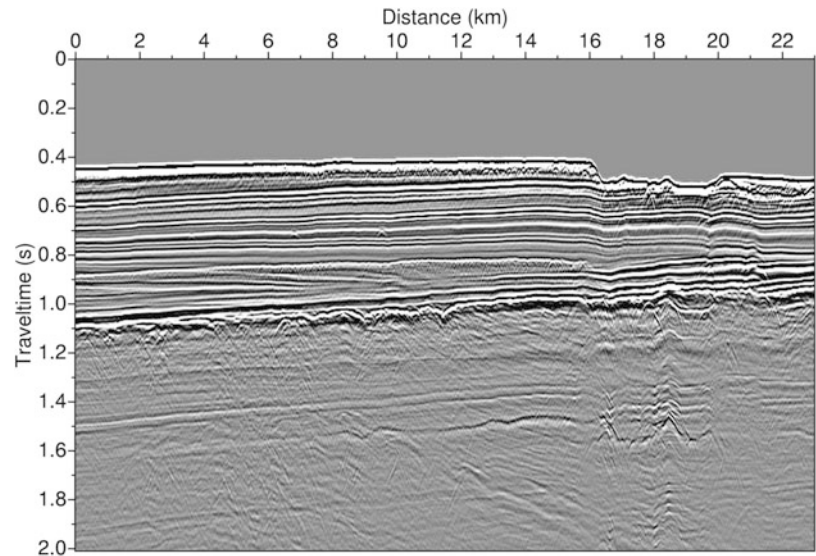
$$G_{3D}(\mathbf{r}, \mathbf{r}_s, \omega) = \frac{1}{4 \pi R} \exp \left(-i \omega \frac{R}{c} \right). \quad (31)$$

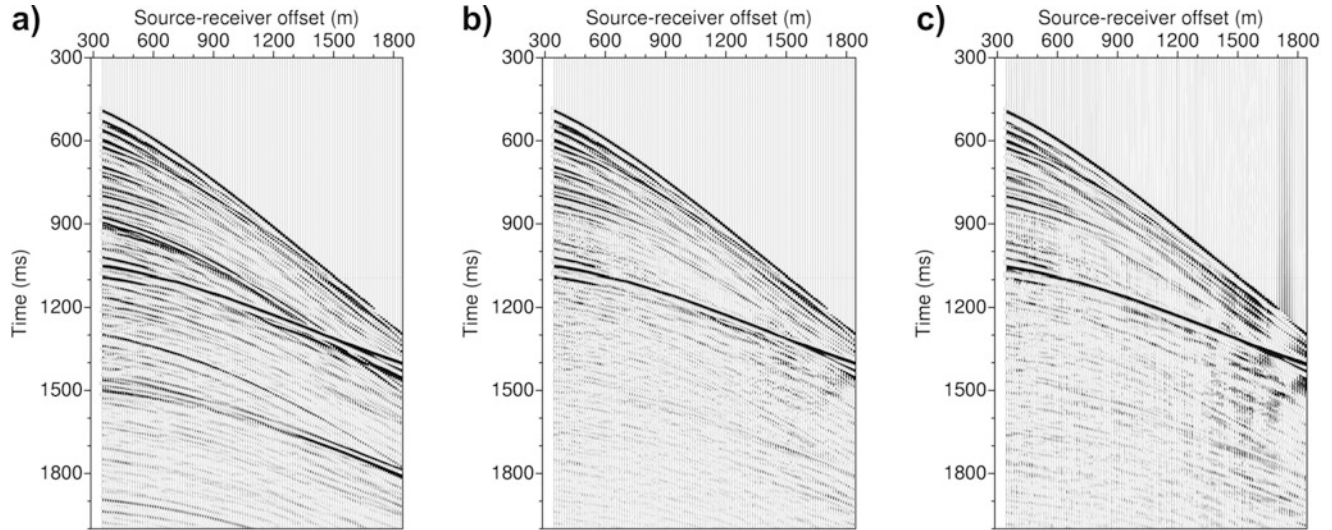
Comparing the 2D and 3D, Green's functions produces a partial compensation operator as

$$W = \sqrt{\frac{2 \pi R c}{i \omega}}. \quad (32)$$

In the time domain, the operator W shows the following behavior (in the far-field)

Seismic, Waveform Modeling and Tomography, Fig. 5 An example profile of marine seismic data which shows the structure along a section. However, the input to waveform tomography is shot records which should be subject to some necessary processing





Seismic, Waveform Modeling and Tomography, Fig. 6 A marine seismic data example. (a) A sample shot record with 120 traces. (b) The shot record after multiple attenuation. (c) The same shot record after partial compensation

$$W(t) = D_{-1/2}(t)\sqrt{2\pi Rc}, \quad (33)$$

where $D_{-1/2}(t)$ is a half integrator, defined as the inverse Fourier transform of $(i\omega)^{-1/2}$ (Deregowski and Brown 1983). For a narrow-offset, reflection geometry, one can assume $2R \propto ct$ and obtain

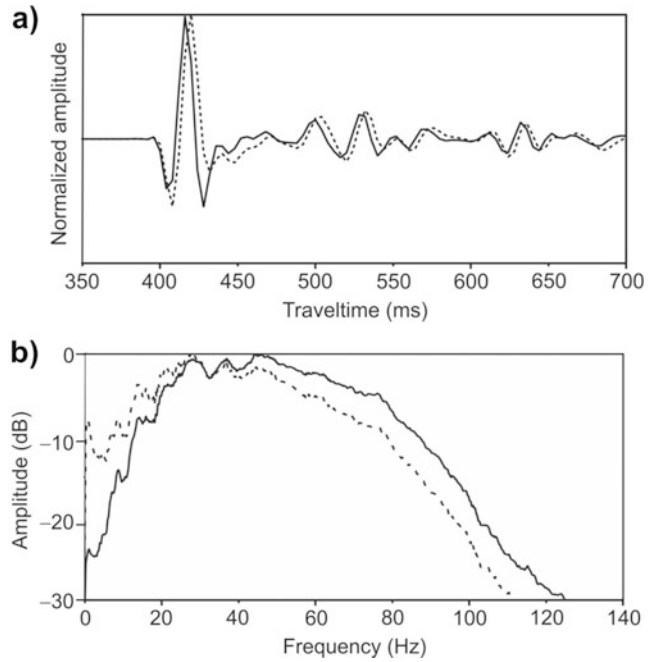
$$W(t) \propto D_{-1/2}(t) c(t)\sqrt{t}. \quad (34)$$

Therefore, this partial compensation may be implemented in two steps: scaling $c(t)\sqrt{t}$ in the time domain and multiplication $(i\omega)^{-1/2}$ in the frequency domain.

Figure 7 closely compares a seismic trace from a point source (solid curve) and a trace after partial compensation (dash curve). With the application of the operator $(i\omega)^{-1/2}$, wavelets in the trace from an equivalent line source are broader than those actually generated from a point source.

Alternatively, one could use a 2.5D wave modeling and inversion scheme for waveform tomography. But as it involves integration along the infinite line-source direction that is perpendicular to the source-receiver profile, a 2.5D scheme would take a much longer running time, compared to the 2D wave modeling and inversion.

The frequency ω can also be complex-valued, to include an exponential function either for the amplitude attenuation or compensation. The attenuation used in wavefield $P(t)$ is for suppressing the wraparound effect in the Fourier transform domain. An opposite sign with an exponential increase in $P(t)$ could also be used to boost the energy of deep reflections. In the latter case, the real data should be balanced with an automatic gain control.



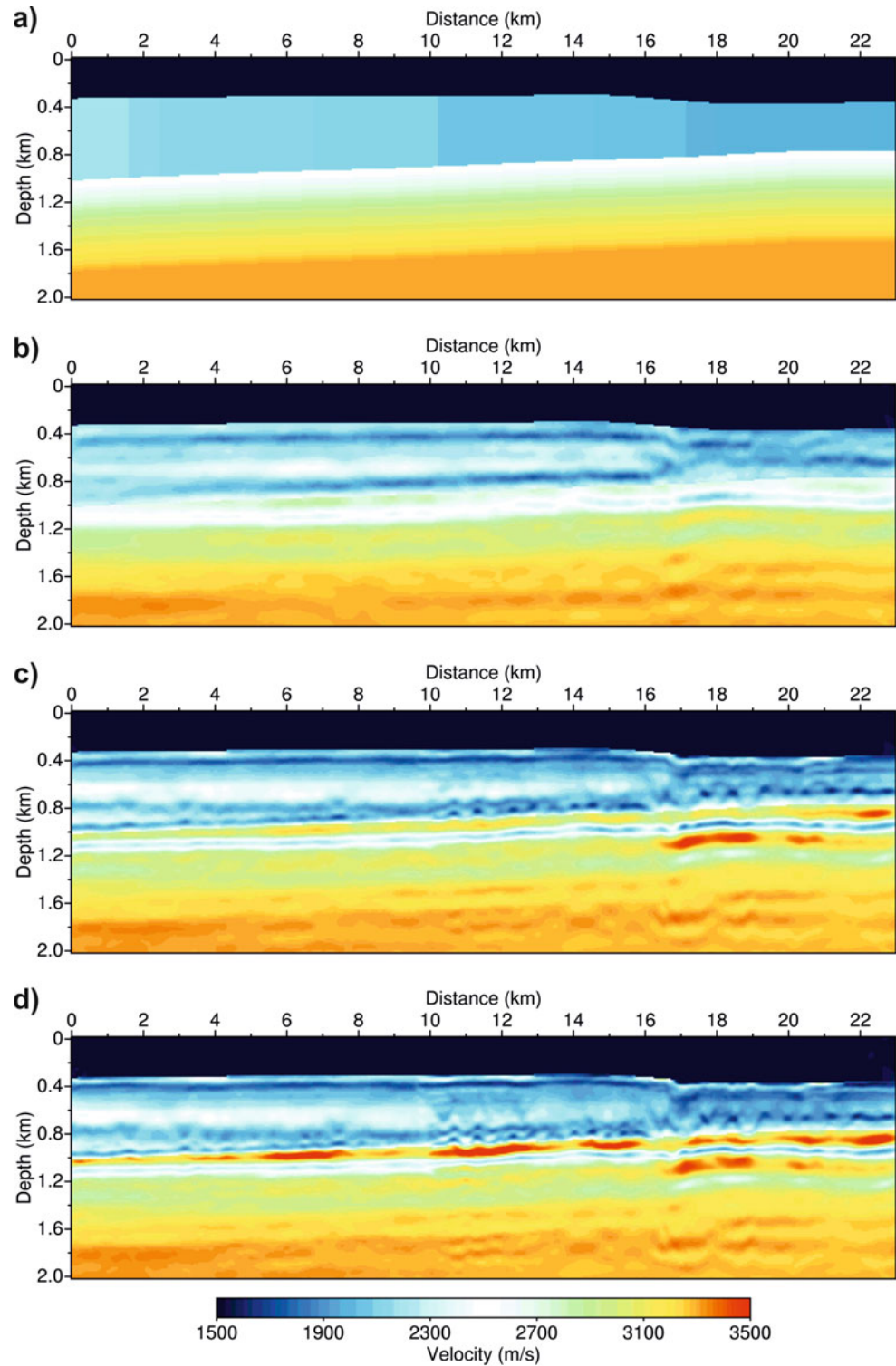
Seismic, Waveform Modeling and Tomography, Fig. 7 (a) Comparison between a seismic trace from a point source (solid) and the trace after partial compensation (dash curve). Wavelets in a trace from a line source (i.e., after partial compensation) are broader than those from a point source. (b) Comparison of amplitude spectra of seismic traces from a point source and a line source

Seismic Reflection Waveform Tomography

In the seismic reflection data, there is no significant energy recorded at low frequencies less than 6 Hz. A travel time inversion is often performed to generate the initial velocity model for the iterative waveform tomography. Referring to

Seismic, Waveform Modeling and Tomography,

Fig. 8 Waveform tomography. (a) The initial velocity model built from travel time tomography. (b) The velocity model of waveform tomography using frequencies in the range of 6.9–7.5 Hz. (c) The velocity model of waveform tomography using frequencies in the range of 6.9–13.8 Hz. (d) The final velocity model obtained from the waveform tomography using all frequencies in the range of 6.9–30 Hz



the stack section (Fig. 5), we pick the travel times of two reflections from the pre-stack seismic data and run a travel time inversion to generate a layered velocity model with two interfaces (Fig. 8a). The first layer is a water layer with

velocity 1500 m/s. The second layer has a linear velocity of 2200 m/s at one end and 2000 m/s at distance 23 km.

In the frequency-domain waveform tomography, we use a group of three frequencies simultaneously in each iterative

inversion. Using a group of neighboring frequencies in the input can suppress the noise effect in the real data, and also more data samples used in an inversion means a much better determined inverse problem. In this example, there are 26 groups of frequencies in total in the range of 6.9–30 Hz with an interval of 0.3 Hz. The first group includes frequencies 6.9, 7.2, and 7.5 Hz and the last 29.4, 29.7, and 30 Hz. Figure 8b, c are the velocity models from waveform tomography using frequencies in ranges of 6.9–7.5 Hz and 6.9–13.8 Hz, respectively, where Fig. 8d is the final velocity model obtained from waveform tomography using all frequencies in the range.

Although the starting model generated by travel time tomography generates a smooth boundary for the water bottom, waveform tomography produces a sharp geometry with a spatial variation close to that shown in the seismic section (Fig. 5). The tomography image shows clearly a stratified structure underneath the water layer. In the second layer, at the distance between 0 and 10 km, there is a high-low-high vertical velocity variation immediately underneath the water bottom, between depth of 350 and 700 m at the left end of the profile, and a high-low-high-low vertical variation above the second interface. At the distance between 17 and 23 km, there is also a low-high-low vertical velocity variation within the second layer.

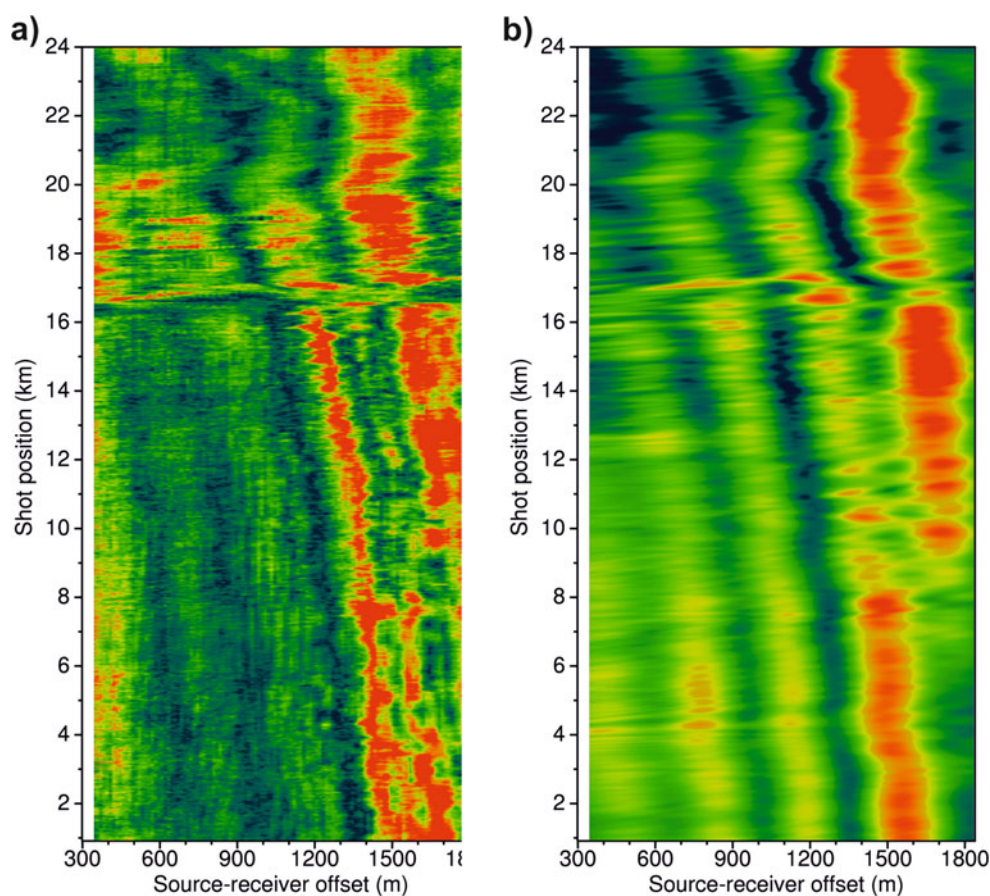
Most significantly, in the third layer, high-low-high velocity channels appear immediately underneath the second interface and cross the entire section. This thin-layer separation could not be generated by any conventional travel time tomography. Beneath that, the velocity pattern in the depth between 1.2 and 1.6 km varies laterally between 0–5, 5–17, and 17–23 km in distance.

Seismic reflection is generated due to the contrast in impedance which is the product of density and velocity. In the inversion example here, however, the density is assumed to be constant. Therefore, for a more accurate quantitative interpretation of the high velocity variation within the top-most thin layer immediately underneath the water layer, a density contrast between the water and water bottom sediment needs to be compensated.

Figure 9 compares a sample frequency slice (at 9.3 Hz) of real data input to waveform tomography and synthetic data generated from the inversion result. The vertical axis is the shot position in the surface, and the horizontal axis is the source-receiver offset. The data displayed are the amplitudes of complex-valued samples in the frequency domain and are normalized based on their RMS values. For the far-offset seismic traces, there are strong amplitudes for post-critical angle reflections, which play an important role in the

Seismic, Waveform Modeling and Tomography, Fig. 9

(a) A sample amplitude slice of real seismic data at frequency 9.3 Hz. (b) The amplitude slice of synthetics at the same frequency, obtained from the waveform tomography result



inversion for the shallow part of model. There is also a good fit at near-offset traces, which influence the high wavenumber perturbations of the velocity field.

Summary

Waveform tomography uses the original waveform recordings and wave equation modeling to extract high-resolution tomographic images from seismic data. Waveform tomography uses a group of frequencies simultaneously in an iterative inversion and proceeds from low to high frequencies. For seismic data with multiple sources, shot encoding which encodes all shots randomly and sums them into one super shot is an effective strategy to reduce the computational cost significantly by reducing the number of wavefield simulations in the inversion. However, this process will induce instability in the iterative inversion regardless of whether it uses a robust L-BFGS algorithm. The restarted L-BFGS algorithm is both stable and efficient. Even for seismic reflection data with limited source-receiver offsets, the waveform tomography also can potentially image subsurface features with detailed spatial variation at sub-wavelength scales.

Cross-References

- [Seismic Data Acquisition and Processing](#)
- [Seismic Imaging, Overview](#)
- [Seismic Ray Theory](#)
- [Traveltime Tomography Using Controlled-Source Seismic Data](#)

Bibliography

- Alford RM, Kelly KR, Boore DM (1974) Accuracy of finite-difference modeling of the acoustic wave equation. *Geophysics* 39:834–842
- Boonyasiriwat C, Valasek P, Routh P, Cao W, Schuster GT, Macy B (2009) An efficient multiscale method for time-domain waveform tomography. *Geophysics* 74:WCC59–WCC68
- Bunks C, Saleck FM, Zaleski S, Chavent G (1995) Multiscale seismic waveform inversion. *Geophysics* 60:1457–1473
- Deregowski SM, Brown SM (1983) A theory of acoustic diffractors applied to 2-D models. *Geophys Prospect* 31:293–333
- Kelly KR, Ward RW, Treitel S, Alford RM (1976) Synthetic seismograms: a finite difference approach. *Geophysics* 41:2–27
- Levander A (1988) Fourth-order finite-difference P-SV seismograms. *Geophysics* 53:1425–1437
- Mattsson K (2012) Summation by parts operators for finite difference approximations of second derivatives with variable coefficients. *J Sci Comput* 51:650–682
- Mattsson K, Nordström J (2004) Summation by parts operators for finite difference approximations of second derivatives. *J Comput Phys* 199:503–540
- Rao Y, Wang Y (2009) Fracture effects in seismic attenuation images reconstructed by waveform tomography. *Geophysics* 74:R25–R34
- Rao Y, Wang Y (2013) Seismic waveform simulation with pseudo-orthogonal grids for irregular topographic models. *Geophys J Int* 194:1778–1788
- Rao Y, Wang Y (2017) Seismic waveform tomography with shot-encoding using a restarted L-BFGS algorithm. *Sci Rep* 7:8494
- Rao Y, Wang Y (2018) Seismic waveform simulation for models with fluctuating interfaces. *Sci Rep* 8:3098
- Štekl I, Pratt RG (1998) Accurate viscoelastic modeling by frequency-domain finite differences using rotated operators. *Geophysics* 63:1779–1794
- Strand B (1994) Summation by parts for finite difference approximations for d/dx . *J Comput Phys* 110:47–67
- Svärd M, Mattsson K, Nordström J (2005) Steady-state computations using summation-by-parts operators. *J Sci Comput* 24:79–95
- Tarantola A (1984) Inversion of seismic reflection data in the acoustic approximation. *Geophysics* 49:1259–1266
- Tarantola A (1987) Inverse problem theory: methods for data fitting and parameter estimation. Elsevier, Amsterdam
- Virieux J (1986) P-SV wave propagation in heterogeneous media: velocity-stress finite-difference method. *Geophysics* 51:889–901
- Wang Y (2003) Seismic amplitude inversion in reflection tomography. Pergamon, Oxford
- Wang Y (2004) Multiple prediction through inversion: a fully data-driven concept for surface-related multiple attenuation. *Geophysics* 69:547–553
- Wang Y (2007) Multiple prediction through inversion: theoretical advancements and real data application. *Geophysics* 72:V33–V39
- Wang Y (2016) Seismic inversion: theory and applications. Wiley-Blackwell, Oxford
- Wang Y, Houseman GA (1994) Inversion of reflection seismic amplitude data for interface geometry. *Geophys J Int* 117:92–110
- Wang Y, Houseman GA (1995) Tomographic inversion of reflection seismic amplitude data for velocity variation. *Geophys J Int* 123:355–372
- Wang Y, Pratt RG (2000) Seismic amplitude inversion for interface geometry of multi-layered structures. *Pure Appl Geophys* 157:1601–1620
- Wang Y, Rao Y (2006) Crosshole seismic waveform tomography, I: strategy for real data application. *Geophys J Int* 166:1237–1248
- Wang Y, Rao Y (2009) Reflection seismic waveform tomography. *J Geophys Res* 114:B03304
- Wang Y, White RE, Pratt RG (2000) Seismic amplitude inversion for interface geometry: practical approach for application. *Geophys J Int* 142:162–172
- Zhang ZG, Zhang W, Chen XF (2014) Complex frequency-shifted multi-axial perfectly matched layer for elastic wave modelling on curvilinear grids. *Geophys J Int* 198:140–153

Seismicity, Intraplate

Paul Bodin

Pacific Northwest Seismic Network (PNSN), University of Washington, Seattle, WA, USA

Definition

Intraplate Seismicity refers to earthquakes that take place on faults other than those that form the boundaries between Earth's tectonic plates.

Historical Context

From the inception of plate tectonics as a simple set of unifying concepts, Earth scientists recognized that the notion of completely rigid plates should not be taken to be rigorous or complete. Deformations of Earth's crust and upper mantle, and the earthquakes that generally accompany them, must take place off of recognized plate interfaces. This fact was evident theoretically because in order for plate tectonics to operate on timescales greater than one Wilson cycle, new plate boundaries must form within plates, so deformation must take place off of existing plate boundaries. In addition, it was unassailable observationally because seismicity off of recognized plate boundaries clearly *does* take place, albeit at a relatively trivial pace compared to Earth's total earthquake production.

Two generations later, despite the fact that they obviously exist, intraplate earthquakes remain enigmatic and challenging; fitting subjects for special study. The dominant questions about these earthquakes not only reveals our incomplete understanding of earthquake physics in general, but also highlights a number of issues specific to active tectonics and seismic hazard in intraplate settings.

Definitions and Classification

We may define interplate earthquakes as being those that rupture parts of faults that connect directly mechanically to the globe-girdling system of recognized boundaries between tectonic plates. Intraplate earthquakes break faults that do not so connect. The difference is one of ultimate slip. Because interplate faults form a network within which new plates may be created and old plates consumed, slip on them is unbounded. On the other hand, the total slip on intraplate faults must terminate at the ends of the fault. So increasing slip comes at the cost of increasing strain in the surrounding rocks which, itself being presumably limited, caps how much any given fault can slip. Therefore, intraplate faults may be expected to slip generally more slowly, and less far, than interplate faults, and this possibly to have some consequence on the characteristics of earthquakes on the two types of faults.

It is customary for Earth scientists to distinguish further between intraplate earthquakes that break faults within *mobile zones* and those that take place within *stable zones*. Mobile zones are regions that, although not directly on recognized plate boundary fault, have been experiencing regional deformation distributed over a broad area near a plate boundary. An example is the active faulting across China, Tibet, and Mongolia, the deformations of which may be clearly related to three-dimensional complexities in the plate boundary framework. Stable zones are not only sufficiently remote from plate boundaries to be insensitive to their geometric complexities, but also are stable in the sense that significant tectonic events

have not affected them during the Cenozoic Era (~66 million years). They are generally expected to be areas of cold, thick continental lithosphere. Earthquakes in stable continental regions, often designated SCR earthquakes, are particularly enigmatic.

Reasons to Distinguish Intraplate Earthquakes

One reason to study intraplate earthquakes (and particularly SCR earthquakes) as a class involves a practical consequence: they strike the most unwary victims. Because of the long recurrence intervals between intraplate earthquakes, the threatened populace is generally unprepared and the structures they build are particularly vulnerable. It has recently been noted that this argument is becoming less restrictive to intraplate earthquakes as population growth within plate boundary regions has taken place so quickly of late that even these regions are becoming less prepared as a whole (Bilham 2009). Nevertheless, the argument remains applicable in plate interiors where the building stock may be not only not up to modern seismic codes, but is generally older.

In part, SCR earthquakes remain especially enigmatic not merely because they are rare, but also because their recurrence may be so long that (1) the sizes and sometimes locations of previous earthquakes are not well known, even for those that have occurred in historical times, and (2) it is usually not possible to ascertain what their recurrence is with certainty. This leads to great uncertainty about the seismic hazards in intraplate regions. Is damaging earthquake shaking possible anywhere (albeit with small probability within any 50-year time period)? Or are even SCR earthquakes restricted to occur in places with discernible (and hence predictable) characteristics? This is tantamount to a question of whether the entirety of Earth's crust at critical stress levels such that only small fluctuations or differences in fault strength control the occurrence of earthquakes.

With such long recurrence intervals, seismological observations are sparse and we rely on the paleoseismic evidence contained within the geological record to try to understand them. It is clear that plate tectonics is accurate to first order and that intraplate deformation takes place very slowly, which makes them a challenge for even modern high-precision geodetic techniques. It seems likely that progress in understanding intraplate earthquakes will remain slow for the time being.

Another reason they are of special interest is that they may reveal clues about elusive earthquake physics. While Earth scientists believe that nearly all tectonic earthquakes are caused by sudden slip on a fault driven by stresses in the surrounding rock (the elastic rebound theory), the specific mechanisms that localize rupture initiation and growth of a rupture remain obscure. Systematic differences between fault,

stress, and geological characteristics of intraplate faults discussed below may be exploited to test hypotheses of the important controls of earthquake fault rupture.

Do Intraplate and Interplate Earthquakes Scale Differently?

Because intraplate earthquakes, and particularly those in stable zones, have longer recurrence intervals, it is often hypothesized that their host faults will have longer time to “heal” between earthquakes (e.g., Scholz 2002). Fault healing here refers to a set of hypothesized fault processes such as the formation of minerals that bind the sides of the fault together making them stronger over time. Thus, intraplate earthquakes may be expected to break under higher loads – at higher stress levels – than their cousins that break more active plate boundary faults. The anticipated result is that they “scale” differently: that is, big earthquakes differ from small earthquakes in systematically different ways in interplate as opposed to intraplate regions. Specifically, intraplate earthquake ruptures will slip more in an earthquake per unit of fault area, relieve more stress (have a larger “stress drop”) and produce more seismic energy for their size, than the earthquakes we are more familiar with at plate boundaries. Thus seismologists expect, and arguably observe, that intraplate earthquakes tend to have higher stress drops than interplate earthquakes. Seismological evidence suggests that as a class their stress drops may be a factor of 2–5 higher than plate boundary earthquakes (Kanamori and Allen 1986). But because observations are so few, much uncertainty remains about this issue.

Source scaling studies are further hindered by systematic differences between seismic wave propagation in intraplate regions and that in more mobile or active zones. Because rock surrounding plate boundaries is often less intact and hotter than in plate interiors, seismic wave energy is more strongly absorbed, or attenuated, than in the cold hard hearts of plates. So, per unit of energy released at their source, the effects of intraplate earthquakes are more widely distributed – they have a longer reach.

What Drives Intraplate Earthquakes?

Probably the most critical enigma of intraplate seismicity concerns the source of driving stresses. It is easy to invoke “left over” or residual stresses originating from geometric irregularities or variations in frictional properties on the master plate boundary faults to explain the origin of mobile zone intraplate earthquakes. Remote from plate boundaries, occasionally sources of localized stresses may be recognized. One clear example is from intraplate regions experiencing uplift from glacial rebound, as ice sheets retreat with global

warming. This mechanism well explains the seismicity pattern of Sweden, for example (e.g., Möner 2009). Swarms of seismicity often accompany geothermal activity, even at sites in the remote interior of tectonic plates (e.g., Ibs-von Seht et al. 2008). Where such sources of localized stress are not evident, however, it is more problematic to assign the provenance of seismic stress. Modeling studies suggest that stresses within plates may be best explained as a combination of resolved stresses originating at plate boundaries (although they may be thousands of kilometers distant) plate bending due to a plate’s travel over a non-spherical Earth, or forces arising from convective patterns in Earth’s mantle.

Many scientists presume that intraplate earthquakes take place at sites where crustal rocks have been significantly weakened compared to those in surrounding rocks. In this model, seismicity should be associated with reactivated fault zones. Big intraplate earthquakes are thought to be restricted to areas that have previously been faulted, say within failed continental rift systems (e.g., Johnson 1996).

One intriguing idea is that the strength of faults could be reduced by high pore-fluid pressure. Fluids (which could include water or even carbon dioxide or hydrocarbons) that occupied the internal portion of a fault could become pressurized and reduce the frictional resistance of the fault to slipping. That would imply that seismically active faults in intraplate regions are those that contain significant quantities of fluids at high confining stress. Sufficient stress to drive earthquakes could indeed be present everywhere within plate interiors, and the conditions to release the stress controlled by something external.

Examples

The moment magnitude (M_w) = 6.8 “Nisqually” earthquake of January 23, 2001 in Washington state, USA, provides an example of an intraplate earthquakes breaking a fault within the crust of the subducting ocean plate (Ichinose et al. 2006). The earthquake resulted from slip on a normal fault driven by bending stresses as the downgoing slab flexes in its descent. On a larger scale, the Cascadia subduction zone provides an interesting example of the interplay between inter- and intraplate earthquakes in that while the master plate boundary megathrust fault is broken about every 500 years with large subduction thrust events (e.g., Satake and Atwater 2007), there appears to be no current seismicity on the plate boundary fault. Rather all earthquakes either historical or instrumentally recorded are either intra-slab events or crustal earthquakes within the overriding North America plate. These events are interplate by our definition. However, they are clearly associated with a nearby subduction interface and are driven by residual stresses that the plate boundary is unable to completely relieve.

The May 13, 2008 $M_w = 7.9$ Wenchuan earthquake in the Szechuan Province of China is an earthquake in the tectonically mobile zone of Asia, and is part of the escape of material from Asia and the uplifting of the Himalaya due to the collision of the Indian subcontinent with Asia. About 300 km of a fault on the boundary between the Longmenshan orogenic belt and the Szechuan basin broke in the earthquake (Burchfiel et al. 2008). With a high overall stress drop of 17 MPa and localized patches of up to 53 MPa (Zhang et al. 2009), the Wenchuan earthquake had large slip and disastrous effects on a populace that was largely unaware of the potential for, and unprepared for, this long-recurrence-interval (2,000–10,000 years) intraplate earthquake.

The January 26, 2001 $M_w = 7.6$ Bhuj earthquake took place ~400 km from the rather diffuse onshore portion of the left-lateral strike-slip Makran boundary between India and the Arabian and Eurasian plates, but within the western part of the peninsular Indian shield. It broke the deeper part (~5 to ~40 km) of a previously unknown reverse fault. It was located within ~100 km of a similar-sized (though shallower) earthquake in 1835 that generated a surface rupture – the Allah Bundh in the Great Rann of Kachchh. The stress drop was a high, but not extreme ~16–20 MPa, but the earthquake was very powerful, with the source lasting only about 7 s. Widespread liquefaction and secondary effects were observed, but the large death toll (~30,000) was apparently not the result of extreme ground motions, which could have been engineered for, but rather of poor construction and preparation.

Controversy

One particularly illustrative and significant example of SCR interplate seismicity concerns the New Madrid Seismic Zone that threatens a 7-state region surrounding the lower Mississippi River valley in the south central USA (e.g., Johnson and Schweig 1996). The region was the site of at least three major earthquakes during the winter of 1811–1812. While at the time thinly populated and rural, a return of these earthquakes today would affect ten million people and put at risk numerous cities, commerce, and transportation and communication infrastructure. The 1811–1812 earthquakes were accompanied by a host of surface effects (liquefaction, landslides, surface deformation) that remain visible in the landscape today. Ground motions from the earthquakes were also reported to be felt as far away as Charleston, South Carolina (~950 km) and even Boston, Massachusetts (~1600 km). There is a complex zone of active microearthquakes producing several $M < 3$ earthquakes each week. Early research based on ground-shaking intensities derived from historical reports proposed that the three largest earthquakes in the series exceeded M_8 . The ~100-km long zone of active microseismicity, which many presume to represent the active faults,

seems much too small to host even a single M_8 earthquake, let alone three, unless stress drops, and therefore slip per unit length, were extremely large. The earthquakes occupy a shallow basin formed atop a failed continental rift, formed in latest Precambrian time and briefly reactivated (but once again failed) in the Cretaceous. The faults are oriented such that they are nearly optimally oriented to be reactivated as thrust and strike-slip fault in the current compressional mid-plate stress regime.

Subsequent reexamination of the historic intensity data, and taking into account the feeble attenuation of seismic waves within rocks in the plate interior, have yielded generally smaller estimates of magnitudes, as low as M_7 for the three largest 1811–1812 earthquakes. In the absence of new information, however, it might be best to admit that while these earthquakes were undoubtedly large, their magnitudes will remain forever moot. Paleoseismic studies of earthquake-induced liquefaction features in surface deposits reveal at least four repeated episodes of strong shaking (as strong as what took place in the early nineteenth century), often in clusters like the 1811–1812 events, with an average recurrence interval of ~500 years. The geological evidence for recurrent large earthquakes is balanced by the lack of primary surface faulting, or even the prominent topographic relief one would expect if much slip had taken place at the rates paleoseismic evidence infers for the past several thousand years. So either the big events are really a good deal smaller than the largest estimates of their size or seismic activity in the zone was initiated only a few thousand years ago.

Geodetic evidence, from repeated campaign surveys and continuous GPS studies show that deformation surrounding the active seismic zone (Smalley et al. 2005) is very slow, with some (e.g., Newman et al. 1999) arguing that it is not different from zero. However, given the lack of a mid-plate fault model it is not clear what deformation pattern one should expect. So there is no general agreement about what the lack of observed active deformation means, and the interpretations remain contentious. Some suggest that either the largest earthquakes to be expected from the seismic zone are not so terrifyingly large as previously thought, or perhaps that, however large they were, they are not being recharged with stress and are not to be feared. Others hold that the paleoseismic, historical, and microseismic evidence point to the likelihood of future damaging earthquakes just like the historical ones, albeit driven by processes that are still beyond our ken.

Summary or Conclusions

Intraplate seismicity is a catchall phrase that encompasses a wide variety of seismogenic phenomena; essentially all earthquakes that do not occur on plate bounding faults. Locales

hosting intraplate earthquakes range from broad deformation zones that surround some plate boundaries to areas that lie within the cores of continents. Larger earthquakes – especially those in what are otherwise tectonically stable cold continental crust – appear to be associated with faults that formed in an earlier tectonic setting but are reactivated by the stress field applied currently. Not all faults are so reactivated, so it is likely that some as-yet-unclear weakening mechanism causes only certain faults to fail. Localized deformation, say glacial rebound or geothermal stresses can be responsible for interplate earthquakes, but does not explain all of them. As a class of faults, intraplate faults usually slip more slowly and less far than interplate faults, and this leads to longer intraplate earthquake recurrence intervals. Also, and probably related to this, intraplate earthquakes may generally be expected to have larger stress drops. Although less frequent than plate boundary earthquakes, intraplate seismicity is very dangerous because the population and infrastructure at risk from them is less prepared than in more seismically active areas. The study of intraplate seismicity is an endeavor that, because of its low rate is best done by integrating across different source regions and combining different Earth science disciplines.

Cross-References

- [Earthquake, Focal Mechanism](#)
- [Earthquake, Magnitude](#)
- [Earthquakes and Crustal Deformation](#)
- [Earthquakes, Strong-Ground Motion](#)
- [Geodesy, Figure of the Earth](#)
- [Geodesy, Ground Positioning, and Levelling](#)
- [Geodynamics](#)
- [Mantle Viscosity](#)
- [Paleoseismology](#)
- [Plate Motions in Time: Inferences on Driving and Resisting Forces](#)
- [Plate Tectonics, Precambrian](#)
- [Plate-Driving Forces](#)
- [Plates and Paleoreconstructions](#)

Bibliography

- Bilham R (2009) The seismic future of cities. *Bull Earthq Eng* 7:839. <https://doi.org/10.1007/s10518-009-9147>
- Burchfiel BC, Royden LH, van der Hilst RD, Hager BH, King RW, Li C, Lü J, Yao H, Kirby E (2008) A geological and geophysical context for the Wenchuan earthquake of 12 May 2008, Sichuan, People's Republic of China. *Geol Soc Am Today* 18(7). <https://doi.org/10.1130/GSATG18A.1>
- Ibs-von Seht M, Plenefisch T, Klinge K (2008) Earthquake swarms in continental rifts – a comparison of selected cases in America, Africa and Europe. *Tectonophysics* 452(1–4):66–77. <https://doi.org/10.1016/j.tecto.2008.02.008>. ISSN 0040-1951

- Ichinose GA, Thio HK, Somerville PG (2006) Moment tensor and rupture model for the 1949 Olympia, Washington, earthquake and scaling relations for Cascadia and Global Intraslab earthquakes. *Bull Seismol Soc Am* 96(3):1029–1037. <https://doi.org/10.1785/0120050132>
- Johnson AC (1996) Seismic moment assessment of earthquakes in stable continental regions II: historical seismicity. *Geophys J Int* 125(3):639–678
- Johnson AC, Schweig ES (1996) The enigma of the New Madrid earthquakes of 1811–1812. *Annu Rev Earth Planet Sci* 24:339–384
- Kanamori H, Allen CR (1986) Earthquake repeat time and average stress drop. In: Das S et al (eds) *Earthquake source mechanics*. AGU geophysics monograph, vol 37. American Geophysical Union, Washington, DC, pp 227–236
- Mörner N-A (2009) Late Holocene earthquake geology in Sweden. *Geol Soc Lond, Spec Publ* 316:179–188. <https://doi.org/10.1144/SP316.11>
- Newman A, Stein S, Weber J, Engeln J, Mao A, Dixon T (1999) Slow deformation and lower seismic hazard at the New Madrid Seismic zone. *Science* 284:619–621. <https://doi.org/10.1126/science.284.5414.619>
- Satake K, Atwater BF (2007) Long-term perspectives on giant earthquakes and tsunamis at subduction zones. *Annu Rev Earth Planet Sci* 35:349–374. <https://doi.org/10.1146/annurev.earth.35.031306.140302>
- Scholz CH (2002) *The mechanics of earthquakes and faulting*. Cambridge University Press, Cambridge, UK
- Smalley R, Ellis MA, Paul J, Vanarsdale R (2005) Space geodetic evidence for rapid strain rates in the New Madrid seismic zone of central USA. *Nature* 435:1088–1090. <https://doi.org/10.1038/nature03642>
- Zhang Y, Feng W, Xu L, Zhou C, Chen Y (2009) Spatiotemporal rupture process of the 2008 great Wenchuan earthquake. *Sci China Ser D Earth Sci* 52(2):145–154

Seismicity, Subduction Zone

Akira Hasegawa

RCPEV, Graduate School of Science, Tohoku University, Sendai, Japan

Definition

- | | |
|------------------------|---|
| Subduction zone | Consuming plate boundary where the oceanic plate subducts into the mantle beneath the continental plate. |
| Interplate earthquakes | Earthquakes that occur along the plate boundary. In case of the subduction zone, they occur at the shallow portion of the boundary between the subducting plate and the overriding plate. |
| Intraslab earthquakes | Earthquakes that occur within the subducted plate, that is, within the slab. |
| Shallow earthquakes | Earthquakes that occur at depths shallower than 60 km. Focal depths of interplate earthquakes in subduction zones are usually shallower than 60 km, and so they are shallow earthquakes. |
| | Earthquakes that occur at depths of 60–300 km. |

Intermediate-depth earthquakes

Deep earthquakes Earthquakes that occur at depths deeper than 300 km. Both intermediate-depth and deep earthquakes occur only within the subducted plate, and so they are intraslab earthquakes.

Introduction

There are seismically active and inactive areas on the Earth. Most of Earth's seismic activity is concentrated in narrow continuous belts that wrap around it. Plate tectonics proposes that the Earth's entire surface is comprised of somewhat more than ten large, rigid, mosaic-like plates, and that the relative movement of these plates causes tectonic activity at their boundaries. Consequently, most earthquakes are concentrated at plate boundaries, forming narrow, belt-shaped seismic zones.

Oceanic plates are generated at mid-ocean ridges by the upwelling of mantle material. Since the Earth's surface area is constant, the surface area does not expand with the generation of the oceanic plate. Instead, the same amount of material moves back down into the mantle. The generation of plates is balanced by the consumption of other plates elsewhere at consuming plate boundaries. Subduction zones are where oceanic plates plunge downward into the mantle. At subduction zones, two plates collide with each other; the heavier plate is overridden by the lighter one and sinks down into the mantle. Thus, the heavier oceanic plate subducts beneath the lighter continental plate. The subduction of the oceanic plate causes high seismic activity. In fact, it is in these subduction zones that most of the world's large earthquakes occur.

Mainly four types of earthquakes take place in subduction zones. The first type is interplate earthquakes that occur along the interface between the subducting oceanic plate and the continental plate (A in Fig. 1). The second type of events occur in the trench-outer rise region of the subducting plate (B in Fig. 1). The third is shallow inland earthquakes that occur in the shallow portion of the overriding continental plate (C in Fig. 1). The last is intraslab earthquakes that occur at depths more than about 60 km within the subducted oceanic plate (D in Fig. 1).

Interplate Earthquakes

In subduction zones, the denser oceanic plate sinks down into the mantle underneath the lighter continental plate. The shallow portion of the plate boundary is locked by friction, and in order to release accumulated stress, occasional slips at the plate boundary occur and cause earthquakes (A in Fig. 1). Most of the world's large, shallow earthquakes are interplate

earthquakes caused in this way. Interplate coupling occurs only within a specific range of depths, which seems to be primarily determined by the temperature of the plate interface.

Oleskevich et al. (1999) estimated the thermal structures of plate interfaces in several subduction zones, and compared them with the updip and downdip depth limits of source areas of large interplate earthquakes. The results show: (1) the updip limit of the depth of interplate earthquakes corresponds with the depth at which the temperature of the plate boundary reaches 100–150 °C and (2) the downdip limit approximately matches the depth at which the temperature reaches 350 °C (Fig. 2). In the shallowest part of plate interfaces, smectite clays, which are clay minerals distributed in the sediments there, prevent interplate coupling by producing stable sliding. The temperature rises with increasing depth, and when it reaches 100–150 °C, dehydration decomposition occurs and the smectite clay turns into illite and chlorite. This transformation is considered the onset of interplate coupling.

Then, deeper, where the temperature exceeds 350 °C, interplate coupling ceases and stable sliding takes place again. Figure 2 shows that the transition zone from unstable sliding to stable sliding occurs at depths where the temperature ranges from 350 °C to 450 °C. In the case where the depth at which the temperature of the plate interface reaches 350 °C is deeper than the depth at which it contacts the Moho of the upper plate, the contact zone with the Moho is considered to be the downdip limit. The reason the downdip is limited to the Moho contact zone is that stable sliding prevails where the mantle of the upper plate contacts the oceanic plate because the mantle may be serpentinized by water supplied by the dehydration of the subducting oceanic plate. It is considered that interplate earthquakes do not occur any more under those conditions (Hyndman et al. 1995). In some subduction zones, such as in northeast Japan, the upper plate's mantle is not serpentinized, since the seismic wave velocities there show normal values of non-serpentinized mantle materials. In such cases, the downdip limit is once again determined by the plate interface temperature of ~350 °C.

The degree of interplate coupling can be indicated by how much interplate slip is accounted for by earthquakes. This index is called "seismic coupling" and denotes the percentage of the seismic slip among all slip amounts occurring at the "seismogenic zone" in Fig. 2. If seismic coupling is 1, all slip occurring in that area is caused by earthquakes. In contrast, if it is 0, it means that all slip occurring in that area is aseismic stable sliding. Seismic coupling differs from one subduction zone to another.

Interplate coupling along the seismogenic zone is heterogeneous both in space and time; the seismogenic zone is not fully locked, as mentioned above, and some portions of it will slide stably during interseismic periods. Seismic waveform inversions of large interplate earthquakes have revealed spatially heterogeneous seismic slip along the seismogenic zone,

Seismicity, Subduction Zone,

Fig. 1 Schematic illustration of subduction zone seismicity. Earthquakes are shown by open circles on a cross-arc vertical cross section of a typical subduction zone. Arrows attached to open circles indicate dominant focal mechanisms; two parallel arrows in opposite directions show seismic slip along the plate interface, two arrows pointing outward show normal faulting and those pointing inward thrust faulting. Low-velocity zones in the crust and the mantle wedge of the upper plate and in the uppermost portion of the subducting plate are shown by shaded areas

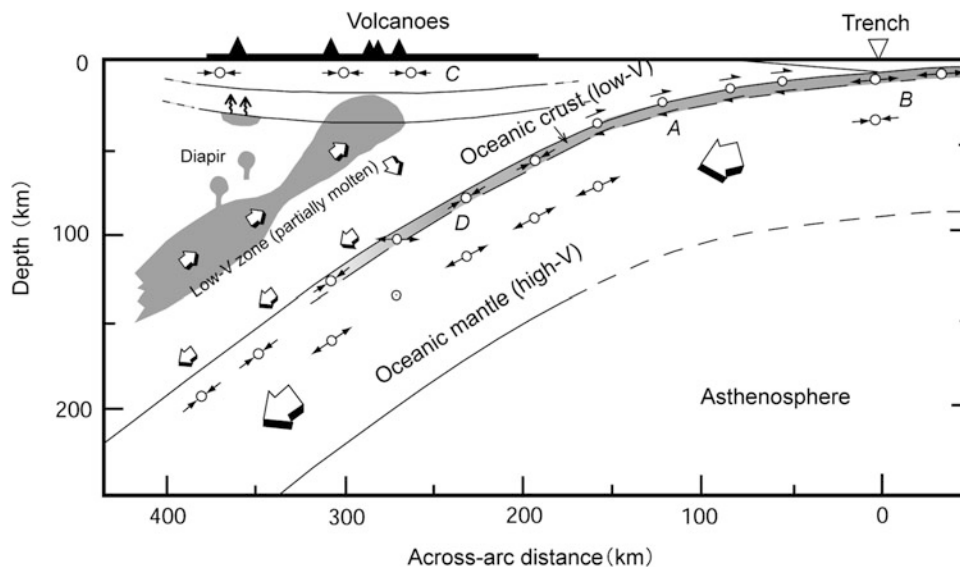
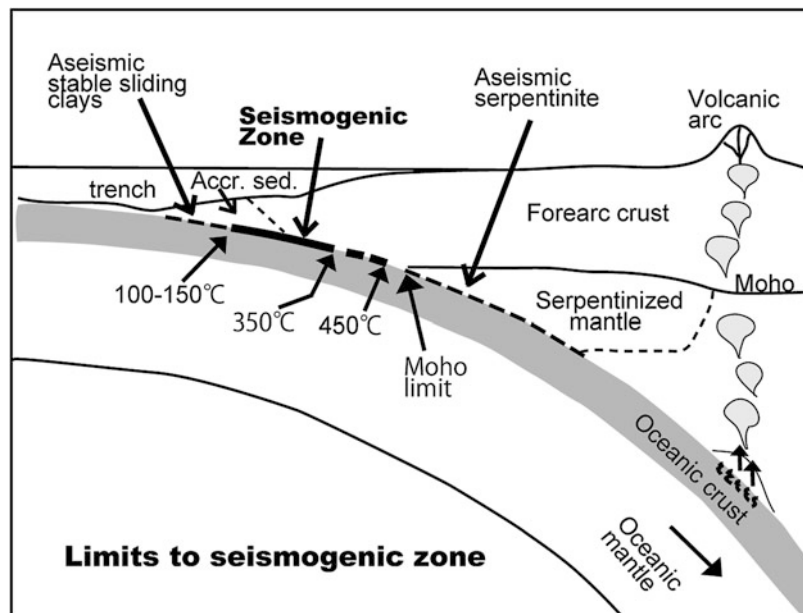
**Seismicity, Subduction Zone,**

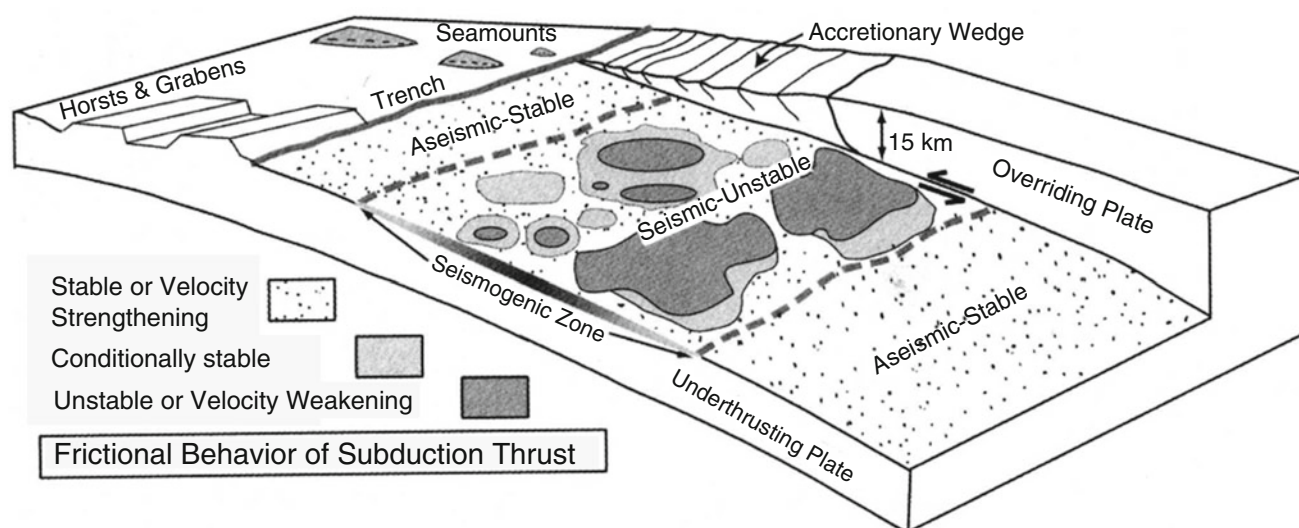
Fig. 2 Sketch of the subduction plate boundary seismogenic zone. (Oleskevich et al. 1999, copyright by the American Geophysical Union)



and backslip inversions of geodetic data, which estimate spatial distribution of interplate coupling, have confirmed spatially heterogeneous strain accumulation along the zone during the interseismic period. A conceptual model of frictional properties on the fault surface based on laboratory-derived fault constitutive laws (Fig. 3) explains the observed heterogeneous interplate coupling along the seismogenic zone. It involves patches of frictionally unstable fault surface surrounded by stable and conditionally stable fault surfaces. Frictionally stable portions of the fault are subject to steady aseismic sliding, whereas frictionally unstable patches (asperities) remain locked during the interseismic period. Then they finally fail, as evidenced by the occurrence of

earthquakes. Conditionally stable portions typically slide stably, but may fail and slide dynamically if loaded at high strain rates by the seismic slip of a neighboring asperity. Earthquakes nucleate only in the unstable patches, and once an earthquake occurs in an unstable patch, its seismic slip propagates throughout the patch, with some penetration into the neighboring conditionally stable areas. This conceptual model indicates that the larger the area occupied by the frictionally unstable patches in the seismogenic zone, the larger the seismic coupling of that subduction zone is.

A comparative study of the recent three great mega-thrust earthquakes, 2004 Sumatra-Andaman (Mw 9.2), 2010 Chile (Mw 8.8), and 2011 Tohoku (Mw 9.0), showed systematically



Seismicity, Subduction Zone, Fig. 3 A conceptual model of frictional properties on the subduction plate boundary. (Originally from Bilek and Lay (2002) and modified by Dixon and Moore (2007), copyright by the Columbia University Press)

varying rupture properties with depth on the plate interface (Lay et al. 2012). At very shallow depths shallower than ~15 km, sediments and high pore pressure cause slow rupture expansion as tsunami earthquakes. At shallow depths from ~15 to ~35 km, large and relatively uniform frictionally unstable patches are distributed, and they generate large earthquake slip but modest amount of short-period seismic radiations upon failure. At ~35 to ~55 km depths, smaller isolated unstable patches are distributed, and these areas are surrounded by conditionally stable areas. When these areas fail, coherent short-period radiations are produced which brings local strong ground motions. Deeper than that, only stable aseismic sliding occurs. However, the cause is not known so far why such a systematic variation of rupture properties with depth is commonly found. Further studies are needed to understand it.

Trench-Outer Rise Intraplate Earthquakes

The oceanic plate is bent downward as it sinks from the trench into the interior of the Earth. In order to release the stress caused by this bending deformation, earthquakes occur in the shallow portion of the subducting oceanic plate (B in Fig. 1). They are called trench-outer rise earthquakes because they occur in the outer rise region seaward of the trench which is a topography high formed with the bending deformation of the oceanic plate.

In the northeast Japan subduction zone, the trench-outer rise earthquakes form two parallel upper and lower planes separated by 25–30 km from each other (Gamage et al. 2009). Earthquakes in the upper plane have normal-fault type focal mechanism, while those in the lower plane show reverse-fault focal mechanism. This is opposite to the direction of the stress

field of the double seismic zone at depths of 70–150 km within the subducted plate (see section “[Intermediate-Depth Earthquakes](#)”). Such formation of the two parallel seismic planes in the outer rise region can be understood if we take finite depth of brittle layer of the oceanic plate into consideration. Since temperature increases with depth, only the upper half of the oceanic plate has brittle properties and the remaining lower half exhibits ductile properties. Therefore, associated with the downward bending deformation of the oceanic plate, a tensile stress is applied at the upper end of the brittle upper half of the plate and a compressive stress at its lower end. This causes the formation of the two parallel seismic planes; one at the upper end of the brittle upper half of the plate with normal-fault focal mechanism (upper plane) and the other at the lower end with reverse-fault mechanism (lower plane).

Shallow normal-fault type earthquakes, that is, upper plane events, are found to occur in outer rise regions of many other subduction zones. In addition, as in the northeast Japan subduction zone, lower plane events with reverse-fault mechanism have been also found in several subduction zones. Seno and Yamanaka (1996) pointed out that, in such subduction zones, intermediate-depth earthquakes form a double seismic zone in a subducted oceanic plate, suggesting a causative relationship between the two parallel shallow seismic planes in the outer rise region before the plate subduction and the double seismic zone of intermediate-depth events after the subduction.

Shallow Inland Intraplate Earthquakes

Convergence of two plates at subduction zones causes stress accumulation also in the overriding continental plate,

producing deformation in it well beyond the trench axis. Usually, this deformation is compressional crustal shortening, which partially accommodates the plate convergence. Shallow earthquakes occur in the crust of the overriding plate (C in Fig. 1), reflecting this ongoing compression. Focal mechanisms of these earthquakes are thrust fault or strike-slip fault with compressional axis oriented nearly in the plate convergence direction. Recurrence intervals of these types of earthquakes are much longer than those of interplate thrust earthquakes. However, once they occur, it often causes serious damages in lives and societies due to the shallow hypocenter and proximity to inhabited areas.

In most cases, these shallow inland earthquakes occur in the depth range between 1–3 and 10–20 km, forming the seismogenic zone mainly in the upper crust of the overriding plate. Deformation is made by aseismic ductile flow at shallow depths above the upper bound because of the presence of unconsolidated granular material, and at large depths below the lower bound because of the onset of plasticity at a critical temperature (Scholz 1998). This critical temperature is considered to be about 300–400 °C, and so the lower bound of the seismogenic zone varies depending on the local thermal gradient. Active faults, which are surface traces of repeated fault movements by earthquakes, show locations where in the seismogenic zone of the overriding plate large shallow earthquakes have occurred in the recent past, the Quaternary Period. Surveys of active faults further provide information on history of earthquakes, including their magnitudes and the time of the last event, which is basic and important for long-term earthquake forecast of shallow inland intraplate earthquakes.

Although we can learn from active faults where large earthquakes have occurred, it is difficult to understand why earthquakes occur in those locations. Recent investigations based on dense GPS and seismic observation network data in Japan have provided some information on this problem. Analyses of GPS data showed the existence of belt-like zones with concentrated crustal shortening in the overriding plate, along which many large shallow earthquakes have occurred or active faults are distributed (Sagiya et al. 2000; Miura et al. 2004). Seismic tomography studies revealed prominent low seismic velocity zones in the lower crust to the uppermost mantle right beneath these concentrated deformation zones (Nakajima et al. 2001; Nakajima and Hasegawa 2007). The low seismic velocities are inferred to reflect the existence of aqueous fluids, which are perhaps supplied from the subducted slabs. These observations suggest that the fluids weaken the surrounding crustal materials and cause the concentrated deformation there, leading to large shallow inland earthquakes (Iio et al. 2004; Hasegawa et al. 2005). Increase in pore fluid pressure along the fault planes should contribute to the occurrence of these shallow inland earthquakes by decreasing the fault strength (Sibson 2009). If this is the

case, places to which aqueous fluids are supplied intensively from the mantle wedge below are considered to be locations where large shallow earthquakes occur.

Intraslab Earthquakes

Earthquakes that occur at depths of more than about 60 km are found only in subduction zones. That is because these earthquakes occur in subducting oceanic plates or slabs (D in Fig. 1). They are distributed in an inclined plane in the mantle, which is called “deep seismic zone” or “Wadati–Benioff zone,” after its discoverers. The frequency distribution of earthquakes for each depth range throughout the world (Fig. 4) shows that depths from 0 to 50–60 km have the greatest occurrence frequency. Proceeding deeper, the frequency decreases nearly monotonically until a depth of about 300 km, then starts to increase with depth from that point, reaching a local maximum at 500–600 km deep. The earthquake depth range can be divided into three zones: shallow (0–60 km), intermediate depth (60–300 km), and deep (deeper than 300 km). Of these, intermediate-depth earthquakes and deep earthquakes occur within the slab, so they are called intraslab earthquakes (or slab earthquakes).

In many cases, the focal mechanisms of intraslab earthquakes are classified as either downdip compression (DC) earthquakes, in which the compressional axis is oriented in the dip direction of the slab, or downdip extension (DE) earthquakes, in which, on the contrary, the tensional axis is oriented in the dip direction of the slab. There is a systematic tendency in the depth distribution of focal mechanisms that depends on the age of the subducting plate. That is, as indicated in Fig. 5: (1) DE-type earthquakes dominate in young plate subduction zones such as in Central America, in which intraslab earthquakes occur only to a depth of about 250 km. (2) Intraslab earthquakes occur to a depth of about 650 km in old plate subduction zones such as Tonga and the Izu-Bonin arc, and the DC-type focal mechanism dominates at all depths in those areas. (3) Subduction zones such as the Kermadec and Kuriles areas have characteristics intermediate between (1) and (2); DE-type earthquakes dominate in shallow reaches while DC-type earthquakes dominate deeper. (4) In subduction zones like Chile and the New Hebrides areas, in which intraslab earthquakes do not occur at depths between 300 and 500 km, the DE-type dominates at shallow depths while the DC-type dominates deeper.

Slab mineralogy studies have shown that depth perturbations of phase transformation boundaries resulting from thermal anomalies within the slabs generate positive and negative buoyancy forces depending on depth, and so are major contributors to stress fields in the slabs (e.g., Bina et al. 2001). Figure 6 shows a diagram of the mineralogy in a subducting slab. Since the slab is colder than the surrounding mantle, the

Seismicity, Subduction Zone,

Fig. 4 Focal depth distribution of global seismicity. Earthquakes with magnitudes greater than 4, 5, and 6 in the EHB catalog (Engdahl et al. 1998) for the period from 1964 to 2004 are plotted

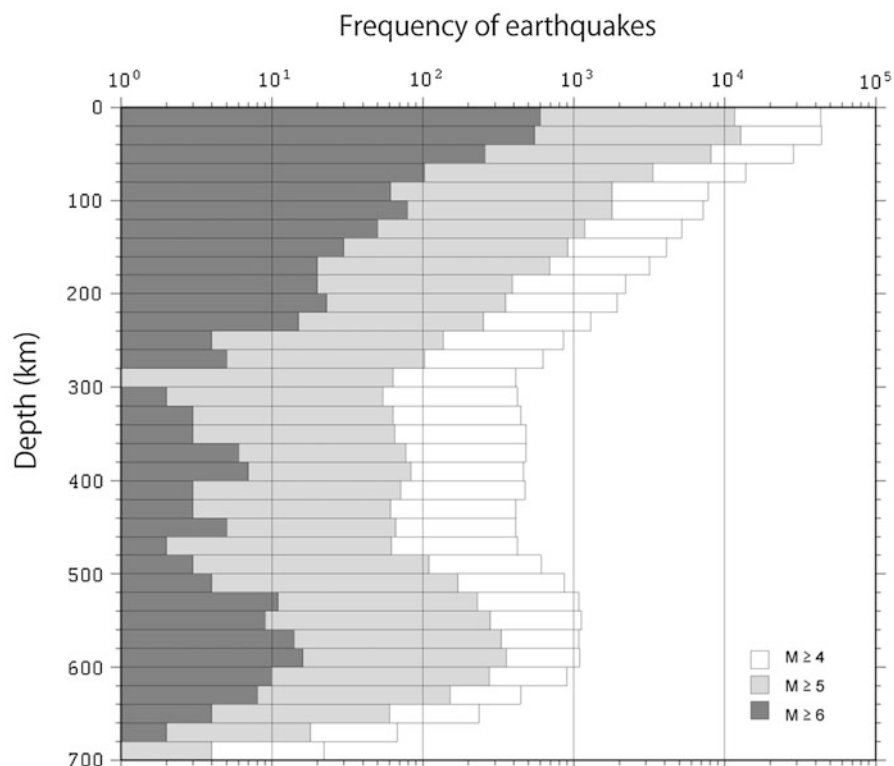
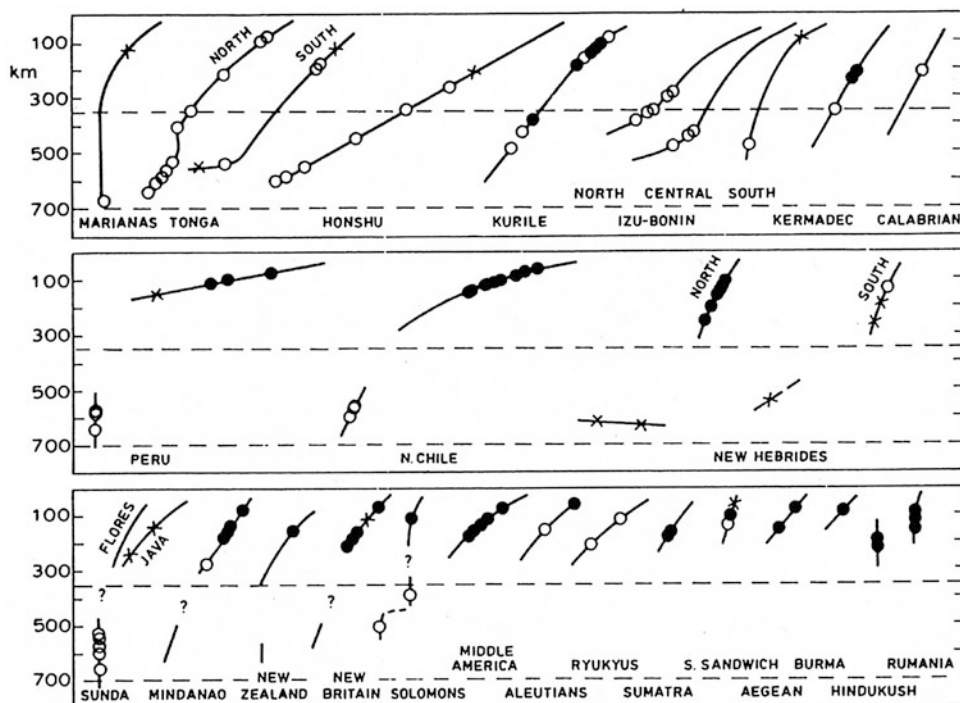
**Seismicity, Subduction Zone,**

Fig. 5 Global summary of stress states within subducting slabs (Isacks and Molnar 1971, copyright by the American Geophysical Union). Open and solid circles show downdip compression and downdip extension, respectively. Crosses indicate that neither the compression nor tension axis is approximately parallel to the slab dip



α -olivine to β -spinel phase transformation at 410 km depth and the β -spinel to γ -spinel phase transformation at 550 km depth become shallower, whereas decomposition of γ -spinel to perovskite+magnesiowustite at 660 km depth becomes

deeper than in the ambient mantle. Moreover, the olivine to β -spinel transition must be kinetically hindered within the cold slab, since the relatively colder temperatures in the slab inhibit reaction rates. A wedge-shaped zone of metastable

peridotite thus formed in the cold core of the slab persists to the depth shown by the shaded area in Fig. 6 (Kirby 1995).

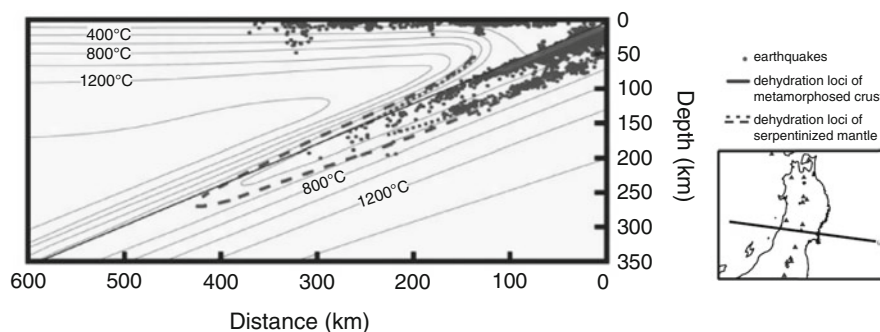
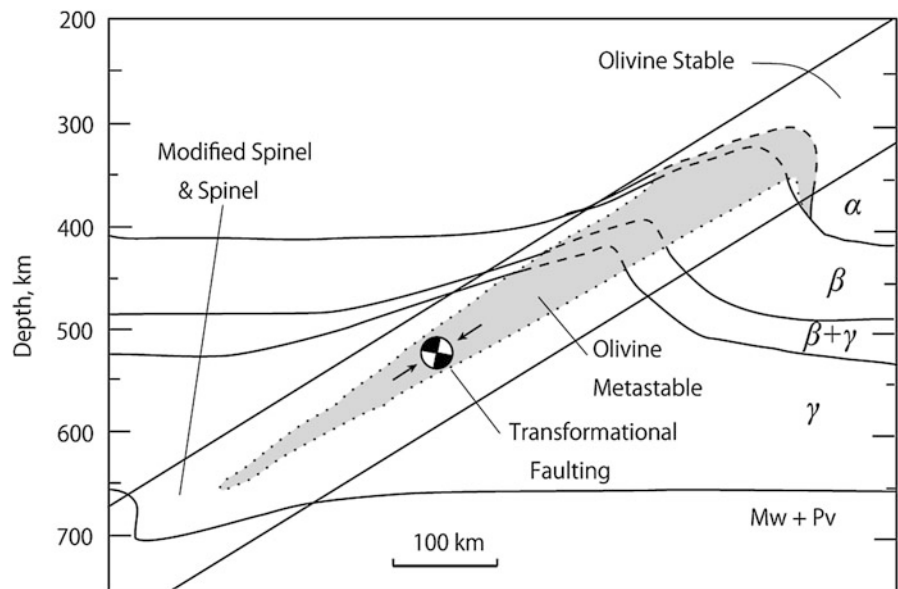
Upward deflections of the olivine to β -spinel transition at 410 km and the β -spinel to γ -spinel transition at 550 km yield negative buoyancy forces by stabilizing the dense higher pressure phases in the slab surrounded by the less dense lower pressure phases in the ambient mantle. On the contrary, the downward deflection of the γ -spinel to perovskite +magnesiowustite transition at about 660 km yields positive buoyancy forces. Moreover, the metastable olivine wedge yields positive buoyancy forces. Downdip compression at depths deeper than about 300 km, observed for all the subduction zones (Fig. 5), can be well explained by the compressional stress generated by the combination of the buoyancy forces produced by the metastable olivine wedge and by the downward deflection of the perovskite-forming transition, and the force resisting slab penetration into the lower mantle

(Chen et al. 2004). It is expected that the downdip compressional stress caused by the metastable olivine wedge is smaller for warm (young) and slow slabs. Chen et al. (2004) found a good correlation between the thermal parameter of the slab and the nature of the downdip stress at depths shallower than about 300 km in accordance with this expectation: warm, slow slabs tend to have tensional stress; and cold, fast slabs tend to have compressional stress.

Intermediate-Depth Earthquakes

It is known that intermediate-depth intraslab earthquakes form a double seismic zone in several subduction zones. This was first clearly found beneath northern Honshu, in northeast Japan (Fig. 7), where the deep seismic zone is composed of two planes at depths of 70–150 km that are

Seismicity, Subduction Zone, Fig. 6 Schematic diagram of the mineralogy within a subducting slab (Kirby 1995, copyright by the American Geophysical Union)



Seismicity, Subduction Zone, Fig. 7 Double seismic zone and dehydration loci of oceanic crust and mantle for central Tohoku, northeast Japan, projected on an across-arc vertical cross section (Yamasaki and Seno 2003, copyright by the American Geophysical Union). The solid

line and broken and dotted lines show the dehydration loci of the crust and mantle, respectively. Dots denote earthquake hypocenters. Medium thick and thin lines are the upper plate interface and isotherm contours, respectively

parallel to each other and separated by about 30 km (Hasegawa et al. 1978). Upper-plane earthquakes, most of which occur in the crust of the slab, have DC-type focal mechanisms, while lower-plane ones, occurring in the mantle of the slab, are characterized as DE-type earthquakes. Engdahl and Scholz (1977) explained the formation of such a double-planed deep seismic zone by the unbending of the subducting oceanic plate at this depth range. Later studies showed, however, that DC-type upper-plane and DE-type lower-plane events are not always seen and that in some subduction zones both the upper and lower planes have DC stress. These observations indicate that the unbending force may not be the main cause of the formation of the double seismic zone, although it must be acting on the plate.

Similar to ordinary shallow earthquakes, most intermediate-depth and deep earthquakes have double couple radiation patterns and little or no isotropic component (Frohlich 2006). However, it has been thought that intermediate-depth and deep earthquakes are mechanically different from shallow earthquakes that are generated by brittle fracture. This is because normal stress on the fault becomes too high to cause ordinary brittle shear faulting. At depths of 60–650 km, the overburden produces confining pressures of 2–23 GPa, yielding the strength of the fault of about the same order of magnitude. Earthquake occurrence requires a shear stress exceeding that strength, but it is unlikely that such a high shear stress is actually produced at such depths in the slab. Even if some process in the slab can produce such a high shear stress, it cannot be maintained because ductile flow will occur. Therefore, some special mechanism is necessary to generate earthquakes at such great depths.

One possible mechanism that decreases the strength of the fault for intermediate-depth earthquakes is dehydration embrittlement (Raleigh and Paterson 1965; Green and Houston 1995). Before its subduction, the oceanic plate contains a considerable amount of water as hydrated minerals. The temperature and pressure of the plate rise as it subducts and decompose the hydrated minerals and discharge water. The water reduces the effective normal stress and enables brittle fracture. Dehydration embrittlement also seems to explain the formation of the double seismic zone as discussed below.

If dehydration embrittlement is the mechanism generating intermediate-depth intraslab earthquakes, intraslab earthquakes should not be expected to occur everywhere in the slab, but only where hydrated minerals exist, particularly at facies boundaries where the water content changes. It has become increasingly clear that intermediate-depth intraslab earthquakes do seem to occur in such areas.

In order to examine whether or not dehydration embrittlement is the cause of intraslab earthquakes, we need to know the location within the slab where the hydrated minerals are decomposed and the water is discharged. Yamasaki and Seno (2003) determined the locations in the slab where the

dehydration decompositions of serpentinized slab mantle and metamorphosed slab crust occur based on experimentally obtained phase diagrams of rocks. They compared these locations with the spatial distribution of intraslab earthquakes. The result for the Pacific slab under northeast Japan is shown in Fig. 7. The estimated dehydration loci of the crust and mantle roughly correspond to the location of the upper and lower planes of the double seismic zone. They obtained the same results for the other five subduction zones. This explains well why double seismic zones are formed and why earthquakes in the lower plane have a planar distribution near the center of the slab mantle.

Recent studies in Japan have shown a more detailed spatial distribution of upper-plane earthquakes and revealed its clear relationship with the slab structure. This, as described below, can be explained well if the cause of intraslab earthquakes is dehydration embrittlement.

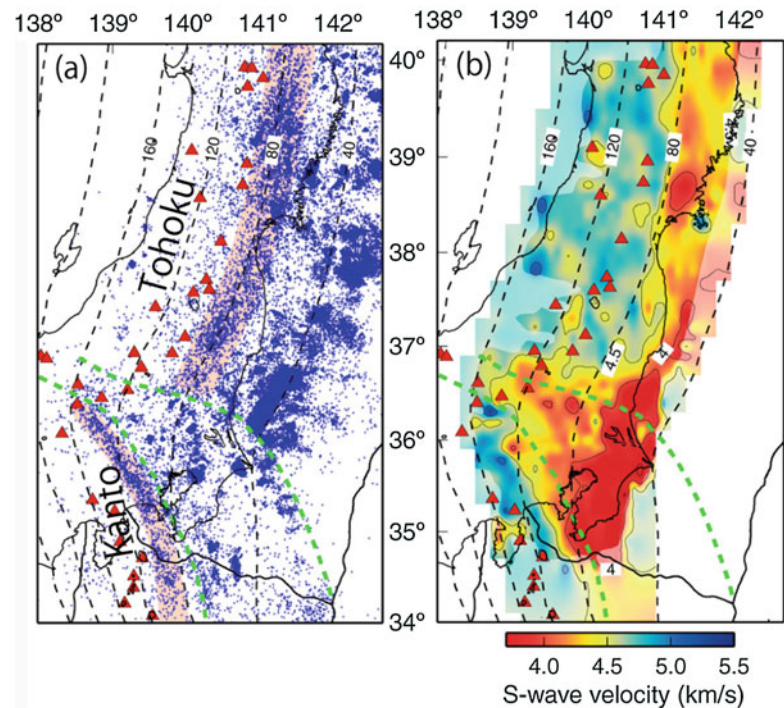
As indicated by the pale-pinked zones in Fig. 8a, earthquakes in the upper plane of the double seismic zone beneath Tohoku, northeast Japan form a remarkable belt-like seismic zone at depths between 70 and 90 km, parallel to iso-depth contours of the upper surface of the Pacific slab. If the cause of intraslab earthquakes is dehydration embrittlement, this concentration of seismicity is expected to correspond to the location in the slab crust where hydrated minerals are present, particularly around the facies boundary where dehydration occurs. Estimated location of the facies boundary based on the phase diagram of mid-oceanic ridge basalt and the geotherm showed that this expectation is almost true (Kita et al. 2006).

This belt-like seismicity indicated by the pale-pinked zones is not parallel to iso-depth contours, but is oblique to them and deepens locally in Kanto (Fig. 8a). This local deepening is thought to be caused by the shield effect of the Philippine Sea slab, which is located immediately above the Pacific slab. The two slabs are in direct contact under Kanto, and the contact zone between the two is indicated as an area enclosed by two broken lines in the figure. The downdip end of the contact zone is parallel to the belt-like seismicity beneath Kanto, which is oblique to the iso-depth contours of the upper plate interface. Such a clear correspondence indicates that under Kanto, the Philippine Sea slab prevents the Pacific slab from being heated up by the mantle wedge and delays the phase transformation in the slab crust, resulting in the locally deepened belt-like seismicity in that area.

Phase transformation of the slab crust causes an increase in seismic wave velocity. Therefore, we can verify from seismic tomography whether the phase transformation is the cause of the belt-like seismicity. Distribution of S-wave velocity in the slab crust (Fig. 8b) shows that the low-velocity layer persists down to about 80 km depth in Tohoku; deeper than that, S-wave velocity becomes high. On the other hand, in Kanto, the depth range of the low-velocity layer deepens locally in the zone of contact with the Philippine Sea slab. This means

Seismicity, Subduction Zone,

Fig. 8 (a) Earthquake epicenter distribution and (b) S-wave velocity distribution in the subducting Pacific slab crust beneath northeast Japan (Nakajima et al. 2009, copyright by the American Geophysical Union). Pale-pinked zones in (a) show the belt-like seismic zone. S-wave velocity perturbations on a curved plane 5 km below the upper surface of the slab are shown in (b) by the color scale at the bottom. Iso-depth contours of the upper surface of the slab are shown by broken curves. The zone of contact with the overlying Philippine Sea slab is enclosed by green broken lines



the distribution of S-wave velocity also confirms the prediction that the phase transformation is delayed.

The above results support the hypothesis that dehydration-derived aqueous fluids cause intermediate-depth intraslab earthquakes. It is also possible that the initiation of earthquake rupture is caused by local excess pore fluid pressure, and subsequent ruptures propagate through shear heating instability (Hasegawa and Nakajima 2017). In either case, slab-derived aqueous fluids play an important role in generating intermediate-depth events. However, generation mechanism for lower-plane earthquakes is still debated. Whether hydrated minerals actually exist at depths where lower-plane earthquakes occur (the deeper part of the mantle that lies at a maximum of about 40 km below the upper slab boundary) is a point of contention. In this regard, further validation is necessary for the generation of lower-plane events.

Deep Earthquakes

As described in previous sections, seismic observations seem to support the dehydration embrittlement hypothesis for the generation of intermediate-depth earthquakes, at least those in the crust of the slab. Focal depth distribution of global seismicity (Fig. 4) shows that the frequency of earthquakes, after a nearly monotonic decrease from shallow depths, recovers at a depth of about 300 km and gradually increases until the local maximum at 500–600 km. This may suggest that a different mechanism is working for the generation of deep earthquakes.

Several hypotheses have been proposed for the generation of deep earthquakes. There exists the difficulty in generating deep earthquakes as already described in section “Intermediate-Depth Earthquakes”. Three mechanisms out of those proposed so far seem to overcome the difficulty. One is the shear heating instability of creep (Griggs and Handin 1960; Ogawa 1987). Plastic creep of minerals produces heat, while high temperatures promote creep. Therefore, positive feedback promoting creep would arise if the creep of a thin layer occurs at a higher rate than the heat produced by the creep that diffuses into the surrounding area. Thus, the creep accelerates, and finally the thin layer of material begins to melt, producing dynamic slip on the layer, that is, occurrence of an earthquake.

The second hypothesis is the transformational faulting (Kirby 1987; Green and Burnley 1989). Olivine in the mantle of the slab transforms to β -spinel associated with the subduction of the slab. The phase transformation does not occur simultaneously but originates sporadically here and there, forming small lens-shaped spinel phases. The phase transformation proceeds as these spinel lenses expand. The spinel lenses will be formed in a direction perpendicular to the maximum compressive stress axis due to volume reduction occurring in the phase transformation. At the same time, compressive stress will arise at the tip of each spinel lens due to volume reduction as the phase transformation proceeds. This compressive stress promotes the phase transformation. In this way, the spinel lens expands as if it were a crack. Green and Burnley called this spinel lens anticrack. At some critical anticrack density, those anticracks that are

aligned on a plane directed obliquely to the maximum compressive axis at an acute angle are connected with each other and become unified, producing a dynamic slip on the plane, that is, the occurrence of an earthquake.

The last one is the dehydration embrittlement. Although this hypothesis seems to work for the generation of intermediate-depth earthquakes as already explained, some researchers argue that the dehydration reaction continues to occur in the mantle of the slab at greater depths and that dehydration embrittlement is responsible also for deep earthquakes (Omori et al. 2004).

It is not clear which of the three mechanisms is actually working, or if another completely different mechanism is responsible for generating deep earthquakes. More investigations are required for an exact understanding of the mechanism by which deep earthquakes are generated.

Summary

Recent investigations have significantly developed our understanding of generation mechanism for earthquakes in subduction zones. Observed heterogeneous interplate coupling along the megathrust is explained by frictional properties on the fault surface based on laboratory-derived fault constitutive laws. Frictionally unstable patches (asperities) are locked during the interseismic period and finally fail as interplate earthquakes. Surrounding conditionally stable areas typically slide stably, but may slide dynamically if loaded by the seismic slip of an adjacent asperity. Also found are systematically varying rupture properties with depth along the megathrust which are produced by heterogeneous spatial distribution of the frictional properties. Dehydration embrittlement model seems to work as the generation mechanism for intraslab intermediate-depth earthquakes, particularly for earthquakes in the slab crust. Aqueous fluids expelled by the dehydration reaction in the subducted plate migrate upward, finally reaching the arc crust. Recent investigations have shown that the fluids thus supplied originally from the subducting plate play an important role in generating shallow inland intraplate earthquakes. Generation mechanism for deep earthquakes is still an open question. More systematic and intensified studies are expected to resolve it.

Cross-References

- [Earthquake, Focal Mechanism](#)
- [Earthquakes and Crustal Deformation](#)
- [Great Earthquakes](#)
- [Lithosphere, Mechanical Properties](#)
- [Plate Motions in Time: Inferences on Driving and Resisting Forces](#)
- [Seismic Tomography](#)

- [Seismic Velocity and Temperature Relationships](#)
- [Seismicity, Intraplate](#)
- [Seismological Networks](#)
- [Seismology, Global Earthquake Model](#)
- [Slow Earthquake](#)
- [Subduction Zones](#)

Bibliography

- Bilek SL, Lay T (2002) Tsunami earthquakes possibly widespread manifestations of frictional conditional stability. *Geophys Res Lett* 29(14):1673. <https://doi.org/10.1029/2002GL015215>
- Bina CR, Stein S, Marton FC, Van Ark EM (2001) Implications of slab mineralogy for subduction dynamics. *Phys Earth Planet Inter* 127:51–66
- Chen P-F, Bina CR, Okal EA (2004) A global survey of stress orientations in subducting slabs as revealed by intermediate-depth earthquakes. *Geophys J Int* 159:721–733
- Dixon TH, Moore JC (2007) The Seismogenic zone of subduction thrust faults, margins theoretical institute and experimental earth science series. Columbia University Press, New York
- Engdahl ER, Scholz CH (1977) A double benioff zone beneath the central aleutians: an unbending of the lithosphere. *Geophys Res Lett* 4:473–476
- Engdahl ER, van der Hilst R, Buland R (1998) Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bull Seismol Soc Am* 88:722–743
- Frohlich C (2006) Deep earthquakes. Cambridge University Press, New York
- Garage S, Umino N, Hasegawa A, Kirby S (2009) Offshore double planed shallow seismic zone in the NE Japan forearc region revealed by sP depth phases recorded by regional networks. *Geophys J Int* 178:195–214
- Green HW, Burnley PC (1989) A new self-organizing mechanism for deep-focus earthquakes. *Nature* 341:733–737. <https://doi.org/10.1038/341733a0>
- Green HW, Houston H (1995) The mechanics of deep earthquakes. *Annu Rev Earth Planet Sci* 23:169–213
- Griggs DT, Handin J (1960) Observations on fracture and a hypothesis of earthquakes. In: Griggs DT, Handin J (eds) *Rock deformation: geological society of America. Memoirs*, 79. Geological Society of America, New York, pp 347–373
- Hasegawa A, Nakajima J (2017) Seismic imaging of slab metamorphism and genesis of intermediate-depth intraslab earthquakes. *Prog Earth Planet Sci*. <https://doi.org/10.1186/s40645-017-0126-9>
- Hasegawa A, Umino N, Takagi A (1978) Double-planed structure of the deep seismic zone in the northeastern Japan arc. *Tectonophysics* 47:43–58
- Hasegawa A, Nakajima J, Umino N, Miura S (2005) Deep structure of the northeastern Japan arc and its implications for crustal deformation and shallow seismic activity. *Tectonophysics* 403(1–4):59–75
- Hyndman RD, Wang K et al (1995) Thermal constraints on the seismogenic portion of the southwestern Japan subduction thrust. *J Geophys Res* 100:15373–15392
- Iio Y, Sagiya T, Kobayashi Y (2004) Origin of the concentrated deformation zone in the Japanese Islands and stress accumulation process of intraplate earthquakes. *Earth Planets Space* 56:831–842
- Isacks BL, Molnar P (1971) Distribution of stresses in the descending lithosphere from a global survey of focal mechanism solutions of mantle earthquakes. *Rev Geophys Space Phys* 9:103–174
- Kirby S (1987) Localized polymorphic phase transformations in high-pressure faults and applications to the physical mechanism of deep earthquakes. *J Geophys Res* 92(B13):789–800. <https://doi.org/10.1029/JB092iB13p13789>
- Kirby S (1995) Intraslab earthquakes and phase changes in subducting lithosphere. *Rev Geophys* 33(S1):287–297

- Kita S, Okada T, Nakajima J, Matsuzawa T, Hasegawa A (2006) Existence of a seismic belt in the upper plane of the double seismic zone extending in the along-arc direction at depths of 70–100 km beneath NE Japan. *Geophys Res Lett* 33:L24310. <https://doi.org/10.1029/2006GL028239>
- Lay T, Kanamori H, Ammon CJ, Koper KD, Hutko AR, Ye L, Yue H, Rushing TM (2012) Depth-varying rupture properties of subduction zone megathrust faults. *J Geophys Res* 117:B04311. <https://doi.org/10.1029/2011JB009133>
- Miura M, Sato T, Hasegawa A, Suwa Y, Tachibana K, Yui S (2004) Strain concentration zone along the volcanic front derived by GPS observations in NE Japan arc. *Earth Planets Space* 56:1347–1355
- Nakajima J, Hasegawa A (2007) Deep crustal structure along the Niigata-Kobe tectonic zone, Japan: its origin and segmentation. *Earth Planets Space* 59:e5–e8
- Nakajima J, Matsuzawa T, Hasegawa A, Zhao D (2001) Three-dimensional structure of Vp, Vs, and Vp/Vs beneath northeastern Japan: implications for arc magmatism and fluids. *J Geophys Res* 106:843–857
- Nakajima J, Tsuji Y, Hasegawa A (2009) Seismic evidence for thermally-controlled dehydration reaction in subducting oceanic crust. *Geophys Res Lett* 36:L03303. <https://doi.org/10.1029/2008GL036865>
- Ogawa M (1987) Shear instability in a viscoelastic material as the cause of deep focus earthquakes. *J Geophys Res* 92:13801–13810
- Oleskevich DA, Hyndman RD, Wang K (1999) The updip and downdip limits to great subduction earthquakes: thermal and structural models of Cascadia, South Alaska, SW Japan, and Chile. *J Geophys Res* 104:14965–14991
- Omori S, Komabayashi T, Maruyama S (2004) Dehydration and earthquakes in the subducting slab: empirical link in intermediate and deep seismic zones. *Phys Earth Planet Inter* 146:297–311
- Raleigh CB, Paterson MS (1965) Experimental deformation of serpentinite and its tectonic implications. *J Geophys Res* 70:3965–3985
- Sagiya T, Miyazaki S, Tada T (2000) Continuous GPS array and present-day crustal deformation of Japan. *Pure Appl Geophys* 157:2303–2322
- Scholz CH (1998) Earthquakes and friction laws. *Nature* 391:37–42
- Seno T, Yamanaka Y (1996) Double seismic zones, compressional deep trench–outer rise events and super plumes. In: Bebout GE et al (eds) *Subduction: top to bottom*, Geophysical monograph series, vol 96. AGU, Washington, DC, pp 347–355
- Sibson RH (2009) Rupturing in overpressured crust during compressional inversion – the case from NE Honshu, Japan. *Tectonophysics* 473:404–416
- Yamasaki T, Seno T (2003) Double seismic zone and dehydration embrittlement of the subducting slab. *J Geophys Res* 108(B4): ESE9–ESE1. <https://doi.org/10.1029/2002JB001918>

Seismogram Interpretation

Ota Kulhanek¹ and Leif Persson²

¹Department of Earth Sciences, Section of Seismology, Uppsala University, Uppsala, Sweden

²Department of Underwater Research, FOI, Stockholm, Sweden

Synonyms

Anatomy of seismograms; Reading seismic records; Structure of seismograms

Definition

Seismogram interpretation. Identification of various phases appearing on seismograms (measurable changes in frequency and/or amplitude), including the recognition of the waves with respect to possible travel paths through the Earth.

Introduction

A seismogram is the recording of Earth motion (displacement, velocity, or acceleration) as a function of time. Its appearance reflects the combined effects of the source, the propagation path, the characteristics of the recording instrument (see ► “[Seismic Instrumentation](#)”), and the ambient noise due to the specific conditions at the particular recording site. To decipher the often complicated form of seismogram traces requires an extensive knowledge of source physics, structure of the Earth, wave propagation, and earthquake geography. In the early years of observational seismology, say, prior to 1960, the analyst was essentially reading short- and long-period analog seismograms made on smoked, photographic, or heat-sensitive paper or by ink-pen recorders. Approximately, in the middle of the 1980s, the situation changed dramatically. The advent and extensive deployment of broadband digital seismometers with large dynamic ranges complemented with an access to fast, powerful computers opened for seismogram interpreters new possibilities never expected before. Today, the interpreter performs easily a set of operations that significantly improve and accelerate the process of phase identification on digital seismic records. These include, for example, rotation of seismogram components and particle motion techniques. Digital multichannel data serve as input for frequency-wave number analysis and polarization filtering. Synthetic seismograms model observations and phase onsets revealed on seismograms can be associated within narrow time windows with theoretically predicted onset arrival times.

Velocity models for radially stratified Earth have been developed to provide travel times for major seismic waves for the purpose of earthquake location and phase identification. In 1987, IASPEI initiated an effort to update the JB Tables (Jeffreys and Bullen 1940), which have served for 50 years, by constructing new global travel-time tables. With access to an extensive ISC data set (1964–1987), this effort resulted in two velocity models: *iasp91* (Kennett and Engdahl 1991) and *sp6* (Morelli and Dziewonski 1993) deduced from empirical smoothed travel-time curves of the main phases. The two models reveal only small differences in predicted travel times. With respect to the older JB Tables, the most significant differences were found in the upper mantle and core (see ► “[Earth’s Structure, Global](#)”). New, *iasp91*, travel-time tables, which include also the ancillary phases, were calculated from the velocity model (Kennett 1991). For teleseismic *P* waves, the new tables are in average 1.8–1.9 s

faster than the JB Tables. The *iasp91* model was slightly modified in 1995, and a new model *ak135* was proposed (Kennett et al. 1995). The model is suited for predicting the arrival times of a wide variety of seismic phases. It is routinely used by major agencies (ISC, NEIC) in event locations and phase identifications.

At first sight, it seems that there is a contradiction between the duration of the rupture at the source, which takes between a fraction of a second and a few minutes (depending on magnitude), and the length of the observed seismogram, which for large and distant shocks can extend over several hours. The length of the seismogram depends primarily on various wave propagation effects such as reflection, refraction, mode conversion, dispersion, etc. and has little to do with the duration of the quake. A seismogram, especially from a distant earthquake, will often show a number of more or less distinct waves, commonly called phases, distributed in time, which have traveled along different propagation paths and which were subjected to different mode conversions. Numerous manuals for seismogram interpretation have been available since the early 1950s. Some of them have been of local/regional importance, but some have gained worldwide recognition. The latter include works of Neumann (1951), Simon (1968), Willmore (1979), Payo (1986), and Kulhanek (1990). The most recent is the manual edited by Bormann (2012).

Essentially, there are two types of seismic waves, *P* and *S* body waves (see ► “[Body Waves](#)”) and Love (*LQ*) and Rayleigh (*LR*) surface waves (see ► “[Surface Waves](#)”). The most important characteristics of body and surface waves, which are invaluable in any seismogram interpretation, are the following:

- Different waves travel with different velocities. At any epicentral distance, *P* is recorded first, followed by *S*, *LQ*, and *LR*.

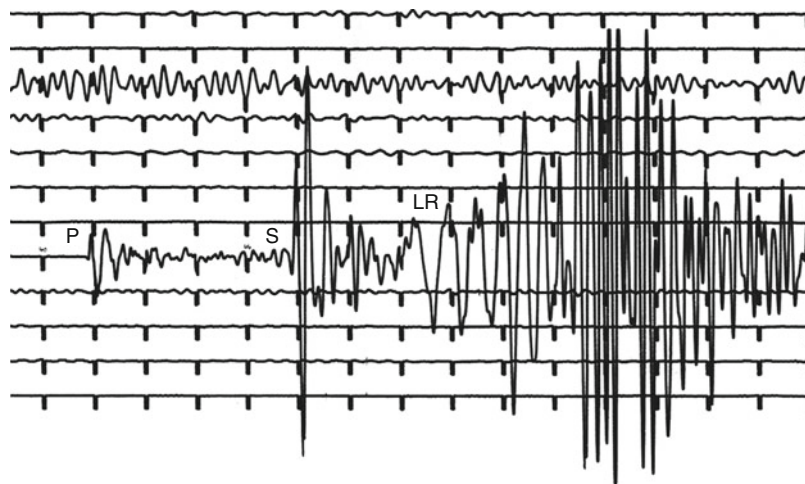
- Body waves produce records in the high-frequency range, from 0.1 to 10 Hz, while surface waves occupy low frequencies of about 0.1 Hz and lower. Exceptions will be described later.
- Due to different polarization, *P* waves are usually better displayed on vertical-component seismograms, while *S* waves are often best on horizontal-component records. Fluids do not sustain shear strain, and therefore *S* waves do not travel through liquid parts of Earth’s interior. *LQ*-waves are exhibited only on horizontal-component records, while *LR*-waves are received by both the vertical and horizontal seismometers. For shallow shocks, surface waves usually dominate the seismogram.
- Surface waves exhibit an important property called velocity dispersion (see ► “[Surface Waves](#)”). In practice, this means that the long-period surface waves approach the station first and are recorded ahead of the “slower” shorter waves. Ideally, the seismogram of *LQ*- or *LR*-waves will start with long-period motion, which gradually, as time increases, will turn into shorter and shorter periods.

Some of these features are demonstrated on the seismogram displayed in Fig. 1. There is a sharp *P* onset followed after approximately 4 min by a clear *S*. About 2 min after *S*, we observe the arriving *LR* (vertical-component seismogram). The clearly dispersive character of the recorded *LR* in this case manifests a continental propagation path signature. The large *LR* amplitudes indicate a shallow-focus depth. Except for the first *P* onset, all later arrivals are contaminated by codas of preceding phases so that, on the records, there is virtually no interval of quiescence between individual phases.

In the following description, we introduce several categories of seismic events. The classification is based on the distance between the event and the recording site, that is, on the epicentral distance, which in turn governs propagation paths along which seismic waves travel through the Earth’s

Seismogram Interpretation,

Fig. 1 Long-period, vertical-component, ink-pen seismogram of the Greek earthquake of February 25, 1981 ($M = 6.3$, $h = 33$ km) made at UPP (Uppsala), Sweden, at an epicentral distance of 22° . There is 1 min between successive time marks (small downward offsets)



interior. Events recorded at distances shorter than about 1° will be called local events. Regional earthquakes are shocks recorded at distances between approximately 1° and 10° . For epicentral distances 0 – 10° , seismic waves propagate through the crust and/or along Moho. Accordingly, we call these waves crustal waves. At recording distances 10 – 103° , the waves travel mainly through the mantle. For distances larger than about 103° , seismic waves enter the core or are diffracted along the core-mantle boundary (so-called core waves). Events observed at distances larger than 20° (or 30°) are collectively called teleseisms.

Crustal Waves: Recording Distances 0 – 10°

In a one-layer continental crust, a source located within the crust radiates both P and S waves, which will be recorded along the Earth's surface. These waves are encoded as Pg and Sg . The subscript g refers to granitic layer. In continental earthquakes it has been often observed that Sg has the largest amplitude. The nomenclature of seismic phases used in the present article is that of the IASPEI Standard (Storchak et al. 2003). A reflected wave is also possible from the outer side of the Moho, and the corresponding reflected P and S waves are labeled as PmP and SmS , respectively. At larger distances (beyond the critical distance), we shall record so-called head waves, that is, P and S waves refracted along the Moho discontinuity and labeled Pn and Sn . Head waves propagate with velocities of the uppermost mantle. Note that Pg and Sg exist for all epicentral distances from $\Delta = 0$ and outwards, while Pn and Sn cannot be observed at distances shorter than the critical distance, which for a continental crust is about 100 km. Waves reflected from Moho are rather scarce and difficult to identify. The best chance to discern PmP or SmS on records is at very short epicentral distances where the contamination by Pg and Pn (or Sg and Sn) is not severe.

At epicentral distances shorter than about 10° , records are complicated because of lateral heterogeneities in the crust. In a two-layer crustal model, which includes the Conrad discontinuity, we record new phases. For epicentral distances larger than about 100 km, these are the refracted P and S , traveling along the Conrad discontinuity. An asterisk in the superscript position, P^* and S^* , indicates this phase. An alternative code is Pb and Sb . The subscript b refers to basaltic layer.

Let us assume a continental crust with propagation velocities for Pg , P^* , and Pn of 6.0, 6.6, and 8.0 km/s, respectively. At short distances, less than about 150 km or so, the first seismic wave arriving at the recording station is Pg . For distances larger than critical but less than about 150 km/s, Pg is followed by P^* and Pn , in this order. P^* and Pn travel with velocities significantly higher than Pg . Therefore, at distances larger than approximately 150–200 km, crustal waves change their order of arrival. For distances larger than

about 200 km, the first arriving phase is Pn , next arrives P^* , and then Pg . This is true only for continental travel paths. Seismograms from earthquakes beneath the sea bottom, made at islands or coastal stations, will not show Pg or Sg since there is no granitic layer in the oceanic crust. Similarly, quakes originating in the lower crust, beneath the Conrad discontinuity, do not produce Pg or Sg phases. Hence, first arrivals on records from these events will be Pn or P^* . With a certain time delay following the P phases, proportional to the epicentral distance, the crustal S waves arrive in the same order as P waves. For local events the order of S onsets will be Sg , S^* , and Sn , while for events from distances larger than about 300 km, we observe first Sn followed by S^* and Sg . As an example, a record from a regional earthquake is displayed in Fig. 2. The first discernible phase is Pn , weakly recorded on the vertical and N-S channels. It is followed after about 6 s by strong Pg and after another 39 s by clear Sg , visible on all three components. Sn expected to precede Sg by 9 s cannot be identified on the seismograms.

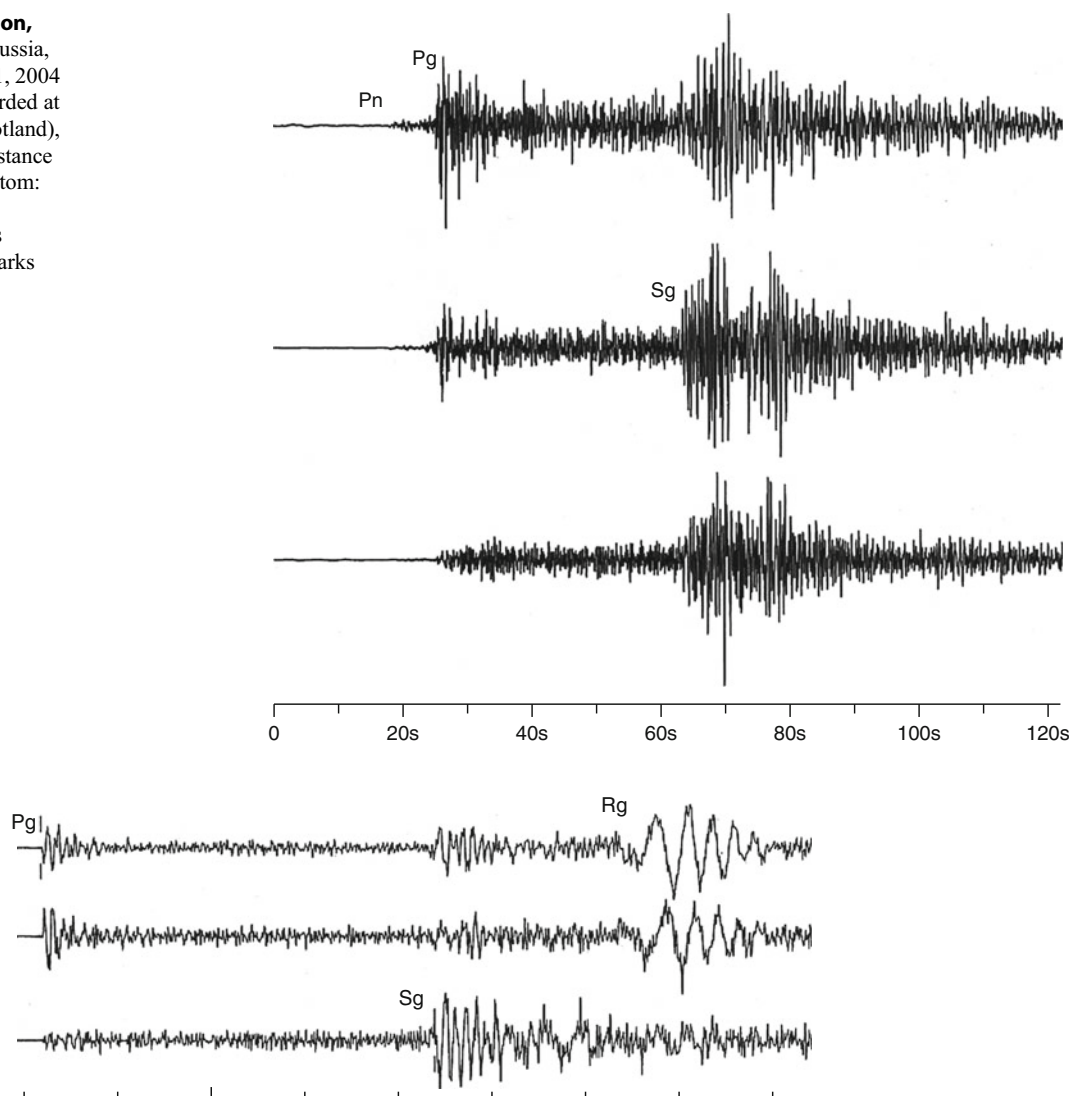
Short-period S waves, multiply reflected between the free surface and Moho or other crustal velocity discontinuities, interfere with each other and give rise to a wave group labeled Lg . There is no clear distinction between Sg and Lg . Lg waves propagate as guided waves, supercritically incident on the Moho and multiply reflected within the (continental) crust with a velocity of about 3.5 km/s. They may dominate the seismograms made usually at a distance of 5° and larger. Lg are best seen on horizontal-component records. Some rare cases are known where Lg propagated over distances of several thousand kilometers (Kulhanek 1990).

Near-surface local and regional events (earthquakes, industrial explosions, rock bursts, etc.) also generate short-period surface wave of Rayleigh type, labeled Rg . The presence of Rg in the seismogram is a reliable indicator of a very shallow event with focal depth of the order of one or a few kilometers. On the other hand, if Rg waves are absent (epicentral distance of several hundred kilometers or less), we are concerned with a deeper natural event, that is, with a crustal quake at a depth most likely between 5 and 25 km, since all types of man-made events can be excluded. Rg waves travel with a velocity of 3 km/s or slightly higher and are usually well recorded up to distances of about 600 km (Båth 1983). These waves are best, but not exclusively, seen on vertical-component records. An example of a recorded Rg wave is displayed in Fig. 3. The Rg phase is well developed (large amplitudes, clear dispersion) indicating a focal depth of 2 km or less.

Island and coastal seismographic stations frequently record so-called T waves (tertiary waves) arriving after P and S (see ► “[T Waves](#)”). They are characterized by propagation within the oceans as ordinary sound waves emitted by earthquakes near the sea bottom or by submarine volcanic eruptions. The propagation of T waves is very efficient, and

Seismogram Interpretation,

Fig. 2 The Kaliningrad, Russia, earthquake of September 21, 2004 ($M = 5.0$, $h = 15$ km) recorded at broadband station GOT (Gotland), Sweden, at an epicentral distance of 335 km. From top to bottom: vertical, N–S and E–W components. There are 10 s between successive time marks



Seismogram Interpretation, Fig. 3 Weak, near-surface event ($M \sim 2$) in southern Sweden, recorded at broadband station NYN (Nynäshamn) at an epicentral distance of 36 km. From top to bottom: vertical, N–S and E–W components. There are strong onsets of Pg (vertical, N–S) and Sg

(E–W) and a clear Rg phase with distinct dispersion, about 2 s after Sg , best developed on the vertical and N–S components. There is 1 s between successive time marks at the bottom. (Note: epicenter almost due north of NYN)

observations at distances as large as about 80° have been reported. Observations of T phases have proved useful in discriminating between underground nuclear explosions, detonated beneath oceanic islands, and tectonic earthquakes (Adams 1979; Okal 2001). On records, T phases often exhibit rather monochromatic oscillations (periods usually less than 1 s) with a gradual increase and decrease of amplitudes of total duration up to several minutes. There is no sharp onset in the T phase group, which creates difficulties when reading the T arrival times. In general, there is a great variety in the appearance of T phases due to the dependence upon the bottom topography in the vicinity of generation, oceanic stratification, and water-land conversion and transmission.

Appearance of seismograms made at local and/or regional distances varies from place to place mainly due to lateral

variations in crustal structure. It is, therefore, difficult to list generally valid clues for record interpretation. Nevertheless, following principles may guide the analyst to read seismograms and evaluate software solutions of local and regional earthquakes.

- Predominant periods of recorded crustal waves Pg , P^* , Pn , Sg , S^* , Sn , etc. are normally less than 1 s. Rg periods are usually not longer than several seconds.
- Often, Sg has the largest amplitude (when large short-period Rg is missing), best on horizontal-component records. At larger distances, Lg may dominate the records.
- For epicentral distances less than about 200 km (depending upon the crustal structure and focal depth), the first arriving phase is Pg . For larger distances, Pn arrives first.

- Near-surface events, from distances less than about 600 km, often generate short-period Rg with clear dispersion, best seen on vertical channels.
- Local and regional events of low or moderate magnitude are characterized by short total record duration, usually not exceeding several minutes.
- Island and coastal seismographic stations frequently record various types of T phases.

Body Waves Traversing the Mantle: Recording Distances 10–103°

Mantle, in contrast to the overlying crust, may, in the first approximation, be considered as laterally homogeneous. At distances around 10°, Pn and Sn become difficult to identify in the records. Instead, P and S phases show on seismograms. Some workers consider the distance range between 10° and 103° as ideal to record not only the direct P and S waves but also the whole family of reflected and converted waves. Travel paths of these waves are dominated by the mantle, and corresponding seismograms are relatively simple. P is usually stronger on the vertical component, while S is more clearly seen on horizontal components. S often exhibits wave trains with longer periods when compared with corresponding P . Large-amplitude S waves are often observed at distances up to about 100°.

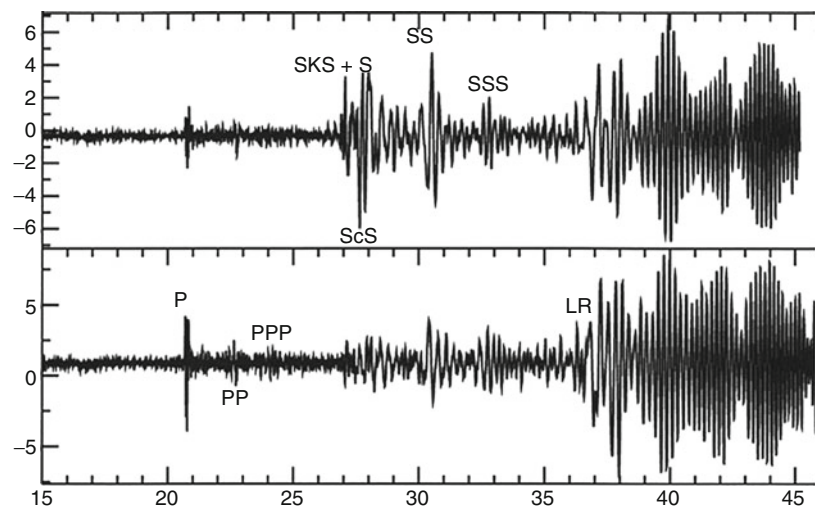
A direct P reflected from the free surface once or twice is called PP or PPP , respectively. In the same way, we have also SS or SSS . For distances larger than about 40°, the free-surface reflected phases become very distinct. At distances around 100° and larger, PP and SS often belong to the largest recorded body waves. Converted waves like PS and SP appear only at distances larger than 40°. Examples of seismograms with free-surface reflections are displayed in Fig. 4.

Core-reflected phases, that is, waves reflected back into the mantle from the outer core-mantle boundary, are labeled PcP and ScS . Including mode-converted waves, we have also PcS and ScP . Because core-reflected phases emerge steeply, ScP is usually stronger on vertical-component records than PcS . Large core-reflected phases are usually recorded at shorter epicentral distances, say at 40° or less. At distances around 39°, ScP and PcS (surface foci) are often contaminated with the arrival of direct S , and the phase separation is difficult. When the ray path of PcP grazes the outer core boundary, the combination of direct P and PcP is called P diffracted and labeled $Pdif$ (older $Pdiff$).

Waves ascending from the focus to the free surface (in continents) or to the sea bottom (under oceans), where they are reflected back into the mantle, are commonly called depth phases. We can list the four possibilities of reflections near the epicenter, which are pP , sP , pS , and sS . The first case, for example, denotes the wave that traveled upward from the focus as P (short leg) and had been reflected back off the free surface again as P (long leg). Depth phases, primarily pP , are the most important phases routinely used in focal-depth estimations. The deeper the focus, the later is the pP phase in relation to direct P . Thus, accurately measured arrival time differences, $pP-P$, are reliable indicators of the depth of the focus. Most teleseismic shallow events (depth of focus around 33 km) will reveal a pP 9–11 s after P (see, e.g., Herrin et al. 1968). In case of a deeper focus, it is sometimes possible to recognize several different reflections/conversions from the free surface, for example, pPP , sPP , pPS , sPS , pSP , etc. Interpretation of depth phases must be done with utmost care since, for example, pP from a deep event can easily be erroneously interpreted as P when the first arrival (P) is weak. Similar difficulty emerges for multiple shocks. Depending on focal orientation and other factors, sP may be stronger than pP and may be mistaken for it. Depth phases are sometimes

Seismogram Interpretation,

Fig. 4 The Costa Rica earthquake of August 20, 1999, ($M = 6.9$, $h = 40$ km). Exhibited seismograms are long period, E–W (upper trace) and vertical-component (lower trace), records made at KONO (Kongsberg), Norway. The epicentral distance is 84°. There are 100 s between two successive time marks at the bottom of the figure. For this distance range, SKS arrives ahead of S , but the separation is difficult. Both displayed channels are dominated by a fundamental LR wave, developed along a mainly oceanic propagation path



stronger than the direct P and may be the first readable phase. We usually require several records made at different epicentral distances to make a reliable identification of the depth phase. We can also mention less common phases like pmP and pwP (first identified by Mendiguren 1971), corresponding to Moho underside reflection and to water surface reflection, respectively (Schenk et al. 1989). A thin layer of low-velocity sediments at the sea bottom may have small impedance contrast. Hence, short-period records may show large pwP when compared with pP reflecting at the sea bottom (Engdahl et al. 1998). pwP can easily be misinterpreted as pP leading to an overestimation of the focal depth.

Bolt (1982) introduced symbols like $P400P$ and $P650P$ to indicate seismic waves reflected at the underside of a secondary discontinuity, in the upper mantle, at a depth of 400 and 650 km, respectively. These phases arrive at the recording station ahead of the expected (calculated) arrival time for the main PP phase and are interpreted as reflections from upper mantle discontinuities, that is, as PdP . When the early arrival time cannot be explained in terms of known discontinuities as PdP , we call these onsets early PP or precursors to PP .

Body Waves Traversing the Core: Recording Distances 103° and Beyond

Due to the diffraction around the core-mantle boundary (see ► “Earth’s Structure, Global”), amplitudes of direct P waves decay dramatically at distances larger than 100° . The short-period P reappears consistently on records first at distances of about 140° . The distance range $103^\circ < \Delta < 140^\circ$ is called the shadow zone. On seismograms, $Pdif$ waves usually show small amplitudes, emergent or gradual onsets, and energy shifts to longer periods. Long-period $Pdif$ are sometimes observed out to distances of 160° or more. S waves are affected at the core-mantle boundary in a similar way.

Waves that traveled through the outer core but do not enter the inner core are labeled PKP (also P'), PKS , SKS (also S'), and SKP . The SKP phase is stronger on vertical components than PKS . These phases have a caustic near 130° , and close to this distance are often the only phase recorded on short-period seismograms. At epicentral distances close to 144° , PKP shows a distinct concentration of energy, that is, large amplitudes on records. The phenomenon may be viewed in terms of two PKP travel-time branches denoted $PKPbc$ and $PKPab$ for the first and the second arrival, respectively. At 144° , the waves from the two branches coincide; the waves reinforce one another, which results in energy concentration near that distance.

P waves that traverse the inner core are denoted by I , giving rise to phases $PKIKP$, $PKIKS$, $SKIKS$, and $SKIKP$, although these are often still simply referred to as PKP , PKS , etc. Phases, with an S leg in the inner core, would

include the letter J , such as $PKJKP$, but these have been difficult to identify on records. Cao et al. (2005) made use of high-quality, broadband seismograms from the Gräfenberg array station, Germany, and identified a phase with arrival time and slowness consistent with predictions of $PKJKP$.

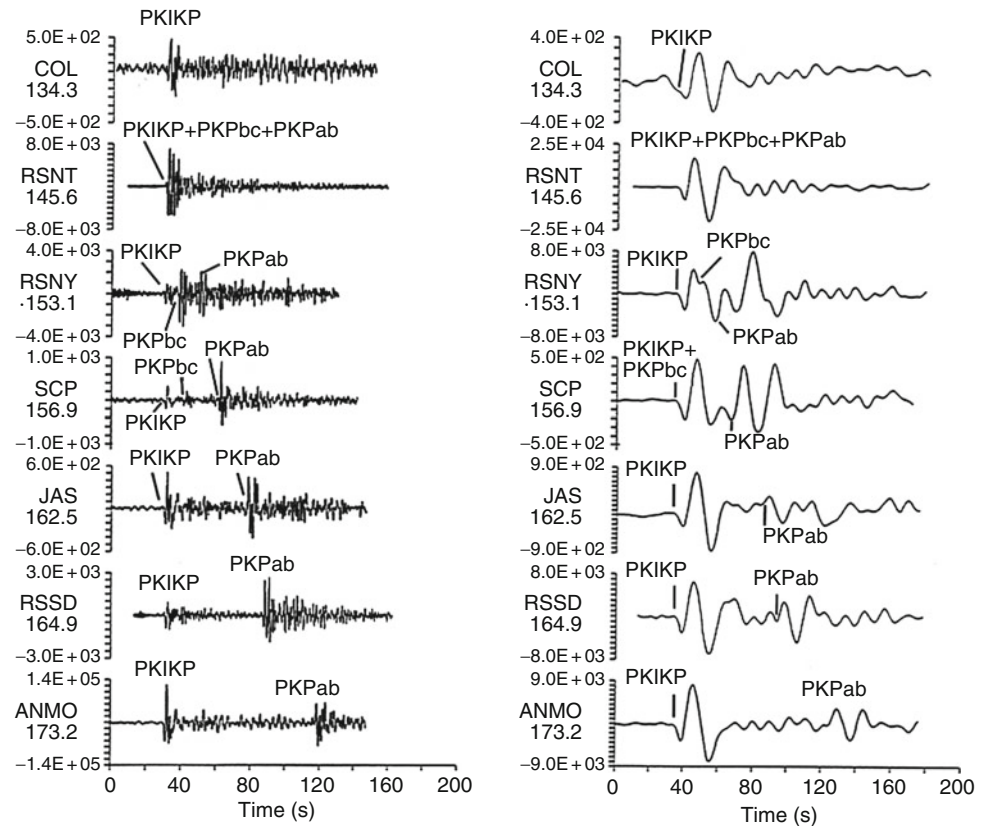
If the studied event is weak, then usually no $Pdif$ is observed in the entire distance interval $\Delta > 103^\circ$, and the first arrival seen on the record will be that of PKP . At epicentral distances 105° – 120° , $PKIKP$ usually provides the first onset discernible on the seismogram. In the region of the caustic, that is, around 144° , the wave train of recorded core phases becomes particularly complicated. It is first at distances beyond the caustic point where observed onsets may be separated into individual PKP branches. The energy distribution changes with the increasing distance. $PKPbc$ is the dominant branch just beyond the caustic, up to about 153° . In records of weaker events ($\Delta = 144^\circ$ – 153°), $PKPbc$ is often the first visible onset since $PKIKP$ (*alt.* $PKPdf$), theoretically preceding $PKPbc$, is too weak to be observed. As the distance increases, $PKPbc$ becomes weaker and vanishes from records at distances of about 160° and larger. For distances beyond, say, 157° , $PKPab$ usually dominates the seismogram. $PKIKP$ in the distance range from about 125° to the caustic is often preceded by early arrivals or precursors, which can arrive many seconds ahead of the main phase. These are often explained by scattering phenomena at or near the core-mantle boundary. Subscripts *ab*, *bc*, and *df* are used in agreement with travel-time charts of Jeffreys and Bullen (1940). Recorded core phases made at a suite of seismographic stations at a distance range from 134° to 173° are shown in Fig. 5.

We may form new symbols for the whole family of waves propagating through the outer core. For example, $PKKP$ is a P wave reflected from the inside of the core-mantle boundary, often very pronounced on records made at distances between 60° and 80° . However, the striking onset may easily be misinterpreted as a first P arrival of another event. P waves trapped inside the Earth’s liquid core and with multiple K legs are called $PNKP$, where $N - 1$ is the number of reflections. Multiple reflections within Earth’s outer core were first observed by Engdahl (1968). He used records from two deep earthquakes and an event in Novaya Zemlya ($h = 0$) and identified recorded seismic waves reflected as many as four times (i.e., $P5KP$) within the Earth’s outer core. Cases like $P4KP$ and $P7KP$ have been reported by Bolt (1982). Waves that traverse the Earth’s interior and are reflected at the outside or inside of the inner core are labeled $PKiKP$ or $PKIIKP$, respectively.

The best chance to observe $PKPPKP$, or for short $P'P'$, is around distances $2 \times 144^\circ = 288^\circ$ or 72° if we take the shortest distance from source to station. $P'P'$ is often well recorded, arriving about 30 min after P , when most of the coda amplitudes of preceding phases have already become faint.

Seismogram Interpretation,

Fig. 5 Seismograms from a Mid-Indian-Rise earthquake of May 16, 1985 ($M = 6.0$, $h = 10$ km), made at a suite of seismographic stations that sample the epicentral distance between 134.3° and 173.2° . Short-period (*left*) and long-period (*right*), vertical-component records are exhibited. Station codes, epicentral distances, and amplitude scaling are given to the left of each trace (G. Choy, personal communication; reproduced from Kulhanek 1990, with permission from Elsevier, Science)



$P'P'$ may in some cases be wrongly interpreted as a new P or PKP . 72° is also equivalent to $3 \times 144^\circ = 432^\circ$, so the phase $P'P'P'$ is also strong at this distance and may be observed, for strong earthquakes, about another 20 min after $P'P'$. In 1969, first observations of precursors to $P'P'$, sometimes called early $P'P'$, were made (Engdahl and Flinn 1969). These were interpreted as $P'dP'$ analogous to PdP , that is, as reflections of PKP at secondary velocity discontinuities within Earth's upper mantle. For example, $P'650P'$ passes through the core twice and on seismograms precedes $P'P'$ by about 2 min.

Similar to PKP and $PNKP$, there are SKS and $SNKS$ waves, respectively. First SKS waves are observed at distances between 60° and 70° , and the range of observations extends to distances of 180° or so. SKS exhibits a caustic point at a distance of about 80° so that the best region to study SKS waves is that between 70° and 90° . The phase identification has to be made with utmost care since SKS recorded at this distance range are often contaminated with direct S . To mistake S for SKS and vice versa will adversely affect the epicentral location. At about 82° (depending upon details in the structural model), SKS begins to arrive ahead of S (Fig. 4). For distances shorter than about 95° , SKS is usually smaller than S ; however, at distances beyond 95° , SKS amplitudes are quite large. S and SKS are best recorded on long-period horizontal-component seismograms. The period of SKS phases may

reach several tens of seconds. Occasionally, these body waves are also seen on short-period records, although the onset time of the later of the two phases is usually very emergent due to the contamination by the coda of the earlier phase.

Body Waves from Intermediate-Focus and Deep-Focus Earthquakes

Later studies confirmed conclusions from Wadati's pioneering work (Wadati 1927) that intermediate-focus and deep-focus shocks produce simpler seismograms with exceptionally well-recorded impulsive body waves while surface-wave amplitudes decrease as the quake becomes deeper. Strong depth phases, such as pP and sS , are also frequently very distinct on records from deep events. However, the duplication of principle phases by surface reflection (e.g., confusion between pP and PP) often complicates the seismogram interpretation. Weak precursors, pmP , (Moho underside reflections) to surface reflections, pP , are discussed in Schenk et al. (1989). They are best observed in long-period records and in continents often produce clear arrivals. In a more retrospective-type interpretation, the absence of aftershocks (see ► "Earthquake, Aftershocks") will support the classification of the shock as a deep-focus or intermediate-

focus event. Fig. 6 shows seismograms from an intermediate-focus earthquake made at an epicentral distance of 41° . Note the impulsive character of the P , pP , PcP , S , and ScS phases; virtually no surface waves have been recorded from this event.

Surface Waves

With exception of very short epicentral distances, long-period surface waves, LQ and LR dominate the seismograms of shallow events. Deep events of the same magnitude will generate abnormally small (insignificant) surface waves. This feature provides the interpreter with a viable tool to discriminate, at first glance, between shallow- and deep-focus earthquakes. (Note a paradox: underground nuclear explosions, being very shallow events, generate only small, sometimes invisible, surface waves.) The analyst has several clues to identify surface waves on the record and to distinguish LQ from LR . While Love waves are best displayed on horizontal-component seismograms, Rayleigh waves are best seen on vertical-component records. Both LQ and LR propagate slower than P or S , but since LQ propagate faster than LR , they are recorded ahead of LR . Due to dispersion, long-period surface waves advance on records toward the beginning of the wave train as it travels through the medium. An experienced analyst will distinguish between recorded surface waves that have traveled along pure oceanic or continental path. Dispersion of oceanic routes gives rise to long wave trains with rather slow and sometimes hardly visible period change over relatively long (10–20 min) record segments (see Fig. 4). In contrast, continental paths generate surface waves with characteristic fast period decrease with time, which is often easily recognized on records (see Fig. 1). Surface waves can also travel by different modes (overtones), which are often seen on records as high-frequency components superimposed on the surface-wave train. We talk then about

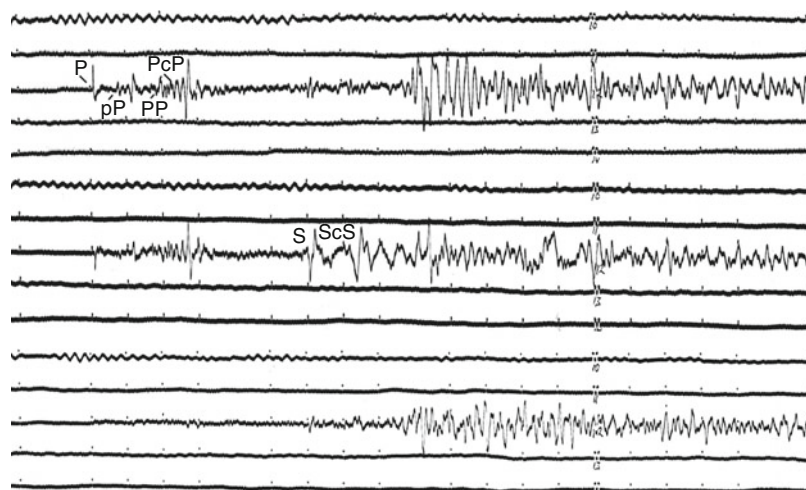
fundamental-mode and higher-mode surface waves (see ► “Surface Waves”). Higher modes are most frequently observed for waves traversing purely continental paths. Nevertheless, higher modes have in some cases been associated with oceanic paths. Surface-wave higher modes disappear when the waves cross the transition between continental and oceanic structures. Higher modes propagate faster than the fundamental mode and are, therefore, recorded ahead of LQ and LR .

Dispersion curves (see ► “Surface Waves”) show a rather complicated pattern with several local minima and maxima. Surface waves traveling with these minima or maxima group velocities are called Airy phases. On seismograms, an Airy phase is characterized by a constant-frequency compact wave train, often with a remarkable amplitude buildup of dispersed surface waves traveling by fundamental-mode propagation. In the period range from approximately 50–200 s, the group velocity of LR (both the oceanic and continental paths) monotonically decreases with increasing period. Physically this means that in this period range, long-period RL waves follow the law of inverse dispersion. Observations of this phenomenon are rather scarce. An example is exhibited in Kulhanek (1990).

Extremely long-period surface waves called mantle waves, with periods from somewhat less than 1 min to several minutes, have been observed from large distant shocks. They can be of either Love or Rayleigh wave type. The former propagates with nearly constant speed of 4.4 km/s and shows an impulsive shape on the seismogram. The latter travels with velocity between 3.6 and 4.1 km/s. Wavelengths of mantle waves vary from several hundreds to more than 1,000 km, so that a large part of Earth’s mantle is affected by these waves. An interesting feature of mantle waves is their repeated appearance on records, which is due to their multiple travel around the Earth. The LQ mantle wave was given the label G (after B. Gutenberg) and the LR -type mantle wave the label R . The older nomenclature sometimes uses W instead for

Seismogram Interpretation,

Fig. 6 Seismograms from an intermediate-focus earthquake ($M = 6$, $h = 199$ km) in the Hindu-Kush region on September 4, 1993, made at UPP, Sweden, at an epicentral distance of 41° . The traces show long-period channels from top to bottom: vertical, E–W and N–S components



R (from German *Wiederkehrewellen*, meaning repeated waves). *G* waves that propagate the direct and anticenter routes are labeled *G1* and *G2*, respectively. Waves that have in addition traveled once around the Earth are denoted *G3* and *G4* and so on. Similarly, we have *R1*, *R2*, *R3*, *R4*, etc. As an exceptional case, we can mention records of the 1960 Chile earthquake, $M = 9.5$. Seismograms made at Uppsala, Sweden, reveal mantle waves *G20* and *R20* that have traveled a total distance equal to that from the Earth to the Moon (Båth 1979).

Volcanic Earthquakes and Unusual Seismic Sources

During eruptive episodes, volcanoes can produce up to thousands of small earthquakes per day, recorded only at short distances, say, less than 50 km, from the volcanoes (see ► [“Earthquakes, Volcanogenic”](#)). Different categorization of volcanic earthquakes may be found in the literature. For example, Minakami (1961) classifies volcanic earthquakes into three groups, in accordance with different generation mechanisms and signal characteristics: (1) A-type earthquakes with focal depth between 1 and 10 km near the volcanic magma system, corresponding records show clear high-frequency *P* and *S* phases; (2) B-type earthquakes with foci at depth of 1 km or less, records reveal low-frequency coda, emergent onsets, and usually no distinct *S*; and (3) explosion-type earthquakes taking place at the very surface of the Earth. Close to active volcanoes, we also frequently detect so-called volcanic tremors, which are due to long-duration, more or less continuous, volcanic vibration.

Implosion earthquakes, impact earthquakes (e.g., the Tunguska, Siberia, impact in 1908), frost actions, low-magnitude icequakes generated by temperature changes in glaciers, earthquakes related to large-scale landslides, etc. are some of the types of events that, together with tectonic earthquakes, volcanic earthquakes, and oceanic microseisms, belong to the category of natural seismic sources. There is also a variety of man-made seismic sources such as industrial and military explosions, cultural noise (traffic, industry work), mining activity, high dams (seismicity, triggered/induced), fluid injections, etc. The source identification, for earthquakes other than tectonic or volcanic, is usually a task in itself. A classical example of source identification is the well-known problem of discriminating underground nuclear explosions from earthquakes (see ► [“Seismic Monitoring of Nuclear Explosions”](#)).

Conclusions

Seismogram interpretation, described in this article, is essentially devoted to the art of identification of various seismic

“arrivals” or wave types visible on seismograms. It is usually followed by seismogram analysis, which may include determination of basic source parameters (origin time, hypocenter location, size) but may also cover more advanced studies, such as waveform modeling, determination of velocity distribution, etc. Obviously, a large part of seismogram analysis is a domain of research. Nevertheless, phase identification is a doorway, and without correct seismogram interpretation, hardly any analysis would be possible. Repeated observations of peculiar phases on seismic records often led to new discoveries of details in Earth’s structure and/or dynamics. Sometimes, theories were developed in advance, to be later confirmed by observations (e.g., free oscillations of the Earth). Seismogram interpretation is a fundamental and, in our view, also fascinating and rewarding part of modern seismology.

Cross-References

- [Body Waves](#)
- [Core-Mantle Coupling](#)
- [Earth’s Structure, Upper Mantle](#)
- [Earthquake, Aftershocks](#)
- [Earthquake, Focal Mechanism](#)
- [Earthquakes, Volcanogenic](#)
- [Seismic Monitoring of Nuclear Explosions](#)
- [Seismic Phase Nomenclature: The IASPEI Standard](#)
- [Seismic, Velocity, and Density Relationships](#)
- [Surface Waves](#)
- [Traveltime Tomography Using Controlled-Source Seismic Data](#)

Bibliography

- Adams RD (1979) T-phase recordings at Rarotonga from underground nuclear explosions. *Geophys J R Astron Soc* 58:361–369
- Båth M (1979) *Introduction to seismology*. Birkhäuser, Basel/Boston/Stuttgart
- Båth M (1983) Earthquake data analysis: an example from Sweden. *Earth Sci Rev* 19:181–303
- Bolt BA (1982) *Inside the earth*. Freeman, Berkeley
- Bormann P (ed) (2012) *New manual of seismological observatory practice (NMSOP-2)*. Deutsches GeoForschungszentrum GFZ; IASPEI, Potsdam
- Cao A, Romanowicz B, Takeuchi N (2005) An observation of *PKJKP*: inferences on inner core properties. *Science* 308:1453–1455
- Engdahl ER (1968) Seismic waves within Earth’s outer core: multiple reflections. *Science* 161:263–264
- Engdahl ER, Flinn EA (1969) Seismic waves reflected from discontinuities within Earth’s upper mantle. *Science* 163:177–179
- Engdahl ER, van der Hilst RD, Buland R (1998) Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bull Seismol Soc Am* 88:722–743
- Herrin E, Arnold EP, Bolt BA, Clawson GE, Engdal ER, Freedman HW, Gordon DW, Hales AL, Lobdell JL, Nuttli O, Romney C, Taggart J,

- Tucker W (1968) Seismological tables for P phases. *Bull Seismol Soc Am* 58:1193–1352
- Jeffreys H, Bullen KE (1940) Seismological tables. British Association for the Advancement of Science, London
- Kennett BLN (1991) IASPEI 1991 seismological tables, vol 3. Research School of Earth Sciences, Australian National University, Canberra, p 122
- Kennett BLN, Engdahl ER (1991) Travel times for global earthquake location and phase identification. *Geophys J Int* 105:429–465
- Kennett BLN, Engdahl ER, Buland R (1995) Constraints on seismic velocities in the earth from travel times. *Geophys J Int* 122: 108–124
- Kulhanek O (1990) *Anatomy of seismograms*. Elsevier, Amsterdam
- Mendigüen JA (1971) Focal mechanism of a shock in the middle of the Nazca plate. *J Geophys Res* 76:3861–3879
- Minakami T (1961) Study of eruptions and earthquakes originating from volcanoes. *Int Geol Rev* 3:712–719
- Morelli A, Dziewonski AM (1993) Body wave travel times and a spherically symmetric P- and S-wave velocity model. *Geophys J Int* 112:178–194
- Neumann F (1951) Principles underlying the interpretation of seismograms. Special publication no 254. U.S. Department of Commerce, Coast and Geodetic Survey, Rockville
- Okal EA (2001) T-phase stations for the international monitoring system of the comprehensive nuclear-test ban treaty: a global perspective. *Seismol Res Lett* 72(2):186–196
- Payo G (1986) *Introducción al Análisis de Sismogramas*. Instituto Geográfico Nacional, Madrid
- Schenk T, Müller G, Brüstle W (1989) Long-period precursors to *pP* from deep-focus earthquakes: the Moho underside reflection *pMP*. *Geophys J Int* 98:317–327
- Simon RB (1968) *Earthquake interpretations*. Colorado School of Mines, Golden
- Storchak DA, Borman P, Schweitzer J (2003) The IASPEI standard seismic phase list. *Seismol Res Lett* 74:761–772
- Wadati K (1927) Existence and study of deep earthquakes (in Japanese). *J Meteorol Soc Jpn Ser* 2(5):119–145
- Willmore PL (1979) *Manual of seismological observatory practice*. Report SE-20, World Data Center for solid earth geophysics. NOAA, Boulder

Seismological Networks

István Bondár¹ and Eric Robert Engdahl²

¹Research Centre for Astronomy and Earth Sciences Geodetic and Geophysical Institute, Kövesligethy Radó Seismological Observatory, Budapest, Hungary

²Center for Imaging the Earth's Interior, Department of Physics, University of Colorado at Boulder, Boulder, CO, USA

Definition

Seismological Networks. Networks of seismographic stations for the recording of ground motions produced by seismic waves propagating from natural and anthropogenic sources.

Introduction

Seismology has its roots in observations of earthquakes and earthquake-generated ground motion. Seismological networks as envisioned by the pioneers in seismology are now a reality, a reality that is frequently upgraded and expanded as technology improves. A global system of broadband seismographs with high dynamic range is now in place. This is supplemented by many national, regional, and local networks capable of high-resolution monitoring of the frequent smaller earthquakes in seismically active places. Parametric data derived from these networks make a major contribution to national and international information services. Because of page limitations, we can only hope to provide the reader with a glimpse of all there is to know about seismographic networks.

Global Seismic Networks

Seismological recordings have been made on Earth for hundreds of years in some form or another. However, global monitoring of earthquakes only began in the 1890s when John Milne created 40 seismic observatories to measure the waves from these events (see Fig. 1). Shortly after the International Geophysical Year (1957–1958), a concerted effort was made to establish and maintain a more modern standardized seismic network on a global scale.

Worldwide Standardized Seismograph Network

In the early 1960s, the World-Wide Standardized Seismograph Network (WWSSN) was established. Between 1961 and 1966, 120 WWSSN stations with identical seismic instruments were deployed in more than 60 countries and islands throughout the world. The WWSSN program included the collection, review, and copying of the recorded seismograms and the distribution of copies to researchers throughout the world. Considered one of the most important advances ever in observational seismology, the WWSSN produced the data needed to support unprecedented progress in earthquake, explosion, and tectonic research. In the years that followed the deployment of the WWSSN, many of the stations were modernized and expanded into regions not initially covered. Eighty of the original WWSSN stations are still in operation and still supported.

International Federation of Digital Seismograph Networks

In the 1980s, the international seismological community recognized the new opportunities within its field for improved understanding of the internal structure and dynamical properties of the Earth provided by recent developments in

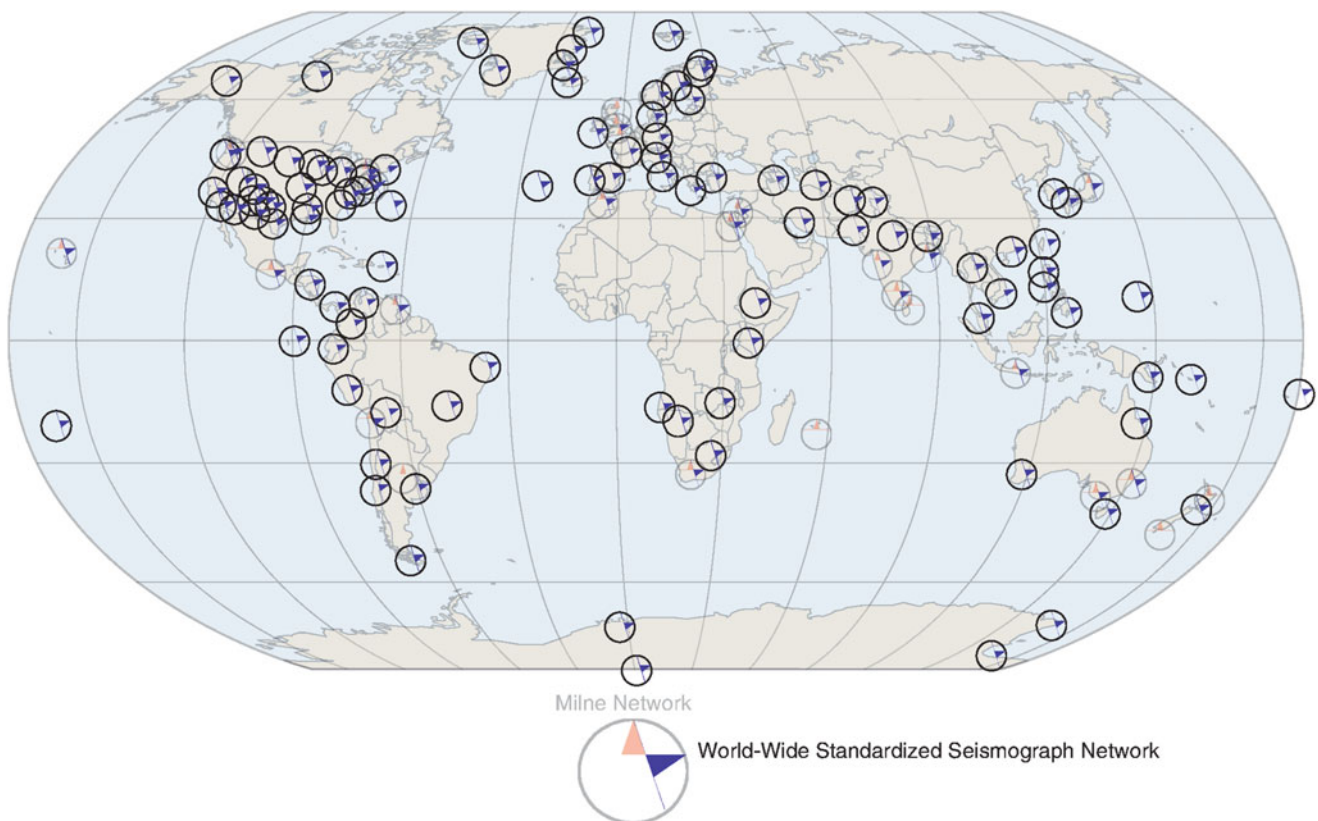
seismograph network technology. It also recognized that rapid access to seismic data from networks of modern broadband digital instruments wherever they might be was now possible. The developments included greatly improved broadband seismographic systems that capture the entire seismic wave field with high fidelity (see Fig. 2), efficient and economical data communications and storage, and widely available, powerful computing facilities.

In view of the above and to take advantage of existing developing global and regional networks, the International Federation of Digital Seismograph Networks (FDSN, <http://www.fdsn.org/>) was formed in 1986 to provide a forum for developing common minimum standards in seismographs (e.g., bandwidth) and recording characteristics (e.g., resolution and dynamic range); developing standards for quality control and procedures for archiving and exchange of data among component networks; and coordinating the siting of additional stations in locations that will provide optimum global coverage.

The FDSN was formed as a global organization, open to all national and international programs committed to the deployment of broadband seismographs and willing to contribute to the establishment of an optimum global system with timely

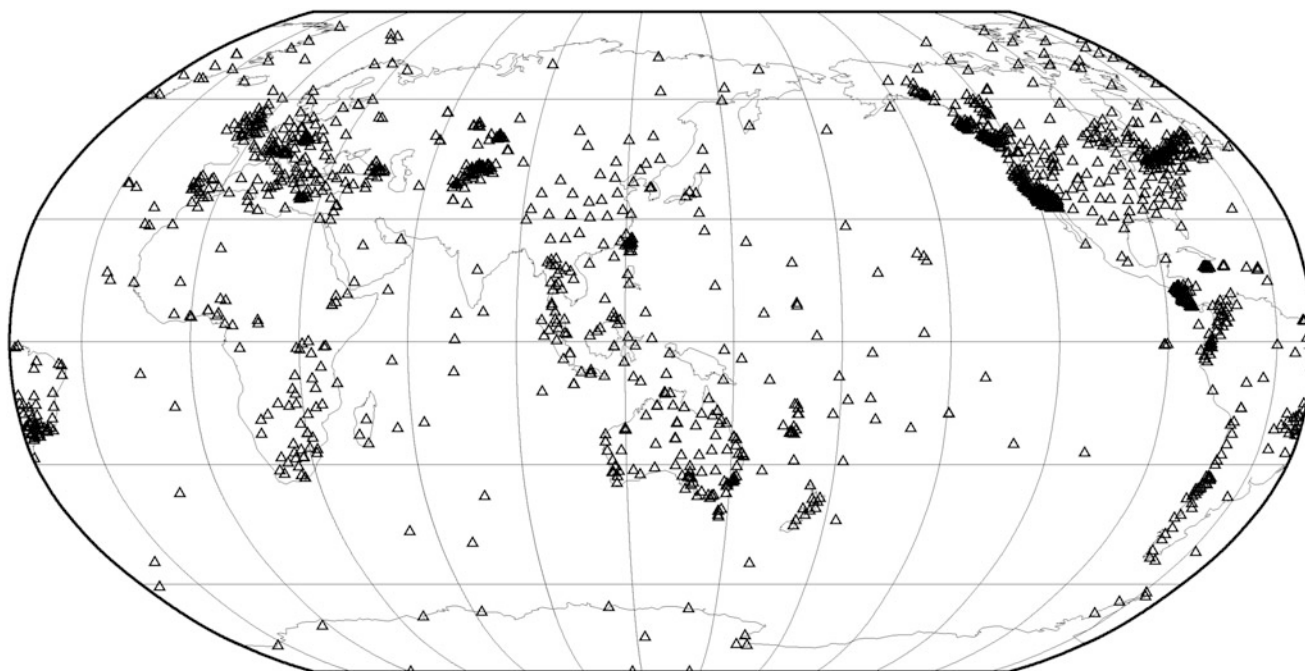
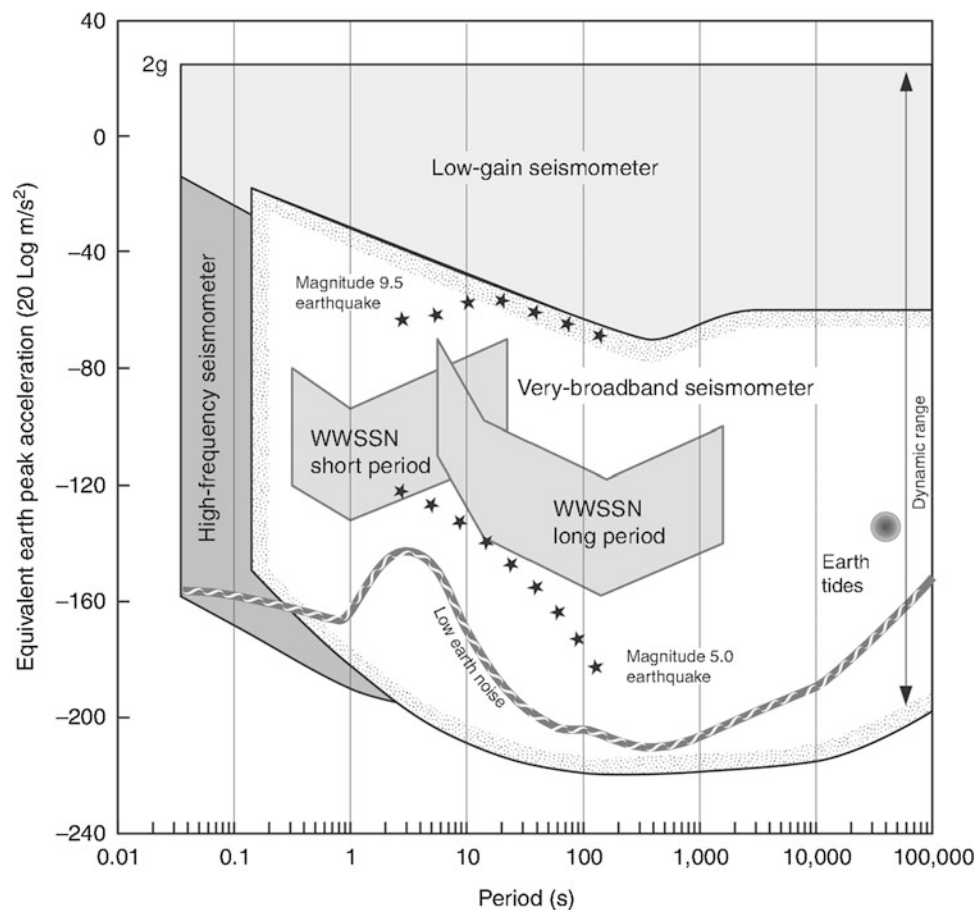
data exchange. Its membership is comprised of groups responsible for the installation and maintenance of seismographs either within their geographic borders or globally (Fig. 3). Membership in the FDSN is open to all organizations that operate more than one broadband station. However, among the most important contributors are the Global Seismographic Network (GSN) of the Incorporated Research Institutions for Seismology (IRIS) and the French GEOSCOPE and German GEOFON global networks. It is important to note that many stations in the FDSN were formerly part of the WWSSN program, some of them having been upgraded to digital recording (DWWSSN). Members agree to coordinate station siting and provide free and open access to their data. This cooperation helps scientists all over the world to further the advancement of earth science and particularly the study of global seismic activity.

The FDSN goals related to station siting and instrumentation are to provide stations with good geographic distribution, recording data with 24 bits of resolution in continuous time series with at least a 20-sample per second sampling rate. The FDSN was also instrumental in development of a universal standard (SEED) for distribution of broadband waveform data and related parametric information. The FDSN system of



Seismological Networks, Fig. 1 Locations of stations in the Milne network and the World-Wide Standardized Seismograph Network. (Courtesy of Kent Anderson, Incorporated Research Institutions for Seismology)

Seismological Networks, Fig. 2 Bandwidth and dynamic range of WWSSN and modern seismographic systems. (Courtesy of Rhett Butler, Incorporated Research Institutions for Seismology)



Seismological Networks, Fig. 3 Locations of stations in the International Federation of Digital Seismograph Networks. As of April 1919, some 3000 permanent stations are included in the FDSN. (Courtesy of Tim Ahern, Incorporated Research Institutions for Seismology)

global network observatories includes contributions from many international partners forming a large backbone of permanent seismological observatories. Currently more than 4000 permanent and temporary stations send waveforms in real time to IRIS data center (Fig. 4). Developments in communications and other technological advances have expanded the role of the FDSN in rapid earthquake analysis, tsunami warning, and nuclear test monitoring. With such long-term observations, scientists are now getting a glimpse of Earth structure changes on human time scales, such as the rotation of the inner core. Continued observations for the next 50 years will enhance our image of the Earth and its processes.

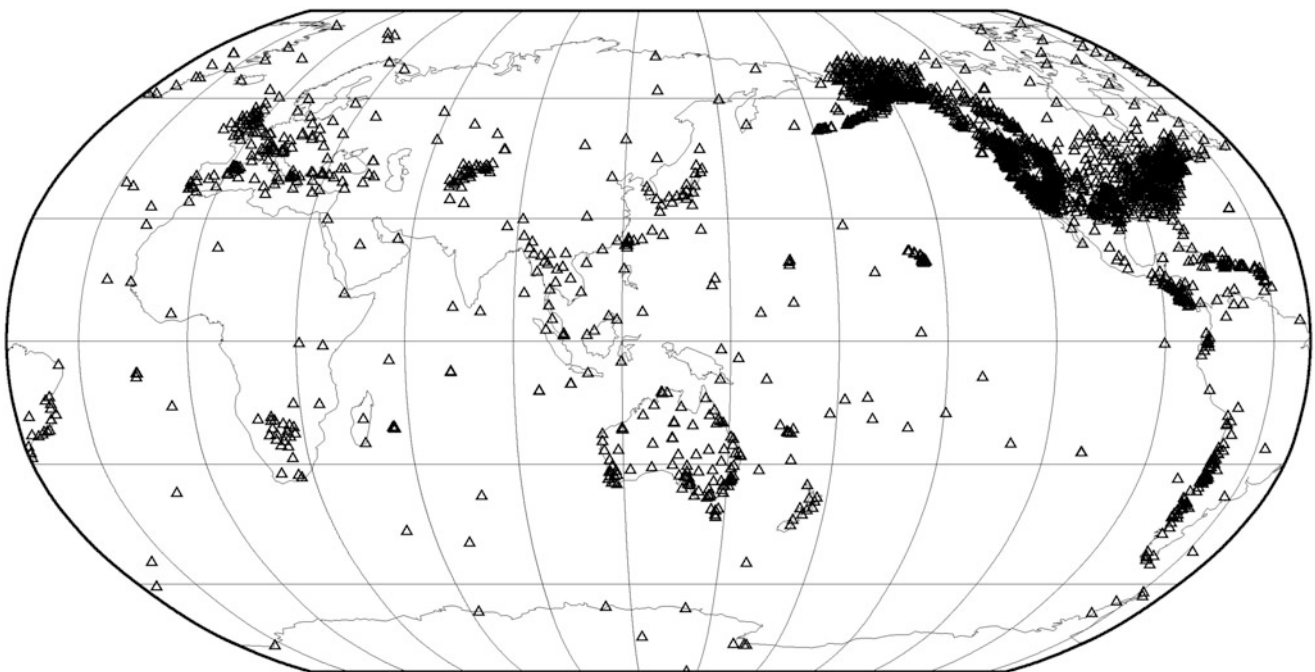
International Monitoring System

Under the Comprehensive Nuclear-Test-Ban Treaty (CTBT, <http://www.ctbto.org/>), it is a requirement that there will be an International Monitoring System (IMS) to detect any clandestine nuclear weapon detonation in any environment – underground, underwater, or above ground. The aim of the IMS is detection, identification, and location of any such test, and the technologies involved will be seismological, hydroacoustic, infrasound, and atmospheric radionuclide monitoring. In support of these technologies, there will be appropriate means of global communication and an International Data Centre (IDC) to which monitoring data will be transmitted.

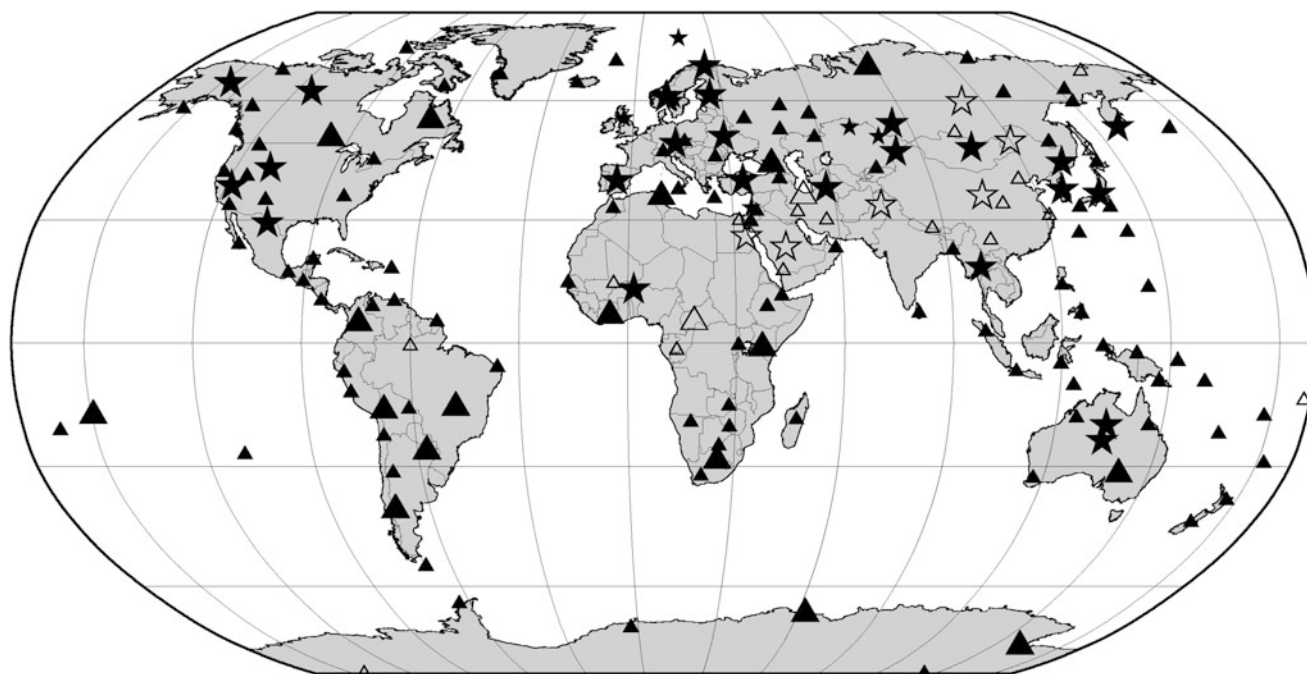
Seismological monitoring is considered the cornerstone of the IMS because of the likelihood that tests would be

conducted underground and out of sight. The seismological technology required for the detection of earthquakes is well known, but the problem for the IMS is to distinguish between signals from weapons detonations, earthquakes, and nonnuclear explosions in quarrying or mining. There are so many earthquakes (>200,000 per year) and mining events worldwide, of similar magnitude to a small nuclear explosion, that data analysis becomes a key problem. The seismological verification program therefore includes intensive study of seismicity and mining operations in order to develop regional backgrounds.

A global network of 50 “primary” stations in 34 countries is planned, with 2 basic equipment packages: “three-component” (3-C) broadband seismometers, which act as single detection systems monitoring ground movement in 3 directions – vertically and horizontally east-west and north-south – or “arrays” which include 3-C systems plus a cluster of narrowly spaced vertical-component short-period sensors distributed to gain optimum signal-to-noise ratios by waveform correlation. Arrays may contain many detectors located over large areas (up to 200 km²). The advantage of an array station is that it allows approximate event location on a stand-alone basis. Historical developments such as the LASA, NORSAR, and Graefenberg arrays have marked important steps in seismological network development that have facilitated new research fields and services in seismology. As Fig. 5 indicates, approximately half of the proposed stations will be arrays.



Seismological Networks, Fig. 4 Permanent and temporary stations providing real-time data to IRIS. (Courtesy of Tim Ahern, Incorporated Research Institutions for Seismology)



Seismological Networks, Fig. 5 Locations of stations in the CTBTO IMS network (Comprehensive Nuclear-Test-Ban Treaty Organization www.ctbto.org). Symbols defined as follows: filled symbols,

operational; open symbols, not yet operational; stars, array stations; triangles, three-component stations; large symbols, primary stations; and small symbols, auxiliary stations

In addition to the primary network, an “auxiliary” network comprising 120 stations distributed among 61 countries is planned (Fig. 5). Stations in this network will mainly consist of existing 3-C stations, which already form part of the host countries’ seismological monitoring operations. The purpose of the auxiliary network is simply to provide additional data to support that from the primary network in order to facilitate signal discrimination and hence event detection and location.

Regional Seismic Networks

National seismographic networks operate within the political boundaries of a country, and their primary mission is to issue rapid alerts to government agencies and the general public for potentially damaging earthquakes. National networks typically exchange data with neighboring countries in order to improve the accuracy of the determination of earthquake parameters in border regions.

Regional networks play an important role in monitoring the seismicity of the Earth within a well-defined region. Regional networks may be part of a national network (e.g., the United States, China, Japan, and Russia) or may be constituted from multinational networks (e.g., Euro-Mediterranean region). The magnitude detection threshold of regional networks is typically lower than that of the global networks, which allows for producing bulletins for a specific region that are more complete than it could be achieved at a global scale.

Local networks focus on a much smaller area than regional networks. Examples of local networks include temporary deployments to capture aftershock sequences of a large earthquake, networks monitoring volcano activities, and dense local networks installed in urban areas with elevated seismic risk.

Euro-Med Region

The Euro-Med region encompasses Europe, North Africa, and the Middle East. The European-Mediterranean Seismological Centre (CSEM/EMSC, <http://www.emsc-csem.org>), an international, nongovernmental organization of more than 70 seismological data centers of more than 50 countries, produced earthquake bulletins in the region from 1998 to July 2012, using parametric data reported by national and local networks in the region. After this date, all contacts and exchange procedures were transferred to the ISC to ensure data will remain collected.

The mission of EMSC is to provide rapid comprehensive seismological information and to disseminate notifications for potentially damaging earthquakes in the region in coordination with national and regional institutions as well as to offer redundancy and backup to national services. The EMSC and the US National Earthquake Information Center (NEIC) routinely exchange data, and NEIC bulletin data for the European-Mediterranean region are incorporated in the EMSC data collection. Earthquake parameters (location, magnitude, phase picks, moment tensors, etc.) provided

by the EMSC and its members are incorporated in the global bulletin prepared by the International Seismological Centre (ISC).

The EMSC developed a multichannel information system (websites, Quakebot on Twitter, and a smartphone app, LastQuake) offering rapid information of felt and damaging earthquakes to eyewitnesses and crowdsourcing observations (felt reports, geo-located pictures, and videos) to contribute to improved situation awareness.

United States

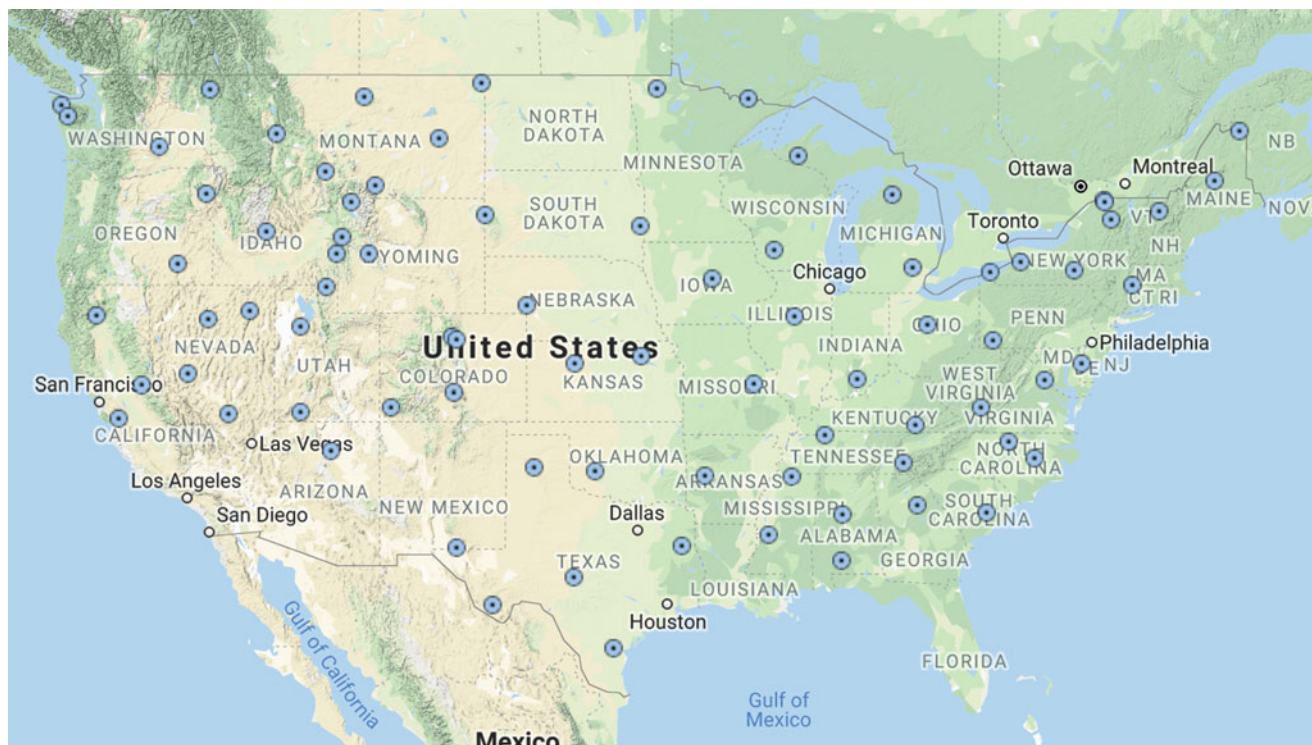
In the United States, the US Geological Survey's National Earthquake Information Center (NEIC, <http://earthquake.usgs.gov/regional/neic>) is the national data center and archive for earthquake information. Its mission is to determine as rapidly and accurately as possible the location and size of all significant earthquakes that occur worldwide. The NEIC operates 24 h a day and issues rapid alerts for earthquakes larger than magnitude 3 in the conterminous states of the United States and for those larger than magnitude 5 globally. The NEIC publishes global earthquake bulletins on a daily, weekly, and monthly basis.

The NEIC serves as the National Operations Center of the Advanced National Seismic System (ANSS), which includes the ANSS backbone network and regional seismic networks across the United States. Figure 6 shows the permanent digital

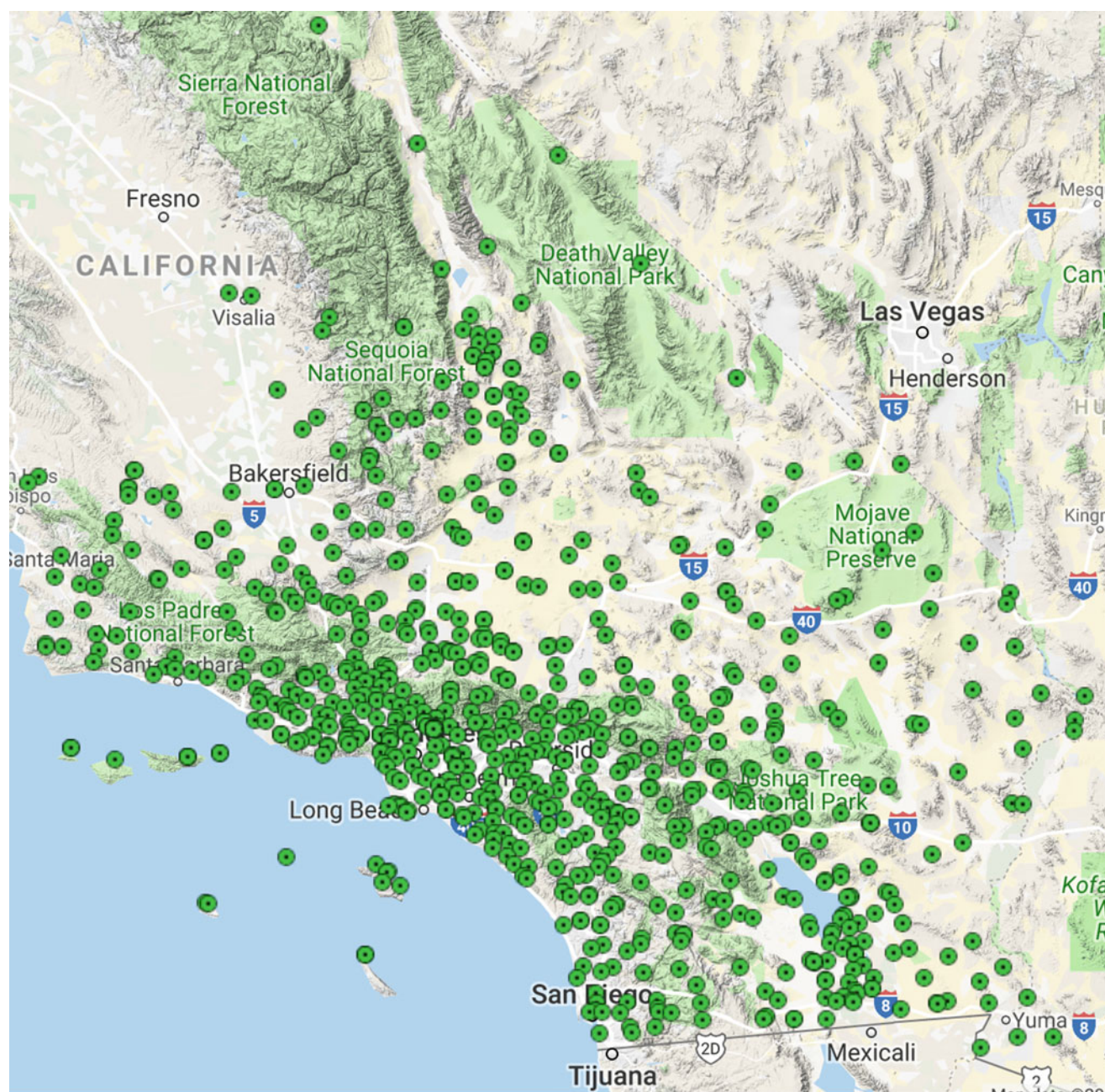
seismic and strong motion stations in the 48 conterminous states of the United States. The strong motion stations (circles) represent an integral part of the ANSS network as first arriving phases are picked on strong motion records and used in the earthquake location procedures. Figure 7 shows the regional Southern California Seismic Network. As in Fig. 6, the map is colored by the peak ground acceleration, a measure of seismic hazard. The network is much denser in urban areas and regions with high seismic hazard.

China

The China Digital Seismograph Network (CDSN) was established in 1986, with nine digital seismic stations. During the past 20 years, the Chinese seismic network, both at the national and regional levels, went through unprecedented improvements. The China Earthquake Administration (CEA) has completed the analog-to-digital transition of the existing analog stations and deployed a large number of new digital stations (Liu et al. 2008). Figure 8 shows the currently operating national network of 162 broadband seismographic stations. Some of these stations also belong to the FDSN. Dense regional networks further support the national network. Each of the 31 provinces, autonomous regions, and municipalities operates regional digital networks. Figure 9 shows the distribution of the 945 regional seismic stations.



Seismological Networks, Fig. 6 USGS backbone permanent network in the 48 conterminous states of the United States. (Courtesy of Incorporated Research Institutions for Seismology, <http://ds.iris.edu/gmap>)



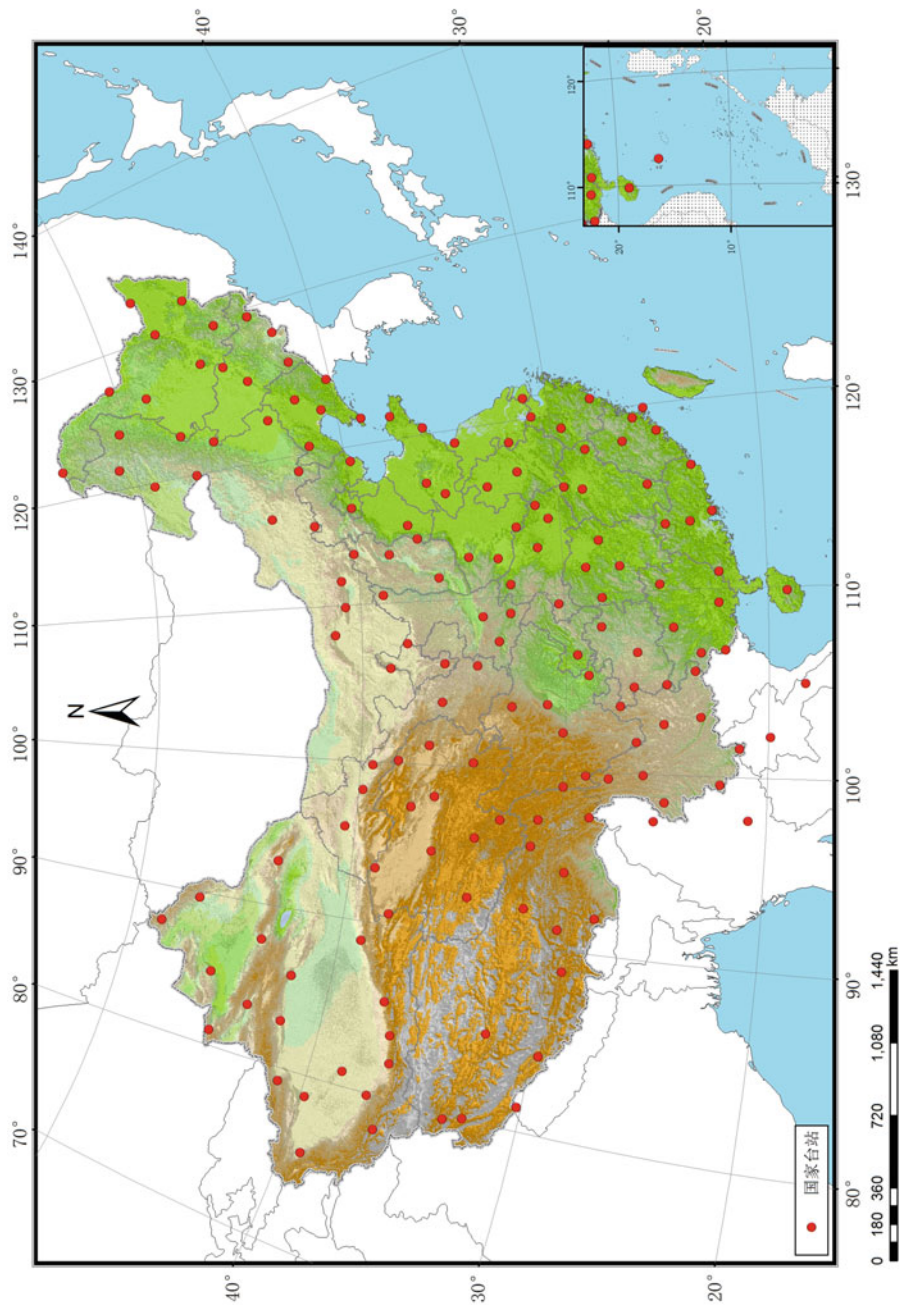
Seismological Networks, Fig. 7 The Southern California Seismic Network: an example for a very dense regional network. (Courtesy of Incorporated Research Institutions for Seismology, <http://ds.iris.edu/gmap>)

China's Earthquake Networks Center (CENC) archives the continuous waveform data from the national and regional networks and performs real-time data processing. The CENC issues rapid earthquake alerts for $M_s > 4.5$ events inside China within 10 min and focal mechanism solutions in no more than 30 min after the earthquake occurred.

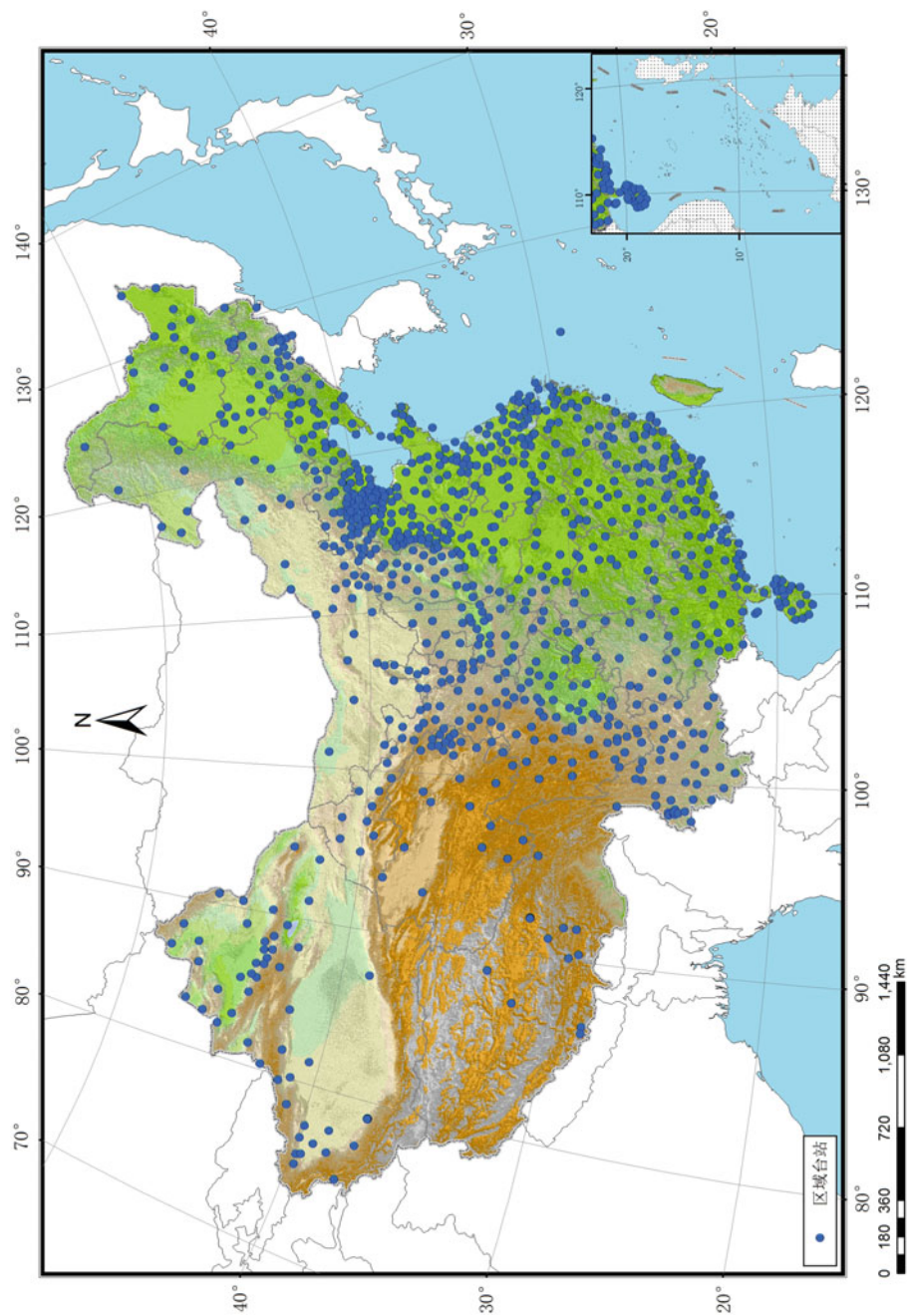
Large-Scale Arrays

The USArray (www.usarray.org), one of the three-component Earthscope projects (www.earthscope.org), has opened a new chapter in seismology. The project began in

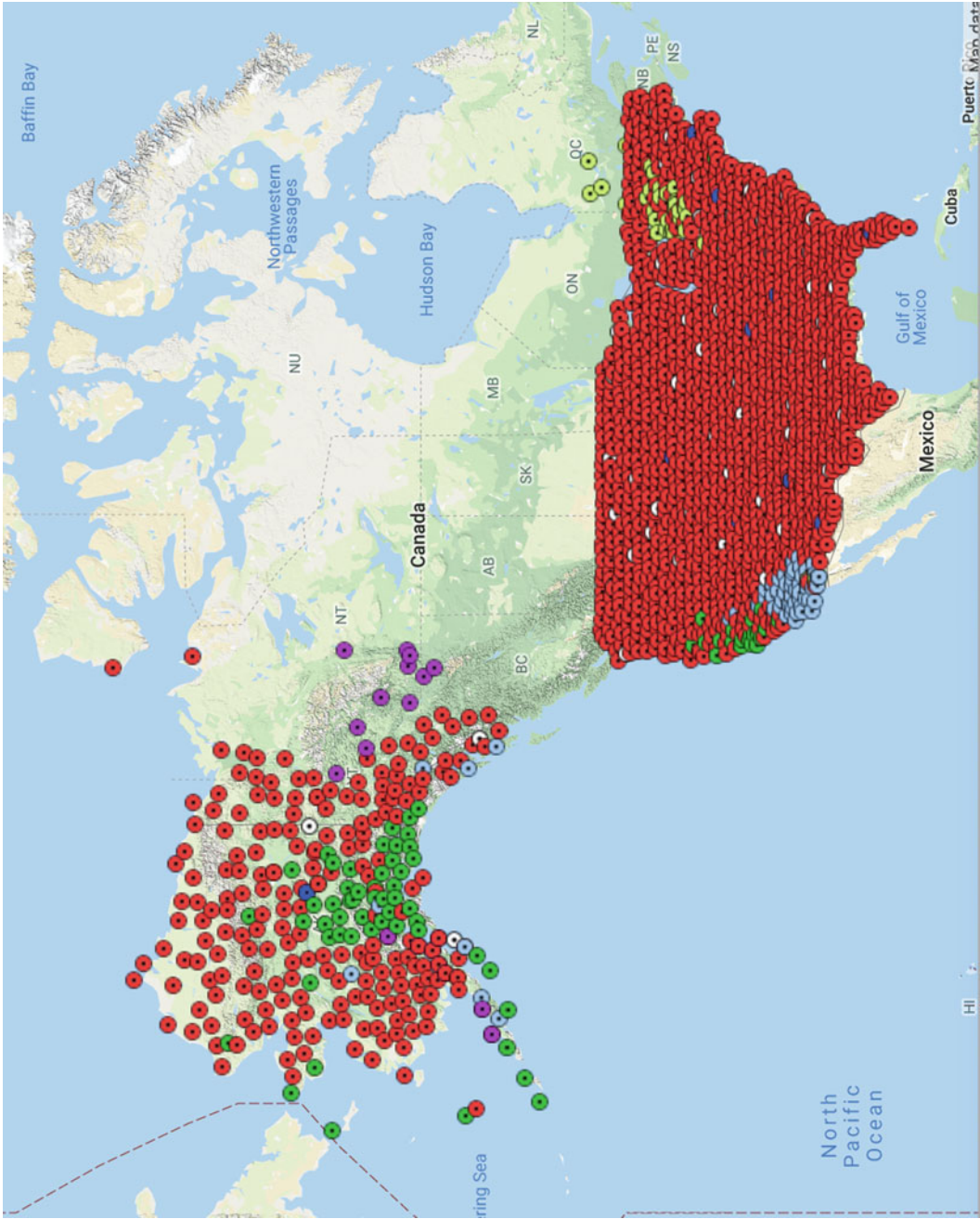
2007 with some 400 high-quality broadband seismographs deployed with 70 km station spacing in the West Coast and marching across the entire mainland United States, reaching the East Coast in 2013. The mainland USArray project finished in 2017, and now it has moved to Alaska (Fig. 10). USArray provided data on the North American lithosphere in unprecedented detail. The success of USArray has inspired large-scale array deployments worldwide, such as the IberArray (iberarray.ictja.csic.es), AlpArray (alparray.ethz.ch), ChinaArray (chinaarray.org), and, most recently, AusArray and AdriaArray.



Seismological Networks, Fig. 8 Locations of stations belonging to the China Digital Seismograph Network. (Courtesy of Baokun Li and Guangbao Du, China Earthquake Networks Center)



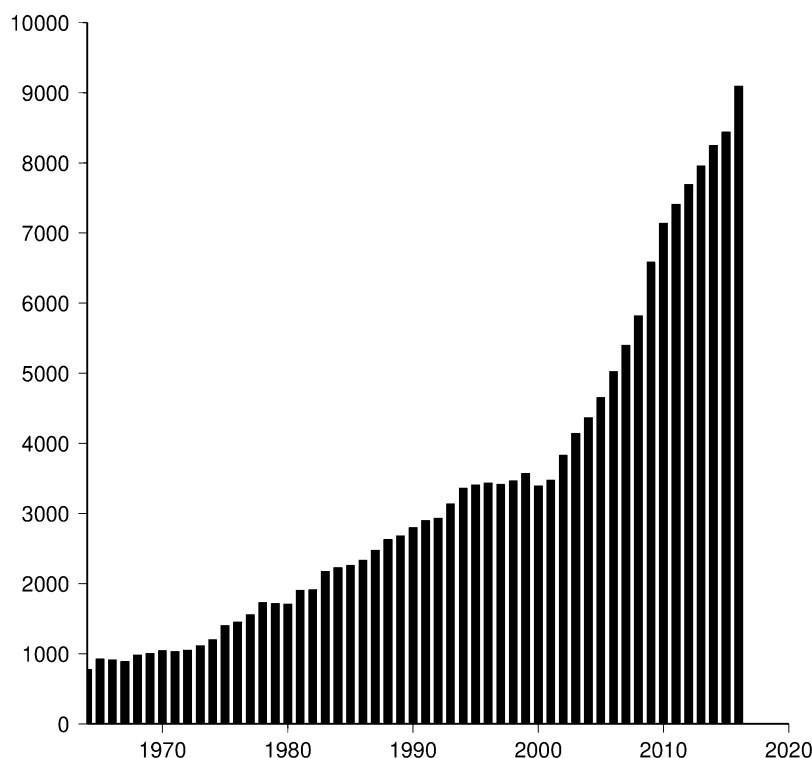
Seismological Networks, Fig. 9 Digital seismographic stations operated by 31 regional networks in China. (Courtesy of Baokun Li and Guangbao Du, China Earthquake Networks Center)



Seismological Networks, Fig. 10 USAarray deployment in the mainland United States and Alaska. (Courtesy of Incorporated Research Institutions for Seismology, <http://ds.iris.edu/gmap>)

Seismological Networks,

Fig. 11 Number of stations reporting to the ISC between 1964 and 2016. (Courtesy of James Harris, International Seismological Centre)



International Registry of Seismograph Stations

Since the humble beginnings of instrumental global seismology in the late nineteenth century, the number of seismographic stations has steadily increased every year. Figure 11 shows the number of stations reporting to the ISC in each year since its operations began. In order to be able to uniquely identify seismographic stations, the ISC and the NEIC, in its capacity as World Data Center A for Seismology, jointly maintain the International Registry of Seismograph Stations (IRSS). The IRSS represents a global catalog of seismic stations (currently with some 27,000 registered stations) and contains information about the station coordinates, instruments, operating networks, and when a station began/stopped operating. To facilitate data exchange between networks and organizations, it is strongly recommended that network operators register their stations and use the international station codes.

Summary

The high-quality data recorded by seismographic networks would be of limited value without international cooperation by all countries worldwide in the acquisition and exchange of seismic measurements and waveforms. This cooperation is essential for the location of earthquakes, for understanding the physics of earthquakes, and for studies of Earth's internal structure, properties, and processes. One of the latest challenges to acquiring, processing, and distributing data from

seismographic networks by data centers globally is the automatic performance of traditional tasks in as close to real time as current technology permits.

Cross-References

- [Earthquake, Aftershocks](#)
- [Earthquake, Magnitude](#)
- [Earthquakes, Location Techniques](#)
- [Seismic Monitoring of Nuclear Explosions](#)
- [Seismology, Monitoring of CTBT](#)

Acknowledgments The authors thank Baokun Li and Guangbao Du (China Earthquake Networks Center), Tim Ahern (Incorporated Research Institutions for Seismology), Ronan LeBras (Comprehensive Nuclear-Test-Ban Treaty Organization), Remy Bossu (European-Mediterranean Seismological Centre), and James Harris (International Seismological Centre) for providing input to the article.

Bibliography

- China Earthquake Networks Center. www.cenc.ac.cn
 Comprehensive Test-Ban-Treaty Organisation. <https://www.ctbto.org>
 European-Mediterranean Seismological Centre. <http://www.emsc-csem.org>
 Incorporated Research Institutions for Seismology. www.iris.edu
 International Seismological Centre. www.isc.ac.uk
 Liu R-F, Gao J-C, Chen Y-T, Wu Z-L, Huang Z-B, Xu Z-G, Sun L (2008) Construction and development of China digital seismological observation network. *Acta Seismol Sin* 21:535–541

Seismology and Environment

Jean-Paul Montagner¹, Anne Mangeney² and
Eléonore Stutzmann²

¹Seismological Laboratory, Institut de Physique du Globe,
University Paris-Diderot, UMR CNRS/7154, Paris, France

²Seismological Laboratory, Institut de Physique du Globe,
Université de Paris, UMR CNRS/7154, Paris, France

Definition

Environmental Seismology	Field using seismic waves for investigating phenomena at the interface of solid and fluid Earth.
Anthropic Seismology	Field using seismic waves for investigating phenomena induced by anthropogenic activity (nuclear explosions, triggered seismic events in industry, mines, oil reservoirs, quarries, traffic, . . .).
Glacial Earthquakes	Earthquakes occurring in glaciers or ice polar caps, often associated with calving of icebergs at glacier terminus or at the discontinuity between bedrock and glaciers (specific class of icequakes with large magnitude).
Hurricanes, typhoon	Violent wind with a circular movement originating in tropical zones.
Landslides	Destabilization of rocks or soils from a sloping area such as a mountain, a cliff, or a volcano.
Microseisms	Small earth tremor whose sources are related to the ocean-wave activity. Primary microseisms are oscillating at the characteristic ocean-wave periods (approximately between 10 and 20 s periods) and result from the interaction of these waves with the shallow sea floor. Secondary microseisms are generated by interactions among ocean waves and have the double of their frequency.
Seismic hum	Continuous excitation of the Earth in the absence of earthquakes (at periods longer than 50 s) created by the interaction of long period ocean waves (infragravity waves) and the sea floor. This signal is weak but nevertheless large enough to excite the earth-free oscillations.
Snow avalanches	Destabilization of a mass of snow from a sloping topography turning into a flow of snow particles or aggregates that stop downslope.

Very long
period noise

For periods larger than 200 s, the noise results from gravitational attraction changes induced by local pressure perturbations.

Introduction

Seismology is usually devoted to the investigation of tectonic earthquakes and imaging of the solid structure of the Earth. But with the improvement of seismometers and the explosion of the continuous recording of the ground motion by broadband seismic networks, other phenomena can be investigated. Since seismometers are sensitive to any kind of ground displacement, small events hidden in the noise can be detected. A large part of these small events is related to the interaction or the coupling of the surface of the solid Earth with its surficial unconsolidated solids, fluid/gas envelopes (ocean and atmosphere). Iceberg calving, glacier dynamics, landslides, snow avalanches, hurricanes, and ocean waves belonging to this kind of environmental events are able to generate observable seismic signals. Microseisms permanently excited by the interaction of gravity waves and swells in oceans, fluvial seismic signals, make it necessary to use statistical approaches for extracting useful information on the source processes in a broad frequency range. In contrast to microseisms, some events are episodic, deterministic; some of them well located in space and time and can be individually investigated. The study of all these natural phenomena giving rise to observable seismic signals constitutes the field of environmental seismology.

Due to the faintness of the associated signals, new methods and techniques have been used. For deterministic events, beamforming methods, time-reversal methods or their less general form, and back-projection technique (see Larmat et al., this issue for all references) have been applied to locate them in space and time. Microseisms were investigated by Wiechert (1904), and physical processes to explain them were proposed many decades ago (Longuet-Higgins 1950; Hasselmann 1963). For a stochastic approach of ambient noise, the cross-correlation methods of signals between two seismic stations are now routinely used and make it possible to retrieve the empirical Green's function (impulse response of the propagating medium) between these two stations (see Campillo et al. this issue). They are particularly well suited for monitoring physical parameters such as variation in the sub-soil temperature, aquifers, and climate and global changes.

Environmental seismology can be distinguished from "anthropic seismology," including the study of seismic signals associated with anthropic activity, such as nuclear, mining explosions, attacks, triggered events, and urban seismology (traffic, footquakes, etc.), even though similar methods can be used to study them.

Broadband Seismic Noise

Seismic noise has several and various origins. It is present in the whole seismic frequency range 50 Hz–1 mHz (or period range 0.02–1000 s). It is classical to divide the frequency (or period) range of seismic noise into four bands:

- Frequency higher than 1 Hz (periods smaller than 1 s), noise is mostly related to local sources, impulsive natural events of various origins, and it also corresponds to the area of expertise of anthropic seismology.
- Frequency band between 1 Hz and 0.05 Hz (period range 1–20 s) corresponds to the classical **microseismic** band.
- Frequency range between 0.05 Hz and 0.01 Hz (period range 20–100 s), sometimes named the noise notch where the noise is minimum.
- Very low frequency range (frequency < 0.01 Hz) or very long-period range (period $T > 100$ s) where the seismic **hum** is the best observed.

The high frequency range ($T < 1$ s), where the anthropic sources are the most visible, is routinely used in urban seismology (Ward and Crawford 1966; Bonnefoy-Claudet et al. 2006; Nakata et al. 2011). It is also the frequency range of fluvial seismology where the seismic noise induced by rivers is well correlated with meteorological and hydrological data (Govi et al. 1993; Burtin et al. 2008; Tsai et al. 2012; Gimbert et al. 2014) and can be used as well to investigate erosion processes and sediment transport dynamics (Roth et al. 2016), avalanches, or rockfall activity (Hibert et al. 2011; Dammeier et al. 2011). This rapidly developing field will not be detailed here. Many natural events (icequakes, landslides) can be investigated in the whole broadband frequency range. At longer periods (smaller frequency), noise is dominated by natural sources. It is the strongest between 4 s and 20 s, and this period range is named the microseismic band, with two prominent peaks at ~7 s (secondary peak) and ~14 s (primary peak) due to the complicated nonlinear interaction between oceanic waves and solid Earth (Longuet-Higgins 1950; Hasselmann 1963; Kedar et al. 2008; Arduin et al. 2011; Stutzmann et al. 2012). At longer period, in the absence of earthquakes, the interaction between the liquid/gas envelopes (ocean, atmosphere) and the solid earth continuously excites the free oscillations (Suda et al. 1998; Kobayashi and Nishida 1998; Arduin et al. 2015). This phenomenon is named seismic hum.

Continuous Monitoring of the Earth Environment

Microseisms have been investigated for many decades and are continuously excited, and it is not possible to treat them as individual events. However, for a homogenous distribution of noise sources, it is possible to derive the empirical Green's function by using stacking contributions of all microseismic

sources to the cross-correlations of seismic records of two stations. Cross-correlation formula CC_{AB} between seismic signals $s_A(t)$, $s_B(t)$ of two stations A and B is given by:

$$CC_{AB}(\tau) = \int s_A(t) s_B(t + \tau) dt$$

and the time derivative of $CC_{AB}(\tau)$ is simply related to the Green's function. Seismic records associated with microseisms and hum are dominated by surface waves, but it is possible to also extract body-wave data from noise data. So far, the application of cross-correlation methods has been the local tomography (Shapiro et al. 2005) or global tomography (Nishida et al. 2009) and the monitoring of seismogenic zones, and volcanic zones (see Campillo et al. this issue and references therein). The multiplicity of sources (microseisms, hum) and their continuous excitation can also be used to investigate environmental challenges (Larose et al. 2015).

Usually considered as noise, microseisms (and more generally environmental events) can also be recognized as a meteorological signal (Gutenberg 1958) as they provide complementary information on ocean activity, for example, and can ultimately provide invaluable information on global warming.

Spatiotemporally Localized Events

It is also possible to extract some specific events from the seismic noise that we can consider as deterministic events. Three examples are presented here: the glacial earthquakes, cyclones, and landslides.

Glacial Earthquakes

Cryoseismology, the field investigating seismic signals associated with the dynamics of glaciers, has exploded during the last two decades, (see the review papers of Podolskiy and Walter 2016, and Aster and Winberry 2017). Glacial earthquakes are specific cases of icequakes but belong to this new class of seismic environmental events of large magnitude which can be as large as ~5. These events have been discovered since the 70s and detected automatically in Greenland on long-period seismograms recorded on the FDSN (Federation of Digital Seismograph Networks) (Ekström et al. 2003).

Most glacial earthquakes occur during the late summer in the northern hemisphere, showing a strong seasonality. Most of them have been detected in Greenland and Alaska. Glacial earthquakes in Antarctica are less well studied, but some of them exhibit several characteristics similar to glacial earthquakes in Greenland. Different mechanisms are in play, and the cumulative stickslip motion of ice stream in West Antarctica is equivalent to an earthquake of moment magnitude ~7 in spite of modest seismic amplitudes owing to the source duration of 20–30 min (Wiens et al. 2008). The origin of large

glacial earthquakes has been first interpreted as resulting from stick-slip motion or sliding at the base of marine terminating glacials (Ekström et al. 2003). However, thanks to mechanical models and more precise visual observations, it has been shown that iceberg calving is the main cause for generating the large glacial earthquakes even though other processes can be invoked for different kinds of icequakes (Tsai et al. 2008; Amundson et al. 2008; Sergeant et al. 2019). Indeed, a strong temporal correlation is observed between the distinct seismic signals of glacial earthquakes and large ice-loss events in which icebergs of cubic-kilometer scale capsize against glacier terminus. Figure 1(left) displays an example of seismogram for a glacial earthquake detected in Alaska. It is characterized by abnormal long-period spectral content compared with a classical tectonic earthquake. Since there are generally no clear identifiable P- or S-wave arrivals, innovative techniques have been used to locate and quantify the glacial earthquakes, such as the cross-correlation technique of Ekström et al. (2006) or the time-reversal method (Larmat et al. 2006, Fig. 1 middle).

The seismogenic process is related to the force exerted by these icebergs on the glacier and the underlying solid earth. Sometimes, a sudden change in glacier speed results from these glacial-earthquake calving events. Figure 1 (right) shows the time variations of the force for a large-scale multiple-iceberg calving episode at the terminus of the Jakobshavn Isbrae on 21 August 2009 (Sergeant-Boy et al. 2016). It was recorded by broadband seismic stations located in Greenland including the seismic records from four permanent stations of the GLISN network (Dahl-Jensen et al. 2010).

The rate of glacial earthquake occurrence varied during the last decades but tends to increase since the beginning of the century (Olsen and Nettles 2017; Sergeant et al. 2019). Seasonal and interannual variations in glacier-terminus position partly account for general characteristics of the temporal variation in earthquake occurrence. By coupling seismological analysis with numerical modeling of iceberg calving, Sergeant et al. (2019) quantified the spatiotemporal evolution of ice loss in Greenland from iceberg calving between 1993 and 2013. They showed that calving was responsible for the loss of 370 gigatonnes of ice, representing 8–21% of the mass lost on glacier termini, the rest attributed to ice melt.

Cyclones (Hurricanes–Typhoons)

Tropical cyclones (hurricanes in the Atlantic Ocean and typhoons in the Pacific ocean) are also sources of seismic signal in the microseismic band. They participate to microseismic source distribution, as particular cases of the atmosphere-ocean system interacting with the solid Earth (Ebeling and Stein 2011). Storms are sensitive to topography and also carry information on the structure of the Earth and on meteorological processes.

Ocean wave-wave interactions are very large in the wake of cyclones. They generate P-waves just below the ocean

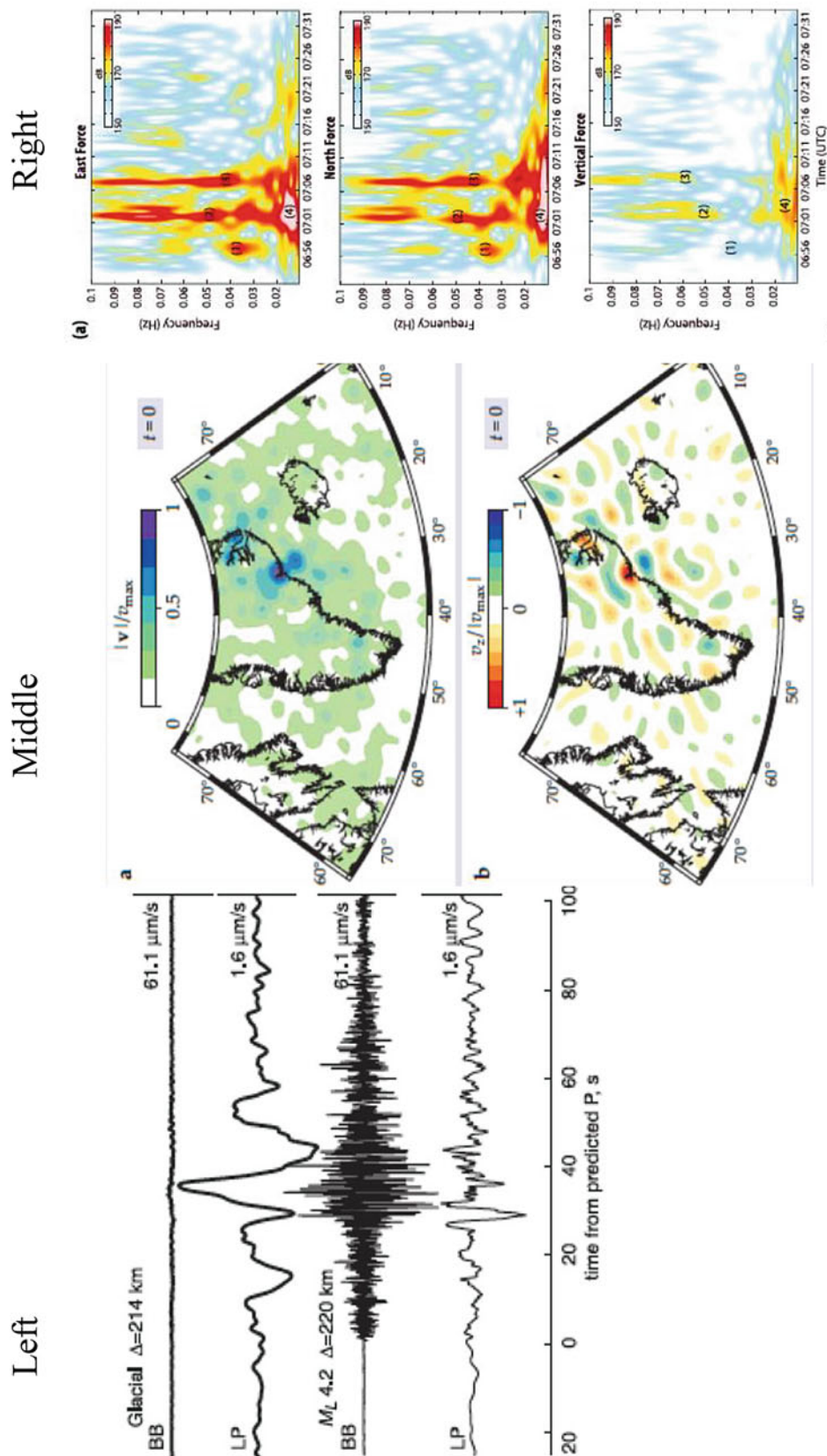
surface that propagate in the ocean and reach the ocean bottom where they are reflected, transmitted, and converted, and the resulting body and surface waves propagate in the Earth (Gerstoft and Bromirski 2016). The *P* waves penetrated deep into the earth for the super Typhoon Ioke and have been recorded by the SCSN (Southern California Seismic Network) (Fig. 2a), allowing source localization with array processing. One application is the quantification of cyclones, by using the amplitude of seismic waves (Fig. 2b, Farra et al. 2016). The temporal discernible trends in the increase of the hurricane frequency or energy during the last decade can be monitored by their seismic signature.

Using machine learning algorithms such as blind source separation technique, it is also possible to measure travel times from individual strong sources in the diffuse seismic wavefield (Meschede et al. 2019). This method enables to determine the mantle structure along paths usually not sampled by tectonic earthquakes and has promising application for ocean bottom networks (Fig. 3).

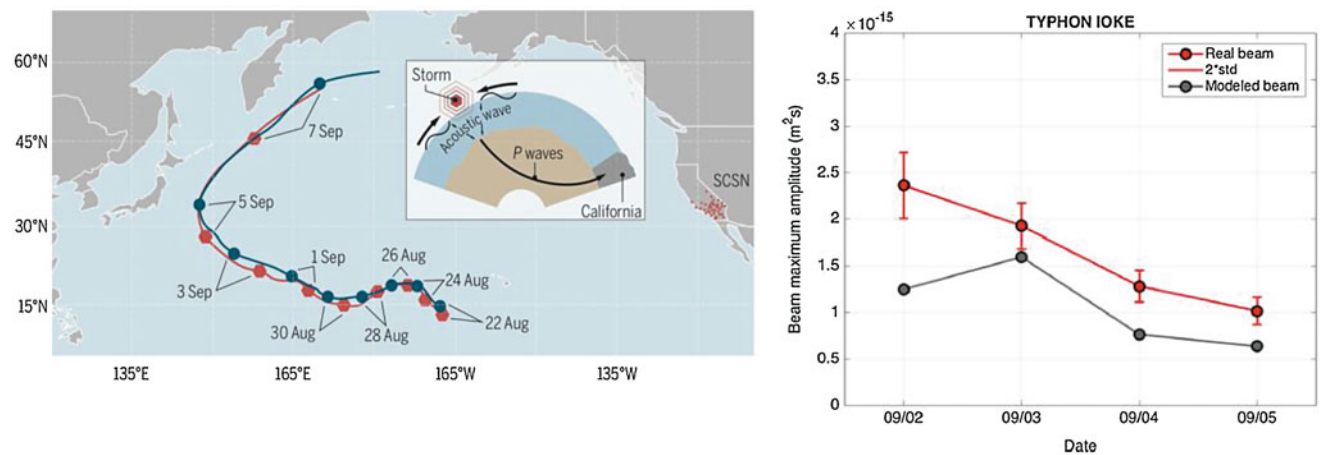
Landslides

Landslides are generally unpredictable and inaccessible processes, but their signature is available in the form of seismic waves (Kanamori and Given 1982; Kawakatsu 1989). The same seismic approach was also followed for avalanches (Lawrence and Williams 1976). Indeed, as landslides travel down the slope, the flowing mass apply a time varying force to the topography that generate waves in a broad frequency range (0.005–50 Hz). These waves are recorded at distances up to 1000 km from the source depending on the landslide volume (m^3 to km^3) and the frequency. Beyond mere detection and localization of landslides, exploitation of this amazing amount of high-accuracy wave data provides unprecedented clues to flow characteristics and dynamics (Hibert et al. 2011; Moretti et al. 2012). This is, however, a highly challenging issue because of the complexity of natural processes and of their subtle and intricate imprint on wave characteristics. The sources of the generated waves are highly sensitive to the dynamic history of landslides, which itself depends on landslide characteristics (volume, topography, fluid content, etc.) and on the physical processes involved (friction of the sliding material, collision of particles with the ground, erosion/deposition, etc.). As a result, detailed information on landslides dynamics can only be obtained by comparing the recorded waves with accurate modeling of the source processes and wave propagation.

The low frequency source of landslide-generated waves ($f < 1$ Hz) results from the mean acceleration/deceleration of the flowing mass (Brodsky et al. 2003), while the high frequencies are related to particle-scale force variations (Farin et al. 2018). Indeed, the huge energy released by a landslide at large scale breaks down into increasingly fragmented mass movements, generating different types of seismic waves. The energy is thus dissipated across a granular medium at

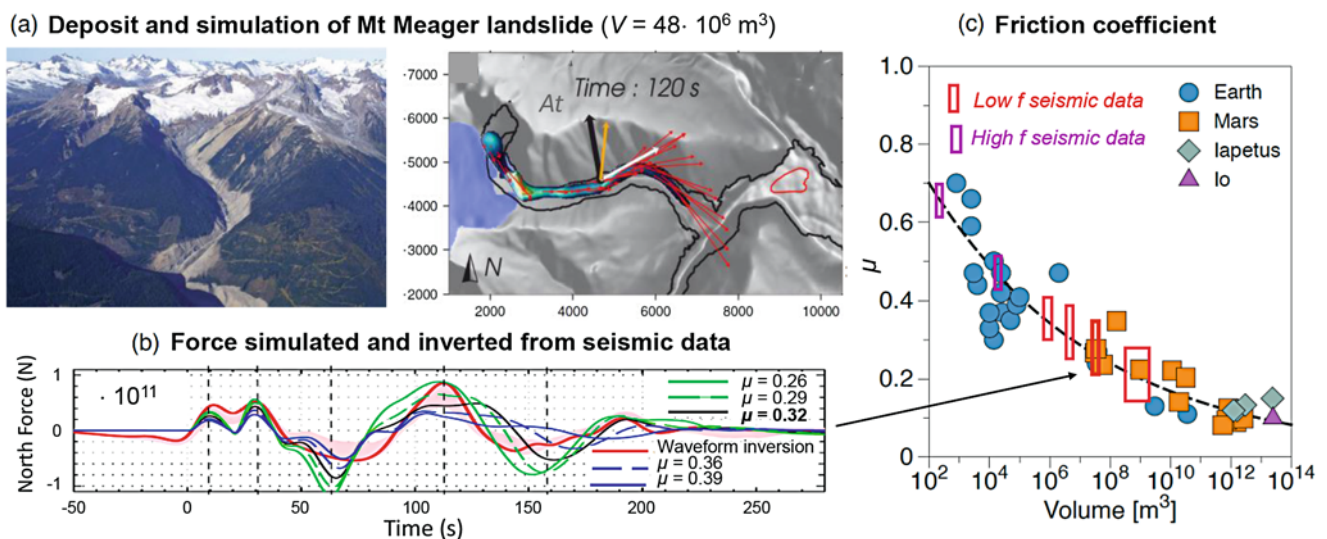


Seismology and Environment, Fig. 1 Left: example of Seismogram from a glacial earthquake in Alaska (4 September 1999, 15:15:20.0) (top two traces) compared to a nearby, tectonic, $M = 4.2$ earthquake (bottom two traces) recorded at comparable distances. For each trace pair, top shows broad-band (BB) (0.00833 to 20 Hz) velocity record, and bottom shows filtered long-period (LP) (0.01 to 0.2 Hz) velocity record. Vertical scale is shown by labeled height of line on right. The seismograms are aligned at the predicted P -wave arrival time (Ekström et al. 2003). Middle: localization of one glacial earthquake in Greenland by time-reversal method from broadband seismic records (Larmat et al. 2008). Right: source time function of glacial earthquake on the three-component derived from broadband seismic network in Greenland (Sergeant-Boy et al. 2016)



Seismology and Environment, Fig. 2 (a) The Super Typhoon Ioke in the Pacific ocean from 22 August to 7 September 2006 has been tracked by the National Hurricane Center (blue line), tracks Ioke-generated P wave sources (red) using the Southern California Seismic Network

(SCSN). Locations of P wave sources (red hexagons) and storm centers (blue circles) are marked every second day (Gerstoft and Bromirski 2016). (b) Observed (red) and modeled (black) amplitude of the P -waves generated by the Super typhoon Ioke (Farra et al. 2016)



Seismology and Environment, Fig. 3 (a) (left) Mt. Meager landslide deposit and (right) numerical simulation of the flowing mass at time 120 s (Moretti et al. 2015). The arrows represent the simulated velocities (in red), the velocity (around 40 m/s) of the center of mass (in white), and the horizontal force that the landslide applied to the ground, simulated and inverted from seismic data (in black and orange respectively). The color scale represents the thickness of the flowing mass, and the Capricorn glacier is shown in violet. (b) Sensitivity of the time evolution of the

force to the friction coefficient used in the model. The red line corresponds to the best inverted force and the pink area is the confidence interval, while the black line corresponds to the best simulated force. (c) Friction coefficient that best fits observed deposits (symbols) of landslides on Earth and other planets and that best reproduces seismic data (rectangles with dimensions corresponding to the estimated error on this friction coefficient) (Lucas et al. 2014)

intermediate scales down to the grain scale. Even if, during a landslide, the energy transmitted into seismic waves E_s is much smaller than the loss of potential energy of the granular mass ΔE_p ($E_s \approx (10^{-3}-10^{-6}) \Delta E_p$), it is high enough to be measured by seismometers (Deparis et al. 2008; Levy et al. 2015).

Extensive research has been carried out to extract detailed information on landslide characteristics from low frequency

seismic waves because their inversion makes it possible to calculate the time history of the force exerted by the landslides to the ground (Kanamori and Given 1982). This is however not possible with high frequency seismic waves because of their high sensitivity to topography and Earth properties variations during their travel from the source to the station. However, empirical relations relating landslide characteristics (volume, friction coefficient, velocity, etc.) to high frequency

seismic signals have been proposed (Hibert et al. 2011; Dammeier et al. 2011). Laboratory experiments on granular flows and generated waves have recently provided insight into the physical origin of these relations (Farin et al. 2019).

Unique strategies for landslide detection, quantification, and monitoring can be designed from seismic data analysis (Chao et al. 2017), making it possible to investigate the link between landslide activity and meteorological, climatic, seismic, or volcanic activity. As an example, based on these seismic approaches, Durand et al. (2018) showed the stability of volcanic craters can be inferred from rockfall activity and that rockfall characteristics may be an indicator of the location of future volcanic eruptions.

New Challenges

The new field of environmental seismology makes it possible to provide some key elements on the global change: For example, the increasing number of glacial earthquakes enables us the quantification of ice loss. The continuous monitoring of physical parameters of landslides and the analysis of tropical cyclone through microseisms can provide new insight in temperature, hydrology, and in processes of shallow subsurface and critical zone (thin outer layer of the Earth's surface). As these seismic signals are recorded continuously, their analysis in terms of physical processes involved, makes it possible to relate natural events' characteristics to external forcing such as meteorological (Stopa et al. 2019), climatic, volcanic, or seismic activity (e. g., Durand et al. 2018). This is particularly important nowadays as climate change strongly increases the risks associated with landslides, and glacier ice mass loss in mountainous, polar, and coastal regions (IPCC).

The distinction between anthropic seismology and environmental seismology is minor, since they are both rapidly growing fields addressing similar issues on the spatial and temporal evolution of our environment and using seismic signals previously considered as noise. Particularly challenging is the joint quantitative interpretation of high frequency and low frequency data. New efforts are made to develop methods making it possible to recover quantitative physical processes from the generated seismic waves. A key element of these efforts relies on the improvement of fluid and solid mechanical models used to interpret seismic data, providing unique opportunities to discriminate the intricate physical processes involved at the source of the waves.

Cross-References

- Body Waves
- Earth's Structure, Upper Mantle
- Free Oscillations of the Earth
- Propagation of Elastic Waves: Fundamentals

- Seismic Instrumentation
- Seismic Tomography
- Seismic, Ambient Noise Correlation
- Surface Waves
- Time-Reversal in Seismology

Bibliography

- Amundson, J. M., M. Truffer, M. P. Lüthi, M. Fahnestock, M. West, and R. J. Motyka (2008) Glacier, fjord, and seismic response to recent large calving events, Jakobshavn Isbræ, Greenland, *Geophys Res Lett*, 35(22), L22,501.
- Ardhuin F, Stutzmann E, Schimmel M, Mangeney A (2011) Ocean wave sources of seismic noise. *J Geophys Res* 115(C9):C09004
- Ardhuin F, Gualtieri L, Stutzmann E (2015) How ocean waves rock the earth: two mechanisms explain microseisms with periods 3 to 300 s. *Geophys Res Lett* 42:765–772. <https://doi.org/10.1002/2014GL062782>
- Aster RC, Winberry JP (2017) Glacial seismology. *Rep Prog Phys* 80(12):126801
- Bonnefoy-Claudet S, Cotton F, Bard P (2006) The nature of noise wavefield and its applications for site effects studies. A literature review. *Earth Sci Rev* 79:205–227
- Brodsky E, Gordeev E, Kanamori H (2003) Landslide basal friction as measured by seismic waves. *Geophys Res Lett* 30(24):2236
- Burtin A, Bollinger L, Vergne J, Cattin R, Nabelek JL (2008) Spectral analysis of seismic noise induced by rivers: a new tool to monitor spatiotemporal changes in stream hydrodynamics. *J Geophys Res* 113:B05301. <https://doi.org/10.1029/2007JB005034>
- Chao WA, Wu YM, Zhao L, Chen H, Chen YG, Chang JM, Lin CM (2017) A first near real-time seismology-based landslide monitoring system. *Sci Rep* 7:43510
- Dahl-Jensen T, Larsen TB, Voss PH, others (2010) Greenland ice sheet monitoring network (GLISN): a seismological approach. *Geol Surv Denmark Greenland Bull* 20:55–58
- Dammeier F, Moore JR, Haslinger F, Loew S (2011) Characterization of Alpine rockslides using statistical analysis of seismic signals. *J Geophys Res* 116:F04024
- Deparis J, Jongmans D, Cotton F, Baillet L, Thouvenot F, Hantz D (2008) Analysis of rock-fall and rock-fall avalanche seismograms in the French Alps. *Bull Seismol Soc Am* 98(4):1781–1796
- Durand V, Mangeney A, Haas F, Jia X, Peltier A, Hibert C, Ferrazzini V, Kowalski P, Lauret F, Brunet C, Satriano C, Wegner K, Delorme A, Bonilla F, Villeneuve N (2018) On the link between external forcings and slope instabilities in the Piton de la Fournaise summit crater, Reunion Island. *J Geophys Res Earth Surf* 123:2422–2442
- Ebeling CW, Stein S (2011) Seismological identification and characterization of a large Hurricane. *Bull Seismol Soc Am* 101(1):399–403
- Ekström G, Nettles M, Abers GA (2003) Glacial earthquakes. *Science* 302:622. <https://doi.org/10.1126/science.1088057>
- Ekström G, Nettles M, Tsai VC (2006) Seasonality and increasing frequency of greenland glacial earthquakes. *Science* 311(5768):1756–1758. <https://doi.org/10.1126/science.1122112>
- Farin M, Mangeney A, De Rosny J, Toussaint R, Trinh PT (2018) Insights into the generated seismic signal and dynamics of granular flows on horizontal and sloping beds. *J Geophys Res Earth Surf* 123(6):1407–1429
- Farin M, Mangeney A, De Rosny J, Toussaint R, Trinh PT (2019) Relations between the characteristics of granular column collapses and that of the generated high-frequency seismic signal. *J Geophys Res Earth Surf*. <https://doi.org/10.1029/2019JF005258>
- Farra V, Stutzmann E, Gualtieri L, Schimmel M, Ardhuin F (2016) Ray-theoretical modeling of secondary microseism P-waves. *Geophys J Int* 206:1730–1739. <https://doi.org/10.1093/gji/ggw242>

- Gerstoft P, Bromirski PD (2016) “Weather bomb” induced seismic signals. *Science* 353(6302):869–870
- Gimbert F, Tsai VC, Lamb MP (2014) A physical model for seismic noise generation by turbulent flow in rivers. *J Geophys Res Earth Surf* 119(10):2209–2238
- Govi M, Maraga F, Moia F (1993) Seismic detectors for continuous bed load monitoring in a gravel stream. *Hydrol Sci J* 38(2):123–132
- Gutenberg B (1958) Microseisms. *Adv Geophys* 5:53–92
- Hasselmann K (1963) A statistical analysis of the generation of microseisms. *Rev Geophys* 1(2):177–210
- Hibert C, Mangeney A, Grandjean G, Shapiro N (2011) Slopes instabilities in the Dolomieu crater, la Réunion island: from the seismic signal to the rockfalls characteristics. *J Geophys Res Earth Surf* 116:F04032
- IPCC Fifth Assessment Report. www.ipcc.ch/
- Kanamori H, Given J (1982) Analysis of long-period seismic waves excited by the May 18, 1980, eruption of Mount St. Helens – a terrestrial monopole. *J Geophys Res* 87:5422–5432
- Kawakatsu H (1989) Centroid single force inversion of seismic waves generated by landslides. *J Geophys Res Solid Earth* 94(B9):12363–12374
- Kedar S, Longuet-Higgins M, Webb F, Graham N, Clayton R, Jones C (2008) The origin of deep ocean microseisms in the North Atlantic Ocean. *Proc R Soc A* 464(2091):777–793
- Kobayashi N, Nishida k (1998) Continuous excitation of free oscillations by atmospheric disturbance. *Nature* 395:357–360
- Larmat C, Montagner J-P, Fink M, Capdeville Y, Tourin A, Clévéde E (2006) Time-reversal imaging of seismic sources and application to the great Sumatra Earthquake. *Geophys Res Lett* 33. <https://doi.org/10.1029/2006GL026336>
- Larmat C, Tromp J, Liu Q, Montagner J-P (2008) Time reversal location of glacial earthquakes. *J Geophys Res* 113:B09314
- Larose E, Carrière S, Voisin C, Bottelin P, Baillet L, Guéguen P, Walter F, Jongmans D, Guillier B, Garambois S, Gimbert F, Massey C (2015) Environmental seismology: what can we learn on earth surface processes with ambient noise? *J Appl Geophys* 116:62–74
- Lawrence WST, Williams TR (1976) Seismic signals associated with avalanches. *J Glaciol* 17(77):521–526
- Levy C, Mangeney A, Bonilla F, Hibert C, Calder E, Smith PJ (2015) Friction weakening in granular flows deduced from seismic records at the Soufrière Hills Volcano, Montserrat. *J Geophys Res Solid Earth* 120(11):7536–7557
- Longuet-Higgins MS (1950) A theory of the origin of microseisms. *Phil Trans R Soc A* 243(857):1–3
- Lucas A, Mangeney A, Ampuero JP (2014) Frictional weakening in landslides on Earth and on other planetary bodies. *Nature Commun* 5(3417)
- Meschede M, Stutzmann E, Schimmel M (2019) Blind source separation of temporally independent microseisms. *Geophys J Int* 216:1260–1275. <https://doi.org/10.1093/gji/ggy437>
- Moretti L, Mangeney A, Capdeville Y, Stutzmann E, Christian Huggel C, Schneider D, Bouchut F (2012) Numerical modeling of the Mount Steller landslide flow history and of the generated long period seismic waves. *Geophys Res Lett* 39:L16402
- Moretti L, Allstadt K, Mangeney A, Capdeville Y, Stutzmann E, Bouchut F (2015) Numerical modeling of the Mount Meager landslide constrained by its force history derived from seismic data. *J Geophys Res Solid Earth* 120(4):2579–2599
- Nakata N, Snieder R, Tsuji T, Larner K, Matsuoka T (2011) Shear wave imaging from traffic noise using seismic interferometry by cross-coherence. *Geophysics* 76(6):SA97–SA106
- Nishida k, Montagner J-P, Kawakatsu H (2009) Global surface wave tomography using seismichum. *Science* 326:112
- Olsen KG, Nettles M (2017) Patterns in glacial-earthquake activity around Greenland, 2011–13. *J Glaciol* 63(242):1077–1089
- Podolskiy EA, Walter F (2016) Cryoseismology. *Rev Geophys* 54(4):708–758
- Roth DL, Brodsky EE, Finnegan NJ, Rickenmann D, Turowski JM, Badoux A (2016) Bed load sediment transport inferred from seismic signals near a river. *J Geophys Res Earth Surf* 121(4):725–747
- Sergeant A, Mangeney A, Yastrebov V, Walter F, Montagner JP, Castelnaud O, Stutzmann E, Bonnet P, Ralaiaisoa JL, Bevan S, Luckman A (2019) Monitoring Greenland ice-sheet buoyancy-driven calving discharge using glacial earthquakes. *Ann Glaciol* 60(79):75–95
- Sergeant-Boy A, Mangeney A, Stutzmann E, Montagner J-P, Walter F, Moretti L, Castelnaud O (2016) Complex force history of a calving-generated glacial earthquake derived from broadband seismic inversion. *Geophys Res Lett* 43(3):1055–1065
- Shapiro N, Campillo M, Stehly L, Ritzwoller MH (2005) High resolution surface-wave tomography from ambient seismic noise. *Science* 307:1615–1618
- Stopa J, Arduin F, Stutzmann E, Lecocq T (2019) Sea state trends and variability: consistency between models, altimeters, buoys, and seismic data (1979–2016). *J Geophys Res Oceans*. <https://doi.org/10.1029/2018JC014607>
- Stutzmann E, Arduin F, Schimmel M, Mangeney A, Patau G (2012) Modelling long-term seismic noise in various environments. *Geophys J Int* 191(2):707–722. <https://doi.org/10.1111/j.1365-246X.2012.05638.x>
- Suda N, Nawa K, Fukao Y (1998) Earth’s background free oscillations. *Science* 279:2089–2091
- Tsai VC, Rice JR, Fahnestock M (2008) Possible mechanisms for glacial earthquakes. *J Geophys Res* 113(F3):F03,014. <https://doi.org/10.1029/2007JF000944>
- Tsai VC, Minchew B, Lamb MP, Ampuero J-P (2012) A physical model for seismic noise generation from sediment transport in rivers. *Geophys Res Lett* 39(2). <https://doi.org/10.1029/2011GL050255>
- Ward H, Crawford R (1966) Wind-induced vibrations and building modes. *Bull Seismol Soc Am* 56:793–813
- Wiechert E (1904) Verhandlungen der zweiten internationalen Seismologischen Konferenz. *Gerlands Beitr Geophys* 2:41–43
- Wiens DA, Anandakrishnan S, Winberry JP, King MA (2008) Simultaneous teleseismic and geodetic observations of the stick-slip motion of an Antarctic ice stream. *Nature* 453:770

Seismology, Global Earthquake Model

John F. Schneider¹, Jephraim Oro¹, Anselm Smolka¹, Peter Suhadolc² and Zhongliang Wu³

¹GEM Foundation, Pavia, Italy

²Department of Geosciences, University of Trieste, Trieste, Italy

³China Earthquake Administration, Beijing, China

Definition

- | | |
|----------------|---|
| Exposure | Elements at risk; an inventory of those people or artifacts (and their characteristics) that is exposed to a hazard. |
| Seismic hazard | A potentially damaging physical event, a phenomenon induced by an earthquake that may cause loss of life or injury, property damage, social and economic disruption, or environmental degradation. Seismic hazards can be single, sequential, or combined in their origin and effects. Each hazard is |

Vulnerability	characterized by its location, intensity, frequency, and probability (https://www.unisdr.org/we/inform/terminology). The conditions determined by physical, social, economic, and environmental factors or processes, which increase the susceptibility of an individual, a community, assets, or systems to the impacts of hazards (https://www.unisdr.org/we/inform/terminology).
Seismic risk	The probability of harmful consequences, or loss of life or injury, property damage, social and economic disruption, or environmental degradation, resulting from interactions between seismic hazards and vulnerable conditions.

Introduction

The assessment and subsequent mitigation of earthquake risks is among the ultimate goals of both seismology and earthquake engineering. During the last two decades, these goals have never been more important and relevant due to rapid urban development and the rise of megacities in earthquake-prone areas which have considerably increased the seismic risk worldwide (Fig. 1).

Seismic risk models are needed for accurate assessment of risks by scientists, engineers, and practitioners, in order to promote risk reduction and mitigating actions, such as

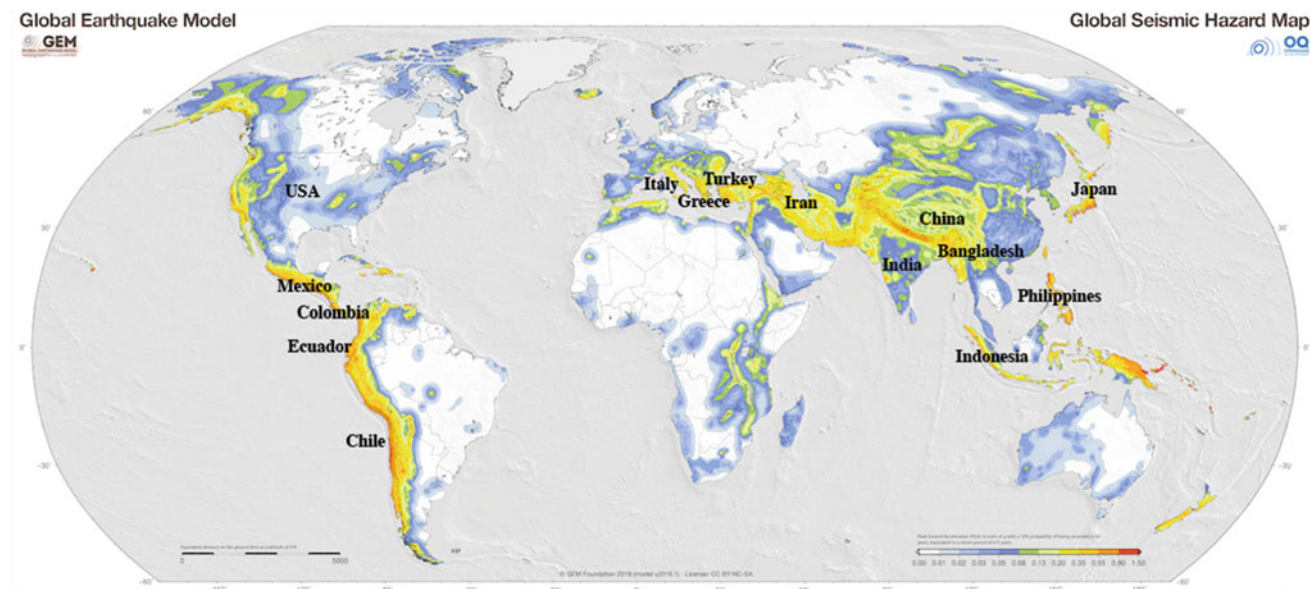
improvement of building codes and construction practices, sustainable land use, emergency response, and protection of critical infrastructures. Risk models are also essential for input to risk financing mechanisms, including catastrophe risk insurance and bonds which are needed to expedite disaster response and recovery efforts and to complement risk reduction and mitigation activities.

The GEM Initiative

In response to this challenge, the Global Earthquake Model (GEM) was put forward in 2006 during the 11th OECD Global Science Forum in Germany. Scientists and engineers realized that such an initiative was needed in order to establish an independent, uniform, and openly accessible standard to calculate and communicate seismic risk worldwide and to develop much-needed tools and software as well as databases for reliable and transparent seismic risk assessment globally. At that time, seismic risk models did not exist for most of the world, and where they did exist, they were not widely or openly accessible.

The GEM Foundation

To ensure that the scientific ambitions and the global seismic risk reduction goals of the GEM initiative are realized, an international community of scientists and professionals bonded together to form the GEM Foundation in 2009 as a nonprofit organization in Pavia, Italy, supported by a public-private sponsorship with a vision to create a world that is resilient to earthquakes.



Seismology, Global Earthquake Model, Fig. 1 Global hazard map with top 15 countries in terms of average normalized annual economic losses (AAL), 2018

In Working Program 1 (2009–2013), GEM focused: (1) on the development of a set of global databases and guidelines to support earthquake hazard and risk assessment; (2) on developing an international collaboration network of scientists and engineers across public, private, and academic sectors; and (3) on developing open earthquake hazard and risk assessment software and tools, and disseminating them globally.

In Working Program 2 (2014–2018), GEM focused on developing major application products, including a global compilation of earthquake hazard and risk models. The OpenQuake suite, launched in Jan 2015, provides free, public access to GEM's open-source earthquake hazard and risk analysis software (OpenQuake Engine) (<https://www.globalquakemodel.org/oq-get-started>), GEM's collection of national and regional-scale open-source earthquake hazard and risk models, and tools for use in building the models. In December 2018, GEM released global earthquake hazard and risk maps, which are based on a mosaic of national and regional models developed and/or implemented using the OpenQuake Engine.

In the third Working Program (2019 onwards), GEM is focused on: continuous improvement and release of the mosaic of national and regional models, including for public and commercial applications; on leveraging its expertise in earthquakes to multi-hazard risk analysis of natural hazards; on extending risk assessment information into decision-making tools for risk reduction and mitigation at national to local scale; and on capacity building assistance with emphasis on developing countries.

Scientific Framework

The GEM scientific framework serves as the underlying basis for constructing hazard and risk models, and consists of interdependent components (Fig. 2: Scientific Framework): seismic hazard; exposure; physical vulnerability;

physical seismic risk; and socioeconomic vulnerability and resilience (SVR).

Physical seismic risk is defined as a product of the following: seismic hazard (the probability of levels of ground shaking, resulting from earthquakes, within a given time span); exposure (the elements at risk – mainly buildings, critical infrastructure, and humans); and seismic vulnerability (the probability of loss given a level of ground shaking).

Earthquakes, however, have an impact that goes beyond physical damage or casualties. Earthquakes can severely damage the economy and influence society and social well-being. The addition of the SVR component aims to analyze and understand the various socioeconomic factors that may exacerbate the impact of an earthquake, including the ability of communities to fully recover.

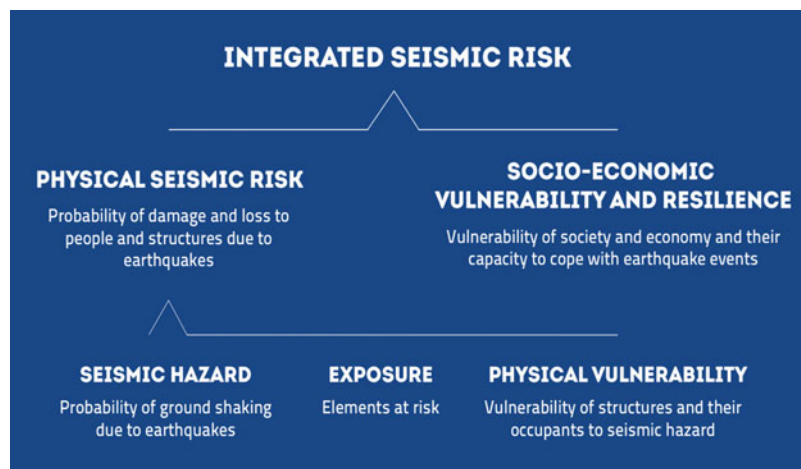
The understanding of physical seismic risk combined with socioeconomic vulnerability and resilience is what GEM calls integrated seismic risk – a holistic understanding of the effects of earthquakes that can directly support decision making on short-term needs after an event (relief, shelter); medium-term needs (recovery and reconstruction); and long-term needs related to policies on risk mitigation.

Institutional Framework

The formation of the GEM Foundation and the global collaboration network to develop the Global Earthquake Model has been a cooperative public-private initiative, composed of national governments, private organizations, research institutions, international organizations and global risk initiatives, NGOs, and individuals.

The partnership includes a number of authoritative global institutions, such as the Organisation for Economic Co-operation and Development (OECD), the International Association of Seismology and Physics of the Earth's Interior (IASPEI), the United Nations Educational, Scientific and

Seismology, Global Earthquake Model,
Fig. 2 GEM scientific framework



Cultural Organization (UNESCO), and the United Nations Office for Disaster Risk Reduction (UNDRR). It includes organizations or ministries representing many national governments, such as Australia, Canada, Great Britain, Colombia, Germany, Japan, Nepal, New Zealand, Singapore, Switzerland, and the United States, and a number of prominent private corporations from the financial, engineering, energy and risk consulting/modelling sectors, as well as academic institutions worldwide (Fig. 3).

Moreover, there are hundreds of institutions, organizations, and individuals involved in GEM that contribute expertise, data, or software, participate in regional programs, and take part in reviews and public assessments. Participation of individuals and institutions worldwide ensures that the earthquake models, information, and tools are owned and trusted by all relevant sectors, as well as global to local communities, and reflects their needs and knowledge (Fig. 4). For example, IASPEI has endorsed the GEM collaboration network and scientific products through formal resolutions in 2009 (<http://www.iaspei.org/documents/resolutions-statements>) and 2015 (<http://www.iaspei.org/documents/resolutions-statements>).

Milestones and Scientific Products

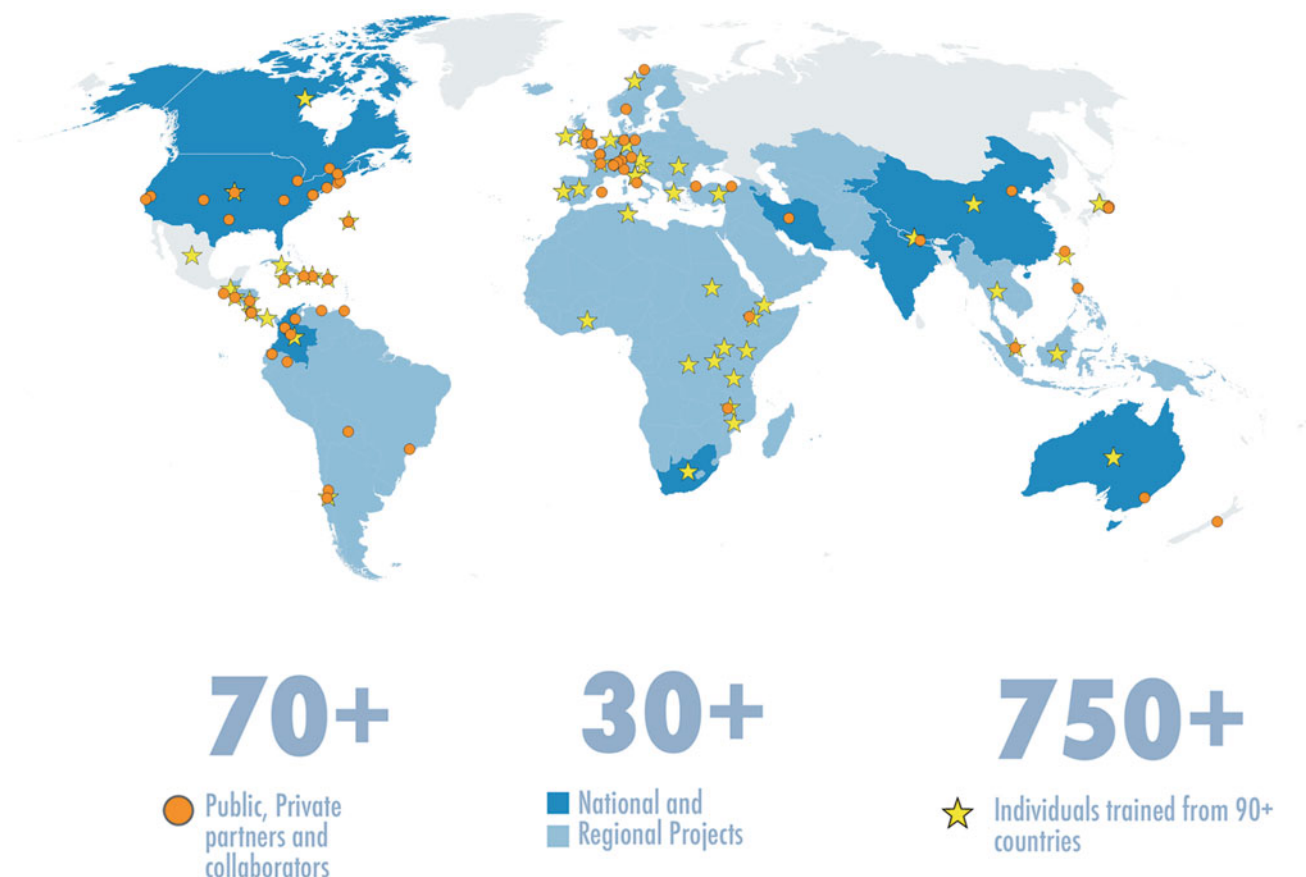
The Global Earthquake Model

The objective of developing the Global Earthquake Hazard and Risk Model was a key priority for GEM under its 2014–2018 work program. The objective is to collaboratively develop a complete set of earthquake data and models, and to deliver a comprehensive global assessment of earthquake risk. GEM released the first version of the model on December 5, 2018. The OpenQuake Engine was used to calculate all of the risk and hazard values for the global maps.

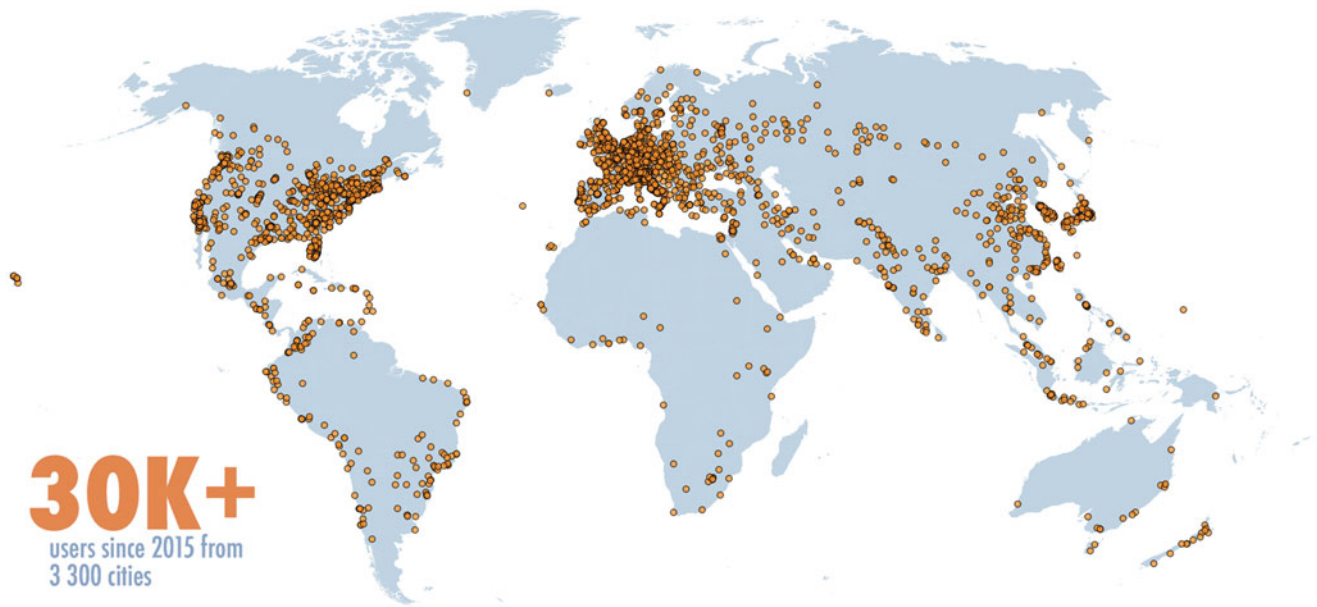
Global Earthquake Hazard

The Global Seismic Hazard Map (<https://www.globalquake-model.org/gem>) (version 2018.1) depicts the geographic distribution of the peak ground acceleration (PGA) with a 10% probability of being exceeded in 50 years, computed for reference rock conditions (shear wave velocity on the top 30 m, V_{30} , of 760–800 m/s).

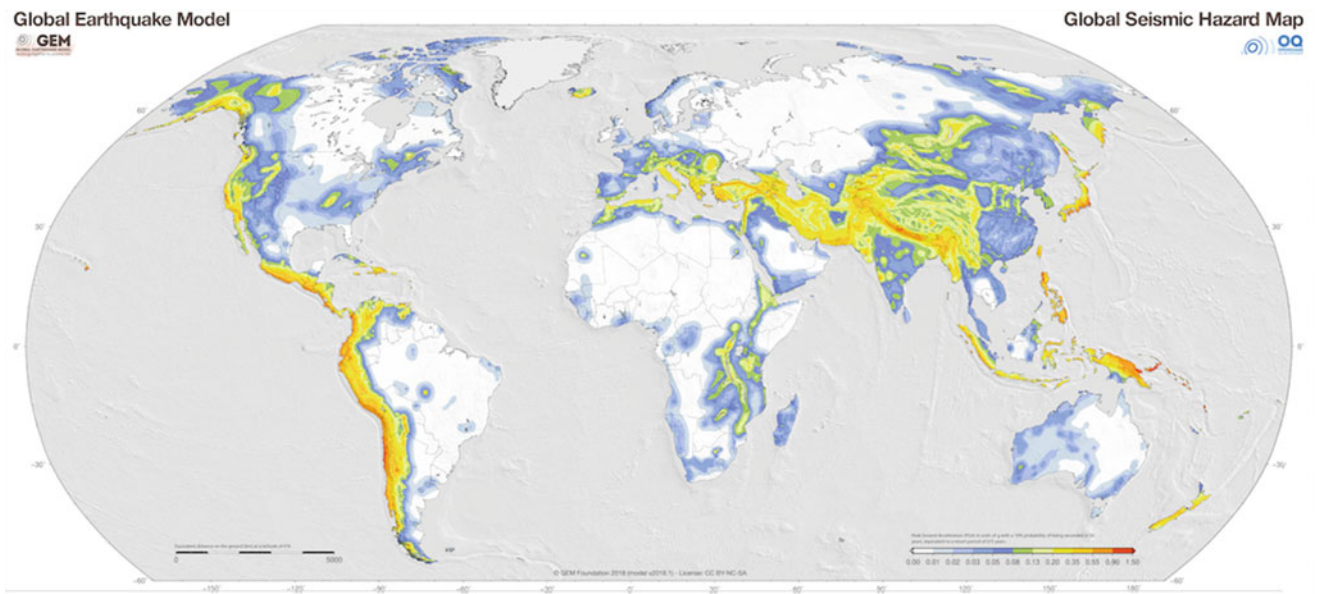
The map was created by collating maps computed using national and regional probabilistic seismic hazard models



Seismology, Global Earthquake Model, Fig. 3 GEM collaboration network, 2018



Seismology, Global Earthquake Model, Fig. 4 OpenQuake Platform users, 2018



Seismology, Global Earthquake Model, Fig. 5 Global earthquake hazard map, 2018

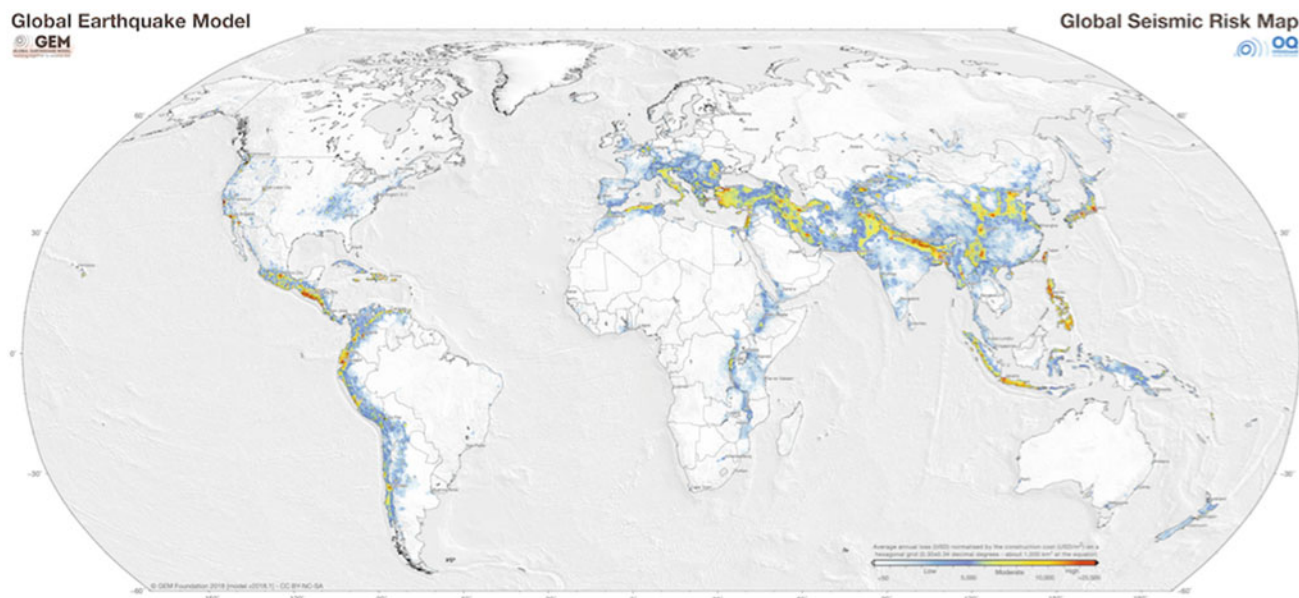
developed by various institutions and projects, and by GEM Foundation scientists (Fig. 5).

Global Earthquake Risk

The Global Seismic Risk Map (<https://www.globalquake-model.org/gem>) (v2018.1) is comprised of four global maps: global seismic hazard (PGA 10% exceedance in 50 years), global exposure, global average annual loss, and global

fatalities. The main map presents the geographic distribution of average annual loss (USD) normalized by the average construction costs of the respective country (USD/m²) due to ground shaking in the residential, commercial, and industrial building stock, considering contents and structural and nonstructural components (Fig. 6).

The seismic hazard, exposure, and vulnerability models employed in these calculations were provided by national



Seismology, Global Earthquake Model, Fig. 6 Global Earthquake Risk Map, 2018

institutions, or developed within the scope of regional programs or bilateral collaborations. These global maps and the underlying databases are based on best available and publicly accessible datasets and models.

The GEM global hazard and risk maps are dynamic products that will be updated when new datasets and models become available. Releases of updated versions of these maps are anticipated approximately every 2 years.

Technical details on the compilation of the hazard and risk maps and the underlying models are available at <http://globalquakemodel.org/gem>.

Global Social Vulnerability and Integrated Risk Map

GEM will release a global map of social vulnerability and resilience indicators in May 2020, which will complement the current hazard and risk maps. GEM continues to work in an interdisciplinary manner to integrate social, economic, and physical factors into a comprehensive risk measurement framework that takes into account quantitative, top-down, and stakeholder-based bottom-up approaches.

Software and Analysis Tools

At the core of the Global Earthquake Model is the development of state-of-the-art modeling capabilities and a suite of software tools that can be utilized worldwide for the assessment and communication of earthquake hazard and risk. Most of the tools are already embedded in the engine and the platform, but others are released as stand-alone applications.

OpenQuake Engine and Built-In Toolkits

The OpenQuake Engine (<https://github.com/gem/oq-engine#installation>) is GEM's state-of-the-art software for seismic

hazard and risk assessment at varying scales of resolution, from global to local. It can be used on a computer cluster, in the cloud or on a laptop. It is open-source, fully transparent, and can be used with GEM's or user-developed models to carry out scenario-based and probabilistic calculations and produce a great variety of outputs.

The Engine comes with the following built-in toolkits essential for model development:

- Hazard Modeller's Toolkit (<http://docs.openquake.org/oq-hazardlib/stable/openquake.hmtk.html>) – A suite of tools used to create PSHA input models, mostly developed by GEM scientists.
- Risk Input Preparation Toolkit (<https://platform.openquake.org/ipt/>) – Tools used to create exposure, fragility, and vulnerability risk input models.
- GMPE Strong Motion Modeller's Toolkit (<https://github.com/GEMScienceTools/gmpe-smtk>) – Python and OpenQuake-based toolkit for analysis of strong motions and interpretation of GMPEs.
- The Risk Modeller's Toolkit (RMtk) (<https://github.com/GEMScienceTools/rmtk>) is a library of Python scripts and IPython notebooks that provide engineers with tools to create exposure, fragility, and vulnerability input models required for running seismic risk calculations using the OpenQuake engine.

Interactive Global Earthquake Map Viewer Tool

The interactive map viewer (<https://maps.openquake.org/>) is an online tool that consists of three global maps: seismic hazard, seismic risk, and exposure. The global maps can be used to assess different levels of seismic hazard exposure

by country and territory. Users can click on a country to display the annual cost for residential, industrial, or commercial buildings and can download the full profile for any country.

OpenQuake Platform

The Platform (<https://platform.openquake.org>) is a website that allows the community to explore, manipulate, and visualize the datasets and models from the OpenQuake Engine and to use tools that GEM produces. The platform also allows users to contribute, share, and discuss new findings and results with the GEM community.

The Platform also hosts hundreds of vulnerability and fragility functions, and the Building Classification Survey site.

Integrated Risk Modeling Toolkit (IRMT)

The IRMT (<http://plugins.qgis.org/plugins/svir/version/3.7.2/>) is a QGIS plugin that was developed by GEM and co-designed by the Center for Disaster Management and Risk Reduction Technology (CEDIM). The plugin allows users to form an integrated workflow for the construction of metrics used to assess characteristics within societies that affect earthquake risk. It provides a GIS-based platform for the construction of indicators that foster comparative assessments.

Building Taxonomy

The Building Taxonomy is an online glossary that explains around 400 terms contained in the GEM Building Taxonomy (<https://www.globalquakemodel.org/publications>) v 2.0 with nearly 700 images. Users can contribute in the form of photographs or images, illustrating glossary terms.

The purpose of the GEM Building Taxonomy is to describe and classify buildings in a uniform manner as a key step toward assessing their seismic risk.

Building Classification Survey

The Building Classification Survey (<https://platform.openquake.org/building-class/>) is an online tool that allows users to share information regarding the building stock of a given region. The purpose of the tool is to create a detailed inventory of the most frequent building typologies in the world. The site is hosted by the OpenQuake Platform.

Inventory Data Capture Tools (IDCT)

IDCT is a Building Data Capture application for Android phone or tablet and is available for installation via Google Play (<https://play.google.com/store/apps/details?id=org.globalquakemodel.org.idctdo>).

Global Databases

Since 2009, GEM has compiled several global databases on seismic hazard and risk in collaboration with national, regional, and international partners.

Global Instrumental Earthquake Catalogue (1900–2009)

The GEM-ISC catalogue (<https://platform.openquake.org/maps/79>) was developed under the leadership of the International Seismological Centre (ISC). The catalogue is composed of earthquakes with homogeneous locations and magnitude estimates, determined using the same tools and techniques to the extent possible. The magnitude determination and location procedures applied represent a synthesis of state-of-the-art methods. The catalogue is maintained by the ISC (<http://www.isc.ac.uk/>).

Global Historical Earthquake Catalogue and Archive (1000–1903)

The catalogue (<https://platform.openquake.org/maps/80>) and archive provide a global starting point for continuous improvement of historical seismicity both for science and concrete disaster risk reduction (DRR) applications, especially in regional contexts.

Updated Active Fault Database

The GEM Global Active Faults (<https://blogs.openquake.org/hazard/global-active-fault-viewer/>) project is a global dataset of active faults (<https://github.com/GEMScienceTools/gem-global-active-faults>) for seismic hazard assessment as well as research, education, and general interest. This work is ongoing, and a significant portion of the world has already been covered.

Updated Vulnerability Database

The Global Vulnerability Database (<https://platform.openquake.org/vulnerability/list>) is a collection of vulnerability and fragility models for the most common building classes at a global scale. The creation of the database was supported by the United States Agency for International Development (USAID) and other GEM partners from across the globe.

The completion of over 300 functions has addressed the lack of reliable fragility and vulnerability models for many regions of the globe, which are fundamental for seismic risk assessment at various scales.

Global Exposure Database (GED4GEM and GED4ALL)

GED4ALL (<https://www.globalquakemodel.org/publications>) is an open exposure database schema for multi-hazard risk assessment. The database is capable of storing different assets (building stock, lifelines, infrastructure, crops, forestry, livestock, and socioeconomic data), while considering relevant attributes for six natural hazards: earthquakes, floods, volcanoes, strong winds, tsunamis, and drought. The GED4ALL schema is based on the level 2 schema present in GED4GEM (<https://www.globalquakemodel.org/publications>) with extensions and modifications to improve support for multiple perils.

Current Status

GEM is currently working to make GEM tools, products, and scientific knowledge developed over the last 10 years useful for a wide range of disaster risk reduction (DRR) stakeholders, i.e., governments, insurance companies, risk management agencies, NGOs, emergency responders, etc.. GEM collaborates closely with UNDRR to develop the Global Risk Assessment Framework (GRAF) as part of its commitment to develop a global, multi-hazard framework for risk assessment, including the implementation of the Sendai Framework for Disaster Risk Reduction. GEM also provides risk modeling support to the Insurance Development Forum (IDF), which seeks to build resilience and provide insurance protection to governments, institutions, and people most vulnerable to disasters.

GEM strives to deliver lasting effects on the ground in the form of reduced earthquake risk and improved earthquake risk management at local and national levels.

To date, governments from more than 30 countries are using the OpenQuake Engine and GEM products to develop or revise national hazard maps, e.g., Australia, New Zealand, Indonesia, Colombia, Italy, India, etc.. In fact, many of the models contained in the Global Earthquake Model mosaic are already used at national level to update building codes and compute risk at national level.

From the private sector, insurance and reinsurance companies are now using the OpenQuake engine, GEM resources, and technical services to enhance in-house capabilities in seismic risk assessment and model development.

Since 2015, more than 750 risk professionals from more than 90 countries have been trained in seismic hazard and risk analysis using the OpenQuake engine with its built-in toolkits.

GEM's training and capacity development program combined with multi-scale risk assessment projects have enhanced global understanding of earthquake risk; standardized and made information more accessible for disaster management; improved information for insurance pricing; and improved skills in scientific seismic risk assessment.

Future Program

GEM will continue to focus on the development of models and tools for earthquake risk assessment, and on their application at global, regional, national, and local levels. Key activities will include improving the OpenQuake engine with its built-in toolkits and databases, and strengthening our capacity building and user support program.

GEM will also extend its activities to address more complex risk issues, and will collaborate more extensively with other communities to make OpenQuake tools and models useful for multi-hazard risk assessment.

More information can be found at www.globalquake.org.

Cross-References

► Seismic Hazard

Acknowledgments The authors would like to thank Vitor Silva, Paul Henshaw, and Marco Pagani at GEM Foundation for their leadership in the development of GEM's models and tools, and for their reviews of this contribution.

Bibliography

- Bevington J, et al (2012) Exposure data development for the Global Earthquake Model: inventory data capture tools (IDCT). In: Conference paper presented at the 15th world conference on earthquake engineering, Lisbon
- Burton C, Silva V (2016) Assessing integrated earthquake risk in OpenQuake with an application to mainland Portugal. *Earthquake Spectra* 32(3):1383–1403
- Douglas J, et al (2012) Compilation and critical review of GMPEs for the GEM-PEER global GMPEs project. In: Conference paper presented at the 15th world conference on earthquake engineering, Lisbon
- Gamba P, et al (2012) The GED4GEM project: development of a global exposure database for the GEM initiative. In: Conference paper presented at the 15th world conference on earthquake engineering, Lisbon
- Pagani M et al (2014) OpensQuake engine: an open hazard (and risk) software for the Global Earthquake Model. *Seismol Res Lett* 85(3):692–702
- Pagani M et al (2015) A summary of hazard datasets and guidelines supported by the Global Earthquake Model during the first implementation phase. *INGV Ann Geophys* 58(1):S0108; <https://doi.org/10.4401/ag-6677>
- Pinho R (2012) GEM: a participatory framework for open, state-of-the-art models and tools for earthquake risk assessment. In: Conference paper presented at the 15th world conference on earthquake engineering, Lisbon
- Silva V, et al (2013) Development of the OpenQuake engine, the Global Earthquake Model's open-source software for seismic risk assessment. *Nat Hazards* 72(3):1409–1427. (2014)
- Silva V, et al (2018) GEM seismic risk model presentation, slide 85. December 5th 2018, Pavia, Italy
- Silva V, et al (2019) Assessing seismic hazard and risk globally for an earthquake resilient world. Contributing paper to the Global Assessment Report (GAR) 2019, UNDRR, Geneva
- Stewart JP, et al (2012) Selection of a global set of GMPEs for the GEM-PEER Global GMPEs project. In: Conference paper presented at the 15th world conference on earthquake engineering, Lisbon
- Vamvatsikos D, et al (2013) Seismic fragility and vulnerability assessment using simplified methods for the Global Earthquake Model. In: Conference paper presented at the 4th ECCOMAS thematic conference on computational methods in structural dynamics and earthquake engineering, Kos Island
- Witze A (2018) Earthquake-risk maps pinpoint world's most vulnerable areas. *Nature.Com*, News Section, 7 Dec 2018, correction 10 Dec 2018. <https://www.nature.com/articles/d41586-018-07705-2>
- Yepes C et al (2016) The Global Earthquake Model physical vulnerability database. *Earthquake Spectra* 32(4):2567–2585

GEM Foundation Website

<https://www.globalquakemodel.org/gem>

<https://www.globalquakemodel.org/oq-get-started>

<https://www.globalquakemodel.org/publications>

GEM OpenQuake Website

<http://docs.openquake.org/oq-hazardlib/stable/openquake.hmtk.html>
<https://blogs.openquake.org/hazard/global-active-fault-viewer/>
<https://maps.openquake.org/>
<https://platform.openquake.org/>
<https://platform.openquake.org/building-class/>
<https://platform.openquake.org/ipt/>
<https://platform.openquake.org/maps/79>
<https://platform.openquake.org/maps/80>
<https://platform.openquake.org/vulnerability/list>

Github Website

<https://github.com/gem/oq-engine#installation>
<https://github.com/gem/oq-engine/#openquake-engine>
<https://github.com/GEMScienceTools/gem-global-active-faults>
<https://github.com/GEMScienceTools/gmpe-smtk>
<https://github.com/GEMScienceTools/rmtk>

Google Apps Website

<https://play.google.com/store/apps/details?id=org.globalquakemodel.org.idctdo>

UNDRR Website

<https://www.unisdr.org/we/inform/terminology>

ISC Website

<http://www.isc.ac.uk/>

IASPEI Website

ftp://ftp.iaspei.org/pub/resolutions/resolutions_2009_cape_town.pdf
ftp://ftp.iaspei.org/pub/resolutions/resolutions_2015_prague.pdf

QGIS Website

<http://plugins.qgis.org/plugins/svir/version/3.7.2/>

Seismology, Monitoring of CTBT

Scientific and Technical Advances in Seismology and Their Relevance to CTBT

Zhongliang Wu¹ and Paul G. Richards²

¹Institute of Earthquake Forecasting, China Earthquake Administration, Beijing, China

²Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA

Definition

Monitoring of the *Comprehensive Nuclear-Test-Ban Treaty* (CTBT) entails the detection, identification, and characterization of atmospheric, underwater, and specifically underground nuclear tests, and the discrimination of nuclear explosions from other artificial and/or natural events such as quarry blasts and earthquakes (CTBTO PrepComm. 2009a).

Introduction

Matching the needs of the monitoring of CTBT deals with seismic, infrasound, and hydro-acoustic, as well as radionuclide monitoring systems. At present the CTBT verification system specified by the treaty has three major components: the International Monitoring System (IMS) with a global network of 297 certified facilities up to the time of writing/revision (CTBTO PrepComm. 2019a); the International Data Centre (IDC) for the processing of observational data; and an On-Site Inspection (OSI) regime that utilizes a series of high-resolution technologies and has many similarities to the emergency on-site observation of disaster areas such as the meizo-seismal regions of large earthquakes. In this verification system, seismic monitoring has been the leading subject, due to three aspects: (1) Since 1980 up to this time of writing/updating (2019), all nuclear tests have been conducted underground, and seismic monitoring is the most efficient way to monitor tests in that environment. (2) Development of seismic monitoring played the leading role in the development of IMS. Compared to other monitoring techniques, seismic monitoring data, characterized by the real-time transmission of broadband/high-frequency continuous waveforms, aided by the automatic/intelligent algorithms for data processing, plays a major role in the whole monitoring system. (3) Important concepts in CTBT monitoring, including evasion scenarios such as decoupling and mine-masking which are most directly a challenge to seismic monitoring, were directly or indirectly developed from seismic monitoring practice.

Entry *seismic monitoring of nuclear explosions* describes in detail the basic properties of explosion and earthquake signals and seismological technologies for the detection and identification of explosions. The main tools for such a monitoring purpose, local, regional, and global seismographic networks, are introduced in the entry *seismological networks*. Sections following focus on recent scientific and technical advances in seismology and their relevance to CTBT monitoring.

Evaluation and Design of the Seismic Monitoring System: A Systems Engineering Perspective

Monitoring and verification require an international observational system composed of the following components: observational stations, communication links from the stations to data center/s, and tools for the analysis of signals recorded for verification purposes. The evaluation and design of the seismic monitoring system from the perspective of systems engineering is one of the significant advances in recent years. It is useful not only for the functioning and sustainability of the CTBT monitoring system but also for the operation of modern seismological networks,

which not only record earthquakes but also have missions such as the fast seismological information service and the managing of huge amount of observational data. In 2009 the Project “International Scientific Studies of the Implementation of the CTBT Verification System” (ISS) was organized by the Preparatory Commission of CTBT Organization (CTBTO PrepComm. 2009b), in which *system performance evaluation* provided useful tools for the evaluation and design of the seismic monitoring system. In the following biennial Science and Technology Conference series since 2011 (CTBTO PrepComm. 2019b), *system performance evaluation* kept being one of the important themes under discussion.

Evaluation is to assess the readiness and capability of the monitoring and verification system, while design is to plan the road-map to ensure the expected readiness and capability. Design and evaluation of a seismological monitoring system includes four perspectives: (1) *physical perspective*, including primary seismographic stations and arrays, auxiliary stations, national data centers, international data center, network maintenance centers, communication links, system for the automatic data processing and review of analyst, database, and data sharing devices; (2) *functional perspective*, including data transmission, automatic data processing, automatic seismic phase picking and event location, interactive phase picking and location by analysts, array processing for location, magnitude determination, determination of earthquake parameters such as moment tensor and radiated energy, discrimination tools, and tools for routine evaluation of the state-of-health; (3) *operational perspective*, including the design, implementation, management, and upgrading of the whole system; and (4) *scientific perspective*, dealing with Earth structure model, travel time database, algorithm for location and phase picking, attenuation model and site correction for magnitude determination, Green functions or attenuation relations for source parameter determination, azimuth correction database for array detection, and event database for discrimination. It may be unnecessary for a seismologist to directly use the concepts and tools of systems engineering. However, methodologies such as specification of the system and its components as well as the relationships among the components, specification of the functions of the system and decomposition of system functional requirements into component functional requirements, and development of measures of performance at system level and component level, respectively, provide a clear concept for the systematic evaluation and design of the seismic monitoring system. This becomes especially important when dealing with a modern seismological monitoring system with many stations, working continuously in a real-time mode, being a technically complex system.

Advances in Seismology, Potentially Relevant to CTBT Monitoring

Recent decades have witnessed the fast development of seismology (*Seismological Grand Challenges* Writing Group 2008). These advances have apparent relevance to CTBT monitoring. Location of earthquake epicenters using conventional methods has the uncertainty of up to 10^1 to 10^2 km. This uncertainty is significantly reduced by the modern methods of earthquake location using waveform cross-correlation and relative location technique (Richards et al. 2006). Investigation reveals that “repeating earthquakes,” identified by waveform cross-correlation, seem to be much more common than expected in seismic activity (e.g., Schaff and Richards 2004), providing the evaluation of location accuracy with an innovative and efficient tool (e.g., Jiang et al. 2014). Recent studies show that even with a very sparse set of stations, the location precision can be obtained even at the level of tens of meters, for a cluster of events that were closely-spaced (Slinkard et al. 2016; Schaff et al. 2018). Moreover, by waveform comparison and relative location, it becomes possible for seismologists to determine the accurate location of new tests if data from previous nearby tests is available (e.g., Wen and Long 2010). As archives grow, new seismic events that appear problematic often can best be resolved by comparison of their signals with those of previously recorded events from the same region.

In the evaluation of location accuracy, one of the key concepts is the *ground truth* (GT) *events* (International Seismological Centre 2019), a database of reference earthquakes and/or explosions for which the epicenter information is known with high confidence (to ± 5 km or better) with seismic signals recorded at regional and/or teleseismic distances. GT events are used as “calibrating information” to assess the location accuracy of the monitoring system. Furthermore, a GT database is useful to facilitate better visualization of Earth structure, better modeling of velocities of seismic waves, and more accurate travel time determinations, which are all essential to the enhancement of monitoring capability. Using and developing the GT event database is thus endorsed by international seismological organizations such as the International Association of Seismology and Physics of the Earth’s Interior (IASPEI 2009).

Seismology is a practice-oriented branch of observational science, in which the automated processing of multiple seismic signals to detect and localize seismic events is a central tool. Recent development of seismology has been manifested by its fast enhancement of the capability of observation and data processing, partly due to the advances in information technology. In this field, great progresses have been made, among which, directly related to CTBT monitoring, is the NET-VISA approach to generating a bulletin of seismicity (Arora et al. 2013). It took about 10 years

(2009–2019) to bring the concept to the point of implementation. The up-to-date results showed that it not only does just about as well as is possible for a bulletin-generating process based upon use of phase picks, but also appears to perform even better than a trained analyst (CTBTO PrepComm. 2019b). Related to this field, in recent years, seismographic networks in different regions developed their own algorithms for automatic phase picking, location, magnitude determination, and moment tensor inversion, respectively. Meanwhile, machine learning (ML) has been playing an increasing role in the data-driven studies in geosciences (Bergen et al. 2019), potentially reshaping the technical skyline of seismograph networks and data centers.

Recent developments of seismology kept changing some of the conventional ideas well accepted in the CTBT monitoring practice. At present, there are several complementary methods for the identification of an explosion from an earthquake, such as the difference between m_b and M_s , and the ratio of high-frequency (>2 Hz) P to S energy (Richards and Zavales 1990; Bowers and Selby 2009). One of the important measures is the minimum size of earthquakes and/or explosions for the discrimination, called the *magnitude threshold*, below which the discrimination tool fails to work. Conventionally the *magnitude threshold* for discrimination is regarded as different from, often 0.5 magnitude unit larger than, the *magnitude threshold* for detection (Hannon 1985). However, new results (CTBTO PrepComm. 2009b) showed that these thresholds appear to be very similar if regional data, adequate for measurement of spectral ratios, is available. For purposes of estimating the location and yield of underground explosions, it is always important to have accurate information on the inhomogeneous structure of the Earth within which seismic waves propagate. New approaches to this problem using seismic noise (e.g., Shapiro et al. 2005) provide an innovative tool for imaging the Earth structure. Increasing of seismic stations and development of inversion techniques kept improving the Earth structure model (*Seismological Grand Challenges* Writing Group 2008). Accordingly, traditionally problematic earthquake parameters such as focal depth and “modern” earthquake parameters such as moment tensor and radiated energy potentially can contribute more to CTBT monitoring, along with improved 3D Earth models, and accurate account of wave propagation in complex 3D media.

With the development of digital seismic waveform analysis techniques, it has also become possible for seismologists to capture the time lapse behavior of Earth medium with considerable accuracy (e.g., Grêt and Snieder 2005). Combining with the recent advances in active source technology, this development in turn has direct impact on the OSI technology. Combination of seismic data with remote sensing images shows potential in revealing more details of the explosion source, such as decoupling (Sykes et al. 1993).

This combination, enabled by modern information technology such as “Digital Earth” (Wu and Chen 2000; International Society for Digital Earth 2019), or more practically Google Earth®, provides an alternative approach that supplements the traditional concepts of OSI and even verification.

The advancement of seismic monitoring has been tested by several cases of underground nuclear tests, such as the series of announced tests from 2006 to 2017 carried out by the Democratic People’s Republic of Korea (DPRK). Recent results from the DPRK tests show that high precision relative location methods using geodetic constraints allow the determination of nuclear test epicenters (latitude and longitude) with precision of the order of 100 m (Xie and Zhao 2018; Walter and Wen 2018); with a precise epicenter known and coupled with satellite imagery, the burial depth, an essential event parameter for yield estimate of a nuclear test, can be accurately determined. Additionally, secondary sources accompanying a test (such as a cavity collapse, and explosion-induced aftershocks) can be characterized (Han et al. 2017).

“Forensic Seismology”: Evidences and Judgments

Seismology is a branch of observational science based on the analysis of signals from different sources, penetrating through the Earth, and at last recorded by seismic networks. The field related to seismological observatory practice (Bormann 2002) and the operation of seismic networks is sometimes called *seismological observation and interpretation* (IASPEI 2019). When seismology is applied to the monitoring of CTBT, more considerations are needed beyond the recording and analysis of seismic signals. Key concepts related to the practical functioning of the seismic monitoring system are *monitoring* and *verification*. The term *monitoring* refers to technical activities and analyses associated with objective data from observational systems that acquire signals from different sources, either explosions or earthquakes, in a continuous and, in some cases, real-time, regime. The term *verification* refers to evaluations that include non-seismologists, to enable authorities and the public to judge whether some detected phenomena had its origin in a nuclear explosion. To this end, *verification* is regarded as a branch of *forensic seismology* (Bowers and Selby 2009), in which the word “forensic” means answering questions of interest to (international) legal system/s.

In concept, *monitoring* is different from *verification* in that *monitoring* basically provides objective observational evidences, while *verification* has to have some subjective decision-making processes based on the objective (but in some cases limited) evidences provided by monitoring. Examples of this distinction arise from the six nuclear tests (from 2006 to 2017) conducted and announced by the DPRK.

Seismic signals can be used to distinguish between earthquakes and underground nuclear test explosions, but not between underground nuclear and chemical explosions if, unusually, all the chemical energy is released at essentially the same time and place. For the DPRK's announced tests, all of them clearly recorded by seismic networks, direct (radionuclide) evidence was lacking for some tests, so whether they were chemical or nuclear events remains a question. Another example is the series of nuclear tests in May 1998 carried out by India and Pakistan, for which seismic monitoring data could not provide persuasive evidence that *all* the tests were conducted as successfully as announced. In fact, however, the monitoring system works well for the clarification of special events, such as the earthquake near Lop Nor of China on March 13, 2003 (Bowers and Selby 2009). In practice, an important task of the seismic monitoring system is to *screen* events of interest, to identify events that are definitely not nuclear tests. Note that in the context of forensic science, to verify that a nuclear test has been conducted somewhere and to verify that no nuclear test has been conducted in a certain place for a specific time duration, even when using the same monitoring system, need different sets of evidence from that system.

Concluding Remarks: Interaction Between Science and CTBT Monitoring

The concept of seismic monitoring of nuclear tests can be traced back to the earliest underground nuclear explosions (Bolt 1976; Richards and Kim 2009). In September, 1957, during the General Assembly of the International Union of Geodesy and Geophysics (IUGG), Toronto, K. E. Bullen gave the address "Seismology in our Atomic Age." Needs of nuclear test monitoring played an important role in the development of global seismic networks in the 1960s (Richards 2002). In the meantime, effective techniques have been developed for processing the data from an array of seismometers spread over a few kilometers (Husebye and Mykkeltveit 1981; Richards 2002), enabling the work of monitoring for explosions significantly smaller than those that can be monitored by conventional stations (Douglas 2002).

In recent years, the international CTBT monitoring system, which includes the use of multiple seismometer arrays, has been in operation as a "big science device" that is similar in scale to accelerators in high-energy physics and satellites in space science. It provides not only experiences and lessons in the practice of monitoring, but also unique datasets that are used for basic research and the science and technology for sustainability. Advances in science and technology also promote the implementation and development of CTBT monitoring practice, and assessments of the overall capability. In the above sections, these relevant

advances are introduced via three components: *engineering*, *science and technology*, and *decision-making*. Not discussed above but importantly for future, new studies of the application of fiber-optic cables (Marra et al. 2018; Jousset et al. 2018) have provided seismological monitoring with a brand new type of sensor and an expanded system of observation. The impact of the innovative technologies on CTBT monitoring is hard to predict but noticeable.

It is worth mentioning that, as indicated by a series of important events, especially the Conference "CTBT: Synergies with Science, 1996–2006 and Beyond" (CTBTO PrepComm. 2006), the ISS Project (CTBTO PrepComm. 2009b), and the Science and Technology Conference series followed (CTBTO PrepComm. 2019b), a new era of cooperation between CTBT monitoring communities and scientific research communities has started, which will in turn contribute both to the CTBT monitoring practice and to the development of geophysical science. Starting from the ISS Project, *science and technology foresight*, a professional practice that emerged in recent years, aiming at identifying today's research and innovation priorities on the basis of scenarios of future developments in science and technology, society, and economy, has been introduced to the scientific assessment of CTBT monitoring, which indicates that the interaction between science and CTBT monitoring has started to be systematic and comprehensive, through a professionally planned road-map.

Cross-References

- Seismic Monitoring of Nuclear Explosions
- Seismological Networks

Bibliography

- Arora NS, Russell S, Sudderth E (2013) NET-VISA: network processing vertically integrated seismic analysis. *Bull Seismol Soc Am* 103:709–729
- Bergen KJ, Johnson PA, de Hoop MV, Beroza GC (2019) Machine learning for data-driven discovery in solid Earth geoscience. *Science* 363:eaau0323. <https://doi.org/10.1126/science.aau0323>
- Bolt BA (1976) Nuclear explosions and earthquakes – the parted veil. W. H. Freeman and Co., San Francisco
- Bormann P (ed) (2002) IASPEI new manual of seismological observatory practice. GFZ, Potsdam
- Bowers D, Selby ND (2009) Forensic seismology and the comprehensive nuclear-test-ban treaty. *Annu Rev Earth Planet Sci* 37:209–236
- CTBTO PrepComm (2006) CTBT: synergies with science, 1996–2006 and beyond. <https://www.ctbto.org/specials/ctbt-synergies-with-science1996-2006-and-beyond/>
- CTBTO PrepComm (2009a) Verification regime. <http://www.ctbto.org/verification-regime/>
- CTBTO PrepComm (2009b) ISS 09 – international scientific studies. <http://www.ctbto.org/specials/the-international-scientific-studies-project-iss/>

- CTBTO PrepComm (2019a) International monitoring system status. <https://www.ctbto.org/>
- CTBTO PrepComm (2019b) Science and technology – the conference series. <https://www.ctbto.org/the-organization/science-and-technology-the-conference-series/>
- Douglas A (2002) Seismometer arrays – their use in earthquake and test ban seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology. Academic Press, Amsterdam, pp 357–368
- Grêt A, Snieder R (2005) Monitoring rapid temporal change in a volcano with coda wave interferometry. *Geophys Res Lett* 32: L06304. <https://doi.org/10.1029/2004GL021143>
- Han LB, Wu ZL, Jiang CS, Liu J (2017) Properties of three seismic events in September 2017 in the northern Korean Peninsula from moment tensor inversion. *Sci Bull* 62:1569–1571
- Hannon W (1985) Seismic verification of a comprehensive test ban. *Science* 227:251–257
- Husebye ES, Mykkeltveit S (eds) (1981) Identification of seismic sources – earthquake or underground explosion. D. Reidel Publishing Co., Dordrecht
- IASPEI (2009) Resolutions at the 2009 scientific assembly, Cape Town, South Africa. ftp://ftp.iaspei.org/pub/resolutions/resolutions_2009_cape_town.pdf
- IASPEI (2019) Commission on seismological observation and interpretation (CoSOI). <http://www.iaspei.org/commissions/commission-on-seismological-observation-and-interpretation>
- International Seismological Centre (2019) IASPEI Ground Truth (GT) reference events. <http://www.isc.ac.uk/gtevents/>
- International Society for Digital Earth (2019) <http://www.digitalearth-isde.org/>
- Jiang CS, Wu ZL, Li YT, Ma TF (2014) ‘Repeating events’ as estimator of location precision: the China National Seismograph Network. *Pure Appl Geophys* 171:413–423
- Jousset P, Reinsch T, Ryberg T, Blanck H, Clarke A, Aghayev R, Hersir GP, Hennings J, Weber M, Krawczyk CM (2018) Dynamic strain determination using fibre-optic cables allows imaging of seismological and structural features. *Nat Commun* 9:2509
- Marra G, Clivati C, Luckett R, Tampellini A, Kronjäger J, Wright L, Mura A, Levi F, Robinson S, Xuereb A, Baptie B (2018) Ultraprecise laser interferometry for earthquake detection with terrestrial and submarine cables. *Science* 361:486–490
- Richards PG (2002) Seismological methods of monitoring compliance with the comprehensive nuclear test ban treaty. In: WHK L, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology. Academic Press, Amsterdam, pp 369–382
- Richards PG, Kim W-Y (2009) Monitoring for nuclear explosions. *Sci Am* 300(3):64–71
- Richards PG, Zavales J (1990) Seismic discrimination of nuclear explosions. *Annu Rev Earth Planet Sci* 18:257–286
- Richards PG, Waldhauser F, Schaff D, Kim W-Y (2006) The applicability of modern methods of earthquake location. *Pure Appl Geophys* 163:351–372
- Schaff DP, Richards PG (2004) Repeating seismic events in China. *Science* 303:1176–1178
- Schaff DP, Richards PG, Slinkard M, Heck S, Young C (2018) Lg-wave cross correlation and epicentral double-difference location in and near China. *Bull Seismol Soc Am* 108:1326–1345
- Seismological Grand Challenges Writing Group (2008) Seismological grand challenges in understanding Earth’s dynamic systems, report of the long range science plan for seismology workshop, September 18–19, 2008. Incorporated Research Institutions for Seismology (IRIS), Washington, D. C. <http://www.iris.edu/hq/lrps/>
- Shapiro NM, Campillo M, Stehly L, Ritzwoller MH (2005) High resolution surface wave tomography from ambient seismic noise. *Science* 307:1615–1617
- Slinkard M, Heck S, Schaff DP, Bonal N, Daily D, Young C, Richards PG (2016) Detection of the Wenchuan aftershock sequence using waveform correlation with a composite regional network. *Bull Seismol Soc Am* 106:1371–1379
- Sykes LR, Deng J, Lyubomirskiy P (1993) Accurate location of nuclear explosions at Azgir, Kazakhstan, from satellite images and seismic data: implications for monitoring decoupled explosions. *Geophys Res Lett* 20:1919–1922
- Walter WR, Wen L (2018) Preface to the focus section on North Korea’s September 2017 nuclear test and its aftermath. *Seismol Res Lett* 89:2013–2016. <https://doi.org/10.1785/0220180281>
- Wen L, Long H (2010) High-precision location of North Korea’s 2009 nuclear test. *Seismol Res Lett* 81:26–29
- Wu ZL, Chen YT (2000) Potential impact of ‘Digital Earth’ on seismological monitoring. *Seism Res Lett* 71:548–552
- Xie X, Zhao L (2018) The seismic characterization of North Korea underground nuclear tests. *Chin J Geophys (in Chinese)* 61:889–904

Seismology, Rotational

William H. K. Lee

U.S. Geological Survey, Menlo Park, CA, USA

Definition

Rotational seismology is an emerging study of all aspects of rotational motions induced by earthquakes, explosions, and ambient vibrations. The subject is of interest to several disciplines, including seismology, earthquake engineering, geodesy, and Earth-based detection of Einstein’s gravitation waves.

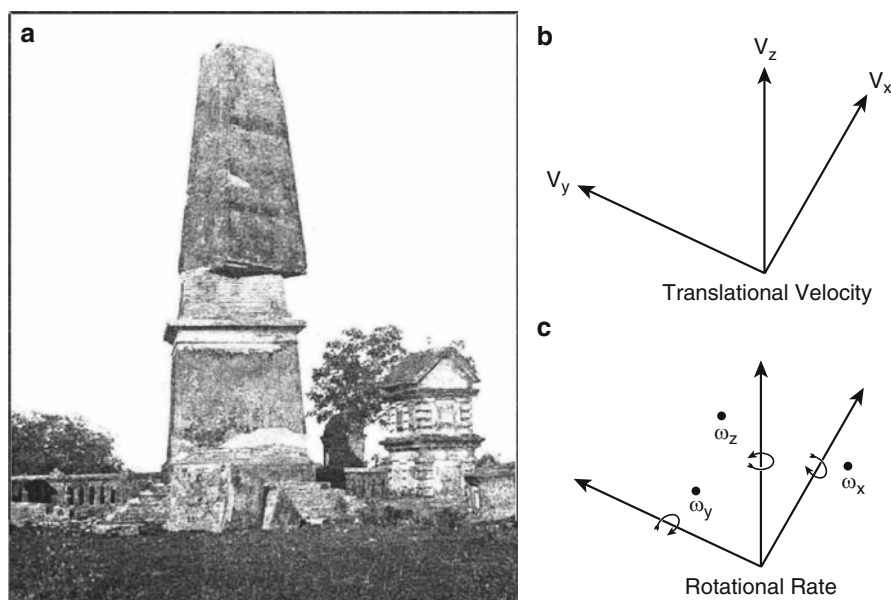
Introduction

Rotational effects of earthquake waves together with rotations caused by soil-structure interaction have been observed for centuries (e.g., rotated chimneys, monuments, and tombstones relative to their supports). A summary of historical examples of observations on earthquake rotational effects is provided by Kozák (2009), including reproduction of the relevant sections from Mallet (1862) and Reid (1910). Figure 1a shows the rotation of the monument to George Inglis (erected in 1850 at Chatak, India) as observed by Oldham (1899) after the 1897 Great Shillong earthquake. This monument had the form of an obelisk rising over 60 ft. high from a base 12 ft. on each side. During the earthquake, the topmost six-foot section was broken off and fell to the south and the next nine-foot section was thrown to the east. The remnant is about 20 ft. in height and is rotated about 15° relative to the base.

A few early authors proposed rotational waves or at least some “vortical” motions. Many different terms were used for

Seismology, Rotational,

Fig. 1 (a) Rotation of the monument to George Inglis (erected in 1850 at Chatak, India) as observed by Oldham (1899) after the 1897 Great Shillong earthquake. (b) Coordinate system for translational velocity. (c) Coordinate system for body-fixed or instantaneous rotational rate



the rotational motion components at this early stage of the field's development. For example, "rocking" is rotation around a horizontal axis, sometimes also referred to as "tilt." Mallet (1862) proposed that rotations of a body on the Earth's surface are due to a sequence of different seismic phases emerging at different angles. Reid (1910) studied this phenomenon, which was observed in the 1906 San Francisco earthquake, and pointed out that the observed rotations are too large to be produced by waves of elastic distortion. Such waves "produce very small rotations, whose maximum amount, . . . is given by the expression $2\pi A = \lambda$, where A is the amplitude and λ the wavelength; with a wave as short as 10,000 feet (3 km) and an amplitude as large as 0.2 of a foot (6 cm), the maximum rotation would only be about 0.25 of a minute of arc [0.0042 degrees, or 7.3 micro-radians (μrad)], a quantity far too small to be noticeable" (Reid 1910, p. 44). A modern analysis of such rotational effects is presented in Todorovska and Trifunac (1990).

Observational seismology is based mainly on measuring *translational* motions because of a widespread belief that *rotational* motions are insignificant. For example, Richter (1958, footnote on p. 213) states that "Theory indicates, and observation confirms, that such rotations are negligible." Richter provided no references, and there were no instruments at that time sensitive enough to either confirm or refute this claim. Recent advances in rotational seismology become possible because sensitive rotational sensors have been developed in aeronautical and astronomical instrumentation. In this article we provide some background information and selected highlights of recent advances in rotational seismology. An extensive list of references to original papers is given because few "synthesis" articles exist.

Early Attempts in Studying Rotational Motions

Ferrari (2006) summarized two models of an electrical seismograph with sliding smoked paper, developed by P. Filippo Cecchi in 1876 to record three-component *translation* motions and also the *torsion* movements from earthquakes. Although these instruments operated for several years, no rotational motion could be recorded because of low transducer sensitivity. Pioneers in several countries attempted to measure rotational motions induced by earthquakes. Nearly a century ago, Galitzin (1912) suggested using two identical pendulums installed on different sides of the same axis of rotation for separate measurement of rotational and translational motion. This was later implemented, for example, by Kharin and Simonov (1969) in an instrument designed to record strong ground motion. Using an azimuthal array of seismographs, Droste and Teisseyre (1976) derived rotational seismograms for rock bursts in a nearby mine. Inspired by Walter Munk, Farrell (1969) constructed a gyroscopic seismometer, and obtained a static displacement of <1 cm and a tilt of <0.5 μrad at La Jolla, California, during the Borrego Mountain earthquake of April 9, 1968 (magnitude 6.5), at an epicentral distance of 115 km.

Early efforts also included studies of explosions. For example, Graizer (1991) recorded tilts and translational motions in the near field of two nuclear explosions using seismological observatory sensors to measure point rotations directly. Nigbor (1994) measured rotational and translational point ground motions directly with a commercial rotational MEMS sensor and found significant near-field rotational motions (660 μrad at 1 km distance) from a one-kiloton explosion.

Rotations and strains of the ground and of response of structures have been deduced indirectly from accelerometer

arrays using methods valid for seismic waves having wavelengths that are long compared to the distances between sensors (e.g., Trifunac 1979, 1982; Oliveira and Bolt 1989; Spudich et al. 1995; Bodin et al. 1997; Huang 2003; Suryanto et al. 2006; Wassermann et al. 2009). The rotational components of ground motion have also been estimated theoretically, using kinematic source models (Bouchon and Aki 1982; Wang et al. 2009) and linear elastodynamic theory of wave propagation in elastic solids (Lee and Trifunac 1985, 1987).

Measuring Rotational Motions

The general motion of the particles or a small volume in a solid body can be divided into three parts: translation (along the X-, Y-, and Z-axes), rotation (about the X-, Y-, and Z-axes), and strain (six components). Figure 1b shows the axes in a Cartesian coordinate system for *translational velocity* measured by the usual seismometers in seismology, and Fig. 1c shows the corresponding axes of *rotation rate* measured by rotational sensors (Evans 2009). These are “body attached” coordinates, those that a seismic instrument would measure at a given instant as the sensors move and rotate through space. Converting an extended record of these body-fixed motions to recover motions in an Earth-fixed, quasi-inertial coordinate system has been performed for decades in “strapped down” inertial navigation systems such as those attached to a moving airplane. Lin et al. (2010) introduce these equations into seismology and earthquake engineering for recovering inertial-frame ground and structure motions.

Rotational ground motions can be measured directly by gyroscopic sensors or inferred indirectly from an array of translational sensors. According to Cochard et al. (2006), in a linear elastic medium the displacement $\mathbf{u}$ of a point $\mathbf{x}$ is related to a neighboring point $\mathbf{x} + \delta\mathbf{x}$ by

$$\mathbf{u}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{u}(\mathbf{x}) + \boldsymbol{\varepsilon} \delta\mathbf{x} + \boldsymbol{\omega} \times \delta\mathbf{x} \quad (1)$$

where $\boldsymbol{\varepsilon}$ is the strain tensor and

$$\boldsymbol{\omega} = 1/2 \nabla \times \mathbf{u}(\mathbf{x}) \quad (2)$$

is a pseudo-vector representing the infinitesimal angle of rigid rotation generated by the disturbance. The three components of rotation about the X-axis, Y-axis, and Z-axis are given by the following equations for such infinitesimal motions:

$$\begin{aligned} \omega_x &= 1/2 (\partial u_z / \partial y - \partial u_y / \partial z), \\ \omega_y &= 1/2 (\partial u_x / \partial z - \partial u_z / \partial x), \\ \omega_z &= 1/2 (\partial u_y / \partial x - \partial u_x / \partial y) \end{aligned} \quad (3)$$

Therefore, rigid rotations can be observed by: (1) an array of translational seismometers indirectly for “cord” rotations

associated with long wave lengths by assuming that contamination of translational signals by rotational motions is small, and that the linear elasticity theory is valid (e.g., Spudich and Fletcher 2008), or (2) rotational sensors directly for “point” body-fixed rotations (e.g., Lee et al. 2009b; Lin et al. 2010).

In the past decade, rotational motions from small local earthquakes to large teleseisms were successfully recorded by sensitive rotational sensors in several countries (e.g., Takeo 1998; McLeod et al. 1998; Igel et al. 2005, 2007; Suryanto et al. 2006; Cochard et al. 2006). In particular, the application of Sagnac interferometry in large ring laser gyros provided greatly improved sensitivity to rotations at teleseismic distances and showed that they are a good match to those estimated from linear elastic wave theory. Such motions provide additional observations that – combined with translations – lead to new approaches to the seismic inverse problem (Bernauer et al. 2009; Fichtner and Igel 2009). Recently Kurrle et al. (2010b) reported the first observations of Earth’s free oscillations using ring laser recordings, and opened up potential applications of rotational seismology at long periods.

In contrast, strong-motion observations near the source in both Japan and Taiwan showed that the amplitudes of these rotations can be one to two orders of magnitude greater than that expected from linear elasticity theory (e.g., Takeo 1998; Lee et al. 2009b). Theoretical work suggests that in shallow granular or cracked continua (e.g., weathered rock at or near Earth’s surface), asymmetries of the stress and strain fields can create rotations separate from and larger than those predicted by classical elastodynamic theory (e.g., Teisseyre et al. 2006, 2008). Unlike the traditional fault-slip model, Knopoff and Chen (2009) consider the case for faulting that takes place on a fault of finite thickness. They show that there is an additional single-couple term in the body-force equivalence and additional terms in the far-field displacement. They also show that the single-couple equivalent does not violate the principles of Newtonian mechanics because the torque imbalance in the single-couple is counterbalanced by rotations within the fault zone, with torque waves being radiated.

Large Ring Laser Gyros

An unexpected advance in studying rotational ground motions came from a different field of geophysics. Recent developments of highly sensitive ring laser gyroscopes to monitor the Earth’s rotation also yield valuable data on rotational motions from large teleseismic events. The most important property that makes such rotation sensors useful for seismology is its very low noise floor and high sensitivity to rotational motions and its insensitivity to translational and cross-rotational motions. The rotation rates expected and

observed in seismology range from the order of 10^{-1} rad/s (e.g., Nigbor 1994; Trifunac 2009) near seismic sources down to order 10^{-11} rad/s for large earthquakes at teleseismic distances (e.g., Igel et al. 2005, 2007). This range spans at least 10 orders of magnitude, 200 dB, much as for translational motions, and it is unlikely that one instrument or one instrumental technology will be capable of providing accurate measurements over such a large range of amplitudes. Ring laser technology is currently the most promising approach to recording the small rotational motions induced by teleseisms, but the primary drawback is its very high cost.

Ring lasers detect the Sagnac beat frequency of two counter-propagating laser beams (Stedman 1997; and Fig. 2c). These active interferometers generally form triangular or square closed loops several meters across and are evacuated. If this instrument is rotating on a platform with respect to inertial space, the effective cavity length between co-rotating and counter-rotating laser cavity differs and one observes frequency splitting resulting in a beat frequency. This beat frequency δf is directly proportional to the rotation rate Ω around the surface normal $\mathbf{n}$ of the ring laser system, as given by the Sagnac equation:

$$\delta f = \frac{4A}{\lambda P} \mathbf{n} \cdot \Omega, \quad (4)$$

where P is the perimeter of the instrument, A its area, and λ the laser wavelength. This equation has three contributions that influence the beat frequency δf : (1) variations in the scale factor ($4A/\lambda P$) have to be avoided by making the instrument mechanically as rigid and stable as possible, (2) changes in orientation $\mathbf{n}$ (tilting relative to Earth's rotation axis) enter the beat frequency via the inner product, and (3) variations in Ω (e.g., changes in Earth's rotation rate and seismically induced rotations). Thankfully, the dominant contribution to δf is Ω . Note that translations do not contribute to the Sagnac frequency unless they affect P or A in some indirect manner.

Ring lasers are sensitive to rotations only, assuming stable ring geometry and lasing. However, for co-seismic observations at the Earth's surface the horizontal components of rotation (i.e., tilts) will contribute to the vertical component of rotation rate. As shown by Pham et al. (2009), the tilt-coupling effect is several orders of magnitude below the level of the earthquake-induced rotational signal unless one is very close to the source (where sensitive ring lasers would not be the appropriate technology).

At present, there are ring laser gyros capable of measuring rotation (induced by small local earthquakes or distant large teleseisms) at four sites: (1) Cashmere cavern, Christchurch, New Zealand (McLeod et al. 1998); (2) Wettzell, Germany (Schreiber et al. 2005); (3) Conway, Arkansas (Dunn et al. 2009); and (4) Piñon Flat, California (Schreiber et al. 2009a).

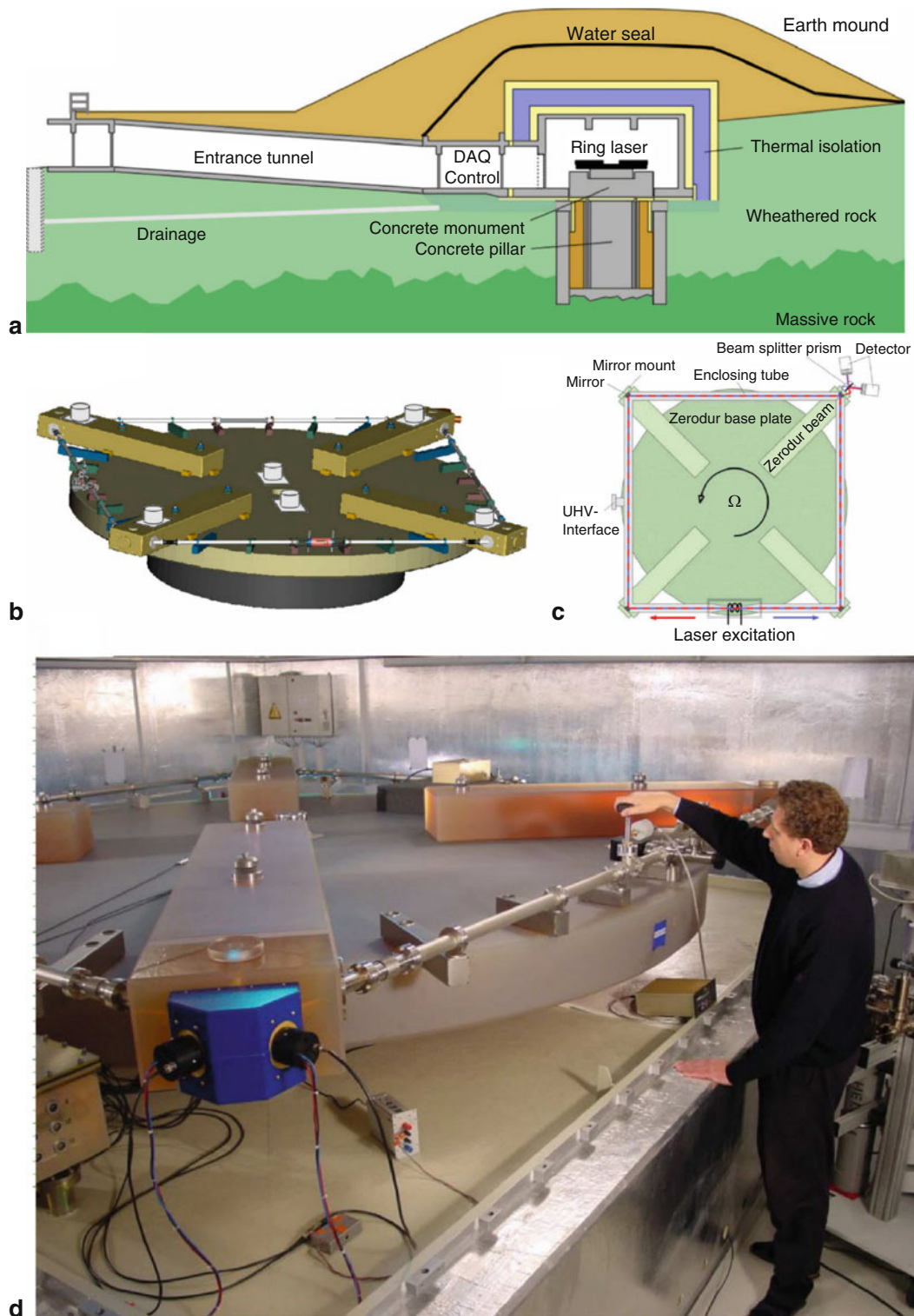
G Ring Laser and Recording Teleseisms

Since 2001, the “G Ring” laser (capable of measuring rotation rate of about 10^{-10} rad/s) has been operating at the primary geodetic station (Fundamentalstation) at Wettzell, in Bavaria, (<http://www.fs.wettzell.de/>). A cross-sectional view of the site of the G Ring laser is shown in Fig. 2a. The instrument is resting on a polished granite table (Fig. 2b) embedded in a 90-t concrete monument. As shown in Fig. 2a, the monument is attached to a massive 2.7-m diameter concrete pillar and this is founded on crystalline bedrock 10 m below. A system of concrete rings and isolation material shields the monument and pillar from adjacent weathered rock to eliminate its deformation and heat-flow contributions. The G Ring laser is protected against external influences by a subsurface installation with passive thermal stability provided by a 2-m layer alternating between Styrofoam and wet clay, this beneath a 4-m soil mound. A lateral entrance tunnel with five isolating doors and a separate control room minimize thermal perturbations during maintenance. After 2 years of thermal adaptation, the average temperature reached 12.2 °C with seasonal variations of less than 0.6 °C. Figure 2c shows the schematic drawing of instrument, and Fig. 2d is a photo of the G Ring laser with its designer, Ulli Schreiber.

Figure 3 is a comparison of direct point measurements of ground rotations around a vertical axis (red lines) to transverse accelerations (black lines, converted to rotation rate for each time window) for the M8.1 Tokachi-oki earthquake, September 25, 2003 (Igel et al. 2005). Figure 3a is a schematic view of the great-circle-path through the epicenter in Hokkaido, Japan, and the observatory in Wettzell, Germany. 3b–e show the superposition of the rotation rate derived from transverse translations (black) and measured directly (red) for various time windows: Fig. 3b the complete signal, Fig. 3c the latter part of the surface wave train, Fig. 3d the direct *S*-wave arrival, and Fig. 3e for the initial part of the surface wave train. These results confirm the expectation from linear elasticity that the waveforms of transverse acceleration and rotation rate (around the vertical axis) should be identical assuming plane harmonic waves. Information on subsurface structure is contained in the ratio between the corresponding motion amplitudes.

Strong-Motion Inertial Angular Sensors

In aerospace, automotive, and mechanical engineering, smaller rotational-motion sensors are common and generically known as gyroscopic or inertial angular sensors. Nigbor (1994) used a MEMS-based Coriolis rotation-rate sensor to measure the rotational components of strong ground motions close to a large chemical explosion. Similar sensors were used by Takeo (1998) to measure rotational motions from an earthquake swarm ~3 km away. However, such sensors do not have

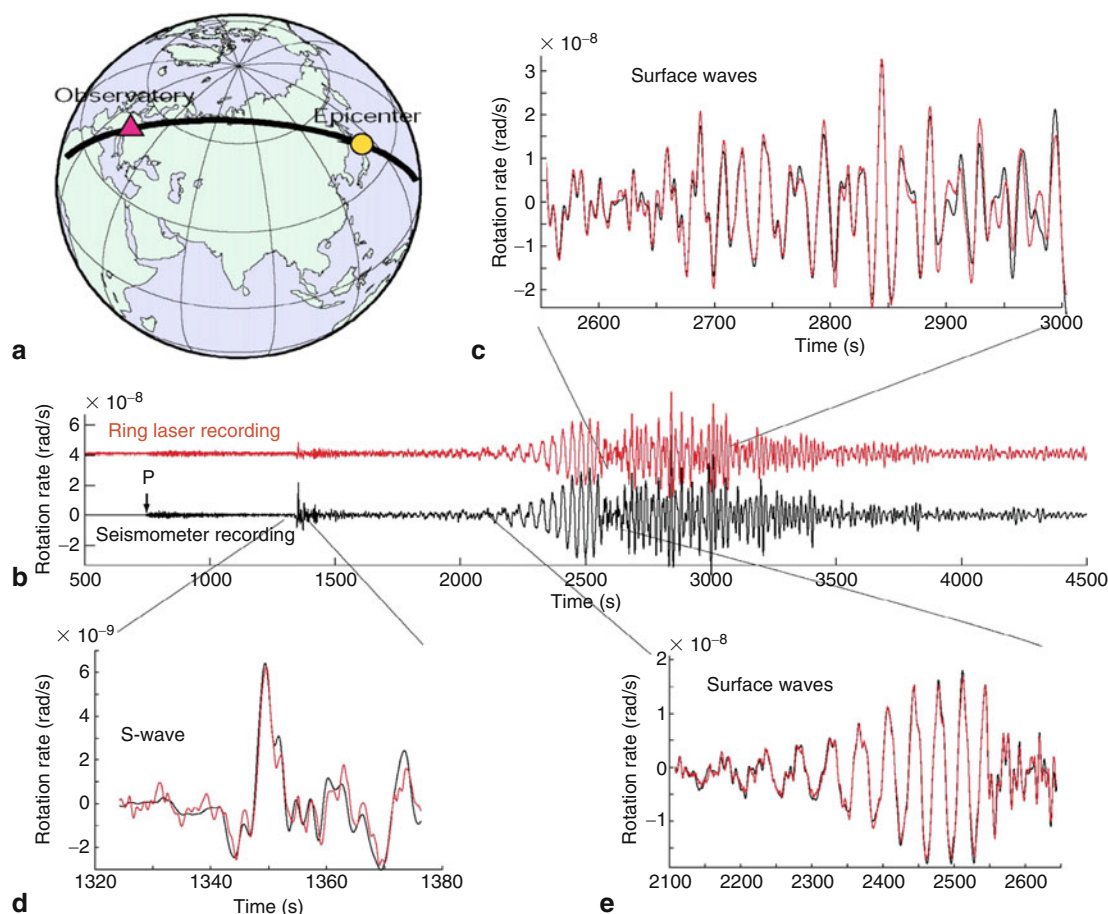


Seismology, Rotational, Fig. 2 G Ring laser gyro at the Wettzell Superstation, Germany. (a) Cross-sectional view of the instrument site. (b) Instrument resting on a granite table. (c) Schematic drawing. (d) Photo of G Ring laser gyro with its designer, Ulli Schreiber

the sensitivity to record rotations from small local earthquakes (magnitude ~ 4) at distances of tens of kilometers.

The eentecTM model R-1TM rotational seismometer is the first modestly priced sensor capable of recording small earthquakes at distances up to several tens of kilometers. It

uses electrochemical technology in which the motion of an electrolytic fluid inside a torus is sensed electronically, yielding a voltage signal proportional to rotational velocity. Nigbor et al. (2009) carried out extensive tests of commercial rotational sensors and concluded that the R-1 sensor



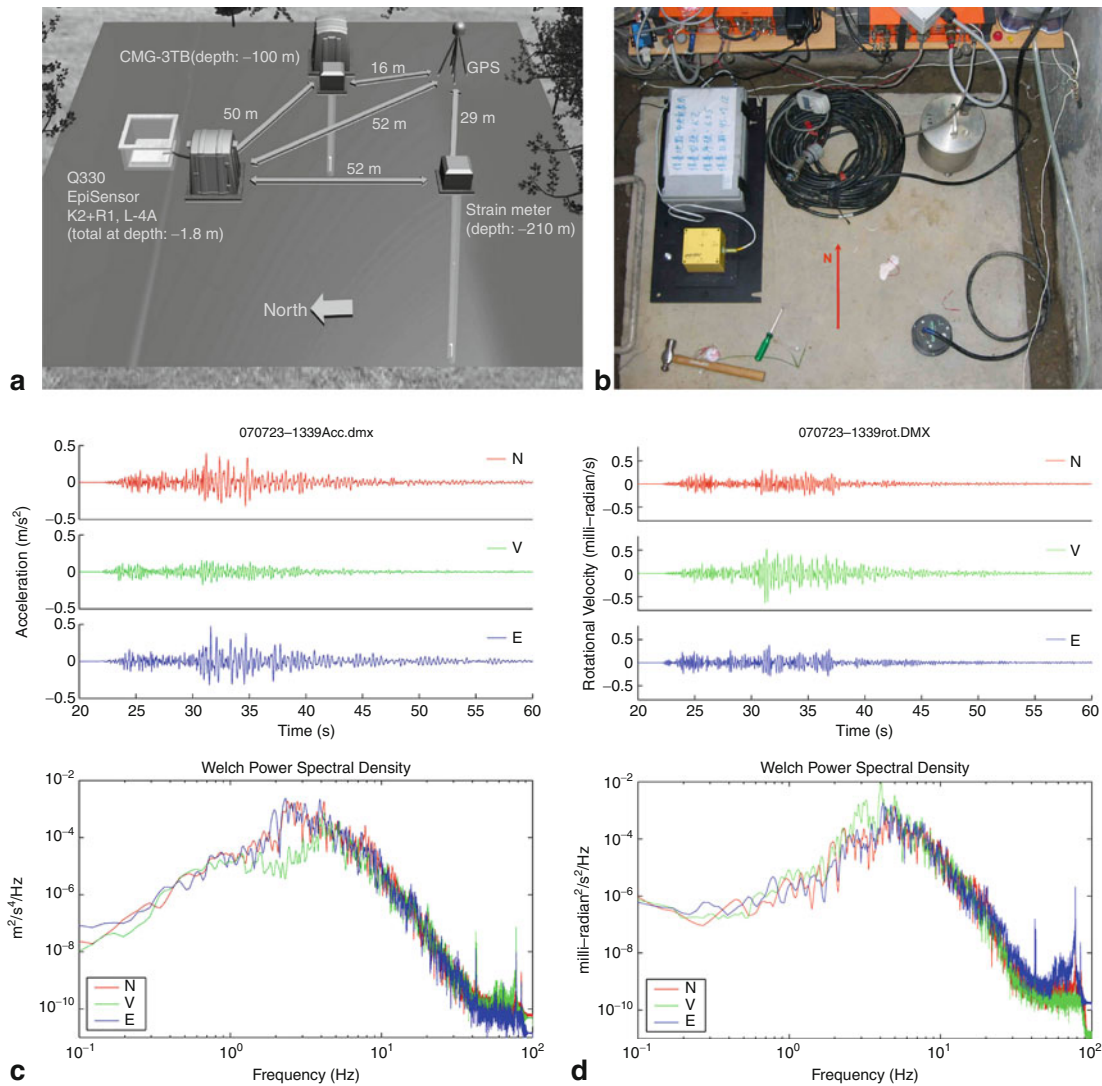
Seismology, Rotational, Fig. 3 Comparison of direct measurements of ground rotational motions around a vertical axis (red lines) with transverse accelerations (black lines, converted to rotation rate for each time window) for the M8.1 Tokachi-oki earthquake, September 25, 2003

generally meets the specifications given by the manufacturer but that clip level and frequency response vary from those specifications and between individual channels enough that more detailed calibrations are warranted for each unit. A typical transfer function for the R-1 can be found at the manufacturer's website (<http://www.eentec.com/>). The instrument response is roughly "flat" from 0.1 to 20 Hz, and its self noise (rms) is $<10 \mu\text{rad/s}$ over the same frequency band.

The R-1 rotational seismometers successfully recorded several hundred local earthquakes and two explosions in Taiwan (Lee et al. 2009b). Figure 4a shows the instruments deployed at station HGSD in eastern Taiwan (Liu et al. 2009). The top frame is a schematic drawing of the various seismic, geodetic, and strain instruments there. The bottom frame shows the subset of the instruments deployed in the shallow vault at the left hand side of the upper drawing; these include a datalogger (Quanterra Q330), an accelerometer (Kinematics Episensor), a six-channel digital accelerograph (Kinematics K2 with an external rotational seismometer, R-1 by eentec), and a short-period seismometer (Mark Products L-4A). The

K2 + R-1 instrument is at the left hand side, and the yellow-color box is the R-1 rotational seismometer.

The largest peak rotational rate recorded at the HGSD station (to early 2008) is from a M_w 5.1 earthquake at a hypocentral distance of 51 km at 13:40 UTC, 23 July 2007. Figure 4c shows the amplitudes and spectra of translational accelerations recorded by the K2's accelerometer. The peak ground acceleration was 0.47 m/s^2 , and the two horizontal components have much higher amplitude than the vertical. Figure 4d shows the amplitudes and spectra for rotational rates recorded from its external R-1 seismometer. The peak rotational rate was $0.63 \times 10^{-3} \text{ rad/s}$ for the vertical component, much more than that for the horizontal components. The spectra in Fig. 4c show that the dominant frequency band in ground acceleration is about 2–5 Hz (horizontal components) while those in Fig. 4d show that the dominant frequency band in ground rotation rate is about 2.5–5.5 Hz for the vertical component. Other studies report observations with the R-1 sensor and compare their point measurements of rotation to array-derived area rotations (e.g., Wassermann et al. 2009).



Seismology, Rotational, Fig. 4 (a) Schematic of instrument layout at station HGSD in Taiwan. (b) Photo showing the instrument vault containing a K-2 accelerograph, an R-1 rotational seismometer (yellow), an EpiSensor accelerometer, and an L-4A velocity

seismometer. (c) Recorded translational accelerations (top) and their spectra (bottom) from an M_w 5.1 earthquake 51 km from this site. (d) Recorded rotational rates (top) and their spectra (bottom) from the same earthquake

Discussions

Many authors have emphasized the benefits of studying rotational motions (e.g., Twiss et al. 1993; Spudich et al. 1995; Takeo and Ito 1997; Teisseyre et al. 2006; Trifunac 2006, 2009; Igel et al. 2007; and Fichtner and Igel 2009). We discuss some basic issues briefly.

Linear and Nonlinear Elasticity

Real materials of the Earth are heterogeneous, anisotropic, and nonlinear, especially in the damage zone surrounding faults and in poorly consolidated sediments, soil, and weathered, fractured rock just beneath seismic instruments, particularly typically installed strong-motion instruments. In the presence of significant nonlinearity we are forced to consider

the mechanics of chaos (Trifunac 2009) and to interpret such complexities must record both the rotational and translational components of strong motion.

Seismology is primarily based on the linear elasticity theory, which is applicable to simple homogeneous materials under infinitesimal strain. “Cord” rotation is defined as the curl of the displacement field in Eq. 2, and in the classical elasticity theory, the rotational components of motion are contained in the S waves. Meanwhile, continuum mechanics has advanced far beyond the classical theory. In particular, the elasticity theory of the Cosserat brothers (Cosserat and Cosserat 1909) incorporates (1) a local rotation of continuum particles as well as the translational motion assumed in classical theory, and (2) a couple stress (a torque per unit area) as well as the force stress (force per unit area). In the constitutive

equation of classical elasticity theory there are two independent elastic constants while in Cosserat elastic theory there are six or more elastic constants. Pujol (2009) provides a tutorial on rotations in the theories of finite deformation and micropolar (Cosserat) elasticity. Twiss (2009) derives an objective asymmetric micropolar moment tensor from a discrete-block model for a deforming granular material. He also investigates seismogenic deformation associated with volumes of distributed seismicity in three different geographic areas, and finds support in the micropolar model for the effects of a granular substructure on the characteristics of seismic focal mechanisms.

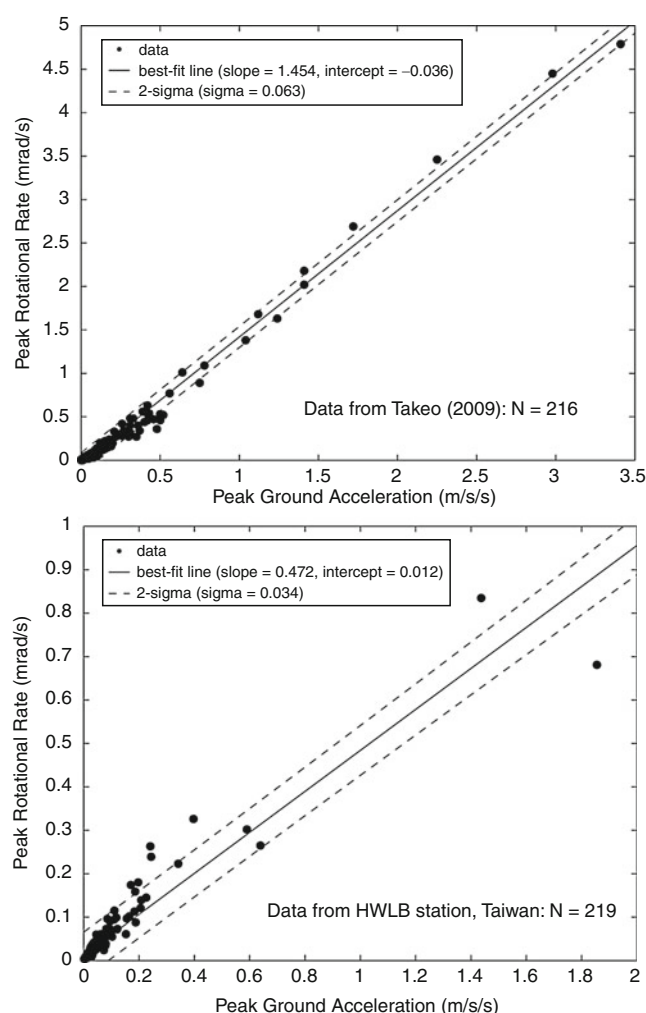
Near-Field Seismology

Although the observed rotational motions agree well with the classical elasticity theory for teleseisms, it is not the case for local earthquakes. As first noted by Takeo (1998) and confirmed by Lee et al. (2009b), observed rotational rates from local earthquakes are much larger than those predicted from the classical elasticity theory. For example, Bouchon and Aki (1982) obtained theoretically a maximum rotational rate of 1.5×10^{-3} rad/s magnitude 6.5 earthquake, whereas a maximum rotational rate of $>1 \times 10^{-3}$ rad/s for several magnitude 4.5–5.5 earthquakes had been observed in Japan and Taiwan. Takeo (1998) reported the largest rotational rate was 2.6×10^{-2} rad/s around the north-south axis from the second largest earthquake (magnitude 5.2) during the 1997 swarm (at 14:09 UTC, 3 March), east of Cape Kawana, offshore Ito, Japan (about 3 km away). As of the end of 2009, the largest rotational rate recorded at the HWLB station (Hualien, Taiwan) was 2.58×10^{-3} rad/s around the east-west axis from a M_w 6.4 earthquake offshore (at UTC 13:02, 19 December 2009) at a hypocentral distance of about 49 km. The peak rotational rate is 1.57×10^{-3} rad/s around the north-south axis, and 0.68×10^{-3} rad/s around the vertical axis. The corresponding peak ground acceleration is 1.16, 1.85, and 0.50 m/s^2 for the east-west, north-south, and vertical component, respectively.

Figure 5 shows the vertical peak rotational rate versus horizontal peak ground acceleration for the earthquake data set from Takeo (2009) (top frame), and for the earthquake data set (22 August, 2008 to 25 December, 2009) recorded at the HWLB station, Taiwan (bottom frame), where N = number of data points. As noted by Takeo (1998, 2009), there is a reasonable linear relationship in such plots. The data scatter is larger in the Taiwan data than the Takeo (2009) data. This may due to the fact that the earthquake sources for the Takeo (2009) data were from a nearby offshore swarm, whereas the Taiwan data were recorded from many different earthquake sources. Several authors noted similar linear relationships before (see e.g., Spudich and Fletcher 2008; Stupazzini et al. 2009; Takeo 2009; and Wang et al. 2009). In particular,

Takeo (2009) showed “a linear correlation between the maximum rotational displacements around vertical axis and the maximum [ground] velocities.” The two plots in Fig. 5 are equivalent to Takeo’s linear relationship, without performing the integration for the measured rotational rate and ground acceleration to obtain rotational displacements and ground velocities. The linear slope in Fig. 5 has the dimension unit of s/km, or unit for slowness. Spudich and Fletcher (2008) interpreted this “slowness” as the inverse of an “apparent velocity,” characterizing the seismic wavefield beneath the recording station.

Spudich and Fletcher (2008) computed peak values of ground strain, torsions, and tilts for the 2004 Parkfield earthquake (M_w 6.0) and four aftershocks (M_w 4.7–5.1) using the



Seismology, Rotational, Fig. 5 Vertical peak rotational rate (PRR) versus horizontal peak ground acceleration (PGA) for the earthquake data set from Takeo (2009) (top frame), and for the earthquake data set (22 August, 2008 to 25 December, 2009) recorded at the HWLB station, Taiwan (bottom frame). N number of data points

data recorded by an array of accelerators. Takeo (2009) noted that his observations of peak rotation values are about 100 times larger for earthquakes with similar magnitudes and distances, and proposed the following explanations: the different spatial scale of rotational motion by a single-point gyro measurement and by an array observation, the effect of topography, and the difference of the degree of maturation between the San Andreas fault and the swarm volume of offshore Ito. Two arrays of both rotational and translational sensors have been deployed by Wu et al. (2009) in Taiwan to study this discrepancy problem, and hopeful a resolution may be found after recording sufficient numbers of earthquakes.

Using Explosions to Study Rotational Motions

Since a large earthquake occurring near a station is rare, explosions have been used to study rotational motions by several pioneers (e.g., Graizer 1991; Nigbor 1994). Lin et al. (2009) deployed an array of 8 triaxial rotational sensors, 13 triaxial accelerometers, and 12 six-channel, 24-bit dataloggers with GPS time receivers to record two explosions in northeastern Taiwan. These instruments were installed at about 250 m (1 station), 500 m (11 stations), and 600 m (1 station) from the explosions. The 11 stations form a “Center Array” with station spacing of about 5 m. The code name for the first shot with 3000 kg explosives is “N3P,” and that for the second shot with 750 kg explosives is “N3.” Although the N3P shot used four times larger amounts of explosives than that used for the N3 shot, the peak ground translational acceleration and rotational velocity at the 13 station sites from the N3P shot are only about 1.5 times larger than that for the N3 shot. Large variations (by tens of %) of translational accelerations and rotational velocities were observed at the very small Center Array. The largest peak rotational rate was observed for the horizontal transverse component: 2.74×10^{-3} and 1.75×10^{-3} rad/s at a distance of 254 m from the N3P and N3 shots, respectively.

The acceleration data from these two explosions were used by Langston et al. (2009) to compute acceleration spatial gradients, horizontal strains and horizontal rotation, and to perform a gradiometric analysis of the strong ground motion wave train. The analysis yields a complex, frequency-dependent view of the nature of seismic wave propagation over short propagation distances that imply significant lateral velocity changes in the near-surface crustal structure. Areal strain and rotation about the vertical have equal amplitudes and suggest significant wave scattering within the confines of the river valley where the experiment was performed and/or significant departure from an axisymmetric explosion source. Gradiometry shows that the P wave arrives at the array 35° off-azimuth clockwise from

the straight-line path and appears to have been refracted from the northern side of the valley. Chi et al. (2011) successfully recovered the first order features of vertical rotational rate ground motions from the translational velocity waveforms in the 0.5–20 Hz bandwidth using the software of Spudich and Fletcher (2009), and strain as large as 10^{-4} was deduced. To fulfill the uniform rotation assumption in the linear elasticity theory, it is necessary to use a small-aperture array. However, inverting data from an array of small spatial dimension requires accurate waveforms of high signal to noise ratio and high sampling rates; since waveforms from adjacent stations are very similar, small noises have strong influence on displacement gradients due to the small station spacing. The Lin et al. (2009) recordings were limited to 200 samples per second, and much higher sampling rate will be necessary to record explosions in the near field.

Processing Collocated Measurements of Translations and Rotations

Processing collocated observations of rotation and translation is routinely performed in the inertial navigation units of aircraft and other vehicles. A similar analysis is possible for various combinations of strain components, rotations, and translations. With the exception of velocity–strain combinations (e.g., Gomberg and Agnew 1996) this terrain was largely unexplored until the work of Lin et al. (2010), who have demonstrated an appropriate set of these equations for earthquake engineering and seismology to recover inertial-frame displacements and rotations. Further, it is already apparent that rotational motions provide useful additional analysis opportunities, simply put, that more data at a site yield more results.

Phase Velocities and propagation directions. A simple calculation for non-dispersive linear-elastic plane waves with transverse polarization shows that the ratio of transverse acceleration to vertical-axis rotation rate is proportional to local phase velocity. This result implies that information on subsurface velocity structure (otherwise only accessible through seismic array measurements and combined analyses) is contained in any single-point measurement that includes rotational sensors. It has been shown that such ratio-derived phase velocities agree with velocities predicted by theory (Igel et al. 2005; Kurrle et al. 2010a). In a recent theoretical study based on full ray theory for Love waves (normal mode summation), Ferreira and Igel (2009) demonstrated that the Love wave dispersion relation also can be obtained by taking the spectral ratio of transverse acceleration to vertical-axis rotation rate. This result implies that seismic shear wave tomography is possible without requiring sub-arrays to determine local mean phase velocities. Information on the direction of propagation also is contained in the azimuth-

dependent phase fit between rotations and translations; this fit is optimal in the direction of propagation, from which back azimuths can be estimated to within a few degrees (Igel et al. 2007). Linking observational translations, strains, and rotations together also is advocated by Langston (2007) to yield a snapshot of the wavefield including wave direction, slownesses, and radial/azimuthal amplitude gradients independently at each such station.

Toward a new kind of tomography. The possibility of deriving local dispersion relations from single-station records leads to the question of what subsurface volume one resolves and to what depth velocity perturbations can be recovered. The method of choice to answer this type of question is the adjoint method (Fichtner and Igel 2009), with which sensitivity kernels (first Fresnel zones) can be calculated to indicate the volume in which the observable (typically travel times) is sensitive to structural perturbations. Fichtner and Igel (2009) introduced a new observable quantity – apparent shear wave velocity – which is a time-windowed ratio of the moduli of translational velocity and rotation angle. It turns out that the sensitivity near the source vanishes, leading to a new type of kernel that shows high sensitivity only in the vicinity of the receiver and in a somewhat smaller portion of that volume than the kernels of translational motions alone. This result implies that a tomographic inversion for near-receiver structures based on rotations and translations is possible and further highlights the potential of rotation measurements. Synthetic tomographic inversions are given in Bernauer et al. (2009).

Scattering properties of the crust: Partitioning of P and S waves. The partitioning of P and S energy and stabilizing the ratio between the two is an important constraint on the scattering properties of a medium. Igel et al. (2007) discovered surprisingly great rotational energy in a time window prior to teleseismic S , containing the P -coda. Detailed analysis of the signals and modeling of wave propagation through three-dimensional random media demonstrate that these signals can be explained with P - SH scattering in the crust with scatterers of very roughly 5-km correlation length and rms perturbation amplitude of 5% (better constrained). This result further illustrates the efficacy of rotation measurements in their own right, for example, as a filter for SH type motion, as noted by Takeo and Ito (1997).

Similar processing steps will be possible for the horizontal components of rotation and the corresponding components of translation, both to reduce the effects of tilt on the horizontal translational sensors and to extract new information. It is conceivable that the combination of these various components might lead to tight constraints on near-receiver structure, results otherwise only available from array measurements.

Conclusion

Seismology has been very successful in the *far field* because large (magnitude >6) earthquakes occur every week somewhere on Earth, and because classical elasticity theory works very well for interpreting the recorded *translational* motions at large distances. Because of this success and limited instrumentation options, most funding for earthquake monitoring historically has gone into global and regional seismic networks using only translational seismometers. However, to improve our understandings of damaging earthquakes we must also deploy rotational and translational instruments in the *near field* of active faults where potentially damaging earthquakes (magnitude >6.5) occur *infrequently*. For strong-motion seismology and engineering, this is a risky business because a damaging earthquake on any given fault may not take place for hundreds of years. Recording ground motions in the near field would require extensive seismic instrumentation along some well-chosen active faults and luck.

Ring laser observations at Wettzell, Germany, and at Piñon Flat, California, demonstrated consistent measurements of rotational ground motions in the *far field*. So far this success can only be demonstrated with one component of rotation. The high cost of present high-precision ring laser gyros makes widespread deployment unlikely. Less expensive and/or less sensitive alternatives are now being pursued by five academic groups (Cowsik et al. 2009; Dunn et al. 2009; Jedlička et al. 2009; Schreiber et al. 2009b; Takamori et al. 2009). As of the end of 2009, only Taiwan has a modest program (Lee et al. 2009b) to monitor both translational and rotational ground motions from local and regional earthquakes at several permanent seismic stations, as well as by two arrays in a building and a nearby free-field site. These two arrays are designed to “capture” a repeat of the 1906 Meishan earthquake (magnitude 7.1) in the *near field* with both translational and rotational instruments (Wu et al. 2009).

Based on the developments described in the BSSA Special Issue on rotational seismology and engineering (Lee et al. 2009a), observation, analysis, and interpretations of both rotational and translational ground motions will soon play a significant role in seismology and earthquake engineering. An international working group on rotational seismology (IWGoRS) was organized in 2006 to promote investigations of rotational motions and their implications, and for sharing experience, data, software, and results in an open Web-based environment (Todorovska et al. 2008). Anyone can join IWGoRS at <http://www.rotational-seismology.org>, subscribe to the mailing list, and contribute to the content (publications, data, links, etc.). Rotational

seismology is also of interest to physicists using Earth-based observatories for detecting Einstein's gravitational waves (e.g., Lantz et al. 2009) because they must correct for the underlying Earth motion.

Cross-References

- [Body Waves](#)
- [Earthquake Rupture: The Inverse Problem](#)
- [Earthquake, Magnitude](#)
- [Earthquakes, Strong-Ground Motion](#)
- [Great Earthquakes](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Hazard](#)
- [Seismic Instrumentation](#)
- [Seismological Networks](#)
- [Surface Waves](#)

Bibliography

- Bernauer M, Fichtner A, Igel H (2009) Inferring earth structure from combined measurements of rotational and translational ground motions. *Geophysics* 74(6):WCD41–WCD47. <https://doi.org/10.1190/1.3211110>
- Bodin P, Gomberg J, Singh SK, Santoyo M (1997) Dynamic deformations of shallow sediments in the valley of Mexico, part I: three dimensional strains and rotations recorded on a seismic array. *Bull Seismol Soc Am* 87:528–539
- Bouchon M, Aki K (1982) Strain, tilt, and rotation associated with strong ground motion in the vicinity of earthquake faults. *Bull Seismol Soc Am* 72:1717–1738
- Chi WC, Lee WHK, Lin CJ, Liu CC (2011) Inverting horizontal translational for vertical rotational ground motions: a test using data from explosions. Submitted to the *Bull Seismol Soc Am*, July, 2010
- Cochard A, Igel H, Schuberth B, Suryanto W, Velikoseltsev A, Schreiber U, Wassermann J, Scherbaum F, Vollmer D (2006) Rotational motions in seismology: theory, observation, simulation. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Heidelberg, pp 391–411
- Cosserat E, Cosserat F (1909) *Théorie des Corps Déformables*. Hermann (available from the Cornell University Library Digital Collections), Paris. (in French)
- Cowsik R, Madziwa-Nussinov T, Wagoner K, Wiens D, Wyssession M (2009) Performance characteristics of a rotational seismometer for near-field and engineering applications. *Bull Seismol Soc Am* 99(2B):1181–1189
- Droste Z, Teisseyre R (1976) Rotational and displacement components of ground motion as deduced from data of the azimuth system of seismograph. *Publ Inst Geophys Pol Acad Sci* 97:157–167
- Dunn RW, Mahdi HH, Al-Shukri HJ (2009) Design of a relatively inexpensive ring laser seismic detector. *Bull Seismol Soc Am* 99(2B):1437–1442
- Evans JR (2009) Suggested notation conventions for rotational seismology. *Bull Seismol Soc Am* 99(2B):1429–1436
- Farrell WE (1969) A gyroscope seismometer: measurements during the Borrego earthquake. *Bull Seismol Soc Am* 59:1239–1245
- Ferrari G (2006) Note on the historical rotation seismographs. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Heidelberg, pp 367–376
- Ferreira A, Igel H (2009) Rotational motions of seismic surface waves in a laterally heterogeneous Earth. *Bull Seismol Soc Am* 99(2B):1073–1075
- Fichtner A, Igel H (2009) Sensitivity densities for rotational ground-motion measurements. *Bull Seismol Soc Am* 99(2B):1302–1314
- Galitzin BB (1912) Lectures on seismometry. Russian Academy of Sciences, St. Petersburg. in Russian
- Gomberg J, Agnew D (1996) The accuracy of seismic estimates of dynamic strains: an evaluation using strainmeter and seismometer data from Piñon flat observatory, California. *Bull Seismol Soc Am* 86:212–220
- Graizer VM (1991) Inertial seismometry methods. *Izvestiya of the USSR Academy of Sciences. Phys Solid Earth* 27(1):51–61
- Huang BS (2003) Ground rotational motions of the 1991 Chi-Chi, Taiwan, earthquake as inferred from dense array observations. *Geophys Res Lett* 30(6):1307–1310
- Igel H, Schreiber U, Flaws A, Schuberth B, Velikoseltsev A, Cochard A (2005) Rotational motions induced by the M8.1 Tokachi-oki earthquake, September 25, 2003. *Geophys Res Lett* 32:L08309. <https://doi.org/10.1029/2004GL022336>
- Igel H, Cochard A, Wassermann J, Schreiber U, Velikoseltsev A, Pham ND (2007) Broadband observations of rotational ground motions. *Geophys J Int* 168:182–197
- Jedlička P, Buben J, Kozak J (2009) Strong-motion fluid rotation seismograph. *Bull Seismol Soc Am* 99(2B):1443–1448
- Kharin DA, Simonov LI (1969) VBPP seismometer for separate registration of translational motion and rotations. *Seism Instrum* 5:51–66. (in Russian)
- Knopoff L, Chen YT (2009) Single-couple component of far-field radiation from dynamical fractures. *Bull Seismol Soc Am* 99(2B):1091–1102
- Kozák JT (2009) Tutorial on earthquake rotational effects: historical examples. *Bull Seismol Soc Am* 99(2B):998–1010
- Kurle D, Igel H, Ferreira AMG, Wassermann J, Schreiber U (2010a) Can we estimate local love wave dispersion properties from collocated amplitude measurements of translations and rotations? *Geophys Res Lett* 37:L04307. <https://doi.org/10.1029/2009GL042215>
- Kurle D, Igel H, Ferreira AMG, Wassermann J, Schreiber U (2010b) First observation of rotational motions from Earth's free oscillations. *Geophys Res Abstr* 12. EGU2010-3281
- Langston CA (2007) Wave gradiometry in two dimensions. *Bull Seismol Soc Am* 97:401–416
- Langston CA, Lee WHK, Lin CJ, Liu CC (2009) Seismic-wave strain, rotation, and gradiometry for the 4 march 2008 TAIGER explosions. *Bull Seismol Soc Am* 99(2B):1287–1301
- Lantz B, Schofield R, O'Reilly B, Clark DE, DeBra D (2009) Review: requirements for a ground rotation sensor to improve advanced LIGO. *Bull Seismol Soc Am* 99(2B):980–989
- Lee VW, Trifunac MD (1985) Torsional accelerograms. *Soil Dyn Earthq Eng* 4(3):132–139
- Lee VW, Trifunac MD (1987) Rocking strong earthquake accelerations. *Soil Dyn Earthq Eng* 6(2):75–89
- Lee WHK, Celebi M, Igel H, Todorovska MI (2009a) Introduction to the special issue on rotational seismology and engineering applications. *Bull Seismol Soc Am* 99(2B):945–957

- Lee WHK, Huang BS, Langston CA, Lin CJ, Liu CC, Shin TC, Teng TL, Wu CF (2009b) Review: progress in rotational ground-motion observations from explosions and local earthquakes in Taiwan. *Bull Seismol Soc Am* 99(2B):958–967
- Lin CJ, Liu CC, Lee WHK (2009) Recording rotational and translational ground motions of two TAIGER explosions in northeastern Taiwan on 4 march 2008. *Bull Seismol Soc Am* 99(2B):1237–1250
- Lin CJ, Huang HP, Liu CC, Chiu HC (2010) Application of rotational sensors to correcting rotation-induced effects on accelerometers. *Bull Seismol Soc Am* 100:585–597
- Liu CC, Huang BS, Lee WHK, Lin CJ (2009) Observing rotational and translational ground motions at the HGSD station in Taiwan from 2004 to 2008. *Bull Seismol Soc Am* 99(2B):1228–1236
- Mallet R (1862) Great Neapolitan earthquake of 1857, vol I and II. Chapman and Hall, London
- McLeod DP, Stedman GE, Webb TH, Schreiber U (1998) Comparison of standard and ring laser rotational seismograms. *Bull Seismol Soc Am* 88:1495–1503
- Nigbor RL (1994) Six-degree-of-freedom ground motion measurement. *Bull Seismol Soc Am* 84:1665–1669
- Nigbor RL, Evans JR, Hutt CR (2009) Laboratory and field testing of commercial rotational seismometers. *Bull Seismol Soc Am* 99(2B):1215–1227
- Oldham RD (1899) Report on the great earthquake of 12th June 1897. *Mem Geol Surv India* 29:379
- Oliveira CS, Bolt BA (1989) Rotational components of surface strong ground motion. *Earthq Eng Struct Dyn* 18:517–526
- Pham DN, Igel H, Wassermann J, Cochard A, Schreiber U (2009) The effects of tilt on interferometric rotation sensors. *Bull Seismol Soc Am* 99(2B):1352–1365
- Pujol J (2009) Tutorial on rotations in the theories of finite deformation and micropolar (Cosserat) elasticity. *Bull Seismol Soc Am* 99(2B):1011–1027
- Reid HF (1910) The mechanics of the earthquake. The California earthquake of April 18, 1906, Report of the state earthquake investigation commission, vol 2. Carnegie Institution of Washington, Washington, pp 43–47
- Richter CF (1958) Elementary seismology. W. H. Freeman, San Francisco
- Schreiber KU, Igel H, Cochard A, Velikoseltsev A, Flaws A, Schuberth B, Drewitz W, Müller F (2005) The GEOSensor project: rotations – a new observable for seismology. In: Flury J, Rummel R, Reigber C, Rothacher M (eds) Observation of the earth system from space. Springer, Heidelberg, pp 1–19
- Schreiber KU, Hautmann JN, Velikoseltsev A, Wassermann J, Igel H, Otero J, Vernon F, Wells J-PR (2009a) Ring laser measurements of ground rotations for seismology. *Bull Seismol Soc Am* 99(2B):1190–1198
- Schreiber KU, Velikoseltsev A, Carr AJ, Franco-Anaya R (2009b) The application of fiber optic gyroscopes for the measurement of rotations in structural engineering. *Bull Seismol Soc Am* 99(2B):1207–1214
- Spudich P, Fletcher JB (2008) Observation and prediction of dynamic ground strains, tilts, and torsions caused by the M_w 6.0 2004 Parkfield, California, earthquake and aftershocks, derived from UPSAR array observations. *Bull Seismol Soc Am* 98:1898–1914
- Spudich P, Fletcher JB (2009) Software for inference of dynamic ground strains and rotations and their errors from short baseline array observations of ground motions. *Bull Seismol Soc Am* 99(2B):1480–1482
- Spudich P, Steck LK, Hellweg M, Fletcher JB, Baker LM (1995) Transient stresses at Parkfield, California, produced by the M 7.4 landers earthquake of June 28, 1992: observations from the UPSAR dense seismograph array. *J Geophys Res* 100(B1):675–690
- Stedman GE (1997) Ring laser tests of fundamental physics and geophysics. *Rep Prog Phys* 60:615–688
- Stupazzini M, De La Puente J, Smerzini C, Kaser M, Igel H, Castellani A (2009) Study of rotational ground motion in the near field region. *Bull Seismol Soc Am* 99(2B):1271–1286
- Suryanto W, Igel H, Wassermann J, Cochard A, Schuberth B, Vollmer D, Scherbaum F, Schreiber U, Velikoseltsev A (2006) First comparison of array-derived rotational ground motions with direct ring laser measurements. *Bull Seismol Soc Am* 96:2059–2071
- Takamori A, Araya A, Otake Y, Ishidoshio K, Ando M (2009) Research and development status of a new rotational seismometer based on the flux pinning effect of a superconductor. *Bull Seismol Soc Am* 99(2B):1174–1180
- Takeo M (1998) Ground rotational motions recorded in nearsource region. *Geophys Res Lett* 25(6):789–792
- Takeo M (2009) Rotational motions observed during an earthquake swarm in April, 1998, at offshore Ito, Japan. *Bull Seismol Soc Am* 99(2B):1457–1467
- Takeo M, Ito HM (1997) What can be learned from rotational motions excited by earthquakes? *Geophys J Int* 129:319–329
- Teisseyre R, Takeo M, Majewski E (eds) (2006) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin/Heidelberg
- Teisseyre R, Nagahama H, Majewski E (eds) (2008) Physics of asymmetric continua: extreme and fracture processes: earthquake rotation and soliton waves. Springer, Berlin/Heidelberg
- Todorovska MI, Trifunac MD (1990) Note on excitation of long structures by ground waves. *J Eng Mech-ASCE* 116(4):952–964
- Todorovska MI, Igel H, Trifunac MD, Lee WHK (2008) Rotational earthquake motions – international working group and its activities. In: Proceedings of the 14th world conference on earthquake engineering, Beijing, October 12–17, Paper ID: S03-02-0031
- Trifunac MD (1979) A note on surface strains associated with incident body waves. *Bull. Eur Assoc Earthq Eng* 5:85–95
- Trifunac MD (1982) A note on rotational components of earthquake motions on ground surface for incident body waves. *Soil Dyn Earthq Eng* 1:11–19
- Trifunac MD (2006) Effects of torsional and rocking excitations on the response of structures. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Heidelberg, pp 569–582
- Trifunac MD (2009) Earthquake engineering, nonlinear problems in. In: Meyers RA (ed) Encyclopedia of complexity and systems science. Springer, New York, pp 2421–2437
- Twiss RJ (2009) An asymmetric micropolar moment tensor derived from a discrete-block model for a rotating granular substructure. *Bull Seismol Soc Am* 99(2B):1103–1131
- Twiss R, Souter B, Unruh J (1993) The effect of block rotations on the global seismic moment tensor and patterns of seismic P and T axes. *J Geophys Res* 98:645–674
- Wang H, Igel H, Gallovic F, Cochard A (2009) Source and basin effects of rotations: comparison with translations. *Bull Seismol Soc Am* 99(2B):1162–1173
- Wassermann J, Lehndorfer S, Igel H, Schreiber U (2009) Performance test of a commercial rotational motions sensor. *Bull Seismol Soc Am* 99(2B):1449–1456
- Wu CF, Lee WHK, Huang HC (2009) Array deployment to observe rotational and translational ground motions along the Meishan fault, Taiwan: a progress report. *Bull Seismol Soc Am* 99(2B):1468–1474

Sentinel Satellites Initiate New Era in Earth Observatories

The European Space Agency's Josef Aschbacher Talks with Eos About the Sentinel Fleet of Earth Observing Satellites

Randy Showstack

American Geophysical Union, Washington, DC, USA

Independent Journalist, Washington, DC, USA

With the European Space Agency's (ESA) Sentinel-1A satellite entering into its operational phase in the summer of 2014, it ushered in a new series of Earth observation satellites. The first generation of the fleet of satellites, which will total 20 when all are launched, will comprise "the most comprehensive Earth observation system worldwide," according to Josef Aschbacher, head of program planning and coordination for ESA's Earth Observation Programs directorate. The satellites are a key part of Copernicus, the European Earth Observation Program that from 1998 to 2012 was called the Global Monitoring for Environment and Security (GMES) program.

Aschbacher spoke with *Eos* about Copernicus, which is a joint initiative of the European Commission and ESA; the Sentinel fleet of satellites; free and open data access; and the relationship between the Sentinels and some other Earth-observing satellites.

He told *Eos* that because of the resolution scale and the way the Sentinels are built, they will be used mostly for environmental applications. "Literally, it is for the benefit of mankind," he said of the Sentinel fleet. "The biggest hope is to have an everlasting observation system based on Earth observation data that provides all these observations, ad infinitum, eternally, in order to serve many disciplines on our planet, from agriculture to marine science to climate change to forestry to security to disaster management and so on."

The Copernicus program includes space, services, and in-situ components, with the space component comprising about 80% of the budget, he explained. Within the space, component are the Sentinels and the use of other data from contributing missions available on the market for free or on a commercial basis.

The Sentinel Satellites

Sentinel-1A, which was launched on 3 April 2014, is a polar-orbiting radar mission with a 5-m ground range resolution and a 12-day orbit repeat cycle at the equator. Its applications

comprise land and ocean monitoring, including sea ice detection, oil spill monitoring, ship detection and tracking, and flood mapping. Even with the satellite still in its commissioning phase, it already was useful for some important monitoring, including providing data that helped map floods in Bosnia and Herzegovina in May. With its twin, Sentinel-1B, the pair has a 6-day repeat cycle.

The Sentinel-1 family helps ensure continuity with ESA's Envisat satellite, which had a 35-day repeat cycle and stopped working in April 2012. Envisat's failure at the end of 10 years – twice its nominal life – "was a bit of a shock for us. It certainly put urgency on the Sentinel development," in particular Sentinel-1 with its critical radar capability, Aschbacher said.

Other members of the fleet, some already launched, include Sentinel-2A and -2B, a high-resolution optical mission for land monitoring; Sentinel-3A and -3B, a medium-resolution imaging and altimetry mission for marine observation; Sentinel-4A and -4B, a geostationary atmospheric chemistry mission; Sentinel-5A, -5B, and -5C, a low Earth orbit atmospheric chemistry mission; and Sentinel-6A and -6B, an altimetry mission. Also, Sentinel-5p is a low Earth orbit atmospheric chemistry precursor mission. In addition, several Sentinels will have recurrent units.

"This really is a suite of sensors in one programmatic context – Copernicus – which doesn't exist anywhere else in that magnitude covering the major domains in terms of instruments and sensors," Aschbacher said. "We are starting to develop a long-term operational program with satellites that are providing measurements for decades from now, up to 2030."

Sentinels and Landsat

Aschbacher discussed the Copernicus program and the Sentinels in the context of other Earth observation systems. He said the program has "extremely good cooperation" with US federal agencies involved with distributing Landsat data. "The plan is that we have both science groups and technical groups to work on the complementarity of Landsat and Sentinel-2 data, for example – because Sentinel-2 and Landsat are quite similar in terms of sensors, resolutions, and purpose."

Aschbacher said that with Landsat and the Sentinel-2 pair working together, global coverage would be more frequent. Landsat along with Sentinel-2 "is a great combination for a number of applications, including for natural resources, agriculture, and forestry," he said.

ESA and US federal agencies involved with Landsat have cooperated for a number of years already, and "we are

working now more intensely on how this can be better exploited, because this would be fantastic to better use Landsat together with Sentinel-2,” Aschbacher said.

The United States is interested in access to data from other Sentinel satellites as well, he added, noting that there are discussions regarding how to better exchange data and allow ESA and US federal agencies involved with Landsat each to have access to the other’s respective data.

“We are very open and very constructive in finding the best ways of giving the best access to [Sentinel] data,” Aschbacher said. “Europe has been the beneficiary of using Landsat. I would hope that in the future, also, America can be a beneficiary of using our Sentinels.”

Free and Open Access to Data

A policy of free and open access to environmental data – with some security and other exceptions – from the Copernicus program, including from Sentinel satellites, in some ways follows on the Landsat open data policy, Aschbacher said. The Landsat policy “was one of our main influences or inspirations for the data policy,” he stated.

With the Sentinel-1 calibration validation completed, “we expect to widely open distribution to everyone,” he said. “I expect that the free and open data will be a big change for the community, and this is not only the specialized Earth observation community, but really the IT [information technology] community at large [that] is combining different sources of information.”

A European Commission regulation on free and open Sentinel data, which came into force in December 2013, noted that the data could help users “to capitalize on the social benefits arising from an increased use” of the data.

Aschbacher said that the adoption of the data policy “was quite a debate.” The debate “was mostly between some commercial data suppliers who are in the domain which is complementary to the Sentinels but maybe close enough in terms of resolution so that they feared that free data from the Sentinels would impact on their market,” he said.

Widening the Market and the Access to Data

“I think what convinced them, at the end, that a free and open data policy may, first of all, not hinder their market but maybe even be supportive to their market is, first of all, with free data – free Sentinel-2 data, for example – you really promote the use of Earth observation at large to many users [who] don’t exist today and which they may not even have access to. It’s also a campaign to widen the market and widen the access to these data,” Aschbacher said.

“Secondly, if they could offer their commercial data in combination with free data from the Sentinels, they have a product which is unique and which may be a good market entry from many segments,” he noted.

Although some companies have expressed concern about the data policy, Aschbacher stressed that commercial and Sentinel systems can work complementarily. He said that commercial systems often have a resolution of a half meter or 1 meter – much better than the Sentinel satellites – that can focus on details for city planning, security, or other applications. Sentinel and commercial systems could, he said, be used “in combination: one for the large overview and the other one really to focus on a certain area [to] zoom in and get information on the particular area of interest that you need and a much better geographic resolution.”

The European Association of Remote Sensing Companies has supported the adoption of a free and open data policy for the Sentinel data as the “most effective way to develop the downstream market (and generate pull on the upstream),” among other reasons, association secretary general Geoff Sawyer noted in a March 2014 presentation about industry’s perspective on Copernicus and big data.

These satellites are providing an enormous amount of data: Whereas Envisat provided 0.3 terabyte (TB) per day, each Sentinel-1 provide 1.8 TB/day, with Sentinel-2 s providing 1.6 TB and Sentinel-3 s providing 0.6 TB, Aschbacher said. Access to the large volume of Sentinel data “will be a bit of a hurdle at this point in time, but with the advancement of technology, this will be figured out,” he said, acknowledging an initial potential bottleneck.

Aschbacher compared the availability and amount of satellite data today with what he worked with as a doctoral student. He used data from NASA’s Nimbus satellite, launched in 1978. Accessing that data on huge magnetic tapes – for which he received special access through his professor and for which he remains grateful to NASA – was cumbersome, he said, adding, “Data were really handled like gold in my age as a student. There, you could only dream of getting free and open access to data.”

Financial Challenges

Another challenge that needed to be overcome was finances for the Copernicus program. The program currently is fully funded through 2020, he said.

Aschbacher said financial discussions were complicated by the number of parties involved, including member states of ESA and the European Union, ESA itself, the European Commission, and the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT), the latter of

which is a partner for Sentinel-3, -4, and -5 operations. He added, “To get the financing organized in times of financial crisis [was] not an easy thing.” However, he said that Copernicus’s usefulness to society and cost-benefit analyses that indicated the program’s economic value helped convince parties to fund the program.

With various hurdles to the program overcome, Aschbacher admitted that “when Sentinel-1A was launched, I had an emotional moment. I thought, goodness me, 16 years ago” – since the 1998 Baveno Manifesto that marked the start of GMES – “when we [began] putting this together, and working hard over all these years to make it happen, and despite arguments against, and despite resistance of people [who] needed to be convinced, now it flies. Of course, one has to see it through [to] the end and the whole thing has to fly as a constellation, but yes, I think it is pretty much an achievement.”

For more information, see <https://sentinel.esa.int/>.

Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting

Stuart Crampin
British Geological Survey, Murchison House, Edinburgh, UK

Synonyms

Birefringence; Double refraction

Definition

Observations of shear-wave splitting in the Earth lead to a new understanding of fluid-rock deformation: a New Geophysics, where earthquakes can be stress-forecast.

Shear-wave splitting	Seismic shear-waves propagating through effectively anisotropic solids, such as aligned crystals or aligned microcracks, split into two phases with different velocities and approximately orthogonal polarizations.
New Geophysics	A new understanding of fluid-rock deformation in distributions of stress-aligned fluid-saturated microcracks where microcracks in the crust are so closely spaced that they verge on fracturing and hence are critical-systems (also known as complex-systems). Critical-systems impose fundamentally new properties on conventional

Earthquake stress-forecasting

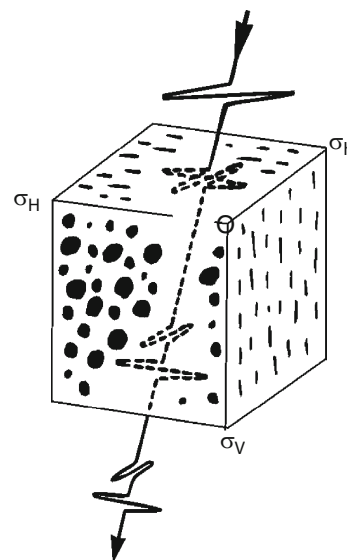
subcritical geophysics and hence are a New Geophysics.

Shear-wave splitting above swarms of small earthquakes can be used as stress-measuring stations to monitor stress-accumulation and stress-relaxation (crack-coalescence) before earthquakes, and hence *stress-forecast* the time, magnitude, and, in some cases, fault-break of impending large earthquakes.

Introduction

Transversely polarized seismic shear-waves propagating through in situ rocks with some form of elastic anisotropy, such as aligned microcracks, split into two orthogonal polarizations which propagate with different velocities, and hence lead to shear-wave splitting. The polarizations are strictly orthogonal for phase-velocity propagation, and approximately orthogonal for group-velocity propagation. Figure 1 is a schematic illustration of shear-wave splitting in the parallel vertical stress-aligned microcracks characteristic of most rocks in the Earth’s crust, once below near-surface weathering and stress-release anomalies. Originally referred to as birefringence or double refraction, shear-wave splitting was known to exist from the properties of anisotropic elastic solids (Love 1927; Nye 1957).

Double refraction has been observed experimentally as two orthogonally polarized shear-waves with different velocities in a stressed rock sample (Nur and Simmons 1969). The



Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting, Fig. 1 Schematic illustration of stress-aligned shear-wave splitting through parallel vertical microcracks (after Crampin 1994). σ_v , σ_H , and σ_h are principal axes of vertical stress, and maximum and minimum horizontal stress, respectively

phenomenon of shear-wave splitting was first recognized (and named) in synthetic seismograms in an anisotropic crystalline Earth model (Keith and Crampin 1977) and in distributions of parallel vertical microcracks (Crampin 1978). Shear-wave splitting was first positively identified in the field (above swarms of small earthquakes) by Crampin et al. (1980) and in seismic reflection surveys in hydrocarbon reservoirs by Alford (1986). An earlier attempt by Gupta (1973) to demonstrate double refraction in shear-waves some 100 km from small earthquakes failed because of misunderstanding the behavior of shear-wave splitting and the shear-wave window (Crampin et al. 1981).

Although variation of velocity with direction is certainly characteristic of seismic propagation in anisotropic media, shear-wave splitting (or the generalized three-dimensional coupling of surface-wave motion), easily identifiable in particle-motion diagrams (hodograms), is highly diagnostic of some form of effective anisotropy. One of the major difficulties is that observations of shear-wave splitting are path-integration phenomena that give no indication of where the anisotropy is located along the ray path, nor its extent or strength. The only exception is in sedimentary basins where the processing seismic surveys in exploration seismology may locate the anisotropy more accurately (for example, in Angerer et al. 2002).

Note that shear-wave splitting, as in Fig. 1, seems quite simple in concept which may tempt authors, such as Gupta (1973), to make unfounded assumptions which lead to incorrect and misleading conclusions. Shear-wave splitting is comparatively straightforward, but requires a jump in comprehension from previous isotropic experience. Two early comprehensive reviews (Crampin 1977, 1981) outlined seismic-wave propagation in anisotropic and cracked elastic media. Crampin and Peacock (2008) recently reviewed the current understanding of shear-wave splitting and identified 17 commonly assumed fallacies in understanding shear-wave splitting that can lead to unjustified conclusions.

It is suggested that the behaviour of shear-wave splitting is now comparatively well-understood. The major new advance is that observations of shear-wave splitting indicate that the stress-aligned fluid-saturated microcracks in almost all rocks are so closely spaced that they verge on failure by fracturing (and earthquakes). Such verging on failure indicates that the distributions of microcracks are critical-systems. Critical systems are a *New Physics* (Davies 1989), hence a *New Geophysics* that imposes a range of fundamentally new properties on the previous subcritical geophysics. This is a major advance in understanding fluid-rock deformation that has applications to many branches of solid-earth geophysics (Crampin and Peacock 2005, 2008; Crampin 2006).

This entry will not discuss the theory of seismic anisotropy and shear-wave splitting which is adequately outlined in

Crampin (1977, 1981), Helbig (1994), Maupin and Park (2007), and elsewhere. Here, we merely outline the behaviour of shear-wave splitting in the crust that leads to the New Geophysics and stress-forecasting earthquakes.

Terminology: references to terminology in the text will be indicated by (qv).

APE: anisotropic poro-elasticity is a model for deformation/evolution of critical-systems of distributions of stress-aligned fluid-saturated microcracks under changing conditions (Zatsepin and Crampin 1997; Crampin and Zatsepin 1997).

Aspect-ratio: Crack thickness over crack diameter.

Band-1 directions: Band-1 ray-path directions are the solid angle between 15° and 45° either side of the crack planes in a distribution of vertical parallel microcracks (Crampin 1999). APE shows that the effect of increasing (or decreasing) stress is to increase (or decrease) the average time-delays in Band-1 directions.

Band-2 directions: Band-2 ray-path directions are in the solid angle $\pm 15^\circ$ to the crack planes (Crampin 1999). Band-2 directions are sensitive to crack density, but APE shows that crack density does not vary consistently for small changes of stress.

Complex-systems: see critical-systems.

Crack density: Dimensionless crack density, CD , is specified by $CD = N a^3$ where N is number of cracks of radius a per unit cube.

Critical-systems: Critical-systems (also known as complex-systems, Davies 1989) are complex heterogeneous interactive systems which verge on fracture-criticality and failure. Extremely common in a huge range of different phenomena, critical-systems impose fundamentally new properties on the previously subcritical physics and subcritical solid-earth geophysics, and hence are a New Geophysics (Crampin 2006).

Earthquake prediction: The prediction of time, magnitude, and location of earthquakes by conventional sub-critical techniques, but not using shear-wave splitting which implies critical techniques.

Earthquake stress-forecasting: Using shear-wave splitting to monitor stress-accumulation and stress-relaxation in the surrounding rock mass that allows the time and magnitude of earthquakes to be stress-forecast, where other precursory phenomena may indicate the fault-break.

EDA: Extensive-Dilatancy Anisotropy is the name given to the distributions of fluid-saturated microcracks pervading most rocks in the crust (Crampin et al. 1984). EDA-cracks, in ostensibly unfractured rocks, are observed to have crack densities between ~ 0.015 and ~ 0.045 : that is shear-wave velocity anisotropy of $\sim 1.5 - 4.5\%$ (Crampin 1994). This comparatively small percentage of velocity anisotropy has massive implications for a huge range of phenomena and leads to the New Geophysics.

Fracture-criticality: The value of crack density, ~ 0.055 , at which stress-aligned microcracks are so closely spaced they verge on fracturing, and hence are critical-systems (qv) (Crampin 1994, 1999; Crampin and Zatsepin 1997).

New Geophysics: Shear-wave splitting indicates that microcracks in the crust are so closely spaced they verge on fracture-criticality and fracturing, and hence are critical-systems. Critical-systems are a *New Physics* (Davies 1989), hence a *New Geophysics*, that imposes a range of fundamentally new properties on conventional subcritical geophysics (Crampin 2006). Most of these properties have been observed in the field (Crampin and Peacock 2005, 2008).

PTL-anisotropy: Distributions of horizontal Periodic Thin Layers lead to a form of anisotropy (hexagonal symmetry, commonly known as transverse isotropy, with a vertical axis of symmetry) that is common in sedimentary basins, including many hydrocarbon reservoirs (Wild and Crampin 1991). Shear-wave velocity anisotropy in PTL-anisotropy may exceed 30% and would be likely to generate cusps in shear-wave velocity sheets. Such cusps have been observed (and modeled) in hydrocarbon reservoirs in the field (Slater et al. 1993).

Seismic anisotropy: Seismic anisotropy describes an elastic solid; these properties vary with direction as in aligned crystals or solids pervaded by aligned microcracks.

Seismic isotropy: Isotropy describes an elastic solid which has the same elastic properties in all directions.

Stress-forecasting earthquakes: see Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting.

TIH- (or HTI-) anisotropy: Transverse isotropy with a horizontal axis of cylindrical symmetry: the typical symmetry of EDA-cracks. Note that although transverse isotropy is specified by five elastic constants, distributions of vertical parallel EDA-cracks are specified by only three: crack density, crack aspect-ratio, and crack strike.

TIV- (or VTI-) anisotropy: Transverse isotropy with a vertical axis of cylindrical symmetry: the typical symmetry of PTL-anisotropy. In TIV-anisotropy, propagation in the horizontal plane is isotropic with no variation with azimuth.

Time-delay: The time-delay between the arrivals of the two split shear-wave phases in seismic anisotropy. Time-delays above earthquakes are frequently normalized by path length to ms/km.

Transverse isotropy: Transverse isotropy (strictly hexagonal symmetry) is the anisotropic symmetry system with five elastic constants, and is isotropic in directions perpendicular to an axis of cylindrical rotational symmetry.

90°-flips in shear-wave polarisations: Abrupt 90° changes in shear-wave polarisations (90°-flips) have been observed and modeled along shear-wave ray paths passing near to point-singularities (see below) (Bush and Crampin 1991; Crampin 1991). 90°-flips can also be caused by critically high pore-fluid pressures locally re-orienting stress-aligned

fluid-saturated EDA-cracks, both in fluid-injection in critically pressurized hydrocarbon reservoirs (Angerer et al. 2002) and adjacent to seismically-active (critically-pressurized) faults (Crampin et al. 2002).

Fundamental Features of Shear-Wave Splitting in the Crust

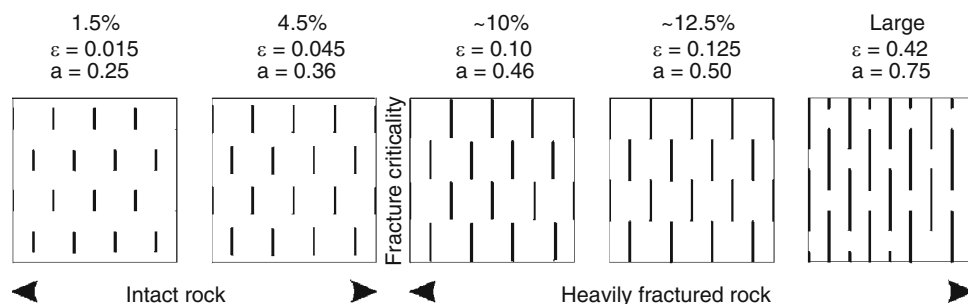
Observations of shear-wave splitting at a horizontal free surface typically display stress-aligned parallel polarisations. Since only transverse isotropy with a horizontal axis of symmetry (TIH-anisotropy) has such parallelism, and only parallel vertical cracks have such symmetry, the splitting is necessarily caused by stress-aligned EDA-microcracks. Microcracks is conformed as temporal changes in time-delays are observed, in fluid-injections (Angerer et al. 2002) and before earthquakes (Crampin and Peacock 2008). Only microcracks have sufficient compliance to allow such changes.

Observations of azimuthally varying stress-aligned shear-wave splitting, in almost all igneous, metamorphic, and sedimentary rocks in the crust, indicate a minimum of $\sim 1.5\%$ shear-wave velocity anisotropy and a maximum in ostensibly unfractured rock of $\sim 4.5\%$ (Crampin 1994, 1999; Crampin and Peacock 2008). Since crack density can be written as a dimensionless quantity $CD = Na^3$, where N is the number of cracks of radius a per unit cube and if Poisson's ratio is 0.25, $CD \approx 1/100$ of the percentage of shear-wave velocity anisotropy, observed shear-wave splitting can be imaged as uniform distributions of parallel vertical dimensionless cracks in Fig. 2 (Crampin 1994; Crampin and Peacock 2008).

The evolution of an initially random distribution of stress-aligned fluid-saturated microcracks under changing conditions can be modeled by the equations of anisotropic poroelasticity, APE, (Crampin and Zatsepin 1997). Figure 3 gives a schematic illustration of APE-deformation. The effect of increasing (or decreasing) stress is to increase (or decrease) the average aspect-ratio for cracks parallel to the increasing stress. APE-deformation is almost without parameters and yet matches a huge range of phenomena in exploration and earthquake seismology (Crampin and Peacock 2008), where the general applicability is because microcracks are so closely spaced that they are critical-systems and lead to the New Geophysics. Note that APE-modeling, because of the universality of critical-systems, is almost without parameters, and yet is widely applicable.

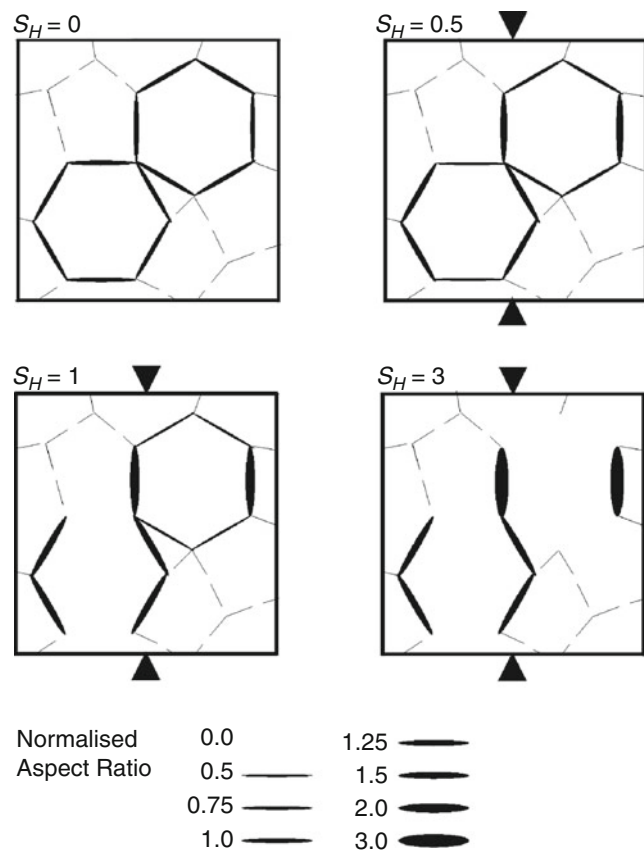
Classes of Anisotropic Symmetry and Shear-Wave Splitting

Shear-wave splitting has characteristic patterns of polarizations and velocities in three-dimensional propagation



Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting, Fig. 2 Schematic illustration of dimensionless three-dimensional crack distributions in situ rocks for a range of percentages

shear-wave velocity anisotropy, where ε is crack density and a is normalized crack radius (after Crampin 1994). Fracture-criticality is at $\sim 5.5\%$ shear-wave velocity anisotropy (Crampin and Zatsepin 1997)



Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting, Fig. 3 Schematic illustration of anisotropic poroelastic (APE) modelling of the effects of increasing stress on aspect-ratios of (6% porosity) microcracks, where aspect-ratios are given below (Crampin and Zatsepin 1995; Crampin 2006)

depending on the class of anisotropic symmetry (Nye 1957; Crampin 1977; Crampin and Kirkwood 1981). Elastic constants allow eight classes of anisotropic symmetry (including isotropic symmetry which describes a material with two elastic constants, usually written λ and μ , and has identical elastic properties in all directions of propagation). It is the characteristic 3D patterns of planes of mirror symmetry which define

the classes of anisotropic symmetry. Four types of anisotropic symmetry are commonly found in the Earth (where the horizontal plane is typically a plane of mirror symmetry):

1. *Orthorhombic symmetry* with nine independent elastic constants has three mutually-orthogonal planes of mirror symmetry. Crystalline olivine, a possible upper mantle constituent has orthorhombic symmetry, where shear-waves split in 3D patterns of polarizations. In addition, simple combinations of EDA- and PTL-anisotropy also have orthorhombic symmetry (Wild and Crampin 1991). The effects of combinations of EDA- and PTL-anisotropies have been observed in hydrocarbon reservoirs (Bush and Crampin 1991).
2. *Transverse isotropy with a vertical axis of cylindrical symmetry* commonly referred to as TIV-anisotropy. In TIV, shear-waves split strictly into phases oscillating in the vertical plane (*SV* waves) and phases oscillating horizontally (*SH* waves). There is no azimuthal variation. TIV is the effective anisotropic symmetry of finely divided horizontal layers found in many hydrocarbon reservoirs. It is also the symmetry of small horizontal platelets in shales. Thomsen (1986) derived three highly important approximations (the Thomsen parameters) for the behaviour of seismic waves in TIV media which are directly analogous to parameters measured in reflection surveys. Thomsen parameters are widely used and highly effective in interpreting exploration surveys. Thomsen parameters cannot be used to describe *azimuthally varying* anisotropy, except in isolated directions of sagittal symmetry.
3. *Transverse isotropy with a horizontal axis of cylindrical symmetry* commonly referred to as TIH-anisotropy. TIH is commonly observed above small earthquakes where the faster split shear-wave is polarized parallel to the direction of maximum horizontal stress. Only TIH-anisotropy can produce such parallelism. Only fluid-saturated microcracks are common to the huge variety of geological materials where TIH is observed, and since temporal changes before earthquakes are observed in the field, only fluid-saturated

microcracks have sufficient compliance to display temporal changes for small changes in parameters. Consequently, the observed parallelism of fast polarizations directly confirms stress-aligned microcracks throughout almost all rocks in the crust (Crampin and Peacock 2008).

4. *Monoclinic symmetry* has 13 elastic constants and is the symmetry of two (or more) sets of intersecting non-orthogonal vertical parallel EDA-cracks. Shear-wave splitting in such monoclinic symmetry has been observed and modeled in the field (Liu et al. 1993a). If two intersecting sets of vertical EDA-cracks are orthogonal, the symmetry is orthorhombic. Liu et al. (1993b) display the theoretical effects of intersecting sets of parallel vertical EDA-cracks for a range of parameters and a range of angles of intersection.

Cusps

In strong shear-wave velocity anisotropy, the direction of group-velocity (energy) propagation of the faster of the two split shear-wave may deviate so strongly that they overlap themselves and cause large amplitude cusps. Cusps have been observed and modeled in hydrocarbon reservoirs (Slater et al. 1993). However, the commonest form of cusps that occur in all anisotropic symmetry systems are the cuspidal lids, fins, and ridges associated with point- and line-singularities in group-velocity shear-wave surfaces (Crampin 1991; Wild and Crampin 1991).

Shear-Wave Window

Shear-waves with *SV* orientations incident on the free surface of an isotropic medium are identical to the incoming shear-wave (but double the amplitude) *only within* the shear-wave window. Outside the window, *S*-wave energy is lost to *SV*-to-*P* conversions so that *SV* waves are heavily distorted and even the arrival of the direct *S* wave may be unreadable. The edge of the shear-wave window is outlined by the critical angle of incidence, $i_c = \sin^{-1}(V_s/V_p)$, where V_p and V_s are the isotropic *P*-wave and *S*-wave velocities, and i_c is the critical angle for *S*-to-*P* conversions (Evans 1984; Booth and Crampin 1985). The conical solid angle of the shear-wave window may be considered as *above* small earthquakes or *below* surface recorders.

Other features of the shear-wave window:

1. The critical angle of incidence i_c is $\sim 35^\circ$ for a uniform isotropic medium with a Poisson's ratio of 0.25, where Lamé parameters $\lambda = \mu$.
2. Because of near-surface low-velocity layers, ray paths tend to curve upwards as they approach the free surface, so that the effective shear-wave window is often out to straight-line source-to-receiver incidence of 45° or greater.
3. Topographic irregularities may seriously distort the shear-wave window and may cause *S*-to-*P* conversions (the Local *SP*-waves) both inside and outside the theoretical

shear-wave window. Local *SP*-waves are frequently observed in the field as large single-sided pulses on the *SV*-component, and may be mistakenly interpreted as anomalous shear-wave splitting (Crampin 1990).

4. The shear-wave window only refers to *SV*-propagation. No energy is lost by *SH* waves at any angle of incidence on a plane horizontal free surface. Consequently, only the *SV*-component of any incident-split shear-wave is affected by the shear-wave window.
5. In anisotropic propagation, the shear-wave window may severely disturb shear-wave splitting.

These various anomalies mean that shear-wave splitting recorded at a free surface needs to be interpreted with caution (Crampin and Peacock 2008).

Shear-Wave Singularities

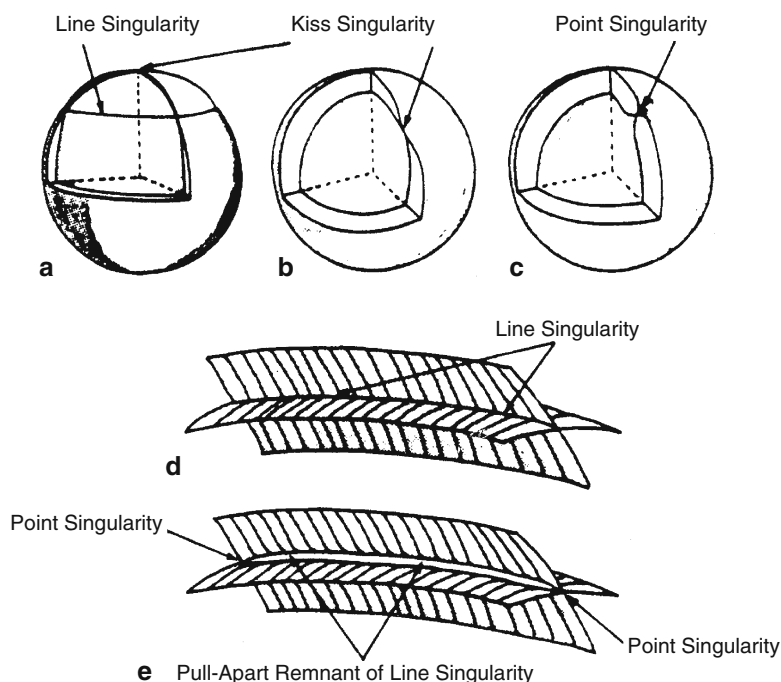
The three body waves in anisotropic solids trace out three velocity surfaces for propagation in three-dimensional propagation: a *P* wave; and two orthogonally-polarized shear-wave surfaces. The phase-velocity shear-wave surfaces from a point of propagation necessarily touch and are continuous in a variety of singularities in three-dimensional patterns of direction characteristic of the anisotropic symmetry class (Crampin and Yedlin 1981; Crampin and Kirkwood 1981; Wild and Crampin 1991). The only way for energy of a split shear-wave arrival to pass from one shear-wave surface to the other is along directions passing directly through a singularity.

Fig. 4 shows the three types of singularity in *phase-velocity surfaces*:

1. *Kiss-singularities* where two shear-wave surfaces touch tangentially (Fig. 4a, b). Kiss-singularities occur in directions of principal crystallographic axes and are restricted to cubic, tetragonal anisotropic symmetry, and along the cylindrical symmetry direction of transversely isotropic symmetry. Kiss-singularities can also occur in orthorhombic and monoclinic anisotropic symmetry when two shear-wave surfaces happen to touch serendipitously.
2. *Point-singularities* (sometimes called conical points) where the two shear-wave surfaces touch at the vertices of convex and concave cones (Fig. 4c). Shear-waves in all classes of anisotropic symmetry possess point-singularities except transverse isotropy (hexagonal symmetry), which only has kiss- and line-singularities (Fig. 4a). Point-singularities are particularly numerous in combinations of PTL- and EDA-anisotropy, where thin layers are pervaded by vertical microcracks (Wild and Crampin 1991).
3. *Line- or intersection-singularities* occur only in transverse isotropy where the two shear-wave phase-velocity surfaces intersect each other in circular line-(intersection-) singularities about the symmetry axis (Fig. 4a, d). In even marginally disturbed transverse isotropy, there are

Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting,

Fig. 4 Schematic illustration of the intersection of two shear-wave phase-velocity surfaces showing topography near: (a) line-singularity or intersection-singularity (with kiss-singularity); (b) kiss-singularity; and (c) point-singularity (after Crampin and Yedlin 1981). (d) A segment of a representation of a line-singularity, and (e) a *pull-apart remnant of a line-singularity* when the transverse isotropy symmetry is disturbed (after Wild and Crampin 1991)



dramatic changes. Since both PTL- and EDA-anisotropy have transversely isotropy symmetry with line-singularities, Wild and Crampin (1991) show that the phase-velocity representations of combinations of PTL and EDA have *pull-apart remnants of line-singularities*. The intersections of the inner and outer shear-wave surfaces separate, and 'pinch' together, in varying degrees of tightness, between a string of point-singularities (Fig. 4e).

Singularities in *group-velocity surfaces* have a variety of behaviour:

1. *Kiss-singularities* have simple group-velocity representation, the shear-wave surfaces again touch tangentially.
2. *Point-singularities* in group-velocity surfaces: the faster shear-wave surfaces transform into open almost circular "holes", whereas the slower surfaces transform into flat cuspidal "lids" which exactly fit into the open holes. The effects of such point-singularities have been observed and modeled in a hydrocarbon reservoir the field (Bush and Crampin 1991). Crampin (1991) modeled synthetic seismograms for ray path directions passing close to point-singularities. The effects can be dramatic with possibly great amplification or attenuation of amplitudes and abrupt 90° changes in shear-wave polarisations (90°-flips) either side of point-singularities.
3. *Line- or intersection-singularities*. In undisturbed transverse isotropy, the group-velocity representations of line-singularities merely intersect with each sheet showing no evidence of intersection, as in the phase-velocity

representation in Fig. 4d. However, group-velocity representations of pull-apart remnants of line-singularities in disturbed transverse isotropy typically display thin cuspidal "fins", "ridges", and "lids" of extraordinary complexity (and beauty) (Wild and Crampin 1991).

There are two conclusions. The effects of pull-apart remnants of line-singularities are sensitive to very small differences in direction of propagation or microcrack geometry, say. The subtleties of shear-wave propagation near the extensive pull-apart remnants of line-singularities are likely to be the reason for many of the complications frequently observed in field records of shear-waves in sedimentary basins. If correctly interpreted, such features can provide crucial information particularly about the relative proportions of PTL- and EDA-anisotropy (Bush and Crampin 1991).

Sources of Shear-Wave Splitting

There are three common areas of investigation into shear-wave splitting (and seismic anisotropy).

1. *Shear-wave splitting in the upper mantle*, which is typically assumed to be caused by strain-aligned crystals such as olivine, and is used to demonstrate directions of stress and flow in plate tectonics (reviewed by Savage 1999).
2. *Shear-wave splitting in hydrocarbon reservoirs* in seismic exploration in industrial seismology, where initially "anisotropy is a nuisance" (Helbig 1994) because many

well-established processing procedures assume isotropic propagation and are no longer valid if the rocks are anisotropic. More recently, however, azimuthally varying shear-wave splitting in TIH-anisotropy is being actively used to determine fracture orientations and improve fluid-flow for hydrocarbon recovery (reviewed by Helbig and Thomsen 2005).

3. *Shear-wave splitting above small earthquakes* initially stimulated development of the theory (reviewed by Crampin 1981) and observation (Crampin et al. 1980; Alford 1986), and is now demonstrating the *New Geophysics* of a crack-critical crust (Crampin 2006), where time-delays directly monitor low-level fluid-rock deformation at levels of deformation well-below those at which fracturing takes place (Crampin and Zatsepin 1997).

Remarkably, there has been very little interaction between these three areas of investigation. Each has developed largely independently where papers in one area seldom cite other areas of research. Shear-wave splitting in these three areas of investigation will be discussed separately.

Shear-Wave Splitting in the Upper Mantle

Seismic anisotropy in the mantle was first recognized from the incompatibility of Rayleigh and Love surface-wave inversions (Anderson 1961) and from the velocity anisotropy of horizontally propagating *Pn*-waves in oceanic basins (Hess 1964; Raitt et al. 1969). Another anisotropic surface-wave phenomenon directly analogous to shear-wave splitting of body waves is the coupling of surface-wave motion of Raleigh and Love modes into Generalized modes varying in three-dimensions (Crampin and King 1977). Just as shear-wave splitting is highly diagnostic of some form of seismic anisotropy, coupled surface-wave motion is highly diagnostic of anisotropy in the mantle. The pronounced coupling between Second Mode Rayleigh and Second Mode Love waves across Eurasia could be caused by as little as 4 km of 4% shear-wave velocity anisotropy at the top of the upper mantle (Crampin and King 1977).

Shear-wave splitting of body waves in the mantle only became observable with advances in digital instrumentation. The first reports of such shear-wave splitting were by Ando et al. (1980) who reported time-delays of ~1 s on 2 s-period shear-waves propagating vertically. This was attributed to 2–4% shear-wave velocity anisotropy in propagation through a magma reservoir, 100–200 km in depth beneath a volcanic area in Japan above the subducting Pacific Plate.

Since that time, shear-wave splitting has been observed extensively in the uppermost 200 km of the mantle and in the D region (Silver and Chan 1988, 1991), reviewed by Savage (1999), and Long and Silver (2009) and others. These observations are principally of shear-wave splitting in *SKS*-phases, which are frequently large amplitude, where the *P-S*

conversion on exiting from the core provides a known *SV* orientation at the start the final shear-wave path to the surface. Frequencies are typically 0.5–0.25 Hz leading to 2–4% shear-wave velocity anisotropy, similar to the crack-induced shear-wave velocity anisotropy in the crust.

The cause of the shear-wave splitting in the mantle is not wholly resolved. The most frequently cited cause is referred as lattice-preferred orientation (LPO) of the crystallographic axes of elastically anisotropic crystals, such as olivine (Karato and Wu 1993) or orthopyroxene (Ben-Ismaïl et al. 2001). These crystals are assumed to be aligned by convective flow where the presence of water, induced by bending of subducting slabs (Faccenda et al. 2009), say, could have major effects on crystal orientations and shear-wave splitting (Karato and Jung 1998; Katayama et al. 2005).

However, the similarities in the degree of shear-wave splitting and the polarization to splitting in the crust which is certainly caused by EDA-cracks suggests the alternative source of anisotropy of fluid-saturated stress-aligned cracks. There is water in much of the upper mantle. Water lowers the melting point of rock, and when crystalline rock first begins to melt, it melts along grain boundaries. Thus the anisotropy in the mantle could be caused by aligned cracks of films of liquid melt (Crampin 2003).

Shear-Wave Splitting in Hydrocarbon Reservoirs

The effective TIV-anisotropy of finely stratified rocks was demonstrated by Postma (1955) and others, where double refraction (shear-wave splitting) only has *SV*- and *SH*-polarisations. Shear-wave splitting of TIV-anisotropy was not much investigated at that time since the predominant single-component vertical seismometers only recorded *SV* waves.

Azimuthally varying shear-wave splitting in hydrocarbon reservoirs burst onto the seismic exploration industry in a series of some dozen papers in two sessions at the 1986 SEG symposium in Houston. Alford (1986) rotated (“Alford rotations”) the polarisations of reflection surveys recorded on two-component horizontal geophones from two-component horizontal vibrators, and showed that seismic shear-wave reflection sections displayed coherent behavior only when both source and receivers were rotated into the preferred (anisotropic) polarizations. These papers stimulated great activity in exploring seismics in order to obtain preferred directions of fluid-flow by locating fracture orientations (reviewed by Helbig and Thomsen 2005). This interest in anisotropy still continues today but is frequently investigated by wide-angle and other *P*-wave surveys because of the expense of three-component shear-wave surveys including instrumentation as well as the tripling of the quantity of data to be recorded and processed.

Note that azimuthally varying shear-wave splitting in hydrocarbon reservoirs has been typically interpreted by the

hydrocarbon industry as being caused by *aligned fractures*, where dimensions of the fractures are unspecified but are expected to be large enough to lead to improved hydrocarbon recovery. In contrast, Crampin and Peacock (2008) suggest that the remarkable compliance of shear-wave splitting to low-level deformation necessarily means that the splitting is caused by stress-aligned fluid-saturated *microcracks*. Since in many circumstances microcracks are parallel to macrofractures, the difference in interpretation is often immaterial. However, the recent demonstration, both in theory (Chapman 2003) and observations (Maultzsch et al. 2003), that anisotropy is frequency-dependent and varies with the dimensions of microcracks may change this interpretation.

Shear-Wave Splitting Above Small Earthquakes

Shear-wave splitting was first positively identified in the crust above a swarm of small earthquakes near the North Anatolian Fault in Turkey in the Turkish Dilatancy Project (TDP1) experiment, (Crampin et al. 1980). This experiment was designed to search for the microcrack-induced shear-wave splitting suggested by Crampin (1978). Similar observations have now been observed above earthquakes worldwide (reviewed by Crampin and Peacock 2008). However observations above earthquakes need to be recorded by three-component seismic networks within the shear-wave window of the earthquakes. This is a severe restriction as swarms of small earthquakes are scarce and intermittent, and seldom observed by seismic networks. Nevertheless, such swarms of small earthquakes have been used as ‘stress-measuring stations’ to monitor the effects of stress changes before impending earthquakes (Crampin and Peacock 2008).

Iceland is an optimum location for studying shear-wave splitting. It is an offset of the Mid-Atlantic Ridge where, uniquely, (two) transform faults run onshore. These transform faults have persistent swarm activity monitored by a state-of-the-art seismic network (Stefánsson et al. 1993) available over the Internet (Volti and Crampin 2003).

The APE model of microcrack deformation (Crampin and Zatsepin 1997) shows that increases of stress increase crack aspect-ratios (crack thicknesses), and increases of aspect-ratio can be recognized by increases of average time-delays in Band-1 directions of the shear-wave window (Crampin 1999), and have been used to stress-forecast earthquakes (Crampin et al. 1999, 2008).

The New Geophysics

The shear-wave velocity anisotropy observed in the crust of the Earth indicates that the distributions of stress-aligned fluid-saturated microcracks (EDA-cracks) are so closely spaced they verge on fracturing and hence are critical-systems

(Fig. 2). Such critical microcracks are the most compliant elements of in situ rock, and indicate profound differences from conventional subcritical solid-earth geophysics. Critical systems are a *New Physics* (Davies 1989), hence a *New Geophysics*, which imposes a range of fundamentally new properties on conventional subcritical physics (Crampin and Peacock 2005; Crampin 2006). All complex heterogeneous interactive systems are critical-systems, and they are extremely common: the weather, clustering of traffic on roads, the life-cycle of fruit flies, and a huge range of physical phenomena from stellar radiation to quantum mechanics. Since the Earth is an archetypal complex heterogeneous interactive system, it is necessarily expected to be a critical system.

The new geophysical properties include:

1. *Monitorability* – effects of criticality (on the geometry of EDA-cracks) can be monitored within the interior of the crust by shear-wave splitting (Crampin 1994, 2006).
2. *Calculability* – the evolution of microcrack geometry can be calculated by APE (Crampin and Zatsepin 1997; Crampin 1999; Angerer et al. 2002).
3. *Predictability* – if changing conditions are known, the effects on crack geometry can be predicted (as in Angerer et al. 2002).
4. *Controllability* – if the intended effects of some operation (fluid-injection, say) are desired (the opening of specific fractures for fluid-flow, say), the effects can be calculated by APE, and the effects can be controlled by feedback by analyzing shear-wave splitting. Crack aspect-ratios, as in Angerer et al. (2002) where the effects of fluid-injections were calculated, could be controlled by feedback.
5. *Universality* – the changes in critical-systems are pervasive over all available space where appropriate conditions are maintained. The coincidence of observations of 1.5–4.5% shear-wave velocity anisotropy with similar stress-oriented polarizations in all types of rock regardless of porosity, rock types, or geology (Crampin 1994; Crampin and Peacock 2005, 2008) is an example of universality in the crack-critical Earth, that is difficult to explain in a conventional sub-critical Earth.
6. *Sensitivity* – extreme (butterfly wings) sensitivity to initial conditions. A prime example of such sensitivity is the SMSITES experiment in Iceland where distinctive variations in *P*- and shear-wave travel-time variations were observed at 70 km from swarm activity with an energy equivalent to a $M = 3.5$ earthquake at 70 km distance. A $M = 3.5$ earthquake is small, so these effects are observed at hundreds of times the likely source diameter in a conventional subcritical earth and are a clear demonstration of critical-system sensitivity (Crampin et al. 2003).

Stress-Forecasting Earthquakes

Shear-wave splitting and the New Geophysics lead to a new understanding of tectonic deformation before earthquakes. Initially, tectonic stress accumulates by interactions at plate boundaries leading to deformation of EDA-cracks which can be monitored by shear-wave splitting. Such stress-accumulations are not initially associated with any particular source-zone and may be observable at great distances from any potential source. As the stress-field approaches, fracture-criticality weaknesses are identified and microcracks begin to coalesce onto the potential slip-plane. There is stress-relaxation, and eventually the impending earthquake occurs (Gao and Crampin 2004). Logarithms of the durations of increases (stress-accumulation) and decreases (crack-coalescence) are each separately proportional (self-similar) to impending earthquake magnitudes (Crampin et al. 2008), similar to the linearity of the Gutenberg–Richter relationship. Consequently, shear-waves above swarms of small earthquakes can be used as “stress-measuring stations” to monitor such stress-accumulation and stress-relaxation before impending earthquakes.

Peacock et al. (1988) were the first to recognize such changes in shear-wave time-delays (now identified with stress-accumulation), using seismic station KNW as a stress-measuring station, before a $M = 6$ earthquake near the San Andreas Fault in California. A few other examples were found worldwide (reviewed by Crampin and Peacock 2008), but the significant advances came when persistent seismicity in Iceland became available for analysis online. Volti and Crampin (2003) displayed increases of Band-1 time-delays before five earthquakes in Iceland monitoring stress-accumulation. The time, magnitude, and fault-plane of a fifth earthquake were successfully stress-forecast 3 days before it occurred in a comparatively tight time/magnitude window (Crampin et al. 1999, 2008).

Crack-coalescence (stress-relaxation), before the impending earthquake occurs when the increasing stress-accumulation abruptly begin to decrease, was first recognized by Gao and Crampin (2004). Since then, stress-accumulation has been recognized before 15 earthquakes worldwide, of which nine had sufficient seismicity beneath stress-measuring stations to also display crack-coalescence (Crampin and Peacock 2008).

Unfortunately, swarms of small earthquakes are far too scarce and irregular to be used as stress-measuring stations for reliable routine stress-forecasting. Reliable stress-forecasting requires three 1–1.5 km-deep borehole Stress-Monitoring Sites (SMSs), where a Downhole-Orbital Vibrator source (DOV) (Leary and Walter 2005) radiate shear-waves to borehole geophones in Band-1 directions. A prototype SMS in non-optimal source-to-geophone geometry between existing

boreholes in Iceland recorded exceptional sensitivity to low-level seismicity equivalent to a $M = 3.5$ earthquake at 70 km-distance (Crampin et al. 2003) demonstrating that SMSs have sufficient sensitivity to stress-forecast times and magnitudes of $M = 5$ earthquakes up to 1000 km from impending source zones (Crampin and Gao 2010).

Summary

Shear-wave splitting in the Earth reveals a fundamentally new understanding of fluid-rock deformation, where APE-modeling shows that modifications to shear-wave splitting directly monitor stress-induced modifications to microcrack geometry. Since if the changing conditions can be specified, the deformation can be calculated and predicted by APE; this is a substantial advance, a New Geophysics, on conventional subcritical solid-earth geophysics. Changes in microcrack geometry can be monitored, calculated, predicted, and in some circumstances potentially controlled. There is extreme sensitivity to initial conditions, and universality, so that effects are widely and uniformly distributed.

The New Geophysics allows stress-accumulation and stress-relaxation (crack-coalescence) before earthquakes to be recognized and impending large earthquakes stress-forecast. There are many important implications and applications. In particular, any solid-earth application that cannot accommodate compliant stress-aligned fluid-saturated EDA-microcracks pervading most in situ rocks is in error and may lead to serious misunderstandings. New Geophysics leading to monitorability; calculability; predictability; potential controllability; universality; extreme sensitivity; and earthquake stress-forecasting, is likely to be the most fundamental advance in solid-earth geoscience for many decades.

Cross-References

- ▶ [Magnetic Anisotropy](#)
- ▶ [Seismic Tomography](#)
- ▶ [Seismic Viscoelastic Attenuation](#)
- ▶ [Seismic Wave Propagation in Real Media: Numerical Modeling Approaches](#)
- ▶ [Seismic Waves, Scattering](#)
- ▶ [Seismic, Receiver Function Technique](#)

Bibliography

- Alford RM (1986) Shear data in the presence of azimuthal anisotropy: dilley, Texas. In: 56th annual international society of exploration geophysicists meeting, Houston, Expanded Abstracts, pp 476–379

- Anderson DL (1961) Elastic wave propagation in layered anisotropic media. *J Geophys Res* 66:2953–2963
- Ando M, Ishikawa Y, Wada H (1980) S-wave anisotropy in the upper mantle under a volcanic area in Japan. *Nature* 286:43–46
- Angerer E, Crampin S, Li X-Y, Davis TL (2002) Processing, modelling, and predicting time-lapse effects of overpressured fluid-injection in a fractured reservoir. *Geophys J Int* 149:267–280. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Ben-Ismaïl W, Barruol G, Mainprice D (2001) The Kaapvaal craton seismic anisotropy: petrophysical analyses of upper mantle kimberlite nodules. *Geophys Res Lett* 28:2497–2500
- Booth DC, Crampin S (1985) Shear-wave polarizations on a curved wavefront at an isotropic free-surface. *Geophys J Roy Astron Soc* 83:31–45
- Bush I, Crampin S (1991) Paris Basin VSPs: case history establishing combinations of matrix- and crack-anisotropy from modelling shear wavefields near point singularities. *Geophys J Int* 107:433–447
- Chapman M (2003) Frequency dependent anisotropy due to meso-scale fractures in the presence of equant porosity. *Geophys Prospect* 51:369–379
- Crampin S (1977) A review of the effects of anisotropic layering on the propagation of seismic waves. *Geophys J R Astron Soc* 49:9–27
- Crampin S (1978) Seismic wave propagation through a cracked solid: polarization as a possible dilatancy diagnostic. *Geophys J R Astron Soc* 53:467–496
- Crampin S (1981) A review of wave motion in anisotropic and cracked elastic-media. *Wave Motion* 3:343–391
- Crampin S (1990) The scattering of shear waves in the crust. *Pure Appl Geophys* 132:67–91
- Crampin S (1991) Effects of point singularities on shear-wave propagation in sedimentary basins. *Geophys J Int* 107:531–543
- Crampin S (1994) The fracture criticality of crustal rocks. *Geophys J Roy Astron Soc* 118:428–438. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Crampin S (1999) Calculable fluid-rock interactions. *J Geol Soc* 156:501–514. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Crampin S (2003) Aligned cracks not LPO as the cause of mantle anisotropy, EGS-AGU-EUG joint assembly, nice. *Geophys Res Abstr* 5:00205. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Crampin S (2006) The new geophysics: a new understanding of fluid-rock deformation. In: Van Cotthem A et al (eds) *Eurock 2006: multiphysics coupling and long term behaviour in rock mechanics*. Taylor and Francis, London, pp 539–544. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Crampin S, Gao Y (2010) A review of earthquake stress-forecasting. In: *Proceedings of the international symposium on earthquake seismology and earthquake prediction, Beijing, 2009*, *Pure Appl Geophys*, (in press)
- Crampin S, King DW (1977) Evidence for anisotropy in the upper mantle beneath Eurasia from generalized higher mode surface waves. *Geophys J R Astron Soc* 49:59–85
- Crampin S, Kirkwood SC (1981) Velocity variations in systems of anisotropic symmetry. *J Geophys* 49:35–42
- Crampin S, Peacock S (2005) A review of shear-wave splitting in the compliant crack-critical anisotropic Earth. *Wave Motion* 41:59–77. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Crampin S, Peacock S (2008) A review of the current understanding of shear-wave splitting and common fallacies in interpretation. *Wave Motion* 45:675–722. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Crampin S, Yedlin M (1981) Shear-wave singularities of wave propagation in anisotropic media. *J Geophys* 49:43–46
- Crampin S, Zatsepin SV (1995) Production seismology: the use of shear waves to monitor and model production in a pororeactive and interactive reservoir. In: 65th annual international society of exploration geophysicists meeting, Houston, Expanded Abstracts, pp 199–202
- Crampin S, Zatsepin SV (1997) Modelling the compliance of crustal rock: II – response to temporal changes before earthquakes. *Geophys J Int* 129:495–506. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Crampin S, Evans R, Üçer B, Doyle M, Davis JP, Yegorkina GV, Miller A (1980) Observations of dilatancy-induced polarization anomalies and earthquake prediction. *Nature* 286:874–877
- Crampin S, Evans R, Doyle M, Davis JP (1981) Comments on papers about shear-wave splitting in dilatancy-induced anisotropy by Gupta, I. N., and by Ryall, A., and Savage, W. U. *Bull Seismol Soc Am* 71:375–377
- Crampin S, Evans R, Atkinson BK (1984) Earthquake prediction: a new physical basis. *Geophys J R Astron Soc* 76:147–156
- Crampin S, Volti T, Stefánsson R (1999) A successfully stress-forecast earthquake. *Geophys J Int* 138:F1–F5
- Crampin S, Volti T, Chastin S, Gudmundsson A, Stefánsson R (2002) Indication of high pore-fluid pressures in a seismically-active fault zone. *Geophys J Int* 151:F1–F5. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Crampin S, Chastin S, Gao Y (2003) Shear-wave splitting in a critical crust: III – preliminary report of multi-variable measurements in active tectonics. *J Appl Geophys* 54:265–277. Special Issue. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Crampin S, Gao Y, Peacock S (2008) Stress-forecasting (not predicting) earthquakes: a paradigm shift? *Geology* 36:427–430. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Davies P (1989) The new physics: a synthesis. In: Davies P (ed) *The new physics*. Cambridge University Press, Cambridge, pp 1–6
- Evans R (1984) Effects of the free surface on shear wavetrains. *Geophys J R Astron Soc* 76:165–172
- Faccenda M, Gerya TV, Burlini L (2009) Deep slab hydration induced by bending-related variations in tectonic pressure. *Nat Geosci* 2:790–793
- Fouch MJ, Rondenary S (2006) Seismic anisotropy beneath stable continental interiors. *Phys Earth Planet Inter* 258:292–320
- Gao Y, Crampin S (2004) Observations of stress relaxation before earthquakes. *Geophys J Int* 157:578–582. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Gupta IN (1973) Premonitory variations in S-wave velocity anisotropy before earthquakes in Nevada. *Science* 182:1129–1132
- Helbig K (1994) Foundations of anisotropy for exploration seismics. In: *Handbook of geophysical exploration, section 1, seismic exploration*. Elsevier Science, Oxford, p 248
- Helbig K, Thomsen L (2005) 75-plus years of anisotropy in exploration and reservoir seismics: a historical review of concepts and methods. *Geophysics* 70:9ND–23ND
- Hess H (1964) Seismic anisotropy of the uppermost mantle under oceans. *Nature* 203:629–631
- Karato S-I, Jung H (1998) Water, partial melting and the origin of the seismic low velocity and high attenuation zone in the upper mantle. *Earth Planet Sci Lett* 157:193–207
- Karato S-I, Wu P (1993) Rheology of the upper mantle: a synthesis. *Science* 260:771–778
- Katayama I, Karata S-I, Brandon M (2005) Evidence of high water content in the deep upper mantle inferred from deformation microstructures. *Geology* 33:613–616

- Keith CM, Crampin S (1977) Seismic body waves in anisotropic media: synthetic seismograms. *Geophys J R Astron Soc* 49:225–243
- Leary PC, Walter LA (2005) Physical model for the downhole orbital vibrator (DOV) – I. Acoustic and borehole seismic radiation. *Geophys J Int* 163:647–662
- Liu E, Crampin S, Queen JH, Rizer WD (1993a) Velocity and attenuation anisotropy caused by microcracks and macrofractures in multiazimuthal reverse VSPs. *Can J Explor Geophys* 29:177–188
- Liu E, Crampin S, Queen JH, Rizer WD (1993b) Behaviour of shear waves in rocks with two sets of parallel cracks. *Geophys J Int* 113:509–517
- Long MD, Silver PG (2009) Shear wave splitting and mantle anisotropy: measurements, interpretations, and new directions. *Surv Geophys* 30:407–461
- Love AEH (1927) A treatise of the mathematical theory of elasticity. Dover, New York. 1944 reprinted
- Maultzsch S, Chapman M, Liu E, Li X-Y (2003) Modelling frequency dependent seismic anisotropy in fluid-saturated rock with aligned fractures: implication of fracture size estimation from anisotropic measurements. *Geophys Prospect* 51:381–392
- Maupin V, Park J (2007) Theory and observations – wave propagation in anisotropic media. In: Schubert G (ed) *Treatise on geophysics*. Elsevier, Amsterdam, pp 289–321. <https://doi.org/10.1016/B978-044452748-6/00007-9>
- Nur A, Simmons G (1969) Stress-induced anisotropy in rock: an experimental study. *J Geophys Res* 74:6667–6674
- Nye JF (1957) *Physical properties of crystals*. Clarendon, Oxford
- Peacock S, Crampin S, Booth DC, Fletcher JB (1988) Shear-wave splitting in the Anza seismic gap, southern California: temporal variations as possible precursors. *J Geophys Res* 93:3339–3356
- Postma GW (1955) Wave propagation in a stratified medium. *Geophysics* 20:780–806
- Raitt RW, Shor GG, Francis TJG, Morris GB (1969) Anisotropy of the Pacific upper mantle. *J Geophys Res* 74:3095–3109
- Savage MK (1999) Seismic anisotropy and mantle deformation: what have we learned from shear wave splitting? *Rev Geophys* 37:65–106
- Silver PG, Chan WW (1988) Implications for continental structure and evolution from seismic anisotropy. *Nature* 335:34–39
- Silver PG, Chan WW (1991) Shear wave splitting and sub-continental mantle deformation. *J Geophys Res* 96:16,429–16,454
- Slater C, Crampin S, Brodov LY, Kuznetsov VM (1993) Observations of anisotropic cusps in transversely isotropic clay. *Can J Explor Geophys* 29:216–226
- Stefánsson R et al (1993) Earthquake prediction research in the South Iceland seismic zone and the SIL project. *Bull Seismol Soc Am* 83:696–716
- Thomsen L (1986) Weak elastic anisotropy. *Geophysics* 51:1954–1966
- Volti T, Crampin S (2003) A four-year study of shear-wave splitting in Iceland: 2 Temporal changes before earthquakes and volcanic eruptions. *Geol Soc Lond Spec Publ* 212:135–149. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>
- Wild P, Crampin S (1991) The range of effects of azimuthal isotropy and EDA-anisotropy in sedimentary basins. *Geophys J Int* 107:513–529
- Zatsepin SV, Crampin S (1997) Modelling the compliance of crustal rock: I – response of shear-wave splitting to differential stress. *Geophys J Int* 129:477–494. Available at file <http://www.geos.ed.ac.uk/homes/scrampin/opinion/>

Single and Multichannel Seismics

Tamás Tóth

Geomega Limited, Budapest, Hungary

Definition

Seismics or Seismic survey	The method of imaging the subsurface with reflected and refracted seismic waves, which are generated by controlled sources and sensed by seismic detectors.
Single-channel seismics	Seismic survey using one detector (channel) for sensing returning seismic waves.
Multichannel seismics	Seismic survey using multiple detectors (channels) at different locations for sensing returning seismic waves generated by each source points.

Introduction

Reflection and refraction seismology has been applied for imaging the subsurface since the 1920s (Karcher 1987; Keppner 1991) and is still one of the most powerful methods of exploration geophysics. Applied technology has evolved considerably over the decades, but the principles have not changed. Imaging is performed by elastic waves generated by controlled sources, such as explosive, vibroseis, air gun, sparker, boomer, or similar sources. Generated waves propagate through the subsurface layers and part of them get reflected or refracted back from layer boundaries. These returning waves are sensed on the surface by seismic detectors, geophones on land, and hydrophones in water.

In case of single-channel seismic recording, elastic waves generated by each shot are recorded by one single detector or a group of detectors. On the other hand, multichannel seismic recording utilizes multiple detectors or groups of detectors located at different positions for recording the wavefield generated by each shot. Multichannel seismics provide several advantages over single-channel seismics; however, simplicity of the recording equipment and high resolution of the recorded seismic profiles make single-channel seismics still a powerful and cost-effective tool for shallow investigations.

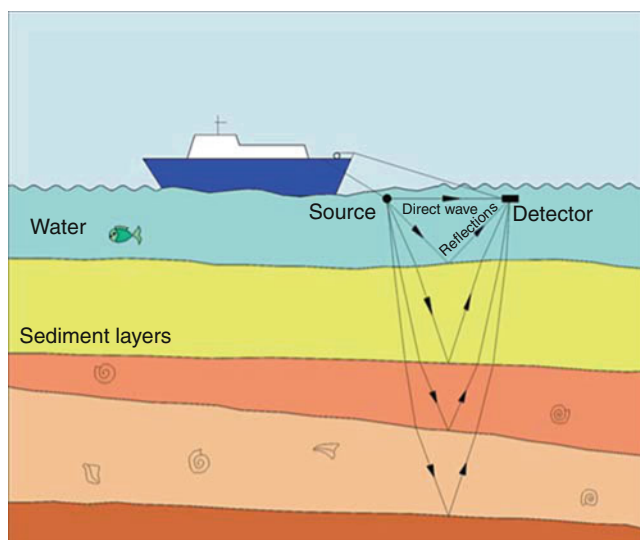
Single-Channel Seismics

Simplest possible configuration for a seismic reflection survey is single-channel recording. Reflected energy from each shot is sensed by only one receiver or one group of receivers, normally positioned at a constant distance from the source (see Fig. 1).

The source–receiver pair is moved, and each shot generates one seismic trace for the profile. In case of singlechannel recording, each shot provides information from one subsurface location only, therefore it is critical to have a low cost and high repetition rate source. Shot generation is more labor intensive on land; hence, single-channel recording is hardly used for onshore surveys. Situation is more favorable for offshore surveys, as easy-to-operate sources with high repeatability are available and make single-channel recording an attractive alternative. This is especially true, when relatively shallow “penetration” is sufficient, but *high resolution* is needed.

Penetration, on the one hand, strongly depends on the seismic source used; on the other hand, it is also determined by the physical properties of the subsurface strata. Seismic *resolution* is a function of the frequency content of the source used. It is important to note that penetration and resolution are not independent parameters. Higher frequency sources ensure maximum resolution, but as high.

frequencies attenuate quicker, they provide less penetration. Highest frequency sources are Compressed High Intensity Radar Pulse (*CHIRP*) and boomers, while

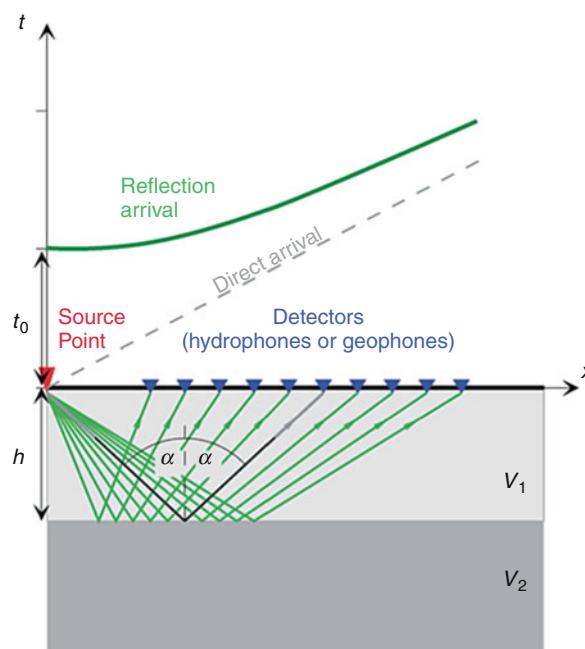


Single and Multichannel Seismics, Fig. 1 During singlechannel recording, seismic energy generated by each source point images one subsurface point of each layer boundary, generating one seismic trace

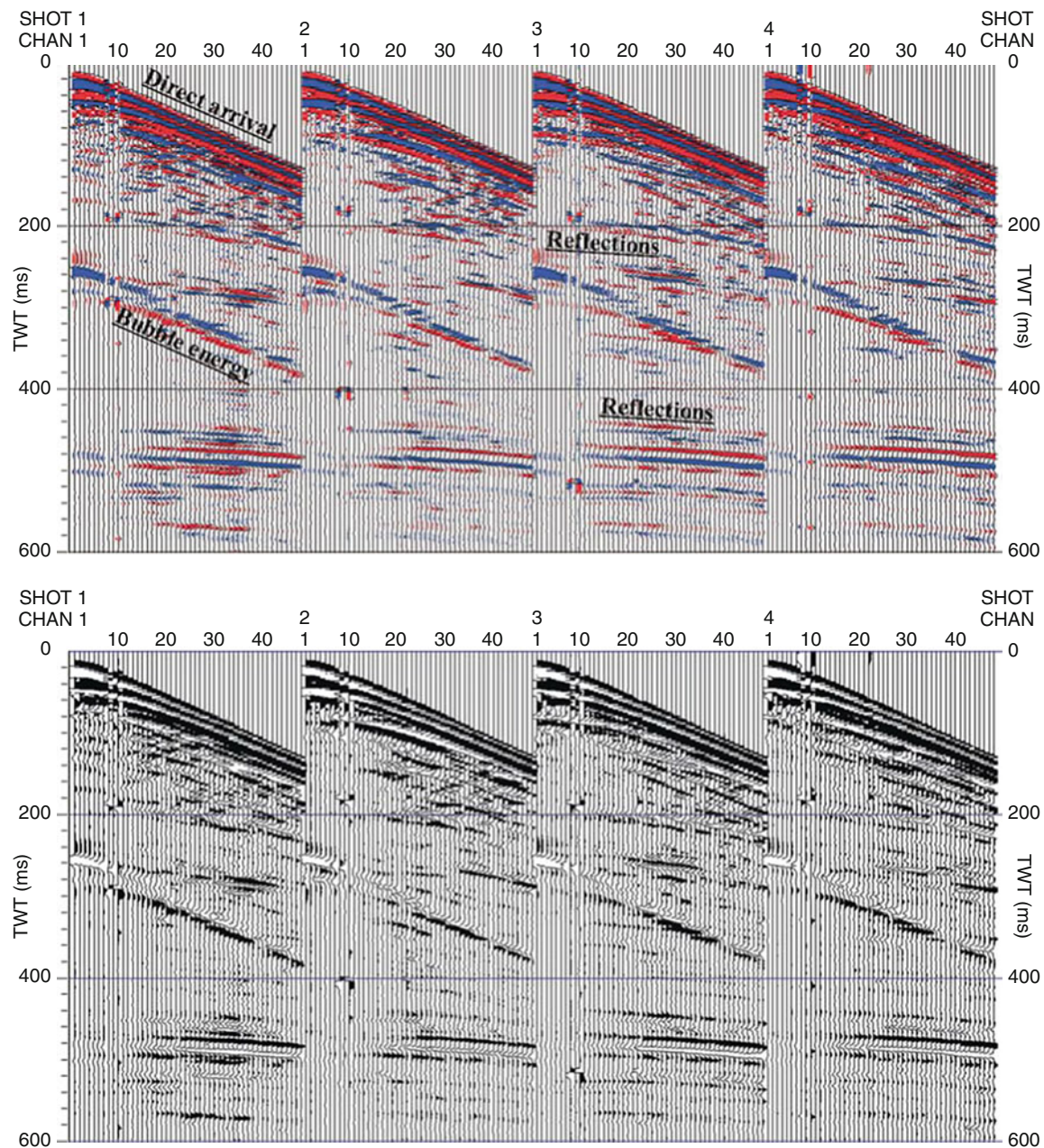
sparkers and small volume air gun sources with their higher energy provide deeper penetration, but generally lower resolution.

Multichannel Seismics

Several advantages are gained by recording the reflected energy of each shot by multiple detectors, which is called multichannel seismic recording (see sketch in Fig. 2 and sample shot gathers in Fig. 3). The most important of these advantages are multifold imaging of the same location, called the common depth point (CDP) and the possibility to estimate seismic velocity from the recorded data itself. Multifold imaging enhances the signal/noise ratio of the final seismic section through stacking. The best possible estimate of velocity field of the subsurface layers is necessary for normal move out (*NMO*) correction and “migration” before and after stacking. This information is derived from the parameters of the hyperbola fitted on the reflections in the CDP gathers, a process called “velocity analysis.”



Single and Multichannel Seismics, Fig. 2 Reflected energy generated by each source point is recorded by multiple detectors. Each detector images a different subsurface point. Reflection at each subsurface point follows Snell’s law. Angles of incidence and reflection (marked α) are equal; hence, distance between the subsurface reflection points is half of the distance between the detectors. Upper part of the Figure shows the theoretical hyperbola of the reflection arrivals and the direct arrival between source point and detectors



Single and Multichannel Seismics, Fig. 3 The figure shows four subsequent shot gathers recorded by a 48-channel recording system. Upper series displays the shot gathers in red-blue color display with wiggle trace overlay, while the lower series shows the same traces in

wiggle trace – varying area (WT/VA) display. Seismic energy was generated by an air gun and the direct arrival, bubble energy and reflections are clearly visible on all of four shot gathers

Commonly Used High-Resolution Single- and Multichannel Seismic Sources

The *CHIRP* systems use sweep signals for generating pressure waves. Electromagnetic sweep signals are commonly used in

commercial and military radar systems, while acoustic sweeps are used in sonar systems for marine applications, and for *vibroseis* on land. *CHIRP* systems (e.g., Schock and LeBlanc 1990; Gutowski et al. 2002) became very popular in marine surveys, while *vibroseis* is the most commonly used

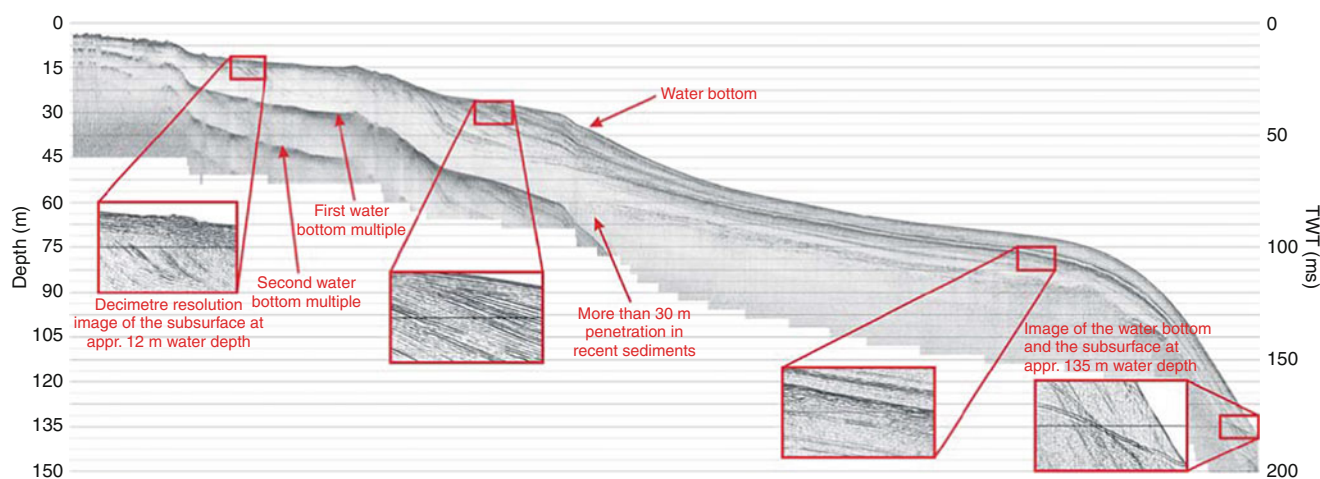
seismic source on land. Marine CHIRP systems typically operate in the 400 Hz–25 kHz frequency range, providing decimeter resolution of the sub-bottom layers down to a few meters or maximum few tens of meters depth in unconsolidated sediments. Advantage of the CHIRP systems is the high repeatability and good S/N (signal-to-noise) ratio. Disadvantage is the loss of phase information of the signal due to the cross-correlation during signal recording. CHIRP sources are used mainly for single-channel seismic recording; however, high-resolution 3D Chirp system has also been developed and used for decimeter-scale object detection (Vardy et al. 2008; Plets et al. 2009).

Boomers are electromagnetically driven sources consisting of a flat coil and a metal plate below the coil (Edgerton and Hayward 1964). Discharge of a high-voltage capacitor bank through the coil generates Eddy-current in the metal plate. The generated Eddy-current violently repels the plate from the coil. Pulling back the plate by strong springs or rubber bands creates a cavitation in the water acting as a sound source. Energy of the source depends on the capacitor bank, which is typically ranging between few 100 and 1,000 J, but several kJ systems have also been used. Boomers generate highly repeatable source signatures in the frequency range of 300 Hz–20 kHz with decimeter resolution and several tens of meters penetration. Innovative system design, such as the IKB-Seistec system utilizing a boomer source and a hydrophone group shielded by a focusing cone (Simpkin and Davis 1993), can further improve S/N ratio of boomer-based systems providing exceptional combination of resolution and penetration, as shown in Fig. 4. Boomers used to

be typical single-channel seismic sources; however, multi-channel systems developed for high-resolution 3D data acquisition are also built around boomer sources (Müller et al. 2009).

In *sparker* sources, similar to the sparking plug of an engine, discharge of a capacitor bank creates a spark between the positive and negative electrodes of the sparker. This spark vaporizes water between the electrodes and generates a pressure impulse. Energy and shape of the sparker wavelet are influenced by the physical design of the sparker, but is mainly controlled by the capacitance and the voltage of the high-voltage capacitor bank, commonly called “shot box.” Typical energy values per shot range between few 100 J and several thousand or even tens of thousands of Joules (Allen 1972). Large energy sparkers are rarely used nowadays as air gun sources mainly replaced them. Small-energy sparkers are still in use for engineering and research applications as they provide an alternative source fitting between boomers and small air guns. Sparkers are cost-effective to run, but shape of the wavelet, directivity, and repeatability are better controlled for other sources. Use of the sparker is also limited by the conductivity of the water. It is normally used in saltwater environment as in fresh water, extra care has to be taken to increase the conductivity of the fluid between the electrodes. This can be achieved, for example, by wrapping the sparker in a plastic bag and adding salt inside.

Air gun is a pneumatic seismic energy source generating pressure waves by the sudden release of high-pressure (70–200 bar) air into the water (Giles 1968). The gun is charged with high-pressure air from a compressor or air



Single and Multichannel Seismics, Fig. 4 Single-channel boomer profile recorded in the Bay of Napoli showing decimeter resolution and more than 30 m penetration. Comparing the left and right inserts, the attenuation of high frequencies and decreasing of seismic resolution with

depth are also visible. High resolution of the seismic combined with gravity coring enabled exceptional dating along the recorded profiles. Details of the research can be found in Sacchi et al. 2009

container and explosively releases the air from its chamber. The release of the high-pressure air produces a shock wave followed by bubble forming and several oscillations resulting from the expansion and collapse of the air bubble. Good example of bubble energy from the first oscillation is shown in Fig. 3.

Energy and frequency content of the air gun are determined by the volume of the gun, the pressure of the air and tow depth of the gun. High-resolution surveys require the smallest volume guns, typically 5–20 in³ (1 in³ = 16.39 cm³); however, gun clusters of 1,000–10,000 in³ are also commonly used during surveys with deeper targets. High-frequency content of small air guns can exceed 1,000 Hz, and large volume guns or gun clusters may provide enough energy for several kilometer deep penetration. Air guns are highly repeatable sources with considerable literature on

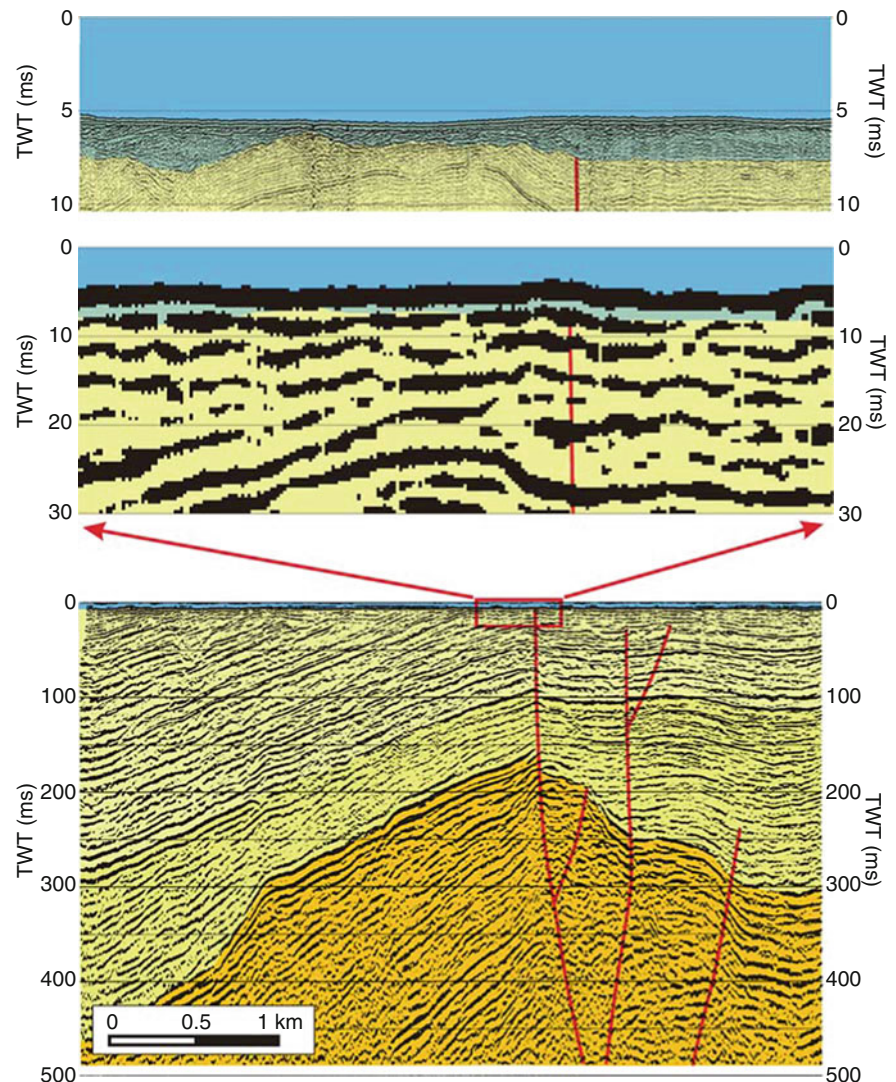
measuring and designing their wavelet (Ziolkowski et al. 1982).

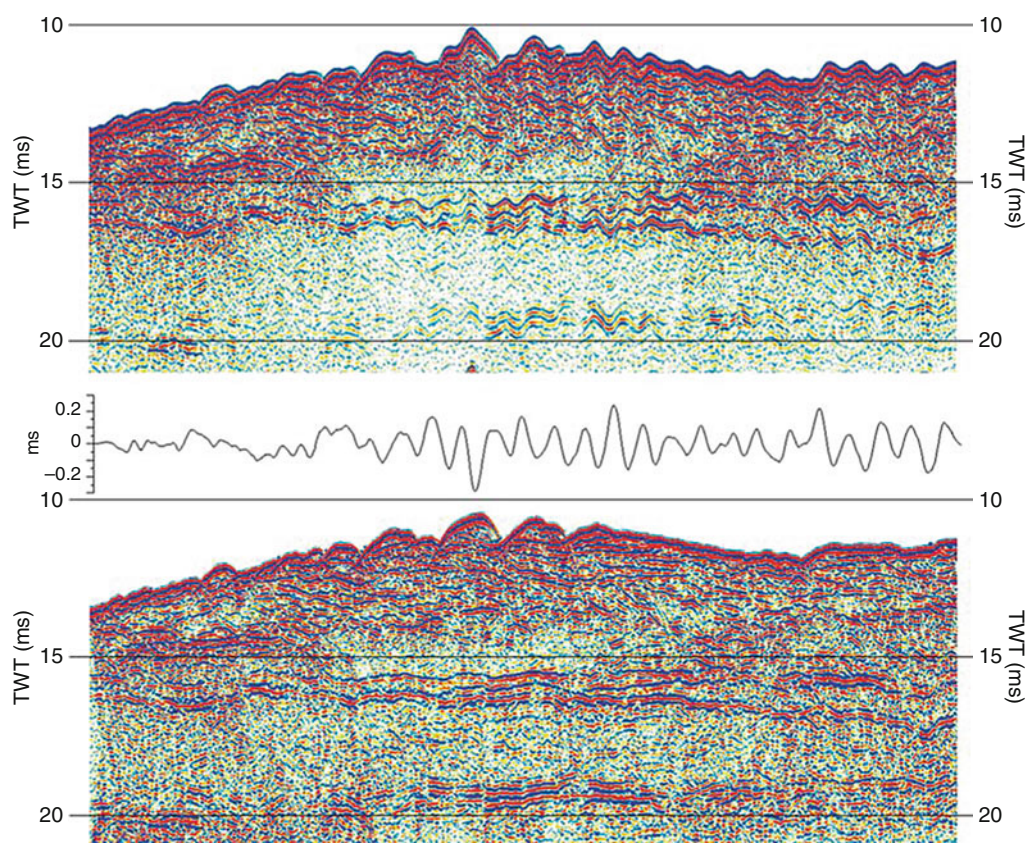
Bubble oscillation can be decreased or even completely diminished by the timely injection of a second volume of high-pressure air. Guns operating under this principle are called *Generated Injection Air Guns* or simply *GI guns*. GI guns have two chambers. The primary chamber (generator) produces the actual pressure impulse, while the secondary chamber (injector) is used to inject a second air volume near the maximum expansion of the first bubble in order to prevent its collapse.

Air gun and *GI* sources due to their higher energy and deeper penetration are mainly used for multichannel recording. A good comparison of single-channel boomer profile and a multichannel air gun profile is shown in Fig. 5.

Single and Multichannel

Seismics, Fig. 5 Bottom image shows a multichannel seismic profile recorded on river Danube, Hungary. Middle image is an optical zoom of the area in the red rectangle. Top image shows the corresponding single-channel boomer profile recorded along the same section of the river. Blue color indicates water, green, recent river sediments, while shades of yellow, Pliocene-Pleistocene sediments. Note the order of magnitude difference in resolution but also in penetration between the two profiles





Single and Multichannel Seismics, Fig. 6 Top image shows single-channel boomer data recorded on river Rhine with several decimeter amplitude waves. Wave action has been estimated from the cross-correlation of the seismic traces and the necessary time shift is displayed

in the middle graph. Lower image shows the seismic profile after correction for surface wave action. Note that the velocity pull-up and push-down effects due to the velocity difference between water and sediments are clearly visible along the reflector just below 15 ms in the lower image

Processing Considerations for High-Resolution Single- and Multichannel Seismic Profiles

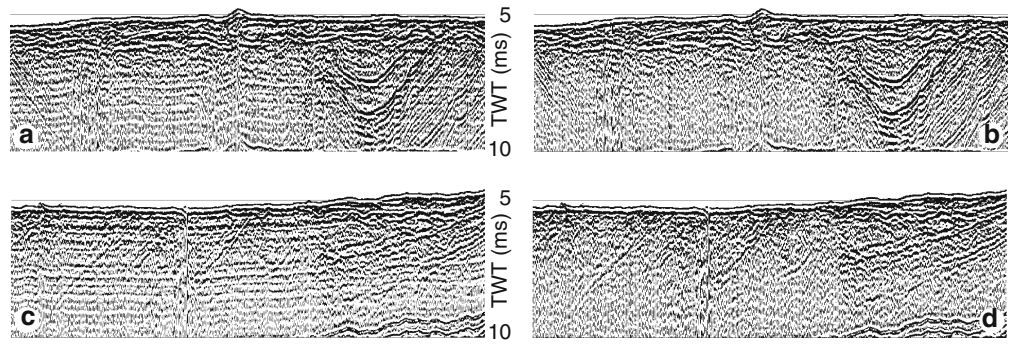
Static correction is always a key issue for proper imaging of the subsurface with seismic waves. This is especially true for high-resolution surveys. Offshore seismic surveys present a special static problem due to the wave motion. This equally affects single- and multichannel surveys, but the higher the resolution of the seismics, the more severe the problem can be. Coherency of the reflections in a decimeter resolution boomer survey can be severely affected by surface waves. Heave motion detectors offer a compensation via measuring the wave motions together with the seismic profile; however, if no heave motion is available for the survey, the necessary correction can be calculated from the data. Example of this is shown in Fig. 6.

High repeatability of many of the sources (boomer, air guns) offers good opportunity for wavelet determination and

application of deterministic *deconvolution* using the calculated wavelet. This is illustrated in Figs. 7 and 8 showing two seismic profiles before and after deterministic wavelet deconvolution and the calculated wavelets and deconvolution operators.

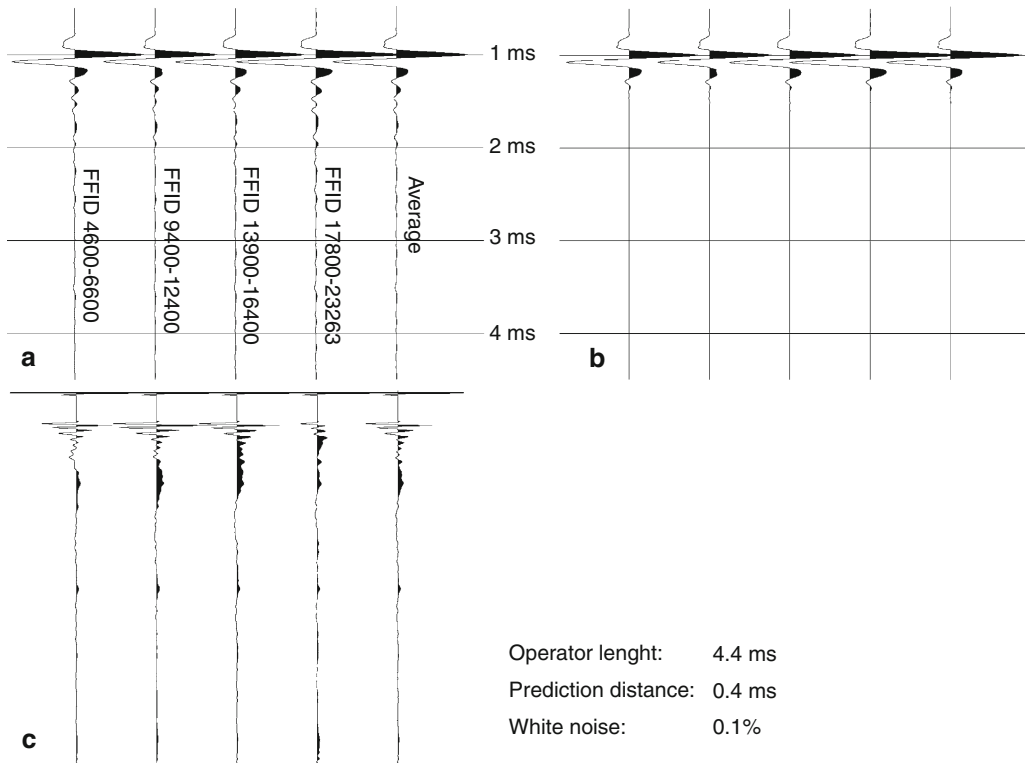
Migration is also a key processing step not only for the multichannel, but also for the single-channel seismic profiles. In the latter case, estimation of the migration velocity is more complicated, as there is no a priori information available from stacking velocity analysis. Velocity estimates from other measurements or migration velocity analysis can be used for migrating single-channel data. An example of this is shown in Fig. 9.

Multiples generated by the free water surface present a significant problem for single-channel seismic profiles. Several processing algorithms exist for multichannel seismic data in order to suppress multiple energy, but almost all of them fail in case of single-channel seismic profiles. Figure 10



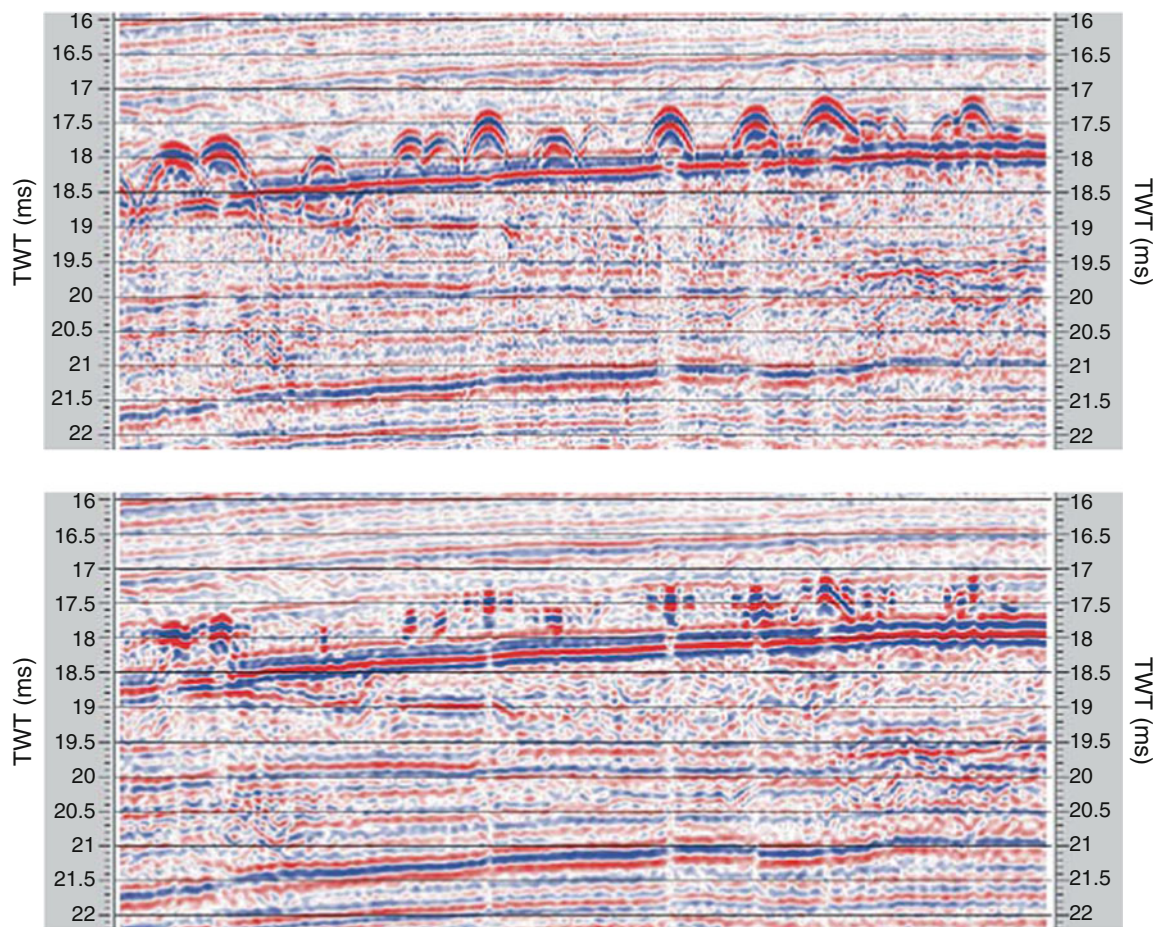
Single and Multichannel Seismics, Fig. 7 Two images on the left show profiles without wavelet deconvolution applied. Note the “ringing” of the wavelet running parallel with the water bottom and intersecting

genuine reflectors below the water bottom. Profiles on the right have been deconvolved using wavelet-based deterministic deconvolution. The ringing of the wavelet is completely eliminated



Single and Multichannel Seismics, Fig. 8 Wavelets before (a) and after (b) deterministic deconvolution applied and the deconvolution operator applied (c)

Operator length:	4.4 ms
Prediction distance:	0.4 ms
White noise:	0.1%



Single and Multichannel Seismics, Fig. 9 Upper and lower images show the same detail of the seismic profile before and after migration, respectively. Note the collapse of the diffraction hyperbolae around

18 ms TWT. Lateral resolution of the profile is significantly improved by the migration process

shows a single-channel profile recorded in lake Balaton, Hungary with significant multiple energy.

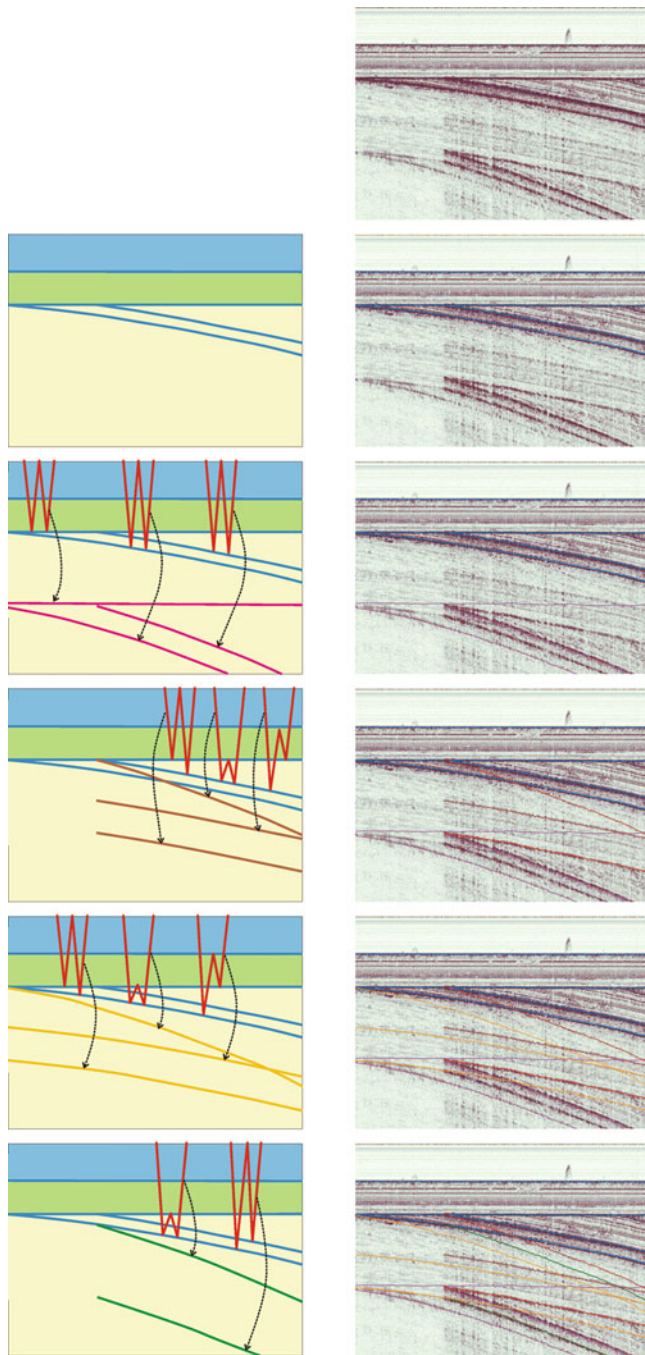
Summary

Seismic surveys provide one of the most detailed image of the subsurface from shallow (few tens of meters) to deep (several kilometers or even tens of kilometers) intervals. Although most of the recent seismic surveys utilize multichannel recording, high-resolution single-channel seismic surveys can also provide a cost-effective solution for offshore surveys. Main advantage of single-channel seismics is high-resolution

imaging and relative simplicity of the survey. Multichannel surveys, on the other hand, can significantly improve the signal-to-noise ratio of the seismic profile by applying multi-fold imaging and at the same time also provide additional information, for example, estimates of the velocity field.

Cross-References

- ▶ [Seismic Data Acquisition and Processing](#)
- ▶ [Seismic Velocity and Temperature Relationships](#)
- ▶ [Seismic, Migration](#)
- ▶ [Seismic, Velocity, and Density Relationships](#)



Single and Multichannel Seismics, Fig. 10 Top right image shows the original seismic profile recorded in lake Balaton, Hungary. In the lake, approx. 4 m water and 4 m soft, unconsolidated mud can be found above the older Pliocene strata. This, together with the dipping strata in the Pliocene, presents an ideal condition for multiple generation. Top sketch marks with blue the interfaces considered for multiple generation. These are: water surface, water bottom, mud bottom, and two dipping intra-Pliocene strata. First, the free surface multiples (marked by pink) are calculated for each of these interfaces. This is followed by different peg-leg multiples calculated and marked by brown, yellow, and green lines on the profile. It is very interesting to observe that very many of the reflections recorded on the seismic profile are actually multiple energies generated by the above mentioned surfaces only

Bibliography

- Allen FT (1972) Some characteristics of marine sparker seismic data. *Geophysics* 37:462–470
- Edgerton HF, Hayward EC (1964) The boomer sonar source for seismic profiling. *J Geophys Res* 69:3033–3042
- Giles BF (1968) Pneumatic acoustic energy source. *Geophys Prospect* 16:21–53
- Gutowski M, Bull J, Hestock T, Dix J, Hogarth P, Leighton T, White P (2002) Chirp sub-bottom profiler source signature design and field testing. *Mar Geophys Res* 23:481–492
- Karcher JC (1987) The reflection seismograph: its invention and use in the discovery of oil and gas fields. *Lead Edge* 6(11):10–19
- Keppner G (1991) Ludgermintrop. *Lead Edge* 10(9):21–28
- Müller C, Woelz S, Ersoy Y, Boyce J, Jokisch T, Wendt G, Rabbel W (2009) Ultra-high-resolution marine 2D-3D seismic investigation of the Liman Tepe/Karantina Island archaeological site (Urla/Turkey). *J Appl Geophys* 68:124–134
- Plets RMK, Dix JK, Adams JR, Bull JM, Henstock TJ, Gutowski M, Best AI (2009) The use of a high-resolution 3D Chirp sub-bottom profiler for the reconstruction of the shallow water archeological site of the Grace Dieu (1439), River Hamble, UK. *J Archeol Sci* 36:408–418
- Sacchi M, Molisso F, Violante C, Esposito E, Insinga D, Lubritto C, Porfido S, Tóth T (2009) Insights into flood-dominated fan-deltas: very high-resolution seismic examples off the Amalfi cliffed coast, eastern Tyrrhenian Sea. *Geol Soc Lond, Spec Publ* 322:33–71
- Schock SG, LeBlanc LR (1990) Chirp sonar: new technology for sub-bottom profiling. *Sea Technol* 31(9):35–43
- Simpkin PG, Davis A (1993) For seismic profiling in very shallow water, a novel receiver. *Sea Technol* 34(9):21–28
- Vardy ME, Dix JK, Henstock TJ, Bull JM, Gutowski M (2008) Decimeter-resolution 3D seismic volume in shallow water: a case study in small-object detection. *Geophysics* 73:B33–B40
- Ziolkowski A, Parkes G, Hatton L, Haugland T (1982) The signature of an air gun array. *Geophysics* 47:1413–1421

Slow Earthquake

Teruyuki Kato

Earthquake Research Institute, The University of Tokyo,
Bunkyo-ku, Tokyo, Japan

Definition

An earthquake is a sudden rupture of a fault embedded in the crust or upper mantle, and radiates elastic waves toward the earth's surface to vibrate the ground. Ground vibration includes wide spectrum of frequencies from tens of hertz to hundredth of seconds or lower. The ground vibration is recorded by a pendulum called a seismograph. Since the pendulum has its own frequency characteristics, components of longer periods becomes less visible and are not recognizable in the longest period.

Some types of fault ruptures have predominant component in long-period seismogram and such earthquakes are called

slow earthquakes. The earthquake that does not radiate enough energy to be recorded on seismogram is called a silent earthquake or slow slip event (SSE). In this definition, observed pre-seismic or post-seismic slip, or creep events may all be categorized in the silent earthquake or SSE.

Early Studies on Slow and Silent Earthquake

Earlier studies of slow earthquake were made by examining difference in magnitude determined by different frequency ranges (e.g., Kanamori 1972; Kanamori and Stewart 1979). Seismic magnitude of large earthquakes estimated from shorter period wave such as body waves are significantly smaller than that determined from longer period seismic waves. Ultimately, magnitude determined by seismic wave analysis is often much smaller than the moment magnitude or that estimated by the geodetic data inversion analysis that includes longest period (or permanent displacement) data. These have been interpreted such that the significant amount of moment release was achieved by the component of fault rupture that is slow enough not radiating much seismic wave. Kanamori and Stewart (1979), for example, showed that a series of such slow events were triggered after the June 6, 1960, Chilean earthquake (Mw 9.5).

One of such difference could be manifested by a type of earthquake called “tsunami earthquake.” The tsunami earthquake is characterized by anomalously high tsunami generation than that expected from body wave or surface wave magnitude. This can be interpreted by source characteristic such that the fault rupture was slow enough not radiating shorter seismic wave but rapid enough to generate tsunami. The 1896 Sanriku, Japan, earthquake (Ms 7.2, Mw 8.0), 1992 Nicaragua earthquake (Ms 7.0, Mw 7.6) are examples (e.g., Kanamori and Kikuchi 1993).

If the rupture is much slower and does not radiate any seismic energy, then it is called a silent earthquake. Beroza and Jordan (1990) tested if such very slow or silent earthquake could be detected using spectra of Earth’s free oscillations. Another trial is the comparison of long-term moment release rate at subduction zones. If all of strain accumulated along plate boundaries is released by seismic energy, then the released moment at the time of earthquake should be equal to the accumulated moment. Kawasaki et al. (2001) examined this hypothesis and found that the total moment released by earthquakes along the plate boundary is far smaller than the totally accumulated moment in an interseismic stage. This means that there are other unknown slow processes that release remaining moment at the plate boundaries.

One of possible mechanisms of slow strain release would be post-seismic transient displacement. Earlier discovery of post-seismic crustal deformation was found due to the 1946 Nankai, Japan, earthquake (Mw 8.1) (e.g., Okada and Nagata

1953). The cause of such a post-seismic crustal deformation was interpreted as a retarded slip after the earthquake (Fitch and Scholz 1971). Kawasaki et al. (1995) asserted that a post-seismic slow rupture was generated after the 1992 Sanriku, Japan, earthquake (Ms 6.9), by examining the strain records observed at the Esashi, Northern Japan, station. However, the question if such post-seismic transients on the records of strain sensors might stem from instrumental drifts after a strong shake was not well resolved.

More direct and convincing finding of slow slip events was brought in 1990s by the advent of the Global Positioning System (GPS). Geographical Survey Institute of Japan (GSI) first deployed nationwide GPS network to monitor crustal deformation of the Japanese Islands starting in 1992. The network was much densified in 1996 with more than 600 permanent GPS sites together with near real-time monitoring facility. The network is called GPS Earth Observation Network System (GEONET). It has been further augmented to more than 1,200 sites all over the Japanese Islands. This nationwide array of continuous GPS observation provided such a powerful tool to discover slow slip events around the Japanese Islands, as described in the following section.

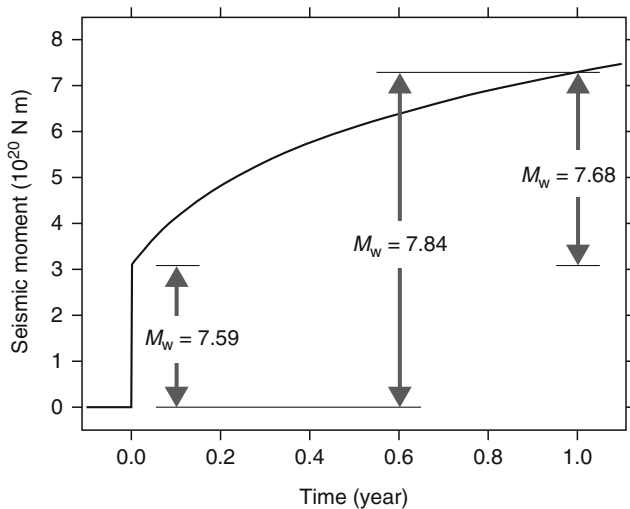
Also, the USA deployed a large array of GPS together with strain and tilt meters along the western coast of the North American continent and is called Plate Boundary Observatory (PBO), which also enabled discoveries of slow slip events.

Rapid developments of GPS arrays all over the world, in particular along the Pacific Rim areas, enabled us to discover that silent earthquakes or slow slip events (SSE) are occurring at various areas around the globe, in particular, along the subducting plate interface. Comprehensive description on where do those slow events are taking place are found in Schwartz and Rokosky (2007) and Schwartz (2009).

Long-Term Slow Slip Events

A variety of slow slip events, including post-seismic slip, whose time durations are months to years have been discovered along the subducting plate boundaries.

The first eminent case of large slow slip events found using GPS network data would be the post-seismic slip due to the 1994 Sanriku-haruka-oki earthquake (Heki et al. 1997). Heki et al. (1997) analyzed the GEONET data and found that the moment released by the post-seismic slow slips is bigger than that released at the time of earthquake (Fig. 1). Then, Hirose et al. (1999) found a slow slip event in the Bungo Channel, western part of Japan, that occurred after the two 1996 Hyuganada earthquakes of Mw 6.6 on October 19 and Mw 6.7 on December 3 that occurred in the south of the region. This slip continued about 1 year in 1997 and was assumed to be due to a slip on the subducting Philippine Sea plate. Their geodetic inversion suggested that the slip occurred on about



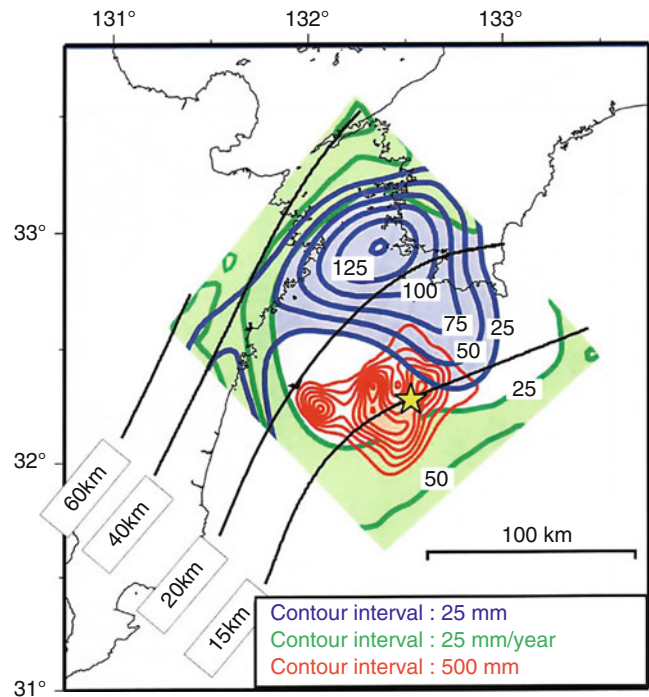
Slow Earthquake, Fig. 1 Moment release rate due to the 1994 Sanriku-Haruka-oki, Japan, earthquake. Figure suggests that the moment released in 1 year after the earthquake is bigger than the co-seismic moment release (Heki et al. 1997)

60 × 60 km surface with the maximum fault slip of about 18 cm, which is equivalent to M_w 6.6 earthquake (Fig. 2).

Other examples of slow events found in the Japanese region include (1) east off the Boso Peninsula where slow slip events have repeated in 1996 and in 2002 and they were followed by earthquake swarms (Sagiya 2004; Ozawa et al. 2003) and (2) a post-seismic transient slip after the 2003 Tokachi earthquake (M_w 8.0) (Miyazaki et al. 2004).

Lowry et al. (2001) found a transient movement at a continuous GPS site established at Cayaco, Guerrero region, Mexico in early 1998. The area has been considered as a seismic gap. The event lasted about several months with 2 mm east, 26 mm south, and 16 mm up of displacements. They suggested that the displacement is consistent with a slip along the subduction interface that propagated from east to west. The transient displacement repeated in 2001–2002 for about 6 months. Kostoglodov et al. (2003) and Yoshioka et al. (2004) further investigated the mechanism of the event. Another repeated transient event in Mexico has been found in Oaxaca segment, hundreds of kilometers southeast of Guerrero segment, with more dense GPS array, and was studied by Brudzinski et al. (2007) and Correa-Mora et al. (2008).

Other long-term slow events have been found in Costa Rica (Protti et al. 2004), Alaska-Aleutian (Ohta et al. 2006), and New Zealand (Douglas et al. 2005). SSEs have also been found along the shallower part of the San Andreas fault (e.g., Wesson 1987; Gladwin et al. 1994; Linde et al. 1996), and in Italy (Crescentini et al. 1999; Amoruso et al. 2002). Readers are asked to refer to Schwartz and Rokosky (2007) and Schwartz (2009) for a detailed review.

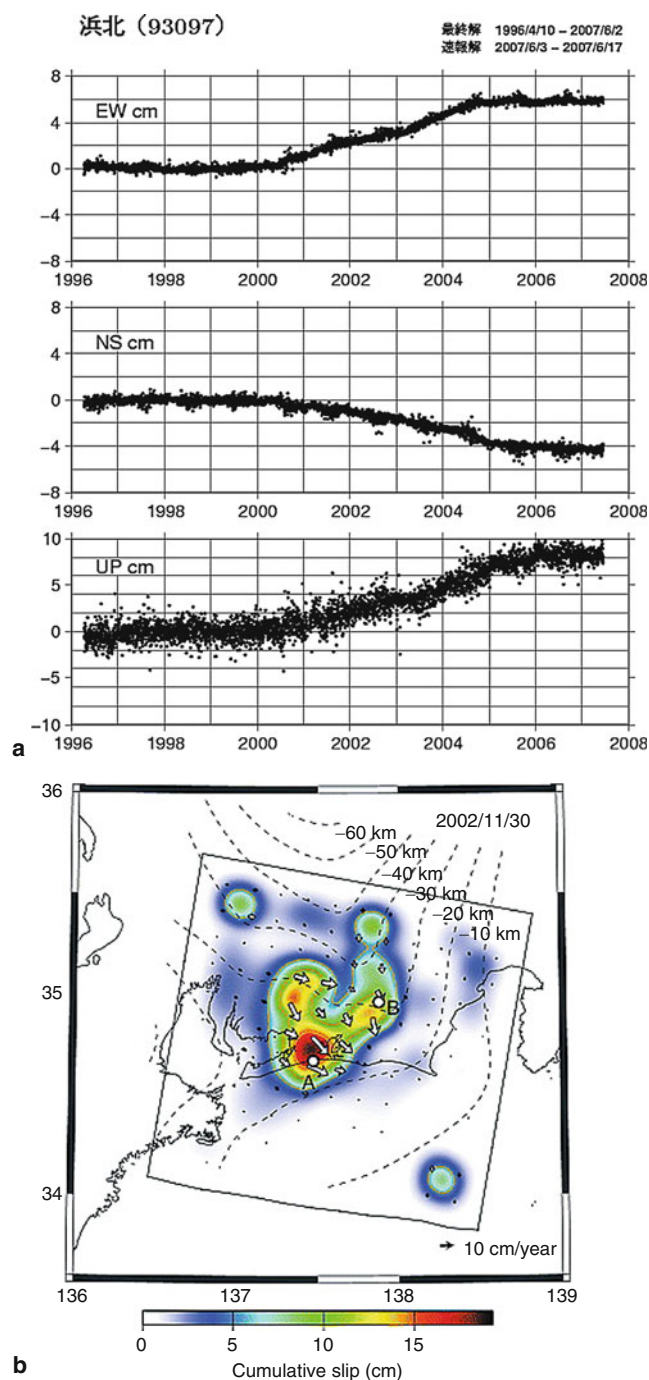


Slow Earthquake, Fig. 2 Slip distributions on the subducting plate interface shown by equi-depth contours, projected on the earth's surface: (Blue) The slip distribution due to the 1997 slow event at the Bungo Channel, Japan (After Yagi and Kikuchi 2003). (Red) the slip distribution due to the 1968 Hyuganada earthquake, and (Green) area of steady slip. (Courtesy by Y. Yagi.)

Yagi et al. (2001), and Yagi and Kikuchi (2003) used GPS data to examine areal distribution of co-seismic slip, post-seismic slip, and slow slip events for two Hyuganada earthquakes of 1996 (M_w 6.6 and M_w 6.7). They found that these areas complementarily share the areas and postulated that this complementarity may come from different constitutive parameters of slip at respective regions (Fig. 2). A similar complementary region of seismic slip and aseismic slip is found also in the northeastern Japan (Yagi et al. 2003). Correa-Mora et al. (2008) indicated that the repeated transients at the Oaxaca segment occurred at the lower extension of the locked part of the seismogenic zone. These lines of evidence may suggest distinct physical rock properties along the different segments of the subducting plate interface.

The largest and the longest slow slip event, that has ever been found, lasted for about 5 years. It was recorded along the northern margin of the subducting Philippine Sea plate, namely, Tokai district (Fig. 3). The event started in the middle of 2000 and continued until mid-2005. Both Ozawa et al. (2002) and Miyazaki et al. (2006) employed a state-space model to delineate temporal evolution of the slow slip event (Segall and Matthews 1997). The slip amounted to be about several tens of centimeters, which is equivalent to

moment magnitude of about 7.2–7.3 earthquake. Figure 3 shows a sample time series of coordinate change at a GEONET station that is located right above the slowly slipping region.



Slow Earthquake, Fig. 3 (a) A sample GPS time series at the Hamakita site for the Tokai slow slip event. Permanent trend for the period from 1996 to 2000 was removed. Slow slip event was attributed in the period from late 2000 to early 2005. (From webpage of the Geographical Survey Institute) (b) Cumulative slip inferred from the extended Network Inversion Filter. Arrows show the direction and magnitude of slip of the upper plate (Miyazaki et al. 2006)

Low-Frequency Tremor and Short-Term Slow Slip

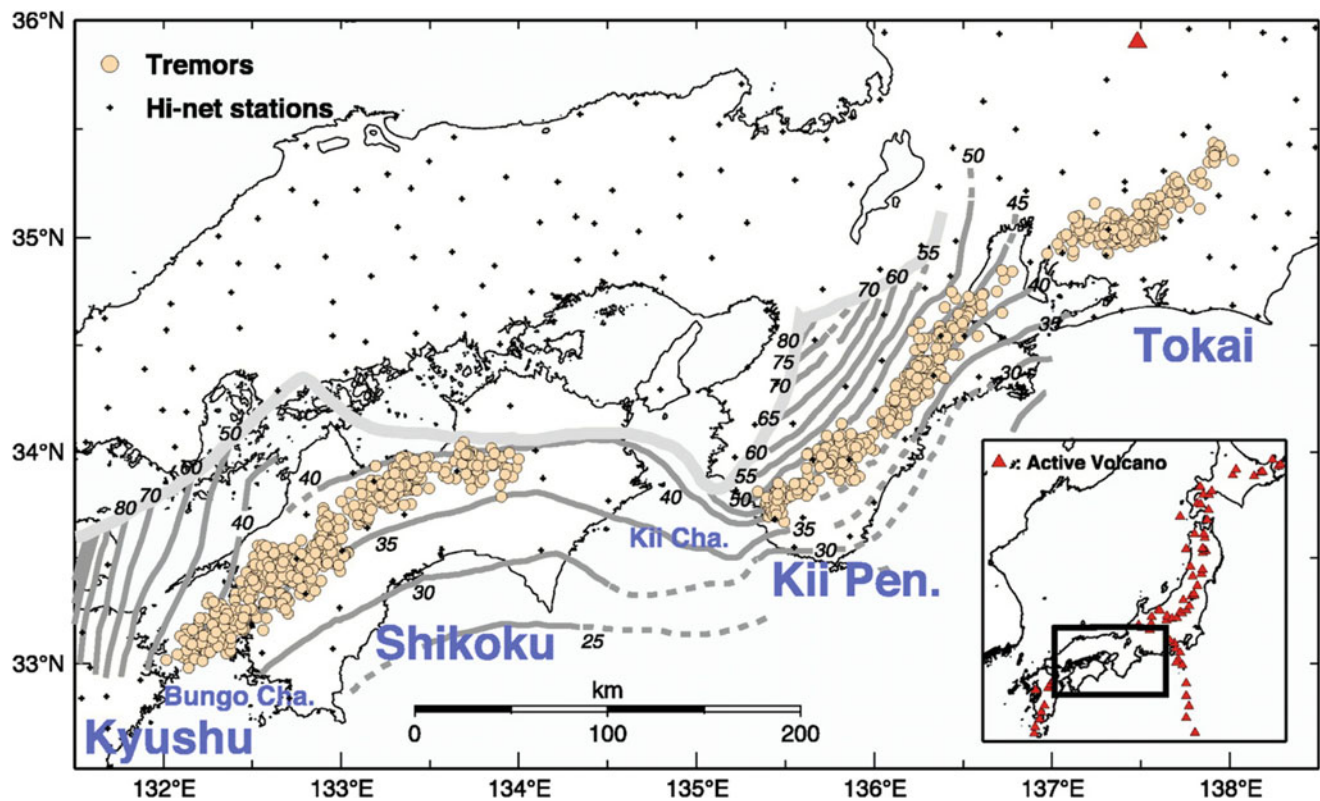
Obara (2002) discovered so-called deep low-frequency tremor using a nationwide seismic array in Japan, which is named as HiNet (High Sensitivity Seismograph Network) constituting of about 600 sites all over Japan. The deep low-frequency tremor is about 1 h of tremor activity, which is located nearly at the plate interface of 35–40 km in depth (Fig. 4). The tremor has 0.2–2 Hz in predominant frequency. And then, the tremor was found to be accompanied by small silent slips of a few to several days (Obara and Hirose 2006; Hirose and Obara 2006).

Similar nonvolcanic tremor accompanied by short-term slow events were found also in the Northwestern Pacific (Dragart et al. 2001; Rogers and Dragart 2003; Miller et al. 2002). Dragart et al. (2001) indicated that the SSE rather regularly repeats with interval of about 14 months and is associated with seismic swarm activities (Fig. 5). The estimated slip is about 3 cm. Such slow events associated with swarm activities are called episodic tremor and slip (ETS) (Rogers and Dragart 2003).

Hirose and Obara (2005) found that mid-term (about a month) and short-term (1 or 2 weeks) of slow slip events are accompanied by low-frequency tremors. The short term silent slip was too small to be monitored by GPS but can be observed on highly sensitive tiltmeters, which are mostly embedded in HiNet boreholes. The estimated amount of slip was only 1–2 cm. They also found repeated similar short-term slow slip events associated with low-frequency tremors at the Tokai region (Hirose and Obara 2006; Fig. 6). Observed slow events were modeled by a slip at the depth of 25 km and the slip amounted only 1.2 cm.

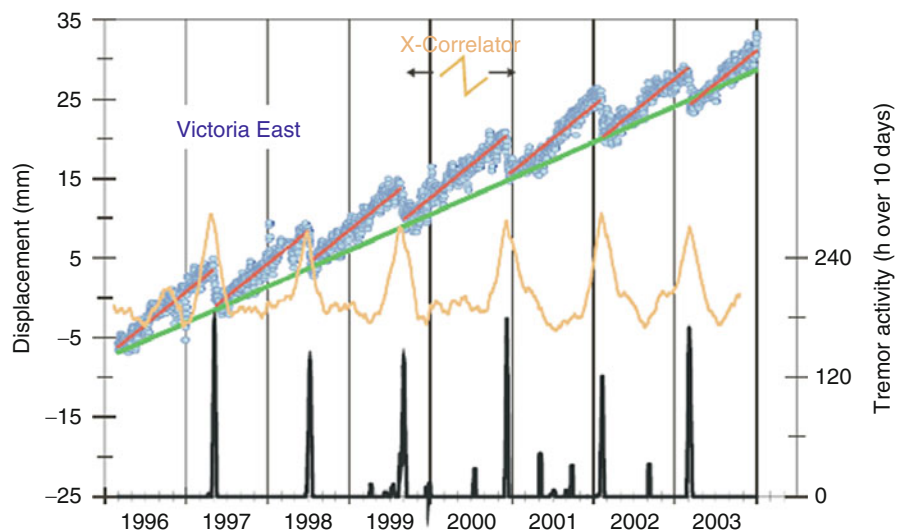
Several lines of evidence from seismological studies suggest that the occurrence of low-frequency tremors and slow slip events may somehow be related to the existence of fluid flow. Shelly et al. (2006) and Obara and Hirose (2006) suggested that episodic tremors that include low-frequency earthquakes are generated by a slow shear slip on the plate interface. Fluids might play a key role for the generation of such low-frequency earthquakes. Matsubara et al. (2009) and Kato et al. (2010) also indicated that the high-pressure fluids released by dehydration from the subducting oceanic crust generates those tremors and slow slip events.

In addition to the low-frequency tremor that occurs at a deeper extension of the seismogenic zone, low-frequency earthquakes (LFE) of a characteristic period of 10–20 s have also been observed at very shallow areas at the updip extension of the seismogenic zones along the subducting plate. They are found along the Nankai trough (e.g., Ishihara 2003; Obara and Ito 2005) and at the junction between the Japan trench and the Kuril trench (Asano et al. 2008). Considering that the LFEs along the Nankai trough are occurring within the accretionary prism and have higher dip angle



Slow Earthquake, Fig. 4 Distribution of low-frequency tremor shown in dots (Obara 2002)

Slow Earthquake, Fig. 5 Frequently and regularly repeating slow slip events and low-frequency tremors along the Cascadia subduction zones (Dragart et al. 2004)

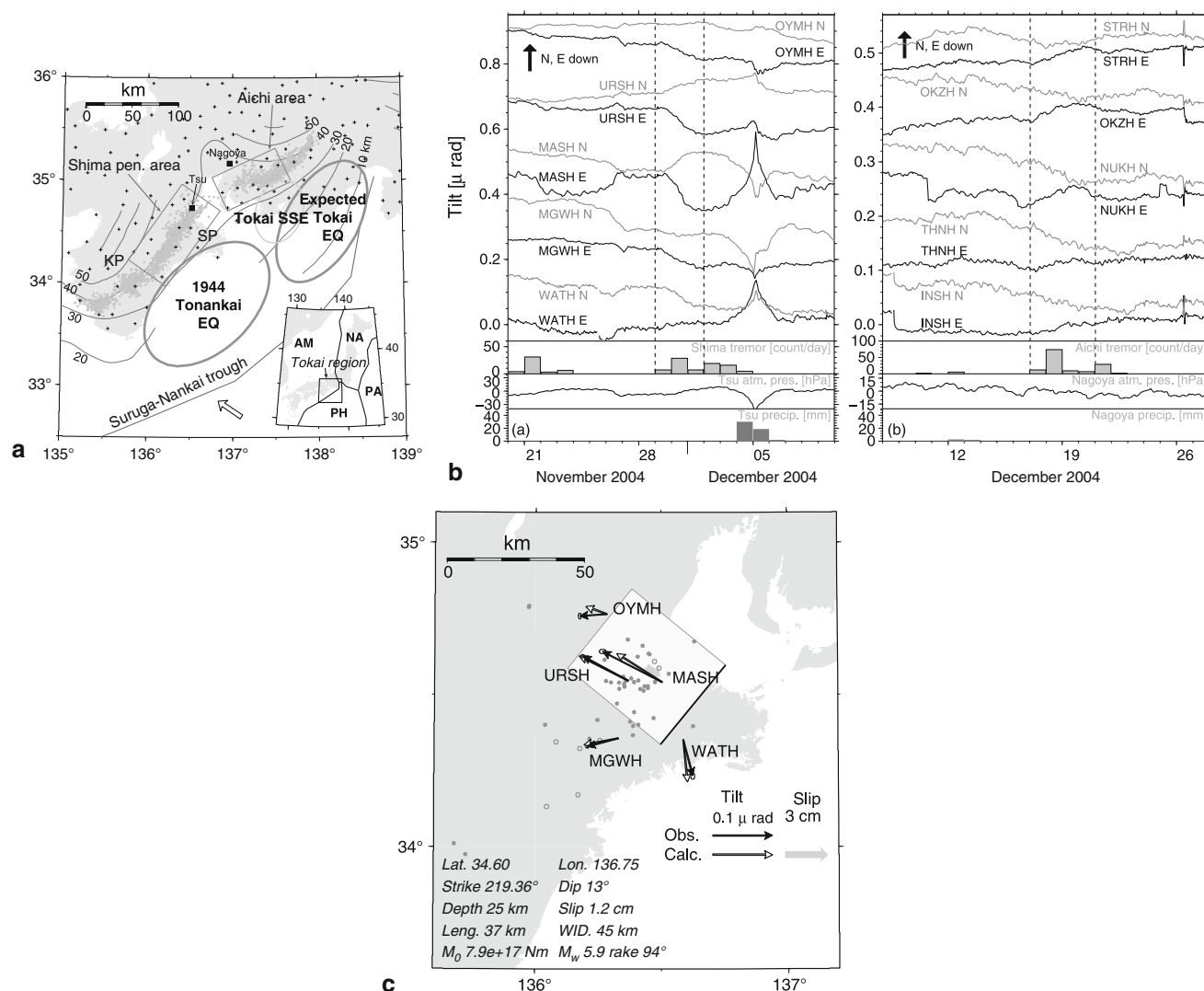


compared with the subducting plate interface, Ito and Obara (2006a) hypothesized that they may occur in the splay faults (or out-of-sequence thrusts) in the wedge.

Existence of fluid would have a key role for generating such low-frequency events, for both deep and shallow low-frequency earthquakes (e.g., Davis et al. 2006; Ito and Obara 2006b).

Pre-Seismic Slip

A number of possible pre-seismic slips have been reported. Roeloffs (2006), among others, compared ten distinct examples of reported pre-seismic deformation rate changes before large earthquakes. For example, anomalous crustal



Slow Earthquake, Fig. 6 (a) Map of studied area. (b) Observed tilt changes in November and December 2004, due to short-term slow slip events in the Tokai region. Periods of slow events are bounded by dashed

lines. Recorded low-frequency tremors are also shown, (c) Modeled source region and estimated parameters. Observed and predicted tilt changes are shown in vectors (Hirose and Obara 2006)

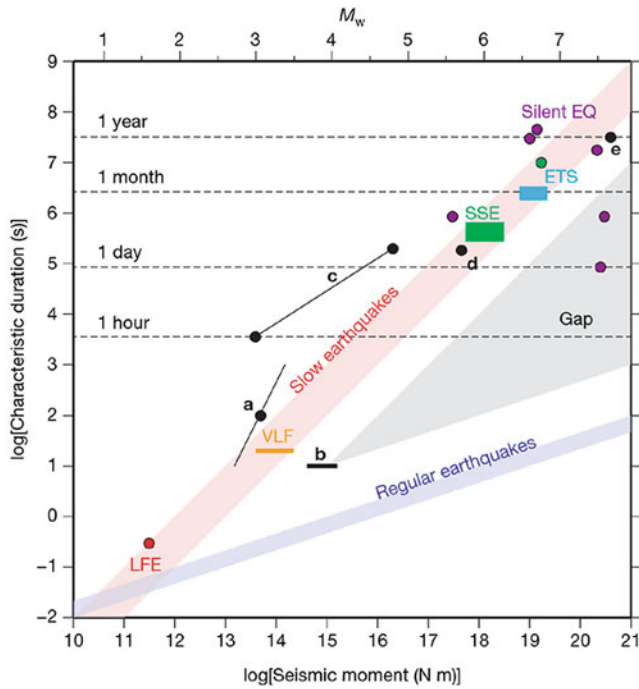
deformations were observed before the 1944 Tonankai and 1946 Nankai earthquakes in Japan. Linde and Sacks (2002) showed that these deformations are consistent with the assumed aseismic slip of about 2 m in the down dip extension of the seismically slipped interface. Kanamori and Cipar (1974) and Linde and Silver (1989) postulated that a slow event had preceded the 1960 Chile earthquake.

It might be pointed out that the mechanism of precursory slow slip would be physically the same as slow slip event. Only the difference would be that while the precursory slip is an accelerating phase toward the rapid rupture, slow slip event is the slip that is not grown to a rapid rupture. Studies to clarify the mechanism that creates such difference are yet to be conducted in the future.

Scaling Relation

An important question raised is if the discovered slow earthquakes or slow events are just the slow version of “regular” or “normal” earthquakes. If this is the case, those slow events may have the similar scaling relations as the other regular earthquakes. Ide et al. (2007), however, suggested that the slow earthquakes including silent earthquake, slow slip, etc., had different scaling relation and spectral behavior compared with normal earthquakes (Fig. 7). They suggested that the characteristic duration (T sec) and the released moment (M_0 Nm) has linear relationship:

$$M_0 \approx T \times 10^{12-13} \quad (1)$$



Slow Earthquake, Fig. 7 Comparison between seismic moment and the characteristic duration of various slow earthquakes. LFE (red), VLF (orange), SSE (green), and ETS (light blue) are low-frequency earthquake, very low frequency earthquake, slow slip event, and episodic tremor and slip, respectively. See Ide et al. (2007) for more details

While this relation for the regular earthquake is

$$M_0 \approx T^3 \times 10^{15-16} \quad (2)$$

Thus, the characteristics of moment rate spectrum are also different between these events.

This finding may tell us that the slow events are not just a slower extension of seismic slip but rather have a physically different nature. They suggest that such differences would be interpreted by “a constant low-stress drop model or a diffusional constant-slip model” (Ide et al. 2007). Though the idea is yet to be tested, this might bring us a new insight for understanding the mechanism of fault slip or “earthquake” itself.

Mechanism of Slow Earthquake

The mechanism of fault slip has been extensively studied in the field of rock mechanics including theoretical treatments. It is inevitable to clarify the constitutive relations of fault ruptures to understand what mechanism controls the occurrence of slow earthquake (or earthquake itself). Numerous

experimental and theoretical studies have been conducted for this purpose in decades. Among various proposed constitutive laws, the rate- and state-dependent friction law has been widely accepted (Dieterich 1979; Ruina 1983). The law is expressed by the following equations:

$$\mu\left(\frac{\tau}{\sigma}\right) = \mu^- + a \ln\left(\frac{V}{V^*} + \theta\right) \quad (3)$$

$$\frac{d\theta}{dt} = -\left(\frac{V}{L}\right) \left[\theta + b \ln\left(\frac{V}{V^*}\right)\right] \quad (4)$$

Where, μ is friction coefficient, τ is shear stress, σ is normal stress, μ^- is friction coefficient at V^* where V^* is constant, L is characteristic slip distance, and θ is so-called state variable that represents the state along the fault surface. Coefficients a and b are constant. When the state θ does not change with time, it is called steady state and the friction coefficient in such steady state μ^{ss} is shown as

$$\mu^{ss} = \mu_0 + (a - b) \ln\left(\frac{V}{V^*}\right). \quad (5)$$

A simple spring-slider model is often used to model the fault slips (Ruina 1983; Rice and Ruina 1983). The system is characterized by the normal stress, shear stress, and the stiffness (or spring constant) k . Studies suggest that the type of slip, either rapid or slow slip, is controlled by several parameters. From Eq. (5), it is suggested that, if $(a - b)$ is negative, the friction coefficient μ_{ss} becomes smaller when slip velocity V increases, which is called velocity weakening. In case of velocity weakening, system stiffness k controls the slip mode; if k is smaller than a critical value k_c which is given by $(b - a) \sigma/L$ (e.g., Dieterich 1979), then the slip is unstable. Even in case of velocity weakening, the slip becomes slow if $(a - b)$ is close to zero. When $(a - b) > 0$, it is called velocity strengthening and the slip is decelerated.

Marone et al. (1991) used the case of velocity strengthening to interpret the post-seismic transient slow slip along the San Andreas fault and estimated the thickness of surface velocity strengthening layer to be 2–5 km. Yoshida and Kato (2003) used two tethered spring-sliders model to investigate fault interactions and examined what kind of conditions controls the occurrence of slow slip events. They showed that the repeated slow slip events may occur if the stiffness of the first block k is smaller than k_c and the second block is such that k is slightly bigger than k_c (which means that the condition is close to the boundary between stable and unstable slip). They also examined the case that an earthquake triggers the slow slip events in the adjacent region as was observed in the case of 2003 Tokachi-oki earthquake.

Further application of the rate- and state-dependent friction law to the numerical simulation of slip along the fault has been developed by a number of researchers. Tse and Rice (1986), for example, applied the relation to the two-dimensional San Andreas fault and suggested that unstable sliding may be limited to the depth shallower than 5–7 km and the steady slow slips prevails below 13–15 km of depth. Three-dimensional application was, for example, done by Rice (1993) and Stuart and Tullis (1995).

Concluding Remarks

Although the slow earthquake has been studied for decades, its rapid progress was realized since the advent of the Global Positioning System observations in the middle of 1990s. Slow slip events would be one of the most important discoveries in seismology in the recent decades, thanks to GPS. In addition, the discoveries of other low-frequency events by high-sensitivity seismic arrays provided another important progress in understanding the mechanism of earthquake generation.

It may be reminded that the whole earthquake cycle was once categorized into four stages, namely, interseismic, pre-seismic, co-seismic, and post-seismic stages, and they had been studied rather independently. However, these stages are now recognized as cross-sectional views of single continuous series of progress of slips along the fault under the unified governing fault constitutive law.

Synthetic image of earthquake from high frequency that has been recorded by seismometers to very low frequency including transient slow events provided seismologists with opportunities of ultimate understanding of what “earthquake” is. In order, however, to realize this, a lot of more advanced researches are to be conducted and the results should be synthesized; for example, detailed deep sounding of structures of the lower crust and upper mantle where brittle and ductile ruptures are taking place should be conducted to clarify the field of earthquakes. Also, laboratory experiments for developing more conclusive governing laws of rapid and slow ruptures are indispensable. In addition, further findings of slow slip events and slow earthquakes at various fields in the world will provide deeper insights on the mechanism of earthquake generation. Such progresses might lead to a further development toward the most difficult seismological problem of earthquake prediction.

Cross-References

- Earthquake Source Theory
- GPS, Data Acquisition, and Analysis
- GPS, Tectonic Geodesy
- Seismic Quiescence and Activation

Acknowledgments The author is indebted to Professor Naoyuki Kato and Professor Kazushige Obara of Earthquake Research Institute, the University of Tokyo, for their helpful discussions and providing information to the author. This study was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, under its Observation and Research Program for Prediction of Earthquakes and Volcanic Eruptions.

Bibliography

- Amoruso A, Crescentini L, Morelli A, Scarpa R (2002) Slow rupture of an aseismic fault in a seismogenic region of central Italy. *Geophys Res Lett* 29(24):2219. <https://doi.org/10.1029/2002GL016027>
- Asano Y, Obara K, Ito Y (2008) Spatiotemporal distribution of very-low frequency earthquakes in Tokachi-oki near the junction of the Kuril and Japan trenches revealed by using array signal processing. *Earth Planets Space* 60:871–875
- Beroza GC, Jordan T (1990) Searching for slow and silent earthquakes. *J Geophys Res* 95:2485–2510
- Brudzinski M, Cabral-Cano E, Correa-Mora F, DeMets C, Márquez-Azúa B (2007) Slow slip transients along the Oaxaca subduction segment from 1993 to 2007. *Geophys J Int* 171:523–538
- Correa-Mora F, DeMets C, Cabral-Cano E, Márquez-Azúa B (2008) Interplate coupling and transient slip along the subduction interface beneath Oaxaca, Mexico. *Geophys J Int* 175:269–290
- Crescentini L, Amoruso A, Scarpa R (1999) Constraints on slow earthquake dynamics from a swarm in central Italy. *Science* 286:2132–2134
- Davis EE, Becker K, Wang K, Obara K, Ito Y, Kinoshita M (2006) A discrete episode of seismic and aseismic deformation of the Nankai trough subduction zone accretionary prism and incoming Philippine Sea plate. *Earth Planet Sci Lett* 242:73–84
- Dieterich JH (1979) Modeling of rock friction. 1. Experimental results and constitutive equations. *J Geophys Res* 84:2161–2168
- Douglas A, Beavan J, Wallace L, Townend J (2005) Slow slip on the northern Hikurangi subduction interface, New Zealand. *Geophys Res Lett* 32(L1):6305. <https://doi.org/10.1029/2005GL023607>
- Dragart H, Wang K, James TS (2001) A silent slip event on the deeper Cascadia subduction interface. *Science* 292:1525–1528
- Dragart H, Wang K, Rogers G (2004) Geodetic and seismic signatures of episodic tremor and slip in the northern Cascadia subduction zone. *Earth Planets Space* 56:1143–1150
- Fitch TJ, Scholz CH (1971) Mechanism of underthrusting in southwest Japan: a model of convergent plate interactions. *J Geophys Res* 76:7260–7292
- Gladwin MT, Gwyther RL, Hart RHG, Brechenridge KS (1994) Measurements of the strain field associated with episodic creep events on the San Andreas fault at San Juan Bautista, California. *J Geophys Res* 99:4559–4564
- Heki K, Miyazaki S, Tsuji H (1997) Silent fault slip following an interplate thrust earthquake at the Japan trench. *Nature* 386:595–598
- Hirose H, Obara K (2005) Repeating short- and long-term slow slip events with deep tremor activity around the Bungo Channel region, southwest Japan. *Earth Planets Space* 57:961–972
- Hirose H, Obara K (2006) Short-term slow slip and correlated tremor episodes in the Tokai region, central Japan. *Geophys Res Lett* 33: L17311. <https://doi.org/10.1029/2006GL026579>
- Hirose H, Hirahara K, Kimata F, Fujii N, Miyazaki S (1999) A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, Southwest Japan. *Geophys Res Lett* 26(21):3237–3240
- Ide S, Beroza GC, Shelly DR, Uchide T (2007) A scaling law for slow earthquakes. *Nature* 447:76–79. <https://doi.org/10.1038/nature05780>

- Ishihara Y (2003) Major existence of very low frequency earthquakes in background seismicity along subduction zone of Southwestern Japan. *Eos Trans AGU* 84(46), Fall Meet Suppl, Abstract S41C-0107
- Ito Y, Obara K (2006a) Dynamic deformation of the accretionary prism excites very low frequency earthquakes. *Geophys Res Lett* 33: L02311. <https://doi.org/10.1029/2005GL025270>
- Ito Y, Obara K (2006b) Very low frequency earthquakes within accretionary prisms are very low stress-drop earthquakes. *Geophys Res Lett* 33:L09302. <https://doi.org/10.1029/2006GL025883>
- Kanamori H (1972) Mechanism of tsunami earthquakes. *Phys Earth Planet Inter* 6:346–359
- Kanamori H, Cipar J (1974) Focal process of the great Chilean earthquake May 22, 1960. *Phys Earth Planet Inter* 9:127–136
- Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake: a slow tsunami earthquake associated with subducted sediments. *Nature* 361:714–716
- Kanamori H, Stewart GS (1979) A slow earthquake. *Phys Earth Planet Inter* 18:167–175
- Kato A, Iidaka T, Ikuta R, Yoshida Y, Katsumata K, Iwasaki T, Sakai S, Thurber C, Tsumura N, Yamaoka K, Watanabe T, Kunitomo T, Yamazaki F, Okubo M, Suzuki S, Hirata N (2010) Variations of fluid pressure within the subducting oceanic crust and slow earthquakes. *Geophys Res Lett* 37:L14310. <https://doi.org/10.1029/2010GL043723>
- Kawasaki I, Asai Y, Tamura Y, Sagiya T, Mikami N, Okada Y, Sakata M, Kasahara M (1995) The 1992 Sanriku-Oki, Japan, ultra-slow earthquake. *J Phys Earth* 43:105–116
- Kawasaki I, Asai Y, Tamura Y (2001) Space-time distribution of interplate moment release including slow earthquakes and the seismo-geodetic coupling in the Sanriku-oki region along the Japan trench. *Tectonophysics* 330:267–283
- Kostoglodov V, Singh SK, Santiago JA, Franco SI, Larson KM, Lowry AR, Bilham R (2003) A large silent earthquake in the Guerrero seismic gap, Mexico. *Geophys Res Lett* 30(15):1807. <https://doi.org/10.1029/2003GL017219>
- Linde AT, Sacks IS (2002) Slow earthquakes and great earthquakes along the Nankai trough. *Earth Planet Sci Lett* 203:265–275
- Linde AT, Silver PG (1989) Elevation changes and the great 1960 Chilean earthquake: support for aseismic slip. *Geophys Res Lett* 16:1305–1308
- Linde AT, Gladwin MT, Johnston MJS, Gwyther RL, Bilham RG (1996) A slow earthquake sequence on the San Andreas fault. *Nature* 383:65–68
- Lowry AR, Larson KM, Kostoglodov V, Bilham R (2001) Transient fault slip in Guerrero, southern Mexico. *Geophys Res Lett* 28:3753–3756
- Marone CJ, Scholz CH, Bilham R (1991) On the mechanism of earthquake afterslip. *J Geophys Res* 96:8441–8452
- Matsubara M, Obara K, Kasahara K (2009) High-Vp/Vs zone accompanying non-volcanic tremors and slow-slip events beneath southwestern Japan. *Tectonophysics* 472:6–17
- Melbourne TI, Webb FH, Stock JM, Reigber C (2002) Rapid postseismic transients in subduction zones from continuous GPS. *J Geophys Res* 107(B10):2241. <https://doi.org/10.1029/2001JB000555>
- Miller MM, Melbourne T, Johnson DJ, Sumner WQ (2002) Periodic slow earthquakes from the Cascadia subduction zone. *Science* 295:2423
- Miyazaki S, Segall P, Fukuda J, Kato T (2004) Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: implications for variations in fault zone frictional properties. *Geophys Res Lett* 31: L06623. <https://doi.org/10.1029/2003GL019410>
- Miyazaki S, Segall P, McGuire JJ, Kato T, Hatanaka Y (2006) Spatial and temporal evolution of stress and slip rate during the 2000 Tokai slow earthquake. *J Geophys Res* 111:B03409. <https://doi.org/10.1029/2004JB003426>
- Obara K (2002) Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science* 296:1679–1681
- Obara K, Hirose H (2006) Non-volcanic deep low-frequency tremors accompanying slow slips in the southwest Japan subduction zone. *Tectonophysics* 417:33–51
- Obara K, Ito Y (2005) Very low frequency earthquakes excited by the 2004 off the Kii peninsula earthquakes: a dynamic deformation process in the large accretionary prism. *Earth Planets Space* 57:321–326
- Ohta Y, Freymueller JT, Hreinsdottir S, Suito H (2006) A large slow slip event and the depth of the seismogenic zone in the south central Alaska subduction zone. *Earth Planet Sci Lett* 247:108–116
- Okada A, Nagata T (1953) Land deformation of the neighborhood of Muroto point after the Nankaido great earthquake in 1946. *Bull Earthquake Res Inst* 31:169–177
- Ozawa S, Murakami M, Kaidzu M, Tada T, Hatanaka Y, Yurai H, Nishimura T (2002) Detection and monitoring of ongoing aseismic slip in the Tokai region, central Japan. *Science* 298:1009–1012
- Ozawa S, Miyazaki S, Hatanaka Y, Imakiire T, Kaidzu M, Murakami M (2003) Characteristic silent earthquakes in the eastern part of the Boso peninsula, Central Japan. *Geophys Res Lett* 30(6):1283. <https://doi.org/10.1029/2002GL016665>
- Protti M, Gonzales V, Kato T, Iinuma T, Miyazaki S, Obana K, Kaneda Y, LaFemina P, Dixon T, Schwartz S (2004) A creep event on the shallow interface of the Nicoya Peninsula, Costa Rica seismogenic zone. *Eos Trans AGU* 85(47), Fall Meet Suppl, Abstract S441D-07
- Rice JR (1993) Spatio-temporal complexity of slip on a fault. *J Geophys Res* 98:9885–9907
- Rice JR, Ruina AL (1983) Stability of steady frictional slipping. *J Appl Mech* 50:343–349
- Roeloffs EA (2006) Evidence for aseismic deformation rate changes prior to earthquakes. *Annu Rev Earth Planet Sci* 34:591–627
- Rogers G, Dragart H (2003) Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science* 300:1942–1943
- Ruina AL (1983) Slip instability and state variable friction laws. *J Geophys Res* 88:10359–10370
- Sagiya T (2004) Interplate coupling in the Kanto district, central Japan, and the Boso Peninsula silent earthquake in May 1996. *Pure Appl Geophys* 161:2327–2342
- Schwartz SY (2009) Episodic aseismic slip at plate boundaries. In: Kanamori H (ed), Schubert G (ed-in-chief), *Earthquake seismology*, vol 4. Elsevier, Oxford, pp 445–472
- Schwartz SY, Rokosky JM (2007) Slow slip events and seismic tremor at circum-Pacific subduction zones. *Rev Geophys* 45:RG3004
- Segall P, Matthews M (1997) Time dependent inversion of geodetic data. *J Geophys Res* 102(B10):22391–22409
- Shelly DR, Beroza GC, Ide S, Nakamura S (2006) Low-frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. *Nature* 442:188–191
- Stuart WD, Tullis TE (1995) Fault model for preseismic deformation at Parkfield, California. *J Geophys Res* 100:24079–24099
- Tse ST, Rice JR (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. *J Geophys Res* 91:9452–9472
- Wesson RL (1987) Modeling aftershock migration and afterslip of the San Juan Bautista, California, earthquake of October 3, 1972. *Tectonophysics* 144:215–229
- Yagi Y, Kikuchi M (2003) Partitioning between co-seismic slip and aseismic slip. *J Geogr* 112(6):828–836. (in Japanese with English abstract)
- Yagi Y, Kikuchi M, Sagiya T (2001) Co-seismic slip, post-seismic slip, and aftershocks associated with two large earthquakes in 1996 in Hyuga-nada, Japan. *Earth Planets Space* 53:793–803
- Yagi Y, Kikuchi M, Nishimura T (2003) Co-seismic slip, post-seismic slip, and largest aftershock associated with the 1994 Sanriku-haruka-oki, Japan, earthquake. *Geophys Res Lett* 30(22):2177. <https://doi.org/10.1029/2003GL018189>

Yoshida S, Kato N (2003) Episodic aseismic slip in a two-degree-of-freedom block-spring model. *Geophys Res Lett* 30(13):1681. <https://doi.org/10.1029/2003GL017439>

Yoshioka S, Mikumo T, Kostoglodov V, Larson KM, Lowry AR, Singh SK (2004) Interplate coupling and a recent aseismic slow slip event in the Guerrero seismic gap of the Mexican subduction zone, as deduced from GPS data inversion using a Bayesian information criterion. *Phys Earth Planet Inter* 146:513–530

Spherical Harmonic Analysis Applied to Potential Fields

Nikolaos K. Pavlis

Geodesy and Geophysics Basic and Applied Research,
National Geospatial-Intelligence Agency (NGA),
Reston, VA, USA

Synonyms

Fourier expansion on the sphere; Spherical harmonic decomposition; Spherical harmonic expansion; Spherical harmonic transform

Definition

A square integrable function, defined on a surface that has a one-to-one correspondence with the unit sphere, may be represented as a linear combination of *Surface Spherical Harmonic functions*. *Spherical Harmonic Analysis* (SHA) is the process by which the coefficients defining this linear combination are determined. These coefficients constitute the *Surface Spherical Harmonic spectrum* of the function. Functions that satisfy Laplace's partial differential equation are called *harmonic*. These can be decomposed into series of *Solid Spherical Harmonic functions*. Although SHA is *not* restricted to functions that represent potential fields, it plays a special role in the determination of such fields through the solution of *Boundary Value Problems* (BVP).

Introduction: Basic Formulas

A scalar function V , representing a potential field such as the Earth's gravitational or magnetic field, satisfies Laplace's partial differential equation (PDE), in the space outside of the sources generating the field. In geocentric spherical polar coordinates (r, θ, λ) , where r is the radial distance, θ is the geocentric colatitude (defined as 90° minus the geocentric

latitude), and λ is the longitude, Laplace's PDE takes the form (Heiskanen and Moritz 1967, Eq. 1–41):

$$\Delta V \equiv \nabla^2 V = \frac{\partial^2 V}{\partial r^2} + \frac{2}{r} \frac{\partial V}{\partial r} + \frac{1}{r^2} \frac{\partial^2 V}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial V}{\partial \theta} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 V}{\partial \lambda^2} = 0. \quad (1)$$

Equation 1 may be solved using the method of separation of variables. For the space outside a sphere of radius $r = a$, its solution may be written in the form (see Heiskanen and Moritz 1967 for a complete derivation):

$$V(r, \theta, \lambda) = \sum_{n=0}^{\infty} \left(\frac{a}{r}\right)^{n+1} \sum_{m=-n}^n v_{nm} Y_{nm}(\theta, \lambda). \quad (2)$$

The functions $[(a/r)^{n+1} Y_{nm}(\theta, \lambda)]$ are called *Solid Spherical Harmonic functions*. The *Surface Spherical Harmonic functions* $Y_{nm}(\theta, \lambda)$ are defined as:

$$Y_{nm}(\theta, \lambda) = \bar{P}_{n|m|}(\cos \theta) \cdot \begin{cases} \cos m\lambda & \text{if } m \geq 0 \\ \sin |m| \lambda & \text{if } m < 0 \end{cases}. \quad (3)$$

$\bar{P}_{n|m|}(\cos \theta)$ is the fully normalized *Associated Legendre function* of the first kind, of degree n and order $|m|$ (Heiskanen and Moritz 1967, Sects. 1–11 and 1–14). *Fully normalized* Associated Legendre functions are commonly used in geodesy. These are related to their unnormalized counterparts $P_{n|m|}(\cos \theta)$ by:

$$\bar{P}_{n|m|}(\cos \theta) = \sqrt{(2 - \delta_{m|0|})(2n+1) \frac{(n-|m|)!}{(n+|m|)!}} \cdot P_{n|m|}(\cos \theta), \quad (4)$$

where:

$$\delta_{m|0|} = \begin{cases} 1 & \text{if } m = 0 \\ 0 & \text{if } m \neq 0 \end{cases}. \quad (5)$$

For the surface spherical harmonic functions Y_{nm} , this normalization implies that:

$$\frac{1}{4\pi} \iint_{\sigma} Y_{nm}^2(\theta, \lambda) d\sigma = 1, \quad (6)$$

with the integration being performed over the unit sphere σ , whose area element is $d\sigma$ ($d\sigma = \sin \theta d\theta d\lambda$). Note that in geomagnetism, the *Schmidt seminormalized* Associated Legendre functions (Blakely 1995, p. 113) $\tilde{P}_{n|m|}(\cos \theta)$ are used instead. These are defined by:

$$\tilde{P}_{n|m}(\cos \theta) = \sqrt{(2 - \delta_{|m|0}) \frac{(n-|m|)!}{(n+|m|)!}} \cdot P_{n|m}(\cos \theta). \quad (7)$$

Surface spherical harmonics constitute a set of *orthogonal* basis functions on the unit sphere, i.e.:

$$\iint_{\sigma} Y_{nm}(\theta, \lambda) Y_{sr}(\theta, \lambda) d\sigma = 0 \text{ if } n \neq s \text{ or } m \neq r \text{ or both.} \quad (8)$$

This property of *orthogonality* permits the determination of the spherical harmonic coefficients v_{nm} , which appear in Eq. 2, as follows. Assume for the sake of this discussion that the function $V(r; \theta, \lambda)$ is observable everywhere on the surface of the sphere $r = a$. Eq. 2 then takes the form:

$$V(a, \theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=-n}^n v_{nm} Y_{nm}(\theta, \lambda). \quad (9)$$

In this case, the orthogonality of surface spherical harmonics implies that:

$$v_{nm} = \frac{1}{4\pi} \iint_{\sigma} V(a, \theta, \lambda) Y_{nm}(\theta, \lambda) d\sigma. \quad (10)$$

The coefficients v_{nm} constitute the *surface spherical harmonic spectrum* of the function $V(r; \theta, \lambda)$, pertinent to the surface of the sphere of radius $r = a$. It should be emphasized that *spherical harmonic expansions*, as the one given in Eq. 9, are *not* restricted to functions satisfying Laplace's equation (*harmonic functions*). Any square integrable function $f(\theta, \lambda)$, defined over a surface S that has a one-to-one correspondence with the unit sphere, may be expanded in surface spherical harmonics as:

$$f^S(\theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=-n}^n f_{nm}^S Y_{nm}(\theta, \lambda), \quad (11)$$

with the coefficients f_{nm}^S given by:

$$f_{nm}^S = \frac{1}{4\pi} \iint_{\sigma} f^S(\theta, \lambda) Y_{nm}(\theta, \lambda) d\sigma. \quad (12)$$

The superscript “S” emphasizes the dependence of the spectrum f_{nm}^S of the function $f(\theta, \lambda)$, on the particular surface S over which its values $f^S(\theta, \lambda)$ are given. Jekeli (1988) carefully distinguishes between the direct Legendre transform (i.e., the spectrum) of an arbitrary square integrable function $f(\theta, \lambda)$, defined on a surface S that has a one-to-one correspondence with the unit sphere, and the special case of the

solution of Laplace's equation when the boundary data reside on the surface of the sphere $r = a$. The former is expressed in Eqs. 11 and 12 above; the latter in Eqs. 9 and 10. Obviously the two quantities take the same form with an appropriate choice of coordinates. Furthermore, note that while rather arbitrary functions can be expanded in surface spherical harmonics, *only* harmonic functions (i.e., functions satisfying Laplace's equation) can be expanded into solid spherical harmonics, within their region of harmonicity.

The total power of the function $f(\theta, \lambda)$, defined over the surface S , is defined to be:

$$M\{[f^S]^2\} = \frac{1}{4\pi} \iint_{\sigma} [f^S(\theta, \lambda)]^2 d\sigma. \quad (13)$$

A generalization of Parseval's theorem permits the computation of $M\{[f^S]^2\}$ from the spectrum f_{nm}^S , by:

$$M\{[f^S]^2\} = \sum_{n=0}^{\infty} f_n^S = \sum_{n=0}^{\infty} \sum_{m=-n}^n [f_{nm}^S]^2. \quad (14)$$

The quantities f_n^S , defined as:

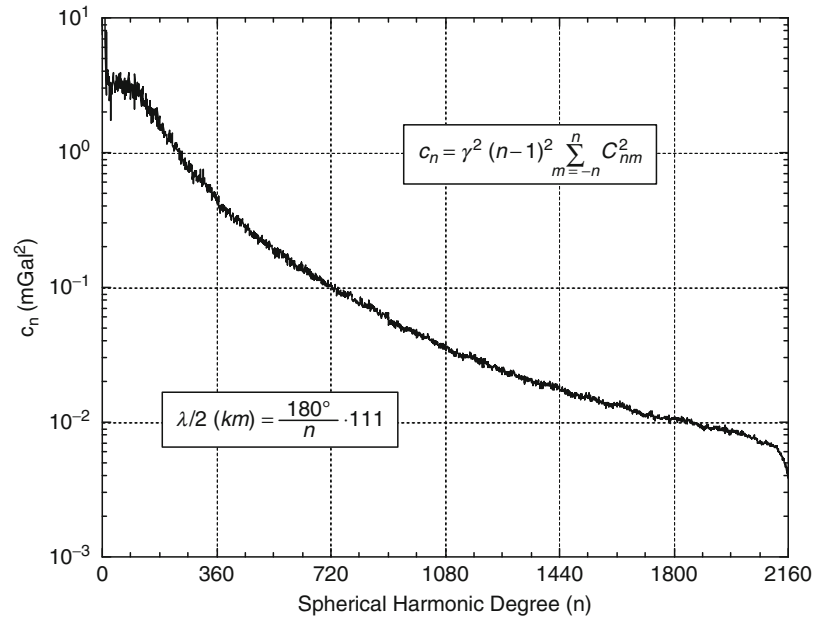
$$f_n^S = \sum_{m=-n}^n [f_{nm}^S]^2, \quad (15)$$

represent the total power (or variance) per spherical harmonic degree. In geodetic literature, they are known as *degree variances*. A direct correspondence exists between spherical harmonic degree and spatial wavelength. Therefore, the degree variances enable the study of the distribution of power within the function $f^S(\theta, \lambda)$, in terms of spatial wavelength. For example, Fig. 1 (Pavlis et al. 2008) shows the degree variances implied by the Earth Gravitational Model 2008 (EGM2008) for the *free-air gravity anomaly* signal.

Spherical harmonic expansions of functions, such as the Earth's topography (heights above and depths below the Mean Sea Level), surface temperature, atmospheric pressure, etc., are very useful since they permit the study of the spectral properties of these functions. Furthermore, the decomposition of these functions into their spectral components permits the efficient application of *convolutions* on the sphere (Driscoll and Healy 1994), and of band-pass filtering of data, with strict control on the filter's spherical harmonic-degree bandwidth. For functions representing potential fields, in addition to these spectral analysis applications, spherical harmonic analysis offers one approach for the solution of the underlying *Boundary Value Problem* that is associated with the specific *functional* of the field that has been observed. We discuss these aspects next.

Spherical Harmonic Analysis Applied to Potential Fields,

Fig. 1 Degree variances (c_n) of the free-air gravity anomaly signal (in mGal^2) implied by the EGM2008 global gravitational model, as a function of spherical harmonic degree n . The degree variances refer to the surface of a sphere of radius $a = 6,378,136.3$ m. The relationship between half-wavelength ($\lambda/2$) resolution and spherical harmonic degree is provided in the *insert*. In SI units, $1 \text{ mGal} = 10^{-5} \text{ ms}^{-2}$



Boundary Value Problems (BVP) of Potential Theory

In general, a BVP consists of a differential equation that is subject to a set of boundary conditions. In potential theory, the unknown (scalar) function to be determined is the potential, V , which satisfies Laplace's PDE, in the space outside of its generating sources. The function V should vanish at infinity as the reciprocal of the distance between the point in question and the generating source element. The boundary conditions, in order to be of practical use, should represent quantities related to V (*field functionals*) that are observable over some surface S . Two general approaches exist for the solution of a BVP:

- (a) Solution using spectral analysis, i.e., spherical harmonics
- (b) Solution using *Greens functions*, i.e., integral formulas

Traditionally, in gravimetric geodesy, the following three BVP of potential theory have been considered:

1. First BVP of potential theory, or *Dirichlet's problem* (Heiskanen and Moritz 1967, Sects. 1–16): given an arbitrary function f on a surface S , determine a function V that is harmonic either inside or outside S and which assumes on S the values of the prescribed function f . If the surface S is a sphere of radius $r = a$, the solution in terms of spherical harmonics, for the region outside S , is given by Eq. 2. The corresponding Greens function solution is given by *Poisson's* integral formula (Heiskanen and Moritz 1967, Eq. 1–89).
2. Second BVP of potential theory, or *Neumann's problem* (Heiskanen and Moritz 1967, Sects. 1–17): here, instead of

the values of the potential V itself, one is given on a surface S the values of its normal derivative $\partial V/\partial n$. The normal derivative is the derivative along the outward-directed surface normal n to S . If the surface S is a sphere of radius $r = a$, the solution in terms of spherical harmonics, for the region outside S , is given by:

$$V(r, \theta, \lambda) = -a \sum_{n=0}^{\infty} \frac{1}{n+1} \left(\frac{a}{r}\right)^{n+1} \sum_{m=-n}^n u_{nm} Y_{nm}(\theta, \lambda), \quad (16)$$

where u_{nm} is the spectrum of the $\partial V/\partial n$ boundary values, given on the sphere $r = a$, i.e.:

$$\begin{aligned} \left(\frac{\partial V}{\partial n}\right)_{r=a} &= \left(\frac{\partial V}{\partial r}\right)_{r=a} \\ &= \sum_{n=0}^{\infty} \sum_{m=-n}^n u_{nm} Y_{nm}(\theta, \lambda). \end{aligned} \quad (17)$$

The values of u_{nm} can be obtained from:

$$u_{nm} = \frac{1}{4\pi} \iint_{\sigma} \left(\frac{\partial V}{\partial r}\right)_{r=a} Y_{nm}(\theta, \lambda) d\sigma. \quad (18)$$

The corresponding Greens function solution to the second BVP is given by *Hotine's* integral formula (Hotine 1969).

3. Third BVP of potential theory (Heiskanen and Moritz 1967, Sects. 1–17): here, a linear combination of V and of its normal derivative $\partial V/\partial n$ is given on the surface S , i.e., $hV + k(\partial V/\partial n)$ is given on S . Again, if the surface S is a

sphere of radius $r = a$, the solution in terms of spherical harmonics, for the region outside S , is given by:

$$V(r, \theta, \lambda) = \sum_{n=0}^{\infty} \frac{1}{h - (k/a)(n+1)} \left(\frac{a}{r}\right)^{n+1} \sum_{m=-n}^n w_{nm} Y_{nm}(\theta, \lambda), \quad (19)$$

where w_{nm} is the spectrum of the $hV + k(\partial V/\partial n)$ boundary values, given on the sphere $r = a$, i.e.:

$$\left(hV + k \frac{\partial V}{\partial n}\right)_{r=a} = \left(hV + k \frac{\partial V}{\partial r}\right)_{r=a} = \sum_{n=0}^{\infty} \sum_{m=-n}^n w_{nm} Y_{nm}(\theta, \lambda). \quad (20)$$

The values of w_{nm} can be obtained from:

$$w_{nm} = \frac{1}{4\pi} \iint_{\sigma} \left(hV + k \frac{\partial V}{\partial r}\right)_{r=a} Y_{nm}(\theta, \lambda) d\sigma. \quad (21)$$

This particular BVP is of great importance to physical geodesy. The determination of the *geoid* given *gravity anomalies* as boundary data – called the BVP of physical geodesy – represents a specific case of this third BVP, where $h = -2/a$ and $k = -1$. With these values for h and k , if the surface S is a sphere of radius $r = a$, the solution in terms of spherical harmonics, for the region outside S , is given by:

$$V(r, \theta, \lambda) = a \sum_{n=0}^{\infty} \frac{1}{n-1} \left(\frac{a}{r}\right)^{n+1} \sum_{m=-n}^n g_{nm} Y_{nm}(\theta, \lambda), \quad (22)$$

where g_{nm} is the spectrum generated from the SHA of the gravity anomaly boundary values that are given on the sphere $r = a$. The corresponding Greens function solution to the BVP of physical geodesy is given by *Stokes'* integral formula (Heiskanen and Moritz 1967, Sects. 2–16).

Other BVP may also be formulated and solved, which correspond to other observable boundary data. For example, one may consider the *tensor* of second-order gradients of the gravitational potential (Moritz 1980, p. 4) (or some subset of the elements of this tensor) as boundary data, which give rise to BVP corresponding to *gravity gradiometry*.

Equations 9, 17, and 20 are all of the same form as Eq. 11. Therefore, provided that the boundary data are given on, or can be reduced to, the surface of a sphere of radius $r = a$, surface SHA may be used to determine the surface spherical harmonic spectrum of the field. This surface spectrum, multiplied by the appropriate radial terms that appear in Eqs. 2, 16, and 19,

provides then a solution in terms of solid spherical harmonics, to the corresponding BVP. The convergence of the series 2, 16, and 19 is guaranteed in the space outside of the sphere $r = a$.

Laplace's PDE can also be expressed in terms of the *ellipsoidal coordinates* (u, δ, λ) , where u is the semi-minor axis of the confocal ellipsoid, δ is the *reduced colatitude* (defined as 90° minus the reduced latitude), and λ is the longitude (Heiskanen and Moritz 1967, Sects. 1–19). The Cartesian coordinates (x, y, z) of a point are related to its spherical (r, θ, λ) and ellipsoidal (u, δ, λ) coordinates as follows:

$$\begin{aligned} x &= r \sin \theta \cos \lambda = (u^2 + E^2)^{1/2} \sin \delta \cos \lambda \\ y &= r \sin \theta \sin \lambda = (u^2 + E^2)^{1/2} \sin \delta \sin \lambda, \\ z &= r \cos \theta = u \cos \delta \end{aligned} \quad (23)$$

where E is the constant linear eccentricity of the family of confocal ellipsoids defining the coordinate system. Expressed in the (u, δ, λ) coordinate system, Laplace's PDE can also be solved using the method of separation of variables. The solution, for the space exterior to a reference ellipsoid with semi-minor axis $u = b$, is given by (Heiskanen and Moritz 1967, Sects. 1–20):

$$V(u, \delta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{Q_{n|m}(i(u/E))}{Q_{n|m}(i(b/E))} z_{nm} Y_{nm}(\delta, \lambda). \quad (24)$$

$Q_{n|m}(i(u/E))$ is the *Associated Legendre function* of the second kind, of degree n and order $|m|$ (Heiskanen and Moritz 1967, Sects. 1–12). i is the imaginary unit ($i = \sqrt{-1}$). On the surface of the reference ellipsoid $u = b$, Eq. 24 becomes:

$$V(b, \delta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=-n}^n z_{nm} Y_{nm}(\delta, \lambda). \quad (25)$$

Equation 25 is of the same general form as Eqs. 9 and 11, with the important distinction that δ now represents the *reduced colatitude*. The surface *ellipsoidal* harmonic spectrum z_{nm} can be determined from:

$$z_{nm} = \frac{1}{4\pi} \iint_{\sigma} V(b, \delta, \lambda) Y_{nm}(\delta, \lambda) d\sigma, \quad (26)$$

in exactly the same fashion as in the case of the surface *spherical* harmonic spectrum in 10 and 12. Note that the integration is again over the unit sphere, whose area element is now defined as $d\sigma = \sin \delta d\delta d\lambda$.

Spherical harmonics are used extensively in geodesy because of their simplicity, and because the Earth, to a first-order approximation, is a sphere. Since, to a second-order approximation, the Earth resembles more an ellipsoid of

revolution, ellipsoidal harmonics are more suitable for the solution of BVP, when the boundary data reside on or near the surface of the Earth. The transformation between ellipsoidal and spherical harmonics can be performed using the exact analytical relationships formulated by Jekeli (1988) and implemented by Gleason (1988). In this fashion, the use of the rather complicated solid ellipsoidal harmonics of Eq. 24 may be avoided altogether.

Spherical Harmonic Analysis (SHA): Numerical Techniques

Let us consider Eqs. 11 and 12. Eq. 11 is representative of Eqs. 9, 17, 20, and 25, while Eq. 12 is representative of Eqs. 10, 18, 21, and 26. Each one of Eqs. 11 and 12 suggests an alternative approach for the estimation of the coefficients f_{nm}^S , as we discuss next.

Least Squares Adjustment. The first approach uses the linear mathematical model of Eq. 11 to set up a system of *observation equations*. With f_{nm}^S representing the observable quantity and $\hat{f}_{obs}^S$ the estimates of the unknown parameters, the observation equations can be written as:

$$v(\theta, \lambda) = \sum_{n=0}^N \sum_{m=-n}^n \hat{f}_{nm}^S Y_{nm}(\theta, \lambda) - f_{obs}^S(\theta, \lambda), \quad (27)$$

where $v(\theta, \lambda)$ is the residual associated with the observation $f_{obs}^S(\theta, \lambda)$. Notice that in Eq. 27 the summation over spherical harmonic degree was truncated to a finite maximum degree N . In vector-matrix notation, Eq. 27 takes the form:

$$\mathbf{v} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{I}. \quad (28)$$

$\mathbf{v}$ is the vector representing the residuals $v(\theta, \lambda)$, $\mathbf{A}$ is the *design matrix* containing the terms that multiply the unknown parameters $\hat{f}_{nm}^S$, $\hat{\mathbf{x}}$ is the vector representing the unknown parameters $\hat{f}_{nm}^S$, and $\mathbf{I}$ is the vector containing the observations $f_{obs}^S(\theta, \lambda)$. In Eq. 28, only the matrix $\mathbf{A}$ and the vector $\mathbf{I}$ are known quantities. The linear system (28) can be solved using the method of *Least Squares Adjustment*. If Σ denotes the variance-covariance matrix of the errors associated with the observations $f_{obs}^S(\theta, \lambda)$, then, minimization of the quadratic form:

$$\phi = \mathbf{v}^T \Sigma^{-1} \mathbf{v}, \quad (29)$$

subject to the condition (28), leads to the *normal equations*:

$$\begin{aligned} \mathbf{N}\hat{\mathbf{x}} &= \mathbf{U}, \text{ where } \mathbf{N} = \mathbf{A}^T \Sigma^{-1} \mathbf{A} \text{ and} \\ \mathbf{U} &= \mathbf{A}^T \Sigma^{-1} \mathbf{I}. \end{aligned} \quad (30)$$

A scaled version of the inverse of the variance-covariance matrix is the weight matrix $\mathbf{P}$. $\mathbf{P} = \sigma_0^2 \Sigma^{-1}$ and σ_0^2 is called the

a priori variance of unit weight. σ_0^2 represents the error variance of an observation whose weight is equal to 1. The solution of the system of normal equations is:

$$\hat{\mathbf{x}} = \mathbf{N}^{-1} \mathbf{U}. \quad (31)$$

The variance-covariance matrix of the errors associated with the estimated values $\hat{\mathbf{x}}$ is given by:

$$\Sigma_{\hat{\mathbf{x}}} = \mathbf{N}^{-1} = \left(\mathbf{A}^T \Sigma^{-1} \mathbf{A} \right)^{-1}. \quad (32)$$

Carl Friedrich Gauss (1777–1855), who is credited with the development of the Least Squares Adjustment technique, used this approach in the analysis of geomagnetic data. Herein, we abbreviate the Least Squares Adjustment technique by LS.

Although in the previous discussion we presented the application of the LS method to a mathematical model that describes a *surface* spherical harmonic expansion, the LS method could be applied equally well to mathematical models describing *solid* spherical harmonic expansions, like those of Eqs. 2, 16, and 19. This means that one may use the LS method to estimate the spectra of potential fields, even when the boundary data are arbitrarily located in the three-dimensional space. In contrast, the *Numerical Quadrature* technique that we discuss next, when used for the determination of potential fields, requires that the boundary data are located over surfaces that are coordinate surfaces in the coordinate system in which Laplace's equation is formulated and solved ($r = \text{const.}$ for spherical harmonics, $u = \text{const.}$ for ellipsoidal harmonics). In theory, these coordinate surfaces must also encompass all the sources generating the field.

Numerical Quadrature. In the above discussion of the LS technique, the distribution, both in terms of location and in terms of number, of the measured values $f_{obs}^S(\theta, \lambda)$ remained unspecified. Consider now an equiangular ($\Delta\theta = \Delta\lambda$) partition of the unit sphere, along meridians ($\lambda = \text{const.}$) and parallels ($\theta = \text{const.}$). Such a partition creates a grid of $L \times 2L$ cells on the unit sphere, where $L = \pi/\Delta\theta$. Let us assume that the function $f(\theta, \lambda)$ was sampled in such a way that, for each one of the $L \times 2L$ cells, a value $f_{obs}^S(\theta, \lambda)$ exists that corresponds, for example, to the center of that cell. Such a sampling suggests a possible *discretization* of the surface integral of Eq. 12, leading to the following *Numerical Quadrature* (NQ) formula (Colombo 1981, Eq. 1.5):

$$\hat{f}_{nm}^S = \frac{1}{4\pi} \sum_{i=0}^{L-1} \sum_{j=0}^{2L-1} f_{obs}^S(\theta_i, \lambda_j) Y_{nm}(\theta_i, \lambda_j) \Delta\sigma_i, \quad (33)$$

where the area element $\Delta\sigma_i$ of the cells residing on the i th “row” is given by:

$$\Delta\sigma_i = \Delta\lambda \int_{\theta_i}^{\theta_{i+1}} \sin \theta d\theta = \Delta\lambda \cdot (\cos \theta_i - \cos \theta_{i+1}). \quad (34)$$

Equation 33 is a simple NQ formula, applicable to data $f_{obs}^S(\theta, \lambda)$ that represent point values.

In general, the estimate of the spectrum obtained either by using LS or by using NQ, will be different from the “true” spectrum f_{nm} of the function $f(\theta, \lambda)$. Furthermore, even when the two techniques are applied to the same data $f_{obs}^S(\theta, \lambda)$, their estimates will be, in general, different, i.e.:

$$\hat{f}_{nm}^S \neq \tilde{f}_{nm}^S. \quad (35)$$

It is therefore appropriate to consider the specific errors that affect $\hat{f}_{nm}^S$ and $\tilde{f}_{nm}^S$, compare the LS and NQ techniques, and determine if it is possible to design and formulate any other SHA technique(s) that would satisfy some desirable optimality criterion. This study was essentially carried out by Colombo (1981), in his treatise of the subject, entitled *Numerical Methods for Harmonic Analysis on the Sphere*. We summarize in the following Colombo’s developments and provide some examples of practical application of his formulations. Before this, however, we need to define and discuss briefly the error sources that affect any estimate of the spectrum of a function that is obtained on the basis of observed data, regardless of the technique that is used to obtain this spectrum.

- **Error Sources.** Two types of errors affect any estimate of the spectrum that is obtained from observations of a function: (a) Sampling error, and, (b) Propagated noise. The sampling error arises due to the fact that in reality, observations can only be made at a finite number of discrete points, while Eqs. 11 and 12 require the function $f(\theta, \lambda)$ to be known at every point on the surface S . Propagated noise, on the other hand, arises due to the fact that observations can never be error-free. Therefore, errors in the observations (both random and systematic) will generally propagate into the spectrum that is estimated from them.

Optimal Estimation. The particular way that sampling errors and propagated noise affect the estimated spectrum depends on the particular technique used to estimate the spectrum. Colombo (1981) designed and formulated a linear estimator of the spectrum, which is optimal in the sense that it minimizes the sum of the squares of the sampling error and the propagated noise. His *Optimal Estimation* (OE) technique is a form of Numerical Quadrature, where the *optimal* quadrature weights are defined using the formalism of *Least Squares Collocation* (LSC) (Moritz 1980). By exploiting the geometry of regular grids (i.e., grids where at least the longitude increment is constant) of data located over surfaces that

have rotational symmetry (such as a sphere or a rotational ellipsoid), and by imposing certain conditions on the *covariance functions* of the signal and the data noise, Colombo (1981) demonstrated that the variance–covariance matrix of such data sets consists of *Toeplitz-circulant* blocks. These matrix blocks can be formed and inverted very efficiently, even when the size of the data set is very large. Colombo’s OE technique was used in the gravitational field expansion developed by Hajela (1984), and in the OSU86C/D global gravitational models developed by Rapp and Cruz (1986a). Additional discussion about the estimation of gravitational spectra using LSC can be found in Tscherning (2001).

Colombo (1981, p. 76) also investigated the use of a semiempirical set of numerical quadrature weights, which depend only on the spherical harmonic degree, and can be evaluated very easily as a function of Pellinen’s smoothing factors (Colombo 1981, p. 85). Although suboptimal, these numerical quadrature weights proved to be quite adequate and were used in several gravitational modeling studies. The OSU86E/F (Rapp and Cruz 1986b) and the OSU89A/B (Rapp and Pavlis 1990) global gravitational models were developed using these semiempirical numerical quadrature weights.

Practical Aspects

Several factors complicate the application of SHA techniques when analyzing real data. These are related on the one hand to the inherent properties of the alternative SHA techniques, and on the other to the properties of the data to be analyzed. The former have to do with the computational capabilities and requirements (i.e., the flexibility and efficiency) of the different SHA techniques. The latter include issues related to the distribution and type (point or area-mean values) of the available data, the geometry of the grid to which the original data are reduced, the spectral properties of the function $f(\theta, \lambda)$ that is being approximated, and the spectral properties of the errors that may affect the observed values $f_{obs}^S(\theta, \lambda)$. We discuss briefly these factors in the following paragraphs.

- **Computational Issues.** An inspection of Equations such as (11) indicates that a spherical harmonic expansion complete to degree and order N contains $(N + 1)^2$ harmonic coefficients f_{nm}^S in total. Therefore, the normal matrix $\mathbf{N}$ of Eq. 30 has dimensions $(N + 1)^2 \times (N + 1)^2$. The structure and characteristics of the matrix $\mathbf{N}$ depends on the geometry of the available data distribution and on the characteristics of the variance–covariance matrix $\mathbf{\Sigma}$ of the data errors. If the data to be analyzed are arbitrarily located in latitude and longitude, as is the case with most raw geophysical data collected in the field (e.g., gravimetric or magnetometric data), then, apart from symmetry, the matrix $\mathbf{N}$ would be fully occupied. The same is true

(in general) if the error variance–covariance matrix Σ is arbitrary. The size of the normal matrix N is already formidable for expansions complete to degree and order 200 or so. Normal matrices for expansions to degree and order 2160, as the one corresponding to the EGM2008 model (Pavlis et al. 2008), are well beyond current computational capabilities. Nevertheless, provided that the size of the normal matrix can be handled computationally, LS adjustment and LSC are the *only* techniques that can handle arbitrarily located data, with arbitrary error variance-covariance matrices. NQ and OE require the data to be input in the form of a grid, and impose restrictions on the stochastic properties of the noise, represented within the matrix Σ . In the following, we assume that the raw data have been preprocessed in some fashion, so that the input data to the various estimators are available in gridded form.

As we mentioned previously, unlike estimators based on orthogonality relations, LS permits also the use of *solid* spherical harmonics in the formation of observation equations. Assume for the sake of this example, that a grid of potential values V is available over the physical surface of the Earth (the topography). One may then form observation equations on the basis of the mathematical model given in Eq. 2. In such a case, even if the V values are given over a regular latitude-longitude grid, these values would still be located arbitrarily in the radial direction. This would then require the formation and inversion of a fully occupied normal matrix, as it is described by Pavlis (1988), who performed such estimations to degree 36, 50, and 70, using equiangular grids of free-air gravity anomalies defined on the surface of the Earth's topography.

- *Data Organization, Types, and Distribution.* In most geodetic and geophysical SHA applications, the data are organized in geographic grids. Equiangular grids are most often used, although grids with cells of equal area are also possible, and have been used in some cases. There is therefore a need to transition from the arbitrarily located point measurements, to “artificial” measurements referring to the nodes of some specified grid. In geodetic practice, this transition is accomplished by predicting what the “artificial” grid measurements would have been, given the available arbitrarily scattered data. The prediction is done using the LSC formalism, which requires some knowledge of the signal and noise covariance functions of the data (Moritz 1980). One usually applies LSC to predict area-mean values for the grid cells, rather than point values. SHA can be applied equally well using either point or area-mean gridded data. In the case of area-mean values, the mathematical models have to be modified

appropriately, so that they accurately represent the area-mean character of the data. This is done by integrating the surface spherical harmonic functions over the specific latitudinal and longitudinal limits of each grid cell. From Eq. 3 one has:

$$\begin{aligned} & \int_{\theta_i}^{\theta_{i+1}} \int_{\lambda_j}^{\lambda_{j+1}} Y_{nm}(\theta, \lambda) \sin \theta d\theta d\lambda \\ &= \int_{\theta_i}^{\theta_{i+1}} \bar{P}_{n|m|}(\cos \theta) \sin \theta d\theta \\ & \int_{\lambda_j}^{\lambda_{j+1}} \begin{cases} \cos m\lambda & \text{if } m \geq 0 \\ \sin |m| \lambda & \text{if } m < 0 \end{cases} d\lambda \end{aligned} \quad (36)$$

Efficient recurrence relationships for evaluating the integrals of the Associated Legendre functions have been derived by Paul (1978). The numerical treatment of the evaluation of Associated Legendre functions of very high degree (e.g., $n > 1,400$ or so) requires special care since some of the values of these functions become exceedingly small, which could cause numerical problems.

Even after the gridding of geophysical data collected on or near the Earth's surface, the radial location of these data, corresponding to the Earth's topography, remains arbitrary. Solutions of BVP that are based on orthogonality relations, formulated in either spherical or ellipsoidal harmonics, require these data to be located over a sphere or an ellipsoid, respectively. This artificial (approximately radial) “movement” of the data can be accomplished using *analytical continuation*. Wang (1987, 1988) discusses some techniques that can be applied to analytically continue gravity anomaly data.

In many geophysical problems where SHA techniques are applied, the available data do not suffice to cover completely the Earth. Data gaps can produce undesirable *leakage* effects in the spectral estimates, which by definition are of global character, and require global coverage. These undesirable effects are different depending on the particular technique used to estimate the spectrum. In LS, the spectral estimator “interprets” a data gap as an undefined data value with infinite error, i.e., zero weight (see Pavlis 1988 for details). In NQ techniques, the estimator “interprets” a data gap as a data value that is equal to zero. In either case, the estimated spectrum would produce undesirable results when evaluated over the region of the data gap. It is therefore preferable to “fill in” data gaps with some reasonable values of the quantity under consideration. These “fill-in” values may be obtained on the basis of other data and/or models. For example, in

gravity modeling, topographic elevations and models of the *isostatic compensation* may be used to “fill in” areas void of actual gravity measurements (cf. Pavlis and Rapp 1990).

- *LS versus NQ techniques.* It is of interest to compare the LS and the NQ techniques from an analytical as well as a numerical perspective. Such studies have been reported by Rapp (1969, 1986), Colombo (1981), and Pavlis (1988). Sneeuw (1994) discussed the two techniques from a historical perspective. A brief review of the conclusions reached by these investigations follows. Detailed derivations and analyses supporting these conclusions can be found in the cited references. NQ is used here to identify the simple numerical quadratures formula with the semi-empirical set of suboptimal quadrature weights proposed by Colombo (1981, p. 76).

1. NQ determines each harmonic coefficient independently of all others. In contrast, LS estimates a correlated set of coefficients; thus, solutions to different maximum degrees will yield different values for the common harmonic coefficients. This will occur even if the input data are uncorrelated, and arises from the loss of orthogonality between the discrete (point or integrated) samples of the Associated Legendre functions (Pavlis 1988, Sect. 4.2.1). Unlike the Legendre functions, discrete (point, as well as integrated) samples of $\cos m\lambda$ and $\sin m\lambda$ preserve their orthogonality in the interval $[0, 2\pi)$, as long as the sampling interval $\Delta\lambda$ is constant. It is this orthogonality property of sines and cosines, along with the equatorial symmetry and the parity properties of Legendre functions, which produce the sparsity patterns of the *Block-Diagonal (BD) Least Squares* technique that we discuss next.
2. NQ cannot account for varying accuracies of the input data, while the LS estimator is capable of accounting for any (positive-definite) error variance–covariance matrix Σ associated with the input data.
3. If $L(= \pi/\Delta\theta)$ denotes the *Nyquist degree* implied by the data sampling interval, then the normal equations formed based on observation equations like (27) become singular if $N \geq L$ (Colombo 1981), where N is the maximum solved-for degree.
4. LS estimation (and LSC) can recover *exactly* a set of coefficients from “synthetic” noiseless data, provided that the data are band-limited and their frequency content does not exceed the Nyquist degree. This property offers also one verification test that any software developed to perform SHA using the LS technique should pass (Pavlis 1988, Sect. 4.2.2). The simple NQ technique is incapable of recovering the original “true” coefficients from which

the “synthetic” data were computed, as Rapp’s (1986) numerical experiments have also demonstrated.

5. Unlike the LS technique, the simple NQ does not involve any matrix formation or inversion. In this regard, the simple NQ is considerably less demanding computationally than LS. In addition, if the sampling interval $\Delta\lambda$ in the longitude direction is constant, then the NQ SHA algorithm can benefit enormously from the application of *Fast Fourier Transform* (FFT) techniques, as it was pioneered by Colombo (1981).
6. In contrast to the simple NQ, the LS technique produces not only the estimates of the coefficients, but also an estimate of their error variance–covariance matrix (see Eq. 32). This matrix can then be used to propagate the errors of the estimated coefficients onto other quantities that may be computed from them.

Block-Diagonal (BD) Least Squares. From the previous discussion, it becomes clear that a SHA technique that would combine the computational efficiency of the simple NQ, with the rigor and flexibility of LS, is highly desirable. Such a technique would attempt to retain the advantages of both techniques, while avoiding their respective limitations. The Block-Diagonal (BD) Least Squares may be viewed as such a “best of both worlds” type of technique. Colombo (1981) has shown that if:

- (a) The data reside on a surface of revolution (e.g., a rotational ellipsoid or a sphere)
- (b) The grid is complete and the longitude increment is constant
- (c) The data weights are longitude-independent
- (d) The data weights are symmetric with respect to the equator

then, the elements of the normal matrix $\mathbf{N}$, corresponding to the coefficients f_{nm}^S and f_{rq}^S , would be zero as prescribed by (see also Pavlis 1988 for details):

$$[\mathbf{N}]_{f_{nm}^S f_{rq}^S} = 0 \text{ if } m \neq q, \quad (37)$$

$$\text{or if } m = q \text{ and } n - r = 2k + 1.$$

Note that in this notation the order subscript is a signed integer, whose sign identifies the type of coefficient (positive: cosine, negative: sine). If condition (d) does not hold true, then:

$$[\mathbf{N}]_{f_{nm}^S f_{rq}^S} = 0 \text{ if } m \neq q. \quad (38)$$

The sparsity patterns implied by Eqs. 37 and 38 will be referred to as BD1 and BD2 respectively. In addition, a BD3 pattern may be considered, defined by:

$$[\mathbf{N}]_{f_{nm}^S f_{rq}^S} = 0 \text{ if } |m| \neq |q|, \quad (39)$$

which admits nonzero off-diagonal elements across coefficients of different type within a given order. It is instructive to consider the computational efficiency implied by these patterns. Table 1, which is taken from Pavlis, pp. 8–5, in Lemoine et al. (1998), provides relevant statistics for an expansion complete from degree and order 0 to degree and order 360, excluding degree $n = 1$ terms. Such an expansion involves 130,318 unknown coefficients and the upper (or lower) triangular part of the symmetric but fully occupied normal matrix $\mathbf{N}$ has 8,491,455,721 elements in total.

The enormous computational savings that can be inferred from Table 1 make the BD approximations very attractive estimation techniques. These savings, however, come at the expense of the rigor exercised in the implementation of the SHA. Any approximation of a normal matrix by a BD structure should be simple enough to permit efficient numerical implementation, and, on the same time, rigorous enough to maintain the most important characteristics of the full matrix. Therefore, one has to consider carefully whether the approximations leading to the various BD sparsity patterns can be tolerated, given the characteristics of the particular data set that is analyzed. For example, in global gravitational modeling applications, the real-world gravity anomaly data to be analyzed comply *only* with the conditions (a) and (b) above. In fact, to comply even with the (a) and (b) conditions, “filling-in” techniques and analytical continuation have to be employed, since the original data set is neither complete, nor residing on any surface of revolution, since it refers to the Earth’s topography. These aspects were carefully considered in the development of the OSU91A (Rapp et al. 1991) and the EGM96 (Lemoine et al. 1998) gravitational models. In both these models, different SHA techniques were used toward the estimation of different spectral portions of the models. This was due to the fact that different spectral portions of these models were determined on the basis of different gravimetric data sets.

In the case of *band-limited* data, one can show that the LS adjustment approach using a priori information is formally equivalent to LSC (Moritz 1980, p. 166; Colombo 1981, Sect.

2.13). In such a case, the BD approaches discussed previously are the Least Squares counterpart of Colombo’s (1981) Optimal Estimation technique.

BD techniques of varying sophistication have been used to develop GPM2 (Wenzel 1985), DGFI92A (Gruber and Bosch 1992), GFZ95A (Gruber et al. 1996), the GPM98A, B, and C models (Wenzel 1998, 1999), and the EGM2008 model (Pavlis et al. 2008).

Some of the BD structures of the normal matrix that were discussed above arise also in the analysis of data from certain configurations of *Satellite-to-Satellite Tracking* (SST) and of *Satellite Gravity Gradiometry* (SGG). This was also recognized and studied by Colombo (1984, 1989). Schuh (1996) provides a detailed study of the numerical solution strategies applicable to the analysis of such observables. SST data (in the “low-low” configuration) are currently available from the GRACE mission (GRACE 1998), and SGG data are currently being collected from the GOCE mission (ESA 1999). The availability of BD normal matrices resulting from the analysis of SST and SGG data is very important when considering the combination of these matrices with corresponding matrices derived from the analysis of global grids of gravimetric data. Such a combination of BD normal matrices permits the efficient development of global gravitational models to a very high degree (2160), as it was done in the case of EGM2008 (Pavlis et al. 2008).

Summary

Spherical Harmonic Analysis (SHA) is a critical element of the solution of Boundary Value Problems associated with potential fields. SHA is applicable not only to potential fields but also to a rather wide class of functions that are sampled over surfaces which have a one-to-one correspondence with the unit sphere. Least Squares (using either fully occupied or block-diagonal normal matrices), Numerical Quadrature, and Optimal Estimation techniques have been used to perform SHA within gravitational and magnetic field modeling efforts. The organization of the data to be analyzed in the form of geographic grids defined by meridians and parallels enables

Spherical Harmonic Analysis Applied to Potential Fields, Table 1 Statistics of normal matrices related to an expansion complete to degree $N = 360$ (excluding degree $n = 1$ coefficients) using different sparsity patterns

Statistic	Sparsity pattern		
	BD1	BD2	BD3
Total number of nonzero elements	7,905,721	15,746,100	31,362,241
Percentage of full matrix elements	0.09	0.19	0.37
Number of blocks	1440	721	361
Number of unknowns in largest block	181	360	718
Number of elements in largest block	16,471	64,980	258,121

the use of highly efficient numerical algorithms for SHA. In particular, grids with constant longitudinal spacing of data permit the application of Fast Fourier Transform techniques, thereby increasing tremendously the efficiency of the SHA algorithms.

Cross-References

- [Geoid](#)
- [Gravity Method, Satellite](#)
- [Gravity, Global Models](#)
- [Gravity, Gradiometry](#)
- [Magnetic Anomaly Map, Global](#)
- [Numerical Methods, Boundary Element](#)

Bibliography

- Blakely RJ (1995) Potential theory in gravity and magnetic applications. Cambridge University Press, Cambridge
- Colombo OL (1981) Numerical methods for harmonic analysis on the sphere. Report 310. Department of Geodetic Science and Surveying, Ohio State University, Columbus
- Colombo OL (1984) The global mapping of gravity with two satellites. Netherlands Geodetic Commission, Publications on Geodesy, New Series, vol 7, no 3. Delft
- Colombo OL (1989) High resolution analysis of satellite gradiometry. In: Rapp RH (ed) Progress in the determination of the Earth's gravity field, extended abstracts from the meeting held in Ft. Lauderdale, Florida, September 13–16, 1988. Report 397, Department of Geodetic Science and Surveying, Ohio State University, Columbus
- Driscoll JR, Healy DM Jr (1994) Computing Fourier transforms and convolutions on the 2-sphere. *Adv Appl Math* 15:202–250
- ESA SP-1233 (1) (1999) The four candidate Earth explorer core missions – gravity field and steady-state ocean circulation mission. ESA Publications Division, ESTEC, Noordwijk
- Gleason DM (1988) Comparing ellipsoidal corrections to the transformation between the geopotential's spherical and ellipsoidal spectrums. *Manuscr Geodaet* 13:114–129
- GRACE (1998) Gravity recovery and climate experiment: science and mission requirements document, revision A, JPLD-15928, NASA's Earth System Science Pathfinder Program
- Gruber T, Bosch W (1992) A new 360 gravity field model, presented at the XVII General Assembly of the European Geophysical Society, Edinburgh
- Gruber T, Anzenhofer M, Rentsch M (1996) The 1995 GFZ high resolution gravity model. In: Rapp RH, Cazenave AA, Nerem RS (eds) Global gravity field and its temporal variations. IAG symposia, vol 116. Springer, Berlin/Heidelberg
- Hajela DP (1984) Optimal estimation of high degree gravity field from a global set of $1^\circ \times 1^\circ$ anomalies to degree and order 250. Report 358. Department of Geodetic Science and Surveying, Ohio State University, Columbus
- Heiskanen WA, Moritz H (1967) Physical geodesy. W.H. Freeman, San Francisco
- Hotine M (1969) Mathematical geodesy. ESSA monograph, vol 2. U.S. Department of Commerce, Washington, DC
- Jekeli C (1988) The exact transformation between ellipsoidal and spherical harmonic expansions. *Manuscr Geodaet* 13:106–113
- Lemoine FG, Kenyon SC, Factor JK, Trimmer RG, Pavlis NK, Chinn DS, Cox CM, Klosko SM, Luthcke SB, Torrence MH, Wang YM, Williamson RG, Pavlis EC, Rapp RH, Olson TR (1998) The development of the joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) geopotential model EGM96. NASA Tech. Publ. TP-1998-206861, 575 pp
- Moritz H (1980) Advanced physical geodesy. Herbert Wichmann, Karlsruhe
- Paul MK (1978) Recurrence relations for integrals of associated Legendre functions. *Bull Géodésique* 52:177–190
- Pavlis NK (1988) Modeling and estimation of a low degree geopotential model from terrestrial gravity data. Report 386. Department of Geodetic Science and Surveying, Ohio State University, Columbus
- Pavlis NK, Rapp RH (1990) The development of an isostatic gravitational model to degree 360 and its use in global gravity modelling. *Geophys J Int* 100:369–378
- Pavlis NK, Holmes SA, Kenyon SC, Factor JK (2008) An earth gravitational model to degree 2160: EGM2008. In: Presented at the 2008 General Assembly of the European Geosciences Union, April 13–18, Vienna
- Rapp RH (1969) Analytical and numerical differences between two methods for the combination of gravimetric and satellite data. *Boll Geofis Teor Appl* XI:108–118
- Rapp RH (1986) Global geopotential solutions. In: Sünkel H (ed) Mathematical and numerical techniques in physical geodesy. Lecture notes in Earth sciences, vol 7. Springer, Berlin/Heidelberg
- Rapp RH, Cruz JY (1986a) The representation of the Earth's gravitational potential in a spherical harmonic expansion to degree 250. Report 372. Department of Geodetic Science and Surveying, Ohio State University, Columbus
- Rapp RH, Cruz JY (1986b) Spherical harmonic expansions of the Earth's gravitational potential to degree 360 using $30'$ mean anomalies. Report 376. Department of Geodetic Science and Surveying, Ohio State University, Columbus
- Rapp RH, Pavlis NK (1990) The development and analysis of geopotential coefficient models to spherical harmonic degree 360. *J Geophys Res* 95(B13):21885–21911
- Rapp RH, Wang YM, Pavlis NK (1991) The Ohio State 1991 geopotential and sea surface topography harmonic coefficient models. Report 410. Department of Geodetic Science and Surveying, Ohio State University, Columbus
- Schuh WD (1996) Tailored numerical solution strategies for the global determination of the Earth's gravity field. Technical report. Institute of Theoretical Geodesy, Technical University of Graz, Austria, Graz
- Sneeuw N (1994) Global spherical harmonic analysis by least-squares and numerical quadrature methods in historical perspective. *Geophys J Int* 118:707–716
- Tscherning CC (2001) Computation of spherical harmonic coefficients and their error estimates using least-squares collocation. *J Geod* 75:12–18
- Wang YM (1987) Numerical aspects of the solution of Molodensky's problem by analytical continuation. *Manuscr Geodaet* 12:290–295
- Wang YM (1988) Downward continuation of the free-air gravity anomalies to the ellipsoid using the gradient solution, Poisson's integral and terrain correction – numerical comparison and the computations. Report 393. Department of Geodetic Science and Surveying, Ohio State University, Columbus
- Wenzel H-G (1985) Hochauflösende Kugelfunktionsmodelle für das Gravitationspotential der Erde. *Wiss. Arb. 137. Fachrichtung Vermess. der Univ. Hannover, Hannover*
- Wenzel G (1998) Ultra high degree geopotential models GPM98A, B and C to degree 1800. <http://www.gik.uni-karlsruhe.de/~wenzel/gpm98abc/gpm98abc.htm>
- Wenzel G (1999) Schwerefeldmodellierung durch ultra-hochauflösende Kugelfunktionsmodelle. *Z Vermessungswesen* 124(5):144–154

Statistical Seismology

David A. Rhoades¹, Annemarie Christophersen¹ and Sebastian Hainzl²

¹GNS Science, Lower Hutt, New Zealand

²GFZ German Research Centre for Geosciences, Potsdam, Germany

Synonyms

Earthquake statistics; Statistical methods in seismology

Definition

Stochastic process	A process evolving in time governed partly by probability distributions.
Point process	Stochastic process of point events in time or space or both.
Renewal process	Point process in which the probability distribution of the time to the next event depends only on the time since the last event.
Stationary Poisson process	Point process in which the numbers of events in distinct time intervals of the same length are independent and have the same distribution.

Introduction

Statistical seismology is a field of research at the interface of statistical and physical modeling of earthquake occurrence. It is concerned with quantifying and understanding the distribution of earthquakes in time, magnitude, and location. It includes empirical analysis of earthquake catalogues, stochastic modeling of earthquake occurrence, estimation of the probability of earthquake occurrence and of earthquake-induced ground-shaking, and testing the forecasting power of physical and statistical models of seismicity (Vere-Jones et al. 2005). It provides input to the management of seismic hazard and risk. Although efforts have so far focused mainly on the study of earthquake catalogues, there is scope to incorporate other geophysical measurements into the analyses. A free online resource provides an overview of the field (Michael and Wiemer 2010).

Empirical Relations

One of the classical empirical relations of statistical seismology is the Gutenberg and Richter (1944) law, according to which the distribution of earthquake magnitudes M within

a large time-space volume has a probability density proportional to 10^{-bM} , where the parameter b is close to 1. Physical upper limits to earthquake size necessitate tapering of the density at high magnitudes (Kagan 1999). For small space-time volumes and in certain tectonic settings, the b -value can differ markedly from 1. Temporal variation in b -value has been proposed as an earthquake precursor. Since the radiated seismic energy E_S of an earthquake is approximately proportional to $10^{1.5M}$, the Gutenberg–Richter law implies that the distribution of radiated seismic energy has density proportional to $E_S^{-(\beta+1)}$, where $\beta = 2b/3$. Thus, the Gutenberg–Richter law can be viewed as a power law in radiated energy. The existence of such power laws is often taken as evidence of fractality and self-organized criticality in the earthquake process (Bak and Tang 1989).

Earthquakes tend to occur in clusters. A common type of cluster is the main shock – aftershock sequence. The Omori–Utsu law, first noted by Fusakichi Omori in 1894, describes how an aftershock sequence decays over time (Utsu et al. 1995). According to this relation, which is also a power law, the rate of aftershock occurrence is proportional to $(t + c)^{-p}$, where t is the elapsed time following the occurrence of the main shock, and c and p are adjustable parameters, typically with $c \ll 1$ day, and $p \approx 1$. In some large aftershock sequences, the relation can be seen to hold for decades after the occurrence of the main shock.

The area occupied by an aftershock sequence is approximately 10^{M-4} km², where M is the main shock magnitude (Utsu 1961). This is consistent with scaling relations derived for earthquake source dimensions – fault length, width, and displacement – all of which are approximately proportional to $10^{0.5M}$ (Wells and Coppersmith 1994). However, some major earthquakes appear to trigger aftershocks at long distances from the earthquake source. These can be accommodated by a power law in which the aftershock density is proportional to $(d^2 + r^2)^{-q}$, where r is the distance from the source, q is an adjustable parameter, typically about 1.5, and d is a function proportional to $10^{0.5M}$.

According to Båth's law (Båth 1965), the largest aftershock is typically about 1.2 magnitude units smaller than the main shock. This indicates that the main shock and its aftershocks do not by themselves conform to the Gutenberg–Richter relation, the main shock being too large relative to the other earthquakes in the set. The existence of other types of earthquake clusters than main shock – aftershock sequences, such as swarms (which have no main shock) and multiplets (which may have several main shocks) is a confounding element in attempts to systematically examine Båth's relation.

Rules for defining and classifying earthquake clusters are necessarily somewhat arbitrary. They include simple window methods, linking algorithms, and stochastic declustering (Michael and Wiemer 2010).

Data quality is an ever-present issue affecting statistical analysis of earthquake catalogues. It is necessary to establish the quality of a catalogue at the outset, because an analysis can be seriously compromised by changes in the quality of the catalogue over time, such as temporal and spatial variation of the magnitude threshold of completeness. On the positive side, catalogue data are generally improving in both quality and quantity as better seismograph networks and earthquake location techniques are employed, and this is creating opportunities for more detailed analyses.

Precursors

Studies aimed at identifying precursors of large earthquake are faced with a dual challenge: Whereas large earthquakes are relatively rare, the number of degrees of freedom available when attempting to identify precursors is large. Therefore, a careful analysis is required before the conclusion is drawn that a proposed precursory phenomenon is real. A retrospective analysis can be used to define a possible precursory phenomenon, but prospective testing is necessary to confirm it.

Many major earthquakes are preceded by an increase in the rate of occurrence and magnitude of minor earthquakes, in an area not much larger than the major earthquake source and over a time period ranging from months to decades, depending on magnitude. This is called the increase precursory scale increase (Evison and Rhoades 2004). Sometimes, it consists of a sequence of swarms. The magnitude of the largest precursor is typically about one unit smaller than that of the main shock, and the precursor time T_p (the time from the onset of the increase to the main shock) and the area in which the precursors, major earthquake, and aftershocks all take place are both approximately proportional to $10^{0.5M}$. Proposed earthquake precursors for which T_p is proportional to $10^{0.5M}$ are known as precursors of the first kind (Rikitake 1982) and include anomalies in tilt, strain and land deformation, and changes in the relative velocities of seismic P and S waves.

Models of earthquakes as a critical phenomenon have been invoked to suggest that an accelerating moment release (AMR) should occur in an area much larger than the earthquake source in the approach of criticality, i.e., an accelerating occurrence of minor earthquakes leading up to a major earthquake. Another much-studied phenomenon is precursory seismic quiescence (PSQ) – a proposed reduction in the rate of occurrence of minor earthquakes shortly before a major event. There is apparent empirical support for both AMR and PSQ, but the statistical significance has been questioned. A problem in resolving such matters is that these phenomena, like many other proposed precursors, have only been defined anecdotally and never identified with an explicit stochastic model of earthquake occurrence.

Stochastic Models of Earthquake Occurrence

Early examples of stochastic modeling of earthquake sequences were given by Vere-Jones (1970). Modeling of earthquake occurrence has benefited from the development of the theory of stochastic point processes (Daley and Vere-Jones 2003).

The idea of successive episodes of stress accumulation and release, as suggested by the elastic rebound theory of Henry Fielding Reid in 1910, is the basis for most modeling of recurrence of major earthquakes on a fault or fault segment. The time sequence of earthquakes is often modeled as a renewal process, assuming a characteristic magnitude for events on the fault. The recurrence-time distribution is variously taken as exponential (for constant hazard), Weibull, lognormal, or Brownian passage-time. The latter three distributions all imply some degree of regularity in the time intervals between earthquakes, and physical arguments have been advanced in support of each. The limited data so far available on earthquake recurrence on fault segments do not support a clear preference for a particular distribution or afford a real opportunity to test how informative renewal models are in this context. It is important to account for data and parameter uncertainties when applying such models to the estimation of earthquake hazard.

The renewal process model does not allow for interactions between neighboring faults or for variation in the size of successive events. Models involving time and magnitude include the time-predictable and slip-predictable models, in which either the time to the next earthquake depends on the magnitude of the last earthquake or the magnitude of the next earthquake depends on the elapsed time. These models are sometimes applied to the earthquakes in a region. In a stress-release model of Zheng and Vere-Jones (1994), also applicable to the earthquakes in a region, the hazard rate steadily increases over time, but drops upon the occurrence of an earthquake by an amount that depends on its magnitude.

The epidemic-type aftershock (ETAS) model of Ogata (1989) is a nonstationary Poisson process model developed originally as a model of temporal clustering of earthquakes in a discrete region. In this model, each earthquake has its own aftershock sequence, which decays over time according to the Omori-Utsu law. The number of aftershocks is governed by a magnitude-dependent productivity function. Some earthquakes occur independently according to a stationary Poisson process. The ETAS model fits earthquake data much better than a stationary Poisson process and better than a model in which only the larger events have aftershocks.

Stochastic models involving location, as well as time and magnitude, allow for a more realistic representation of the earthquake process, including clustering and long-range interactions between earthquakes.

The ETAS model was extended to include the spatial variable (Ogata 1998) by introducing a distribution for after-shock location, usually a power law of distance similar to that described above. In this form, it can be used to describe the space-time clustering of earthquakes and as a diagnostic tool for physical changes affecting the earthquake process. As a short-term forecasting model, this model performs far better than smoothed seismicity models, which capture the space and magnitude distributions of past earthquakes but are time-invariant. It is successful at forecasting aftershocks, as well as the minor proportion of major earthquakes which are preceded by a foreshock sequence. For a given location, the rate density of earthquake occurrence in this model can fluctuate abruptly over several orders of magnitude. Changes in the parameters of the spatial ETAS model have been used as indirect indicators of stress-changes in regions of the crust.

The Every Earthquake a Precursor According to Scale (EEPAS) model (Rhoades and Evison 2004) is a space-time-magnitude model designed for medium-term forecasting of the major earthquakes. In this model, every earthquake contributes to the future distribution of earthquake occurrence in time, magnitude, and location on a scale determined by its magnitude, through predictive scaling relations associated with the precursory scale increase phenomenon. The EEPAS model rate density can fluctuate slowly in time over about 1.5 orders of magnitude. The EEPAS model outperforms smoothed seismicity models in well-catalogued inter-plate regions such as California, Japan, and New Zealand.

Physics-Based Models

In contrast to models based only on earthquake catalogues, physics-based models of earthquake occurrence also take stress evolution into account. Stress changes can be estimated from earthquake slip models or aseismic sources such as tectonic and reservoir loading and fluid injections. The most prominent model of this type is the Coulomb Rate-State model which accounts for the earthquake nucleation rate in fault populations where each fault obeys the constitutive rate- and state-dependent friction law derived from laboratory experiments (Dieterich 1994; Heimisson and Segall 2018). Applications show the model performs well when the large uncertainties of the involved stress calculations are taken into account (Cattania et al. 2014).

Numerical earthquake simulators generate synthetic earthquake catalogues based on known physical elements, such as stress transfer between faults, friction laws, stress accumulation, and the detailed geometry of fault surfaces (Tullis 2012). Statistical analysis of such catalogues can reveal under what

physical conditions a given statistical model is likely to be appropriate and contribute to physical understanding of empirical relations derived from real earthquake catalogues (Christophersen et al. 2017).

Testing of Forecasting Methods

The stationary Poisson model with characteristic earthquake magnitudes on faults and spatially distributed background seismicity conforming to the Gutenberg–Richter magnitude relation are the basis for most seismic hazard models in practical use. However, an international effort by the Collaboratory for the Study of Earthquake Predictability (CSEP) is providing transparent, prospective testing of time-varying earthquake occurrence models within certain gridded testing regions in California, New Zealand, Europe, Japan, China, the northeast Pacific, southwest Pacific and, at high magnitudes, the entire globe. To be testable, a model must provide estimates of the expected number of earthquakes for future time windows within grid cells delimited by location and magnitude.

The performance of a probabilistic model of earthquake occurrence on a target set of earthquakes in a catalogue independent of the one to which it was fitted is conveniently measured by the information gain per earthquake, defined as the increase in the log likelihood of the earthquake catalogue under the model compared with that under a reference model, divided by the number of targeted earthquakes.

CSEP uses likelihood-based consistency tests, which can identify significant differences between the expected and observed number of targeted earthquakes, and the distributions in time, magnitude, and location of cell expectations and those of the targeted earthquakes. It is also developing consistency tests for models represented only by simulated catalogues of future events (Michael and Werner 2018).

Some earthquake forecasting methods, such as the M8 algorithm, are alarm-based rather than probabilistic. For alarm-based methods, the error diagram (Molchan 1990), in which the proportion of time or space-time for which a certain level of alarm is exceeded is plotted against the proportion of unpredicted earthquakes, is a useful assessment tool.

Operational Earthquake Forecasting

Some forecasting models are sufficiently well tested to be used for official earthquake warnings. There is often a demand for authoritative forecasts of triggered events in the aftermath of a major damaging earthquake (Jordan et al. 2011). Aftershock models incorporating the Omori-Utsu

law, such as ETAS or simpler alternatives, are generally used to provide such short-term forecasts, with a time horizon of weeks, months, or perhaps a year. Hybrid forecasting models, based on expert judgement informed by statistical testing, and incorporating short-, medium- and long-term components, have been used to extend the time horizon out to several decades or longer (Gerstenberger et al. 2016), as is required for the planning of recovery operations following a major earthquake. An ideal operational earthquake forecasting system would be integrated with seismic hazard and risk tools and would be capable of automatically providing forecasts of earthquake occurrence within any desired region for any specified future time-period based on the best-performing models available.

Conclusion

Statistical seismology is concerned with modeling the empirical observations of earthquake occurrence and testing models against observations while trying to throw light on the physics of earthquake generation. A rapid improvement in seismological databases has been a stimulus to recent developments in the field, and this state of affairs is likely to continue. Improved earthquake catalogues and other systematically collected data, such as earth deformation data derived from global navigation satellite systems, offer opportunities for enhanced physical understanding of the earthquake process and more informative stochastic models. At the same time, increased computing power allows for increased detail and complexity in physics-based modeling. The major challenge for statistical seismology is to use the new data, together with physical insights from empirical studies and detailed physical modeling, to develop ever more informative stochastic models of earthquake occurrence that can be applied to mitigating earthquake hazard.

Cross-References

- Artificial Water Reservoir-Triggered Earthquakes, with Special Emphasis on Koyna, India
- Characteristic Earthquakes and Seismic Gaps
- Earthquake Precursors and Prediction
- Earthquake Prediction, M8 Algorithm
- Earthquake, Magnitude
- Earthquakes and Crustal Deformation
- Earthquakes, Early and Strong Motion Warning
- Earthquakes, Energy
- Seismic Data Acquisition and Processing
- Seismic Quiescence and Activation

Bibliography

- Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. *J Geophys Res* 94:15635–15637
- Båth M (1965) Lateral inhomogeneities of the upper mantle. *Tectonophysics* 2:483
- Cattania C, Hainzl S, Wang L, Roth F, Enescu B (2014) Propagation of Coulomb stress uncertainties in physics-based aftershock models. *J Geophys Res* 119:7846–7864
- Christophersen A, Rhoades DA, Colella HV (2017) Precursory seismicity in regions of low strain rate: insights from a physics-based earthquake simulator. *Geophys J Int* 209:1513–1525
- Daley DJ, Vere-Jones D (2003) An introduction to the theory of point processes, vol 1, 2nd edn. Springer, New York
- Dieterich JH (1994) A constitutive law for rate of earthquake production and its application to earthquake clustering. *J Geophys Res* 99:2601–2618
- Evison FF, Rhoades DA (2004) Demarcation and scaling of long-term seismogenesis. *Pure Appl Geophys* 161:21–45
- Gerstenberger MC, Rhoades DA, McVerry GH (2016) A hybrid time-dependent probabilistic seismic-hazard model for Canterbury, New Zealand. *Seismol Res Lett* 87(6):1311–1318
- Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. *Bull Seismol Soc Am* 34:185–188
- Heimisson ER, Segall P (2018) Constitutive law for earthquake production based on rate-and-state friction: Dieterich 1994 revisited. *J Geophys Res* 123:4141–4156
- Jordan TH, Chen Y-T, Gasparini P, Madariaga R, Main I, Marzocchi W, Papadopoulos G, Sobolev G, Yamaoka K, Zschau Y (2011) Operational earthquake forecasting: state of knowledge and guidelines for utilization. *Ann Geophys* 54(4):315–391
- Kagan YY (1999) Universality of the seismic moment-frequency relation. *Pure Appl Geophys* 155:537–573
- Michael AJ, Werner MJ (eds) (2018) Focus section on the Collaboratory for the Study of Earthquake Predictability (CSEP). *Seismol Res Lett* 89(4):1226–1304
- Michael AJ, Wiemer S (eds) (2010) Community online resource for statistical seismology analysis. <http://www.corssa.org/en/home/>
- Molchan GM (1990) Strategies in strong earthquake prediction. *Phys Earth Planet Inter* 61:84–98
- Ogata Y (1989) Statistical models for standard seismicity and detection of anomalies by residual analysis. *Tectonophysics* 169:159–174
- Ogata Y (1998) Space-time point-process models for earthquake occurrences. *Ann Inst Stat Math* 50(2):379–402
- Rhoades DA, Evison FF (2004) Long-range earthquake forecasting with every earthquake a precursor according to scale. *Pure Appl Geophys* 161:47–71
- Rikitake T (1982) Earthquake forecasting and warning. Center for Academic Publications, Tokyo/Reidel/Dordrecht, 402 p
- Tullis TE (ed) (2012) Focus section on earthquake simulators. *Seismol Res Lett* 83(6):957–1006
- Utsu T (1961) A statistical study on the occurrence of aftershocks. *Geophys Mag* 30:521–605
- Utsu T, Ogata Y, Matsu'ura RS (1995) The centenary of the Omori formula for a decay law of aftershock activity. *J Phys Earth* 43:1–33
- Vere-Jones D (1970) Stochastic models for earthquake occurrence. *J R Stat Soc Ser B* 32:1–62
- Vere-Jones D, Ben-Zion Y, Zuniga R (eds) (2005) Statistical seismology, pageoph topical volumes. *Pure Appl Geophys* 162(6/7):1023–1396
- Wells DL, Coppersmith KJ (1994) New empirical relations among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bull Seismol Soc Am* 84:974–1002
- Zheng X, Vere-Jones D (1994) Further applications of stress release models to historical earthquake data. *Tectonophysics* 229:101–121

Subduction Zones

Geoffrey A. Abers

Lamont-Doherty Earth Observatory, Columbia University,
Palisades, NY, USA

Definition

Subduction zones are convergent plate boundaries involving at least one oceanic plate. The oceanic plate descends beneath the other into the deep mantle, forming a deep-sea trench, often a volcanic arc subparallel to it, an inclined Wadati–Benioff seismic zone, and other subparallel deforming belts.

Introduction

Subduction zones are a fundamental manifestation of planetary convection at the Earth's surface, and can be viewed as regulating two critical systems:

1. They control the primary material flux from the surface to the Earth's interior. Material at the seafloor subducts to depth, some of which emerges in volcanic arcs. These processes control the long-term budgets of volatiles (H_2O , CO_2 , etc.), and the major accretion of material to continental crust.
2. Forces at subduction zones result in much of the planet's deformation, including the generation of the largest earthquakes and explosive volcanic eruptions over time scales of seconds, to the formation of the deepest seafloor and the growth of continents over millions of years. The negative buoyancy of subducting lithosphere probably represents the largest driver of global plate motions.

All of these processes ultimately result from the Earth releasing heat, and the deformation and thermal structure that results. At the largest scale, subduction regulates the mixing of material throughout the mantle of the lifetime of the planet, and the onset and termination of subduction represent profound geologic events. This article follows subduction zones from trench to depth, emphasizing the two themes of material flux and of large-scale deformation. It focuses on active systems from outer rise to sub-arc depths.

Morphology and Nomenclature

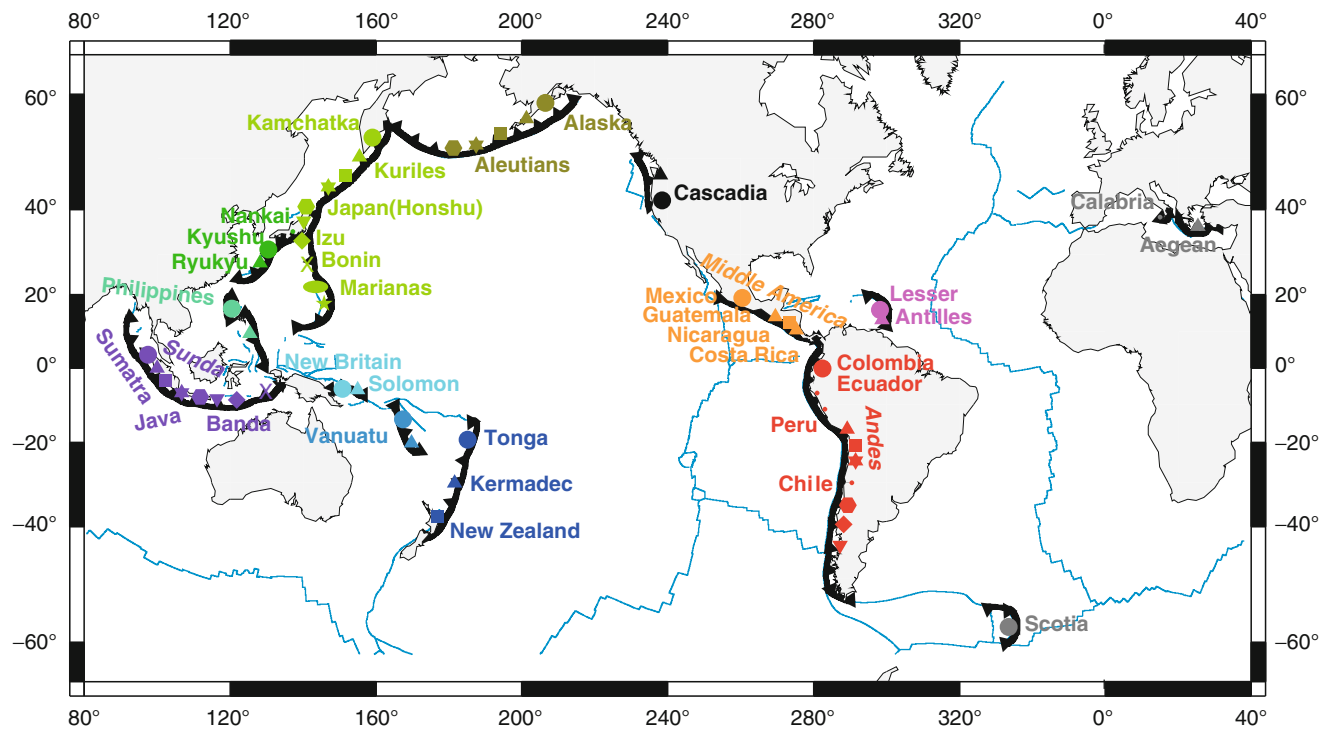
The global subduction oceanic system is over 40,000 km long (Stern 2002), largely but not entirely subducting oceanic plates of the Pacific basin (Fig. 1). For the most part, any

part of this subduction zone exhibits a common morphology (Fig. 2): a downgoing, oceanic plate flexes into a deep-sea trench, beneath a submarine forearc that sometimes includes an accretionary prism. As the plate interface reaches 10–50 km depth, plate motion is accommodated on a shallow-dipping thrust zone, which in many places can generate great earthquakes. Landward of the thrust zone lies a volcanic arc, typically with a sharp front that overlies the region where the Wadati–Benioff zone seismicity reaches 75–180 km depth, with occasional volcanism further in the back-arc. The upper plate beneath and behind the arc varies between those that generate oceanic back-arc basins, exemplified by the Mariana–Izu–Bonin and Tonga systems, and those that undergo long-term convergence and compression, exemplified by the Andes, with a spectrum of behaviors in between (e.g., Lallemand et al. 2005). This characteristic morphology is a consequence of a regular, repeatable balance between forces that drive plate motions, buoyancy of oceanic lithosphere, metamorphic devolatilization, and its rheological/magmatic consequences.

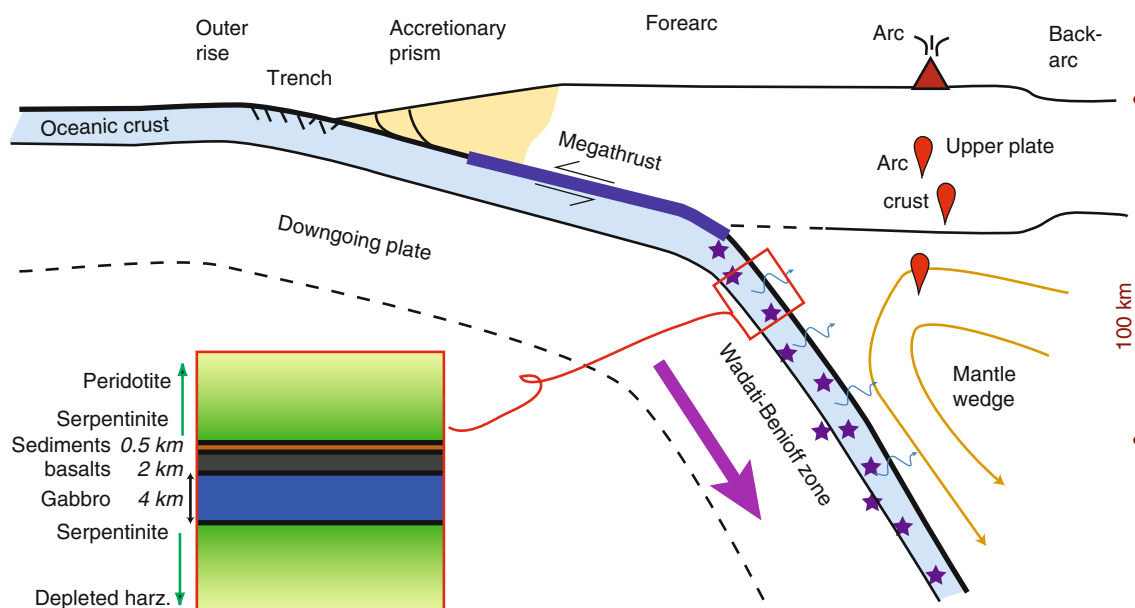
Outer Rise, Trench, and Forearc

The shallow part of subduction systems, where temperatures are relatively cold, are controlled by the mechanics of elastic and brittle deformation. The trench occurs because the downgoing plate is flexed downward, and the shape of the plate seaward of the trench resembles closely that expected for flexure of a thin elastic surface (e.g., Watts and Talwani 1974). The physics of elastic flexure also explains the outer rise – a region of slightly (100–500 m) elevated seafloor that lies seaward of the trench, typically 50–300 km distant (Parsons and Molnar 1976). Strong bending leads to faulting in the outer rise, typically manifest as a series of normal faults with <1,000 m offset, seen on the seafloor and generating earthquakes (Chapple and Forsyth 1979). These normal faults are found seaward of many if not most trenches, and are the most seaward indication of deformation within the subduction zone. They may play an important role in the material cycle as well, since the basins formed by them serve as traps for sediment that can be subducted to great depths (Hilde 1983), and the faults themselves may act as conduits for water to reach into the mantle of the subducting plate, altering the mineralogy (e.g., Ranero et al. 2003). Such alteration of seafloor in the open ocean plays an important role in preconditioning the downgoing plate prior to subduction: the later metamorphic devolatilization of hydrous sediments and altered oceanic lithosphere releases fluids that lubricate the plate boundary, weaken the overlying material, and ultimately drive arc volcanism.

The sediment supply to the downgoing plate varies enormously between trenches near major continental river



Subduction Zones, Fig. 1 The global distribution of subduction zones. (Modified from Syracuse et al. 2010)



Subduction Zones, Fig. 2 Common features of subduction zones

systems and those remote from such sources, leading to several variations in subduction zone behavior and characteristics. In the most extreme cases sediment supply overwhelms the rate at which material is subducted and accreted, so the bathymetric trench becomes a sediment-filled basin. Examples include the Lesser Antilles, Makran, and Cascadia

subduction zones. A first-order distinction exists between subduction zones considered “accretionary,” where a substantial amount of sediment has been offscraped to form an active accretionary prism, and “non-accretionary” subduction zones where all sediment is being subducted bypassing the forearc, and generally sediment supply is low (von Huene and Scholl

1991). The accretionary prisms typically deform internally to form a critical-taper wedge, a geometry in which gravitational stresses created by the bathymetric slope are balanced by internal friction, similar to wedges in front of bulldozers or snowplows (Davis et al. 1983). Accreted material includes deep-sea sediment, but also fragments of bathymetric features such as seamounts on the incoming plate; subduction of relatively rigid seamounts can have a profound effect on the local morphology of the forearc, and may contribute to heterogeneity along the thrust zone at greater depth.

Thrust Zone

As of mid-2010, the nine largest recorded earthquakes, and 22 of the largest 25, occurred on subduction thrust zones or “megathrusts,” including all earthquakes with magnitudes of 9.0 and larger (U.S. Geol. Surv. Web site, <http://earthquake.usgs.gov/>). These great earthquakes also generate most of the largest tsunamis, because they can displace large portions of the seafloor (see entry on ► “Tsunami”). Great earthquakes occur here because the plate-boundary fault system can sustain large rupture areas, and earthquake size scales with rupture area. There are two reasons for the large rupture areas. First, at subduction zones cold material advects downward quickly compared with rates of heat conduction, as discussed below, so temperatures along the fault zone can remain below that of the brittle-ductile transition (or downdip limit of unstable sliding; Scholz 1998) to greater depths than other environments, typically 30–50 km (Tichelaar and Ruff 1993). Second, subduction zone thrusts dip at 5–25°, compared with steeper dips for faults in many other tectonic settings, so a greater fault area exists above the brittle-ductile transition. Both of these reasons allow for faults that can exceed 100 km in the downdip direction (e.g., for the Gulf of Alaska megathrust). Along strike, individual ruptures can extend for more than 1,000 km in some cases, most notably Chile (1960; $M_w = 9.5$) and Sumatra (2004; $M_w = 9.2$). Magnitudes here (M_w) are moment magnitudes (see entry on ► “Earthquake, Magnitude”).

The fault area, and hence the ultimate size of earthquakes, is controlled by updip and downdip changes in material properties. At its updip end, the thrust zone undergoes a transition from aseismic slip on a decollement (subhorizontal plate-boundary fault) to stick-slip behavior or earthquake rupture, at depths of a few kilometers. At these depths the plate interface likely lies at temperatures of 75–150 °C. The causes of this transition likely have to do with compaction, lithification, and low-temperature mineral phase transitions that change the frictional properties of this fault surface (Saffer and Marone 2003). Similarly, the downdip limit can be in many places approximated by a threshold temperature, in this case near 350–500 °C, with notable

exceptions (e.g., Hyndman et al. 1997). Both transitions must be a consequence of change in frictional properties at this depth, with aseismic creep taking place at greater depths (e.g., Scholz 1998). In between, the megathrust fault zone shows a variety of slip behaviors from purely creeping or “decoupled” to fully seismic or “coupled.” In this usage, “coupling” refers to the ratio of fault slip inferred from summing the seismic moment of earthquakes over some time interval, to the total plate motion expected across that boundary over the same time interval. Coupled or “locked” thrust zones can be identified with geodesy, which shows the elastic strain accumulation building up between large earthquakes.

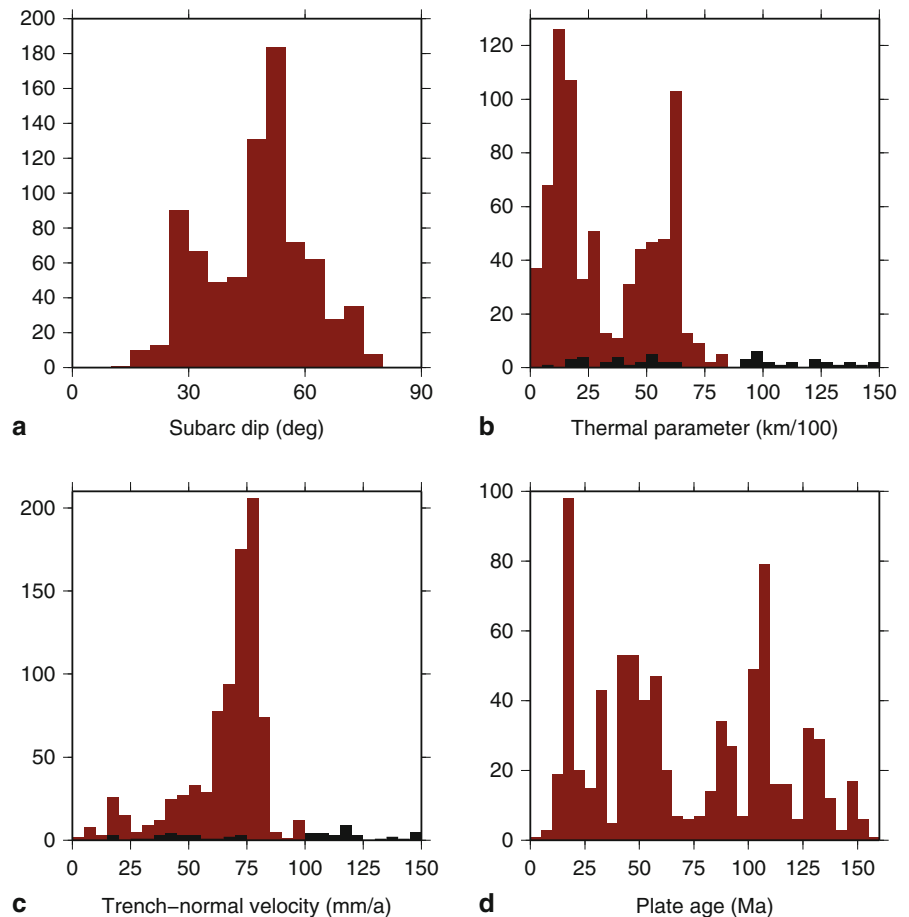
Recent discoveries of slow, silent creep events indicate that the downdip transition (and perhaps the updip one) may be much more complicated (e.g., Rogers and Dragert 2003; Ito et al. 2007; Obara 2002). Continuous geodetic observations have shown that patches of the thrust zone can slip over periods of days to months, in some cases accounting for a large fraction of the plate motion within the patch that slips (usually, adjacent to the locked zone). In some cases these events repeat quasi-periodically at time scales of months to years, although periodicity is not yet established for many regions of slow slip (Schwartz and Rokosky 2007). Often, the slow slip is accompanied by a variety of exotic seismic phenomena, including nonvolcanic tremor, low-frequency earthquakes, very-low-frequency earthquakes, and creep (Ide et al. 2007). In most cases, the tremor seems to come from the plate boundary or just above it and shows motion consistent with thrusting on the plate boundary; these observations are subject of much current research (Gomberg 2010). The physics of these quasi-periodic slip transients remain poorly understood; excess fluid pressure is suspected to play a role.

Kinematics

Subduction zones can be divided between those with little upper-plate deformation, with upper-plate compression, and exhibiting back-arc opening. The highest subduction rates, and indeed the highest relative plate motion rates anywhere, reach 240 mm/year at the northern end of the Tonga subduction system (Bevis et al. 1995). Major-plate motion between Australia and the Pacific Plate is only ~80 mm/year, the remainder being accommodated by spreading in the Lau back-arc system. Subduction can in principle be arbitrarily slow, although the lowest rates associated with an active arc probably occur in the Lesser Antilles (20–25 mm/year convergence; DeMets 2001). A wide range of convergence rates lie in between these end-members with a sharp global peak near 70 mm/year such that half of all subduction zones have arc-normal convergence rates of 53–76 mm/year (Fig. 3).

Subduction Zones,

Fig. 3 Histograms showing common parameters at subduction zones. Each is calculated at arc volcanoes or the nearest point on the plate boundary to them. (Based on data from Syracuse and Abers 2006)



While spreading at mid-ocean ridges is orthogonal to their axes, at subduction zones convergence commonly trends obliquely to trench or arc (e.g., McCaffrey 1992; Fitch 1972). The obliquity, ϕ , is the angle between the normal to trench axis and the plate convergence direction. At low obliquities, convergence occurs in the direction of plate motion but commonly at higher obliquities strain is partitioned between a trench-normal component, represented by thrust faulting on the megathrust, and an along-strike component taken up by internal deformation. The along-strike component may be taken up within the forearc, as arc-parallel strike-slip systems, as sequences of strike-slip faults oblique to the trench axis, or by other means. At many subduction zones the obliquity changes along strike as both plate orientation and location relative to Euler pole change; for example, convergence is trench-normal ($\phi \sim 0$) in the Alaska Peninsula segment of the Aleutian subduction zone, $\phi \sim 45^\circ$ from the trench in the central Aleutians, and plate motion is essentially parallel to the trench ($\phi \sim 90^\circ$) in the far western Aleutians (DeMets et al. 1990). The latter case implies that material no longer descends into the mantle in the far western Aleutians, but translates along strike. A primary consequence of obliquity is

that the trench-normal component of convergence can be significantly less than plate motion rates.

The kinematics of plates at asthenospheric depths may be more complex. While it is convenient to draw flow lines of subducting plates parallel to their dips, assuming steady-state slab geometry, it is not obvious that is the case. In many geodynamic models (e.g., Billen 2008) slabs sink vertically faster than they would if they followed their trajectories, in other words, slabs roll back. Such rollback is probably required to accommodate back-arc opening in many places, although few if any constraints exist on these kinematics. Flow kinematics are complicated by inferences of large along-strike flow inferred from seismic anisotropy (e.g., Russo and Silver 1994; Hoernle et al. 2008; see entry on [“Magnetic Anisotropy”](#)), not easily explained by simple models of wedge flow.

Thermal Structure, Plate Buoyancy, Mantle Flow

At subduction zones, cold oceanic lithosphere advects downward relatively quickly compared with the rate of heat

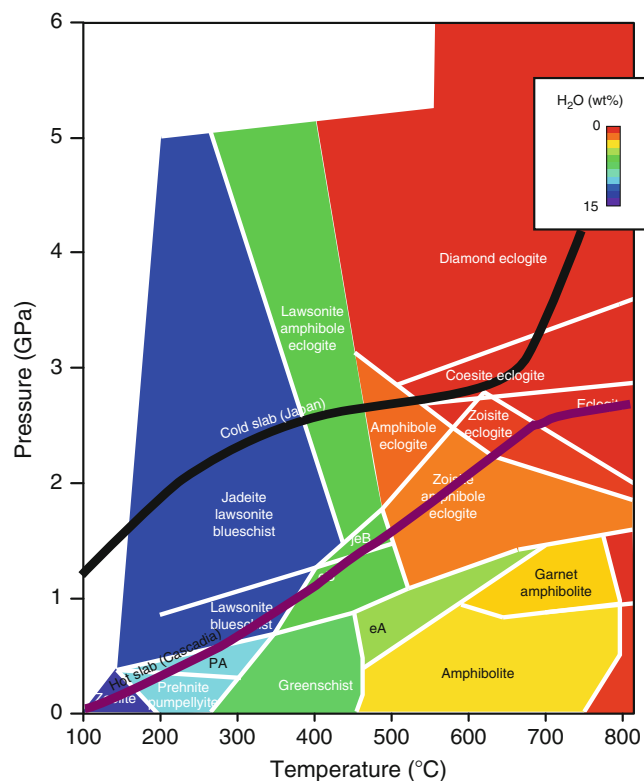
conduction, leading to some of the largest lateral temperature gradients in the planet. Old, fast-subducting slabs can retain temperatures $<600^{\circ}\text{C}$ in their cores to depths exceeding 100 km, where ambient mantle temperatures are otherwise $\sim 1,400^{\circ}\text{C}$ (e.g., Syracuse et al. 2010; Wada and Wang 2009). The low temperatures lead to relatively large density differences, giving oceanic lithosphere significant negative buoyancy and probably supplying one of the major forces that drive plate motions (e.g., Forsyth and Uyeda 1975). Because temperatures are low, mineral phases that hold significant H_2O are stable to much greater depths than elsewhere, making subduction zones major conveyors of volatiles to the deep earth. Still, volcanism occurs in nearly all oceanic subduction zones, indicating that parts of the system remain warm enough to produce melting. As a slab descends, it induces corner flow in the overlying mantle wedge, which draws warm material from the distal back-arc toward the slab corner (Fig. 4). This advection of warm material ultimately heats the top of the slab, leading to very strong temperature gradients near the slab surface, and driving critical material-transport processes there.

Where the slab surface lies at depths less than 50 km, plate-boundary motion is mostly localized along the megathrust

that defines the plate boundary, so the upper-plate deformation accommodates little of the plate motion (at least the downdip component; oblique slip partitions strike-parallel motion into the upper plate). As a consequence, heat transport in this region can be approximated as the balance between conduction through an overlying, stationary upper plate, and downward advection of cold material along the thrust zone (Molnar and England 1990). Shear heating along the thrust zone may elevate temperatures but its effect seems to be minor in the few places where abundant heat-flow measurements have been made (Cascadia, Japan, Tonga), equivalent to that generated by a fault with shear stress of a few tens of MPa at most (Hyndman and Wang 1995; Von Herzen et al. 2001). Thus, interplate thrust faults are relatively weak, perhaps due to high fluid pressures. One consequence is that metamorphism of subducted material follows a low-temperature high-pressure trajectory through this region, to blueschist facies (Fig. 4).

At greater depth, the overlying material flows and transports heat from the back-arc mantle into the wedge. As this hot material is entrained downward by viscous drag, it rapidly heats the top of the downgoing plate, resulting in dehydration of the top of the downgoing plate. Such heating must happen under or trenchward of the volcanic arc, since many arcs erupt basaltic lavas with chemistry that requires temperatures in excess of $1,200^{\circ}\text{C}$ beneath the arc (Kelemen et al. 2003a). The temperatures predicted for the top of the slab depend upon the assumed rheology of the mantle wedge, but for power-law flows typical of olivine dislocation creep, a narrow ($<25\text{--}50$ km wide) thermal boundary layer should rapidly develop at the top of the slab, leading to slab-surface temperatures of $700\text{--}1,000^{\circ}\text{C}$ directly beneath the volcanic front (Kelemen et al. 2003a; van Keken et al. 2002; Syracuse et al. 2010). This heating appears to be sufficient to drive off most mineral-bound volatiles in the altered oceanic crust and sediment. More complicated modeling that allows secondary flow of buoyant slab-surface material into the wedge (e.g., Gerya and Yuen 2003) might lead to lower temperatures, if such flows are sustainable.

Heat supplied by the mantle wedge continues to conduct into the slab, gradually warming its interior. Simple thermal models (e.g., McKenzie 1969; Molnar et al. 1979) predict that the maximum depth of isotherms should scale with $\Phi = AV\sin\delta$, where A is the age of the incoming lithosphere, V is its downdip velocity, and δ is the slab dip. The quantity Φ , sometimes called the “thermal parameter” (e.g., Kirby et al. 1996), describes in a simple way the overall thermal structure of slabs at great depth, and correlates well with the maximum depth of earthquakes over some depth ranges (Gorbatov and Kostoglodov 1997), indicating a thermal control for the processes that cause deep earthquakes. Ultimately, the negative buoyancy of cold slabs and their ability to descend into the lower mantle must depend upon Φ as well. The negative



Subduction Zones, Fig. 4 Pressure-temperature trajectories for the top of slabs in hot (Cascadia) and cold (Japan) subduction zones, and metamorphic facies predicted for hydrated metabasalts, color coded by H_2O content. (Facies from Hacker et al. 2003; trajectories from Syracuse et al. 2010)

thermal buoyancy of the slab counterbalances the positive buoyancy of subducting crust, since gabbro is much less dense than peridotite, such that the net buoyancy is a function of plate age and thickness of subducting crust (Cloos 1993). For normal (5–8 km) thickness of oceanic crust, lithosphere attains negative buoyancy if more than 10–30 Ma old, while crust thicker than 15–25 km will always resist subduction for any observed age of oceanic crust.

The transition from localized slip along a plate boundary to distributed flow appears to correspond to where the slab reaches 70–90 km depth in many settings (Wada and Wang 2009; Abers et al. 2006; Syracuse et al. 2010). Evidence for this transition comes from heat-flow observations at a small number of arcs (northern Japan and Cascadia), which show a step from very low forearc heat flow to relatively high sub-arc and back-arc heat flow, at this point. Also, seismic attenuation shows a sharp step at this depth, in a manner consistent with a cold “nose” of mantle material overlying the shallower portion of the slab, transitioning rapidly to hot mantle beneath the arc. The cause of this transition likely has to do with the maximum depth at which plate motion can be accommodated by a localized shear zone, and hence rheology of that shear zone; the location of arcs relative to the slab indicates a transition at temperatures about 550 °C (Conder 2005; Syracuse et al. 2010).

Fate of the Downgoing Plate: Devolatilization and Metamorphism

The subducting oceanic lithosphere can be divided into several layers (Fig. 2, inset):

- (a) Deep-sea sediment, of widely variable thickness and composition,
- (b) Variably altered oceanic crust (AOC) – largely the upper, basaltic part where hydrothermal circulation has the largest effect,
- (c) Less altered gabbroic lower crust, and
- (d) Subducting mantle lithosphere some of which may be hydrated to varying degrees.

Although some of this package will be scraped off in accretionary margins, much probably survives to subduct to subarc depths. Arc geochemistry includes signatures of subducted sediment, with lava chemistry that varies systematically in some elements in ways that correlate with variations in sediment chemistry (Plank and Langmuir 1998). This observation requires that most of the sediment, especially that deposited away from the trench, subduct to sub-arc depths. For example, lavas at several arcs include trace amounts of ^{10}Be , a radionuclide with ~1.6 Ma half-life formed in the atmosphere and deposited with the shallowest

sediment (Tera et al. 1986), indicating that the shallowest deep-marine sediment subducts intact and the material reaches the earth’s surface within a few Ma (a couple of half-lives). Finally, seismological evidence also supports the deep subduction of oceanic crust relatively intact, to generate images seen in receiver functions (e.g., Rondenay et al. 2008; Kawakatsu and Watada 2007), high-frequency guided waves (Abers 2000), and other mode conversions (e.g., Helffrich 1996).

In its entirety, this package subducts on the order 2×10^9 Tg/Ma of water at the trench (Hacker 2008), about half of which enters as pore water that is expelled by shallow compaction, and the remainder bound in minerals. Numerous drilling and sampling studies constrain well the H_2O content (and CO_2) of the subducting sediment and AOC, but lower sections are relatively less well known. Perhaps the largest uncertainty in these estimates lies in the water content of subducted mantle lithosphere, a potentially vast reservoir. Peridotite will hydrate to serpentinite and serpentine can hold ~14 wt% H_2O , giving this very thick layer potential to dominate the global water budget. However, it is not yet known how much H_2O reaches mantle depths; probably mantle hydration at mid-ocean ridges is minor because near-ridge mantle temperatures there should exceed that of serpentine stability (~600 °C) but later faulting such as at fracture zones and the outer rise (see “[Thermal Structure, Plate Buoyancy, Mantle Flow](#)”) could be major conduits. Less is known about concentrations of CO_2 and other volatiles degassed by metamorphism. Depending upon serpentinite contributions, the volatile outputs of volcanic arcs either roughly match the inputs, implying near-complete volatile recycling, or are significantly less, implying a long-term flux of H_2O from the surface ocean to deep mantle.

Anhydrous oceanic crust can be expected to undergo a series of phase transitions from gabbro to eclogite. These result in a major density increase; gabbro has a density ~10% lower than that of peridotite at similar temperatures and pressures, so unmetamorphosed oceanic crust has buoyancy that resists subduction, while eclogite (dominated by garnet and pyroxenes) is significantly denser than peridotite. The density increase, in transforming from gabbro to eclogite, may contribute significantly to the negative buoyancy of slabs that drives plate tectonics. However, reaction rates for garnet growth (a major step in eclogite formation) within coarse-grained, anhydrous gabbro may be relatively slow, such that gabbro may persist metastably to considerable depth within subduction zones (Ahrens and Schubert 1975). The presence of water should substantially enhance that reaction rate (e.g., Hacker 1996).

If hydrated, basalt/gabbro (and metasediment) are expected to undergo a lengthy sequence of metamorphic reactions that, generally, release H_2O (Fig. 4). As a consequence, H_2O should be released over a wide range of depths from the thrust zone to

sub-arc mantle, hydrating the overlying mantle wedge (Schmidt and Poli 1998; Hacker 2008). At typical to cold subduction conditions, oceanic crust can be expected to transition through greenschist and blueschist facies, capable of holding up to ~5 wt% H₂O, with major dehydration occurring in the transformation to eclogite during high-pressure heating (Hacker et al. 2003). In warmer subduction zones, defined as those with fairly young incoming plates and slow subduction rates (e.g., Cascadia, southern Japan), descending crust should follow a higher-temperature path perhaps through amphibolite facies, dehydrating as amphibole breaks down at ~2 GPa. Most thermal models with realistic mantle-wedge viscosity structure predict that subducting crust should dehydrate by depths of 100–200 km (e.g., Syracuse et al. 2010), at least in its shallow parts.

The mantle within the subducting plate potentially holds hydrous minerals as well, primarily serpentine but also chlorite, as might entrained material lying above the downgoing plate. These layers should also dehydrate, at temperatures of 600–700 °C for serpentines at <200 km depth (e.g., Schmidt and Poli 1998) with chlorite breaking down at temperatures about 200° higher. Serpentine dehydration temperatures are probably reached in all subduction zones on the slab surface and, hence, above it, although it is possible that a weakly hydrated zone of chlorite stability may persist past slab depths (e.g., Grove et al. 2009). Subducting mantle lithosphere in most subduction zones should be cold enough for serpentine to be stable past the arc (Syracuse et al. 2010), and may dehydrate gradually past that. Water persisting to depths past 200 km has the potential to remain within the mantle to considerable depth, in the “alphabet” phases of dense hydrous magnesium silicates, which have broad stability fields at near-slab conditions (Thompson 1992). Water reaching the transition zone (410–660 km depth) may remain for a long time, as the dominant minerals ringwoodite and wadsleyite can retain wt% water as defects in their nominally anhydrous structure (Kohlstedt et al. 1996). This reservoir has potential to hold several times more water than the current surface ocean; thus, the ability of serpentinized mantle lithosphere to transport H₂O to great depths may be a dominant control on the Earth’s long-term water budget.

As water is released, it ascends into the overlying mantle wedge, crust, back along the thrust zone, and through the forearc (Peacock 1990), potentially hydrating it (Hyndman and Peacock 2003). These pathways provide major controls on rheological properties in all these regions and on the behavior and composition of arc volcanoes. Also, it is often suggested that the dehydration process generates earthquakes at intermediate depths (70–150 km) in subduction zones (Green and Houston 1995; Kirby et al. 1996; Hacker et al. 2003). These processes are discussed extensively in the entry on ► [“Seismicity, Subduction Zone”](#).

Subduction Channel

In many collisional mountain belts and accreted terranes, high-pressure, low-temperature rocks have reached the Earth’s surface after having descended to great depths. In some “ultra-high pressure” (UHP) terranes, mineral assemblages or textures in these rocks record pressures in excess of 3–4 GPa (depths >100 km), indicating that these rocks descended deeply in subduction zones before returning to the surface. In part to explain these observations, a “subduction channel” has been often proposed (Cloos and Shreve 1988). The subduction of weak, hydrated metasediments and altered oceanic crust, and their subsequent shearing and interaction with metamorphic fluids, suggests the presence of a heterogeneous weak zone at the top of the downgoing plate. Initially, material in this channel will be dragged downward by coupling with the subducting plate, but it should be buoyant relative to overlying mantle, and in the right conditions a return flow could develop delivering high-grade metamorphic rocks to the surface. A weak zone is also indicated by the presence of a “cold nose” in many subduction zones in the forearc mantle wedge, a feature that probably requires a weak decoupling layer to depths of 70–90 km on the top of the downgoing plate (e.g., Wada and Wang 2009). The presence of UHP rocks at the earth’s surface indicates that the return-flow process or something similar probably happens in some instances, but it remains unclear if the subduction channel is a common feature of subduction zones.

Melting and the Volcanic Arc

Reviews elsewhere describe the geochemical evidence for the nature of melting beneath volcanic arcs and the generation of arc crust (e.g., Stern 2002; Kelemen et al. 2003b), which are briefly summarized here for their relevance to geophysical observations. Magmas emplaced and erupted at subduction zones span a wide range of compositions. Basalts are found at most subduction zones and indicate that primary melting occurs within the mantle beneath the volcanic arc, since basalts primarily form from melting of peridotites (e.g., Gill 1981). These basalts often differ from those found at mid-ocean ridges or hot spots in several important chemical trends; in particular, many elements that are strongly abundant in subducting sediment and altered crust are found in abundance in many volcanic arc rocks in a manner that correlates with sediment chemistry (Plank and Langmuir 1998). In some arcs there is strong evidence for andesites forming under high pressure, perhaps requiring melting of downgoing crust (e.g., Kelemen et al. 2003b).

Magmatic differentiation within the crust and mixing with preexisting arc basement result in a wide spectrum of volcanic

output. Typically, seismic velocities within arcs show an increase with depth consistent with more silicic, felsic rocks at shallow depths ($V_p = 6.0\text{--}6.4$ km/s) underlain by more gabbroic rocks in the middle-to-lower crust ($V_p = 7.0\text{--}7.4$ km/s) (e.g., Suyehiro et al. 1996; Shillington et al. 2004). Since they are petrologically fractionated from more basaltic parents, the andesitic or granitic rocks observed at mid-to-upper crustal levels should be balanced by more mafic silica-poor residuals, perhaps in the lower crust. However, such lower-crustal cumulates are not very abundant, as interpreted from seismic velocities, exposed arc sections, and within continents formed by arc accretion. In general, there is good reason to think that continents grew from accretion of arcs since they share many geochemical characteristics (Rudnick 1995), but overall continents are significantly more silicic than primary arc basalts. This discrepancy probably requires a significant mass loss in the complementary, mafic cumulates that should form in arc lower crust. Where and how this mass loss occurs remains unclear; it may be that the lower crust in arcs occasionally founders or delaminates back into the mantle, since its density may be higher than that of peridotite, but many other possibilities exist. Geophysical observations of the lower crust and upper mantle have potential to place constraints on the formation and evolution of arc crust, and hence the origins of continental material (e.g., Kodaira et al. 2007; Tatsumi et al. 2008).

Primary magmas emerging from the mantle contain several wt% H_2O and elevated CO_2 , several times higher than in mid-ocean ridge basalts (e.g., Hirschmann 2006). Mantle melting beneath arcs often occurs in the presence of H_2O (Grove et al. 2006), leading to the frequent speculation that dehydration reactions within the subducting slab trigger melting in the overlying mantle wedge, and hence the location of dehydration within the slab controls the location of arc volcanism (e.g., Tatsumi 1986). However, experiments on hydrous metabasalts show that dehydration occurs over a wide range of depths beginning much shallower than the sub-arc mantle wedge (e.g., Schmidt and Poli 1998), so that H_2O should flux the mantle at many depths, not just under the volcanic arc. Also, primary, dry basalts have been found in some arcs, so H_2O is not always required for melting. It may be that the return flow (the incoming limb of wedge corner flow) sufficiently decompresses hot asthenosphere to trigger melting. The relative importance of wet (flux) melting and dry (decompression) melting within arcs remains poorly understood, as do the pathways that magmas take from their locus of formation to volcanic arcs. In particular, it is not obvious why the arc front is a narrow, well-defined feature while the conditions for melting should be more widely spread. The depth to slabs beneath the arc front averages 110 km but varies coherently by a factor of 2 from arc to arc, suggesting some other controlling process probably related to thermal structure (England et al. 2004; Syracuse and Abers 2006).

Seismic Imaging of the Subduction Zone

Although subduction zones exhibit a characteristic surface morphology of trench, forearc, volcanic arc, etc., the evidence that these associations represent subduction of deep material comes primarily from geophysical imaging of deep structure. Early observations of seismic wave propagation in Tonga showed that the Wadati–Benioff zone lies in a cold, low-attenuation structure unlike mantle elsewhere on the planet, and provided some of the first evidence that these were sites of descending cold lithosphere (Oliver and Isacks 1967). Since then, methods of seismic tomography have improved greatly (see ► “[Seismic Tomography](#)”) as has the global data set of earthquake travel times, leading to a clear picture of P-wave high-velocity anomalies that show cold slabs descending through the upper mantle and in some cases into the lower mantle (e.g., Van der Hilst et al. 1997). These data show slabs as seismic anomalies 4–10% faster than surrounding mantle (see compilation by Lay 1997), consistent with thermal models that predict temperatures 400–800 °C colder than surroundings. These images, based on global earthquake travel time data sets, show that in some cases slabs reach the mantle transition zone and flatten out, as if reaching a partial barrier to flow (e.g., Japan) while in other cases they pass through the transition zone delivering subducted material to the lower mantle.

Seismic recordings from networks and arrays in subduction zones provide strong constraints on the subduction zone structure of a scale relevant to slab metamorphism, arc petrogenesis, and wedge flow. For example, as dehydration occurs, densities increase as do seismic velocities; basalts are 10–15% slower than surrounding peridotites while eclogites have similar velocities. Furthermore, serpentinized peridotite can be tens of percent slower than peridotite particularly for shear waves, depending upon the extent of serpentinization (hydration). As a result, seismic imaging has potential to show where and how metamorphism within slabs occurs. For example, with migration of P-to-S scattered waves (see ► “[Seismic, Receiver Function Technique](#)”) subducting crust has been observed to ~45 km depth in Cascadia and depths exceeding 100 km in Alaska and Japan, consistent with expected depths of basalt dehydration and differences in thermal structure (Rondenay et al. 2008). Studies of earthquake travel times and signal attenuation also provide information on structure of the mantle wedge where melting initiates, and have shown a clear hot mantle wedge, with a variety of structures suggesting melt transport pathways. Probably the densest long-lived monitoring effort exists in Japan, although a variety of portable array experiments have illuminated subduction zones in parts of the Andes, Central America, Tonga, the Marianas, Alaska, Cascadia, Greece, and other locales.

One of the more complex suites of seismic observations comes from the anisotropic fabric within the mantle of subduction zones, which results in “splitting” of shear waves into orthogonally polarized fast and slow pulses (see ► [“Magnetic Anisotropy”](#)). Mantle flow in the dislocation creep regime leads to lattice-preferred orientation of olivine, and olivine is moderately (~ 10%) anisotropic, leading to an association between seismic anisotropy and mantle flow. However, observations in many subduction zones lie counter to expectation (e.g., Russo and Silver 1994). Simple models of mantle flow in subduction zones predict “corner flow,” in which mantle on either side of the descending plate is entrained via viscous drag (Fig. 2). To fill the potential void left by the downward advection, mantle far from the subducting slab flows inward toward it, eventually turning and flowing downward. While details of this flow depend upon the viscosity structure and its sensitivity to temperature, stress, water, and other variables, the overall pattern of flow is one in which the maximum strain or stretch should be perpendicular to trench and arc, in the direction of flow. In many subduction zones, anisotropic fabric does not show this pattern, and instead the fast direction for shear wave propagation is more often than not parallel to the trench, at least below and behind the volcanic arc (Fischer et al. 1998). More complicated fabrics are also observed, with sharp 90° changes in fabric between sub-arc mantle and the shallower part of the forearc (e.g., in North Japan and Alaska), and the pattern varies between different subduction zones. Explanations for these unexpected patterns remain a matter of much debate, and include strong along-strike flow, secondary flows within the wedge, influence of melt on fabric, activation of unusual olivine slip systems at some subduction zone conditions, abundance of hydrous, foliated minerals in the forearc, extensive fabric within the downgoing slab, and unexpectedly strong upper-plate fabrics.

Subduction Zone Hazards

The forces present in subduction zones also produce some of the most violent natural hazards known. The interplate thrust zone generates all of the largest earthquakes, which produce strong shaking but also submarine landslides, meter-scale uplift and subsidence, and most of the largest tsunamis. The 2004 Sumatra earthquake represents the primary example of the last 50 years; a magnitude 9.2 earthquake ruptured over 1,000 km of plate boundary, producing an ocean-crossing tsunami that directly resulted in over 150,000 deaths and huge devastation to property. These earthquakes are devastating because subduction zone thrust faults are very large, because they are submarine but can break the sea surface, and because they occur in places with often oversteepened bathymetry. Subduction zones are also site to many of the most violent volcanic eruptions. Large volatile (especially water) contents of subduction zone magmas, compared with

most other settings, lead to magmas rich in gas that exsolve at shallow depths. Magmatic differentiation leads to relatively viscous, felsic magmas that rise slowly and do not degas easily. The combination leads to explosive and sometimes very large eruptions; in the last half century, Mt. St Helens (Cascadia), Pinatubo (Philippines), and Montserrat (Lesser Antilles) all produced violent and well-documented eruptions in subduction zones, among many others. Worldwide, about 500 volcanoes have known historic eruptions, roughly two-thirds in subduction zones.

Summary

Subduction zones represent the primary return of material from the earth’s surface to its interior. In the process of doing so, great earthquakes occur, violent volcanoes erupt, and the largest lateral temperature gradients in the Earth’s interior are generated. At shallow depths, elastic and brittle processes control deformation including most of the planet’s large earthquakes, while at greater depths the subduction system deforms by large-scale ductile flow. Along with oceanic crust, sediment and volatiles enter the trench, lubricating the plate boundary at shallow depths, and then are metamorphosed where slabs descend below hot mantle, releasing H₂O and other volatiles into the overlying mantle wedge. While some of this material may mechanically ascend back to the surface, in the form of high-pressure metamorphic rocks, much feeds explosive volcanism in the volcanic arc or descends into the deep mantle. Melting at sub-arc depths is controlled by a combination of advection via flow of hot mantle from the back-arc, and fluxing by slab-derived fluids, a combination of which produces a sharp volcanic front. Magmas ascend to form arc crust and differentiate, some fraction of which persists to form the basis of continental crust. This transfer of energy and material leads to a rich set of phenomena that play a major role in creating the geologic record.

Cross-References

- [Earthquake, Magnitude](#)
- [Magnetic Anisotropy](#)
- [Seismic Tomography](#)
- [Seismic, Receiver Function Technique](#)
- [Tsunami](#)

Bibliography

- Abers GA (2000) Hydrated subducted crust at 100–250 km depth. *Earth Planet Sci Lett* 176:323–330
- Abers GA, van Keken PE, Kneller EA, Ferris A, Stachnik JC (2006) The thermal structure of subduction zones constrained by seismic imaging: implications for slab dehydration and wedge flow. *Earth Planet Sci Lett* 241:387–397

- Ahrens TJ, Schubert G (1975) Gabbro-eclogite reaction rate and its geophysical significance. *Rev Geophys Space Phys* 13:383–400
- Bevis M et al (1995) Geodetic observations of very rapid convergence and back-arc extension at the Tonga arc. *Nature* 374:249–251
- Billen MI (2008) Modeling the dynamics of subducting slabs. *Annu Rev Earth Planet Sci* 36:325–356
- Chapple WM, Forsyth DW (1979) Earthquakes and bending of plates at trenches. *J Geophys Res* 84:6729–6749
- Cloos M (1993) Lithospheric buoyancy and collisional orogenesis; subduction of oceanic plateaus, continental margins, island arcs, spreading ridges, and seamounts. *Geol Soc Am Bull* 105:715–737
- Cloos M, Shreve RL (1988) Subduction-channel model of prism accretion, melange formation, sediment subduction, and subduction erosion at convergent plate margins: 1. Background and description. *Pure Appl Geophys* 128:455–500
- Conder JA (2005) A case for hot slab surface temperatures in numerical viscous flow models of subduction zones with an improved fault zone parameterization. *Phys Earth Planet Inter* 149:155–164
- Davis D, Suppe J, Dahlen FA (1983) Mechanics of fold-and-thrust belts and accretionary wedges. *J Geophys Res* 88:1153–1172
- DeMets C (2001) A new estimate for present-day Cocos-Caribbean plate motion; implications for slip along the central American volcanic arc. *Geophys Res Lett* 28:4043–4046
- DeMets C, Gordon R, Argus D, Stein S (1990) Current plate motions. *Geophys J Int* 101:425–478
- England P, Engdahl ER, Thatcher W (2004) Systematic variations in the depths of slabs beneath arc volcanoes. *Geophys J Int* 156:377–408
- Fischer KM, Fouch MJ, Wiens DA, Boettcher MS (1998) Anisotropy and flow in Pacific subduction zone back-arcs. *Pure Appl Geophys* 151:463–475
- Fitch TJ (1972) Plate convergence, transcurrent faults and internal deformation adjacent to southeast Asia and the western Pacific. *J Geophys Res* 77:4432–4460
- Forsyth DW, Uyeda S (1975) Relative importance of driving forces of plate motion. *Geophys J R Astron Soc* 43:163–200
- Gerya TV, Yuen DA (2003) Rayleigh-Taylor instabilities from hydration and melting propel ‘cold plumes’ at subduction zones. *Earth Planet Sci Lett* 212:47–62
- Gill JB (1981) *Orogenic andesites and plate tectonics*. Springer, New York, p 390
- Gomberg J (2010) Slow-slip phenomena in Cascadia from 2007 and beyond: a review. *Geol Soc Am Bull* 122:963–978
- Gorbatov A, Kostoglodov V (1997) Maximum depth of seismicity and thermal parameter of the subducting slab: general empirical relation and its application. *Tectonophysics* 277:165–187
- Green HW, Houston H (1995) The mechanics of deep earthquakes. *Annu Rev Earth Planet Sci* 23:169–213
- Grove TL, Chatterjee N, Parman SW, Médard E (2006) The influence of H₂O on mantle wedge melting. *Earth Planet Sci Lett* 249:74–89
- Grove TL, Till CB, Lev E, Chatterjee N, Médard E (2009) Kinematic variables and water transport control the formation and location of arc volcanoes. *Nature* 459:694–697
- Hacker BR (1996) Eclogite formation and the rheology, buoyancy, seismicity, and H₂O content of oceanic crust. In: Bebout GE, Scholl D, Kirby S, Platt JP (eds) *Subduction: top to bottom*. AGU monograph series, vol 96. American Geophysical Union, Washington, DC, pp 337–346
- Hacker BR (2008) H₂O subduction beyond arcs. *Geochem Geophys Geosyst* 9:Q03001
- Hacker BR, Abers GA, Peacock SM (2003) Subduction factory 1: theoretical mineralogy, density, seismic wavespeeds, and H₂O content. *J Geophys Res* 108:2029
- Helffrich G (1996) Subducted lithospheric slab velocity structure: observations and mineralogical inferences. In: Bebout GE, Scholl D, Kirby S, Platt JP (eds) *Subduction: top to bottom*. AGU monograph series, vol 96. American Geophysical Union, Washington, DC, pp 215–222
- Hilde TWC (1983) Sediment subduction versus accretion around the Pacific: convergence and subduction. *Tectonophysics* 99:381–397
- Hirschmann MM (2006) Water, melting, and the deep earth H₂O cycle. *Annu Rev Earth Planet Sci* 34:629–653
- Hoernle K, Abt DL, Fischer KM, Nichols H, Hauff F, Abers GA, van den Bogaard P, Heydolph K, Alvarado G, Protti M, Strauch W (2008) Arc-parallel flow in the mantle wedge beneath Costa Rica and Nicaragua. *Nature* 451:1094–1098
- Hyndman RD, Peacock SM (2003) Serpentinization of the forearc mantle. *Earth Planet Sci Lett* 212:417–432
- Hyndman RD, Wang K (1995) The rupture zone of Cascadia great earthquakes from current deformation and the thermal regime. *J Geophys Res* 100:22133–22154
- Hyndman RD, Yamano M, Oleskevich DA (1997) The seismogenic zone of subduction thrust faults. *Island Arc* 6:244–260
- Ide S, Beroza G, Shelly DR, Uchide T (2007) A scaling law for slow earthquakes. *Nature* 447:76–79
- Ito Y, Obara K, Shiomi K, Sekine S, Hirose H (2007) Slow earthquakes coincident with episodic tremors and slow slip events. *Science* 315:503–506
- Kawakatsu H, Watada S (2007) Seismic evidence for deepwater transportation in the mantle. *Science* 316:1468–1471
- Kelemen PB, Rilling JL, Parmentier EM, Mehl L, Hacker BR (2003a) Thermal structure due to solid-state flow in the mantle wedge beneath arcs. In: Eiler JM (ed) *Inside the subduction factory*. Geophysical monograph, vol 138. AGU, Washington, DC, pp 293–311
- Kelemen PB, Hanghøj K, Greene AR (2003b) One view of the geochemistry of subduction related magmatic arcs, with emphasis on primitive andesite and lower crust. In: Rudnick RL (ed) *The crust. Treatise on geochemistry*, vol 3. Elsevier, Oxford
- Kirby S, Engdahl ER, Denlinger R (1996) Intermediate-depth intraslab earthquakes and arc volcanism as physical expressions of crustal and uppermost mantle metamorphism in subducting slabs. In: Bebout GE, Scholl D, Kirby S (eds) *Subduction: top to bottom*. Geophysical monograph, vol 96. American Geophysical Union, Washington, DC, pp 195–214
- Kodaira S, Sato T, Takahashi N, Miura S, Tamura Y, Tatsumi Y, Kaneda Y (2007) New seismological constraints on growth of continental crust in the Izu-Bonin intra-oceanic arc. *Geology* 35:1031–1034
- Kohlstedt DL, Keppler H, Rubie DC (1996) Solubility of water in the alpha, beta and gamma phases of (Mg, Fe)₂SiO₄. *Contrib Mineral Petrol* 123:345–357
- Lallemand S, Hueuret A, Boutelier D (2005) On the relationships between slab dip, back-arc stress, upper plate absolute motion, and crustal nature in subduction zones. *Geochem Geophys Geosyst* 6:Q090006
- Lay T (1997) *Structure and fate of subducting slabs*. Academic, New York, p 185
- McCaffrey R (1992) Oblique plate convergence, slip vectors, and forearc deformation. *J Geophys Res* 97:8905–8915
- McKenzie DP (1969) Speculations on the consequences and causes of plate motions. *Geophys J R Astron Soc* 18:1–32
- Molnar P, England PC (1990) Temperatures, heat flux, and frictional stress near major thrust faults. *J Geophys Res* 95:4833–4856
- Molnar P, Freedman D, Shih JSF (1979) Lengths of intermediate and deep seismic zones and temperatures in downgoing slabs of lithosphere. *Geophys J R Astron Soc* 56:41–54
- Obara K (2002) Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science* 296:1679–1681
- Oliver J, Isacks B (1967) Deep earthquake zones, anomalous structures in the upper mantle, and the lithosphere. *J Geophys Res* 72:4259–4275
- Parsons B, Molnar P (1976) The origin of outer topographic rises associated with trenches. *Geophys J R Astron Soc* 45:707–712
- Peacock SM (1990) Fluid processes in subduction zones. *Science* 248:329–337

- Plank T, Langmuir CH (1998) The chemical composition of subducting sediment and its consequences for the crust and mantle. *Chem Geol* 145:325–394
- Ranero C, Morgan J, McIntosh K, Reichert C (2003) Bending-related faulting and mantle serpentinization at the middle America trench. *Nature* 425:367–373
- Rogers G, Dragert H (2003) Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science* 300:1942–1943
- Rondenay S, Abers GA, van Keken PE (2008) Seismic imaging of subduction zone metamorphism. *Geology* 36:275–278
- Rudnick RL (1995) Making continental crust. *Nature* 378:571–578
- Rudnick RL, Fountain DM (1995) Nature and composition of the continental crust: a lower crustal perspective. *Rev Geophys* 33:267–309
- Russo RM, Silver PG (1994) Trench-parallel flow beneath the Nazca plate from seismic anisotropy. *Science* 263:1105–1111
- Saffer DM, Marone C (2003) Comparison of smectite- and illite-rich gouge frictional properties: application to the updip limit of the seismogenic zone along subduction megathrusts. *Earth Planet Sci Lett* 215:219–235
- Schmidt MW, Poli S (1998) Experimentally based water budgets for dehydrating slabs and consequences for arc magma generation. *Earth Planet Sci Lett* 163:361–379
- Scholz CH (1998) Earthquakes and friction laws. *Nature* 391:37–42
- Schwartz SY, Rokosky JM (2007) Slow slip events and seismic tremor at circum-pacific subduction zones. *Rev Geophys* 45:RG3004
- Shillington D, Van Avendonk HJA, Holbrook WS, Kelemen PB, Hornbach MJ (2004) Composition and structure of the central Aleutian island arc from arc-parallel wide-angle seismic data. *Geochem Geophys Geosyst* 5:Q10006
- Stern RJ (2002) Subduction zones. *Rev Geophys* 40:1012
- Suyehiro K, Takahashi N, Arie Y, Yokoi Y, Hino R, Shinohara M, Kanazawa T, Hirata N, Tokuyama H, Taira A (1996) Continental crust, crustal underplating, and low-Q upper mantle beneath an oceanic island arc. *Science* 272:390–392
- Syracuse EM, Abers GA (2006) Global compilation of variations in slab depth beneath arc volcanoes and implications. *Geochem Geophys Geosyst* 7:Q05017
- Syracuse EM, van Keken PE, Abers GA (2010) The global range of subduction zone thermal models. *Phys Earth Planet Inter* 183:73–90
- Tatsumi Y (1986) Formation of the volcanic front in subduction zones. *Geophys Res Lett* 13:717–720
- Tatsumi Y, Shukuno H, Tani K, Takahashi N, Kodaira S, Kogiso T (2008) Structure and growth off the Izu-Bonin-Mariana arc crust: 2. Role of crust-mantle transformation and the transparent Moho in arc crust evolution. *J Geophys Res* 113:B02203
- Tera F, Brown L, Morris J, Sacks IS, Klein J, Middleton R (1986) Sediment incorporation in island-arc magmas: inferences from ^{10}Be . *Geochim Cosmochim Acta* 50:535–550
- Thompson AB (1992) Water in the Earth's upper mantle. *Nature* 358:295–302
- Tichelaar BW, Ruff LJ (1993) Depth of seismic coupling along subduction zones. *J Geophys Res* 98:2017–2037
- Van der Hilst RD, Widyantoro S, Engdahl ER (1997) Evidence for deep mantle circulation from global tomography. *Nature* 386:578–584
- van Keken PE, Kiefer B, Peacock SM (2002) High-resolution models of subduction zones: implications for mineral dehydration reactions and the transport of water into the deep mantle. *Geochem Geophys Geosyst* 3:1056
- Von Herzen R, Ruppel C, Molnar P, Nettles M, Nagihara S, Ekstrom G (2001) A constraint on the shear stress at the Pacific-Australia plate boundary from heat flow and seismicity at the Kermadec forearc. *J Geophys Res* 106:6817–6833
- von Huene R, Scholl DW (1991) Observations at convergent margins concerning sediment subduction, subduction erosion, and the growth of continental crust. *Rev Geophys* 29:279–316
- Wada I, Wang K (2009) Common depth of slab-mantle decoupling: reconciling diversity and uniformity of subduction zones. *Geochem Geophys Geosyst* 10:Q10009
- Watts AB, Talwani M (1974) Gravity anomalies seaward of deep-sea trenches and their tectonic implications. *Geophys J R Astron Soc* 36:57–90

Surface Waves

Barbara Romanowicz

Berkeley Seismological Laboratory, Berkeley, CA, USA

Collège de France, Paris, France

Definition

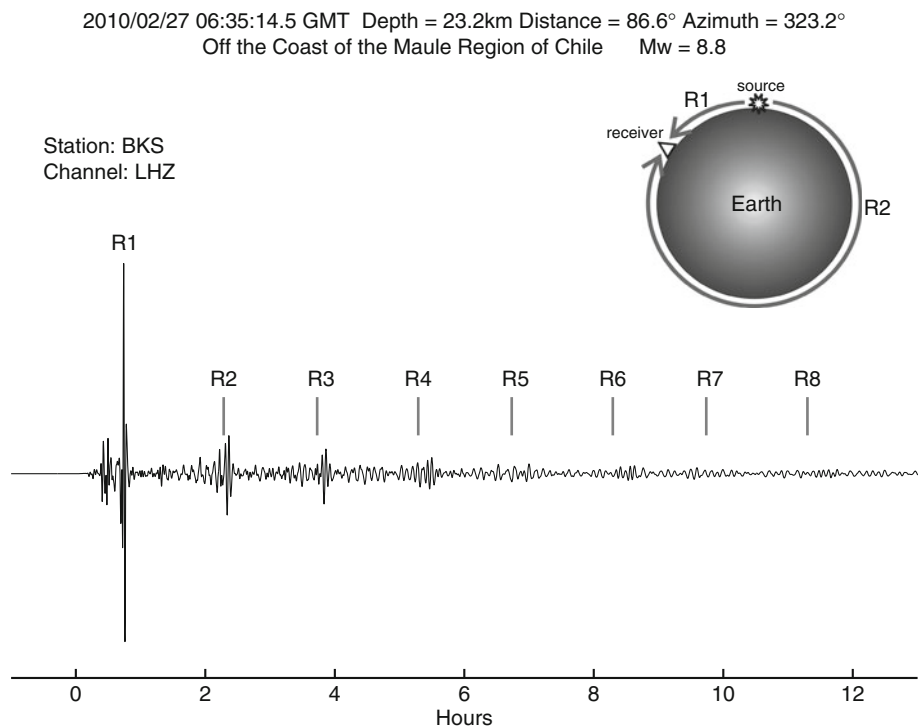
Surface waves are elastic waves which propagate along the surface of the earth and whose energy decays exponentially with depth.

Surface waves contain most of the long period energy (periods greater than 20 s) generated by earthquakes and recorded at teleseismic distances. Most prominent on records of moderate ($M \geq 5.5$) earthquakes are the fundamental mode wave trains that have propagated along the direct great circle path between the epicenter and the station. The dispersive and attenuative properties of these wave trains have been used extensively, since the 1950s, to infer crust and upper mantle structure at the regional scale. For earthquakes of magnitude 7 or larger, successive, earth-circling surface wave trains can be followed for many hours (Fig. 1) and are either analyzed individually or, at the longest periods ($T > 250$ s), they are combined over time lengths of tens of hours or days to produce a spectrum of Earth's free oscillations (see ► [“Free Oscillations of the Earth”](#)).

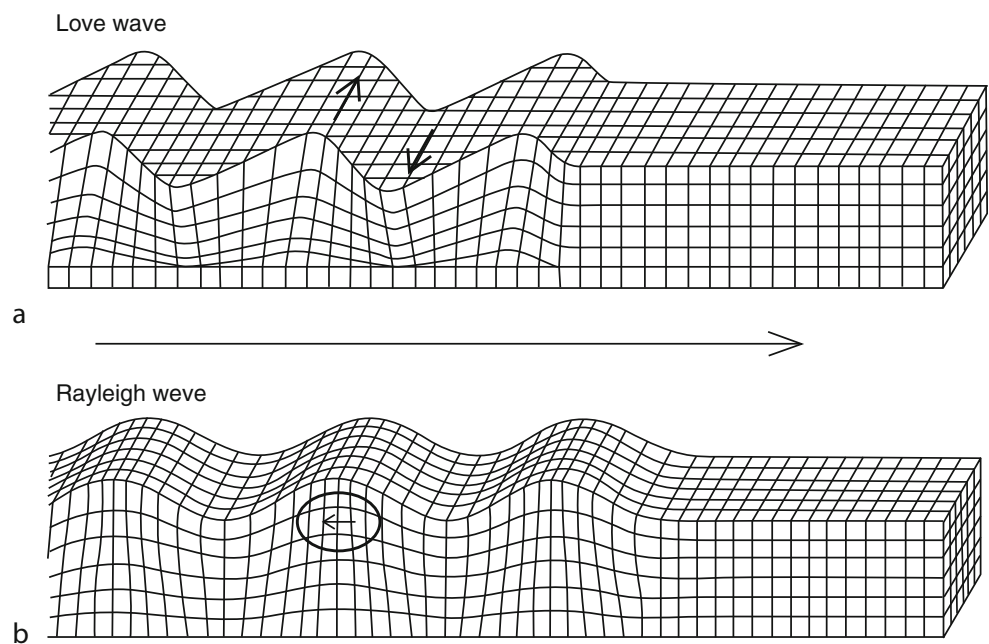
In an isotropic medium, there are two types of surface waves which propagate independently (Fig. 2). Rayleigh waves, polarized in the vertical plane containing the source and the receiver (P-SV energy), have a retrograde elliptical particle motion. These waves have generally good signal-to-noise ratio on the vertical component records and have been studied the most. On the other hand, the analysis of Love waves, polarized horizontally in a direction perpendicular to the propagation direction (SH energy), suffers from the more complex data processing required, and from higher levels of long period background noise on the horizontal component records, due primarily to the effect of atmospheric pressure variations, inducing ground tilts. Love waves, unlike Rayleigh waves, can exist only in a heterogeneous medium (i.e., a layered medium).

The most striking property of surface waves is their dispersive character, with, in general, waves of longer periods

Surface Waves, Fig. 1 Example of vertical component record showing earth-circling Rayleigh wave trains (marked R1...R8) following the M8.8 Maule Chile earthquake of February 27, 2010. This record is from station BKS of the Berkeley Digital Seismic Network (BDSN). The successive wavepackets are well separated from each other at this epicentral distance (86.6°). The long period surface waves are often referred to as “mantle waves.” (Courtesy of Shan Dou)



Surface Waves, Fig. 2 Particle motion for Rayleigh and Love waves. Rayleigh waves have retrograde particle motion confined to the vertical plane of propagation. Love waves have purely transverse motion in the horizontal plane. (After Bolt (1976))



traveling faster, directly reflecting the increase of elastic velocity with depth in the earth's crust and mantle. Indeed, studies of crustal and upper mantle structure progressed rapidly in the 1950s and early 1960s, as the tools developed to measure group and phase velocity dispersion and interpret them in terms of layered mantle and crust models (e.g., Ewing et al. 1957).

In a spherical earth, the theory of generation and propagation of surface waves is intimately linked to that of Earth's free oscillations, leading to the distinction of various modes. The lowest mode is termed fundamental mode, followed by higher modes, or overtones. Fundamental mode Love and Rayleigh waves are generally well separated from other seismic phases on the seismograms, and are well excited by

shallow, crustal earthquakes, while higher modes (i.e., overtones) travel at higher group velocities and appear as compact packets of mixed overtones. They are better excited by deeper earthquakes (Fig. 3).

Surface waves recorded at teleseismic distances contain information about both the characteristics of the earthquake source and the structure of the earth's crust and mantle along the source-station path. Separating the two effects has been one of the long-standing challenges faced by seismologists.

Group and Phase Velocity

The velocity of propagation of surface waves can be described in terms of either group velocity or phase velocity. Phase velocity is the velocity of a specific frequency component of the wave train. Phase velocity derives from the expression of a plane wave of wavenumber k , and angular frequency ω , propagating in the direction x , for example, in one dimension:

$$f(x, t) \sim \exp i(\omega t - kx) \quad (1)$$

It is defined as the velocity of the phase of the component of frequency ω :

$$C(\omega) = \omega/k \quad (2)$$

where ω is the angular frequency and k is the wavenumber. Frequency is related to period of the wave, T , by $\omega = 2\pi/T$, while $k = 1/\lambda$ where λ is wavelength.

Group velocity is the velocity with which the energy travels. The group velocity is derived from considering a wave train containing multiple frequencies:

$$u(\omega, t) = \int A(\omega) \exp i(\omega t - kx) d\omega \quad (3)$$

and finding the point (x, t) where the phase is stationary:

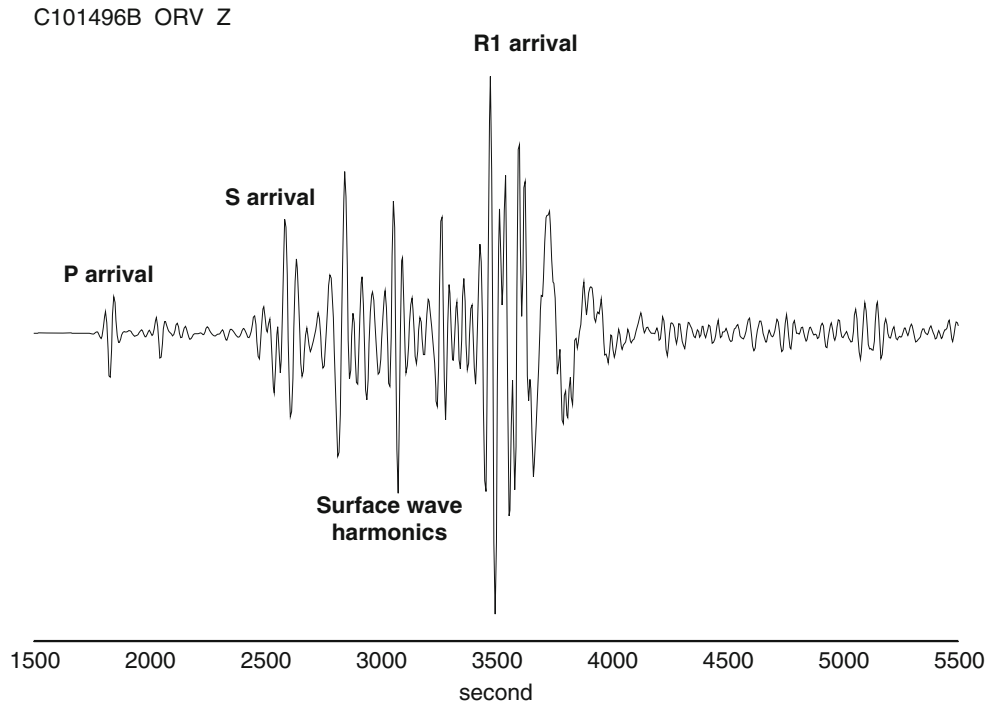
$$d(\omega t - kx)/d\omega = 0 \quad \text{hence} \quad t = xdk/d\omega \quad (4)$$

This leads to the definition of group velocity U as:

$$U(\omega) = d\omega/dk \quad (5)$$

In general, at any given period, the group velocity of surface waves in the earth will be less than the phase velocity. The minimum in the group velocity curve is termed Airy phase, a major feature of many surface wave records. Note that waves with periods shorter than the period of the Airy phase will exhibit reverse group velocity dispersion, with the shorter periods traveling faster. Phase velocity, on the other hand, always increases with period.

Surface Waves, Fig. 3 Example of vertical component record for the October 14, 1996, Solomon Islands earthquake (depth = 19 km, $M_w = 6.7$) at BDSN station ORV showing the arrivals of multiply reflected body wave phases forming a higher-mode Rayleigh wave train in front of the fundamental mode (R1). The Airy phase, corresponding to the group velocity minimum around 230 s, is well visible at the end of the R1 train. The seismogram has been bandpass filtered with cutoff frequencies at 35 and 400 s. (Courtesy of Yuancheng Gung)



Surface Wave Generation and Propagation

The theoretical formulation for the excitation of surface waves and normal modes of the earth, an eigenvalue problem, was developed in the 1960s (e.g., Haskell 1964), much stimulated by the occurrence of the great Chilean earthquake of May 22, 1960 (M 9.5). A computational method, following the original theoretical approach of Saito (1967) based on Runge-Kutta matrix integration method, has long been the main reference for the practical calculation of surface waves and normal modes in laterally homogeneous, elastic, flat, or spherical earth models. Today, a widely used approach for spherical geometry, efficient to relatively short periods (10 s), is based on a propagator matrix method in which minors of sets of solutions are used effectively (Woodhouse 1988). The association of a normal mode formalism to compute dispersion and excitation of surface waves (and complete seismograms), with a moment tensor formalism to describe the earthquake source (e.g., Backus and Mulcahy 1976) led to the rapid development of source studies based on surface waves in the 1980s.

To obtain the frequency spectrum of a single mode surface wave train from the expression of a seismogram obtained by summation of normal modes in a spherically symmetric earth, one uses Poisson's formula (e.g., Aki and Richards 1980), which decomposes the modes into infinite trains of propagating surface waves traveling in opposite directions around the earth. In this process, a high frequency approximation is used, in which the phase velocity (see below) of a surface wave is related to the corresponding normal mode frequency by Jeans' formula:

$$C(\omega) = \frac{a\omega_l}{l + 1/2} \quad (6)$$

where l is the angular order of the mode and ω_l its eigenfrequency, and a is the radius of the earth. In this high-frequency approximation, surface waves propagate along the great circle path between the epicenter and the station and are sensitive to structure only along this great circle.

The spectrum of a single mode propagating surface wave at distance Δ , azimuth θ , and angular frequency ω can be expanded as follows:

$$U(\Delta, \theta, \omega) = U_s(\theta, \omega)S(\Delta)U_p(\Delta, \theta, \omega)F(\omega, \theta_0)D(\omega)I(\omega) \quad (7)$$

where U_s is the source spectrum, U_p contains propagation effects, I is the instrument response, $S(\Delta)$ is the geometrical spreading term, and F and D describe the source process.

The propagation term U_p can be expressed as (e.g., Romanowicz 2002):

$$U_p(\Delta, \theta, \omega) = \frac{1}{(\sin \Delta)^{1/2}} \exp(i\pi/4) \exp(im\pi/2) \times \exp[-i\omega\Delta/C(\omega, \theta)] \exp[-\eta(\omega, \theta)\Delta] \quad (8)$$

where m denotes the number of polar passages and $C(\omega, \theta)$, $\eta(\omega, \theta)$ are, respectively, the average phase velocity and attenuation coefficient along the source-station path.

On the other hand, the source term $U_s(\theta, \omega)$ can be expressed as a linear combination of the moment tensor elements M_{ij} of the source.

Structure Studies Using Surface Waves

Fundamental Mode Studies (Dispersion)

Fundamental mode surface waves are well suited for the study of crust and upper mantle elastic structure, which can be deduced from their group and/or phase dispersion properties. They allow the sampling of vast areas of the globe that are otherwise devoid of seismic stations, such as the oceans.

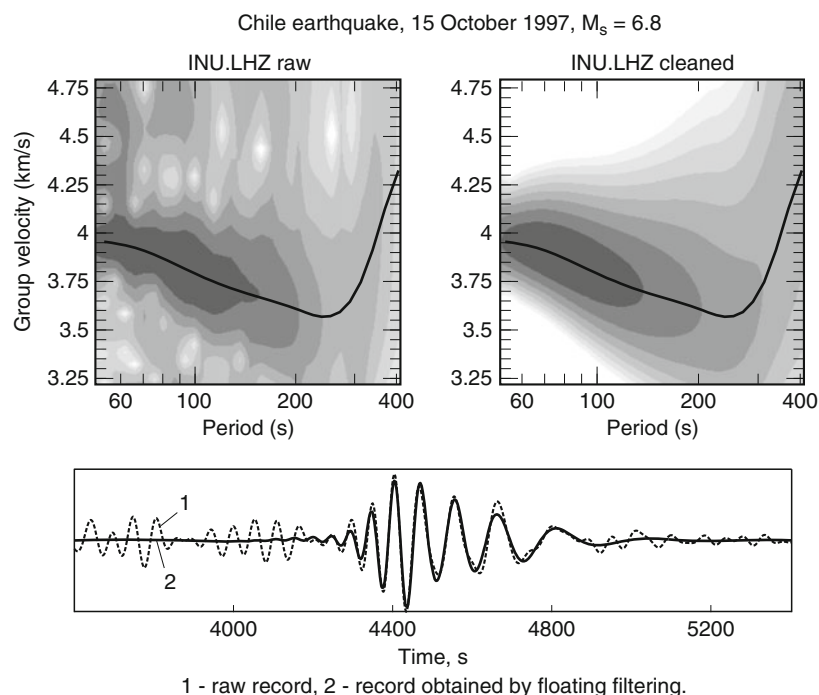
Many early studies documented the correlation of seismic velocity variations with surface tectonic features, using regional measurements of phase and group velocities of fundamental mode Love and Rayleigh waves in the period range 20–100 s (e.g., Knopoff 1972) or at longer periods, reaching deeper into the mantle (e.g., Toksöz and Anderson 1966). Since the early 1970s, the computation of group velocity has relied on the multiple filtering technique introduced by Dziewonski et al. (1969). This approach, later perfected as the FTAN method (e.g., Lander 1989), involves two steps. In the first step, an “energy diagram” is formed by windowing the seismogram over group arrival time and filtering over an array of specified center frequencies. The resulting amplitudes are then plotted as a function of time and frequency and their values contoured. The group velocity curve is obtained by tracing the loci of maximum amplitude as a function of period. In a second step, the time domain seismogram is filtered using multiple filters centered on the group velocity curve (e.g., Fig. 4).

Phase velocity is obtained from the phase Φ of the Fourier spectrum of a dispersed wave train that has been corrected for the contribution of the source and the instrument:

$$C(\omega) = \frac{X}{t_0 - [\Phi - N - (m/4) - 1/8]/\omega} \quad (9)$$

where t_0 is the start time of the Fourier window with respect to the event's origin time, m is the number of polar passages, and N is an integer arising from the 2π indeterminacy of the phase. This integer is determined first at long periods to obtain reasonable values of phase velocity compatible with well-

Surface Waves, Fig. 4 Example of group velocity dispersion diagrams obtained for the fundamental Rayleigh wave mode on the vertical record of the Chile Ms. 6.8 earthquake of October 15, 1997, at Geoscope station INU, using the FTAN multiple filtering approach. (a) Dispersion diagram before filtering; (b) same after filtering. The group velocity dispersion curve obtained follows the maxima of energy as delineated by the gray scale contours. (c) Corresponding time domain seismogram before (broken line) and after (solid line) variable filtering. The bandpass is 50–400 s. (Reprinted from Romanowicz (2002))



constrained global earth models. The phase velocity curve is then successively continued to shorter periods at fixed N , so as to obtain a smooth curve. This can become a problem at periods shorter than 30 s, where small variations in phase velocity correspond to rapid cycling of the phase.

The contribution of the earthquake source to the phase needs to be known accurately for the computation of phase velocity, less so for group velocity. This can be circumvented by making measurements using the “two-station method,” in which dispersion is measured between two stations approximately aligned with the epicenter on a great circle, thus eliminating the common source phase. If, in addition, the two stations are located within a relatively homogeneous geological province, such measurements are called “pure path” and lead directly to the determination of elastic velocity structure beneath that province. This method has been extended to “many stations” to infer structure beneath an array of stations spanning a geologically homogeneous region.

Noise Cross-Correlation Methodology

A radically different methodology was introduced in the early 2000’s. The medium’s “Green’s function” between two seismic stations can be extracted from cross-correlations of records at both stations over long time periods, without the need for using earthquake data (see ► [“Seismic, Ambient Noise Correlation”](#)). These “noise cross-correlations” have been particularly effective for the measurement of the fundamental mode dispersion between two stations in and around the microseismic frequency band (5–30 s), where the background seismic noise is the most energetic. It has been shown

that this approach leads to results that are as good as those obtained using standard methods, in regions where seismicity is strong and provides good illumination of the structure using “standard” approaches (e.g., Shapiro et al. 2005). This has opened up new horizons for the study of regional crustal structure, especially in seismically quiet regions. The measurement of fundamental mode dispersion by noise cross-correlation, combined with inversion for elastic structure (“Ambient-noise tomography”) has now been applied in many continental regions of the world, leading to higher resolution regional crustal models, especially when combined with information from converted phases at crustal discontinuities, which results in locally layered models of absolute shear velocity in the crust. Combining the noise-derived dispersion data with earthquake-based derived ones allows inversion for shear velocity structure down to upper mantle depths. It has been found that it is important to take into account biases introduced by the nonuniform distribution of noise sources around the study region. This is because most of the noise sources that contribute to the reconstruction come from the interaction of ocean waves generated by storms with the seafloor and have strong seasonal fluctuations (see section [“The Earth’s Background Noise Spectrum: Hum and Microseisms”](#)).

Global Studies of Structure Using Surface Waves

With the advent of digital recording in the mid-1970s and the expansion of global digital long-period and later broadband networks (see ► [“Seismological Networks”](#)), the processing

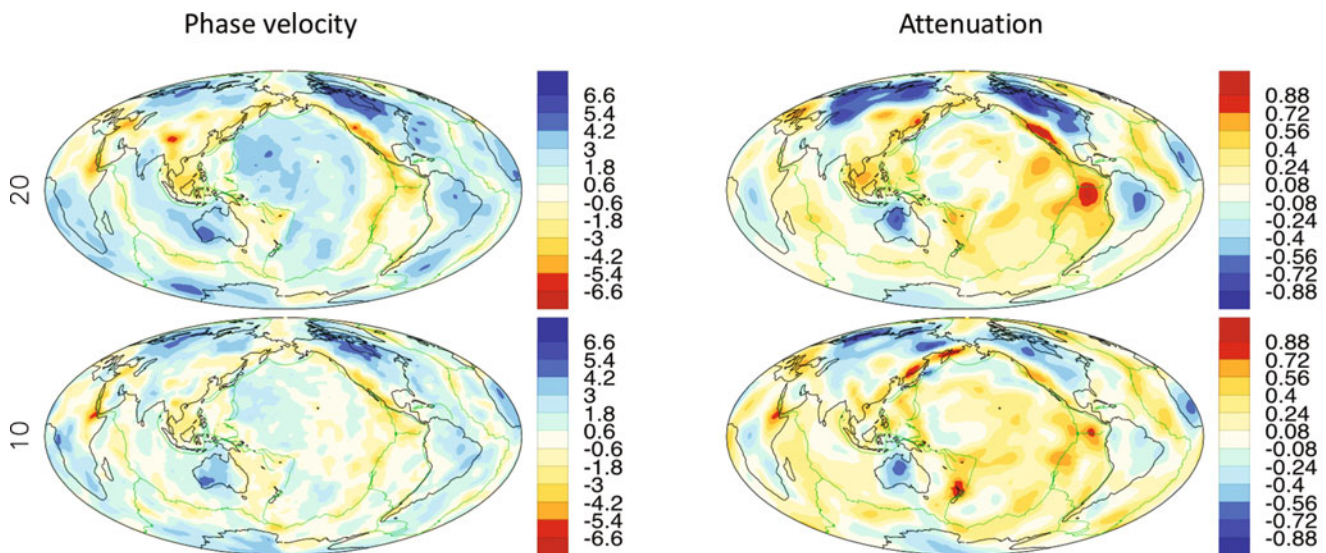
of the relatively long time series needed to measure surface waves became much easier and opened the way, in the 1980s to large-scale and global studies of upper-mantle structure.

Global studies first proceeded according to a regionalization scheme, in which it was assumed that the depth variation of seismic velocities is the same across each tectonic province. These studies confirmed and extended to longer periods (and hence larger depths) early results on the age dependence of structure in the oceans (e.g., Romanowicz 2002; see also ► [“Earth’s Structure, Upper Mantle”](#)). The constraint of regionalization was soon relaxed, replaced by expansion of lateral heterogeneity into a global basis of spherical harmonics. This revealed, in particular, that the correlation of structure with tectonics disappears below depths of 200–300 km (e.g., Ekström et al. 1997). Dispersion measurements have focused primarily on the fundamental mode, generally in the period range 35–150 s, and thus are generally limited in resolution to depths shallower than ~300–400 km. This approach has led to several generations of global phase velocity maps at discrete frequencies, with increasing spatial resolution (e.g., Fig. 5). In these studies, lateral variations of dispersive properties are now well resolved down to wavelengths of ~1,000–2,000 km. At the long period end, the period range is limited by the difficulty of separating consecutive wave trains, and at the short period end, by the increased complexity of surface wave propagation in the strongly heterogeneous crust and uppermost mantle, resulting in lateral refractions and multipathing. The latter are not taken into account in the simple high frequency, great circle propagation assumptions underlying the construction of dispersion maps.

Inversion of Surface Wave Dispersion Data

Inversion of surface wave dispersion data generally involves two steps. By combining measurements at a given frequency for many different paths crossing a particular region, or, at longer periods ($T > \sim 40$ s) over the entire globe (e.g., Fig. 5), dispersion maps can be obtained by a standard linear inversion procedure, yielding point by point dispersion curves over the region considered.

These dispersion curves can then be inverted to obtain the local variation of elastic parameters with depth, by matching observed dispersion curves to theoretical curves calculated for layered earth models, using inverse theory (e.g., Thurber and Ritsema 2015; see also ► [“Inverse Theory, Linear”](#)). The models thus obtained suffer from some degree of nonuniqueness, as the problem is underdetermined. In particular, surface waves cannot by themselves resolve strong gradients of velocity with depth, so that, in general, some a priori constraints are applied, such as fixing the depth of the Moho, or limiting the extent of low-velocity layers. A major advance was the theoretical contribution of Backus and Gilbert (1968), which made it possible to calculate the resolving kernels and errors associated with the various least squares solutions of the inverse problem. Surface wave dispersion is most sensitive to shear velocities, so that, in general, a priori scaling factors between shear velocity and compressional velocity, on the one hand, and density, on the other, are assumed. These scaling factors rely on laboratory measurements of rock properties and are most reliable for crustal and upper mantle depths.



Surface Waves, Fig. 5 Examples of global Rayleigh wave phase velocity maps (left) and attenuation maps (right) at periods of 50 s (20 mHz, top) and 100 s (10 mHz, bottom), highlighting the diminishing lateral variations of structure with period (i.e., with depth) and the

correlation with tectonics in both velocity and attenuation. Phase velocity is shown in relative perturbations (percent) with respect to the global mean at each period. Attenuation is presented in relative perturbations of Q^{-1} . (After Ma et al. (2016). Courtesy of Zhitu Ma)

Crustal Corrections

A significant issue regarding the inversion of long period fundamental mode surface waves (40–200 s) for upper mantle structure is that of crustal corrections. Indeed, surface waves are sensitive to shallow structure even at long periods, but not sensitive enough for inversion, unless more difficult to measure short periods (down to ~ 10 s) are included. Until recently, most studies simply performed crustal corrections in the framework of linear perturbation theory. However, Montagner and Jobert (1988) showed that the effect of the large variations in crustal thickness is nonlinear, and proposed a more accurate correction procedure based on a tectonic regionalization. Developing accurate crustal models worldwide (see ► “Earth’s Structure, Core”) remains a challenge for large-scale surface wave inversions for structure of the upper mantle, especially in the era of numerical computation of waveforms. In this respect, the availability of large dispersion datasets from noise correlations are very helpful in constraining crustal contributions, especially in densely instrumented regions, such as the continental United States (e.g., Ekström 2017).

Inversion Using Overtones

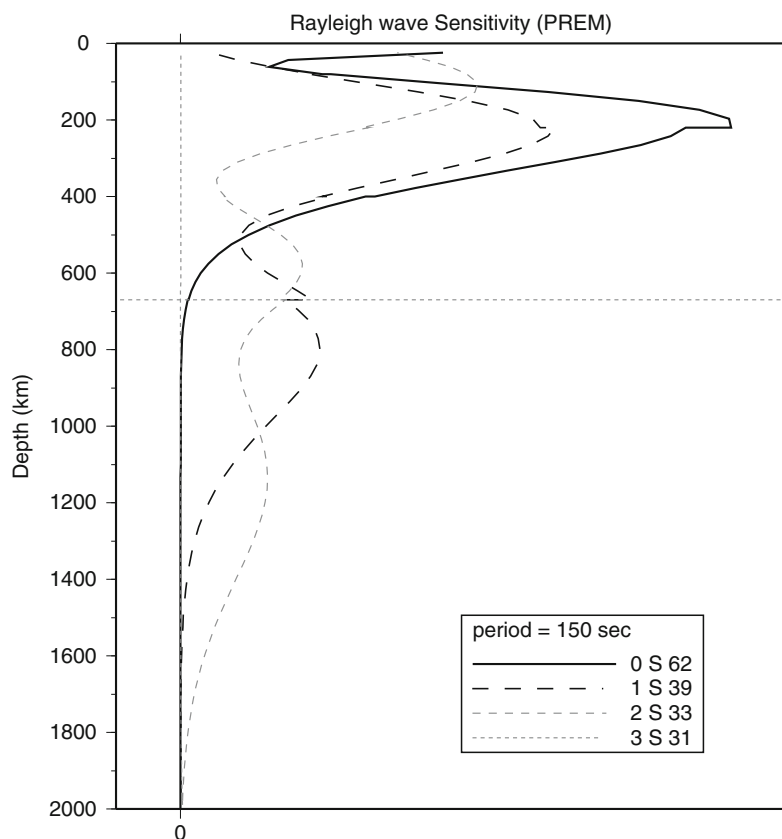
While they are well separated in the time domain from other mode branches, and therefore well suited for frequency domain single mode analysis techniques, fundamental mode

surface waves have several shortcomings: at intermediate periods (~ 20 – 150 s), their sensitivity to structure below about 200 km is poor, whereas longer period mantle waves, which reach down to the top of the upper mantle transition zone, have poor spatial resolution. In any case, resolving structure in the transition zone, which is also poorly sampled by body waves, requires the inclusion of higher-mode surface waves, whose sensitivity is larger at these depths (e.g., Fig. 6). They are also a powerful tool for investigating structures where low velocity zones may be present (e.g., Kovach and Anderson 1964). In general, however, higher-mode surface waves overlap in the time-frequency domain, and single mode dispersion methods therefore cannot be applied.

In the 1970s, array methods were developed independently by Nolet (1975) and Cara (1978) to extract information from overtones, and applied in the period range 20–100 s to paths across Eurasia and the Pacific Ocean, respectively. These methods require a linear regional array of stations approximately aligned with the epicenter (and not in a nodal direction of the source radiation pattern), in order to separate modes in the (ω, k) domain, where k is the wavenumber. They are limited in application to a few regions of the world with relatively dense, linear arrays. A waveform-based method involving the comparison of observed and synthetic seismograms at a single station was proposed by Lerner-Lam and Jordan (1983). In this approach, branch

Surface Waves,

Fig. 6 Sensitivity with depth of Rayleigh wave fundamental mode (solid line) and first two overtones (broken lines) at a period of 100 s, computed in the PREM model (Dziewonski and Anderson 1981). The fundamental mode sensitivity peaks around 200 km whereas the overtones have significant sensitivity down to the top of the lower mantle. (Courtesy of Yuancheng Gung)



cross-correlation functions (bccf's) are formed between a particular single-mode synthetic and the observed seismograms and used to invert for an average structure along each source-station path. This approach lends itself to the derivation of secondary observables such as group or phase velocity. The main drawback of this methodology is the contamination of the single mode objective function by interference from other mode branches. More recently, van Heist and Woodhouse (1997) proposed a “mode-branch stripping” method based on a bccf approach combined with a frequency-stepping procedure, yielding single mode dispersion and amplitudes. A global multimode dispersion dataset assembled in this fashion has been used in the construction of several generations of global shear velocity models (e.g., Ritsema et al. 2011). See also ► [“Earth’s Structure, Upper Mantle”](#).

Time Domain Surface Wave Inversions

Extracting single mode dispersion information directly from spectra of windowed wave trains is only effective for the fundamental mode, which is well separated on the seismogram, and even in this case, overlapping with overtone wave trains or, at long periods, other orbits of the fundamental mode, can be a problem. A waveform matching approach is therefore desirable, where observed seismograms are compared to synthetics computed for a given reference earth model, and the model is perturbed to improve the fits of the observed and synthetics in the target time window of observation, which can be optimized. In particular, such an approach is suitable both for fundamental mode and overtone surface waves.

For many years, the computation of long period synthetic seismograms relied on asymptotic approximations to normal mode first order perturbation theory, as developed in the 1970s. The most widely used approximation is the “path average approximation,” introduced for waveform modeling by Woodhouse and Dziewonski (1984). The key steps that led to the various formalisms routinely used today are described in Romanowicz et al. (2008). See also ► [“Seismic, Waveform Modeling and Tomography”](#).

Working under the assumption of PAVA, Nolet (1990) introduced “the partitioned waveform inversion” (PWI), in which inversion for elastic structure proceeds in two steps, more closely related to the two-step inversion using dispersion data. Path integral parameters are defined and retrieved by nonlinear waveform fitting over each path, and in a second step, inverted linearly for elastic structure. The advantage of this approach is that it reduces the number of parameters to be fit in the nonlinear part of the inversion. The PWI also includes progressive filtering of data starting at the longest periods (see ► [“Seismic Tomography”](#)). It has been applied for the construction of continental scale as well as global scale models of the upper mantle (e.g., Schaeffer and Lebedev 2013).

Because it includes only coupling of modes along single dispersion branches, the PAVA is rigorously only suitable for single-mode wave trains such as the fundamental mode. The corresponding sensitivity kernels to structure are 1D kernels, that is, they only depend on the laterally averaged structure in the vertical plane between the source and the receiver. When applied to overtones and body waves, coupling across mode branches needs to be included. For a review of these methodologies, see Romanowicz et al. (2008) and Thurber and Ritsema (2015).

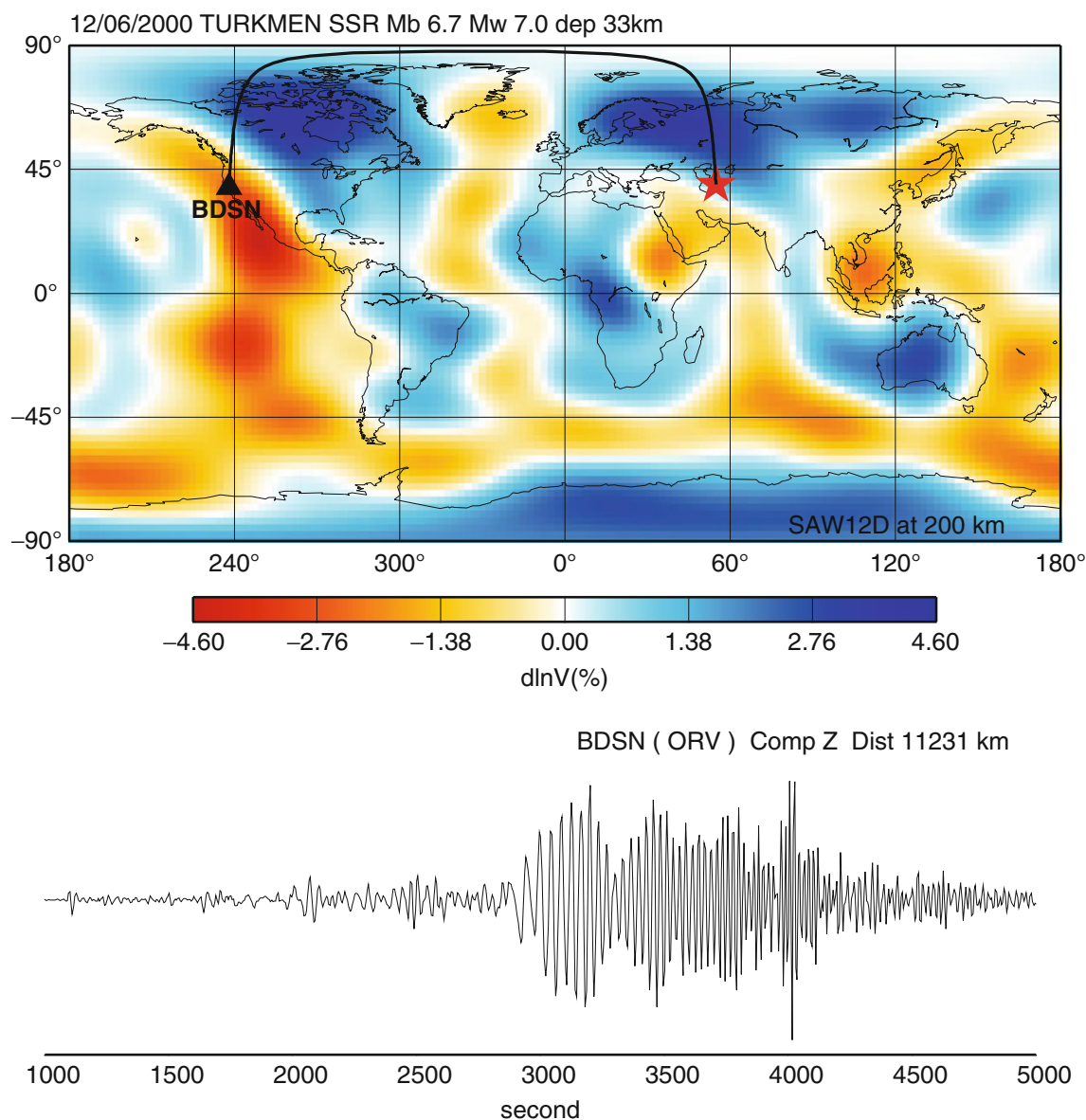
Effects of Scattering and the Born Approximation

Until recently, most regional and global models of upper mantle structure derived from surface waves were based on the standard “path average” approximation (PAVA). This is valid only if the wavelength of lateral variations of structure is long with respect to that of the surface waves considered. And it is not strictly appropriate for overtones.

Many observations indicate that lateral heterogeneity is strong enough to cause departures from PAVA. For example, 20 s surface waves sensitive to shallow crustal structure consistently show multipathing (e.g., Capon 1970 or Fig. 7). At longer periods ($T > 100$ s), later arriving trains often show larger amplitudes than the ones preceding them which cannot be explained by lateral variations of attenuation.

Several approaches have been developed to account for these effects and exploit them to obtain better constraints on lateral variations of structure. Woodhouse (1974) introduced the concept of local modes. These are the surface wave modes of a laterally homogeneous model, which locally has the depth distribution of the laterally varying model. If the medium is laterally smooth, the local mode branches propagate as independent wave trains, and their dispersion is modified according to the evolution of the local modes. However, if lateral variations are sharp (e.g., in the presence of a structural discontinuity such as an ocean-continent boundary), the coupling of the local modes cannot be neglected and its strength depends on the width of the structural transition zone (Kennett 1972). For a review see Romanowicz (2002). Kennett (1984) derived a formalism for mode-coupling in 2D slowly laterally varying structures, which was extended to the 3D scattering case, in Cartesian coordinates (e.g., Bostock and Kennett 1992).

In the framework of scattering theory, the use of tools based on the single-scattering Born approximation has gained popularity in the last 5 years. Born scattering is well suited for inversion since the scattered wavefield depends linearly on structural perturbations (e.g., Snieder 1988). The use of Born kernels is widely advocated to take into account the effects on dispersion of off-great circle sensitivity of surface waves (e.g., Zhou et al. 2006). However, the relative merits of



Surface Waves, Fig. 7 Example of vertical component record at BDSN station ORV for an M_w 7.0 earthquake in Turkmenistan showing multipathing of 20–30 s surface waves. The record has been bandpass filtered between 0.001 and 0.03 Hz. The great circle path is shown on the

top plot with a shear wave tomographic model in the background. It arrives to the station from the north, following the coast and along major structural boundaries in the crust that are likely the cause of the multipathing. (Courtesy of Yuancheng Gung)

single-scattering Born sensitivity kernels versus simpler averaging kernels, such as “fat rays” has been debated (e.g., Boschi et al. 2006).

Born is also a poor approximation in the case of strong heterogeneity, or when the region of scattering is large (e.g., Friederich et al. 1993).

Most recently, the introduction to global seismology of the Spectral Element Method (Komatitsch and Vilotte 1998), which provides an accurate representation of the teleseismic wavefield in arbitrary 3D media up to the second surface wave orbit, has provided a new powerful tool for the study of Earth structure, and is progressively being implemented for

waveform modeling of crust and upper mantle structure at the local (e.g., Tape et al. 2010), regional (e.g., Fichtner et al. 2010), and global scale (e.g., French et al. 2013), using relatively long period surface waves (periods longer than 40 s). These studies show that the use of a more accurate wave propagation theory results in the improved retrieval of amplitudes of lateral variations of shear wave velocity, and, in particular, significantly lower values in low velocity zones. For applications at the global scale, this promising new tool presents several challenges: a considerable computational cost, and the necessity of accurate representation of the 3D structure in the crust. Such an approach has more recently

been extended to whole mantle tomography, adding shorter period body waveform information (French and Romanowicz 2014; Bozdag et al. 2017).

Upper Mantle Anisotropy from Surface Wave Studies

Early studies of surface wave dispersion showed that, in many regions, it is not possible to find a single isotropic model that satisfies both fundamental mode Love and Rayleigh wave dispersion simultaneously. This provided some of the earliest evidence for the presence of anisotropy in the crust and upper mantle (e.g., Anderson 1961). This discrepancy can be explained by introducing a transversely isotropic medium with a vertical symmetry axis (radial anisotropy), down to at least 200 km depth. It is in this context that the still widely used Preliminary Reference Earth Model (PREM) was constructed (Dziewonski and Anderson 1981). Gung et al. (2003) showed that radial anisotropy is also present at depths greater than 200 km beneath the lithospheric roots of stable continents and cratons, and explains discrepancies in lithospheric thickness obtained using surface waves of different polarization, thus reconciling lithospheric thickness obtained from seismology with estimates from other geophysical and xenolith data.

Surface wave dispersion also varies with azimuth of propagation, an indication of “azimuthal anisotropy” (e.g., Forsyth 1975). The global azimuthal variations of Rayleigh and Love wave dispersion at long periods (100–250 s) were first mapped by Tanimoto and Anderson (1985), who showed that the fast direction appears to correlate with flow directions in the mantle. Montagner and Tanimoto (1991) developed the first global model of radial and azimuthal anisotropy in the upper mantle based on fundamental mode surface wave data. While some questions remain about trade-offs between lateral heterogeneity and anisotropy in this type of inversion, Montagner and collaborators have shown that they can explain their datasets with fewer parameters when azimuthal anisotropy is considered than when it is ignored.

To first order, asymptotically, the azimuthal variation of phase velocity (Love or Rayleigh waves) is of the form (Smith and Dahlen 1973):

$$C(\omega, \theta) = A_0 + A_1 \cos(2\theta) + A_2 \sin(2\theta) + A_3 \cos(4\theta) + A_4 \sin(4\theta) \quad (10)$$

where θ is the azimuth of the wavenumber vector defined clockwise from north. The coefficients $A_i(\omega)$ depend linearly on the elastic tensor at every point in the medium. Theoretical expressions and how they can be applied for crust and upper mantle studies are reviewed in Romanowicz (2002) and

Montagner (2015). Because Rayleigh waves are sensitive both to shallow crustal and deeper mantle anisotropy, it is important to use a wide frequency range to resolve the depth dependence of anisotropy using surface waves. Montagner and Nataf (1988) introduced “vectorial tomography.” By combining radial and azimuthal dispersion terms, under the assumption that the material possesses a symmetry axis (orthotropic medium), they showed that the 3D model can be described using seven elastic parameters (plus density): the five parameters describing transverse isotropy, and two angles describing the orientation in space of the axis of symmetry. Combining long period waveforms and SKS splitting measurements, Yuan and Romanowicz (2010) showed that azimuthal anisotropy provides a powerful tool to detect layering in the continental lithosphere. More recently, several studies have provided evidence for strong layering in the oceanic lithosphere and upper mantle from radial and azimuthal anisotropy in tomographic inversions of surface wave dispersion (Burgos et al. 2014; Beghein et al. 2014).

Using different seismological tools, including surface wave overtones (e.g., Trampert and van Heijst 2002), it has also been suggested that azimuthal anisotropy is present in the transition zone. To date, the mapped patterns are not well resolved. A recent study proposes the presence of significant anisotropy in the vicinity of the 410 and 660 km discontinuities (Yuan and Beghein 2013).

Other diagnostic effects of anisotropy in long period surface waves are waveform anomalies caused by Rayleigh-Love coupling, which generates “quasi-Love” waves on vertical components and “quasi-Rayleigh” waves on transverse (e.g., Maupin and Park 2015). See also ► [“Magnetic Anisotropy”](#).

Surface Wave Attenuation

Surface wave measurements provide the primary constraints on attenuation structure in the crust and uppermost mantle. In the absence of perturbing effects due to scattering and focusing, the amplitude spectrum of a single mode wave train can be written as:

$$A(\omega) = A_0(\omega) \exp(-\eta(\omega)X) \quad (11)$$

where X is the epicentral distance in kilometers and $A_0(\omega)$ represents the amplitude at the source. The attenuation coefficient $\eta(\omega)$ is related to the quality factor Q through (e.g., Aki and Richards 1980):

$$\eta(\omega) = \frac{\omega}{2C(\omega)Q(\omega)}$$

where C is the phase velocity.

Regional studies of amplitudes of fundamental mode surface waves in the period range 5–100 s and of Lg waves have long established the presence of large variations of Q correlated with tectonic provinces and in particular with the age of the oceans, and with time elapsed since the latest tectonic activity on continents. For a recent review, see Romanowicz and Mitchell (2015). Recently, attenuation structure in the vicinity of the East Pacific Rise and in the back-arcs and wedges of subduction zones obtained from a combination of surface wave and body wave data have revealed highly attenuating zones underlying volcanic arcs, down to the slab surface. See also ► “[Seismic Viscoelastic Attenuation](#)”.

Lateral variations in attenuation at long periods can be an order of magnitude larger than those in elastic velocity. Progress in constraining global 3D anelastic structure of the upper mantle has been slow because of the inherent difficulty of measuring attenuation in the presence of focusing and scattering effects that can be as large as anelastic ones and depend strongly on the short-wavelength details of the elastic structure. The long wavelength 3D attenuation models obtained thus far indicate that lateral variations in attenuation in the first 200–250 km of the mantle are correlated with tectonics. In the transition zone, the long wavelength pattern shifts to one correlated with hotspots and with the structure at the base of the mantle, with high attenuation overlying the two large low shear velocity provinces under Africa and in the Pacific (e.g., Romanowicz and Mitchell 2015). At the global scale, mapping of fundamental mode Rayleigh wave attenuation, taking into account effects of focusing have confirmed a correlation of attenuation in the earth’s upper mantle (Fig. 5; Ma et al. 2016). Most recently, several new global models of 3D anelastic structure of the upper mantle have been produced, using high quality broadband data and improved techniques to account for the scattering and focusing effects on surface wave amplitudes, but the models are only consistent at the longest wavelengths, due to persistent trade-offs with elastic structure.

Source Studies Using Surface Waves

Fundamental mode surface wave spectra contain information about the source moment tensor, source depth (centroid), source process time, and, under favorable circumstances for very large earthquakes, source directivity. Tsai and Aki (1971) first showed that the amplitude spectrum of Rayleigh waves contained the signature of source depth, in the form of a “hole” in the spectrum in the period range 10–100 s, which appears at a period depending on depth, and which also depends on the source mechanism.

In order to correct for propagation, different approaches need to be taken depending on the size of the earthquake and

the period range considered. At very long periods ($T > 180$ s), and for large earthquakes ($M > 6.5$), propagation effects can be accounted for approximately using a spherically symmetric reference earth model and ellipticity corrections. At shorter periods, and for smaller earthquakes, corrections on individual source-station paths need to be known much more accurately (for a review, see Romanowicz 2002). In the last 15 years, the availability of increasingly accurate global 3D tomographic models of the upper mantle has made it possible to extend the period range to shorter periods (down to ~ 120 s) and to efficiently make use of methodologies, based on time-domain waveform inversion that are no longer restricted to the fundamental mode, thus providing more accurate estimation of the source depth. Such a waveform approach was first introduced by Dziewonski et al. (1981), who combined waveforms of mantle waves at periods greater than 120 s with overtone waveforms at periods greater than 80 s. This forms the basis of a now routine procedure that serves to construct the widely used Harvard centroid moment tensor (CMT) catalog (Now “global CMT catalog”, <https://globalcmt.org/>).

The Earth’s Background Noise Spectrum: Hum and Microseisms

The presence of background peaks in the frequency range 2–7 mHz, corresponding to the fundamental spheroidal free oscillations of the earth, in the absence of large earthquakes, was discovered 20 years ago by Japanese scientists (e.g., Kobayashi and Nishida 1998). It was later determined that the vertical component long-period seismic noise spectrum was made up of Rayleigh waves (Nishida et al. 2002) and that the main source of the “hum” was located in the oceans (Rhie and Romanowicz 2004), resulting from nonlinear interactions between the atmosphere (oceanic storms), the oceans, and the solid earth. The “hum” has also been detected on the horizontal components, corresponding to Love waves.

The largest background seismic noise is found in the period range 2–20 s and corresponds to the so-called microseismic peak. Microseisms were studied extensively in the 1950s and determined to be surface waves generated in the oceans. There are two types of microseisms: “primary” and “secondary.” Secondary microseisms are the strongest. Also called “double-frequency microseisms,” they are generated at frequency f by a nonlinear mechanism involving two opposing ocean waves of frequency $f/2$ (Longuet-Higgins 1950). In the last decade, there has been renewed interest in the study of microseisms, in the context of the use of noise cross-correlations for the study of Earth structure, as seen above. Some success has recently been obtained in the prediction of the low frequency hum by modelling of the propagation of ocean infragravity waves and their interaction with the ocean floor (Ardhuin et al. 2015). See also ► “[Seismic Noise](#)”.

Summary

Surface waves propagate along the surface of the earth with exponentially decreasing energy with depth, with waves of longer period reaching greater depths. They contain most of the elastic energy generated by earthquakes at periods greater than 20 s and are dispersive. There are two types of surface waves, Rayleigh and Love waves, with different polarization properties. The fundamental mode surface waves are well separated from other energy arrivals in the time domain and provide the best constraints to date on continental scale and global scale upper mantle structure. They play a particularly important role in constraining the upper mantle structure in the ocean basins where few seismic stations have been installed. While the dispersion of fundamental mode surface waves is a classical tool for the investigation of upper mantle structure, higher modes, or overtones provide constraints at transition zone depths (400–700 km) and deeper. They require more sophisticated analysis tools because they cannot be distinguished from each other readily on the seismogram. Waveform inversion approaches are increasingly favored to handle both fundamental mode and overtone interpretation, and hold increasing promise as numerical methods for the computation of the seismic wavefield in arbitrary 3D structures are being implemented, and can account for complex scattering and focusing effects accurately. In addition to isotropic shear velocity, surface waves provide constraints on the distribution of polarization and azimuthal anisotropy in the upper mantle, as well as on anelastic attenuation. When path effects have been corrected for, surface waves provide robust constraints on the source depth and radiation pattern, as expressed by its moment tensor.

The earth's continuous background noise consists primarily of surface waves, at least at periods longer than 5 s. Notable is the microseismic noise peak around 6–7 s. Most of this energy is generated in the oceans by nonlinear interactions involving wind-driven ocean waves and the seafloor. Recently, a method based on the cross-correlation of noise records has gained popularity and has been applied successfully to resolve crust and uppermost mantle structure. It is particularly useful in seismically quiet regions.

Cross-References

- [Continental Crustal Structure](#)
- [Earth's Structure, Upper Mantle](#)
- [Earthquake Source Theory](#)
- [Free Oscillations of the Earth](#)
- [Inverse Theory, Linear](#)
- [Lithosphere, Continental](#)
- [Magnetic Anisotropy](#)
- [Propagation of Elastic Waves: Fundamentals](#)

- [Seismic Noise](#)
- [Seismic Tomography](#)
- [Seismic Viscoelastic Attenuation](#)
- [Seismic, Ambient Noise Correlation](#)
- [Seismic, Waveform Modeling and Tomography](#)
- [Seismological Networks](#)

Bibliography

- Aki K, Richards PG (1980) Quantitative Seismology. In: Theory and methods. W.H. Freeman, San Francisco
- Anderson DL (1961) Elastic wave propagation in layered anisotropic media. *J Geophys Res* 66:2953–2963
- Ardhuin F, Gualteri L, Stutzmann E (2015) How ocean waves rock the earth: two mechanisms explain microseisms with periods 3 to 300 s. *Geophys Res Lett* 42:765–772
- Backus G, Gilbert F (1968) The resolving power of gross earth data. *Geophys J R Astron Soc* 16:169–2050
- Backus G, Mulcahy M (1976) Moment tensors and other phenomenological descriptions of seismic sources I- continuous displacements. *Geophys J R Astron Soc* 46:341–371
- Beghein C, Yuan K, Schmerr N, Xing Z (2014) Changes in seismic anisotropy shed light on the nature of the Gutenberg discontinuity. *Science* 343:1237–1240
- Bolt BA (1976) Nuclear explosions and earthquakes. W. H. Freeman, USA
- Boschi L, Becker TW, Soldati G, Dziewonski AM (2006) On the relevance of Born theory in global seismic tomography. *Geophys Res Lett* 33:L06302. <https://doi.org/10.1029/2005GL025063>
- Bostock MG, Kennett BLN (1992) Multiple scattering of surface waves from discrete obstacles. *Geophys J Int* 108:52–70
- Bozdag E, Peter D, Lefebvre M, Komatitsch D, Tromp J, Hill J, Podhorszki N, Pugmire D (2017) Global adjoint tomography: a first generation model. *Geophys J Int* 207:1739–1766
- Burgos G, Montagner J-P, Beucler E, Capdeville Y, Mocquet A, Drilleau M (2014) Oceanic lithosphere-asthenosphere boundary from surface wave dispersion data. *J Geophys Res Solid Earth* 119:1079–1093
- Capon J (1970) Analysis of Rayleigh-wave multipath propagation at LASA. *Bull Seismol Soc Am* 60:1701–1731
- Cara M (1978) Regional variations of higher Rayleigh-mode phase velocities: a spatial filtering method. *Geophys J R Astron Soc* 54:439–460
- Dziewonski AM, Anderson DL (1981) Preliminary reference Earth model. *Phys Earth Planet Inter* 25:297–356
- Dziewonski AM, Bloch S, Landisman M (1969) A new technique for the analysis of transient seismic signals. *Bull Seismol Soc Am* 59:427–444
- Dziewonski AM, Chou AT, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J Geophys Res* 86:2825–2852
- Ekström G (2017) Short-period surface-wave phase velocities across the conterminous United States. *Phys. Earth Planet. Inter* 270:168–175
- Ekström G, Tromp J, Larson EWF (1997) Measurements and global models of surface wave propagation. *J Geophys Res* 102:8137–8157
- Ewing WM, Jardetsky WS, Press F (1957) Elastic waves in layered media. McGraw Hill, New York
- Fichtner A, Kennett B, Igel H, Bunge H-P (2010) Full waveform tomography for radially anisotropic structure: new insights into present and past states of the Australasian upper mantle. *Earth Planet Sci Lett* 290:270–280
- Forsyth DW (1975) The early structural evolution and anisotropy of the oceanic upper mantle. *Geophys J R Astron Soc* 43:103–162

- French S, Romanowicz B (2014) Whole-mantle radially anisotropic shear-velocity structure from spectral-element waveform tomography. *Geophys J Int* 199:1303–1327
- French SW, Lekic V, Romanowicz B (2013) Waveform tomography reveals channelled flow at the base of the oceanic asthenosphere. *Science* 342:227–230
- Friederich W, Wielandt E, Stange S (1993) Multiple forward scattering of surface waves: comparison with an exact solution and Born single-scattering methods. *Geophys J Int* 112:264–275
- Gung YC, Panning M, Romanowicz B (2003) Anisotropy and thickness of the lithosphere. *Nature* 422:707–711
- Haskell B (1964) Radiation pattern of surface waves from point sources in a multi-layered medium. *Bull Seismol Soc Am* 54:377
- Kennett BLN (1972) Seismic waves in laterally heterogeneous media. *Geophys J R Astron Soc* 27:301–325
- Kennett BLN (1984) Guided-wave propagation in laterally varying media – I. Theoretical development. *Geophys J R Astr Soc* 79:235–255
- Knopoff L (1972) Observation and inversion of surface-wave dispersion. *Tectonophysics* 13:497–519
- Kobayashi N, Nishida K (1998) Continuous excitation of planetary free oscillations by atmospheric disturbances. *Nature* 395:357–360
- Komatitsch D, Vilotte J-P (1998) The spectral-element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures. *Bull Seismol Soc Am* 88:368–392
- Kovach RL, Anderson DL (1964) Higher mode surface waves and their bearing on the structure of the Earth's mantle. *Bull Seismol Soc Am* 54:161–182
- Lander AV (1989) Frequency-time analysis. In: Keilis-Borok VI (ed) *Seismic surface waves in a laterally inhomogeneous earth*. Kluwer Academic, Dordrecht, pp 153–163
- Lerner-Lam A, Jordan TH (1983) Earth structure from fundamental and higher-mode waveform analysis. *Geophys J R Astron Soc* 75:759–797
- Longuet-Higgins M (1950) A theory of the origin of microseisms. *Philos Trans R Soc Lond* 243:1–35
- Ma Z, Masters G, Mancinelli N (2016) Two-dimensional global Rayleigh wave attenuation model by accounting for finite-frequency focusing and defocusing effect. *Geophys J Int* 204:631–649
- Maupin V, Park J (2015) Theory and observations – wave propagation in anisotropic media. In: Romanowicz B, Dziewonski AM (eds) *Treatise on geophysics*, vol 1. Elsevier, Oxford, pp 277–305
- Montagner JP (2015) Upper mantle structure: global isotropic and anisotropic elastic tomography. In: Romanowicz B, Dziewonski AM (eds) *Treatise on geophysics*, vol 1. Elsevier, Oxford, pp 613–639
- Montagner JP, Jobert N (1983) Variation with age of the deep structure of the Pacific Ocean inferred from very long-period Rayleigh wave dispersion. *Geophys Res Lett* 10:273–276
- Montagner JP, Jobert N (1988) Vectorial tomography II. Application to the Indian Ocean. *Geophys J Int* 94:309–344
- Montagner JP, Nataf HC (1988) Vectorial tomography – I. Theory. *Geophys J Int* 94:295–307
- Montagner JP, Tanimoto T (1991) Global upper mantle tomography of seismic velocities and anisotropy. *J Geophys Res* 96(20):20337–20351
- Nishida K, Kobayashi N, Fukao Y (2002) Origin of Earth's ground noise from 2 to 20 mHz. *Geophys Res Lett* 29:1413. <https://doi.org/10.1029/2001GL013862>
- Nolet G (1975) Higher-Rayleigh modes in western Europe. *Geophys Res Lett* 2:60–62
- Nolet G (1990) Partitioned waveform inversion and two-dimensional structure under the network of autonomously recording seismographs. *J Geophys Res* 95:8499–8512
- Rhie J, Romanowicz B (2004) Excitation of Earth's free oscillations by atmosphere-ocean-seafloor coupling. *Nature* 431:552–555
- Ritsema J, Deuss A, van Heijst HJ, Woodhouse JH (2011) S40RTS: a degree-40 shear-velocity model for the mantle from new Rayleigh wave dispersion, teleseismic traveltime and normal-mode splitting function measurements. *Geophys J Int* 184:1223–1236
- Romanowicz B (2002) Inversion of surface waves: a review. In: Lee WHK (ed) *Handbook of earthquake and engineering seismology*, part A. IASPEI, Acad. Press, NY, pp 149–174
- Romanowicz B, Mitchell BJ (2015) Q in the earth from crust to core. In: Romanowicz B, Dziewonski AM (eds) *Treatise of geophysics*, vol 1. Elsevier, Oxford, pp 789–827
- Romanowicz B, Panning M, Gung Y, Capdeville Y (2008) On the computation of long period seismograms in a 3D earth using normal mode based approximations. *Geophys J Int* 175:520–536
- Saito M (1967) Excitation of free oscillations and surface waves by a point source in a vertically heterogeneous Earth. *J Geophys Res* 72:3689
- Schaeffer AJ, Lebedev S (2013) Global shear speed structure of the upper mantle and transition zone. *Geophys J Int* 194(1):417–449
- Shapiro NM, Campillo M, Stehly L, Ritzwoller MH (2005) High-resolution surface wave tomography from ambient seismic noise. *Science* 307:1615–1618
- Smith ML, Dahlen FA (1973) The azimuthal dependence of love and rayleigh wave propagation in a slightly anisotropic medium. *J Geophys Res* 78:3321–3333
- Snieder R (1988) Large-scale waveform inversions of surface waves for lateral heterogeneity, I. Theory and numerical examples. *J Geophys Res* 93:12055–12065
- Tanimoto T, Anderson DL (1985) Lateral heterogeneity and azimuthal anisotropy of the upper mantle: Love and Rayleigh waves 100–250s. *J Geophys Res* 90:1842–1858
- Tape C, Liu J, Maggi A, Tromp J (2010) Seismic tomography of the Southern California crust based on the spectral-element and adjoint methods. *Geophys J Int* 180:433–462
- Thurber C, Ritsema J (2015) Seismic tomography and inverse methods. In: Romanowicz B, Dziewonski AM (eds) *Treatise on geophysics*, vol 1. Elsevier, Oxford, pp 307–337
- Toksöz MN, Anderson DL (1966) Phase velocities of long period surface waves and structure of the upper mantle, 1, Great circle Love and Rayleigh wave data. *J Geophys Res* 71:1649–1658
- Trampert J, van Heijst HJ (2002) Global azimuthal anisotropy in the transition zone. *Science* 296:1297–1299
- Tsai YB, Aki K (1971) Amplitude spectra of surface waves from small earthquakes and underground nuclear explosions. *J Geophys Res* 75:5729
- van Heist H, Woodhouse J (1997) Measuring surface-wave overtone phase velocities using a mode-branch stripping technique. *Geophys J Int* 131:209–230
- Woodhouse JH (1974) Surface waves in laterally varying structure. *Geophys J Roy Astron Soc* 37:461–490
- Woodhouse JH (1988) The calculation of eigenfrequencies and eigenfunctions of the free oscillations of the earth and the sun. In: Doornbos DJ (ed) *Seismological algorithms*. Academic, San Diego, pp 321–370
- Woodhouse JH, Dziewonski AM (1984) Mapping the upper mantle: three dimensional modelling of the earth structure by inversion of seismic waveforms. *J Geophys Res* 89:5953–5986
- Yang Y, Forsyth DW (2006) Regional tomographic inversion of the amplitude and phase of Rayleigh waves with 2-D sensitivity kernels. *Geophys J Int* 166:1148–1160
- Yu Y, Park J (1994) Hunting for azimuthal anisotropy beneath the Pacific Ocean region. *J Geophys Res* 99:15399–15421
- Yuan H, Romanowicz B (2010) Lithospheric layering in the north American craton. *Nature* 466:1063–1069
- Yuan K, Beghein C (2013) Seismic anisotropy changes across upper mantle phase transitions, Earth Planet. Sci Lett 374:132–144
- Zhou Y, Nolet G, Dahlen F, Laske G (2006) Global upper-mantle structure from finite-frequency surface-wave tomography. *J Geophys Res* 111:B04304. <https://doi.org/10.1029/2005JB003677>

T Waves

Emile A. Okal
Department of Earth and Planetary Sciences, Northwestern
University, Evanston, IL, USA

Definition and Introduction

T phases are defined as seismic recordings of signals having traveled an extended path as acoustic waves in the water body of the oceans. This is made possible by the “sound fixing and ranging” (SOFAR) channel, a layer of minimum sound velocity acting as a wave guide at average depths of 1000 m (Okal 2007). It allows the efficient propagation of extremely small signals over extremely long distances, in practice limited only by the finite size of the ocean basins. The existence of the SOFAR channel results from the dependence of the velocity of sound in water on temperature, pressure, and salinity. As a result, the detailed structure of the channel (including the value of the minimum velocity) varies both geographically (mainly with latitude) and seasonally, but at first order, the SOFAR channel can be regarded as a quasiuniversal feature of the world’s oceans.

At the shoreline of a continent or island, the acoustic wave is converted into a seismic wave which can be recorded by a seismometer or, occasionally, felt by humans. Efficient propagation inside the waveguide requires short wavelengths, in practice frequencies above 3 Hz, and thus T phases generally attenuate fast once converted, although in exceptional geometries they have been detected up to 1000 km inside stable continental shields. Even though records of T waves were reported as early as 1927, they were not identified as hydroacoustic signals until the late 1940s, when results from antisubmarine warfare during World War II were declassified.

Sources of T Waves

T waves can be generated by a diversity of underwater sources, as long as their energy becomes trapped in the SOFAR channel. Most earthquakes in oceanic margins generate strong T phases, since in the presence of a sloping ocean floor, multiple downslope reflections can lead to efficient penetration of the SOFAR (Johnson et al. 1963). However, their generation by earthquakes in abyssal plains deprived of large-scale bathymetry defies the laws of geometrical optics and remains to this day a major theoretical challenge; it is generally thought that abyssal T phases result from scattering along small random heterogeneities of the ocean floor (Yang and Forsyth 2003). Also, T waves are occasionally generated by nonshallow earthquakes, even the deepest ones, even though their source is de facto distant from the ocean column. This excitation is made possible by travel through the subducting oceanic slabs which, being cold, propagate high frequencies with limited attenuation and can deliver them to the water interface for conversion to acoustic waves (Okal 2001a).

Underwater landslides are also T-wave generators. As compared with earthquakes, such sources, which move generally smaller segments of ocean floor over much larger distances, have a considerably longer duration, a lower-frequency spectrum, and thus result in T waves of longer duration but smaller amplitudes (Okal 2003).

By contrast, explosions at sea, which are intrinsically high-frequency, are particularly efficient sources of T waves; in most instances, they involve a shallow source located inside the SOFAR channel, thus providing optimal feeding of acoustic energy into the channel. In particular, seismic reflection experiments are routinely recorded thousands of kilometers away.

Underwater volcanic eruptions are also regular generators of T waves. Because of the diversity of processes involved

during a volcanic sequence, these signals may feature a wide variety of characteristics. The opening of cracks during the formation of magmatic plumbing systems is essentially in the nature of an earthquake, while magmatophreatic explosions at the contact between a lava conduit and the oceanic column are reminiscent of chemical explosions. The steady output of lava into the ocean takes the form of a prolonged noise of low amplitude which may be compared to underwater slumping (Caplan-Auerbach et al. 2001). Finally, the oscillation of magma inside the plumbing system, known to cause volcanic tremor long recorded seismically on subaerial volcanoes, generates harmonic T-wave signals. Supercritical hydrothermal venting during volcanic events has also been documented as the source of monochromatic hydroacoustic signals (Talandier and Okal 1996).

Among the most remarkable T waves ever reported, signals from the eruption of Macdonald Volcano (29°S; 140°W) on 29 May 1967 led to the eventual discovery of this uncharted underwater seamount, and its interpretation as the active member of the Cook-Austral hotspot chain (Johnson 1970).

The collision of large icebergs, during which their masses can rub against each other, has also been documented as a source of T waves, which can exhibit a harmonic character during episodes of stick-and-slip friction between tabular icebergs (MacAyeal et al. 2008). Hydroacoustic waves are also generated during the disintegration of large icebergs, and thus could be used to monitor and assess any possible change in climatic conditions (Li and Gavrilov 2006).

Finally, marine mammals can talk into the SOFAR channel and correspond across hundreds of kilometers, using signals in the range of a few tens of hertz, but specific to each species.

Human Perception

T waves of sufficient amplitude can be felt by shoreline populations. For example, the underwater nuclear test WIG-WAM on 14 May 1955 off the coast of Southern California was felt in Hawaii and even in Japan, 8000 km away. Large events such as the Alaskan earthquake of 10 July 1958 and the deep Bolivian shock of 9 June 1994 were felt in Hawaii, and T waves from the 2004 Sumatra earthquake were felt in the Maldives and possibly on Zanzibar; they could have provided a warning of the impending tsunami in these otherwise aseismic areas, but their populations were unprepared.

Use in Comprehensive Nuclear-Test Ban Treaty Monitoring

Because hydroacoustic waves can detect small sources (especially intra-oceanic ones) at great distances, they have

been included as one of the four technologies used by the International Monitoring System (IMS) of the Comprehensive Nuclear-Test Ban Treaty. Their spectacular efficiency in propagation allows a full coverage of all the Earth's oceans using only 11 receiver sites, 6 of which are hydrophones deployed within the SOFAR channels, and 5 special seismic stations close to shorelines ("T-phase stations"), instrumented for high-frequency seismic recording in the 50–100 Hz range (Okal 2001b).

Acoustic Thermometry of Ocean Climate

This project consisted of using acoustic sources in the 57 Hz range fired in the vicinity of Heard Island, South Indian Ocean, a location allowing the illumination of all three major oceans (Munk et al. 1994). Detection at extreme ranges of up to 17,000 km allowed precise measurement of acoustic velocities, and it was envisioned to repeat these experiments over several decades, to monitor any evolution of sound velocities, as a proxy to changes in temperature, possibly in the context of global warming. Unfortunately, the project had to be halted in the late 1990s under pressure from the environmental lobby.

Quantification

Like other seismic phases, quantitative measurements of T waves can provide information on source parameters. However, absolute measurements (especially of amplitudes) at seismic receivers are difficult to interpret due to the complexity and site specificity of the conversion processes. In addition, the high-frequency nature of the phase results in amplitude saturation at relatively low magnitudes. In this context, quantification of T waves has emphasized duration of wavetrains over amplitude, and relative measurements under common or comparable receiver conditions. Okal et al. (2003) have introduced the concept of T-phase energy flux, *TPEF*, reminiscent of the evaluation of seismic energy from body waves, and which should scale with seismic moment M_0 for regular seismic sources. Any deviation in the ratio $\Gamma = \text{TPEF}/M_0$ identifies anomalous source properties; in particular, slow earthquakes exhibit deficient Γ , a property with some potential in tsunami warning.

The comparison of duration and amplitude of T wavetrains can efficiently discriminate between earthquakes and explosions at sea, based on the strong differences in the scaling of their source processes (Talandier and Okal 2001), although small earthquakes located inside steep volcanic edifices remain a challenge in this respect.

A Few Geophysical Applications

As most of their path is hydroacoustic, T phases propagate at a velocity close to that of sound in water, 1.5 km/s, that gives them late arrivals on seismic records, hence their name “T” for “third.” Such slow velocities allow a remarkable resolving power for source locations, after accounting for small geographic and seasonal variations in SOFAR velocities, and for the contribution of the converted [seismic] segments of the phase.

In this context, the superior detection capabilities of T phases have been used to improve considerably our coverage of low-magnitude seismicity in the oceanic environment, and especially our understanding of small-scale processes at mid-oceanic ridges, notably in conjunction with swarms of volcanic activity (Dziak et al. 1995).

The slowness of T waves also allows to very precisely beam a receiving array (e.g., triangles of hydrophones at sites of the International Monitoring System [IMS]) to back-track the source of T phases, or even of individual fragments of their wavetrains; this procedure was used to obtain a detailed source tomography of the great 2004 Sumatra-Andaman event (e.g., Tolstoy and Bohnenstiehl 2005).

Conclusion

Because of their high-frequency character, T waves provide a valuable complement to more traditional phases in the field of body-wave seismology. Their incorporation into the IMS allows their more systematic recording and opens up a number of new research opportunities.

Cross-References

- [Seismic Monitoring of Nuclear Explosions](#)
- [Seismology, Monitoring of CTBT](#)

Bibliography

- Caplan-Auerbach J, Fox CG, Duennebie FK (2001) Hydroacoustic detection of submarine landslides on Kilauea volcano. *Geophys Res Lett* 28:1811–1813
- Dziak RP, Fox CG, Schreiner AE (1995) The June–July 1993 seismo-acoustic event at CoAxial segment, Juan de Fuca Ridge: evidence for a lateral dike injection. *Geophys Res Lett* 22:135–138
- Johnson RH (1970) Active submarine volcanism in the Austral Islands. *Science* 167:977–979
- Johnson RH, Northrop J, Eppley R (1963) Sources of Pacific T phases. *J Geophys Res* 68:4251–4260
- Li B, Gavrilov AN (2006) Hydroacoustic observation of Antarctic ice disintegration events in the Indian Ocean. In: *Proceedings of first Australasian Acoustical Societies conference*, Christchurch, pp 479–484

- MacAyeal DR, Okal EA, Aster RC, Bassis JN (2008) Seismic and hydroacoustic tremor generated by colliding icebergs. *J Geophys Res* 113(F03):F03011
- Munk WH, Spindel RC, Baggeroer A, Birdsall TG (1994) The Heard Island feasibility test. *J Acoust Soc Am* 96:2330–2342
- Okal EA (2001a) “Detached” deep earthquakes: are they really? *Phys Earth Planet Inter* 127:109–143
- Okal EA (2001b) T-phase stations for the international monitoring system of the Comprehensive Nuclear-Test Ban Treaty: a global perspective. *Seismol Res Lett* 72:186–196
- Okal EA (2003) T waves from the 1998 Papua New Guinea earthquake and its aftershocks: timing the tsunamigenic slump. *Pure Appl Geophys* 160:1843–1863
- Okal EA (2007) The generation of T waves by earthquakes. *Adv Geophys* 49:1–65
- Okal EA, Alasset P-J, Hyvernaud O, Schindelé F (2003) The deficient T waves of tsunami earthquakes. *Geophys J Int* 152:416–432
- Talandier J, Okal EA (1996) Monochromatic T waves from underwater volcanoes in the Pacific Ocean: ringing witnesses to geyser processes? *Bull Seismol Soc Am* 86:1529–1544
- Talandier J, Okal EA (2001) Identification criteria for sources of T waves recorded in French Polynesia. *Pure Appl Geophys* 158:567–603
- Tolstoy M, Bohnenstiehl DR (2005) Hydroacoustic constraints on the rupture duration, length, and speed of the great Sumatra–Andaman earthquake. *Seismol Res Lett* 76:419–425
- Yang Y, Forsyth DW (2003) Improving epicentral and magnitude estimation of earthquakes from T phases by considering the excitation function. *Bull Seismol Soc Am* 93:2106–2122

Tectonic Geomorphology

Sampat Kumar Tandon and Vinee Srivastava
Department of Earth and Environmental Sciences, IISER
Bhopal, Bhopal, India

Definition

“Tectonic Geomorphology” involves two major disciplines of the Earth sciences, tectonics and geomorphology. Tectonics defines the boundary conditions for topographic development as deformation progresses, and prolonged climate-dependent Earth surface processes interact with the tectonic processes to generate modern landscapes (Burbank and Anderson 2011). Tectonic geomorphology relies on data gathered from numerous Earth science disciplines ranging from stratigraphy, structural geology, tectonics, seismology, Quaternary climate change, geochronology, geodesy, remote sensing, and geomorphology.

Tectonic geomorphology in conjunction with seismotectonic and palaeoseismological data helps understand the geological and geophysical aspects of hazard-prone terrains. It helps assess the seismic hazard and seismic risk by identifying active fault zones, and their possible impact on the society and infrastructure in the region. Together with other Earth science disciplines and tools, a

better understanding of the seismic terrain can be obtained, and eventually, the integrated outcomes can be used for disaster management in these tectonically active regions.

Tectonically Generated First-Order (10^3 – 10^2 Km Scale) Landscapes

Geodynamic processes along tectonic boundaries (~1000 km), where horizontal and vertical motion(s) of tectonic plates result in strike-slip, extension, and compression, govern the dynamics of large scale characteristic landscapes as discussed below.

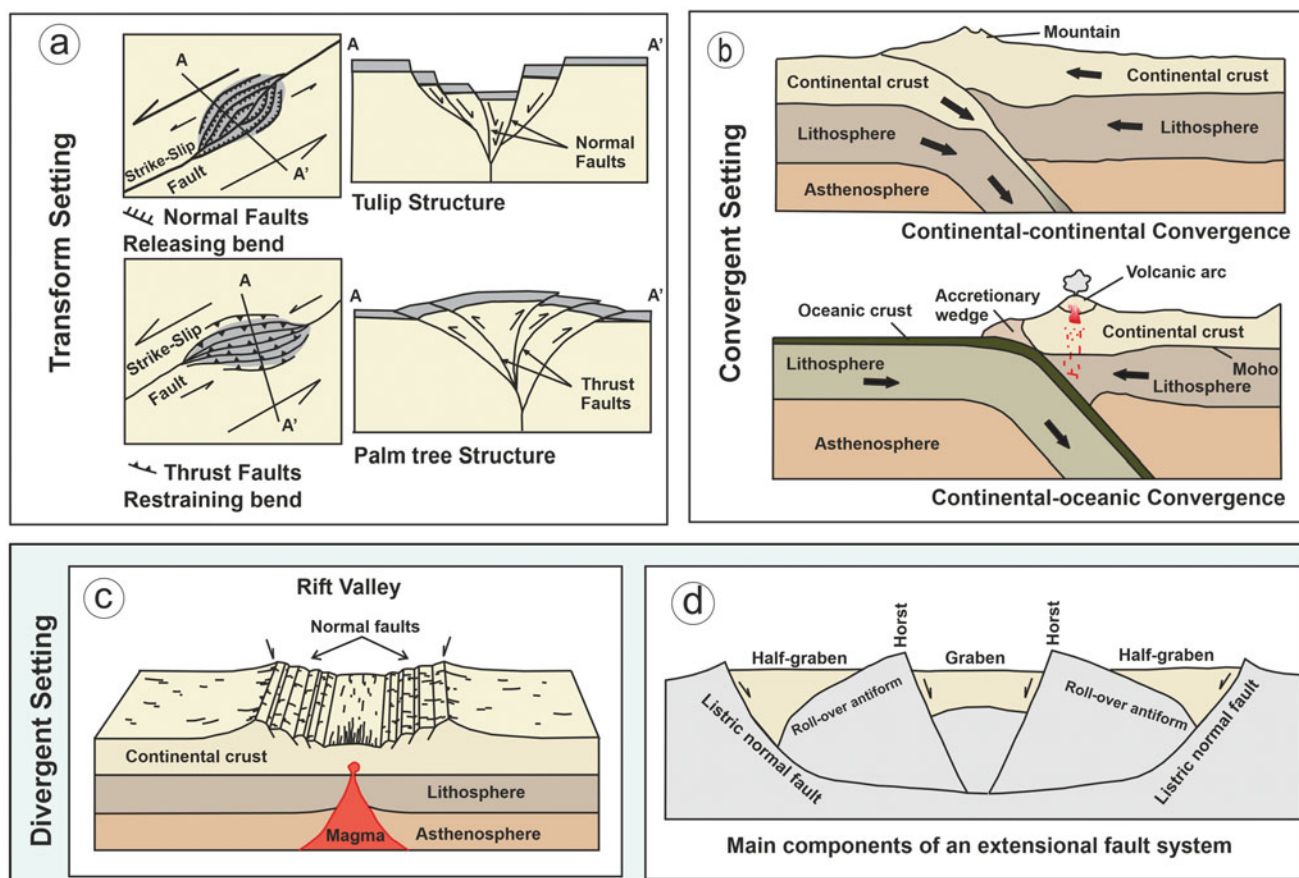
Transform Settings

In transform settings, two plates slide past each other horizontally along strike-slip faults defining a conservative boundary where the lithosphere is neither destroyed nor created. This setting is dominated by fault-parallel simple shear in the horizontal plane. However, a deviation from regional

simple shear occurs near stepovers or bends in the major strike-slip fault (Fig. 1a). Depending on the sense of shear and the direction of bend or step-over along the major fault, zones of transpression or local contraction and transtension or local extension develop in these regions resulting in extensional or contractional structures (Fossen 2010; Cunningham and Mann 2012). Restraining or transpressional bends are associated with pressure ridges, palm tree or positive flower structure, and associated landforms. Releasing or transtensional bends form pull-apart basins, tulip structure or negative flower structure, and associated landforms (Fossen 2010; Cunningham and Mann 2012; Bierman and Montgomery 2013).

Convergent Settings

In convergent settings, two plates collide with each other defining a destructive plate boundary where the lithosphere is subducted and destroyed. The plate collisions can be ocean-ocean, ocean-continent, or continent-continent and are characterized by thrust faulting and distinctive landscape features



Tectonic Geomorphology, Fig. 1 Tectonic landforms associated with tectonic boundaries. (a) “Negative flower or Tulip structure”, “Positive flower structure or Palm Tree structure” in transform setting (modified after Bierman and Montgomery 2013); (b) mountain belts associated with continent-continent collisional setting and trench,

volcanic arc with ocean-continent collision in convergent/compressional setting dominated by thrust faults; (c) rift valley; (d) horst and graben, half graben landforms associated with divergent/extensional setting characterized by normal faults

associated with each such collisional setting (Bierman and Montgomery 2013). In ocean-ocean and ocean-continent, high density oceanic crust subducts beneath less dense oceanic or continental crust respectively resulting in a trench at the boundary. Volcanic island arcs are characteristic landforms associated with ocean-ocean collision, e.g., Japan and many South Pacific islands. Continental volcanic arcs with deformed accretionary wedges are associated with ocean-continent collision such as the Cascade Range in western North America, and Andes Mountains at the western edge of South America. The most studied continent-continent collision example is the Himalaya where collision of Indian-Eurasian plates created a ~2500 km long, wedge-shaped, fold and thrust belt related mountain range (Fig. 1b; Fossen 2010).

Divergent Settings

In divergent settings, two plates diverge away from each other and define a constructive plate boundary where new oceanic lithosphere is created. This setting is characterized by an array of normal faults and consists of distinct landscape features such as rift valleys, horst and graben, and half graben (Fig. 1c, d). Rifting in oceanic lithosphere creates a spreading center or mid-oceanic ridges where new lithosphere is created, e.g., mid-Atlantic ridge separating the South American and African plates (Bierman and Montgomery 2013). Similarly, rifting of continental plates forms a horst and graben structure, e.g., Basin and Range Province, western USA and southwestern Mexico (Fossen 2010).

Signatures of Morphotectonic Landscapes

A landscape can form by tectonic processes ranging from first-order plate scale (10^3 to 10^2 km) features described above to regional scale (10^2 to <1 km) features. New landscapes can be created by deformation, and a preexisting landscape or feature can be modified by deformation leaving behind evidences of recent deformational events. Such evidences can be further subdivided into the following two categories.

Direct Evidences

Direct evidences comprise geomorphic signatures that can be observed, identified, and mapped in the field in a specific tectonic setting. Some characteristic signatures are discussed below.

Geomorphic Markers of Transform Setting

Transform settings characterized by strike-slip faults generate distinctive landforms such as offset streams; these can also be detached from their source at places, creating beheaded streams. Releasing bends and stepovers along a major

strike-slip fault or weakening of rock in a fault or damage zone creates characteristic features such as sag ponds. This can be the result of captured drainage courses or small depressions along the trace of the strike-slip fault due to transtension-related normal faulting (Fig. 2a; Bierman and Montgomery 2013). Similarly, restraining bends or stepovers along a major strike-slip fault can form a topographic ridge known as pressure ridge due to transpression-related thrust faulting.

Geomorphic Markers of Divergent Setting

Normal faults in extensional settings produce distinctive geomorphic features such as offset mountain fronts with bedrock, upland and lowland alluvial plains, and well-marked range fronts. The fault surface is also exhumed as triangular bedrock facets, and hourglass-shaped valleys are formed extending into the range front between the facets (Fig. 2b; Bierman and Montgomery 2013).

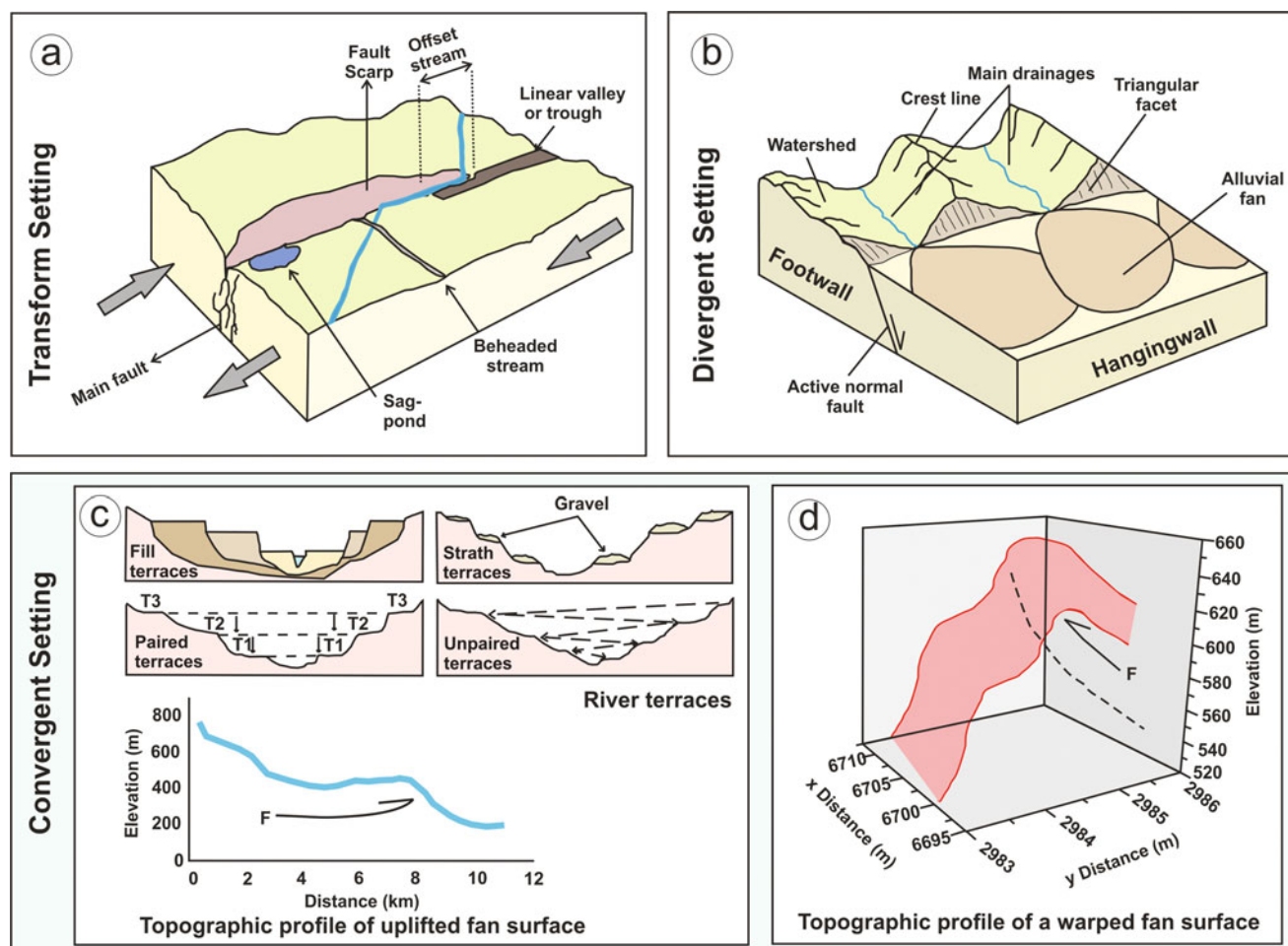
Geomorphic Markers of Convergent Setting

Convergent settings constitute a compressional regime characterized by thrust or reverse faults. Thrust faults are also accompanied by folding via fault-bend folding or fault-propagation folding and eventually generate mountains and valleys within fold and thrust belts (Keller and Pinter 2002). These faults and folds are responsible for generating new landforms or uplifting preexisting ones (unpaired or/and tilted terraces, warped topography) (Fig. 2c, d). Reverse faulting can generate steep mountain fronts with fault and fold scarps.

Geomorphic Markers in Multiple Settings

Some geomorphic landforms and features are found in multiple settings:

- (a) **Unpaired strath terraces:** River terraces are geomorphic features which can be either aggradational or degradational. Fill or aggradational terraces are commonly paired and formed due to aggradation of the river-transported sediments along the river course (Keller and Pinter 2002). Strath or degradational paired terraces are formed due to deep incision by a river. However, deformation or uplift of the preexisting terrace due to thrusting or reverse faulting as well as subsidence due to normal faulting can cause recurrent changes in the course of rivers and form unpaired strath terraces (Fig. 2c; Keller and Pinter 2002).
- (b) **Uplifted/tilted terraces and fan surfaces:** A fold and thrust belt e.g. the Himalaya is marked by geomorphic features such as river terraces and alluvial fan surfaces. These geomorphic features can be deformed because of sequential thrusting or reverse faulting which leads to deviation from their natural slope direction. The terraces



Tectonic Geomorphology, Fig. 2 Geomorphic features associated with tectonic boundaries. (a) Transform setting. (Modified after Bierman and Montgomery 2013) (b) Divergent setting. (Modified after Strak et al.

2011). (c, d) Convergent setting. (Modified after Burbank and Anderson 2011; Srivastava et al. 2017). “F” represents a blind fault

and the fan surfaces uplifted by these faults will slope in the regional dip direction of the faults opposite to their natural slope direction (Fig. 2c; Srivastava et al. 2017). Similarly, in an extensional setting, terraces can be tilted due to normal faulting. Terraces or fan surfaces within the transpressional or transtensional zones of strike-slip regime will be either uplifted or tilted.

- (c) **Warped terraces and fan surfaces:** Terraces or fan surfaces can be warped or folded due to activity along a blind thrust or reverse fault, and fault-related folding such as fault propagation or fault-bend folding. In fault propagation folds where propagation velocity of a fault is less than the slip velocity of overlying strata, slip accumulates at the fault tip and generates topography by warping the surface (Fig. 2d; Srivastava et al. 2017). Warping of the surface can also be formed as roll-over anticline where folding is associated with downward movement along a normal fault. Similarly, deformation in transpressional or transtensional zones in strike-slip setting can also warp terraces or fan surfaces.

Indirect Evidences

Geomorphic parameters/indices discussed below are used to assess subtle tectonic activity in all tectonic settings, and formulae used to quantitatively describe them are given in Table 1. These are best used together with direct evidences but can also be used without direct evidence of tectonic activity.

Geomorphic Indices

- (a) **Hypsometric curve and Hypsometric Integral (HI):**

Hypsometric curve describes the elevation distribution over a particular area for a drainage basin. Relative area (a/A) against relative height (h/H) of a portion of a basin over the total basin area is used to construct the curve (Keller and Pinter 2002). The shape of the curve obtained points to the geomorphic processes operating in a basin. Convex curve indicates the dominance of tectonic activity in the basin, convex-concave shape shows equilibrium between tectonic activity and erosional processes, and concave curve shows dominance of fluvial erosional

Tectonic Geomorphology, Table 1 Geomorphic indices used to assess subtle tectonic activity in a basin (compiled from Keller and Pinter 2002 and Schumm 1956)

S. no.	Geomorphic indices	Values	Interpretations
(a)	Hypsometric integral HI = $H_{mean} - H_{min} / H_{max} - H_{min}$ Hmean = mean elevation Hmin = minimum elevation Hmax = maximum elevation Normalized	HI = 0 – Mature HI = 1 – Youthful	Old basin Tectonically active
(b)	Mountain front sinuosity $Smf = Lmf/Ls$ Lmf = the length of the mountain front Ls = straight line length of the mountain front Normalized	Smf = 1.0–1.5 Smf = 1.5–3.0 Smf > 3	Highly active front, Moderately active front Inactive front
(c)	Asymmetry factor/Drainage basin Asymmetry AF = $100 (A_r/A_t)$ Ar = area of the basin to the right of the trunk stream while facing downstream At = total area of the drainage basin Normalized	AF < 50 AF = 50 AF > 50	Basin tilt to the right Symmetric basin Basin tilt to the left
(d)	Basin elongation ratio (Re) $Re = 2 \sqrt{AT/\pi} / L$ AT = diameter of the circle having the same area as the basin L = maximum basin length Normalized	Circular (Re > 0.9) Oval (0.8 < Re < 0.9) Less elongated (0.7 < Re < 0.8) Elongated (Re < 0.7)	Re > 0.75 = Tectonically inactive Re = 0.5–0.75 = Slightly active Re = <0.5 = Tectonically active
(e)	Stream length gradient index $SL = (\Delta H / \Delta L) L$ ΔH = difference in elevation, ΔL = the length of the stream reach where the index is to be computed. L is the channel length from the drainage divide to the center of the reach. Non-normalized	High SL-index, low-resistance rocks High SL-index, high-resistance rocks Low SL-index, low-resistance rocks	Tectonics Lithological
(f)	Valley floor width-to-height ratio $Vf = 2Vfw / [(Eld - Esc) + (Erd - Esc)]$ Erd = elevation of the right valley divide Eld = elevation of the left valley divide Esc = the valley floor elevation Vfw = the valley floor width Normalized	Vf < 1 – youthful (V-shaped) Vf > 1 – mature (U-Shaped)	Tectonically active

processes in a mature basin. Hypsometric integral (HI) indicates a basin's relief and is used to identify tectonically active regions together with the hypsometric curve (Table 1; Keller and Pinter 2002).

- (b) **Mountain Front Sinuosity (Smf):** Mountain front sinuosity (Smf) distinguishes between straight and sinuous mountain front using the interplay of tectonic and/or erosional processes in mainly convergent/divergent settings (Table 1; Keller and Pinter 2002).
- (c) **Asymmetry Factor (AF):** Asymmetry factor (AF) assesses the symmetry of a basin. Symmetry of a basin and its drainage network depends on a stable setting and uniform lithology. In tectonically active and differentially tilted regions, the AF is used to identify tilted asymmetric basins (Table 1; Keller and Pinter 2002).
- (d) **Basin Elongation Ratio (Re):** Basin elongation ratio quantifies the variation of the shape of a basin from

circular, oval to elongated and its relationship with the tectonic or erosional processes (Table 1; Keller and Pinter 2002).

- (e) **Stream Length Gradient Index (SL):** The stream length gradient index (SL) is used to assess the uplift in a tectonically affected region mainly in convergent settings. The stream length gradient index (SL) determines the relationship between stream power, tectonics, and rock resistance (Keller and Pinter 2002). Anomalously high or low SL values are the result of changes in stream power, slope induced by tectonic deformation or rock resistance due to lithological variation, and results in the occurrence of knick-points along river profiles (Table 1; Keller and Pinter 2002).
- (f) **Valley Floor Width-to-Height Ratio (Vf):** The U- or V-shape of a valley is quantified using valley floor width-to-height ratio. The V-shaped valleys indicate the

Tectonic Geomorphology, Table 2 Global case studies

Divergent setting	
Topal et al. (2016)	Tectonic geomorphology of a large normal Akşehir fault, SW Turkey, indicates presence of triangular facets along the mountain fronts. Subtle tectonic activity has been assessed along the Akşehir fault using uplift patterns computed by vertical slip rate on the fault scarps based on the facet slope in the footwall along the various segments of Akşehir fault which are consistent with the patterns of geomorphic indices along each segment
Copley et al. (2014)	Active faulting within peninsular India represents a significant seismic hazard in central India due to reactivation of Mesozoic Tapi normal Fault as thrust fault as observed through offset alluvial fan surfaces. This indicates one or more magnitude 7.6–8.4 thrust-faulting earthquakes on the Tapi Fault, Maharashtra, India, during the Holocene. Modified drainage pattern in central India and formation of the large west-flowing catchment of the Tapi river are due to downward movement of the footwall of the Tapi Fault
Convergent setting	
Lavé and Avouac (2001)	Geomorphic analysis, active fault bend folding, and upliftment of strath terrace along the Bagmati and Bakeya Rivers indicate that the MFT is the only active structure south of the Lesser Himalaya in central Nepal
Talebian et al. (2016)	Geomorphic evidences, such as warped and incised Quaternary alluvium, and fault scarps were associated with tectonically active blind Pardisan thrust in the central part of Tehran. Rate of uplift was measured using optically stimulated luminescence (OSL) dating on a rare exposed remnant of uplifted alluvial deposits in the hanging wall of Pardisan thrust, Tehran
Srivastava et al. (2016, 2018)	Emergent Main Frontal thrust's (MFT) half-a-kilometer fault zone with well-developed fault core and damage zone together with uplifted gravel cap and MFT induced Mohand Range topography indicated direct evidence of neotectonic deformation along the MFT in the Dehradun recess, NW Himalaya, India
Srivastava et al. (2017)	Geomorphic indices along with their uncertainties computed from real-time kinematic global navigation satellite systems (RTKGNSS), corrected SRTM data together with direct field evidence such as tilted terraces, and warped fan surfaces indicate neotectonic deformation along the Ramgarh thrust in the frontal Gorubathan recess, Darjiling Himalaya, India
Divyadarshini and Singh (2017) and Divyadarshini et al. (2020)	Geomorphic features such as terraces and fan surfaces that were incised, tilted, folded, or uplifted together with stream length index (SL) indicated reactivation/out-of-sequence thrusting signatures in the hanging wall of the MFT. New faults were also identified based on field observations, Hack profile, and SL-indices in Chitwan dun, central Nepal. OSL dating derived ages of the landforms support distinct climate-tectonic coupling events responsible for landscape evolution in Chitwan dun
García-Delgado and Velandia (2020)	Structural and geomorphic indices-based study indicates greater tectonic activity and uplift rates for the northern San Lucas Range, in the Northern Andes
Transform setting	
Arrowsmith and Zielke (2009)	Field evidences, paleoseismic analysis, and high-resolution LiDAR DEM were used to trace the active Cholame segment of San Andreas Fault (SAF) to understand tectonic geomorphology of SAF
Softa et al. (2018)	Transpressional deformation zone, the Eastern Pontides (EP), is the active deformation zone and has a “push-up” geometry in conjunction with the North Anatolian Fault high. The EP exhibits high tectonic activity based on geomorphic indices and more than 0.5 mm/yr upliftment along the thrust faults of the Black Sea Fault (BSF) and Borjomi-Kazbegi Fault (BKF) in the active mountain belt

presence of areas affected by tectonic uplift/subsidence in a given setting. The U-shaped valleys indicate the attainment of erosion at the base level (Table 1; Keller and Pinter 2002).

Tectonic Geomorphology and Paleoseismological Studies of Tectonic-Landforms

Tectonic geomorphology studies are often integrated with palaeoseismological investigations as geomorphic features, studied in the former, are the result of interaction of multiple paleo-seismic events and prolonged Earth surface processes. Paleoseismology is the study of prehistorical earthquakes in terms of location, size, and time; and its aim is to define the earthquake potential of a fault zone and constrain it

chronologically (McCalpin 2009). Evidences of past earthquakes are studied in sediments and rocks in tectonically active regions. Trenching technique is used wherein a trench is dug along an active fault, and geological logs exhibiting characteristics of the rock layers are prepared (cf. Fig. 2a.1, McCalpin 2009). Active faulting traces recognized as offsets of sediment layers due to earthquakes can be observed from the trench walls (McCalpin 2009). If there are traces of multiple faults or fault activity, they can be dated in absolute terms if dateable carbon or human artifacts occur in order to determine the relative age of movements on each fault. Ideal location for trenching along an active fault is near streams, slopes, and ponds which are the loci of sedimentation. Luminescence dating technique includes

thermoluminescence (TL), and optically stimulated luminescence (OSL) methods (Jayagondaperumal et al. 2018 and references therein). It is one of the important methods for dating sediments in paleoseismic research in addition to methods based on radiocarbon and cosmogenic radionuclides. These dating methods are used to determine the timing of fault displacement, calculate slip rates, and estimate earthquake recurrence intervals.

Global Case Studies

Research in Tectonic Geomorphology has evolved rapidly in order to study the interrelationship between tectonics and climate/erosion in different tectonic settings across the globe. An interdisciplinary approach to assessing crustal movements, mechanisms and rates of Earth surface processes, and advances in dating techniques to constrain the ages of these events that form a landscape has furthered exploration in the field of tectonic geomorphology as summarized in Table 2.

Summary

Tectonic geomorphology is evolving fast as new quantitative techniques are being developed that provide the basis to understanding climate/erosion and tectonic interactions in divergent, convergent, and transform settings. It allows the examination of how the tectonic and Earth surface processes interact over a period of time with each other to generate the present-day landscape. Also, the tectonically generated landscapes are useful for making an assessment of the neotectonic/active deformation in orogens like the Himalaya and their densely populated bordering plains. “Tectonic Geomorphology” provides a direct window to assess the hazard in tectonically active regions. Also, with other relevant disciplines such as earthquake geology and seismotectonics, it contributes to seismic hazard risk assessment by identifying active faults and their probable impacts on the population and the infrastructures in the region.

Acknowledgments Authors acknowledge with gratitude the constructive criticism and suggestions by an anonymous reviewer that helped significantly improve the quality of the manuscript. We also thank Tanishka Soni for her help in drafting the figures.

Bibliography

Arrowsmith JR, Zielke O (2009) Tectonic geomorphology of the San Andreas Fault zone from high resolution topography: an example from the Cholame segment. *Geomorphology* 113(1–2):70–81. <https://doi.org/10.1016/j.geomorph.2009.01.002>

Bierman PR, Montgomery DR (2013) Key concepts in geomorphology. Macmillan Learning, p 500. ISBN-13: 978-1-4292-3860-1

Burbank DW, Anderson RS (2011) Tectonic geomorphology. Wiley, p 468. ISBN: 978-1-444-33886-7

Copley A, Mitra S, Sloan RA, Gaonkar S, Reynolds K (2014) Active faulting in apparently stable peninsular India: rift inversion and a Holocene-age great earthquake on the Tapti Fault. *J Geophys Res Solid Earth* 119:6650–6666. <https://doi.org/10.1002/2014JB011294>

Cunningham DW, Mann P (2012) Tectonics of strike-slip restraining releasing bends in continental oceanic settings. *Geol Soc Lond, Spec Publ* 290:1–12. <https://doi.org/10.1144/SP290.1>

Divyadarshini A, Singh V (2017) Identifying active structures in the Chitwan Dun, Central Nepal, using longitudinal river profiles and SL index analysis. *Quat Int* 462:176–193. <https://doi.org/10.1016/j.quaint.2017.07.008>

Divyadarshini A, Singh V, Jaiswal MK, Rawat M (2020) Exploring the roles of climate and tectonics in the geomorphic evolution of the Chitwan Intermontane valley, Central Himalaya. *Geomorphology* (in press)

Fossen H (2010) Structural geology, vol 480. Cambridge University Press, Cambridge. ISBN: 9780511777806

García-Delgado H, Velandia F (2020) Tectonic geomorphology of the Serranía de San Lucas (Central Cordillera): regional implications for active tectonics and drainage rearrangement in the Northern Andes. *Geomorphology* 349:106914. <https://doi.org/10.1016/j.geomorph.2019.106914>

Jayagondaperumal R, Thakur VC, Jeevivek V, Rao Singh P, Gupta AK (2018) Active tectonics of Kumaun Garhwal Himalaya. Springer. ISBN 978-981-10-8243-6

Keller EA, Pinter N (2002) Active tectonics: earthquakes, uplift and landscapes, 2nd edn. Prentice Hall, p 362. ISBN-10: 0130882305

Lavé J, Avouac JP (2001) Fluvial incision and tectonic uplift across the Himalayas of central Nepal. *Journal of Geophysical Research: Solid Earth* 106:26561–26591. <https://doi.org/10.1029/2001jb000359>

McCalpin JP (2009) Paleoseismology, 2nd Edition: International Geophysics Series, Academic Press-Elsevier 629 p

Schumm SA (1956) Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. *Geol. Soc. Am. Bull.* 67(5):597–646

Softa M, Emre T, Sözbilir H, Spencer JQG, Turan M (2018) Geomorphic evidence for active tectonic deformation in the coastal part of Eastern Black Sea, Eastern Pontides, Turkey. *Geodin Acta* 30:249–264. <https://doi.org/10.1080/09853111.2018.1494776>

Srivastava V, Mukul M, Barnes JB (2016) Main Frontal Thrust deformation and topographic growth of the Mohand Range, northwest Himalaya. *J Struct Geol* 93:131–148. <https://doi.org/10.1016/j.jsg.2016.10.009>

Srivastava V, Mukul M, Mukul M (2017) Quaternary deformation in the Gorubathan recess: insights on the structural and landscape evolution in the frontal Darjiling Himalaya. *Quat Int* 462:138–161. <https://doi.org/10.1016/j.quaint.2017.05.004>

Srivastava V, Mukul M, Barnes JB, Mukul M (2018) Geometry and Kinematics of Main Frontal thrust related fault propagation folding in the Mohand Range, northwest Himalaya. *J Struct Geol* 115:1–18. <https://doi.org/10.1016/j.jsg.2018.06.022>

Strak V, Dominguez S, Petit C, Meyer B, Loget N (2011) Interaction between normal fault slip and erosion on relief evolution: insights from experimental modelling. *Tectonophysics* 513:1–19. <https://doi.org/10.1016/j.tecto.2011.10.005>

Talebian M, Copley AC, Fattahi M, Ghorashi M, Jackson JA, Nazari H, Sloan RA, Walker RT (2016) Active faulting within a megacity: the geometry and slip rate of the Pardisan thrust in Central Tehran, Iran. *Geophys J Int* 207:1688–1699. <https://doi.org/10.1093/gji/ggw347>

Topal S, Keller E, Bufer A, Koçyiğit A (2016) Tectonic geomorphology of a large normal fault: Akşehir fault, SW Turkey. *Geomorphology* 259:55–69. <https://doi.org/10.1016/j.geomorph.2016.01.014>

Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat

Christoph Clauser

Institute for Applied Geophysics and Geothermal Energy,
RWTH Aachen University, Aachen, Germany

Synonyms

Enthalpy of fusion; Enthalpy of solidification; Specific heat capacity; Thermal capacity; Volumetric heat capacity

Definition

Specific heat capacity c . Physical property defining the amount of sensible heat that can be stored in or extracted from a unit mass of rock per unit temperature increase or decrease, respectively. Isobaric and isochoric specific heat capacities are defined at constant pressure and volume, respectively; dimension: $\text{J kg}^{-1} \text{K}^{-1}$.

Thermal capacity (also volumetric heat capacity) ρc . It is the product of isobaric specific heat capacity and density. Physical property defines the amount of sensible heat that can be stored in or extracted from a unit volume of rock per unit temperature increase or decrease, respectively; dimension: $\text{J m}^{-3} \text{K}^{-1}$.

Latent heat L (also enthalpy of fusion or solidification H). The enthalpy H or latent heat L of fusion or solidification defines the heat energy needed for the corresponding phase change of any amount of substance while the temperature remains constant during the process. When it is referenced to a unit of mass, it is called the “specific” heat of fusion h , while the molar heat of fusion h_{mol} refers to the enthalpy change per amount of substance.

Thermal Storage Properties

The thermal regime of the Earth is defined by its heat sources and sinks, the heat storage and transport processes, and their corresponding physical properties. The storage properties are discussed following. The transport properties, thermal conductivity and thermal diffusivity, are dealt with in this volume in the companion entry ► “Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity”.

Heat can be stored as sensible heat and enthalpy of transformation. In the Earth, sensible heat is defined by the heat capacity of rocks, and the enthalpy of transformation by their latent heat of fusion.

Heat is transmitted within the Earth mainly by diffusion (in the steady state: conduction), advection, and radiation. Generally, heat diffusion dominates heat radiation within the lithosphere of the Earth, at temperatures below 1000 K. For sufficiently large flow rates, convection-driven heat advection, however, can be equally or even more efficient than diffusion. The heat diffusion–advection equation for a porous medium is:

$$\frac{\partial(\rho c_p T)}{\partial t} = \nabla \cdot \left(\underbrace{\lambda \nabla T}_{q_{\text{dif}}} - \underbrace{(\rho c_p)_f T \mathbf{v}}_{q_{\text{adv}}} \right) + \underbrace{A}_{\text{generation}}, \quad (1)$$

where T is temperature in K, t time in s, ρ density in kg m^{-3} , c isobaric specific heat capacity in $\text{J kg}^{-1} \text{K}^{-1}$, λ thermal conductivity in $\text{W m}^{-1} \text{K}^{-1}$, $\mathbf{v}$ specific fluid discharge (volume flow rate per unit cross section area) in m s^{-1} , and A (radiogenic) heat generation rate in W m^{-3} ; subscript “f” refers to fluid properties. Diffusive bulk storage and transport properties for heat in rocks governing this equation are thermal capacity ρc_p , also referred to as volumetric heat capacity, and thermal conductivity λ . Advection is governed by the thermal capacity and specific discharge of the flowing fluid, $(\rho c_p)_f$ and $\mathbf{v}$, respectively.

Heat advection does not require very large flows for becoming as or more efficient than heat diffusion. This is often the case in sedimentary basins (see, e.g., Clauser et al. (2002) for a literature review). But fluid-driven heat advection may be also important in crystalline rocks and on a crustal scale (e.g., Clauser 1992). The nondimensional Péclet and Nußelt numbers, Pe and Nu , are ratios of specific heat flows q_{adv} and q_{dif} (Eq. 1) indicating the efficiency of advective heat transport (driven by fluid flow) versus heat conduction for given flow geometries. For instance, assuming a flow of magnitude v over a distance L and a temperature difference $T_1 - T_0$ one obtains:

$$\begin{aligned} Pe &= \frac{q_{\text{adv}}}{q_{\text{dif}}} = \frac{(\rho c)_f v (T_1 - T_0)}{\lambda T_1 - T_0 / L} = \frac{(\rho c)_f v L}{\lambda}; \\ Nu &= \frac{q_{\text{adv}} + q_{\text{dif}}}{q_{\text{dif}}} = \frac{q_{\text{adv}}}{q_{\text{dif}}} + 1 = Pe + 1. \end{aligned} \quad (2)$$

Thus, advection or diffusion (in the steady state: conduction) dominates for $Pe > 1$ or $Pe < 1$, respectively (in terms of the Nußelt number for $Nu > 2$ or $Nu < 2$, respectively). At temperatures above 1000 K, heat is propagated increasingly as radiated electromagnetic waves, and heat radiation, begin to dominate diffusion (see ► “Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity”, this entry).

A great amount of petrophysical data has been published in the past by individual researchers interested in particular properties. Usually, these data collections are sorted

according to petrographically defined rock types. But rock type, unfortunately, is a weak descriptor for physical properties: Any rock type shows significant variability in characteristics which influence its physical properties, such as (but not limited to), porosity, cementation, and mineral content. This causes also the physical properties to vary accordingly, some by factors others even by orders of magnitude. Thermal rock properties, generally, vary by a factor within the same order of magnitude.

This variability limits the value of individual, rock-type-specific data measured on single samples, unless they originate exactly from the formation (or even location) of interest. In all other cases, a statistical approach may help to account for this intrinsic variability by characterizing physical rock properties by the mean values and standard deviations of populations of a sufficiently large number of individual measurements. Recently, a large number of petrophysical data were compiled into a publically available, digital petrophysical property data base (Bär et al. 2020). This data, while still incomplete in spite of its large number of more than 75 000 individual measurements, may serve both as a source of individual data from specific locations and for calculating means and standard deviations for larger volumes of rock.

Heat Capacity

Heat can be stored and delivered as sensible heat or as latent heat required or liberated by phase changes. This and the next two paragraphs are concerned with sensible heat. Following this, latent heat will be discussed.

Sensible heat capacity C is defined as the ratio of heat ΔQ required to raise the temperature of a mass M of rock by ΔT . For each molecule of mass m , this temperature increase requires an energy of $(f/2) k \Delta T$, where f is the number of degrees of freedom of the molecule and $k = 1.380\,650\,4(24) \times 10^{-23} \text{ J K}^{-1}$ is the Boltzmann constant. For a body of mass M , a temperature increase of ΔT requires an energy of $\Delta Q = (M/m) (f/2) k \Delta T$. Thus the heat capacity of the body at constant volume is:

$$C_V = \frac{\Delta Q}{\Delta T} = \frac{f}{2} k \frac{M}{m} = \frac{f}{2} k N_A = \frac{f}{2} R \quad (\text{J K}^{-1}), \quad (3)$$

where the Avogadro number $N_A = 6.022\,141\,79(30) \times 10^{23} \text{ mol}^{-1}$ equals the number of molecules or atoms in one mole of substance and $R = k N_A = 8\,314\,472(15) \text{ J mol}^{-1} \text{ K}^{-1}$ is the molar gas constant (numerical values for N_A , R , k , and all other physical constants used in this entry are from CODATA 2006). For solids, $f = 6$, corresponding to the three degrees of freedom of potential and kinetic lattice vibration energy in each space direction. Accordingly, the heat capacity of one mole of substance (the molar heat capacity at constant volume) is constant:

$$C_{V,\text{mol}} = 3kN_A = 3R = 24.94 \quad (\text{J mol}^{-1} \text{ K}^{-1}). \quad (4)$$

Isobaric heat capacity $C_{p,\text{mol}}$ is larger than isochoric heat capacity $C_{V,\text{mol}}$ because additional work is required for volume expansion. Both are related by:

$$C_{p,\text{mol}} = C_{V,\text{mol}} + R = \frac{f}{2} + \frac{2}{2} R \quad (\text{J mol}^{-1} \text{ K}^{-1}). \quad (5)$$

With $C_{V,\text{mol}}$ from Eq. 4 and assuming, as above, $f = 6$, this yields:

$$\begin{aligned} C_{p,\text{mol}} &= 3R + R = 4R = 33.26 \quad (\text{J mol}^{-1} \text{ K}^{-1}) \\ &\text{or :} \\ R &= C_{p,\text{mol}} - C_{V,\text{mol}} \end{aligned} \quad (6)$$

Equation 4, the *Dulong–Petit law*, is satisfied well for heavy elements. In contrast, molar heat capacities of lighter elements remain below this upper limit, and the lower the temperature the smaller $C_{V,\text{mol}}$.

Below the Debye temperature, Θ_D , heat capacity varies with temperature. Θ_D tends to zero as T^3 as absolute temperature approaches zero. Θ_D falls in the range between 85 K and 450 K for most substances and between 200 K and 1000 K for most minerals (Stacey and Davis 2008, Table 1). Therefore, heat capacity in the Earth can be well explained by classical Debye theory, in particular in the mantle and except for a thin crustal layer near the Earth's surface. There are, however, exceptions such as beryllium ($\Theta_D = 1440 \text{ K}$) and diamond ($\Theta_D \approx 1800 \text{ K}$). These are caused by the so-called freezing of

Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Table 1 Debye temperature Θ_D and mass number A of selected elements (Kittel 2004)

Element	Mass number A	Θ_D (K)
Carbon	6	2230
Aluminum	13	428
Silica	14	645
Titanium	22	420
Chromium	24	630
Manganese	25	410
Iron	26	470
Nickel	28	450
Copper	29	343
Zinc	30	327
Silver	47	215
Cadmium	48	209
Tin	50	200
Tantalum	73	240
Tungsten	74	400
Platinum	78	240
Gold	79	165
Lead	82	105

vibrational or rotational degrees of freedom, which, at low temperature, cannot absorb heat any more. Therefore heat capacity tends to zero close to absolute zero.

Isobaric and Isochoric Specific Heat Capacity

Specific heat capacity c of a substance is defined as heat capacity C related to unit mass:

$$c = \frac{\Delta Q}{M \Delta T} = \frac{f}{2} \frac{k}{m} = \frac{f}{2} \frac{k}{A_r m_u} \quad (\text{J kg}^{-1} \text{K}^{-1}), \quad (7)$$

where $m_u = 1.660\,538\,782(83) \times 10^{-27}$ kg is the atomic mass constant, defined as $1/12$ of the atomic mass of the carbon isotope ^{12}C , and A_r the atomic mass of a substance relative to m_u . Isobaric specific heat capacity c_p is larger than isochoric specific heat capacity c_v because additional work is required for volume expansion. Their ratio, the adiabatic exponent, is:

$$c_p/c_v = (f + 2)/f. \quad (8)$$

Alternatively, isobaric specific heat capacity c_p can be expressed by enthalpy $H(T, p) = U + pV$, a state function of temperature and pressure, where U , p , and V are internal energy, pressure, and volume, respectively. In a closed system, the change in internal energy dU is the sum of the change in heat dQ and the work dW delivered: $dU = dQ + dW$. If we only consider volume expansion work, $dW = -p dV$, the change in enthalpy dH becomes:

$$\begin{aligned} dH(T, p) &= dE + p dV + V dp = dQ + V dp \\ &= \left(\frac{\partial H}{\partial T} \right)_p dT + \left(\frac{\partial H}{\partial p} \right)_T dp. \end{aligned} \quad (9)$$

Comparing coefficients, we obtain:

$$\frac{dQ}{dT} = \left(\frac{\partial H}{\partial T} \right)_p = c_p. \quad (10)$$

Thus, Eq. 10 defines isobaric specific heat capacity c_p as the first derivative of enthalpy with respect to temperature. Comparison of Eqs. 7 and 10 shows that both expressions are equivalent for $dQ = \Delta Q/M$, and the isobaric enthalpy change ΔH is equal to the specific heat content $\Delta Q/M$.

Isobaric- and isochoric-specific heat capacity are related to compressibility $\beta = \Delta V/(V \Delta p)$ and its inverse, incompressibility or bulk modulus $K = V \Delta p/\Delta V$, by $c_p/c_v = \beta_T/\beta_S = K_S/K_T$ (e.g., Stacey and Davis 2008), where subscripts T and S indicate isothermal and adiabatic conditions, respectively. Inserting the thermodynamic relation $\beta_T = \beta_S + \alpha^2 T/(\rho c_p)$ (e.g., Birch 1966) between isothermal and adiabatic

compressibility yields the relative difference between isobaric and isochoric specific heat capacity:

$$c_p/c_v = 1 + \alpha \gamma T, \quad (11)$$

where $\alpha = \Delta V/(V \Delta T)$ is the volume expansion coefficient,

$$\gamma = \frac{\alpha K_S}{\rho c_p} = \frac{\alpha K_T}{\rho c_v}, \quad (12)$$

the dimensionless Grüneisen parameter, and ρ density. Inserting the expressions for α and K into Eq. 12 yields:

$$\gamma = \frac{1}{\rho c_p} \frac{\Delta V}{V \Delta T} \frac{V \Delta p}{\Delta V} = \frac{\Delta p}{\rho c_p \Delta T}. \quad (13)$$

Thus, the Grüneisen parameter γ is the relative pressure change in a material heated at constant volume.

For solids, i.e., $f = 6$, the absolute difference between isobaric and isochoric specific heat capacity follows from Eqs. 11 and 8:

$$c_p - c_v = \frac{K_T \alpha^2 T}{\rho} = \frac{3}{4} \frac{K_S \alpha^2 T}{\rho}. \quad (14)$$

For crustal rocks ($\gamma = 0.5$; $\alpha = 20 \mu\text{K}^{-1}$; $T < 10^3$ K; $\rho = 2\,600$ kg m $^{-3}$; $K_S < 75$ GPa (Dziewonski and Anderson 1981; Stacey and Davis 2008)), and the difference between isobaric and isochoric specific heat capacity is less than $9 \text{ J kg}^{-1} \text{K}^{-1}$ or 1 % according to Eqs. 14 and 11, respectively. Thus, the distinction between isobaric and isochoric specific heat capacity is negligible for crustal rocks at temperatures below 1000 K. However, it need be made for mantle rocks. From here on, “specific heat capacity” will always refer to isobaric specific heat capacity, denoted simply by the letter c without the subscript “p.”

This classical treatment of heat capacity is sufficient for temperatures above the Debye temperature. In the Earth, temperature exceeds the Debye temperature everywhere except in the crust (Stacey and Davis 2008). Therefore, in experiments at room temperature and atmospheric pressure, we observe deviations from the values predicted by Eqs. 3 to 14, which are based on the classical Dulong–Petit theory. The lower the temperature, lighter the element, and stronger the lattice bonding become, the larger are these deviations. Clearly, interpretation of heat capacity below the Debye temperature is beyond classical mechanics and requires quantum mechanical treatment. This is, however, beyond the scope of this entry, and interested readers are referred to standard physics textbooks (e.g., Tipler and Mosca 2007). Therefore, heat capacity at crustal temperatures should not be calculated from Eqs. 3 and 14 but rather be measured or calculated from appropriate, quantum mechanical equations.

Čermák and Rybach (1982) compiled data on isobaric specific heat capacity for different rock-forming minerals and different igneous, metamorphic, volcanic, and sedimentary rocks as well as the corresponding variations with temperature.

Measuring Techniques

Specific heat capacity c can be measured directly or derived as the isobaric derivative of enthalpy H with respect to temperature. Specific heat capacity of rocks varies with temperature, pressure, porosity, and saturants. Accordingly, in situ values may deviate from laboratory data according to temperature, pressure, and type and content of pore fluid.

Numerous steady-state and transient calorimetric methods are available for measuring specific heat capacity. The most popular are mixing or drop calorimeters and heat flux differential scanning (DSC) calorimeters. The first method yields an absolute value; the second one is a comparative method. All of these methods and their details are discussed in the literature (e.g., Hemminger and Cammenga 1989; Brown 2001; Haines 2002) to which interested readers are referred. The isobaric enthalpy change (or specific heat content) ΔH of solids may be determined by the method of mixtures using a Bunsen-type calorimeter, in which the unknown isobaric enthalpy change of a sample relative to a base temperature, e.g., 25 °C, is compared to the corresponding known isobaric enthalpy change of platinum (Kelley 1960; Somerton 1992). Popov et al. (2016) provide a recent appraisal of different methods for thermal property testing in the laboratory at atmospheric pressure, in particular for specific heat capacity

Calculated Heat Capacity

When no direct measurements can be performed, the isobaric enthalpy change and specific heat capacity of rocks can be calculated according to Kopp's law, Eq. 15, as the arithmetic mean of the individual mineralogical and fluid contributions weighted by the mass fractions n_i of the N individual phases relative to total rock mass:

$$\Delta H = \sum_{i=1}^N n_i \Delta H_i; \quad c = \sum_{i=1}^N n_i c_i; \quad 1 = \sum_{i=1}^N n_i. \quad (15)$$

Based on data for various minerals (e.g., Kelley 1960; Berman and Brown 1985; Somerton 1992; Waples and Waples 2004), the isobaric enthalpy change ΔH or specific heat capacity c can be computed from Eq. 15 for any rock consisting of an arbitrary number of minerals with given volume fractions.

Temperature Dependence

Derived from measured variation of isobaric enthalpy change ΔH with temperature of various oxides, Kelley (1960) suggested a second-order polynomial for fitting ΔH from which $c_p = (\partial H / \partial T)_p$ can be easily calculated. Somerton (1992) and Clauser (2006) report ΔH and c values of various rock-forming oxides and pore fluids. An alternative approach is fitting heat capacity measured at different temperatures directly to polynomials of various degrees (e.g., Maier and Kelley 1932; Berman and Brown 1985; Fei and Saxena 1987; Holland and Powell 1996; Robertson and Hemingway 1995). Waples and Waples (2004) provide a discussion of the various approaches. The polynomial proposed by Berman and Brown (1985),

$$C_{p,\text{mol}} = k_0 + k_1 T^{-0.5} + k_2 T^{-2} + k_3 T^{-3} \quad (T \text{ in K}), \quad (16)$$

is valid for a large temperature range and yields no values incompatible with the Dulong–Petit law for high temperatures. Table 2 lists values for the temperature coefficients k_i in Eq. 16 determined from fits of heat capacity of selected minerals measured at different temperatures.

As an alternative, Waples and Waples (2004) propose a statistical approach describing the general temperature dependence of all rocks and minerals, which can be rescaled easily for individual rocks and minerals. To this end, measured specific heat capacity c_p was normalized by the corresponding normalizing value $c_{p,n}$ at 200 °C (473.15 K), a temperature at or near which data was available. The resulting polynomial regressions yielded much better coefficients of determination R^2 for data measured on nonporous rock ($R^2 = 0.93$) than for those measured on minerals ($R^2 = 0.62$) while the trends were similar. The regression on the combined data for minerals and nonporous rocks yields an expression for the normalized specific heat capacity of a mineral or nonporous rock at arbitrary temperature T with a coefficient of determination $R^2 = 0.65$:

$$c_{p,n}(T) = 0.716 + 1.72 \times 10^{-3} T - 2.13 \times 10^{-6} T^2 + 8.95 \times 10^{-10} T^3 \quad (T \text{ in } ^\circ\text{C}). \quad (17)$$

Equation 17 can be rescaled for any mineral or nonporous rock at any temperature T_2 provided a value $C_p(T_1)$ measured at temperature T_1 is available, for instance from any of the compilations of Berman and Brown (1985), Fei and Saxena (1987), Berman (1988), Holland and Powell (1996), or Robertson and Hemingway (1995):

$$c_p(T_2) = c_p(T_1) \frac{c_{p,n}(T_2)}{c_{p,n}(T_1)}. \quad (18)$$

Additionally, Waples and Waples (2004) consider the variation of specific heat capacity with lithology, where

Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Table 2 Coefficients for calculating isobaric molar heat capacity $C_{p,\text{mol}}$ ($\text{J mol}^{-1} \text{K}^{-1}$) (From Eq. 16; Berman 1988)

Mineral	Chemical composition	k_0 ($\text{J mol}^{-1} \text{K}^{-1}$)	$k_1 \times 10^{-2}$ ($\text{J mol}^{-1} \text{K}^{-1/2}$)	$k_2 \times 10^{-5}$ ($\text{J mol}^{-1} \text{K}$)	$k_3 \times 10^{-7}$ ($\text{J mol}^{-1} \text{K}^2$)	T (K)
Albite	$\text{NaAlSi}_3\text{O}_8$	393.64	-24.155	-78.928	107.064	250–1373
Almandine	$\text{Fe}_3\text{Al}_2\text{Si}_3\text{O}_{12}$	573.96	-14.831	-292.920	502.208	420–997
Andalusite	Al_2SiO_5	236.48	-11.029	-75.268	93.644	254–1601
Anorthite	$\text{CaAl}_2\text{Si}_2\text{O}_8$	439.37	-37.341	0.0	31.702	292–1373
Calcite	CaCO_3	178.19	-16.577	-4.827	16.660	257–1200
Cordierite	$\text{Mg}_2\text{Al}_4\text{Si}_3\text{O}_{18}$	954.39	-79.623	-21.173	-37.021	256–1652
Corundum	Al_2O_3	155.02	-8.284	-38.614	40.908	250–2300
Dolomite	$\text{CaMg}(\text{CO}_3)_2$	328.48	-25.544	-46.885	79.038	250–650
Enstatite (ortho)	MgSiO_3	166.58	-12.066	-22.706	27.915	254–1273
Fayalite	Fe_2SiO_4	248.93	-19.239	0.0	-13.910	255–1370
Forsterite	Mg_2SiO_4	238.64	-20.013	0.0	-11.624	253–1807
Hematite	Fe_2O_3	146.86	0.0	-55.768	52.563	258–1757
Kaolinite	$\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$	523.23	-44.267	-22.443	9.231	256–560
Kyanite	Al_2SiO_5	262.68	-20.014	-19.997	-6.318	252–1503
Lime	CaO	58.79	-1.339	-11.471	10.298	250–1176
Magnesite	MgCO_3	162.30	-11.093	-48.826	87.466	288–750
Magnetite	Fe_3O_4	207.93	0.0	-72.433	66.436	257–1825
Muscovite	$\text{KAl}_3\text{Si}_3\text{O}_{10}(\text{OH})_2$	651.49	-38.732	-185.232	274.247	257–967
Periclase	MgO	61.11	-2.962	-6.212	0.584	250–1798
Potassium feldspar	KAlSi_3O_8	381.37	-19.411	-120.373	183.643	250–997
Quartz	SiO_2	80.01	-2.403	-35.467	49.157	250–1676
Sillimanite	Al_2SiO_5	256.73	-18.827	29.774	25.096	253–1496
Sphene	CaTiSiO_5	234.62	-10.403	-51.183	59.146	255–1495
Spinel	MgAl_2O_4	235.90	-17.666	-17.104	4.062	256–1805
Talc	$\text{Mg}_3\text{Si}_4\text{O}_{10}(\text{OH})_2$	664.11	-51.872	-21.472	-32.737	250–639
Wollastonite	CaSiO_3	149.07	-6.903	-36.593	48.435	251–1433

interested readers find a specific discussion regarding coals of different carbon content or maturity.

Mottaghy et al. (2005) used a second-order polynomial in temperature to fit the variation of isobaric specific heat capacity with temperature measured on a suite of metasedimentary, volcanic, magmatic, and metamorphic rocks:

$$c_p(T) = A_0 + A_1 T^1 + A_2 T^2 \quad (c_p \text{ in } \text{J kg}^{-1} \text{K}^{-1}, 1^\circ\text{C} \leq T \leq 100^\circ\text{C}). \quad (19)$$

The average values for the coefficients A_0 , A_1 , and A_2 determined from a regression of $c_p(T)$ measured over a temperature range of $1^\circ\text{C} \leq T \leq 100^\circ\text{C}$ on 26 samples from seven boreholes are: $\bar{A}_0 = 747.177 \text{ J kg}^{-1} \text{K}^{-1}$; $\bar{A}_1 = 1.862 \text{ J kg}^{-1} \text{K}^{-2}$; and $\bar{A}_2 = -0.002 \text{ J kg}^{-1} \text{K}^{-3}$.

Based on a composition of 30 % quartz, 60 % feldspar (albite), and 10 % phyllosilicates (5 % phlogopite and 5 % annite), Whittington et al. (2009) suggest average “bulk crustal” molar-specific heat capacity equations based on end-

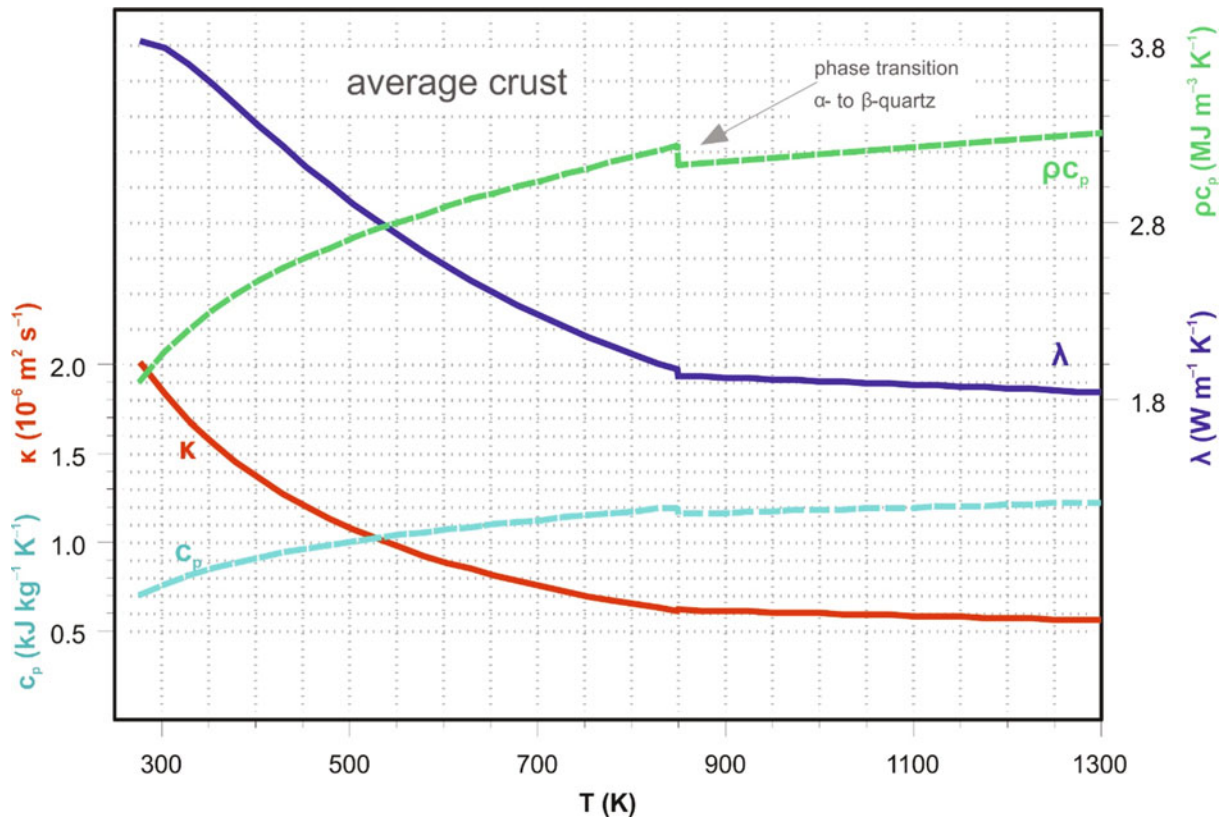
member mineral data for two temperature ranges, separated by the transition at 846 K ($\sim 573^\circ\text{C}$) between α - and β -quartz:

$$C_{p,\text{mol}}(T) = \begin{cases} 199.50 + 0.0857 T - 5.0 \times 10^{-6} T^{-2}; & T \leq 846 \text{ K} \\ 229.32 + 0.0323 T - 47.9 \times 10^{-6} T^{-2}; & T > 846 \text{ K} \end{cases} \quad (C_{p,\text{mol}} \text{ in } \text{J mol}^{-1} \text{K}^{-1}) \quad (20)$$

Assuming an average molar mass of $0.22178 \text{ kg mol}^{-1}$, this yields the variation of isobaric specific heat capacity c_p with temperature shown in Fig. 1.

Volumetric Heat Capacity: Thermal Capacity

When heat capacity is related to unit volume rather than to unit mass or unit amount of substance, it is referred to as volumetric heat capacity or thermal capacity. It can be calculated as the product of specific heat capacity c and density ρ or as the ratio of thermal conductivity λ and thermal diffusivity κ by



Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Fig. 1 Variation of specific heat capacity c_p , the phonon components of thermal diffusivity and thermal conductivity κ_p and λ_p , respectively, and thermal capacity $\rho c_p = \lambda_p/\kappa_p$ with temperature in an average crust according to Eq. 20, and Eq. 31 in

► “Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity” (this entry) for c_p and κ , respectively, assuming molar mass = 0.221 78 kg mol⁻¹ and $\rho = 2\,700$ kg m⁻³ (after Whittington et al. 2009)

$$\rho c = \lambda / \kappa . \quad (21)$$

Again, Kopp’s law yields the rock’s bulk thermal capacity $(\rho c)_b$ as

$$(\rho c)_b = (1 - \phi) (\rho c)_s + \phi \sum_{i=1}^N S_i (\rho c)_i, \quad (22)$$

where ϕ is porosity and $(\rho c)_s$ thermal capacity of the rock skeleton. S_i and $(\rho c)_i$ are saturation and thermal capacity of the i^{th} fluid phase in the pore space. The skeleton thermal capacity itself may be calculated again from Kopp’s law for a given mineral assemblage and the corresponding volume fractions of the solid phase from Eq. 15. Because of the low density of air and gas – about three orders of magnitude lower than that of water and rock – the contribution of the gas phase to thermal capacity can often be ignored. In this case, $N = 2$ for the fluid phases water and oil or $N = 1$ for water only. Expressions for the density of various fluids are reported in Clauser (2006). Data on density of various minerals and rocks are listed, e.g., in Wohlenberg (1982a, b) or Olhoeft and Johnson (1989). Using Eqs. 17 and 18, Waples and Waples

(2004) analyzed a substantial collection of density and specific heat capacity data from various authors and transformed specific heat capacity and thermal capacity to a uniform reference temperature of 20 °C (Table 3).

The mean thermal capacity of “impervious” rocks was found at 2.300(46) MJ m⁻³ K⁻¹ by Roy et al. (1981). This is acceptably close to the mean of 2.460(65) MJ m⁻³ K⁻¹ found by Waples and Waples (2004) for inorganic minerals.

Based on density, specific heat capacity, and thermal conductivity measured at room temperature, Mottaghy et al. (2005) determined thermal capacity as inverse slope of a regression of diffusivity on thermal conductivity according to Eq. 21. All values fell well within ± 20 % of the average of 2.3 MJ m⁻³ K⁻¹ recommended by Beck (1988).

Mottaghy et al. (2008) determined average values for thermal capacity according to Eq. 21 for metamorphic and magmatic crystalline rocks as the inverse slope of a linear regression of values of thermal diffusivity versus thermal conductivity measured at temperatures in the range 20 °C $\leq T \leq$ 300 °C:

$$\kappa(T) = \frac{\lambda(T)}{m + nT}, \quad (T \text{ in } ^\circ\text{C}). \quad (23)$$

Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Table 3 Typical values or ranges for density ρ , isobaric specific heat capacity c_p , and thermal capacity ρc_p of selected rocks at 20 °C (Waples and Waples 2004, Petrunin et al. 2004)

Rock	ρ (kg m ⁻³)	c_p (J kg ⁻¹ K ⁻¹)	ρc_p (kJ m ⁻¹ K ⁻¹)
Albite	2540–2560	755–780	1922–1991
Amphibole	3010	700–1134	2110–3410
Anhydrite	2950–2960	590–940	1740–2780
Anorthite	2740	800	2202
Basalt	2870	880–900	2526–2583
Clay	2680	860	2300
Coal	1350	1300	1760
Diabase	2790	731–860	2040–2400
Dolomite	2800	900	2520
Gabbro	2970–3000	650–1000	1950–2970
Gneiss	2700	770–979	2080–2640
Granite	2620–2650	600–1172	1590–3070
Gypsum	2370	1010	2390
Limestone	2760–2770	680–880	1880–2430
Peridotite	2740–3190	705–1005	1930–3210
Pyroxenite	3190–3240	660–1000	2140–3190
Quartzite	2640	731–1013	1930–2670
Rock salt	2160	880	1900
Sandstone	2640	775	2050
Schist	2770–2900	790–1096	2190–3180
Serpentinite	2270–2540	730–1005	1660–2550
Siltstone	2680	910	2449
Slate	2770–2780	740–1113	2060–3080
Syenite	2820	460	1300
Talc	2780	1000	2780
Tuff	2750	1090	3000

Regression of data measured on seven samples collected along a profile crossing the Eastern Alps from north to south and on nine samples from the northern rim of the Fennoscandian Shield near the Kola ultra-deep borehole SG-3 yielded $m = 2\,066(70) \text{ kJ m}^{-3} \text{ K}^{-1}$, $n = 2.2(4) \text{ kJ m}^{-3} \text{ K}^{-2}$, $R^2 = 0.97$ and $m = 2\,404(91) \text{ kJ m}^{-3} \text{ K}^{-1}$, $n = 3.6(5) \text{ kJ m}^{-3} \text{ K}^{-2}$, and $R^2 = 0.92$, respectively. This yields a range of thermal capacity for the Alpine and Fennoscandian data of about $2.1 \text{ MJ m}^{-1} \text{ K}^{-1}$ to $2.7 \text{ MJ m}^{-1} \text{ K}^{-1}$ and $2.4 \text{ MJ m}^{-1} \text{ K}^{-1}$ to $3.5 \text{ MJ m}^{-1} \text{ K}^{-1}$, respectively, in the temperature range $20^\circ\text{C} \leq T \leq 300^\circ\text{C}$.

The product of an average density of $2\,700 \text{ kg m}^{-3}$ for the crust and isobaric-specific heat capacity calculated for an average molar mass of $0.221\,78 \text{ kg mol}^{-1}$ according to Eq. 20 (Whittington et al. 2009) yields the variation of thermal capacity ρc_p (Eq. 21) with temperature shown in Fig. 1. Its increase with temperature by about a factor of 1.7 in a temperature interval of 1 000 K demonstrates that the effect of temperature is stronger for the phonon component of thermal diffusivity κ than for phonon thermal conductivity λ_p due to the increase of specific heat capacity. Assuming a constant density throughout the crust implies that the increase

and decrease in density due to the increase in pressure and temperature, respectively, partly cancel each other and that these changes are small compared to those of specific heat capacity and thermal diffusivity.

Latent Heat

Solidification of magma and melting of rocks as well as freezing and thawing of water in soils or rocks liberate or consume heat, respectively. The like applies to mineral phase changes such as those associated with the seismic discontinuities at 410 km, 520 km, and 660 km in the transition zone from 400 km to 600 km between the upper and lower mantle. These mineral phases are chemically identical but differ with respect to crystal structure and therefore elastic properties. This is why this transition is seen in the seismic wave field. Phase transitions require a certain pressure p and temperature T , but also a specific relation between these two state variables expressed by the “Clapeyron slope” $dp/dT = \Delta S(p,T)/\Delta V(p,T)$ (the inverse of the Clausius–Clapeyron equation), where S and V are entropy and volume,

Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Table 4 Comparison of isobaric specific heat capacity c_p and latent heat of melting L of granite, basalt, seawater, and freshwater (Stacey and Davis 2008 supplemented)

	Granite	Basalt	Seawater	Freshwater	Dry air (15 °C, 1000 hPa)
c_p (kJ kg ⁻¹ K ⁻¹)	0.83	0.88	3.99	4.18	1.006
L (kJ kg ⁻¹)	420	420	335	333.55	196

respectively. This means that the depth where a certain phase transition occurs varies with the ambient temperature in the crust. Positive and negative values for the Clapeyron slope are associated with exothermic and endothermic reactions, respectively.

Phase changes generally consume or deliver much more latent heat than can be stored or delivered as sensible heat: It requires a temperature increase of more than 500 K to equal by sensible heat the amount of latent heat required to melt 1 kg of granite, and still an increase of more than 80 K to equal by sensible heat the amount of latent heat required to melt 1 kg of sea ice (Table 4).

The *discontinuity at 410 km* is generally associated with the transition in olivine from α -olivine to β -spinel, also named wadsleyite (e.g., Stacey and Davis 2008). At expected pressure and temperature of 14 GPa and 1600 K, respectively, corresponding values for the Clapeyron slope vary between 2.9 MPa K⁻¹ and 4.8 MPa K⁻¹ (Bina and Helffrich 1994; Katsura et al. 2004; Stacey and Davis 2008). The *discontinuity at 520 km* is associated with the transition from β -spinel (wadsleyite) into γ -spinel, also named ringwoodite (e.g., Stacey and Davis 2008). At an expected pressure of 18 GPa, a temperature increase, such as by an ascending plume, would require a higher pressure for this transition according to inferred Clapeyron slopes of 4.0 MPa K⁻¹ to 5.3 MPa K⁻¹ (Helffrich 2000; Deuss and Woodhouse 2001). A second transition occurs between garnet and calcium-perovskite (CaSiO₃), where the iron in garnet goes into Ilmenite, and its CaSiO₃-component into calcium-perovskite. This reaction has a negative Clapeyron slope. The two slopes of different signs may shift the depth for two transitions into opposite directions, which is observed as a splitting of the 520 km discontinuity. The *discontinuity at 660 km* defines the transition into the lower mantle. It is caused by the transition of γ -spinel (ringwoodite) into magnesium-perovskite (MgSiO₃) and ferrous periclase (magnesiowüstite, (Fe,Mg)O). At an expected pressure of 23.5 MPa, this endothermic transition is associated with a Clapeyron slope of -2.8 MPa K⁻¹ (Stacey and Davis 2008).

The latent heat L that corresponds to these additional heat sources and sinks can be elegantly combined with the specific sensible heat capacities of the liquid and solid rock, c_l and c_s , respectively, into an effective bulk specific heat capacity c_{eff} . This effective specific heat capacity then accounts for the

entire enthalpy change, including latent heat. In this approach, the latent heat effects are assumed to occur between the solidus and liquidus temperatures T_1 and T_2 , respectively. The heat liberated by a solidifying (“freezing”) liquid phase is obtained by weighting the corresponding contributions by the mass fractions of liquid and solid phases, ϕ_{liquid} and ϕ_{solid} , respectively. The change of specific enthalpy $h = H/m$, where m is the mass of the rock volume then becomes

$dh_{\text{freezing}} = \{(\phi c)_{\text{liquid}} + (\phi c)_{\text{solid}}\} dT + L d\phi_{\text{liquid}}$, and the effective heat capacity c_{eff} is:

$$c_{\text{eff}}^{\text{freezing}} = \frac{dh}{dT} = (\phi c)_{\text{liquid}} + (\phi c)_{\text{solid}} + L \frac{d\phi_{\text{liquid}}}{dT}. \quad (24)$$

Conversely, when considering melting the solid phase, the enthalpy change of the rock volume is $dh_{\text{melting}} = \{(\phi c)_{\text{liquid}} + (\phi c)_{\text{solid}}\} dT + L d\phi_{\text{solid}}$, and the effective heat capacity in this case c_{eff} is:

$$c_{\text{eff}}^{\text{melting}} = \frac{dh}{dT} = (\phi c)_{\text{liquid}} + (\phi c)_{\text{solid}} + L \frac{d\phi_{\text{solid}}}{dT}. \quad (25)$$

Summary

Understanding the thermal regime of the Earth requires appreciation of properties and mechanisms for storage, transport, and generation of heat within the Earth. Both experimental and indirect methods are available for inferring specific or molar heat capacity and, with additional information on density, also thermal capacity (or volumetric heat capacity). In the temperature and pressure ranges where they occur, phase changes consume or deliver characteristic amounts of latent heat (or enthalpy of fusion or solidification), depending on whether they are endothermic or exothermic. Latent heat may be combined with specific heat capacity into an effective heat capacity.

Cross-References

- [Geothermal Record of Climate Change.](#)
- [Heat Flow, Continental.](#)
- [Heat Flow, Seafloor: Methods and Observations.](#)
- [Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity.](#)

Acknowledgments This contribution benefitted from insightful comments by an anonymous reviewer and editorial help by Dr. Sukanta Roy, Hyderabad.

Bibliography

- Bär K, Reinsch T, Bott J (2020) P³ – PetroPhysical property database – a global compilation of lab measured rock properties. *Earth Syst Sci Data Discuss.* <https://doi.org/10.5194/essd-2020-15>, in review, 2020. Database available at <https://doi.org/10.5880/GFZ.4.8.2019.P3>. Retrieved 21 March 2020
- Beck AE (1988) Methods for determining thermal conductivity and thermal diffusivity. In: Hänel R, Rybach L, Stegena L (eds) *Handbook of terrestrial heat flow density determination*. Kluwer, Dordrecht, pp 87–124
- Berman RG (1988) Internally-consistent thermodynamic data for minerals in the system Na₂O-K₂O-CaO-MgO-FeO-Fe₂O₃-Al₂O₃-SiO₂-TiO₂-H₂O-CO₂. *J Petrol* 29(2):445–522
- Berman RG, Brown TH (1985) Heat capacity of minerals in the system Na₂O-K₂O-CaO-MgO-FeO-Fe₂O₃-Al₂O₃-SiO₂-TiO₂-H₂O-CO₂. *Contrib Mineral Petrol* 89:168–183
- Bina CR, Helffrich G (1994) Phase transition Clapeyron slopes and transition zone seismic discontinuity topography. *J Geophys Res* 99(B8):15853–15860. <https://doi.org/10.1029/94JB00462>
- Birch F (1966) Section 7: compressibility; elastic constants. In: Clark SP Jr (ed) *Handbook of physical constants*, Revised edition, Memoir 97. Geological Society of America, Washington, pp 97–173
- Brown ME (2001) *Introduction to thermal analysis: techniques and applications (hot topics in thermal analysis and calorimetry)*. Kluwer, Dordrecht
- Čermák V, Rybach L (1982) Thermal conductivity and specific heat of minerals and rocks. In: Angenheister G (ed) *Landolt-Börnstein: numerical data and functional relationships in science and technology*. New series, vol V(1a). Springer, Berlin, pp 305–343
- Clauser C (1992) Permeability of crystalline rocks. *EOS Trans Am Geophys Union* 73(21):233, 237. (for an update see also: Clauser C (2001) Update of the permeability of crystalline rocks. Report, Institute for Applied Geophysics and Geothermal Energy, E.ON Energy Research Center, RWTH Aachen University. <http://www.eonerc.rwth-aachen.de/go/id/tsm/>. Retrieved 16 Oct 2010)
- Clauser C (2006) Geothermal energy. In: Heinloth K (ed) *Landolt-Börnstein, Group VIII: advanced materials and technologies. Energy technologies*, Subvol. C: Renewable energies, vol 3. Springer, Heidelberg, pp 480–595
- Clauser C, Griesshaber E, Neugebauer HJ (2002) Decoupled thermal and mantle helium anomalies – implications for the transport regime in continental rift zones. *J Geophys Res* 107(B 11):2269. <https://doi.org/10.1029/2001JB000675>
- Committee on Data for Science and Technology (CODATA) (2006) CODATA internationally recommended values of the fundamental physical constants. International Council of Scientific Unions (ICSU), Paris. <http://physics.nist.gov/cuu/Constants/index.html>. Retrieved 9 Sept 2010
- Deuss A, Woodhouse J (2001) Seismic observation of splitting of the mid-transition zone discontinuity in Earth's mantle. *Science* 294:354–357
- Dziewonski AM, Anderson DL (1981) Preliminary reference earth model. *Phys Earth Planet Inter* 25:297–356
- Fei Y, Saxena SK (1987) An equation for the heat capacity of solids. *Geochim Cosmochim Acta* 52(2):251–254
- Haines P (ed) (2002) *Principles of thermal analysis and calorimetry*. The Royal Society of Chemistry, Cambridge, UK
- Helffrich G (2000) Topography of the transition zone seismic discontinuities. *Rev Geophys* 38:141–158
- Hemminger WF, Cammenga HK (1989) *Methoden der Thermischen Analyse*. Springer, Berlin
- Holland TJB, Powell R (1996) An internally consistent thermodynamic data set for phases of petrological interest. *J Metamorph Geol* 16:309–343
- Katsura T, Yamada H, Nishikawa O, Song M, Kubo A, Shinmei T, Yokoshi S, Aizawa Y, Yoshino T, Walter MJ, Ito E, Funakoshi K-i (2004) Olivine-wadsleyite transition in the system (Mg, Fe)₂SiO₄. *J Geophys Res* 109:B02209. <https://doi.org/10.1029/2003JB002438>
- Kelley K (1960) Contributions to the data on theoretical metallurgy: XIII high-temperature heat-content, heat-capacity, and entropy data for the elements and inorganic compounds, U. S. Bureau of Mines Bulletin 584. U. S. Government Printing Office, Washington, DC
- Kittel C (2004) *Introduction to solid state physics*, 8th edn. Wiley, Hoboken
- Maier CG, Kelley KK (1932) An equation for the representation of high temperature heat content data. *J Am Chem Soc* 54(8):3243–3246
- Mottaghy DC, Schellschmidt R, Popov YA, Clauser C, Kukkonen IT, Nover G, Milanovsky S, Romushkevich RA (2005) New heat flow data from the immediate vicinity of the Kola super-deep borehole: vertical variation in heat flow confirmed and attributed to advection. *Tectonophysics* 401(1–2):119–142. <https://doi.org/10.1016/j.tecto.2005.03.005>
- Mottaghy D, Vosteen H-D, Schellschmidt R (2008) Temperature dependence of the relationship of thermal diffusivity versus thermal conductivity for crystalline rocks. *Int J Earth Sci* 97:435–442
- Olhoeft GR, Johnson GR (1989) Densities of rocks and minerals. In: Carmichael RS (ed) *Practical handbook of physical properties of rocks and minerals*. CRC Press, Boca Raton, pp 139–176
- Petrinin GI, Popov VG, Il'in IA (2004) Conductive heat transfer in plagioclases. *Izv Phys Solid Earth* 40(9):752–759
- Popov Y, Beardsmore G, Clauser C, Roy S (2016) ISRM Suggested methods for determining thermal properties of rocks from laboratory tests at atmospheric pressure. *Rock Mech Rock Eng* 49(10):4179–4207
- Robertson EC, Hemingway BS (1995) Heat capacity and heat content of rocks, open-file report 95-622. U. S. Geological Survey, Reston
- Roy RF, Beck AE, Touloukian YS (1981) Thermophysical properties of rocks. In: Touloukian YS, Judd WR, Roy RF (eds) *Physical properties of rocks and minerals. McGraw-Hill/CINDAS data series on material properties*, vol II-2. McGraw-Hill, New York, pp 409–502
- Somerton WH (1992) *Thermal properties and temperature related behavior of rock/fluid systems*. Elsevier, Amsterdam
- Stacey FD, Davis PM (2008) *Physics of the earth*, 4th edn. Cambridge University Press, Cambridge
- Tipler PA, Mosca GP (2007) *Physics for scientists and engineers: extended version*. Palgrave Macmillan, Basingstoke
- Waples DW, Waples JS (2004) A review and evaluation of specific heat capacities of rocks, minerals, and subsurface fluids. Part 1: minerals and nonporous rocks. *Nat Resour Res* 13(2):97–122
- Whittington AG, Hofmeister AM, Nabelek PI (2009) Temperature-dependent thermal diffusivity of the Earth's crust and implications for magmatism. *Nature* 458:319–321. <https://doi.org/10.1038/nature07818>
- Wohlenberg J (1982a) Density of minerals. In: Hellwege K-H (ed) *Group V: geophysics vol 1: physical properties of rocks*, Subvol. A. Landolt-Börnstein, vol 1. Springer, Berlin, pp 66–113
- Wohlenberg J (1982b) Density of rocks. In: Hellwege K-H (ed) *Group V: geophysics vol 1: Physical properties of rocks*, Subvol. A. Landolt-Börnstein, vol 1. Springer, Berlin, pp 113–120

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

Christoph Clauser

Institute for Applied Geophysics and Geothermal Energy,
RWTH Aachen University, Aachen, Germany

Synonyms

Heat conductivity; Thermal conductivity; Thermal diffusivity

Definition

Thermal conductivity (also: heat conductivity) λ : Physical property governing heat diffusion in the steady state. It defines how much heat flows across a unit cross section of rock along a unit distance per unit temperature decrease per unit time; dimension: $\text{W m}^{-1} \text{K}^{-1}$.

Thermal diffusivity κ : Physical property governing transient heat diffusion. It is defined by the ratio of thermal conductivity and thermal capacity, i.e., by the ratio of heat flowing across the face of a unit volume and the heat stored in the unit volume per unit time; dimension: $\text{m}^2 \text{s}^{-1}$.

General Remark

A great amount of petrophysical data has been published in the past by individual researchers interested in thermal rock properties. Usually, these data collections are sorted according to petrographically defined rock types. But rock type, unfortunately, is a weak descriptor for physical properties: Any rock type shows significant variability in characteristics which influence its physical properties, such as (but not limited to) porosity, cementation, mineral content, etc. This causes also the physical properties to vary accordingly, some by factors and others even by orders of magnitude. Thermal rock properties, generally, vary by a factor within the same order of magnitude.

This variability limits the value of individual, rock type-specific data measured on single samples, unless they originate exactly from the formation (or even location) of interest. In all other cases, a statistical approach help accounting for this intrinsic variability by characterizing physical rock properties by the mean values and standard deviations of populations of a sufficiently large number of individual measurements. An example of this approach is the nationwide

survey performed in the Republic of Korea where thermal conductivity measurements on 1 560 rock samples were combined into statistical moments for ten, eight, and seven major types of igneous, metamorphic, and sedimentary rocks, respectively (Shim et al. 2010).

Recently, a large number of petrophysical data were compiled into a publically available, digital database (Bär et al. 2020) comprising thermal, mechanical, electrical, and magnetic rock properties. Each data point is supplemented by a reference (if available) and information on sample location, petrographic analysis, and chronostratigraphic age. This data, while still incomplete in spite of its large number of more than 75 000 individual measurements, is another valuable source of individual data from specific locations but also useful for calculating means and standard deviations for larger volumes of rock.

Thermal Conductivity

Fourier's law of heat conduction defines the vector of specific heat flow q_i , i.e., heat flow normalized by area, as the product of the thermal conductivity tensor λ_{ij} and the temperature gradient vector $\partial T/\partial x_j$:

$$q_i = -\lambda_{ij} \frac{\partial T}{\partial x_j}. \quad (1)$$

Temperature measurements are usually performed along vertical profiles in boreholes, yielding only the vertical component of the temperature gradient. To a good approximation, thermal conductivity in some, particularly volcanic and plutonic, rocks is isotropic. In this case, heat will flow predominantly vertically, and it is sufficient to consider only the vertical component of Eq. 1. In contrast, thermal conductivity of many sedimentary and metamorphic rocks is strongly anisotropic, and lateral heat flow may be significant. Hence, information on anisotropy is often required, demanding laboratory measurements in different directions. Anisotropy occurs over a wide range of scales, from microscopic, over laboratory and macroscopic, to tectonic.

Heat transport in most of the Earth's crust and mantle is diffusive and caused by (1) scattering of quantized lattice vibrations, the phonons and (2) by diffusive (as opposed to ballistic) radiation of photons. These two processes are described by phonon thermal conductivity λ_p and radiative thermal conductivity, λ_r , respectively, the sum of which is often termed effective thermal conductivity, λ_{eff} . In most polycrystalline material, heat radiation starts to dominate at temperatures above 2 500 K. In single crystals and glasses with little or no scattering (e.g., obsidian), however, radiation

may become important at much lower temperatures of 500 K–1 000 K. In the metallic core, scattering of electrons provides a third heat transfer mechanism (Fig. 1).

The following text deals mostly with phonon thermal conductivity. Radiative heat transfer is discussed in the paragraph on radiative thermal conductivity and in the chapter on thermal diffusivity.

Measuring Techniques

Thermal conductivity can be measured in the laboratory on rock samples, i.e., cores or cuttings or in situ either in boreholes or with marine heat flow probes with a shallow penetration of three to twenty meters. There are numerous steady-state and transient techniques available for measuring thermal conductivity, the most prominent being the “divided bar,” “needle probe,” “optical scanning,” and “(laser) flash,” all of which are suitable to determine also the anisotropy of thermal conductivity. These methods are described in, for instance, Parker et al. (1961), Kappelmeyer and Haenel (1974), Beck (1988), Davis (1988), Somerton (1992), Schilling (1999), Popov et al. (1999b), Beardsmore and Cull (2001), and Blumm and Lemarchand (2002). Among these techniques, the transient ones, e.g., line-source methods (needle probes and half-space line-sources) and optical scanning, are also suitable for determining thermal diffusivity. Popov et al. (2016) provide an appraisal of different methods for thermal property testing in the laboratory at atmospheric pressure, in particular for thermal conductivity and thermal diffusivity.

As with most other petrophysical properties, in situ thermal conductivity may deviate significantly from laboratory values, even if the effects of temperature, pressure, and pore fluid are accounted for. This scale dependence involves different aspects: In situ measurements represent an

average over a much larger rock volume than laboratory measurements performed on small samples and cannot resolve small-scale variations. A subsequent upscaling may be necessary to identify the appropriate representative elementary volume (REV) for which reasonable transport parameter averages can be defined. Jorand et al. (2013) discuss results of some frequently used averaging methods and, based on a detailed laboratory study, compares them with simulated values obtained from numerical thermal petrophysics.

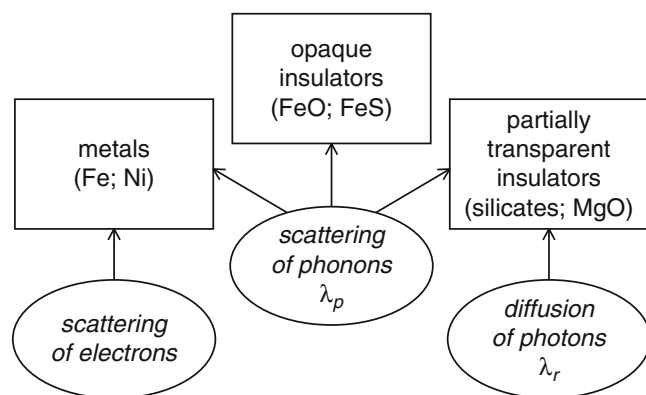
Indirect Methods

When no data are available or no direct measurements can be performed, thermal conductivity can be inferred indirectly, either from mineralogical composition and saturating fluids or from correlations with other physical properties. While some of these methods are based on well-defined physical models, others are purely empirical.

Estimation from mineralogical composition and saturating fluids: Thermal conductivity of rocks may be estimated from their mineral and fluid content. Thermal conductivity of minerals varies much less than in rocks, due to their well-defined composition. As the bulk thermal conductivity of porous rocks varies with different saturants, it may be also of interest to know rock thermal conductivity for other saturants than those used in the laboratory measurement. Numerous models based on volume fractions of the individual mineral and fluid phases have been proposed, all with specific advantages and disadvantages: Some overestimate while others underestimate systematically the true bulk thermal conductivity. Most of them are valid only within a specific range of volume fractions (or porosities) and yield unreasonable results outside. To overcome this problem, additional parameters may be introduced in order to incorporate rock structure into a mixing law (see below).

The parallel and series models for thermal resistance of layered media are easy to understand, but have the disadvantage of being rather special cases, applicable mostly to bedded sediments. They correspond to the well-known weighted arithmetic and harmonic means, λ_{ari} and λ_{har} , Eqs. 2a and 2b, and are related to heat flowing parallel or perpendicular to bedding, respectively. They define upper and lower limits for all other models, sometimes also referred to as the Voigt upper bound and Reuss lower bound, respectively. Thus they constrain the maximum variance of possible predictions. The arithmetic mean between the two is known as the Voigt-Reuss-Hill average, λ_{VRH} (Eq. 2c). It is useful for estimating an effective thermal conductivity, while the arithmetic and harmonic means are used to define a range for possible values.

The weighted geometric and square root means λ_{geo} and λ_{V} (Eqs. 2d and e; e.g., Beardsmore and Cull 2001),



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Fig. 1 Different types of material in the Earth and associated mechanisms of heat transport (Redrawn after Hofmeister et al. 2009)

respectively, are associated with a mixture of different fluid and mineral phases of unspecified geometry. In many cases, both means successfully explain the data and are therefore widely used for lack of more specific information on the geometric arrangement of the individual volume fractions. In the special case where $N = 2$ and $n_1 = n_2 = 0.5$, the arithmetic, harmonic, and geometric means are related by $\lambda_{geo} = \sqrt{\lambda_{ari} \lambda_{har}}$.

Effective medium theory (Bruggeman, 1935) provides the effective medium mean λ_{eff} (Eq. 2f) which is useful for macroscopically homogeneous and isotropic rocks consisting of randomly distributed grains and pores.

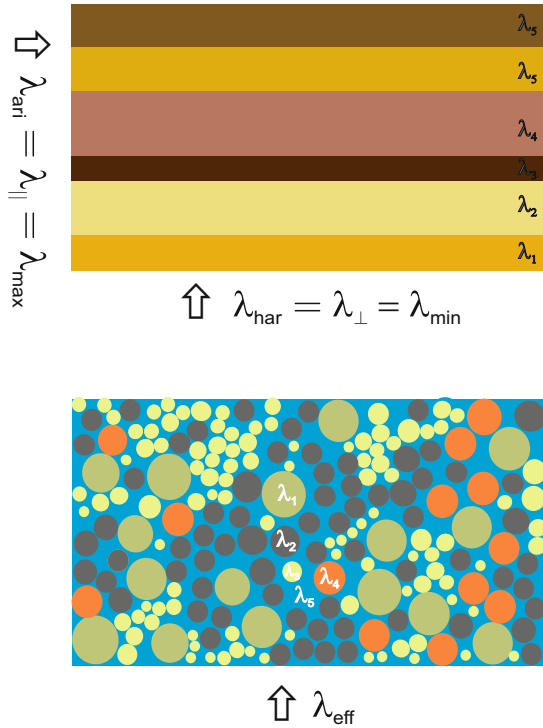
The upper and lower Hashin-Shtrikman bounds λ_{HS}^+ and λ_{HS}^- (Eq. 3; Hashin and Shtrikman 1962), respectively, provide tighter constraints for the predictions of different models other than the arithmetic and harmonic means. The arithmetic average of both defines the Hashin-Shtrikman mean (Eq. 2g). Geometrically, the lower Hashin-Shtrikman bound λ_{HS}^- corresponds to a rock model consisting of grains suspended in a fluid and is close to the geometric mixing law. In contrast, the upper Hashin-Shtrikman bound λ_{HS}^+ may be related to a well-lithified rock with spherical, fluid-filled pores and is close to

the square root mean. Figure 2 illustrates the geometries corresponding to these different models.

If λ_i denotes the thermal conductivity and n_i the volume fraction of the i -th phase relative to the total volume ($1 = \sum n_i$), these seven weighted means are defined by:

$$\begin{aligned} (a) \quad & \lambda_{\max} = \lambda_{\text{ari}} = \lambda_{\parallel} = \sum_{i=1}^N n_i \lambda_i; \\ (b) \quad & \lambda_{\min} = \lambda_{\text{har}} = \lambda_{\perp} = \left(\sum_{i=1}^N \frac{n_i}{\lambda_i} \right)^{-1}; \\ (c) \quad & \lambda_{\text{VRH}} = \frac{1}{2} (\lambda_{\parallel} + \lambda_{\perp}); \quad (d) \quad \lambda_{\text{geo}} = \prod_{i=1}^N \lambda_i^{n_i}; \\ (e) \quad & \lambda_{\sqrt{}} = \left(\sum_{i=1}^N n_i \sqrt{\lambda_i} \right)^2; \quad (f) \quad \lambda_{\text{eff}}^{-1} = \sum_{i=1}^N \frac{3 n_i}{2 \lambda + \lambda_i}; \\ (g) \quad & \lambda_{\text{HS}} = \frac{1}{2} (\lambda_{\text{HS}}^+ + \lambda_{\text{HS}}^-); \end{aligned} \quad (2)$$

where:



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Fig. 2 Geometrical arrangement of layers, mineral grains, and pores assumed in different models for calculating mean bulk thermal conductivity of a composite medium:

arithmetic (λ_{ari}), harmonic (λ_{har}), geometric (λ_{geo}), and square root ($\lambda_{\sqrt{}}$) means, Hashin-Shtrikman upper and lower bounds, λ_{HS}^+ and λ_{HS}^- , respectively, and effective medium mean λ_{eff}

$$\lambda_{HS}^+ = \lambda_{\max} + \frac{A_{\max}}{1 - \alpha_{\max} A_{\max}},$$

$$\text{with : } A_{\max} = \sum_{i=1; \lambda_i \neq \lambda_{\max}}^N \frac{n_i}{\alpha_{\max} + 1/(\lambda_i - \lambda_{\max})},$$

$$\lambda_{\max} = \max(\lambda_1, \dots, \lambda_N) \quad \alpha_{\max} = \frac{1}{3\lambda_{\max}};$$

$$\lambda_{HS}^- = \lambda_{\min} + \frac{A_{\min}}{1 - \alpha_{\min} A_{\min}},$$

$$\text{with : } A_{\min} = \sum_{i=1; \lambda_i \neq \lambda_{\min}}^N \frac{n_i}{\alpha_{\min} + 1/(\lambda_i - \lambda_{\min})},$$

$$\lambda_{\min} = \min(\lambda_1, \dots, \lambda_N), \quad \alpha_{\min} = \frac{1}{3\lambda_{\min}}.$$

(3)

For a two-component system consisting of pore fluid and

solid rock with thermal conductivities λ_f and λ_s , respectively, Eq. 3 simplifies to (Hashin and Shtrikman 1962; Horai 1971):

$$\lambda_{HS}^+ = \lambda_s + \frac{\phi}{\frac{1}{\lambda_f - \lambda_s} + \frac{1-\phi}{3\lambda_s}}; \quad \lambda_{HS}^- = \lambda_f + \frac{1-\phi}{\frac{1}{\lambda_s - \lambda_f} + \frac{\phi}{3\lambda_f}}. \quad (4)$$

Generally, for a two-component system consisting of pore fluid and solid rock with thermal conductivities λ_f and λ_s , respectively, the implicit definition of λ_{eff} in Eq. 2f can be resolved:

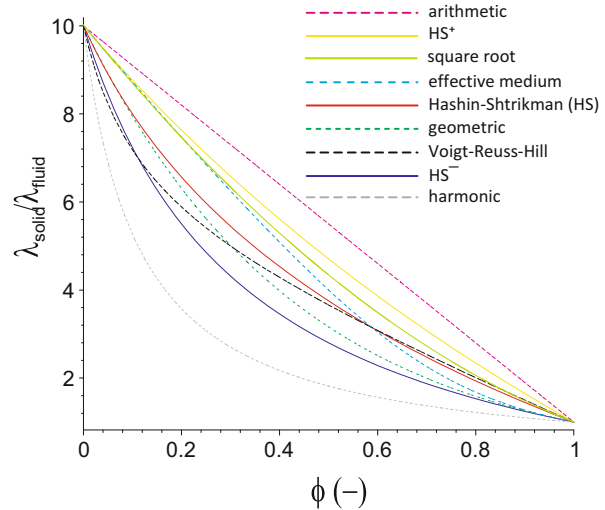
$$\lambda_{\text{eff}} = \frac{1}{4} \left\{ 3 \phi (\lambda_f - \lambda_s) + 2 \lambda_s - \lambda_f + \sqrt{9 \phi^2 \lambda_s^2 + 18 \phi \lambda_s \lambda_f - 18 \phi^2 \lambda_s \lambda_f - 12 \phi \lambda_s^2} \right. \\ \left. + \lambda_f^2 - 6 \phi \lambda_f^2 + 4 \lambda_s \lambda_f + 9 \phi^2 \lambda_f^2 + 4 \lambda_s^2} \right\}. \quad (5)$$

Figure 3 compares the variation of thermal conductivity of a two-component system with volume fraction for these different mixing laws defined by Eqs. 2 to 5. By and large, and in particular for such a two-component system, thermal conductivity of a multiphase rock determined according to these models can be ordered as:

$$\lambda_{\perp} = \lambda_{\text{har}} < \lambda_{HS}^- < \lambda_{\text{VRH}} < \lambda_{\text{geo}} < \lambda_{HS} < \lambda_{\text{eff}} \\ < \lambda_{\sqrt{}} < \lambda_{HS}^+ < \lambda_{\text{ari}} = \lambda_{\parallel}. \quad (6)$$

While only these nine models are presented and discussed here, various other mixing models are available which take into account additional factors, such as the shape of grains and voids. Several models assume spheroidal pores specifying the aspect ratio of the spheroids (Korvin 1978, 1982; Schulz 1981; Zimmerman 1984, 1989; Buntebarth and Schopper 1998; Popov et al. 2003). These models require information on the geometry of the internal rock structure and differ in the way of averaging over a representative elementary volume. Horai (1991) tested the results of predictions from several different mixing models on a remarkable data set in which porosity virtually varies from 0 % to 100 %. As can be expected, most of the models tested were valid only for certain porosity ranges. Only the two-phase models of Fricke-Zimmerman (Fricke 1924; Zimmerman 1989) and Schulz (1981) treating pores as spheroidal inclusions in a homogeneous and isotropic material are valid over the entire range of porosity. However, they require additional

information on the spheroids' aspect ratio or orientation. Given the typical ratios of rock and fluid conductivities we observe in nature, i.e., less than 10, most of the conductivity models tested work to within an accuracy of 10 % to 15 %.



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Fig. 3 Variation of thermal conductivity λ of a two-phase rock with porosity ϕ according to the means in Eqs. 2–5 for solid and fluid thermal conductivities of $\lambda_{\text{solid}} = 6 \text{ W m}^{-1} \text{ K}^{-1}$ and $\lambda_{\text{fluid}} = 0.6 \text{ W m}^{-1} \text{ K}^{-1}$, respectively: arithmetic (λ_{ari}); Hashin-Shtrikman upper bound (λ_{HS}^+); square root ($\lambda_{\sqrt{}}$); effective medium (λ_{eff}); Hashin-Shtrikman (λ_{HS}); geometric (λ_{geo}); Voigt-Reuss-Hill average (λ_{VRH}); Hashin-Shtrikman lower bound (λ_{HS}^-); harmonic (λ_{har})

Based on 1 325 individual measurements on sedimentary rocks and oceanic basalts in dry and saturated condition (λ_{dry} , λ_{sat}), Fig. 4 compares the variation of the ratio $\lambda_{\text{dry}}/\lambda_{\text{sat}}$ with porosity measured with that predicted by the arithmetic, upper Hashin-Shtrikman, square root, geometric, lower Hashin-Shtrikman, and harmonic mixing laws, λ_{ari} , λ_{HS}^+ , λ_{v} , λ_{geo} , λ_{HS}^- , and λ_{har} , respectively (Eqs. 2 to 4). It is evident that no single mixing law applies equally well to all rock types. But it appears that the geometric and square root model, on average, provide the best general fits to the data. It is also clear that the upper and lower Hashin-Shtrikman means λ_{HS}^+ and λ_{HS}^- , respectively, provide much tighter bounds than the arithmetic and harmonic means, λ_{ari} and λ_{har} , respectively.

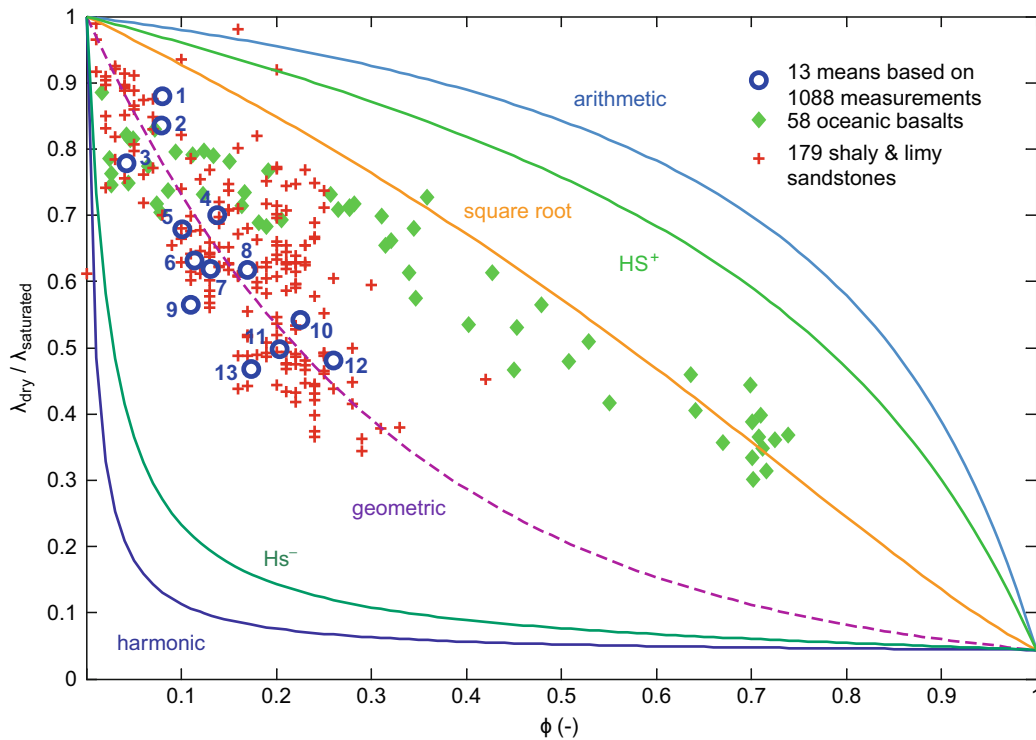
Correlations with other physical properties: Physical properties measured in well-logs can be used to infer estimates for in situ thermal conductivity. The approach is based on an extension of the mixing-model approach to the borehole scale: The volume fractions n_i of the N different mineral (or fluid) phases are either taken directly from induced gamma ray spectroscopy logs (Williams and Anderson

1990) or determined from a joint analysis of a suitable number J of geophysical logs, such as gamma ray (GR), sonic slowness (DT, the inverse of velocity), gamma density (DEN), and neutron porosity (NPHI) (e.g., Hartmann et al. 2005; Goutorbe et al. 2006). Let $\mathbf{x}$ and $\mathbf{b}$ be vectors composed of the N volume fractions n_i and the J theoretical log responses R^j with respect to the N different phases, respectively. Then, each element R^j of vector $\mathbf{b}$ is the cumulative response of the j -th log to all phases weighted by their volume fractions:

$$R^j = \sum_i^N n_i R_i^j, \quad \text{where: } \sum_i^N n_i = 1 \quad \text{and} \quad (7)$$

$$\mathbf{x} = \begin{pmatrix} n_1 \\ \vdots \\ n_N \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} R^1 \\ \vdots \\ R^J \end{pmatrix}.$$

Arranging the specific responses R^j of each log to the N rock phases as rows of a matrix $\mathbf{A}$:



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Fig. 4 Variation of thermal conductivity ratio $\lambda_{\text{dry}}/\lambda_{\text{sat}}$ (measured in dry and saturated condition) with porosity ϕ for different rock types. Numbered open circles represent means based on measurements on 1 058 sedimentary rock samples (Kobolev et al. 1990; Popov et al. 1995, 1999a; see also Clauser 2006): (1) 21 limestones; (2) 54 limestones; (3) 13 quartz sandstones; (4) 44 quartz siltstones; (5) 35 conglomerates; (6) 141 quartz sandstones; (7) 33 claystones; (8) 99 polymictic sandstones; (9) 30 quartz sandstones;

(10) 22 claystones; (11) 65 quartz siltstones; (12) 99 quartz siltstones; and (13) 241 quartz siltstones. Shown for comparison are data measured on 58 oceanic basalts (diamonds) and 179 shaly and limy sandstones (crosses). Curves labeled arithmetic, HS^+ , square root, geometric, HS^- , and harmonic, correspond to the arithmetic, upper Hashin-Shtrikman, square root, geometric, lower Hashin-Shtrikman, and harmonic mixing laws, λ_{ari} , λ_{HS}^+ , λ_{v} , λ_{geo} , λ_{HS}^- , and λ_{har} , respectively (Eqs. 2–4) (Clauser 2006)

$$\mathbf{A} = \begin{bmatrix} R_1^1 & \cdots & R_N^1 \\ \vdots & \ddots & \vdots \\ R_1^J & \cdots & R_N^J \end{bmatrix}, \quad (8)$$

the direct and inverse problems can be written as

$$\mathbf{Ax} = \mathbf{b} \text{ and } \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}, \quad (9)$$

respectively. Thus, in the direct problem, the log response vector $\mathbf{b}$ is computed from the volume fraction vector $\mathbf{x}$ and the specific log response matrix $\mathbf{A}$. In the inverse problem, in contrast, the volume fractions $\mathbf{x}$ are computed from the log responses $\mathbf{b}$ and the inverse of the specific log response matrix $\mathbf{A}^{-1}$. Thus, for N solid rock constituents, solving the inverse requires $N - 1$ logs. The void or fluid-filled space does not count here because it is the difference of one and the sum of the solid rock volume fractions. If more logs are available, making the problem overdetermined, the inverse problem must be solved in a least-squares sense. Once the volume fractions are known and assigned appropriate, representative thermal conductivities, an appropriate mixing model can be applied to compute rock thermal conductivity. Generally, the geometric and square root means, Eqs. 2d and 2e, have often turned out useful, but other mixing models may be more appropriate in specific cases.

Assigning representative thermal conductivities to the solid rock constituents is not trivial. Tabulated values of rock thermal conductivity should be used only if they characterize specimens taken from the logged formations. In all other cases, these formations or their outcrops need to be sampled and these specimens tested in the laboratory. If measurements are performed at ambient conditions, the values need to be corrected for the effect of temperature, and in some cases also for pressure.

In general, the effect of temperature is more pronounced than that of pressure. However, for greater depth and little or less consolidated rocks, it also needs to be accounted for. If commercial log interpretation software is used to perform the inversion, the theoretical log responses R_i^j with respect to the N different rock constituents are usually supplied by the software. Alternatively, values for the theoretical log responses R_i^j can be obtained from the literature (e.g., Crain 1986).

Variability: By combining this approach with a sufficient number of laboratory measurements, statistical moments of thermal conductivity, such as mean, standard deviation, median, quartiles, etc., can be derived for geological units (see, e.g., Jorand et al. 2015). These should comprise similar kinds of rock in a volume that can be reasonably characterized by the available number of samples and borehole logs. This way, they characterize the variability of thermal conductivity in rocks of the same kind and can be used as input in stochastic simulations of heat transport.

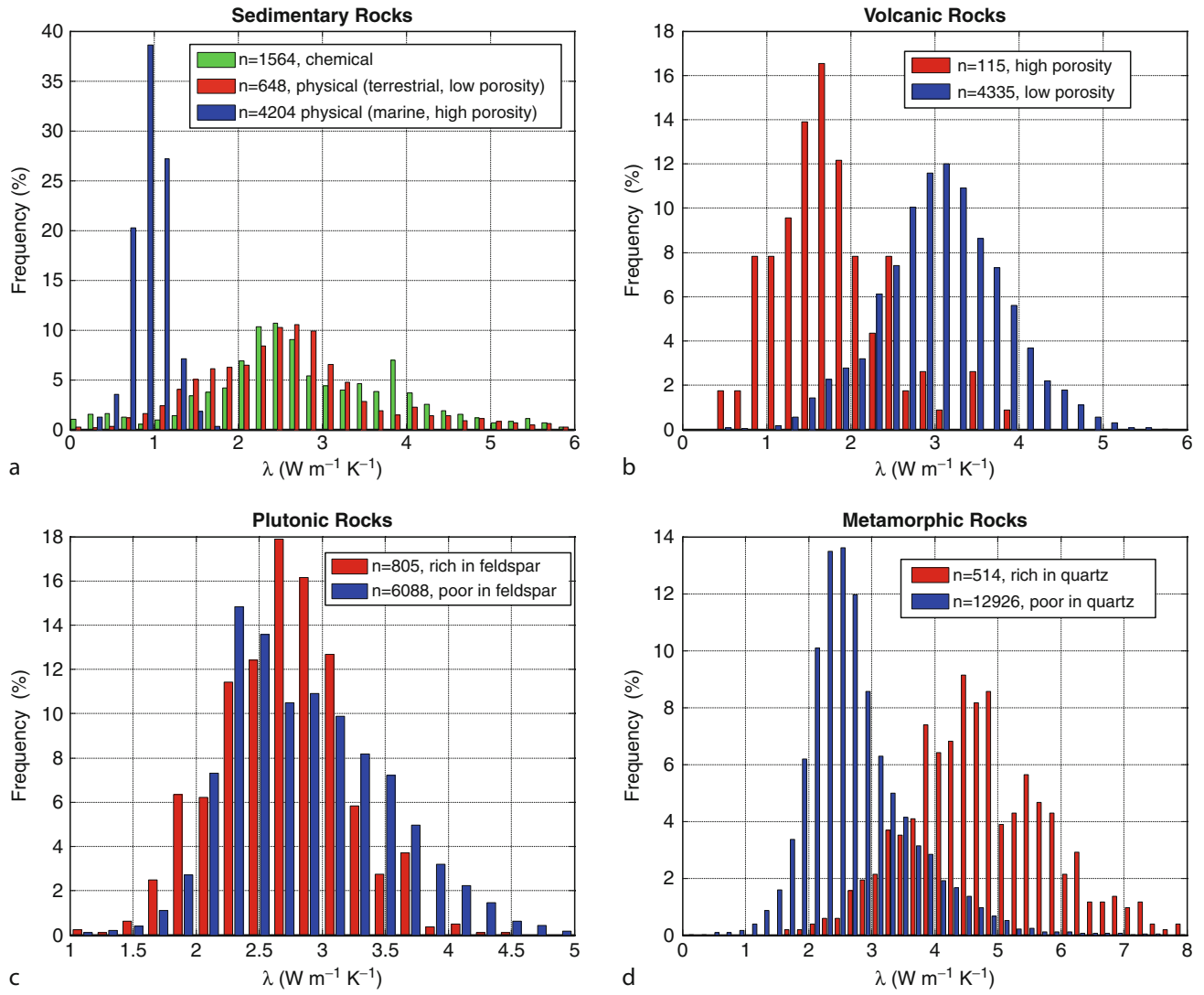
A complementary approach is based on the dependence of thermal conductivity on mineral content and porosity. If the volume fractions of each solid and fluid component in the rock are known, application of a suitable mixing model yields bulk thermal conductivity within an acceptable error bound. This approach, however, does not yield information on the anisotropy of thermal conductivity and, therefore, is particularly suited for isotropic rocks. In a recent study applying this approach to low-permeability, isotropic plutonic rocks, Fuchs et al. (2018) found the harmonic mean to be superior compared to others.

Thermal Conductivity of Minerals and Rocks

Thermal conductivity of minerals is much better constrained than that of rocks, due to the well-defined crystal structure and chemical formula for each mineral. Substantial collections of mineral thermal conductivities were compiled by Birch (1942), Clark (1966), Horai and Simmons (1969), Dreyer (1974); Roy et al. (1981), Čermák and Rybach (1982), Carmichael (1984), Popov et al. (1987), Diment and Pratt (1988); Somerton (1992), Clauser and Huenges (1995), and Clauser (2006).

Rocks are less well defined. In fact, rock type is a rather poor descriptor for physical properties as it rarely characterizes the dominating factors for any given property. With respect to thermal conductivity of any rock type, these comprise mineral content, porosity, pore fluid, saturation, and anisotropy. As these factors vary within each rock type, the variation of rock thermal conductivity was characterized in a statistical manner in Clauser (2006, 2009) according to the four main diagenetic classes of rocks: sedimentary, volcanic, plutonic, and metamorphic. This discussion is summarized here.

For *sedimentary rocks* (Fig. 5a), thermal conductivity is mainly controlled by porosity and type of sediment, i.e., clastic or chemical. For *volcanic rocks* (Fig. 5b), porosity is the controlling factor on thermal conductivity: Both mean and median of the high- and low-porosity histograms differ by nearly a factor of two, and the high-porosity distribution is skewed toward low conductivities. Plutonic and metamorphic rocks are generally much less porous. Here, the controlling factor is the dominant mineral phase. For *plutonic rocks* (Fig. 5c), the feldspar content determines the shape of the histogram: Rocks enriched in feldspar (i.e., >60 %) display a nearly symmetrical conductivity distribution about a lower mean conductivity than rocks with low feldspar content. In spite of these differences, the means and medians for both distributions are nearly identical. For *metamorphic rocks* (Fig. 5d), the quartz content controls thermal conductivity: Both mean and median of the distributions for high and low quartz content differ by nearly a factor of two, similar as for *volcanic rocks* (Fig. 5b) with regard to porosity. While the histogram for high quartz content rocks (mostly quartzites) is nearly symmetrical, the low quartz content histogram is strongly skewed toward low conductivities.



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Fig. 5 Histograms of thermal conductivity for (a) sedimentary, (b) volcanic, (c) plutonic, and (d) metamorphic rocks (Clauser 2009)

Radiative Thermal Conductivity

In addition to heat conduction, heat radiation emitted from hot bodies propagates through a sufficiently transparent and little absorbing medium as electromagnetic waves with velocity

$$c_m = \frac{c_0}{n} = \Lambda_m \nu_m, \quad (10)$$

where Λ_m and ν_m are radiation wavelength and frequency in the medium, and $n = c_0/c_m$ is the ratio of the speed of light in

vacuum and in the medium, i.e., the real part of the index of refraction (Table 1).

Radiation theory is based on Planck's distribution law defining the spectral radiance L_Λ or L_ν of a black body in the wavelength or frequency interval $d\Lambda_m$ or $d\nu_m$ in the medium in terms of the corresponding wavelength Λ or frequency ν in vacuum, respectively:

$$L_\Lambda(\Lambda_m, T) d\Lambda_m = \frac{2h c_0^2 n^2(\Lambda, T)}{\Lambda^5 (e^{h c_0/(\Lambda k T)} - 1)} d\Lambda^{-1}, \quad (W m^{-2} sr^{-1} m^{-1} = W m^{-3} sr^{-1}),$$

$$L_\nu(\nu_m, T) d\nu_m = \frac{2h \nu^3 n^2(\nu, T)}{c_0^2 (e^{h \nu/(k T)} - 1)} d\nu, \quad (W m^{-2} sr^{-1} Hz^{-1} = W s m^{-2} sr) \quad (11)$$

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 1 Refractive index n of selected substances

Substance	$n = c_0/c_m$ (at 589 nm)
Air	1.0003
Water (H ₂ O)	1.33
Magnesium fluoride (MgF ₂)	1.38
Calcium fluoride (CaF ₂)	1.43
Rock salt (NaCl)	1.54
Quartz (SiO ₂)	1.54
Glass	1.5–1.6
Carbon disulfide (CS ₂)	1.63
Corundum (Al ₂ O ₃)	1.76
Zircon (ZrSiO ₄)	1.92
Sulfur (S)	2.00
Zinc sulfide (ZnS)	2.37
Diamond	2.42
Carborundum (SiC)	2.55
Rutile (TiO ₂)	3.10
Galena (PbS) (at 590 nm)	3.90

where $h = 6.626\ 068\ 96(33) \times 10^{-34}$ J s is the Planck constant, $k = 1.380\ 650\ 4(24) \times 10^{-23}$ J K⁻¹ the Boltzmann constant, and $d\Lambda = n\ d\Lambda_m$. Radiance is the radiant energy flux emitted per unit solid angle in a given direction per unit projected area of a source. According to Wien's displacement law, the maximum of spectral radiance increases with temperature and decreases with wavelength (or increases with frequency) of radiation. Hofmeister (2005) estimated the variation of $n(v, T)$ with temperature and frequency and found both to be very small suggesting that n is approximately constant in the visible frequency range.

The intensity of radiation I is related to the incident intensity I_0 , radiation path x , and opacity ε , by $I = I_0 e^{-\varepsilon x}$, where opacity $\varepsilon = 1/\ell$ is the inverse of the mean free path ℓ of radiation which defines the average distance that photons travel before they are absorbed or scattered. In general, opacity is a function of the radiation wavelength. Opacity in an absorbing and scattering medium comprises contributions from both of these processes:

$$\varepsilon = \alpha + \varsigma, \quad (12)$$

where α and ς are absorption and scattering coefficients, respectively. The scattering coefficient ς is usually identified with the inverse of the grain size, suggesting orders of magnitude ranging from 10^2 m⁻¹ to 10^6 m⁻¹. The absorption coefficient α may be as large as $\alpha = 7\ 000$ m⁻¹, but values below $\alpha = 1\ 500$ m⁻¹ are reported for absorption coefficients in single crystal olivines with different proportions of forsterite (Mg₂SiO₄) and fayalite (Fe₂²⁺ SiO₄) (Fo₉₄Fa₆ – Fo₈₆Fa₁₄) in the two pass bands at 0.5 μ m and 0.5 μ m–6.0 μ m (e.g., Shankland et al. 1979; Clauser 1988; Hofmeister 2005). The width and level of these pass bands

depend critically on the iron content in the minerals. The complex index of refraction m of all materials is defined by $m = \sqrt{\varepsilon_r \mu_r} = n - i \cdot k$, where ε_r and μ_r are (complex) relative electrical permittivity and relative magnetic permeability, respectively, $n = c_0/c_m$ is the real and k the imaginary part of the index of refraction (sometimes also called extinction coefficient), and $i = \sqrt{-1}$ is the imaginary unit. The absorption coefficient α is related to the complex index of refraction k by:

$$\alpha = \frac{4\pi}{\Lambda} k, \quad (13)$$

where Λ is the radiation wavelength (Aronsson et al. 1970).

Radiation therefore becomes important for rocks with a larger free mean path of radiation corresponding to smaller values of opacity or absorption coefficient and scattering coefficient. This holds in particular at larger wavelength, in the infrared part of the absorption spectrum's transmission window between about 0.5 μ m and 6.0 μ m.

Radiated heat is diffused when photons emitted by mineral grains are scattered or absorbed by neighboring grains. If the mean free path of radiation ℓ is small compared to the distance to material discontinuities (like grain boundaries) and for moderate temperature gradients (i.e., no large anisotropy in the intensity of radiation), an effective thermal conductivity

$$\lambda_{\text{eff},ij} = \lambda_{p,ij} + \lambda_{r,ij} \quad (14)$$

replaces λ_{ij} in Eq. 1, where $\lambda_{p,ij}$ and $\lambda_{r,ij}$ are phonon and radiative thermal conductivities, respectively.

Rosseland (1930) and Clark (1957) defined the radiative thermal conductivity based on the temperature derivative of the spectral radiance of a blackbody. Following practice in engineering, Hofmeister (2005) accounts in this expression additionally for the emissivity $\eta \leq 1$ of a grainy material by applying the temperature derivative and the subsequent integration to the product of emissivity η and spectral radiance L_v , yielding

$$\lambda_r = \frac{4\pi}{3} \int_0^\infty \frac{1}{\varepsilon(v, T)} \frac{\partial(\eta(v, T) L_v(v, T))}{\partial T} dv. \quad (15)$$

Different practical expressions were derived depending on which parameters can be assumed as independent of temperature or frequency (see, e.g., Clark 1957; Shankland et al. 1979; Schärmeli 1982; Clauser 1988; Hofmeister 1999, 2005; Hofmeister et al. 2009). Based on grain size δ and attenuation,

$$A = \frac{\ln(I_0) - \ln(I_t)}{\delta} \approx \alpha, \quad (16)$$

where α is the absorption coefficient and I_0 and I_t are incident and transmitted intensities, respectively, Hofmeister (2005)

suggests the following expressions for emissivity and opacity:

$$\eta = 1 - e^{-\delta \cdot \alpha(v)} = 1 - e^{-(\ln(I_0) - \ln(I_r))}, \quad (17)$$

$$\varepsilon = \frac{1 + \delta \cdot \alpha}{\delta},$$

and estimates values for the product $\delta \cdot \alpha$ based on the interface reflectivity $R(\%)$ between two neighboring grains. Values for $\delta \cdot \alpha$ considered realistic for the mantle range between 5 and 10 (corresponding to a range for R between 0.07 % and 0.05 %) with a preferred value of 7 (i.e., $R = 0.01$ %).

The “gray body” approximation assumes opacity and emissivity as finite, constant, and independent of radiation wavelength. If the real part of the refractive index n and the spectral radiance L_λ or L_ν are also independent of temperature and wavelength, Eq. 15 simplifies into:

$$\lambda_r = \frac{4\pi\delta}{3} \int_0^\infty \frac{1 - e^{-\delta \cdot \alpha(v)}}{1 + \delta \cdot \alpha(v)} \frac{\partial(L_\nu(v, T))}{\partial T} dv = \frac{8\pi\delta n T^3 k^4}{3 c_0^2 h^3} \sum_{\text{lower}}^{\text{upper}} \int \frac{1 - e^{-\delta \cdot \alpha(v)}}{1 + \delta \cdot \alpha(v)} \left(\frac{n e^x x^4 + 2 T x^3}{e^x - 1} \frac{\partial n}{\partial T} \right) dx \quad (\text{W m}^{-1} \text{ K}^{-1}) \quad (19)$$

where $x = h \nu / (k T)$ and $dx = h d\nu / (k T)$, and the summation extends over the transparent regions above and below the strong absorption bands in the visible part of the spectrum. The second term of the sum in the integrand vanishes if $\partial n / \partial T \approx 0$, as suggested by Hofmeister (2005). Because α varies nonlinearly with frequency or wavelength, and because the cutoff frequencies in the integral in Eq. 19 depend on δ and α , Hofmeister (2005) evaluated the integral numerically for polynomials in T whose exponents ranged from 0 to 6. This yielded a relationship between radiative thermal conductivity and temperature differing from the purely cubic one in Eq. 18.

Variation with Temperature

Since the pioneering experiments of Eucken (1911), thermal conductivity of minerals and rocks is known to decrease with temperature, generally with its inverse. Eucken’s empirical result was corroborated theoretically by Peierls (1929) based on Debye’s (1914) theory of phonon scattering. Eucken (1911) observed in his experiments in contrast to crystals an increase in thermal conductivity of amorphous siliceous glass, a clear indication of radiative heat transfer.

The decrease is primarily due to the decrease of phonon (or lattice) thermal conductivity λ_p with temperature and to a smaller degree to thermal cracking. Since the thermal expansion coefficient increases with temperature (but differently for all minerals), differential expansion may create contact resistances between mineral grains. The effect of contact resistance is less pronounced in water-saturated than in dry rocks,

$$\lambda_r = \frac{16 \eta \sigma n^2 T^3}{3 \varepsilon}, \quad (18)$$

where $\sigma = 5.670 400(40) \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$ is the Stefan-Boltzmann constant. An order of magnitude of λ_r can be estimated from Eq. 18 by (i) assuming an emissivity of $\eta = 0.99$, (ii) neglecting scattering and identifying opacity with the absorption coefficient $1000 \text{ m}^{-1} < \alpha < 1 500 \text{ m}^{-1}$ of olivine (Fo92Fa8) at 1 700 K, and (iii) assuming a typical silicate value of $n = 1.6$ for the refractive index (Table 1). This yields a range for radiative thermal conductivity at 1 700 K of $2.5 \text{ W m}^{-1} \text{ K}^{-1} < \lambda_r < 3.8 \text{ W m}^{-1} \text{ K}^{-1}$.

Inserting Eqs. 16 and 17, and the temperature derivative of the spectral radiance (11) as provided by Shankland et al. (1979) into Eq. 15 yields radiative thermal conductivity as:

the condition in which most rocks are tested at elevated temperatures. For single-mineral aggregates, a linear relationship between temperature and thermal resistivity λ^{-1} discriminates between contributions which depend on temperature T and others which don’t, such as micro-cracks, grain boundaries, shape, and orientation of crystals and their fragments:

$$\lambda^{-1}(T) = c_1 + c_2 T, \quad (20)$$

where λ is in $\text{W m}^{-1} \text{ K}^{-1}$ and T is in K. By measuring thermal conductivity λ and plotting its inverse (thermal resistivity λ^{-1}) versus temperature, constants c_1 and c_2 are obtained from intercept and slope of a linear regression. Table 2 provides values for the constants c_1 and c_2 in Eq. 20 which may be used to infer the temperature dependence of thermal resistivity for some single-mineral aggregates (Clark 1969).

Based on measurements on 113 samples of metamorphic rocks from the KTB research borehole in Germany in the temperature range 50 °C–200 °C, Buntebarth (1991) determined mean values for the constants c_1 and c_2 in Eq. 20 for gneissic and metabasitic rocks. The arithmetic means determined from measurements on 66 gneiss samples are $\bar{c}_1 = 0.16(3) \text{ W}^{-1} \text{ m K}$ and $\bar{c}_2 = 0.37(14) \times 10^{-3} \text{ W}^{-1} \text{ m}$. The corresponding means determined from measurements on 36 metabasite samples are $\bar{c}_1 = 0.33(3) \text{ W}^{-1} \text{ m K}$ and $\bar{c}_2 = 0.22(14) \times 10^{-3} \text{ W}^{-1} \text{ m}$.

In contrast to phonon conductivity λ_p , the radiative contribution λ_r to thermal conductivity generally increases with

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 2 Values of c_1 and c_2 in Eq. 20 for single-mineral aggregates. (Data: Clark (1969))

Mineral	T (°C)	$c_1 \times 10^3$ (W ⁻¹ m K)	$c_2 \times 10^3$ (W ⁻¹ m)
Halite, NaCl	0–400	–52.55	0.788
Periclase, MgO	100–800	–21.50	0.127
Corundum, Al ₂ O ₃	100–800	–28.66	0.155
Quartz, SiO ₂ ^a	100–400	62.10	0.387
Spinel, MgAl ₂ O ₄	100–1000	19.11	0.122
Zircon, ZrSiO ₄	100–800	131.37	0.093
Forsterite, Mg ₂ SiO ₄	100–600	85.98	0.282
Enstatite, Ferrosilite, (Mg ₂ ,Fe ₂)SiO ₃	100–300	200.63	0.222

^aSingle SiO₂ crystal, heat flow $\perp$ to optical axis

the cube of temperature (see above). Thus, measurements of thermal conductivity as function of temperature generally first decrease with temperature until, from about 1000 °C–1200 °C onward, the radiative component balances and sometimes even reverses this decrease.

The temperature dependence of rock thermal conductivity was characterized in a statistical manner in Clauser (2006, 2009) according to the four main diagenetic classes of rocks: sedimentary, volcanic, plutonic, and metamorphic. This discussion is summarized here (Fig. 6).

For *sedimentary rocks* (Fig. 6a) up to 300 °C, there is a reduction by nearly a factor of two, both for physical and chemical sediments. Above 300 °C, while the decrease in thermal conductivity is less, it is more pronounced for chemical sediments than for physical sediments. However, there are very few data for this temperature range, which makes this last observation statistically weak. Above 300 °C, the mean thermal conductivity of sediments varies between 1.0 W m⁻¹ K⁻¹ and 1.5 W m⁻¹ K⁻¹.

Volcanic rocks (Fig. 6b) vary quite differently with temperature depending on their opacity, i.e., on how well they transmit thermal energy by radiation. Due to this additional “radiative thermal conductivity,” volcanic glasses and rocks with small iron content show an increase in thermal conductivity for temperatures above 800 °C–1 000 °C (e.g., Clauser 1988; Hofmeister et al. 2009). In contrast, thermal conductivity of opaque rocks, particularly those with a high iron content, decreases with temperature. An inversion of this trend is indicated by few available high-temperature measurements (above 1 300 °C) but with too few measurements for a statistical appraisal. At about 1 000 °C, thermal conductivity of these rocks is only about 50 % of their room temperature value. Again, there are few data points above 700 °C.

Plutonic rocks (Fig. 6c) show no strong radiative contribution. At temperatures above 600 °C, thermal conductivity decreases only very little. However, in these rocks, the variation of thermal conductivity with temperature depends on their feldspar content. For rocks enriched in feldspar, thermal conductivity up to 300 °C decreases little, while for those poor in feldspar, the decrease is stronger, becoming more

gentle above 300 °C, decreasing further by 20 % over the next 1 000 K. The different behavior of rocks with high feldspar content is due to the increase in thermal conductivity with temperature of some plagioclase feldspars (e.g., Höfer and Schilling 2002; Petrunin et al. 2004) which compensates the decrease in thermal conductivity with temperature observed for most other minerals and rocks. Other notable exceptions are fused silica as well as volcanic and silica glasses.

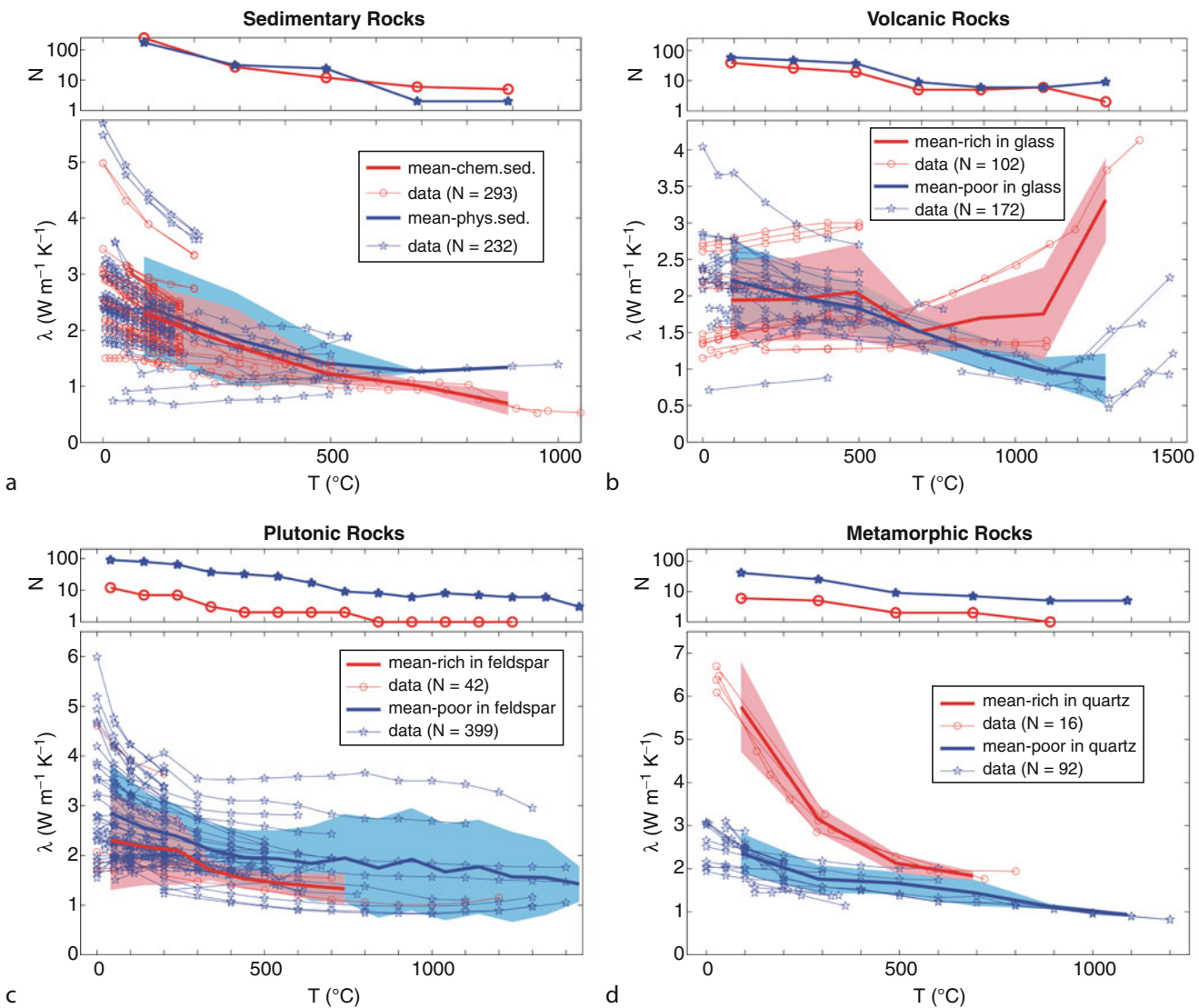
For *metamorphic rocks* (Fig. 6d), the decrease of thermal conductivity with temperature depends on the content in the dominant mineral phase, similar as for plutonic rocks. For quartzites, the decrease is strong, by nearly a factor of three up to a temperature of about 500 °C with only a very mild further decrease above this temperature. For rocks poor in quartz, the decrease in conductivity is not quite as strong, amounting, up to 300 °C, to about one-third of their room temperature value. Then it remains roughly constant up to 500 °C and decreases again to about one-third of their room temperature value up to 750 °C.

In summary, for moderate temperatures, thermal conductivity of rocks is well described by a linear relationship with the inverse of temperature, similar as in Eq. 20. For this temperature range, several approaches are available for inferring thermal conductivity at elevated temperatures. Based on the analysis of available tabulated data of thermal conductivity as function of temperature, Zoth and Hänel (1988) suggested the following form:

$$\lambda(T) = A + \frac{B}{350 + T}, \quad 0^\circ\text{C} \leq T \leq 800^\circ\text{C}, \quad (21)$$

where average values of the coefficients A and B for different rock types are given in Table 3.

Linear relationships between temperature and thermal resistivity, such as Eqs. 20 and 21, discriminate between temperature-dependent contributions and other factors, which are independent of temperature, such as micro-cracks, grain boundaries, pore volume, mineralogical composition, shape, and orientation of crystals and their fragments.



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Fig. 6 Variation of thermal conductivity with temperature for (a) sedimentary, (b) volcanic, (c) plutonic, and

(d) metamorphic rocks. Color shading indicates a range defined by plus and minus one standard deviation, and N is the number of data at each temperature (Clauser 2009)

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 3 Values for constants A and B in Eq. 21 for different rock types (Zoth and Hänel 1988)

Rock type	$T (^{\circ}\text{C})$	A ($\text{W m}^{-1} \text{K}^{-1}$)	B (W m^{-1})
1. Rock salt	-20 - 0	-2.11	2 960
2. Limestones	0-500	0.13	1073
3. Metamorphic rocks	0-1 200	0.75	705
4. Acidic rocks	0-1 400	0.64	807
5. Basic rocks	50-1 100	1.18	474
6. Ultrabasic rocks	20-1 400	0.73	1 293
7. Rock types (2)-(5)	0-800	0.70	770

Sass et al. (1992) and Vosteen and Schellschmidt (2003) distinguish between the effects of composition and temperature. They propose a general empirical relation for $\lambda(T)$, the thermal conductivity in $\text{W m}^{-1} \text{K}^{-1}$ at temperature T in $^{\circ}\text{C}$, as a function of λ_0 , the thermal conductivity at 0°C :

$$\lambda(T) = \frac{\lambda_0}{a + \left(b - \frac{c}{\lambda_0}\right)T} \quad \text{or} \quad \frac{\lambda_0}{\lambda(T)} = \underbrace{a}_{\text{intercept}} + \underbrace{\left(b - \frac{c}{\lambda_0}\right)}_{\text{slope}} T.$$

(22)

For different rock types, the slopes and intercepts of this equation can be determined from linear regressions of Eq. 22 yielding a mean intercept $\bar{a}$ and its uncertainty Δa .

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 4 Coefficients a , b , and c in Eq. 22 and associated uncertainties Δa and σ_b , σ_c ; Δa is the error of the mean intercept $\bar{a}$ for all rock types of the linear regressions of the

normalized thermal resistance $\lambda_0/\lambda(T)$ as a function of temperature T ; σ_b and σ_c are the errors defined by the linear regression of the slopes ($b-c/\lambda_0$) as a function of the thermal resistance $1/\lambda_0$ (see Eq. 22)

Rock type	$\bar{a}$ (—)	Δa (%)	$b \times 10^3$ (K ⁻¹)	$\sigma_b \times 10^3$ (K ⁻¹)	$c \times 10^3$ (W m ⁻¹ K ⁻²)	$\sigma_c \times 10^3$ (W m ⁻¹ K ⁻²)	T (°C)	References
Basement rocks I (from felsic gneiss to amphibolite)	1.007	—	3.6	—	7.2	—	0–250	Sass et al. (1992)
Basement rocks II (magmatic and metamorphic)	0.99	1	3.0	1.5	4.2	0.6	0–500	Vosteen and Schellschmidt (2003)
Sediments	0.99	1	3.4	0.6	3.9	1.4	0–300	Vosteen and Schellschmidt (2003)

Coefficients b and c and associated uncertainties σ_b and σ_c are determined from a second linear regression of the different slopes ($b - c/\lambda_0$) as a function of $1/\lambda_0$ (Table 4).

Since thermal conductivity is usually measured at room temperature, λ_0 is expressed as a function of λ_{25} , the room temperature thermal conductivity, by Sass et al. (1992) for crystalline rocks (felsic gneiss to amphibolite) as:

$$\lambda_0 = \lambda_{25} \left(1.007 + 25 \left(0.0037 - \frac{0.0074}{\lambda_{25}} \right) \right). \quad (23)$$

Vosteen and Schellschmidt (2003) find for magmatic and metamorphic rocks:

$$\lambda_0 = 0.53 \lambda_{25} + 0.5 \sqrt{1.13 \lambda_{25}^2 - 0.42 \lambda_{25}}, \quad (24)$$

and for sedimentary rocks:

$$\lambda_0 = 0.54 \lambda_{25} + 0.5 \sqrt{1.16 \lambda_{25}^2 - 0.39 \lambda_{25}}. \quad (25)$$

Hofmeister (1999) provided a detailed analysis of heat transport based on an analysis of phonon lifetimes obtained from infrared reflectivity. It accounts for the variation of the phonon contribution λ_p to thermal conductivity with both temperature T and pressure p as well as for the pressure-dependent radiative contribution λ_r . It allows approximating the thermal conductivity at conditions in the Earth's mantle if:

$$\lambda_r = \begin{cases} 0.01753 - 1.0365 \times 10^{-4} T + 2.2451 \times 10^{-7} T^2 - 3.407 \times 10^{-11} T^3 & \text{(a)} \\ 8.5 \times 10^{-11} T^3 & \text{(b)} \end{cases}, \quad (27)$$

where T is in Kelvin. These two relationships are shown in Fig. 7.

Figure 1 in the companion entry ► “Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent

(1) $K'_0 = dK_T/dp$, the pressure derivative of the isothermal bulk modulus K_T is constant; (2) the variations of the bulk modulus with temperature and pressure are mutually independent; (3) the pressure derivative of the thermodynamic Grüneisen parameter γ (Eq. 12 or 13 in the entry ► “Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat”, this volume) is constant: $d\gamma/dp = f$. For mantle material, γ varies from 1 to 1.4, K'_0 from 4 to 5, and $f \approx 0$ vanishes. According to Hofmeister (1999) and within the uncertainty of these parameters, thermal conductivity in the mantle is:

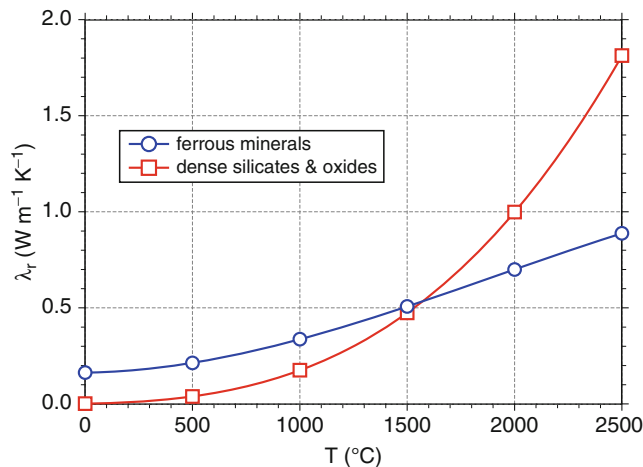
$$\lambda(T, p) = \lambda_r + \lambda(298K, 101.33 \text{ kPa}) \cdot \left(\frac{298}{T} \right)^a \cdot \frac{1 + K'_0 p}{K_T} \cdot e^{-\frac{(4\gamma+1/3)}{298} \int_{298}^T \alpha(T') dT'}, \quad (26)$$

where T is absolute temperature, $\lambda(298 \text{ K}, 101.33 \text{ kPa})$ is thermal conductivity at room temperature and atmospheric pressure, $\alpha(T)$ is volume coefficient of thermal expansion as a function of temperature, and the exponent a is the fitting parameter. The radiative contribution λ_r in (26) may be approximated by Eqs. 15, 18, or 19. Hofmeister (1999) provides alternative expressions for λ_r (in W m⁻¹ K⁻¹) for ferrous minerals (Eq. 27a) and dense silicates and oxides (Eq. 27b):

Heat” (this volume) shows the variation of phonon thermal conductivity λ_p with temperature in an average crust of density $\rho = 2700 \text{ kg m}^{-3}$ molar mass of $0.22178 \text{ kg mol}^{-1}$. It is derived from $\lambda = \kappa \rho c$ as the product of thermal capacity ρc_p

and thermal diffusivity κ calculated according to Eq. 20 (in the entry ► “Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat”, this volume; Whittington et al. 2009). Due to the balancing increase of specific heat capacity with temperature, the decrease in thermal conductivity with temperature is less than in thermal diffusivity by a factor of about 1.3. Assuming a constant density throughout the crust, this implies that the increase and decrease in density due to the increase in pressure and temperature, respectively, partly cancel each other and that these changes are small compared to those of specific heat capacity and thermal diffusivity.

For mantle minerals, such as olivine and its high-pressure polymorphs, the β - and γ -spinel wadsleyite and ringwoodite, respectively, Xu et al. (2004) fitted phonon thermal conductivity measured to 1 373 K and 20 GPa to an exponential equation in temperature yielding values for the exponent between $-0.406(35)$ and $-0.537(11)$ suggesting that fitted phonon thermal conductivity varies with $T^{-1/2}$ (Table 5).



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Fig. 7 Variation of radiative thermal conductivity λ_r of ferrous minerals, and dense silicates and oxides with temperature according to Eq. 27 (Hofmeister 1999)

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 5 Reference values λ_{298} and κ_{298} at 298 K, pressure coefficients a and a' , and fitting functions for

Mineral	p (GPa)	$\lambda = \lambda_{298} (298/T)^{1/2} (1 + a \cdot p)$		$\kappa = \kappa_{298} (298/T)^n (1 + a' \cdot p)$		
		λ_{298} (W m ⁻¹ K ⁻¹)	a (GPa ⁻¹)	$\kappa_{298} \times 10^6$ (m ² s ⁻¹)	n	a' (GPa ⁻¹)
Olivine	4–10	4.13(11)	0.032(3)	1.31(5)	0.681(22)	0.036(4)
	4	4.49(4)	–	1.29(5)	0.563(35)	–
	7	5.19(4)	–	1.74(4)	0.720(26)	–
	10	5.56(4)	–	1.84(5)	0.723(31)	–
Wadsleyite	14	8.10(4)	0.023	2.55(3)	0.721(13)	–
Ringwoodite	20	9.54(5)	0.022	3.09(4)	0.793(17)	–

Variation with Pressure

The effect of pressure on phonon thermal conductivity λ_p is different in two distinct pressure ranges. First, fractures and micro-cracks (developed during stress release after sampling) begin to close with increasing pressure. This reduces thermal contact resistance as well as porosity which is usually filled with a low-conductivity fluid. This process ends when a pressure of about 15 MPa is reached. A compilation of measurements on various sedimentary, volcanic, plutonic, and metamorphic rocks (Clauser and Huenges 1995) indicates that this effect accounts for an increase of about 20% relative to thermal conductivity at atmospheric pressure. A further pressure increase to 40 MPa does not affect thermal conductivity significantly. If pressure is increased further, however, a second process becomes effective, the reduction of intrinsic porosity, i.e., voids which are not created by stress release. For granite and for metamorphic rocks, data indicate an increase of thermal conductivity by about 10% within the pressure range 50 MPa–500 MPa.

For mantle minerals, such as olivine and its high-pressure polymorphs, the β - and γ -spinel wadsleyite and ringwoodite, respectively, Xu et al. (2004) determined values for the pressure coefficient between 0.022 GPa^{-1} and 0.032 GPa^{-1} for phonon thermal conductivity measured to 1373 K and 20 GPa (Table 5). A table of numerical values for the pressure derivatives of phonon thermal conductivity measured by a variety of authors was compiled by Hofmeister et al. (2009). Most values of the pressure coefficient fall into the range 0.04 GPa^{-1} – 0.36 GPa^{-1} , exceeded only by values of 0.69 GPa^{-1} and 0.5 GPa^{-1} for sulfur and quartz_{⊥c} (measured perpendicular to the optical c -axis), respectively. Osako et al. (2004) fitted thermal conductivity measured to 1100 K and 8.3 GPa on isotropic, single-crystal garnet and anisotropic olivine (Fo₉₃Fa₇; in three crystallographic directions) as upper and lower mantle constituents, respectively, to a linear equation in pressure (Table 6). A pressure dependence of garnet and olivine was found to be on the order of 4% per GPa to 5% per GPa and 3% per GPa to 4% per GPa, respectively.

In contrast, radiative thermal conductivity λ_r was found to vary much less with pressure than temperature (Clark 1957).

the variation of thermal conductivity λ and thermal diffusivity κ of lower mantle minerals with absolute temperature T and pressure p (Xu et al. 2004)

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 6 Coefficients and fitting functions for the variation thermal conductivity λ and thermal diffusivity κ of garnet and olivine with absolute temperature T and pressure p (Osako et al. 2004)

Garnet				Olivine				
$\lambda = C_0 + C_1/T$		$\lambda = A_0 + A_1 \cdot p$			$\lambda = C_0 + C_1/T$		$\lambda = B_0 \exp(B_1 \cdot p)$	
C_0 (W m ⁻¹ K ⁻¹)	C_1 (W m ⁻¹)	A_0 (W m ⁻¹ K ⁻¹)	A_1 (W m ⁻¹ K ⁻¹ GPa ⁻¹)		C_0 (W m ⁻¹ K ⁻¹)	C_1 (W m ⁻¹)	B_0 (W m ⁻¹ K ⁻¹)	B_1 (GPa ⁻¹)
2.01(8)	704(43)	3.48(33)	0.160(26)	[100]	1.91(28)	2 088(163)	6.61(13)	0.038(5)
				[010]	0.84(36)	1 377(157)	3.98(15)	0.042(5)
				[001]	2.08(38)	1 731(86)	5.91(25)	0.034(5)
$\kappa = c_0 + c_1/T$		$\kappa = a_0 + a_1 \cdot P$			$\kappa = c_0 + c_1/T$		$\kappa = b_0 \cdot e^{b_1 \cdot P}$	
$c_0 \times 10^6$ (m ² s ⁻¹)	$c_1 \times 10^6$ (m ² s ⁻¹ K)	$a_0 \times 10^6$ (m ² s ⁻¹)	$a_1 \times 10^6$ (m ² s ⁻¹ GPa ⁻¹)		$c_0 \times 10^6$ (m ² s ⁻¹)	$c_1 \times 10^6$ (m ² s ⁻¹ K)	$b_0 \times 10^6$ (m ² s ⁻¹)	b_1 (GPa ⁻¹)
0.29(6)	374(31)	1.19(6)	0.046(1)	[100]	−0.06(11)	938(46)	2.50(4)	0.033(5)
				[010]	−0.13(8)	626(45)	1.52(6)	0.040(7)
				[001]	−0.03(17)	832(98)	2.16(14)	0.035(3)

In particular, this holds once the spectral radiance overlaps the infrared pass band in the absorption spectrum, i.e., at temperatures above 1900 K (Hofmeister 2005).

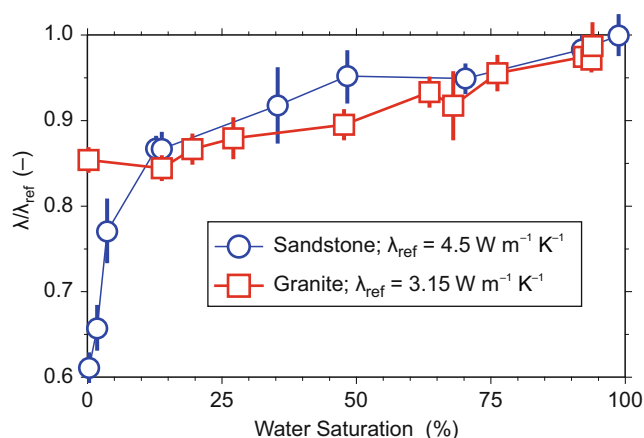
Variation with Other Factors

Apart from temperature and pressure, thermal conductivity also varies with porosity, pore fluid, saturation, dominant mineral phase, and anisotropy. These effects are summarized here from the detailed discussion in Clauser (2006).

For large *porosity* (i.e., $\phi > 1\%$), thermal conductivity of the saturating fluid affects significantly the bulk rock thermal conductivity. The influence varies with the thermal conductivity of the saturants, e.g., water, oil, natural gas, air. The resulting bulk thermal conductivity can be estimated from a suitable mixing model, e.g., Eqs. 2–4.

The effect of partial *saturation* is different for porous or fractured rocks. In porous rocks, porosity comprises both bulk pore space and bottlenecks formed by the contact between individual grains. Dry bottlenecks act as thermal contact resistances between grains, while the bulk pore volume contributes proportionally to the effective rock thermal conductivity. In fractured rocks, in contrast, there are no bottlenecks between grains as in porous rocks, and the small void volume in the fractures corresponds to the bulk pores space of porous rocks.

Figure 8 illustrates these two effects for a water-saturated medium-porosity sandstone and a low-porosity granite: Starting completely dry with an unsaturated conductivity of about 60 % of the saturated value, a level of 85 % is reached for the sandstone at about 10 % saturation. The 15 % conductivity residual is then spread almost linearly over the remaining 90 % of saturation. Physically this observation indicates that the filling of intergranular bottlenecks, accounting for only about 10 %–15 % of the total porosity, significantly reduces the contact resistances between the individual grains. The replacement of low conductivity air by the more conductive fluid in the major part of the pore volume accounts



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Fig. 8 Variation of thermal conductivity with partial saturation for a sandstone (circles; $\phi = 18\%$) and granite (squares; $\phi = 1\%$) saturated with water and standard deviations (bars); values normalized by reference thermal conductivities shown in legend (Clauser 2006 based on data of Reibelt 1991)

for the second effect. If only fractures contribute to the total porosity, as in the granite, there are no bottlenecks and only the second effect is observed: Starting completely dry with an unsaturated conductivity of about 85 % of the saturated conductivity, a quasi linear increase is observed due to the replacement of low conductivity air until the 100 % level is reached for complete saturation.

Anisotropy of sedimentary and metamorphic rocks is due to the conditions during their formation. Anisotropy occurs on several scales: (1) On the microscopic scale, many minerals are intrinsically anisotropic; (2) on the laboratory scale, thermal conductivity of many rocks is also anisotropic. However, even if rocks are composed of anisotropic minerals, random orientation of the crystals within the rock may render the rock's bulk thermal conductivity isotropic on a macroscopic scale; (3) on a still larger scale, if rocks

are exposed to folding, orogenic, or other tectonic processes, thermal conductivity of the resulting rock formation may be anisotropic.

As a result, thermal conductivity parallel to the direction of layering or foliation, $\lambda_{\parallel}$, is greater than thermal conductivity in the perpendicular direction, $\lambda_{\perp}$. The anisotropy factor, the ratio $\lambda_{\parallel}/\lambda_{\perp}$, generally falls into the range 0.9–3, with most values between 1 and 2 (e.g., Clauser and Huenges 1995; Popov and Mandel 1998; Popov et al. 1999a, b; Clauser 2006; Davis et al. 2007). For sedimentary rocks, a general trend has been reported of decreasing $\lambda_{\perp}$ with anisotropy factor $\lambda_{\parallel}/\lambda_{\perp}$, but no such trend was identified for metamorphic rocks (Clauser 2006).

Thermal Diffusivity

Thermal diffusivity is required in the analysis of transient heat transfer. A sizable compilation of room temperature data of phonon thermal diffusivity κ_p measured on various mantle minerals is given by Hofmeister et al. (2009) from measurements by various groups.

If both conductivity and thermal capacity are known, thermal diffusivity κ can be calculated from $\kappa = \lambda/(\rho c)$. As for steady-state thermal conduction, transient heat diffusion in most of the Earth's crust and mantle is caused by scattering of quantized lattice vibrations, the phonons, and by diffusive (as opposed to ballistic) radiation of photons. These two processes are described by phonon thermal conductivity λ_p and radiative thermal conductivity λ_r , respectively, the sum of which is often termed effective thermal conductivity, λ_{eff} . As thermal diffusivity is the ratio of thermal conductivity and thermal capacity, it is also influenced by the variation of density and specific heat capacity. This is of particular interest with respect to the variation with temperature.

Measuring Techniques

All of the transient laboratory methods used to determine thermal conductivity are useful to determine thermal diffusivity as well. More recently, heat pulse (Schilling 1999; Höfer and Schilling 2002; Gibert et al. 2003) and laser flash methods (Parker et al. 1961; Blumm and Lemarchand 2002; Hofmeister 2006) have been used for measurements at high temperature. Compared to other methods, their advantage lies in a reduction or even complete absence of physical contacts between samples on the one hand and temperature sensors and heat sources on the other hand. While most methods measure the effective diffusivity comprising contributions from phonon conduction and diffused heat radiation, the laser flash method yields the diffusive component without contributions from heat radiation. This difference becomes important particularly at high temperatures (for a critical assessment, see, e.g., Hofmeister et al. 2009).

Variation with Temperature

Thermal diffusivity κ of rocks varies even more strongly with temperature than thermal conductivity λ . This is caused by the opposite behavior of thermal conductivity and thermal capacity (ρc) with respect to temperature. Because of several self-compensating factors, thermal capacity (ρc) with few exceptions generally varies within $\pm 20\%$ of $2.3 \text{ MJ m}^{-3} \text{ K}^{-1}$ for the great majority of minerals and rocks (Beck 1988). This is confirmed by a linear regression of thermal diffusivity on thermal conductivity which was measured on a suite of meta-sedimentary, volcanic, magmatic, and metamorphic rocks together with density and specific heat capacity (Mottaghy et al. 2005):

$$\kappa = \frac{\lambda}{\rho c} = \frac{\lambda}{2.3} \\ = 0.44 \lambda \quad (\lambda \text{ in } \text{W m}^{-1} \text{ K}^{-1}; \kappa \text{ in } 10^{-6} \text{ m}^2 \text{ s}^{-1}). \quad (28)$$

A linear regression of thermal capacity as a function of temperature yields also a linear relationship. This allows to determine thermal diffusivity $\kappa(T)$ at any temperature, based only on the known variation of thermal conductivity $\lambda(T)$ with temperature (Mottaghy et al. 2005):

$$\kappa(T) = \frac{\lambda(T)}{2.1340 + 0.0044 T} \quad (T \text{ in } ^\circ\text{C}; \kappa \text{ in } 10^{-6} \text{ m}^2 \text{ s}^{-1}). \quad (29)$$

Thus, thermal diffusivity can be derived from thermal conductivity and vice versa. For the suite of rocks studied by Mottaghy et al. (2005), thermal conductivity decreased by 4 %–7 % in the range 1–100 °C while thermal diffusivity decreased by 18 %–22 %.

Thermal diffusivity was measured at temperatures of up to 550 °C by Ray et al. (2006) on a suite of 16 samples comprising Archean granulitic rocks which are considered a major component of the middle and lower continental crust. Phonon scattering was found to dominate heat transport with radiative diffusion of photons setting in at 450 °C for most, but not all rocks. Based on their measurements, Ray et al. (2006) propose an equation by which thermal diffusivity $\kappa(T)$ at elevated temperature below 450 °C can be derived from room temperature values κ_{rt} :

$$\kappa(T) = 0.7 + 144 \frac{\kappa_{\text{rt}} - 0.7}{T - 150} \quad (T \text{ in K}; \kappa \text{ in } 10^{-6} \text{ m}^2 \text{ s}^{-1}). \quad (30)$$

They also proposed an additional cubic term T^3 in Eq. 30 by which the radiative contribution above 450 °C is fitted. However, this assumes a similar variation with temperature of thermal conductivity and thermal diffusivity and neglects the additional variation of thermal capacity. This, in fact, makes

thermal diffusivity vary stronger with temperature than thermal conductivity, which is why the T^3 -term is omitted here. Whittington et al. (2009) measured the phonon component κ_p of thermal diffusivity at temperatures of up to 1260 K on garnet schist, leucogranite, and welded rhyolitic ash-flow tuff using laser flash analysis. This characterizes purely the phonon heat transfer component without any radiative contribution, in contrast to the values discussed above. Below and above the transition between α - and β -quartz at 846 K ($\sim 573^\circ\text{C}$), the data are reasonably fitted by:

$$\kappa_p(T) = \begin{cases} -0.062 + \frac{567.3}{T}; & T \leq 846 \text{ K} \\ 0.732 - 0.000135T; & T > 846 \text{ K} \end{cases} \quad (T \text{ in K}; \kappa \text{ in } 10^{-6} \text{ m}^2\text{s}^{-1}) \quad (31)$$

assuming an average molar mass of $0.22178 \text{ kg mol}^{-1}$ and an average density of 2700 kg m^{-3} for the crust (see Fig. 1 in the entry ► “Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat”, this volume). The leucogranite and rhyolite samples were homogeneous and isotropic. The schist was anisotropic owing to alternating mica- and quartz-rich layers, which required testing in the directions parallel and perpendicular to foliation.

For mantle minerals, such as olivine and its high-pressure polymorphs, the β - and γ -spinel wadsleyite and ringwoodite, respectively, Xu et al. (2004) fitted phonon thermal diffusivity measured to 1373 K and 20 GPa to an exponential equation in temperature yielding values for the exponent between

$-0.563(35)$ and $-0.793(17)$ suggesting that fitted phonon thermal conductivity varies with temperature as $T^{-1/2}$ to T^{-1} (Table 5).

Pertermann and Hofmeister (2006) measured thermal diffusivity on oriented single crystals and polycrystalline samples of olivine-group minerals with the laser-flash method at temperatures of up to about 1500°C . They fitted the data by a second-order polynomial in T :

$$\kappa_p(T) = a + bT^{-1} + cT^{-2} \quad (T \text{ in K}). \quad (32)$$

Values for the fitting coefficients a , b , and c are shown in Table 7.

Osako et al. (2004) fitted thermal diffusivity measured to 1100 K and 8.3 GPa on isotropic, single-crystal garnet and anisotropic olivine ($\text{Fo}_{93}\text{Fa}_7$; in three crystallographic directions) as upper and lower mantle constituents, respectively, to a linear equation in inverse temperature (Table 6). They found a strong anisotropy in olivine which they assume to prevail throughout the olivine stability field in the mantle down to 410 km.

Variation with Pressure

Tommasi et al. (2001) measured thermal diffusivity in the crystallographic [100] and [010] directions parallel and perpendicular to a strain-induced foliation, respectively, as a function of temperature at atmospheric pressure on spinel lherzolites and spinel harzburgite. These rocks are considered representative for the subcontinental and suboceanic mantle,

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 7 Coefficients for fitting the variation of thermal diffusivity ($10^{-6} \text{ m}^2 \text{ s}^{-1}$) with temperature (K) according to Eq. 32 (Pertermann and Hofmeister 2006)

Sample		Chemical composition	$a \times 10^6$ ($\text{m}^2 \text{ s}^{-1}$)	$b \times 10^6$ ($\text{m}^2 \text{ s}^{-1} \text{ K}$)	$c \times 10^6$ ($\text{m}^2 \text{ s}^{-1} \text{ K}^2$)	T_{max} ($^\circ\text{C}$)
<i>Single crystals</i>						
Olivine	Fo[001]	Mg_2SiO_4	0.3081	679.6	213492	985
	FoCo[001]	$\text{Mg}_{1.99}\text{Co}_{0.01}\text{SiO}_4$	0.2347	587.8	172482	1477
	[010]	„	0.2415	115.4	165515	1181
	Needles[100]	$\text{Mg}_{1.84}\text{Fe}_{0.16}\text{SiO}_4$	0.7088	57.7	202533	985
	[010]	„	0.3100	100.6	86470	739
	[001]	„	0.3805	381.3	79703	886
	Sumput[010]	$\text{Mg}_{1.87}\text{Fe}_{0.13}\text{SiO}_4$	0.3135	127.2	73824	983
Sinhalite	[010]	MgAlBO_4	0.5546	−128.1	432571	741
Chrysoberyl	[100]	BeAl_2O_4	0.5366	551.7	566872	990
	[010]	„	0.3516	415.2	388978	989
	[001]	„	0.6371	428.2	543382	990
<i>Polycrystalline samples</i>						
Dunite	#1	$\sim\text{Mg}_{1.8}\text{Fe}_{0.2}\text{SiO}_4$	0.2291	290.0	92938	888
	#2	„	0.3563	178.4	93356	1083
Monticellite-bearing rock		$\text{Ca}_{1.15}\text{Mg}_{0.79}\text{Mn}_{0.06}\text{SiO}_4$	0.3816	153.0	23706	985
Hortonolite-bearing rock		$\text{Mg}_{1.2}\text{Fe}_{0.8}\text{SiO}_4$	0.2826	301.2	12638	705
Fayalite-bearing slag		$\sim\text{Fe}_{1.98}\text{Mn}_{0.02}\text{SiO}_4$	0.2637	83.9	21216	886
Fayalite-bearing rock		$\sim\text{Fe}_{1.84}\text{Mn}_{0.02}\text{Mg}_{0.14}\text{SiO}_4$	0.1798	265.3	15688	887

Fo: forsterite, FoCo: Co-doped forsterite.

respectively. They found an anisotropy in thermal diffusivity on the order of 25 % with the maximum aligned in the direction of strain. Support of their experimental findings was provided by corroborating petrophysical modeling.

For mantle minerals, such as olivine and its high-pressure polymorphs, the β - and γ -spinel wadsleyite and ringwoodite, respectively, Xu et al. (2004) fitted phonon thermal diffusivity measured to 1 373 K and 20 GPa to a linear equation in pressure yielding a pressure coefficient of $0.036(4) \text{ GPa}^{-1}$ (Table 5).

Osako et al. (2004) fitted thermal diffusivity measured to 1 100 K and 8.3 GPa on isotropic, single-crystal garnet and anisotropic olivine ($\text{Fo}_{93}\text{Fa}_7$; in three crystallographic directions) as upper and lower mantle constituents, respectively, to an exponential equation in pressure (Table 6). The pressure dependence of garnet and olivine was found to be on the order of $4 \% \text{ GPa}^{-1}$ to $5 \% \text{ GPa}^{-1}$ and $3 \% \text{ GPa}^{-1}$ to $4 \% \text{ GPa}^{-1}$, respectively.

Variation with Other Factors

Micro-cracks and grain boundaries give rise to increased thermal resistance and to a reduction of the mean free path of radiation due to scattering of radiation. It is somewhat debated below which grain size the effect is negligible: Based on the agreement between diffusivities measured on minerals and rocks, Gibert et al. (2003) concluded that the effect of grain boundaries, thermal cracking, and secondary phases is negligible. Along the same lines, Seipold (1998) argued that grain size is much larger than the phonon mean free path and therefore grain boundaries should not interfere with heat diffusion. Branlund and Hofmeister (2008) confirm this for quartzites, but find diffusivities measured on chert, agate, and chalcedony to be lowered by grain boundaries. They propose that grain sizes greater than $1 \mu\text{m}$ should not affect heat transfer.

Additional advective heat transfer was identified by Seipold and Schilling (2003) due to the release of water adsorbed at the inner surfaces of voids (i.e., pores and cracks) in rocks at about 450 K and by dehydration of serpentinite at 850 K. Both processes create high local overpressures which are relieved by cracking when the overpressure exceeds the tensile strength of the rock. The resulting flow is then accompanied by a corresponding advective heat transfer. This phenomenon was observed and studied in laboratory experiments (Seipold and Schilling 2003) but has implications for the lower crust and upper mantle with respect to recrystallization processes involving the discharge of fluids. These “crustal burps” provide the only conceivable way how some fluids from the mantle or lower crust may find their way to the Earth’s surface. The example discussed by Seipold and Schilling (2003) involves liberating water of crystallization during the conversion of serpentinite into forsterite and talc, followed by the formation of enstatite. As this involves heat advection as a

separate heat transfer mechanism, this process is better addressed separately and not parameterized into some sort of “effective” heat transport property not directly linked to a physical process.

Summary

Understanding the thermal regime of the Earth requires appreciation of properties and mechanisms for storage, transport, and generation of heat within the Earth. Both experimental and indirect methods are available for inferring the corresponding rock properties. Steady-state heat conduction or transient heat diffusion are the dominant transport processes in the Earth’s crust, except when fluid flow is sufficient for significant heat advection. For most crustal and mantle rocks, heat radiation sets in at temperatures above about 450°C and becomes significant only at temperatures above $1\,200^\circ\text{C}$. At temperatures above $2\,500^\circ\text{C}$, heat radiation becomes a dominant mechanism.

Cross-References

- [Heat Flow, Continental](#)
- [Heat Flow, Seafloor: Methods and Observations](#)
- [Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat](#)

Acknowledgments Dr. Sukanta Roy kindly commented on this contribution.

Bibliography

- Aronsson JR, Bellotti LH, Eckroad SW, Emslie AG, McConnell RK, von Thüna PC (1970) Infrared spectra and radiative thermal conductivity of minerals at high temperatures. *J Geophys Res* 75(17):3443–3456
- Bär K, Reinsch T, Bott J (2020) P³ – PetroPhysical property database – a global compilation of lab measured rock properties. *Earth Syst Sci Data Discuss.* <https://doi.org/10.5194/essd-2020-15>, in review, 2020. Database available at <https://doi.org/10.5880/GFZ.4.8.2019.P3>. Retrieved 21 Mar 2020
- Beardsmore GR, Cull JP (2001) *Crustal heat flow*. Cambridge University Press, Cambridge
- Beck AE (1988) Methods for determining thermal conductivity and thermal diffusivity. In: Hanel R, Rybach L, Stegena L (eds) *Handbook of terrestrial heat flow density determination*. Kluwer, Dordrecht, pp 87–124
- Birch F (1942) Thermal conductivity and diffusivity. In: Birch F, Schairer JF, Spicer HC (eds) *Handbook of physical constants*, Special paper 36. Geological Society of America, New York, pp 243–266
- Blumm J, Lemarchand S (2002) Influence of test conditions on the accuracy of laser flash measurements. *High Temp High Pres* 34:523–528
- Branlund JM, Hofmeister AM (2008) Factors affecting heat transfer in natural SiO_2 solids. *Am Mineral* 93:1620–1629

- Bruggeman DAG (1935) Berechnung verschiedener Konstanten von heterogenen Substanzen – I. Dielektrizitätskonstanten und Leitfähigkeiten der Mischkörper aus isotropen Substanzen. *Ann Phys* 24:636–679
- Buntebarth G (1991) Thermal properties of the KTB-Oberpfalz VB core samples at elevated temperature and pressure. *Sci Drill* 2:73–80
- Buntebarth G, Schopper JR (1998) Experimental and theoretical investigations on the influence of fluids, solids and interactions between them on thermal properties of porous rocks. *Phys Chem Earth* 23:1141–1146
- Carmichael RS (1984) CRC handbook of physical properties of rocks III. CRC Press, Boca Raton
- Čermák V, Rybach L (1982) Thermal conductivity and specific heat of minerals and rocks. In: Angenheister G (ed) *Landolt-Börnstein: numerical data and functional relationships in science and technology. New series, V(1a)*. Springer, Berlin, pp 305–343
- Clark SP Jr (1957) Radiative transfer in the Earth's mantle. *EOS Trans Am Geophys Union* 38:931–938
- Clark SP Jr (1966) Thermal conductivity. In: Clark SP Jr (ed) *Handbook of physical constants. Memoir 97*. Geological Society of America, New York, pp 459–482
- Clark SP Jr (1969) Heat conductivity in the mantle. In: Hart PJ (ed) *The Earth's crust and upper mantle. Geophysical monograph, vol 13*. American Geophysical Union, Washington, DC, pp 622–626
- Clauser C (1988) Opacity – the concept of radiative thermal conductivity. In: Hänel R, Rybach L, Stegena L (eds) *Handbook of terrestrial heat flow density determination*. Kluwer, Dordrecht, pp 143–165
- Clauser C (2006) Geothermal energy. In: Heinloth K (ed) *Landolt-Börnstein, Group VIII: advanced materials and technologies*. Springer, Heidelberg/Berlin. *Energy technologies, Subvol. C: Renewable energies, vol 3*, pp 480–595
- Clauser C (2009) Heat transport processes in the Earth's crust. *Surv Geophys* 30:163–191. <https://doi.org/10.1007/s10712-009-9058-2>
- Clauser C, Huenges E (1995) Thermal conductivity of rocks and minerals. In: Ahrens TJ (ed) *Rock physics and phase relations – a handbook of physical constants*. American Geophysical Union, Washington, DC. *AGU Reference Shelf, vol 3*, pp 105–126
- Crain ER (1986) *The log analysis handbook. Quantitative log analysis methods series, vol 1*. Pennwell Publishing, Tulsa
- Davis EE (1988) Oceanic heat-flow density. In: Hänel R, Rybach L, Stegena L (eds) *Handbook of terrestrial heat flow density determination*. Kluwer, Dordrecht, pp 223–260
- Davis GM, Chapman DS, Van Wagoner TM, Armstrong PA (2007) Thermal conductivity anisotropy of metasedimentary and igneous rocks. *J Geophys Res* 112:B05216. <https://doi.org/10.1029/2006JB004755>
- Debye P (1914) Zustandsgleichung und Quantenhypothese mit einem Anhang über Wärmeleitung. In: *Vorträge über die kinetische Theorie der Materie und der Elektrizität, Mathematische Vorlesungen an der Universität Göttingen, VI*. Teubner, Leipzig
- Diment WH, Pratt HR (1988) Thermal conductivity of some rock-forming minerals: a tabulation. *Open File Report 88-690*. US Geological Survey, Denver
- Dreyer W (1974) *Materialverhalten Anisotroper Festkörper: Thermische und Elektrische Eigenschaften*. Springer, Wien
- Eucken A (1911) Über die temperaturabhängigkeit der wärmeleitfähigkeit fester nichtmetalle. *Annalen der Physik*, 4 Folge 34(2):185–220
- Fricke H (1924) A mathematical treatment of the electric conductivity and capacity of disperse systems. *Phys Rev* 24:575–587
- Fuchs S, Förster H-J, Braune K, Förster A (2018) Calculation of thermal conductivity of low-porous, isotropic plutonic rocks of the crust at ambient conditions from modal mineralogy and porosity: a viable alternative for direct measurement? *J Geophys Res* B 123:8602–8614. <https://doi.org/10.1029/2018JB016287>
- Gibert B, Seipold U, Tommasi A, Mainprice D (2003) Thermal diffusivity of upper mantle rocks: influence of temperature, pressure, and the deformation fabric. *J Geophys Res* 108(B8):2359. <https://doi.org/10.1029/2002JB002108>
- Goutorbe B, Lucazeau F, Bonneville A (2006) Using neural networks to predict thermal conductivity from geophysical well logs. *Geophys J Int* 166(1):115. <https://doi.org/10.1111/j.1365-246X.2006.02924.x>
- Hartmann A, Rath V, Clauser C (2005) Thermal conductivity from core and well log data. *Int J Rock Mech Min Sci* 42:1042–1055. <https://doi.org/10.1016/j.ijrmms.2005.05.015>
- Hashin Z, Shtrikman S (1962) A variational approach to the theory of the effective magnetic permeability of multiphase materials. *J Appl Phys* 33(10):3125–3131
- Höfer M, Schilling FR (2002) Heat transfer in quartz, orthoclase, and sanidine at elevated temperature. *Phys Chem Miner* 29:571–584
- Hofmeister AM (1999) Mantle values of thermal conductivity and the geotherm from phonon lifetimes. *Science* 283:1699–1706
- Hofmeister AM (2005) Dependence of diffusive radiative transfer on grain-size, temperature, and Fe-content: implications for mantle processes. *J Geodyn* 40:51–72
- Hofmeister AM (2006) Thermal diffusivity of garnets at high temperature. *Phys Chem Miner* 33:45–62
- Hofmeister AM, Branlund JM, Pertermann M (2009) Properties of rocks and minerals – thermal conductivity of the Earth. In: Price GD (ed) *Mineral physics, treatise on geophysics, vol 2*. Elsevier, Amsterdam, pp 543–577
- Horai K (1971) Thermal conductivity of rock-forming minerals. *J Geophys Res* 76(5):1278–1308
- Horai K (1991) Thermal conductivity of Hawaiian basalt: a new interpretation of Robertson and Peck's data. *J Geophys Res* 96(B3):4125–4132
- Horai K, Simmons G (1969) Thermal conductivity of rock-forming minerals. *Earth Planet Sci Lett* 6(5):359–368
- Jorand R, Vogt C, Marquart G, Clauser C (2013) Effective thermal conductivity of heterogeneous rocks from laboratory experiments and numerical modeling. *J Geophys Res Solid Earth* 118(10):5225–5235. <https://doi.org/10.1002/jgrb.50373>
- Jorand R, Clauser C, Marquart G, Pechign R (2015) Statistically reliable petrophysical properties of potential reservoir rocks for geothermal energy use and their relation to lithostratigraphy and rock composition: the NE Rhenish Massif and the Lower Rhine Embayment (Germany). *Geothermics* 53:413–428
- Kappelmeyer O, Haenel R (1974) *Geothermics – with special reference to application*. Bornträger, Berlin/Stuttgart
- Kobolev VP, Kutas RI, Popov YA (1990) Method and results of research of thermal properties of Ural region rocks with laser scanning. *Geophys J* 12(4):29–37, Naukova Dumka, Kiev (in Ukrainian)
- Korvin G (1978) The hierarchy of velocity formulae: generalized mean value theorems. *Acta Geod Geophys Montan* 13:211–222
- Korvin G (1982) Axiomatic characterization of the general mixture rule. *Geos exploration* 19:267–276
- Mottaghy DC, Schellschmidt R, Popov YA, Clauser C, Kukkonen IT, Nover G, Milanovsky S, Romushkevich RA (2005) New heat flow data from the immediate vicinity of the Kola super-deep borehole: vertical variation in heat flow confirmed and attributed to advection. *Tectonophysics* 401(1–2):119–142. <https://doi.org/10.1016/j.tecto.2005.03.005>
- Osako M, Ito E, Yoneda A (2004) Simultaneous measurement of thermal conductivity and thermal diffusivity for garnet and olivine under high pressure. *Phys Earth Planet Inter* 143–144:311–320
- Parker WJ, Jenkins RJ, Butler CP, Abbott GI (1961) Flash method of determining thermal diffusivity, heat capacity, and thermal conductivity. *J Appl Phys* 32(9):1679–1684
- Peierls R (1929) Zur kinetischen Theorie der Wärmeleitung in Kristallen. *Annalen der Physik*, 5 Folge 3:1055–1101
- Pertermann M, Hofmeister AM (2006) Thermal diffusivity of olivine-group minerals at high temperature. *Am Mineral* 91:1747–1760

- Petrinin GI, Popov VG, Il'in IA (2004) Conductive heat transfer in plagioclases. *Izv Phys Solid Earth* 40(9):752–759
- Popov YA, Mandel AM (1998) Geothermal study of anisotropic rock masses. *Izv Phys Solid Earth* 34(11):903–915
- Popov YA, Berezin VV, Solov'yev GA, Romushkevich RA, Korostelev VM, Kostyurin AA, Kulikov IV (1987) Thermal conductivity of minerals. *Izv Phys Solid Earth* 23(3):245–253
- Popov YA, Pevzner LA, Romushkevich RA, Korostelev VM, Vorob'ev MG (1995) Thermophysical and geothermal sections obtained from Kolvinskaya well logging data. *Izv Phys Solid Earth* 30(9):778–789
- Popov YA, Pevzner LA, Pimenov VP, Romushkevich RA (1999a) New geothermal data from the Kola superdeep well SG-3. *Tectonophysics* 306(3–4):345–366
- Popov YA, Pribnow DFC, Sass JH, Williams CF, Burkhardt H (1999b) Characterization of rock thermal conductivity by high-resolution optical scanning. *Geothermics* 28:253–276
- Popov Y, Tertychnyi V, Romushkevich R, Korobkov D, Pohl J (2003) Interrelations between thermal conductivity and other physical properties of rocks: experimental data. *Pure Appl Geophys* 160:1137–1161
- Popov Y, Beardmore G, Clauser C, Roy S (2016) ISRM suggested methods for determining thermal properties of rocks from laboratory tests at atmospheric pressure. *Rock Mech Rock Eng* 49(10):4179–4207
- Ray L, Förster H-J, Schilling FR, Förster A (2006) Thermal diffusivity of felsic to mafic granulites at elevated temperatures. *Earth Planet Sci Lett* 251:241–253
- Reibelt M (1991) Study on the influence of surface structure and fluid saturation of rocks on the determination of thermal conductivity by a half-space line source. Unpublished diploma thesis (in German). Institut für Angewandte Geophysik, Petrologie und Lagerstättenforschung, Technische Universität Berlin
- Rosseland S (1930) The principles of quantum theory. In: Eberhard G, Kohlschütter A, Ludendorff H (eds) *Handbuch der Astrophysik*, vol 3(1). Springer, Berlin, pp 452–474
- Roy RF, Beck AE, Touloukian YS (1981) Thermophysical properties of rocks. In: Touloukian YS, Judd WR, Roy RF (eds) *Physical properties of rocks and minerals*. McGraw-Hill/CINDAS Data series on material properties, vol II-2. McGraw-Hill, New York, pp 409–502
- Sass JH, Lachenbruch AH, Moses TH Jr, Morgan T (1992) Heat flow from a scientific research well at Cajon Pass, California. *J Geophys Res* 97(B4):5017–5030
- Schärmeli G (1982) Anisotropy of olivine thermal conductivity at 2.5 GPa up to 1500 K measured on optically non-thick sample. In: Schreyer W (ed) *High-Pressure Researches in Geoscience*. Schweizerbart, Stuttgart, pp 349–373
- Schilling FR (1999) A transient technique to measure thermal diffusivity at elevated temperatures. *Eur J Mineral* 11:1115–1124
- Schulz B (1981) Thermal conductivity of porous and highly porous materials. *High Temp High Press* 13:649–660
- Seipold U (1998) Temperature dependence of thermal transport properties of crystalline rocks – a general law. *Tectonophysics* 291(1–4):161–171
- Seipold U, Schilling RR (2003) Heat transport in serpentinites. *Tectonophysics* 370:147–162
- Shankland TJ, Nitsan U, Duba AG (1979) Optical absorption and radiative heat transport in olivine at high temperature. *J Geophys Res* 84(B4):1603–1610
- Shim BO, Park JM, Kim HC, Lee Y (2010) Statistical analysis on the thermal conductivity of rocks in the Republic of Korea. In: *Proceedings World Geothermal Congress 2010 Bali, Indonesia*, 25–29 April 2010. International Geothermal Association (IGA), Bonn. <https://www.geothermal-energy.org/pdf/IGAstandard/WGC/2010/3126.pdf>
- Somerton WH (1992) Thermal properties and temperature related behavior of rock/fluid systems. Elsevier, Amsterdam
- Tommasi A, Gibert B, Seipold U, Mainprice D (2001) Anisotropy of thermal diffusivity in the upper mantle. *Nature* 411:783–786
- Vosteen H-D, Schellschmidt R (2003) Influence of temperature on thermal conductivity, thermal capacity, and thermal diffusivity for different types of rock. *Phys Chem Earth* 28(9–11):499–509
- Whittington AG, Hofmeister AM, Nabelek PI (2009) Temperature-dependent thermal diffusivity of the Earth's crust and implications for magmatism. *Nature* 458:319–321. <https://doi.org/10.1038/nature07818>
- Williams CF, Anderson RA (1990) Thermophysical properties of the Earth's crust: in situ measurements from continental and ocean drilling. *J Geophys Res* 95(B6):9209–9236
- Xu Y, Shankland TJ, Linhardt S, Rubie DC, Langenhorst F, Klasinski K (2004) Thermal diffusivity and conductivity of olivine, wadsleyite and ringwoodite to 20 GPa and 1373 K. *Phys Earth Planet Inter* 143–144:321–336
- Zimmerman RW (1984) The effect of pore structure on the pore and bulk compressibilities of consolidated sandstones. Ph.D. thesis, University of California, Berkeley
- Zimmerman RW (1989) Thermal conductivity of fluid-saturated rocks. *J Pet Sci Eng* 3(3):219–227
- Zoth G, Hänel R (1988) Appendix. In: Hänel R, Rybach L, Stegena L (eds) *Handbook of terrestrial heat flow density determination*. Kluwer, Dordrecht, pp 449–466

Time-Dependent Seismic Hazard Assessment

Zhongliang Wu

Institute of Earthquake Forecasting, China Earthquake Administration, Beijing, China

Definition

Seismic hazard refers to the potential for earthquake-related phenomena, resulting from a combination of source, path, and site effects on seismic vibrations. *Time-dependent seismic hazard assessment* takes into account the temporal variations associated with all these factors, especially the source factor related to the preparation and occurrence of earthquakes.

Time-Dependent Seismic Hazard

In a general sense, *seismic hazard* refers to any physical phenomenon associated with an earthquake (such as ground motion, ground failure, liquefaction, and even tsunamis). In a specific sense, *seismic hazard* is defined as the potential for earthquake-related phenomena, resulting from a combination of source, path, and site effects on seismic vibrations (Main 1996). Entry *seismic hazard* describes in detail the scientific basis and practical issues of seismic hazard assessment, which also deals with the issues related to time-dependent seismic hazard.

The time-dependent features of seismic source, path of seismic wave propagation, and site effects all contribute to the necessity and difficulty of time-dependent seismic hazard assessment. But in seismology, the most important contribution to such necessity and difficulty is from the complexity of earthquake preparation and occurrence, that is, the complexity of seismicity.

Traditional seismic hazard assessment, for example, the global seismic hazard map (GSHAP, Giardini et al. 2003), did not consider time-dependent factors. On the other hand, however, seismic hazard assessment has to be time-dependent since seismicity patterns are characterized by a variety of fluctuations, as described by the entries *Characteristic Earthquakes and Seismic Gaps*; *Seismic Quiescence and Activation*; *Earthquake, Foreshocks*; and *Earthquake, Aftershocks*, respectively. It has been shown for a long time by geological, paleoseismic, and historical records that two modes of seismicity, namely, power-law distribution of frequency-magnitude relation (the Gutenberg-Richter's law) and characteristic earthquakes, can switch to each other over time. It has been shown theoretically that seismic response of a fault system to steady tectonic loading can exhibit persisting fluctuations in the form of self-driven switching between the characteristic earthquake mode and the truncated power law distribution mode (Ben-Zion et al. 1999). Accordingly, tools have been developed and tested for time-dependent seismic hazard assessment.

Time-Dependent Seismic Hazard: Practical Tools and Case Examples of Application

Probabilistic Approach. Probabilistic seismic hazard assessment (PSHA), the quantification of probability that a specific level of ground motion will be exceeded in a given window of space and time, assumes that earthquake occurrence follows a Poisson process. When considering the variation of seismicity, the log-normal model of earthquake occurrence (Nishenko and Buland 1987) and the Brownian passage time (BPT) renewal model (Matthews et al. 2002) are generally used for probabilistic seismic hazard analysis. In the Uniform California Earthquake Rupture Forecast (UCERF, Field et al. 2014) which was time-independent, time-dependent factors were considered by using the BPT model (Field et al. 2015). For shorter time scales in which clustering of earthquakes has to be taken into account, the Epidemic Type Aftershock Sequence (ETAS) model (Ogata 1989) was used for time-dependent seismic hazard assessment (Field et al. 2017). Tools of statistical seismology, potentially capable of describing the time-dependent behavior of seismicity, play an essential role in such expansion from time-independent to time-dependent. Details of *Statistical seismology* are described by the corresponding entry.

Neo-deterministic Approach. In an approach to deterministic seismic hazard assessment (DSHA), referred to as neo-

deterministic seismic hazard assessment (NDSHA, Panza et al. 2001; XeRiS 2019), knowledge of seismic source, seismic wave propagation within heterogeneous Earth medium, and site effects on seismic ground motion is used for the quantification of earthquake scenarios. Considering the temporal variation of seismicity, the original DSHA is combined with intermediate-term earthquake prediction (Peresan 2018). In this approach, intermediate-term (i.e., several months) and middle-range (i.e., few hundred kilometers scale) predictions of mainshocks above a pre-assigned threshold – based on seismicity “alarms” generated by predictive algorithms like the CN algorithm and the M8 algorithm (Keilis-Borok and Soloviev 2003) – are used as the input of time-dependent seismic hazard assessment. The quintessence of the related predictive algorithms lies in the capturing of the “time of increased probability of earthquakes” by analyzing the state of seismicity, which determines the range of applicability (as well as the limits) of these algorithms. As a further development, the time-dependent seismicity is constrained by geodetic data (Panza et al. 2013) in which velocity variations and related strain accumulations detected by GNSS observation are used to reduce the geographic extent of the “alarms,” developing from “intermediate-term medium-range forecast” to “intermediate-term narrow-range forecast.”

Hybrid/Empirical Approach: The difficulty of time-dependent seismic hazard assessment is entangled with the difficulty of earthquake prediction (Kanamori 2003) which is described, from different perspectives, by the entries *Earthquake Precursors and Prediction*, and *Earthquake Prediction, M8 Algorithm*, respectively. In recent years, significant progress has been made in the study of earthquake predictability (Keilis-Borok 2002; Schorlemmer et al. 2018), providing time-dependent seismic hazard assessment with more solid scientific basis. Empirical approaches are still needed when dealing with multidisciplinary information and sometimes “big data.” An example of such empirical or hybrid approach is the time-dependent seismic hazard assessment in China. This is implemented by a series of scientific practices, represented by corresponding scientific products, with different time scales from the decade scale “Key Regions for Enhanced Monitoring and Preparedness” to the 1-year scale “Annual Consultation on the Likelihood of Earthquakes” (Wu 2018). In such practices the “alarm” areas are determined by a combination of statistics-based, precursor-based, and physics-based models, and the experiences of experts, based on panel discussion (called as “consultation”) oriented at the decision-making considering the (incomplete and multidisciplinary) information, play an important role. In the past, these practices were more related to earthquake forecast. But in recent years, such forecasts have been interconnected to seismic hazard and the engineering/social countermeasures for seismic risk reduction.

Seismic hazard assessment, no matter time-independent or time-dependent, needs test against real earthquakes (e.g.,

Nekrasova et al. 2014). The complexity of such test lies in three aspects: (1) the understanding of the quantities to be assessed (e.g., maximum ground motion with a certain probability of exceedance) as well as the “probabilities” expressed in the assessment; (2) the working assumption of “background” distribution of earthquakes and/or ground motion parameters; and (3) the statistical tools for evaluating the performance of an assessment or forecast based on small number of “real” earthquakes (generally a single earthquake).

Conclusions and Discussion

In the perspective of disaster resilience, *seismic disaster risk* is determined by the intersection of seismic hazard, exposure, and vulnerability. Seismic hazard assessment is basically the characterization of the “seismic hazard input” of seismic disaster risk. Due to the complexity of earthquake preparation and occurrence, the complexity associated with the propagation and attenuation of seismic waves, and the characteristic frequency range within which the seismic ground motion and engineering structures interact, seismic hazard assessment is both a scientific and a technical challenge in seismology (Somerville and Morivaki 2003). When time-dependent factors are taken into consideration, the challenge is even greater.

This entry focuses on the concept and recent progress of the study on *time-dependent seismic hazard*. By case examples the probabilistic, deterministic, and hybrid/empirical approaches are briefly introduced. In the related field, there is still much to be done toward the readiness of time-dependent seismic hazard assessment in answering the increasing needs of social sustainability.

However, in the perspective of seismic disaster risk reduction, time-dependent seismic hazard assessment does provide a new horizon to the study of earthquake forecast and prediction, a notoriously difficult scientific problem in seismology (AAAS 2005). As a matter of fact, quite a few approaches to time-dependent seismic hazard assessment (e.g., Field et al. 2017) have logical connection to the concept “operational earthquake forecast” (Jordan et al. 2011). Seen from the representation of seismic disaster risk by seismic hazard, exposure, and vulnerability, traditional earthquake prediction tried to identify the increase of seismic hazard and provide warnings to the society. The society in turn reduces the exposure (by organized evacuation, such as the case of the 1975 Haicheng earthquake) so that seismic disaster risk may be significantly mitigated. Thinking this problem in an alternative way, when seismic hazard with different spatiotemporal scales is assessed, countermeasures can be taken more efficiently to reduce either the exposure (such as in the planning of land use, with decade time scale) or the vulnerability (such as engineering reinforcement, with annual time scale). In the context of time-dependent seismic hazard, therefore, seismic disaster risk reduction for disaster resilience becomes

practical (or “operational”) based on the presently accessible, although still limited, knowledge of earthquake physics and capability of forecast of the behavior of earthquakes.

Cross-References

- ▶ [Characteristic Earthquakes and Seismic Gaps](#)
- ▶ [Earthquake Precursors and Prediction](#)
- ▶ [Earthquake Prediction, M8 Algorithm](#)
- ▶ [Earthquake, Aftershocks](#)
- ▶ [Earthquake, Foreshocks](#)
- ▶ [Seismic Hazard](#)
- ▶ [Seismic Quiescence and Activation](#)
- ▶ [Statistical Seismology](#)

Bibliography

- AAAS (2005). <http://www.sciencemag.org/site/feature/misc/webfeat/125th/>
- Ben-Zion Y, Dahmen K, Lyakhovsky V, Ertas D, Agnon A (1999) Self-driven mode switching of earthquake activity on a fault system. *Earth Planet Sci Lett* 172:11–21
- Field EH, Arrowsmith RJ, Biasi GP, Bird P, Dawson TE, Felzer KR, Jackson DD, Johnson KM, Jordan TH, Madden C, Michael AJ, Milner KR, Page MT, Parsons T, Powers PM, Shaw BE, Thatcher WR, Weldon RJ II, Zeng YH (2014) Uniform California Earthquake Rupture Forecast, version 3 (UCERF3) – the time-independent model. *Bull Seismol Soc Am* 104:1122–1180
- Field EH, Biasi GP, Bird P, Dawson TE, Felzer KR, Jackson DD, Johnson KM, Jordan TH, Madden C, Michael AJ, Milner KR, Page MT, Parsons T, Powers PM, Shaw BE, Thatcher WR, Weldon IIRJ, Zeng YH (2015) Long-term, time-dependent probabilities for the Third Uniform California Earthquake Rupture Forecast (UCERF3). *Bull Seismol Soc Am* 105:511–543
- Field EH, Milner KR, Hardebeck JL, Page MT, van der Elst NJ, Jordan TH, Michael AJ, Shaw BE, Werner MJ (2017) A Spatiotemporal clustering model for the third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS): toward an operational earthquake forecast. *Bull Seismol Soc Am* 107:1049–1081
- Giardini D, Gruenthal G, Shedlock K, Zhang P (2003) The GSHAP global seismic hazard map. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*. Academic Press, London, pp 1233–1239
- Jordan TH, Chen YT, Gasparini P, Madariaga R, Main I, Marzocchi W, Papadopoulos G, Sobolev G, Yamaoka K, Zschau J (2011) Operational earthquake forecasting: state of knowledge and guidelines for utilization. *Ann Geophys* 54:316–391
- Kanamori H (2003) Earthquake prediction: an overview. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*. Academic Press, London, pp 1205–1216
- Keilis-Borok VI (2002) Earthquake prediction: state-of-the-art and emerging possibilities. *Annu Rev Earth Planet Sci* 30:1–33
- Keilis-Borok VI, Soloviev AA (eds) (2003) *Non-linear dynamics of the lithosphere and earthquake prediction*. Springer, Heidelberg, pp 1–337
- Main I (1996) Statistical physics, seismogenesis, and seismic hazard. *Rev Geophys* 34:433–462
- Matthews MV, Ellsworth WL, Reasenberg PA (2002) A Brownian model for recurrent earthquakes. *Bull Seismol Soc Am* 92:2233–2250

- Nekrasova A, Kossobokov V, Peresan A, Magrin A (2014) The comparison of the NDSHA, PSHA seismic hazard maps and real seismicity for the Italian territory. *Nat Hazards* 70:629–641
- Nishenko SP, Buland R (1987) A generic recurrence interval distribution for earthquake forecasting. *Bull Seismol Soc Am* 77:1382–1399
- Ogata Y (1989) Statistical models for standard seismicity and detection of anomalies by residual analysis. *Tectonophysics* 169:159–174
- Panza GF, Romanelli F, Vaccari F (2001) Seismic wave propagation in laterally heterogeneous anelastic media: theory and applications to seismic zonation. *Adv Geophys* 43:1–95
- Panza GF, Peresan A, Magrin A, Vaccari F, Sabadini R, Crippa B, Marotta AM, Splendore R, Barzaghi R, Borghi A, Cannizzaro L, Amodio A, Zoffoli S (2013) The SISMA prototype system: integrating geophysical modeling and Earth observation for time-dependent seismic hazard assessment. *Nat Hazards* 69:1179–1198
- Peresan A (2018) Recent developments in the detection of seismicity patterns for the Italian region. In: Ouzounov D, Pulinet S, Hattori K, Taylor P (eds) *Pre-earthquake processes: a multidisciplinary approach to earthquake prediction studies*. Geophysical monograph, vol 234. American Geophysical Union/John Wiley, Washington, DC, pp 149–172
- Schorlemmer D, Werner MJ, Marzocchi W, Jordan TH, Ogata Y, Jackson DD, Mak S, Rhoades DA, Gerstenberger MC, Hirata N, Liukis M, Maechling PJ, Strader A, Taroni M, Wiemer S, Zechar JD, Zhuang J (2018) The Collaboratory for the Study of Earthquake Predictability: achievements and priorities. *Seismol Res Lett* 89:1305–1313
- Somerville P, Morivaki Y (2003) Seismic hazards and risk assessment in engineering practice. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*. Academic Press, London, pp 1065–1080
- Wu ZL (2018) Multi-scale seismic hazard in the context of the reduction of earthquake disaster risk: a case study on the earthquake forecast practice in China. In: Madu CR, Kuei C-H (eds) *Handbook of disaster risk reduction & management*. World Scientific, New Jersey, pp 433–449
- XeRis (2019). <http://www.xeris.it/Hazard/index.html>

Time-Reversal in Seismology

Carène Larmat and Clarence S. Clay (Deceased)
Geophysics Group, EES-17, MS D542, Los Alamos National Laboratory, Los Alamos, NM, USA

Definition

Time-reversal Imagine filming the ripples formed when a pebble is dropped into a pond and playing the movie backward. You will observe the ripples converging back to the original impact. This is the fundamental concept behind time-reversal. It relies on the invariance of the wave propagation equation. In the absence of loss terms, the wave equation only contains second-order space and time derivatives. This

implies that if $u(x, t)$ is a wavefield solution, then $u(x, -t)$, $u(-x, t)$, and $u(-x, -t)$ are also solution of the equation; therefore, the wavefield can be reversed. The reversal can be achieved either by instant reversing of the wavefield over the entire space (Loschmidt 1876) or by using a **time-reversal mirror** (TRM). The latter refers to a surface surrounding the original source, on which the signal is recorded during the forward propagation, and then, from which, this signal is rebroadcast. The signal is read backward to impose the reversal. The wavefield created will be an approximate reverse version of an original wave propagation movie and converges back onto the source point. In practice, the mirror corresponds to a set of receivers. The necessary number of receivers and type of signal for a successful reversal is discussed later in “Modern Time-Reversal” and “Time-Reversal and Complexity.”

Time-reversal applications

In geophysics, time-reversal is typically used to image the source of the wavefield. However, time-reversal has also been used for applications requiring wave energy to be focused onto a specific point in space and time, such as lithotripsy, medical imaging, etc. Another way to interpret time-reversal is the establishment of a favored path between a pair of points in space (the source and the receiver), so it can be used to establish communication lines in complex media or allow nonlinear imaging that requires a strong acoustic signal to be applied on the imaged area.

Reciprocal time-reversal

The principle of source-receiver reciprocity states that if the source and a receiver were to be interchanged, the same waveform will be observed. There are applications that require the focus of energy on the receiver rather than the source. In this case, a pulse is first sent from a source point and recorded on a receiver. This signal is reversed and rebroadcast from the source point and observed to focus on the receiver.

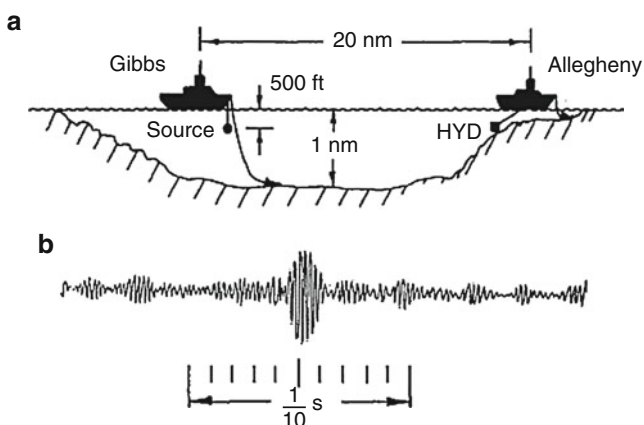
History

Time-Reversal Before Modern Electronics

The story of experimental time-reversal starts in the 1950–1960s in a cluttered room of the Hudson Laboratory of Columbia University where Antares Parvulescu tested his

idea of time-reversed transmissions. A microphone was placed about 25 ft from a loudspeaker, from which a series of short impulses were sent one second apart. A sequence of the signal due to the room reverberations were received at the microphone and tape-recorded. The tape was then played backward to obtain “time-reversed signals” called then “matched signals.” It was played by the loudspeaker (so this is a reciprocal time-reversal experiment). The sound created was an indiscernible rushing noise over the experiment room, except at the microphone. The scope at the microphone displayed a signal that looked like the auto correlation of the reverberation signal. When the scientists (including Clarence Clay) indeed brought their ears to the microphone, they heard a simple “click,” indicating that localization in space and time of the wave energy did occur! This result was first published in a terse abstract (Parvulescu 1961) and is further documented in Parvulescu 1995 and in Chap. 7 of the book of Tölstoý and Clay (1966 and its revision in 1987).

In 1961–1962, Parvulescu and colleagues decided to take this experiment to sea despite the thought at the time that transmission in the ocean was not stable enough to allow time-reversal. Two oceanographic research ships, the Gibbs and Allegheny, with deep anchoring capability were stationed in the oceanic trench, the Tongue of the Ocean. Fig. 1a shows the setting of the experiment. The received signals at the Allegheny, lengthened by multiple reverberations, were radioed back to the Gibbs to be transmitted from the source (this was again a reciprocal time-reversal experiment). Fig. 1b shows the peak of pressure created on the hydrophone. The report of this first “field” time-reversal experiment is documented in Parvulescu and Clay (1965). Details of further experimentations can be found in Chaps. 4–8 of Tölstoý and Clay (1966).



Time-Reversal in Seismology, Fig. 1 (a) Ship positions and instruments used for the marine experiment. Gibbs has the source, and Allegheny has the receiver (HYD). (b) Signal received at the hydrophone when performing a reciprocal time-reversal between the two ships in an oceanic trench. (Figures from Parvulescu and Clay 1965; Tölstoý and Clay 1966)

Modern Time-Reversal

In the late 1980s, advances in micro-electronics, storage capacity, and automated signal processing led to the development of transducer arrays which were used as TRM by Mathias Fink and colleagues at the Laboratoire Ondes et Acoustique (LOA) of the University of Paris VII (Fink 1992). Finite-size arrays of transmit receive transducers served as TRM and have been the basis of countless studies by this group (e.g., Fink et al. 2000; Fink 2006). Their work is the experimental and theoretical roots of time-reversal. Since, the number of applications of time-reversal is staggering, ranging from nondestructive evaluation (e.g., Ulrich et al. 2008a, b) and communication (e.g., Lerosey et al. 2004) to medical imaging (e.g., Fink and Tanter 2010).

Computers and scientific computing have also been a driver to the development of modern time-reversal. In the 1980s, George McMechan and colleagues at the University of Texas modified imaging techniques developed for exploration geophysics. They changed the algorithms to image the primary source of the wavefield instead of imaging the geologic reflectors and independently rediscovered time-reversal. They use a 2D numerical model, purely acoustic, representing a 15 by 12 km vertical slice of a California formation and successfully imaged the source of a small 1983 aftershock earthquake in California using data from a local dense network (McMechan 1982).

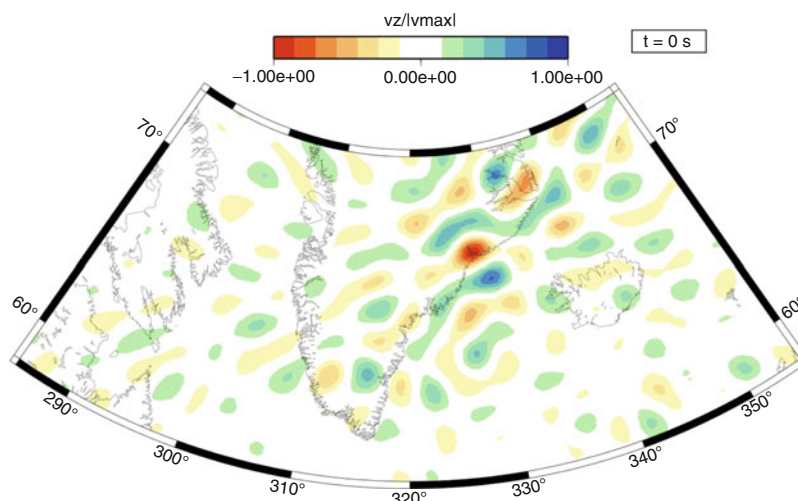
For references and more technical details about time-reversal, such as theoretical limit, one-channel time-reversal, and one-bit time-reversal, see the reviews by Fink et al. 2000; Fink 2006; Anderson et al. 2008; and Larmat et al. 2010.

Location of Seismic Sources

Since the late 1990s, several groups have performed time-reversal location of seismic events (see Anderson et al. 2008; Larmat et al. 2010), harnessing the developing computing power. These applications have demonstrated that contrary to what was thought by McMechan, time-reversal does work with “sparse” rebroadcasting networks (i.e., with receivers far more than one wavelength apart). Fig. 2 shows the time-reversal result from a 2001 magnitude-5 glacial earthquake in Greenland by Larmat et al. (2008). Glacial earthquakes were discovered in 2003 by Ekström et al. (2003) as long-period signal barely above the noise level. The signal from 146 stations scattered in the Northern hemisphere were filtered between 0.01 and 0.02 Hz, time-reversed and rebroadcast into a 3D Earth model. The figure shows the snapshot of the vertical component of the time-reversed displacement at the moment of the focus (actually the vertical component normalized by the maximum value reached on the snapshot). The focus appears as a two-lobe pattern of opposite polarity indicating the origin of the seismic signal. The two-lobe

Time-Reversal in Seismology,

Fig. 2 One hundred and forty-six vertical displacement signals of a glacial earthquake were time-reversed and sent back in a 3D Earth model. The vertical component (v_z) of the created time-reversal signal at the moment of the focus displays a two-lobe pattern of opposite polarity consistent with a single-force vector as the source mechanism. (Adapted from Larmat et al. 2008)



pattern reveals that the nature of the source excitation was a single-force vector, which is consistent with recent source analysis by Tsai et al. (2008). Had the source been a simple monopole, the focus would have manifested itself as a single point (see details in Larmat et al. 2008).

The signal generated by glacial earthquakes is in general embedded in the noise. Therefore, this example is a demonstration of the insight of Brian Kennett (1983) who foresaw the utility of time-reversal as a location method for source of very faint or transient signal or complex sources. Time-reversal has recently been used to locate tremors (Steiner et al. 2008; Larmat et al. 2009; Lokmer et al. 2009). Time-reversal is currently a very active research topic for monitoring microseismicity, either natural or man-made. Time-reversal analysis is computationally intensive for such applications and authors have been developing new imaging conditions to extract the signal of interest and reduce efficiently the four-dimensional space (time and space) that needs to be monitored. Auto-correlation, cross-correlation, and hybrid conditions involving correlation and stacking have been proposed (e.g., Nakata and Beroza 2016; Artman et al. 2010).

Related Seismic Methods

Other location methods are based on the core idea of sending back some recorded signals to a given point in space and time. All **back-projection**-related methods can be seen as simplified time-reversal procedures for which the backward propagation is reduced to shifting the time-series by predicted arrival times (e.g., Kao et al. 2005) instead of performing the full waveform rebroadcast that is time-reversal.

Time-reversal ideas are also used in subsurface imaging. The exploration methods used by McMechan have evolved into **reverse-time migration** (e.g., Baysal et al. 1983). The **adjoint methods** (Tromp et al. 2005) use backpropagation to

compute finite-frequency kernels which represent the sensitivity of waves to the subsurface structure.

Seismic interferometry is another research topic which uses the concept of time-reversal. Under certain conditions, the cross-correlation of the signal recorded at two receivers is the superposition of the Green function between the two receivers and its time-reversed version (see ▶ [“Seismic, Ambient Noise Correlation”](#) and the review by Snieder and Wapenaar 2010, and the pioneer paper of Claerbout 1968). Time-reversal can also be expressed as a cross-correlation of two wavefields, typical of a **matched filter**, whose output is optimal in some sense. Whatever the Green function, the time-reversal wavefield is maximal at the focus time and at the source point (Fink et al. 2000, p. 1946). Applications derived from this interpretation of time-reversal consist in locating seismic sources, such as trapped miners, by comparing live recorded signals with a database of pre-recorded Green functions (Hanafi et al. 2009).

Time-Reversal and Complexity

The earth and oceans are complex environments in which transmitted signals quickly become difficult to interpret: for example, instead of a single pulse for the first arrival, the P-wave is accompanied by a coda train due to the scattering within the crust. Contrary to many location methods based on simple paradigms, time-reversal thrives with complexity. In a homogeneous and purely acoustic medium, only one phase will be recorded on a single receiver from a source point. Introduce a reflecting interface and two phases will be recorded: the primary and a copy of the source impulse due to the reflection on the interface. The two combined phases contain more information about the source location than the primary phase alone. All these different phases will converge and add up onto the source point when traveling backward. Several studies (e.g., Dowling and Jackson 1992; Fink et al.

2000; Blomgren et al. 2002) have actually demonstrated the width of the focus is smaller in randomly heterogeneous media than in the homogeneous case (implying better resolution on the location). This is referred as *super-resolution* (Fink 2006; Lerosey et al. 2007). The second important point for complex media is that any arbitrary time-segment will return to the source. Time-reversal doesn't need any interpretation of the signals.

These unique facets of time-reversal explains the wide use of the method in very different disciplines handling with complex media: e.g., target detection method and communication in underwater acoustics, nondestructive evaluation (i.e., location of defects in intact specimens), improved imaging and destruction of stones/tumors in medicine, and source location and imaging techniques in geophysics.

Summary

Seismic source study has been subject to a shift from event-based temporary observation to continuous monitoring based on dense seismic networks. Seismologists now study a wealth of signals due to less energetic sources than earthquakes such as tremors and microseismicity. Glacial earthquakes, landslides, Earth hum, and meteorological phenomena such as hurricanes belong to a group of events that defines “environmental seismology,” a term coined by the French seismologist Jean-Paul Montagner. Microseismic noise is now used for imaging the Earth. Due to their emergent nature, these signals represent a challenge to classical location methods based on triangulation. Time-reversal as a method and time-reversal ideas are emerging as a promising framework to foster future location and imaging methods. The success of time-reversal as a source location method shall rely on the coverage of seismic networks, the accuracy of high-resolution velocity models, and the performance of numerical schemes to be used for the backward propagation. Still open questions are the convergence and resolution power of time-reversal under various conditions. Our ability to observe and model scattering phenomena will be part of the answer.

Cross-References

- [Earthquake, Focal Mechanism](#)
- [Earthquakes, Location Techniques](#)
- [Propagation of Elastic Waves: Fundamentals](#)
- [Seismic Data Acquisition and Processing](#)
- [Seismic Tomography](#)
- [Seismic, Ambient Noise Correlation](#)
- [Seismic, Migration](#)
- [Seismic, Waveform Modeling and Tomography](#)
- [Seismological Networks](#)

Acknowledgment Carène Larmat wishes to thank Paul Johnson, Robert Guyer, and Kevin Boland for their suggestions for this manuscript as well as the many colleagues she has met for enlightening discussions. The author gratefully acknowledges the input and the support of the US Department of Energy through the LANL/LDRD Program for this work.

Carène Larmat and Clarence S. Clay gratefully acknowledge the support of the US Department of Energy through the LANL/LDRD Program for this work.

Bibliography

- Anderson B, Griffa M, Larmat C, Ulrich TJ, Johnson PA (2008) Time reversal. *Acoust Today* 4(1):5
- Artman B, Podladtchikov I, Witten B (2010) Source location using time-reverse imaging. *Geophys Prospect* 58(5):861–873. <https://doi.org/10.1111/j.1365-2478.2010.00911.x>
- Baysal E, Kosloff DD, Sherwood JWC (1983) Reverse time migration. *Geophysics* 48(11):1514–1524
- Blomgren P, Papanicolaou G, Zhao H (2002) Super-resolution in time-reversal acoustics. *J Acoust Soc Am* 111(1):230–248
- Claerbout JF (1968) Synthesis of a layered medium from its acoustic transmission response. *Geophysics* 33:264–269
- Dowling DR, Jackson DR (1992) Narrow-band performance of phase-conjugate arrays in dynamic random media. *J Acoust Soc Am* 91:3257–3277
- Ekström G, Nettles M, Abers G (2003) Glacial earthquakes. *Science* 302:622–624
- Fink M (1992) Time-reversal of ultrasonic fields – Part I: basic principles. *IEEE Trans Ultrason Ferroelectr Freq Control* 39(5):555–566
- Fink M (2006) Time-reversal acoustics in complex environments. *Geophysics* 71(4):SI151–SI164
- Fink M, Tanter M (2010) Multiwave imaging and super resolution. *Phys Today* 63(2):28–33
- Fink M, Cassereau D, Derode A, Prada C, Roux P, Tanter M, Thomas J-L, Wu F (2000) Time-reversed acoustics. *Rep Prog Phys* 63(12):1933–1995
- Hanafi SM, Cao W, McCarter K, Schuster GT (2009) Using super-stacking and super-resolution properties of time-reversal mirrors to locate trapped miners. *Lead Edge* 28(3):302–307
- Kao H, Shan S-J, Dragert H, Rogers G, Cassidy JF, Ramachandran K (2005) A wide depth distribution of seismic tremors along the northern Cascadia margin. *Nature* 436:841–844
- Kennett BLN (1983) A new way to estimate source parameters. *Nature* 302:659–660
- Larmat C, Tromp J, Liu Q, Montagner J-P (2008) Time reversal location of glacial earthquakes. *J Geophys Res* 113:B09314
- Larmat CS, Guyer RA, Johnson PA (2009) Tremor source location using time reversal: selecting the appropriate imaging field. *Geophys Res Lett* 36:L22034
- Larmat CS, Guyer RA, Johnson PA (2010) Time-reversal methods in geophysics. *Phys Today* 63:31–35
- Lerosey G, De Rosny J, Tourin A, Derode A, Montaldo G, Fink M (2004) Time reversal of electromagnetic waves. *Phys Rev Lett* 92(19):34
- Lerosey G, De Rosny J, Tourin A, Fink M (2007) Focusing beyond the diffraction limit with far-field time reversal. *Science* 315(5815):1120–1122
- Lokmer B, O'Brien GS, Stich D, Bean CJ (2009) Time reversal imaging of synthetic volcanic tremor sources. *Geophys Res Lett* 36:L12308
- Loschmidt J (1876) *Sitzungsber. Kais. Akad. Wiss. Wien, Math. Naturwiss. Classe* 73:128–142
- McMechan GA (1982) Determination of source parameters by wavefield extrapolation. *Geophys J R Astron Soc* 71:613–628

- Nakata N, Beroza GC (2016) Reverse time migration for microseismic sources using the geometric mean as an imaging condition. *Geophysics* 81:KS51–KS60
- Parvulescu A (1961) MESS processing. *J Acoust Soc Am* 33:1674
- Parvulescu A (1995) Matched-signal (MESS) processing by the ocean. *J Acoust Soc Am* 98:943–960
- Parvulescu A, Clay CS (1965) Reproducibility of signal transmissions in the ocean. *Radio Eng Electron Phys* 29(4):223–228
- Schuster GT (2009) *Seismic interferometry*. Cambridge University Press, Cambridge/New York
- Snieder R, Wapenaar K (2010) Imaging with ambient noise. *Phys Today* 63(9):44–49
- Steiner B, Saenger EH, Schmalholz SM (2008) Time reverse modeling of low-frequency microtremors: application to hydrocarbon reservoir localization. *Geophys Res Lett* 35:L03307
- Tölstoy I, Clay CS (1966) *Ocean acoustics, theory and experiment in underwater sound*. McGraw-Hill (Translated into Russian and published, 1969. Revised-edition 1987, the Acoustical Society of America – the American Institute of Physics), New York
- Tromp J, Tape C, Liu Q (2005) Seismic tomography, adjoint methods, time reversal, and banana-doughnut kernels. *Geophys J Int* 160:195–216
- Tsai VC, Rice JR, Fahnstock M (2008) Possible mechanisms for glacial earthquakes. *J Geophys Res* 113:F03014
- Ulrich TJ, Sutin AM, Guyer RA, Johnson PA (2008a) Time reversal and non-linear elastic wave spectroscopy (TR NEWS) techniques. *Int J Non Linear Mech* 43:209–216
- Ulrich TJ, Sutin AM, Claytor T, Pallin P, Le Bas P-Y, TenCate JA (2008b) The time reversal elastic nonlinearity diagnostic applied to the evaluation of diffusion bonds. *Appl Phys Lett* 93:151914

TOPO-EUROPE: From the Deep Earth to the Surface of Continental Europe and Its Margins

Sierd Cloetingh and TOPO-EUROPE Working Group
Tectonics Group, Department of Earth Sciences, Utrecht
University, Utrecht, The Netherlands

Definition

Linking different spatial and temporal scales in coupled deep Earth and surface processes is a prime objective of the multidisciplinary international research program TOPO-EUROPE. The research approach of TOPO-EUROPE integrates active collection of new data, reconstruction of the geological record, and numerical and analog modeling. The results of the program focus on closely interrelated topics: deep Earth, lithospheric structure, mantle-lithosphere interactions, and sedimentary basins and geo-resources. Quantitative understanding of topographic evolution in space and time requires study of processes from the upper mantle through the lithosphere and crust and acting on the Earth's surface. The overview presented here demonstrates the opportunities for further understanding of topography through integrated studies of the full Earth system across space and timescales.

Introduction

One of the important developments in Solid-Earth science over the past decade has been recognition of the importance of linking deep Earth dynamic processes with surface and near-surface geologic processes (e.g., Cloetingh et al. 2007, 2013). Deep Earth research, encompassing fields such as seismology and mantle geodynamics, has traditionally operated distinctly from fields focusing on dynamics of the Earth's surface, such as sedimentology and geomorphology. However, these endeavors have in common the study of Earth's topography and the prediction of its origin and rates of change. Observables from surface studies, such as basin stratigraphy, geomorphology of landscapes, changes in surface elevation, and changes in sea level, provide some of the principal constraints on geodynamic and tectonic models. Conversely, deep geodynamic processes give rise to the topography, erosion, and sediment generation that are the basis of surface geology. The lithosphere, due to its stratified rheological structure, acts as a nonlinear “filter” for deeper sources, attenuating long deformation wavelength and creating new, shorter wavelength deformation, giving a surface response more complex than that of the mantle source.

Surface manifestations of these deep geodynamic processes modified by multifaceted mantle-lithosphere interactions have societal impact by (1) creating natural hazards, such as earthquakes and mass movements, and (2) controlling the distribution of natural resources such as fossil fuels or geothermal energy. The relevance of research conducted in both the deep Earth and surface regimes is thus enhanced through a focus on their interaction.

Long-term processes such as mantle flow drive the system of mantle-surface interaction. However, short-term processes such as erosion and climate tune the response. The continental lithosphere transforms deformation generated at its base by mantle flow, but surface processes control many of the observables such as topography, sediment, and exhumation rates, and thus further filter the source and even tune the response through cyclic processes such as climate. However, little is known on how far these interactions may go, or how these different processes are coupled or feed back into the dynamic system. Many modelling or theoretical studies show that erosion and sedimentation should have an impact on the subsurface evolution of crustal deformation, changing flow patterns and magnitudes in the ductile crust and mantle through changing gravitational stresses or kinematics.

This leads to many questions that remain to be tested. For instance, what is the impact of erosion and surface mass redistribution on mantle flow? What is the influence of protracted precipitation regimes on mountain building? and what is the budget of subduction in terms of sediments and water recycling? What is the sediment and tectonic load incidence on thermicity and sediment blanketing and on foreland

flexure? How and on what timescales does the build-up of topography impact atmospheric circulation processes? How does the creation of topography modify shallow crustal stress fields and how does this impact seismicity and fault kinematics? How can we differentiate between geodynamic, tectonic, and surficial displacements of the Earth's surface if all processes are active? These and many other questions are still unsolved.

Tectonic, surface, and external forcing processes are responsible for the growth and decay of continental topography and sedimentary basins, i.e., the interplay between sediment supply and mass (re)distribution with the full range of deep Earth to surface processes. However, the mechanisms that link exhumation, formation of topography, and sedimentation are still poorly understood because of a lack of insight into the variability of the rates and scales of the underlying processes. TOPO-EUROPE explores the dynamics of these processes and the role played by large-scale lithospheric stresses. To this end, lateral variations of tectonic and erosion/sedimentation evolution through time are incorporated into dynamic models of topography evolution by integrating deposition in sedimentary basins with topography formation and evolution in neighboring source areas. This will provide new opportunities to analyze and quantify the interplay between deep Earth and surface processes, which is critical for novel evaluation of geo-resources and natural hazards.

Scientifically, the role of topography in the Earth's system is as a dynamic, active permeable boundary layer between solid Earth processes occurring at different depths and atmospheric processes. Clearly, understanding the processes and hence predicting rates and signals of topographic change over the planet is a first-order challenge to the Earth Sciences.

Underpinning many of the pure and applied issues concerning resource and hazards is the basic question of which processes define present topographic change, and at what spatial and temporal scales they are acting. Processes such as so-called dynamic topography (e.g., Faccenna et al. 2014; Colli et al. 2018) driven by mantle flow, lithospheric partitioning of stress and strain, and the thermal structure of the mantle and overlying crust remain only partly understood.

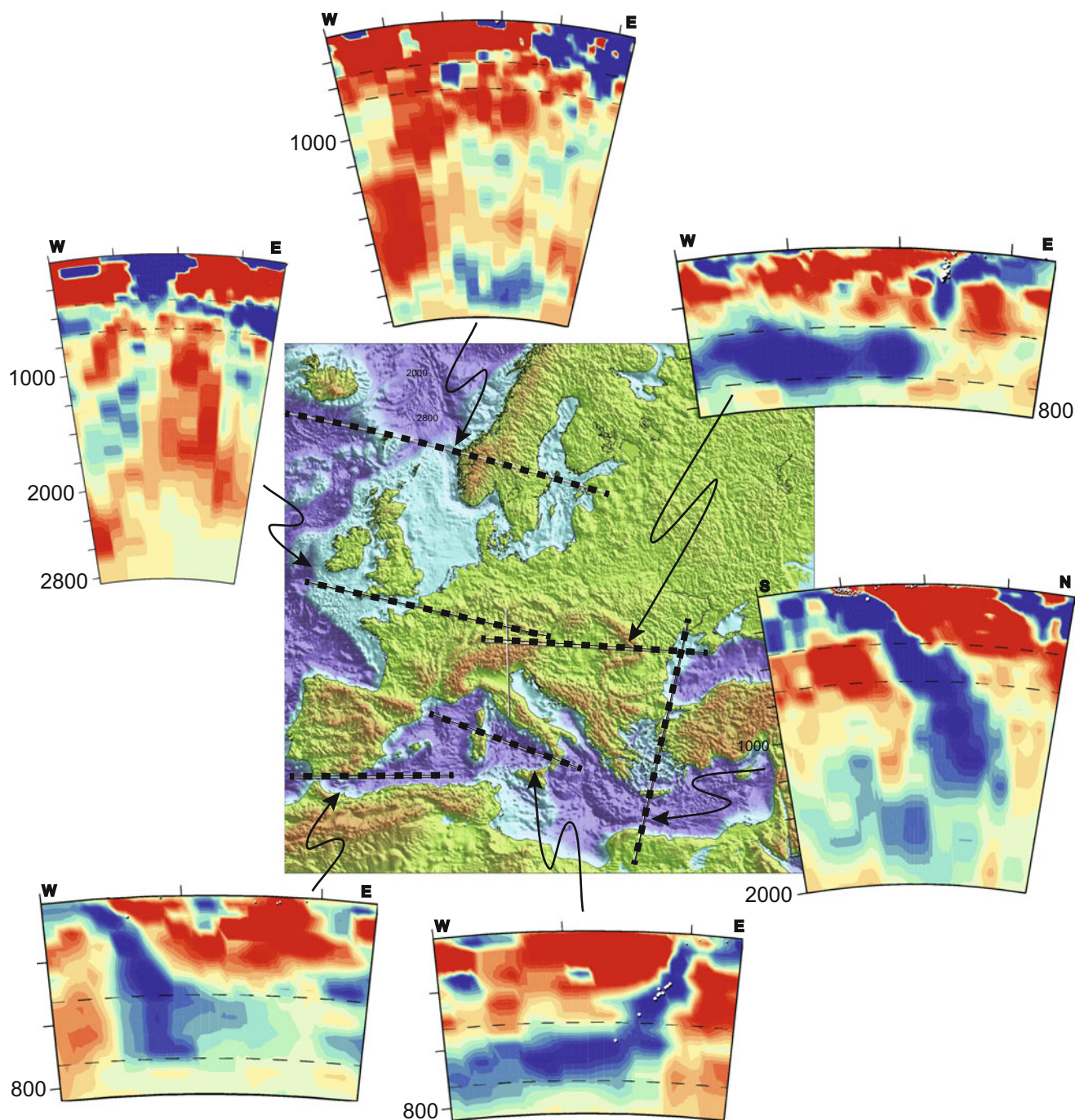
The TOPO-EUROPE program has been designed to serve as a platform for integration of deep Earth and surface processes studies, taking the Europe continent and its margins as a natural laboratory. TOPO-EUROPE has focused on collaborative research projects in specific areas such as the Norwegian continental margin, the Pyrenees, the Alpine-Carpathian system, the western Mediterranean, the Anatolian Plateau, and Europe as a whole. Results of TOPO-EUROPE research have been published in a number of special topical volumes and publications therein (Cloetingh and Tibaldi 2012; Cloetingh et al. 2009, 2011, 2018). TOPO-EUROPE has been building on a rich database on the deep

crustal structure of key sectors of the European continent, and on data obtained from industrial and deep seismic reflection profiling programs (e.g., ECORS, CROP, NFP, DEKORP, and BIRPS), which were set up to characterize the structure and architecture of sedimentary basins and the crust (Roure et al. 2010; Laubscher 2010). Here we present a brief review of the rationale of TOPO-EUROPE as well as some examples of research approaches and research highlights.

Deep Earth

Plate reconstructions and absolute plate motions based on seismic tomography from research within the context of the TOPO-EUROPE project have led to a new reference frame to determine absolute plate motions from remnants of past subduction, as imaged by seismic tomography (Fig. 1). This novel approach is crucial for future understanding of the deep Earth controls on dynamic topography, formed in response to the interaction of the lithosphere and crust with thermal perturbations and convective movements in the Earth's mantle. Several studies (e.g., Faccenna et al. 2014) have explored dynamic topography of the Mediterranean area, demonstrating its sensitivity to adopted crustal models.

Formation of sedimentary basins is an intrinsic corollary to the Earth's planetary evolution, particularly of its lithosphere (Cloetingh et al. 2015). At present, knowledge on the evolution of the Earth's lithosphere is assembled in kinematic reconstructions at regional to global scales. The state-of-the-art in the fields of tectonic reconstructions and coupling deep Earth processes to surface evolution is of high international level. TOPO-EUROPE researchers have combined efforts in regional scale reconstruction with global reconstructions, utilizing Gplates (freeware, see www.gplates.org). These modern quantitative reconstructions provide an interface with thermo-mechanical modelling of plate tectonics coupled to the underlying mantle. Important advances have been made by integrating mantle structure, as a memory of Earth's dynamic evolution, with crust-lithosphere processes and tectonic reconstructions of complex plate boundary regions (Wortel and Spakman 2000) and with global tectonic reconstructions linked to mantle evolution (van der Meer et al. 2010). Crucial is the knowledge of precise mantle structure obtained from tomography. Tectonic transitions are considered to be of key importance in controlling, for instance, the preservation of hydrocarbon reservoirs and the formation of ore deposits as they lead to strong changes in the crustal stress-field, local thermal pulses from the asthenosphere, and mantle-driven crustal flow of mineral-rich fluids. In contrast to its importance, process-based understanding of basin and plate boundary deformation or tectonic



TOPO-EUROPE: From the Deep Earth to the Surface of Continental Europe and Its Margins, Fig. 1 Tomographic cross-sections for the upper mantle below Europe, illustrating heterogeneity in the upper mantle. Depth scale 600 km. Blue and red colours correspond to areas where seismic P-wave velocities are respectively higher and lower than the standard reference velocity model. The sections display distinct patterns of down-going slabs in convergent zones of Europe overlain

by lithosphere with reduced seismic velocities, corresponding with areas of high heat flow and high potential for geothermal energy exploration. The overall topography of Europe, depicted in the central panel, is characterised by elevated areas (including the Alps, Dinarides, Apennines, Carpathians, Pyrenees and the Anatolian Plateau) not only in the convergent settings, but also in intraplate settings such as Iberia and the southern Scandes. (Tomographic sections: Courtesy W. Spakman)

transitions is one of the weak links in current tectonic reconstructions.

Crucial for creating detailed process-based models of natural laboratories coupled to deep processes is the capability to embed regional modelling in a global dynamic

lithosphere-mantle framework. This provides the necessary (realistic) constraints on regional mantle flow and the stress field resulting from global plate-mantle interaction. Global and regional tomography models can now be used to estimate mantle composition, temperatures, and mantle flow, resolving

deep Earth structure in great detail and demonstrating the existence of side lobes of the Iceland Plume under intraplate areas in northwestern Europe with elevated topography (Fig. 2) (e.g., Rickers et al. 2013; Koptev et al. 2017). This potentially can be coupled to tectonic reconstructions in an absolute plate motion frame (Torsvik et al. 2010). To this aim, 3D numerical mantle modelling assimilating models of lithospheric age and thickness and incorporating evolving subduction systems, ocean spreading, and other plate boundary processes is required. The strategy of establishing a global dynamic reconstruction model is essential for predicting the impact of global plate-tectonic and mantle processes on the evolution of local subduction and other plate boundary systems as well as past tractions and temperatures at the base of lithosphere plates crucial for lithosphere-stress modelling.

Lithosphere Structure

On a completely different scale, new approaches have been developed to quantify plate rigidity on a full continental scale, resolving controversies regarding end-member rheological models for continental lithosphere (Tesauro et al. 2009). This work has also led to the insight that lower crustal flow in Europe's lithosphere and mechanical coupling between individual crustal layers is crucial for linking deep Earth processes and mantle-lithosphere interactions to surface deformation (Fig. 3) (Burov et al. 2007). Subsequent work has enabled a comparison of inferences for the effective elastic thickness of the lithosphere from admittance functions with spatial variations in integrated strength based on rock-mechanics data – on a global scale. The resulting strength maps display a robust pattern of variations in lithospheric strength, crucial for understanding localization of deformation in the interior of plates and basins, and for an assessment of stress propagation away from plate boundaries into these continental interiors.

In the context of TOPO-EUROPE, temporal changes in stress regime have also been investigated in terms of transient changes in tectonic regime in the Pyrenees (see also Erdős et al. 2014) and at the southern margin of the European continent where it interacts with the northern African margin (Faccenna et al. 2014). These models have demonstrated strong mechanical control on sedimentary basin systems and their intrinsically polyphase behavior. Stress and strain relationship and rheology are important ingredients in this context.

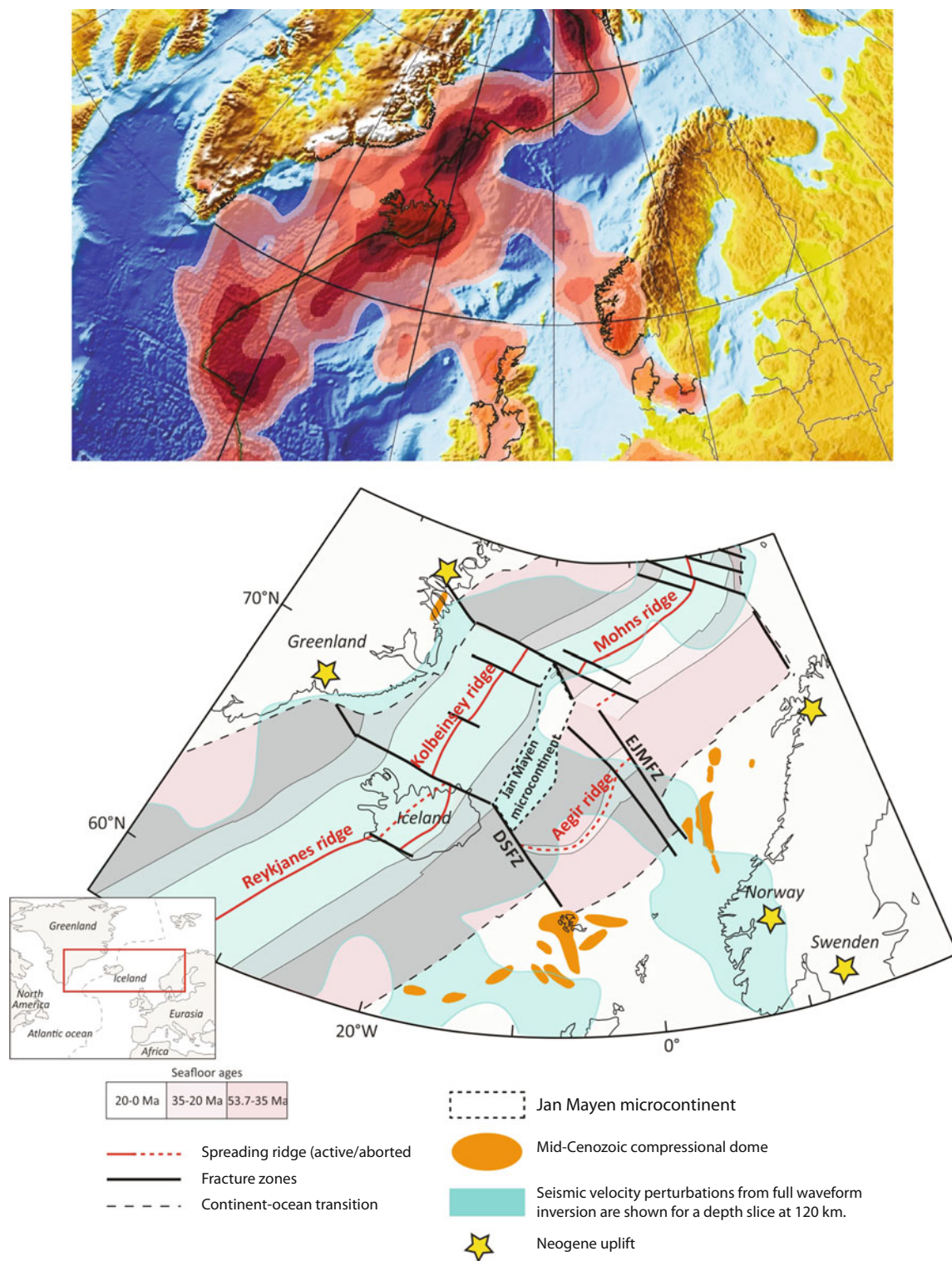
Mantle-Lithosphere Interactions

New fully consistent models of mantle-lithosphere interactions (Fig. 3) and stable and even unstable lithospheric

deformation (such as lithospheric folding) adopting realistic rheologies and integration into sedimentary basin studies (Cloetingh et al. 2015) have recently been developed. Advances in tectonic modelling and its integration with observational studies have allowed quantification of the time constants involved in the temporal succession of mantle/lithosphere interactions and of rifting and lithospheric folding by stress accumulation in intraplate continental crust and lithosphere. The results have been shown to be of critical importance for quantitatively assessing thermal regime, the record of differential vertical motions, and the faulting characteristics of super-deep basins created by the lithospheric folding and its interaction with phase changes in the crust. Key in this approach has been the integration of realistic models for Earth rheology in numerical studies of large-scale deformation and faulting and its role in intraplate structure and processes. Fully consistent 2D and 3D geodynamic numerical models incorporating thermomechanical, thermodynamic, and surface process (via implementation of free surface boundary or similar solutions) will play a key role in testing hypotheses on the topographic impact of deep processes and in separating tectonic components of deformation from dynamic topography. The latter task becomes extremely challenging since it is evident that tectonic and deep Earth processes are so interlinked that the notion of the dynamic topography will probably undergo significant changes in the nearest future.

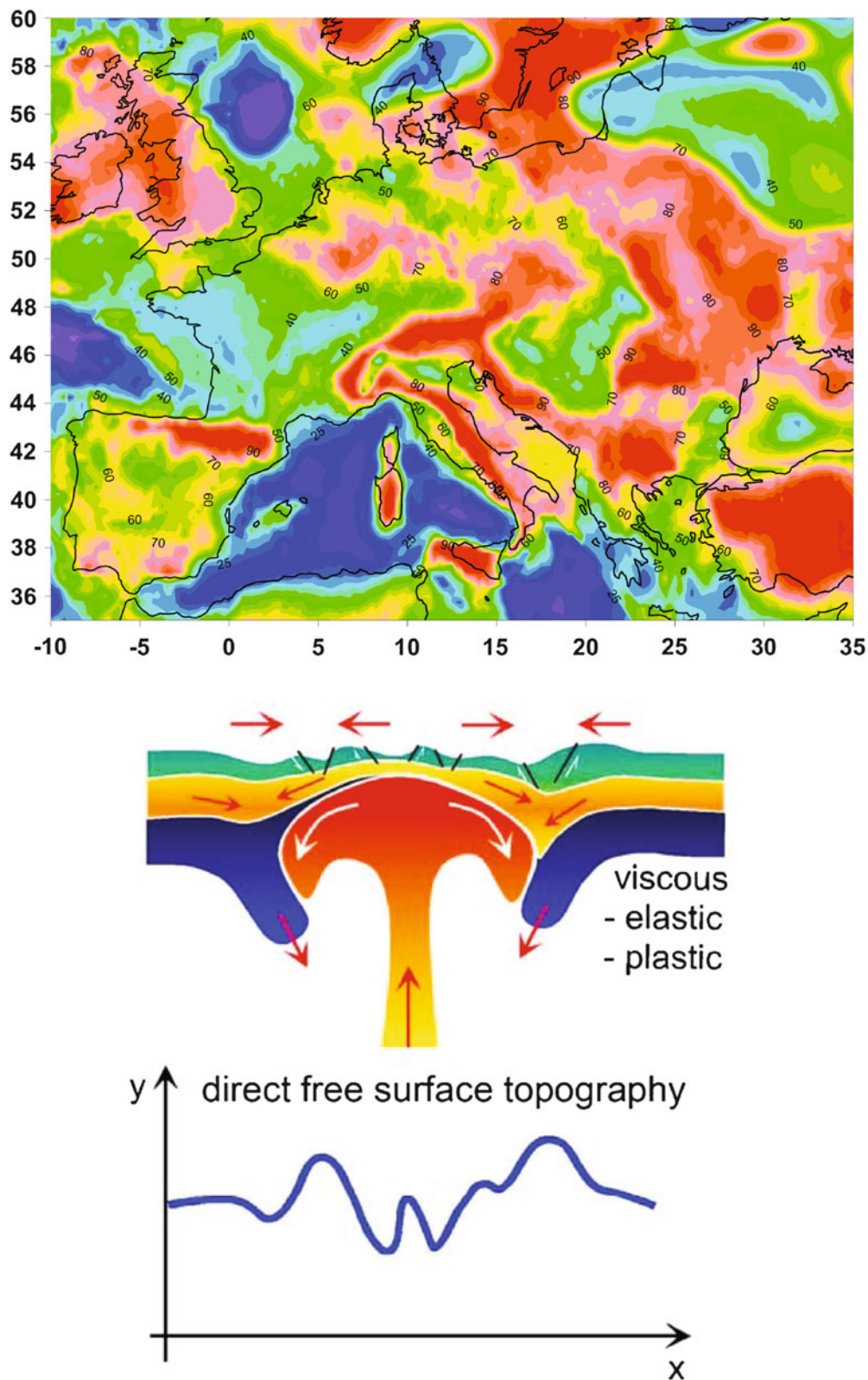
Sedimentary Basins and Geo-Resources

By their nature, sedimentary basins develop and evolve in a plate tectonic setting, characterized by a strong mechanical coupling between plate boundaries and plate interiors. Of particular relevance in this context is the subduction factory, being the prime engine for build-up and release of stresses in the crust and lithosphere (Faccenna et al. 2014). Crucial in this context is the need to quantify the mechanisms for far-field stress propagation, away from plate boundaries and the localization of the resulting deformation in critical zones for resource formation in the interior of continents and their margins. At a regional scale, lithospheric and crustal stresses are responsible for the reactivation of preexisting faults and affect fluid transfers and reservoir characteristics, in particular across areas of contrasting strength and mechanical properties. The second so far underestimated factor significantly influencing subduction factories refers to the presence of structural heritage in the subducting and overriding plate (e.g., subduction of old terrains causing poly-phase subduction and metamorphic exhumation). The interplay between deep Earth and surface processes has distinct consequences for the record of vertical motions and sedimentary basin architecture (e.g., Cloetingh et al. 2015). As a result, the impact of these processes is a key for a quantitative



TOPO-EUROPE: From the Deep Earth to the Surface of Continental Europe and Its Margins, Fig. 2 (*top panel*) Horizontal cross-section through a high resolution S-velocity model of the North Atlantic region, derived from full-waveform tomography. Highlights of the model in the upper mantle include a well-resolved Mid-Atlantic Ridge and two distinguishable strong low-velocity regions (indicated by the red colors) beneath Iceland and beneath the Kolbeinsey Ridge west of Jan Mayen. A sub-lithospheric low-velocity layer is imaged beneath much of the oceanic lithosphere, consistent with the long-wavelength bathymetric high of the North Atlantic. The low-velocity layer extends locally

beneath the continental lithosphere of the southern Scandinavian Mountains, the Danish Basin, part of the British Isles and eastern Greenland. (After Rickers et al. 2013). (*bottom panel*) Tectonic setting of North Atlantic region and low velocity anomaly at 120 km-depth. The side lobes of the Iceland plume (*top panel*) and plume channeling through fracture zones explain the joint occurrence of Late Neogene uplift in the southern Scandes and northwestern part of the British Isles and compressional intraplate doming in the offshore part of the Northern Atlantic volcanic rifted margin system (Koptev et al. 2017)



TOPO-EUROPE: From the Deep Earth to the Surface of Continental Europe and Its Margins, Fig. 3 (top) Map of Europe with colour contours indicating the relative contribution of the crust to the rheological strength of the entire lithosphere (after Tesauro et al. 2009). Dark blue colours indicate that the strength distribution is dominated by the mantle lithosphere, whereas red and yellow colours indicate lithosphere where strength is primarily concentrated in the crust ("crème brûlée" model). Green and light blue coloured areas indicate a stratified

lithosphere with rheologically strong layers in both crust and upper mantle lithosphere ("jelly sandwich" model). (middle and bottom) Models that include a plate rheology and multilayer lithosphere structure predict a complex surface topography with several short-wavelengths generated by intraplate deformation, tectonic-style deformation at surface and strong lithosphere mantle erosion at depth. The short wavelengths are controlled by the thickness of the lithosphere and intra-lithospheric layers. (After Burov et al. 2007)

assessment of resource and storage potential (i.e., aquifers, ore deposits, geothermal energy, hydrocarbons, and greenhouse gases). Sedimentary fluxes are often characterized by episodic moments of rapid changes in tectonic, drainage network distribution, oceanic circulation, or connectivity between basins. The latter are, however, commonly studied in isolation and require constraints on their temporal evolution frequently beyond the resolution of existing dating methods. As a result of the current mismatch between their intrinsic temporal and spatial scales, incorporation of these elements in predictive models for geo-resources is far from complete.

Analogue Tectonic Modeling

European laboratory groups involved in the EU TOPOMOD international training network have all contributed at the international forefront of developments in their respective activities at the associated length scales. Using physical analogue, or scale model experiments, these researchers have delivered major contributions to understanding tectonic processes and faulting at the basin scale, providing a basis for unravelling geological and geophysical observations on basin and crustal structure, and for testing and improving numerical models of basin evolution (e.g., Cloetingh et al. 2015).

Novel tectonic concepts and their implementation in coupled numerical and analogue modelling have opened new approaches to study the thermomechanical behavior of the Earth's lithosphere and to assess the role and interaction of parameters such as intraplate stress, rock rheology, and lithosphere structure. Advanced numerical methods have been implemented for the quantitative analysis of dynamic models that address the link between lithosphere deformation processes and vertical and horizontal motions, including constraints from GPS data (see, e.g., Serpelloni et al. 2013) in space and time. Analogue modelling has been shown to be particularly useful in qualitative and quantitative studies of lithosphere deformation in settings with a complex 3D geometry, such as present in the European Alps where seismic tomography interpretations have yielded evidence for a flip in the polarity of the down-going slab (Fig. 4) (Luth et al. 2013). An extensive database compiled in the TOPO-EUROPE project for the Alpine uplift and erosion history is now available (Fig. 5) (Fox et al. 2015) to further test different models and scenarios proposed for Late Neogene tectonics and climate of this key mountain system.

Coupling of experimental predictions from analogue models into numerical models has led to a better

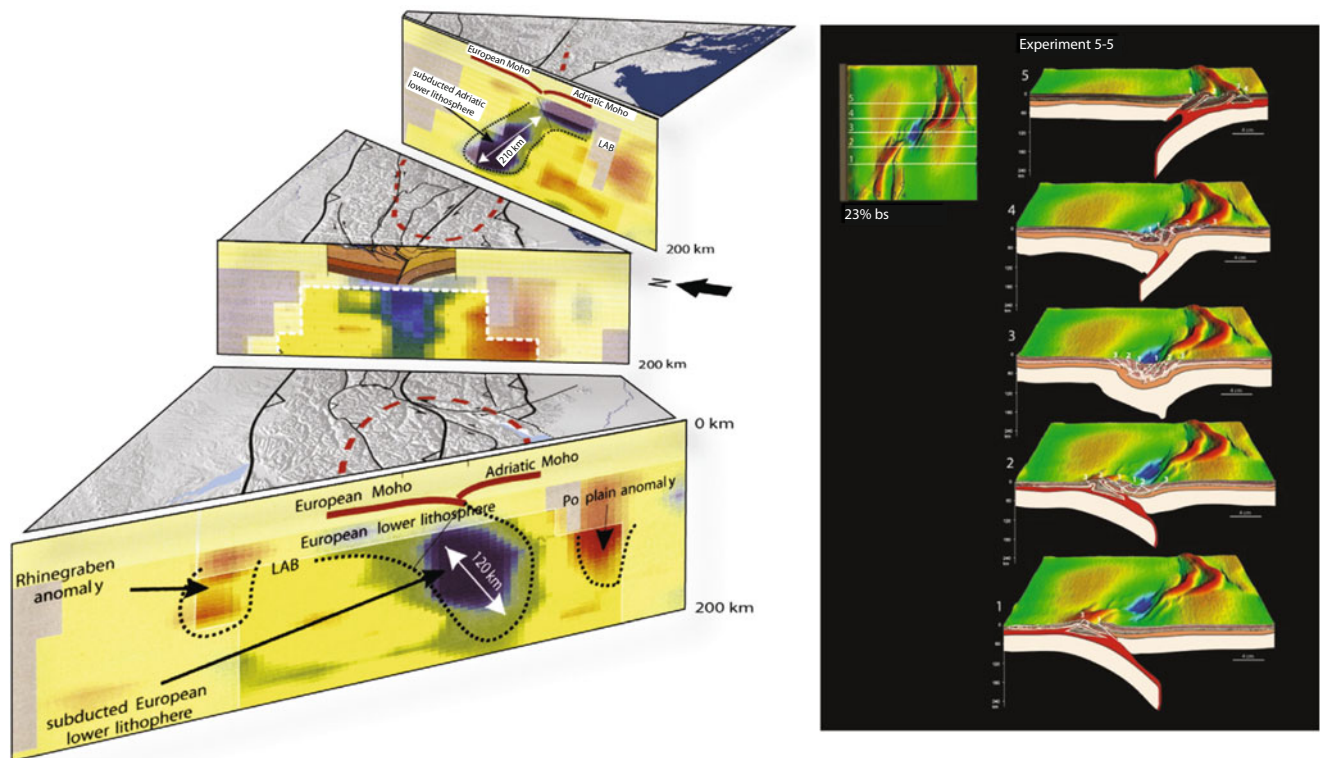
understanding of the interplay between the formation and evolution of sedimentary basins and of the mechanisms that control stress and strain distribution at crustal and lithospheric scale. This has resulted in novel predictions of depositional system behavior and hydrocarbon reservoir architecture.

It has been clearly demonstrated that sedimentary basin systems cannot be treated in isolation but should be treated by connecting their basin fill architecture to the evolution of adjacent basins and topographic highs, accounting for their hydrological connectivity (Matenco and Andriessen 2013). This research has allowed quantification of how different types of basin connectivity affect budgets and pathways of sedimentary fluxes, including source magnitudes and mechanisms of sediment depocenter shift across basins. Spatial variations in the rheology of the underlying crust appear to be key to basin connectivity and the sediment storage capacity of individual basins. These findings are also important for assessing the impact of differential vertical motions in highly populated areas affected by flooding events, regional landslides, and active seismicity (Matenco and Andriessen 2013).

Summary and Conclusions

Research on coupled deep Earth and surface processes has developed as a frontier in integrated solid Earth science. The European continent and its margins provide an excellent natural laboratory to test in particular process-oriented models, such as for instance the effects of spatial variations in lithosphere strength on the mode and surface expression of mantle-lithosphere interactions. This research is by its nature multiscale and requires a close integration of observations, monitoring, and modelling.

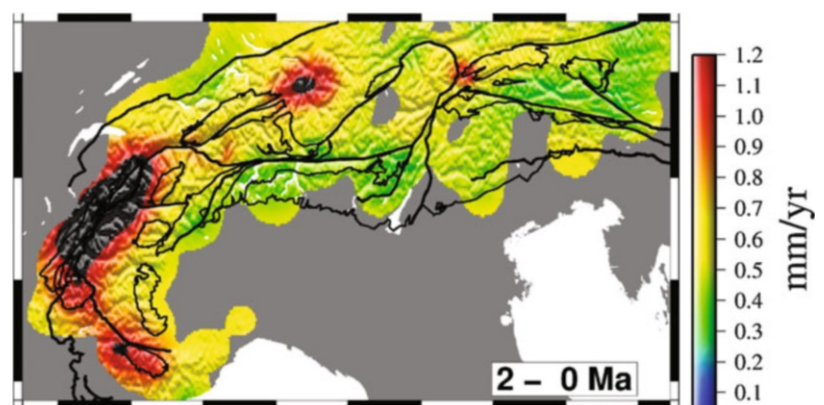
TOPO-EUROPE research has demonstrated, amongst others, an important control by side-lobes of the Iceland mantle plume on anomalous topography of the Southern Scandes of Norway and the offshore Northern Atlantic compressional domes. Slab detachment and slab tearing processes exert pronounced effects on the 4D evolution of the Mediterranean, such as for instance a polarity swap in subduction beneath the Alps. In Iberia, lithospheric folding is a key contributor to differential topography with an average anomalous elevation due to the interaction of intraplate deformation with plate edge driven thermal anomalous in the upper mantle. This research has important spin-offs in the domains of geo-hazards (i.e., volcanism, earthquakes, and landslides) and geo-resources, in particular the exploration and exploitation of geothermal energy, highly relevant for the present era of energy transition to a more sustainable world.



TOPO-EUROPE: From the Deep Earth to the Surface of Continental Europe and Its Margins, Fig. 4 (left) Tele-seismic tomographic cross-sections display along-strike variations in continental subduction polarity. Blue and red colors refer to relatively high and low P wave velocities, respectively. Top views show the major tectonic boundaries as black lines and the outline of the subducted slabs between 135 and 165 km depth in dashed red lines transposed on a DEM. (a) (bottom) Lower lithosphere transects through western Central Alps imaging in blue the SE-dip of subducted European lower lithosphere. Bold dashed line indicates lithosphere–asthenosphere boundary. (b) (middle) TRANSALP crustal section transposed on a deep tomographic transect

disclosing a steep dipping slab. (c) (top) Lower lithosphere transect through Eastern Alps imaging NE-dip of subducted Adriatic lower lithosphere. (right) Cross-sections located along white profile lines shown in top view inset. Within the cross-sections white material corresponds to mantle lithosphere, beige to strong lower crust, red to weak zone, and colored sand to upper crust. Black line represents Moho. White lines are interpreted faults, which are chronically numbered. A DEM overlies the cross-sections (both panels after Luth et al. 2013). See Laubscher (2010) and Roure et al. (2010) for constraints on crustal structure from deep seismic reflection programs across the Alps

TOPO-EUROPE: From the Deep Earth to the Surface of Continental Europe and Its Margins, Fig. 5 Alpine exhumation rates over the past 2 My, inferred by analysis of over 2500 thermo-chronometric ages. (After Fox et al. 2015)



Acknowledgments This review is based on the dedicated research efforts of the TOPO-EUROPE community. Funding by 23 national research councils in Europe and the European Community is gratefully acknowledged. ESF provided valuable coordination during the TOPO-EUROPE EUROCORES program. The International Lithosphere Program and the Academia Europaea are thanked for continuous support during various phases of the lifespan of TOPO-EUROPE. Constructive criticism by François Roure is greatly appreciated. Wim Spakman and Fred Beekman are acknowledged for suggestions and technical support.

Bibliography

- Burov E, Guillou-Frottier L, d'Acremont E, Le Pourhiet L, Cloetingh S (2007) Plume head – lithosphere interactions near intra-continental plate boundaries. *Tectonophysics* 434:15–38
- Cloetingh S, Tibaldi A (eds) (2012) Coupled deep Earth and surface processes in system Earth: monitoring, reconstruction and process modeling. *Glob Planet Change* 90–91, 158 pp
- Cloetingh S, Ziegler P, Bogaard P, Andriessen P, Artemieva I, Bada G, van Balen R, Beekman F, Ben-Avraham Z, Brun J-P, Bunge H-P, Burov E, Crabonell R, Faccenna C, Friedrich A, Gallart C, Green A, Heidbach O, Jones A, Matenco L, Mosar J, Oncken O, Pascal C, Peters G, Sliampa S, Soesoo A, Spakman W, Stephenson R, Thybo H, Torsvik T, de Vicente G, Wenzel F, Wortel M (2007) TOPO-EUROPE: the geoscience of coupled deep Earth-surface processes. *Glob Planet Change* 58:1–118
- Cloetingh S, Thybo H, Faccenna C (eds) (2009) TOPO-EUROPE: studying continental topography and Deep Earth – surface processes in 4D. *Tectonophysics* 474, 416 pp
- Cloetingh S, Vicente G de, Gallart J (eds) (2011) TOPO-EUROPE II – from Iberia to the Carpathians and analogues. *Tectonophysics* 502 (1–2), 266 pp
- Cloetingh S, Willett S, Torsvik T, Werner S (eds) (2013) TOPO-EUROPE III – understanding of the coupling between the deep Earth and continental topography. *Tectonophysics* 602, 384 pp
- Cloetingh S, Ziegler PA, Beekman F, Burov EB, Garcia-Castellanos D, Matenco L (2015) Tectonic models for the evolution of sedimentary basins. *Treatise on geophysics*, 2nd edn. Elsevier, Oxford
- Cloetingh S, Tibaldi A, Dobrzynetska L, Matenco L, Fadi N, Wijck de Vries, B (eds) (2018) From the deep Earth to the surface – a multiscale approach. *Glob Planet Change* 171, 322 pp
- Colli L, Ghelichkhan S, Bunge H-P, Oeser J (2018) Retrodictions of mid-Paleogene mantle flow and dynamic topography in the Atlantic region from compressible high resolution adjoint mantle convection models: sensitivity to deep mantle viscosity and tomographic input model. *Gondwana Res* 53:252–272
- Erdős Z, Huismans RS, Van der Beek P, Thieulot C (2014) Extensional inheritance and surface processes as controlling factors of mountain belt structure. *J Geophys Res Solid Earth* 119:9042–9061
- Faccenna C, Becker TW, Auer L, Billi A, Boschi L, Brun JP, Capitano FA, Funicello F, Horvath F, Jolivet L, Piromallo C, Royden L, Rossetti F, Serpelloni E (2014) Mantle dynamics in the Mediterranean. *Rev Geophys* 52:283–332
- Fox M, Herman F, Kissling E, Willett SD (2015) Rapid exhumation in the Western Alps driven by slab detachment and glacial erosion. *Geology* 43:379–382
- Koptev A, Cloetingh S, Burov E, François T, Gerya T (2017) Long-distance impact of Iceland plume on Norway's rifted margin. *Sci Rep* 7:10408. <https://doi.org/10.1038/s41598-017-07523-y>
- Laubscher H (2010) Jura, Alps and the boundary of the Adria subplate. *Tectonophysics* 483:223–239
- Luth S, Willingshofer E, Sokoutis D, Cloetingh S (2013) Does subduction polarity change below the Alps? Inferences from analogue modeling. *Tectonophysics* 582:140–161
- Matenco L, Andriessen P (2013) Quantifying the mass transfer from mountain ranges to deposition in sedimentary basins: source to sink studies in the Danube Basin – Black Sea system. *Glob Planet Change* 103, 260 pp
- Rickers F, Fichtner A, Trampert J (2013) The Iceland Jan Mayen plume system and its impact on mantle dynamics in the North Atlantic region: evidence from full-waveform inversion. *Earth Planet Sci Lett* 367:39–51
- Roure F, Cloetingh S, Scheck-Wenderoth M, Ziegler PA (2010) Achievements and challenges in sedimentary basin dynamics. A review. In: Cloetingh S, Negendank J (eds) *New frontiers in integrated solid Earth sciences*. Springer, pp 145–233. <https://www.springer.com/gp/book/9789048127368>
- Serpelloni E, Faccenna C, Spada G, Dong D, Williams SDP (2013) Vertical GPS ground motion rates in the Euro-Mediterranean region: new evidence of velocity gradients at different spatial scales along the Nubia-Eurasia plate boundary. *J Geophys Res Solid Earth* 118:6003–6024
- Tesauro M, Kaban MK, Cloetingh S (2009) How rigid is Europe's lithosphere? *Geophys Res Lett* 36:1–6, L16303
- Torsvik TH, Steinberger B, Gurnis M, Gaina C (2010) Plate tectonics and net lithosphere rotation over the past 150 My. *Earth Planet Sci Lett* 291:106–112
- Van der Meer DG, Spakman W, Van Hinsbergen DJJ, Amaru ML, Torsvik TH (2010) Towards absolute plate motions constrained by lower-mantle slab remnants. *Nat Geosci* 3:36–40
- Wortel R, Spakman W (2000) Subduction and slab detachment in the Mediterranean-Carpathian region. *Science* 290:1910–1917

Total Electron Content in Seismo-ionospheric Studies

J. K. Catherine and R. Rajesh
CSIR-NGRI, Hyderabad, India

Synonyms

Ionospheric seismology

Definition

Total electron content (TEC): It is a quantifiable parameter of Earth's ionosphere, which appear in five layers (D, E, Es, F1, and F2) between altitudes ~50–600 km. It represents integrated number of electrons between satellite and receiver along the ray path in a vertical column of unit cross section measured in multiples of the TEC unit defined as $1\text{TECU}=10^{16}\text{el/m}^2$. Although TEC accounts for integrated electron content along the ray path, it is mainly sensitive to the electron content of F2 layer of ionosphere, where the maximum ionization takes place.

Seismo-ionospheric: The processes involved in lithosphere–ionosphere relationship are believed to be associated with tectonic forces acting on the Earth's crust (e.g., seismic

activity/earthquakes and volcanic eruptions). The release of large amount of energy in such processes (acoustic/mechanical) disturb the ionospheric electron concentration. Persuasive correlations are observed between the seismic activities and the ionospheric disturbances preceding and during earthquakes (Davies and Baker 1965; Ryu et al. 2014).

Shock Acoustic Waves (SAW): Ionospheric Response of the Shock Wave

Co-seismic Ionospheric Disturbances (CID): Disturbance in the ionospheric electron content caused by the atmospheric pressure waves that propagate upward into the ionosphere due to co-seismic vertical motion of the ground during an earthquake.

Travelling Ionospheric Disturbances (TIDs): Manifestation of atmospheric waves excited by earthquakes, in the ionosphere.

Ionospheric Pierce Point (IPP) or Seismo-ionospheric Point: The point of intersection of the signal transmitted from satellite to the receiver with the thin ionospheric shell.

Sub-ionospheric Point (SIP): The footprint/projection of the seismo-ionospheric point on to the Earth's surface.

Slant TEC: Integrated TEC along the path connecting ground-based receiver to a GPS satellite

Vertical TEC: The vertical component of the slant TEC at the intercept of the slant path on the ionospheric surface

Plasma: An ionized gas of zero net electric charge comprising positively charged ions and free electrons that exist either at low pressure or high temperature conditions.

Geomagnetic Cyclotron Frequency: The frequency of a charged particle moving perpendicular to the direction of Earth's magnetic field.

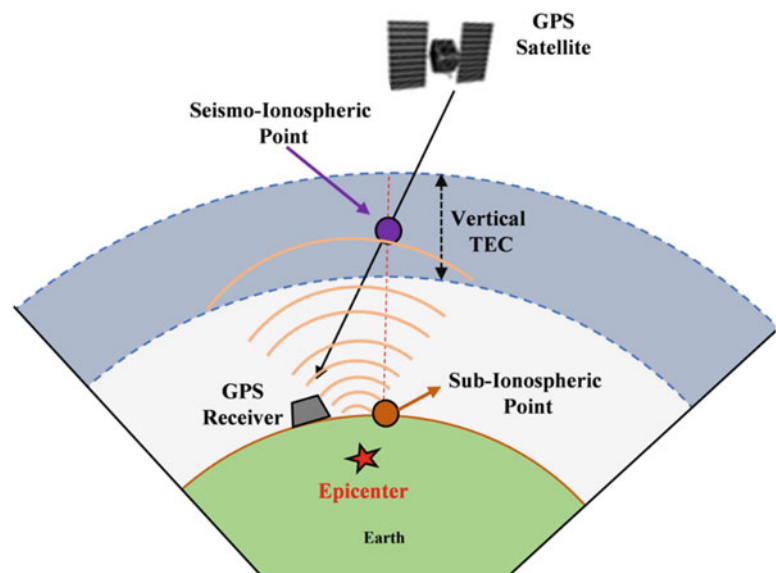
Introduction

The ionosphere, a shell of ionized gas plasma, is a dispersive medium that delay GPS (Global Positioning System) signals. The ionospheric delay along the path from the GNSS (Global Navigation Satellite System) satellite to a GPS receiver is effectively eliminated by a dual-frequency GPS receiver ($f_1 = 1575.42$ MHz on the L1 band and $f_2 = 1227.60$ MHz on the L2 band). This delay is proportional to the TEC, which depends on the elevation angle, azimuth of the Line-of-Sight (LOS), altitude, local time, intensity of solar activity, season, and receiver position and helps to quantify ionospheric disturbances.

The atmospheric infrasonic waves excited by the natural (earthquakes, tsunamis, volcanic eruptions, polar aurora, convective storms) or artificial (rocket launches, powerful explosions) sources causing large vertical near-field displacement of the Earth's surface generate laterally and upward propagating acoustic and gravity oscillations with 10–15 min delay as Shock Acoustic Waves (SAW) away from the source with a velocity equal to the speed of sound (Blanc 1985) in the ionospheric F2layer (Fig. 1). The amplitude of the SAW is inversely proportional to the density which decreases exponentially with altitude and amplified by a factor of 10^4 at ionospheric heights, where the coupling between neutral and ionized atmosphere results in fluctuations of the ionospheric electron density (Calais and Minister 1995, 1998). In response to the large amplitude impulsive forcing by SAW, an N shaped waveform is generated in the ionospheric plasma (Zel'dovich and Raizer 1966) in which the compression is followed by the “recovery” of the medium through rarefaction

Total Electron Content in Seismo-ionospheric Studies,

Fig. 1 Schematic diagram illustrating vertical motion of seismic waves inducing atmospheric acoustic gravity waves



(Landau and Lifshitz 1995). Such disturbances in the ionosphere due to the mechanical coupling between the solid earth and the atmosphere are interpreted as Co-seismic Ionospheric Disturbances (CID) (Watada and Kanamori 1994).

It is generally perceived that dip-slip earthquakes of magnitude $M > 7$ occurring in subduction zones are expected to induce TEC fluctuations (Calais and Minister 1995; Heki and Ping 2005; Rolland et al. 2013). Although the static vertical crustal displacements in strike-slip earthquakes are rather small, still there are evidences of clear CIDs (Astafyeva et al. 2014; Catherine et al. 2015). Analyses of CIDs of earthquakes from different regions with magnitudes from 4.1 to 8.8 suggest that the earthquakes of magnitude $M_w \geq 6.5$ only elicit CIDs (Perevalova et al. 2014).

The expansion of regional and international GNSS networks of closely spaced GPS stations with continuous recordings with accurate phase measurements, generally at an interval of 30 s, provide an opportunity to investigate characteristics of ionospheric perturbation such as apparent propagation velocities, directionality, waveform, and initial motion of CID with high spatio-temporal resolution both near the epicenter and far away from it. Beginning from the great Alaskan earthquake (1964) (Davies and Baker 1965; Leonard and Barnes 1965), several studies have been carried out to detail CIDs due to natural and artificial sources (Bolt 1964; Row 1967; Yuen et al. 1969; Tanaka et al. 1984; Calais and Minister 1995, 1996; Calais et al. 1998; Afraimovich et al. 2002; Ducic et al. 2003; Liu et al. 2004; Artru et al. 2005; Heki and Ping 2005; Heki et al. 2006; Otsuka et al. 2006; Astafyeva et al. 2009, 2014; Ozeki and Heki 2010; Afraimovich et al. 2010; Astafyeva and Heki 2011; Rolland et al. 2010; Liu et al. 2010a, b; Galvan et al. 2011; Liu et al. 2011a; Heki and Enomoto 2013; Cahyadi and Heki 2013; Su et al. 2013; Rolland et al. 2013; Catherine et al. 2015, 2017; Zhang and Tang 2015; Reddy et al. 2016; Bagiya et al. 2018, etc.). These studies serve as a proxy to investigate the coupling and energy transfer processes in the Lithosphere, Atmosphere, and Ionosphere (LAI) coupled system and earthquake source mechanism characteristics. Analyses of period, velocity, and propagation direction of the CIDs help to distinguish the ionospheric disturbances linked with geomagnetic storms, solar flares, etc. Only those perturbations without significant solar and geomagnetic activity may be attributed to earthquakes. The general dynamics of the ionosphere can be studied from TEC Global maps and their computation algorithm is presented in Mannucci et al. (1998) and Schaer et al. (1998).

Methodology for Estimating Total Electron Content (TEC) from GPS Data

In case of electro-magnetic wave propagation in a plasma medium like ionosphere, the frequency of the wave is much

higher than both geomagnetic cyclotron frequency and Earth's ionospheric plasma frequency. The excess time delay due to propagation in the dispersive medium can be approximately expressed as

$$t_{ion} = \frac{e^3}{8\pi^2 m \epsilon_0 c f^2} \int_{\text{receiver}}^{\text{satellite}} N(z) dz \quad (1a)$$

where e and m are the charge and mass of electrons, ϵ_0 is the dielectric constant of vacuum, c is the speed of light, f is the radio wave frequency, and the integral of $N(z)dz$ is the Integrated electron content (IEC) in the propagation path (Sekido et al. 2003) between the satellite and the receiver. We can appropriately write the equation (1a) as follows

$$t_{ion} = \frac{40.3}{c f^2} * IEC \quad (1b)$$

Based on the equation 1, there are different approaches for TEC computation (Calais and Minister 1995; Afraimovich et al. 2001; Carrano and Groves 2009). Dual frequency measurements (1.2 and 1.5 GHz) from GPS satellites located ~20,000 km above the earth's surface provide integrated information of the ionosphere by computing the differential code and carrier phase measurements recorded by ground-based GPS receivers. These measurements compensate for the ionospheric delay by taking advantage of dispersive nature of the ionosphere.

The basic equation for the integrated electron content and its relation with phase ($\emptyset$) and code (p) measurements is given by

$$IEC_{\emptyset} = \frac{1}{40.3} \left[\frac{f_1^2 f_2^2}{f_1^2 - f_2^2} \right] (L_1 \lambda_1 - L_2 \lambda_2) + B \quad (2)$$

$$IEC_p = \frac{1}{40.3} \left[\frac{f_1^2 f_2^2}{f_1^2 - f_2^2} \right] (P_2 - P_1) + B \quad (3)$$

Here, L_1 and L_2 are the carrier phase measurements, P_1 and P_2 are code measurements, and λ_1 and λ_2 are signal wavelengths corresponding to the frequencies f_1 and f_2 . The product, $L_i \lambda_i$ (where 'i' takes the value 1 and 2, respectively, corresponding to f_1 and f_2) is the additional path of radio signal caused by the phase delay in the ionosphere and B is an unknown constant related to initial phase path caused by the cumulative result of indefinite number of phase rotations along the line of sight (Afraimovich et al. 2001; Carrano and Groves 2009). The relative IEC (IEC_R) can be obtained after removing the noise in CODE and ambiguity in the phase measurements by applying a phase leveling along the phase connected arc (Vijayan et al. 2013; Catherine et al. 2015).

$$IEC_R = IEC_{\emptyset} + \langle IEC_p - IEC_{\emptyset} \rangle_{arc} \quad (4)$$

Finally, the TEC can be obtained by scaling the IEC_R by elevation mapping function (emf_{θ}) as follows

$$TEC = \frac{IEC_R}{emf_{\theta}} \quad (5)$$

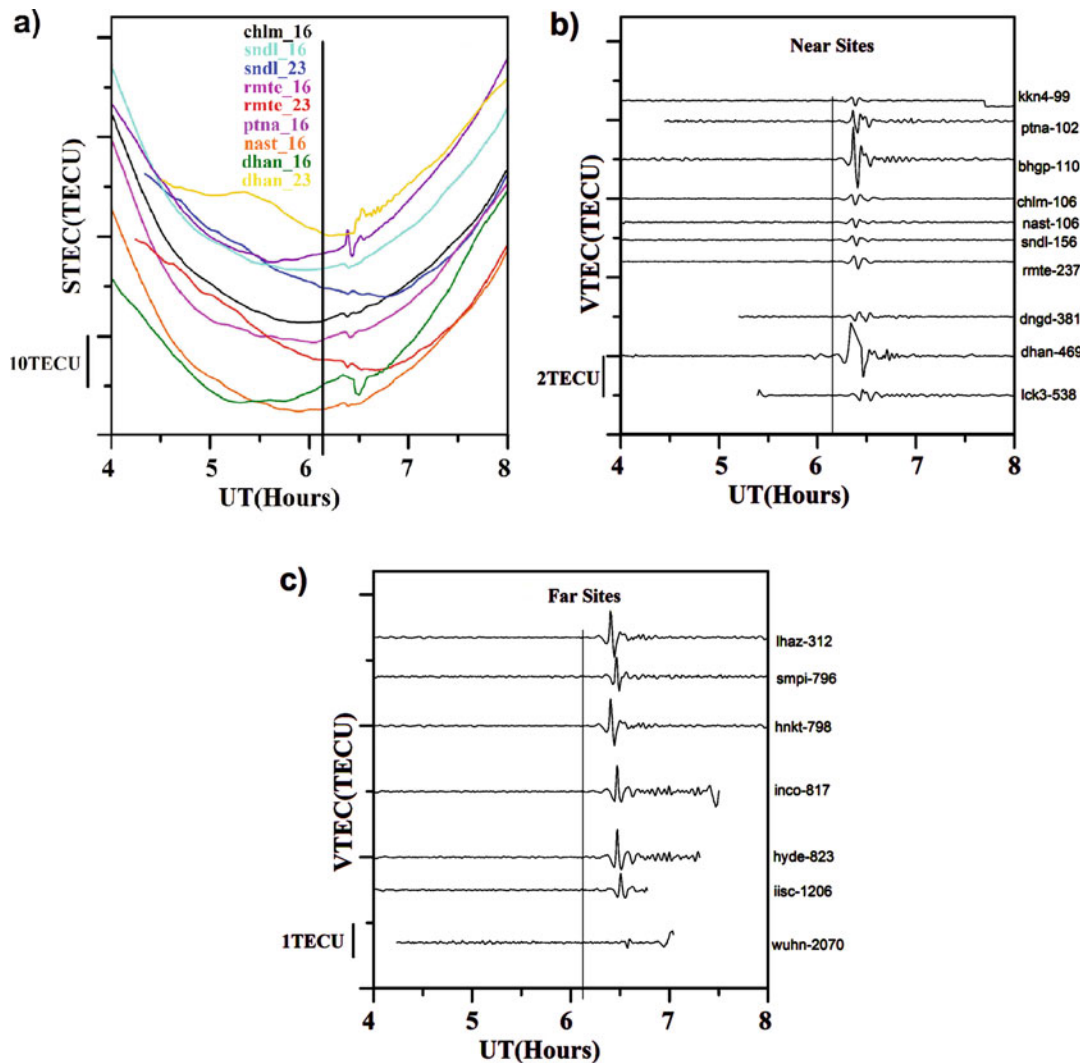
$$emf_{\theta} = \sec \left[\sin^{-1} \left(\frac{R}{R + h_{max}} \right) \cos \theta \right]. \quad (6)$$

$R \rightarrow$ radius of the Earth; $h_{max} \rightarrow$ height of maximum ionization.

This scaling helps to account for the discrepancy in the ray path length through the ionosphere as a function of GPS

satellite elevation angle θ during the course of an orbital pass. As the TEC represents the integrated electron content along the path, it is not possible to determine meaningfully the height of TEC anomaly. However, the TEC variations are mainly sensitive to the height (h_{max}) of the maximum ionization, that is, in F2 layer. The TEC represents a point of intersection of a line-of-sight within the thin F2 layer.

The CIDs are calculated from TEC by smoothing the noise of phase measurements and detrending the variations of regular ionosphere using band pass filtering of 1–10 min period. An example of raw slant TEC (Fig. 2a) time series recorded by commonly visible satellites for April 25, 2015, Gorkha earthquake, Nepal (Catherine et al. 2017) along with filtered TEC time series of sites in the near (Fig. 2b) and far (Fig. 2c) fields of the epicenter are shown in Fig. 2.



Total Electron Content in Seismo-ionospheric Studies, Fig. 2 (a) Time series of the raw slant TEC changes for the April 25, 2015, Gorkha earthquake, Nepal. Traces of band pass filtered (1–10 min) coseismic TEC variation (b) at sites close to the epicenter and (c) far from the

epicenter, as observed by GPS satellites at some of the stations. The vertical line indicates occurrence of the Mw 7.8 Gorkha earthquake (06:11 UT). Distance of the SIPS of GPS sites from the epicenter is indicated on the right side

Propagation Characteristics of CIDs

An anomalous signal is observed on TEC time series of sites that are both in the near and far fields of the source of disturbance, between 10 and 30 min after the earthquake. This signal shows time delays that increase with distance from the epicenter, consistent with the velocity of the propagating perturbation. Irrespective of the source (natural or artificial), the TEC anomalies show the character of a classical N-wave of period 180–360 s, for 30 min to 1 h with amplitude above the background variations under moderate geomagnetic conditions by a factor of 2–5 as a minimum. Such organized perturbation will be absent in the preceding and following days of the earthquake. The spatiotemporal characteristics of CIDs give some information about the source of disturbances.

Amplitudes of CID

The CID spreads out from the epicenter with a quasi-circular propagation pattern with a time lag equivalent to the time required for acoustic waves to reach IPP from the surface. The absolute and relative amplitude of CID increases with the increasing magnitude. The amplitude of acoustic mode of CID decreases quickly with time and distance of SIP's location and the epicenter up to 500 km. The surface wave mode of amplitudes increase from 500 to 1000 km (Fig. 2c). Multi-frequency anomalies are induced after the earthquake in the near and far fields (Fig. 2). Acoustic waves, Rayleigh waves (3–7 mHz) tsunami-generated gravity waves (<3 mHz) (Matsumura et al. 2011; Occhipinti et al. 2013), and also high frequencies signals (7–15 mHz) can be identified in the GPS TEC measurements. Negative or positive CID anomalies indicate either decrease or increase in TEC of ionospheric plasma medium by the rarefaction or compression induced by the acoustic wave front respectively for normal or reverse fault mechanisms. CIDs of negative anomalies are unstable and only visible for large earthquake (Astafyeva and Heki 2009). The polarities of ionospheric disturbance waveforms (Fig. 1b, c in Astafyeva and Heki 2009) are mostly consistent with the coseismic vertical ground motion and focal mechanism of earthquakes. The observed TEC response has N-wave and inverted N-wave for uplift and subsidence of the ground after the earthquake respectively. CID polarities are clear for a dip-slip earthquake, but in the case of strike-slip earthquake, visibility of the polarity depends on the magnitude of the vertical displacement generated by an earthquake.

However, discrepancies in the amplitude and waveforms of CIDs are plausible due to factors like complexity of rupture, distance from GPS station to epicenter, sampling interval, azimuth of the SIP, geometry of LOS, geomagnetic conditions (Afraimovich et al. 2001; Astafyeva and Heki 2009; Heki and Ping 2005) earthquake magnitude (Astafyeva et al. 2013a) attendant differences in the incidence

angles (shallow or deep) between the corresponding LOS and the wave front (Cahyadi and Heki 2013). Further, the difference in the apparent CID amplitudes at the same site for different satellites can also stem from the three-dimensional structure of the CID in the ionosphere of finite thickness (Heki and Ping 2005).

Generally, other than earthquakes, solar eclipses can cause an ionospheric anomaly of 0.15–15 TECU, tropical cyclone can perturb ionosphere at period 1–150 min with anomaly of ~2.5 TECU, geomagnetic storms can perturb ionosphere about 10–14% of its TEC, and rocket launching can cause nearly 0.2–2% change in ionospheric TEC at periods 3–8 min (Afraimovich et al. 2013). Studies suggest that the Earthquakes and tsunamis can cause 1.8–6% of TEC variation at 4–5 min period and 0.15TECU change around 9–14 min (Heki et al. 2006; Occhipinti et al. 2013; Astafyeva et al. 2014; Grawe and Makela 2015). Few examples of the earthquakes of different mechanism, Tsunamis, Solar Eclipse, rocket launch for which the CIDs were observed are charted in Table 1.

Propagation Speed of CIDs

The apparent velocity of CID can be obtained from the slope of the best fit line of travel time diagram, a plot between arrival times of CID at different subionospheric points versus the distances of these points from the center of crustal uplift along the great circle path.

The CIDs are classified into three types based on their characteristics and source of generation

- (i) The long period CIDs induced by large coseismic vertical crustal movement around the epicenter travels with speeds ~0.3 km/s (Bolt 1964) are the signatures of acoustic gravity waves (Davies 1990). They induce quasi-periodic ionosphere disturbances that are seen in the far-field of an earthquake without significant damping and changes of oscillation spectrum (Astafyeva and Afraimovich 2006; Liu et al. 2006b).
- (ii) The second type is the most general and obvious CIDs detected 10–15 min after strong earthquakes in the near fields caused by shock acoustic waves (SAW). They propagate with velocity 0.9–1.2 km/s from the earthquake location in a narrow cone of zenith angles (Calais and Minster 1995; Afraimovich et al. 2001; Heki and Ping 2005; Jung et al. 2006).
- (iii) The third type is short period CIDs induced by surface Rayleigh waves. They propagate along the Earth's surface with a velocity 2.6–3.6 km/s and recorded at the far-field stations (Yuen et al. 1969; Davies 1990; Ducic et al. 2003; Artru et al. 2004; Liu et al. 2006a, b).

Figure 3 shows the CID propagation patterns after the 11 April 2012 Indian ocean earthquake (Mw 8.6) induced

Total Electron Content in Seismo-ionospheric Studies, Table 1 Ionospheric disturbances caused by various natural and artificial sources

Sl. No	Event	Date and time (UTC)	Faulting	CID-AT	CID in TECU	References
1	Northridge (CA) earthquake (Mw 6.7)	17-01-1994 12:30	T	~10 min	~1 to 3% (change)	Calais and Minster (1995)
2	Kunlun earthquake (Mw 7.8)	14-11-2001 09:26	SS	~13 min	~0.66	Astafyeva et al. (2014)
3	Denali Earthquake (M 7.9)	03-11-2002 22:12	SSS	~11 min	~0.05	Ducic et al. (2003)
4	Tokachi-oki Earthquake (Mw 8.0)	25-09-2003 19:50	ST	~10 mins	~0.1–0.2	Heki and Ping (2005)
5	Sumatra earthquake (Mw 9.3)	26-12-2004 00:58	T	~10–15 min	~1.6–6.9	Liu et al. (2006a, b), Occhipinti et al. (2006)
6	Sichuan earthquake, China (Mw 7.9)	12-05-2008 06:28	SOR	~7.9 min	~1	Song et al. (2015)
7	Tohoku Earthquake (Mw 9.0)	11-03-2011 05:46	ST	~7 min	~0.7–2.3	Tsugawa et al. (2011), Astafyeva et al. (2011), Liu et al. (2011a)
8	Indian Ocean Earthquake (Mw 8.6)	12-04-2012 08:38	SS	~10 min	~0.02–0.7	Catherine et al. (2015)
9	Pakistan Earthquake (Mw 7.7)	24-09-2013 11:29	SS	~12 min	~1.4	Astafyeva et al. (2014)
10	Gorkha earthquake (Mw 7.8)	25-04-2015 06:11	T	5–10 min	~0.06–1.55	Catherine et al. (2017)
11	Illapel earthquake, Chile (Mw 8.3)	26-09-2015 22:54	T	~10 mins	~1.4	Reddy et al. (2016)
12	Kaikoura earthquake, Newzealand (Mw 7.8)	13-11-2016 11:02	SOR	8 min	~0.22	Lee et al. (2018), Li et al. (2018), Bagiya et al. (2018)
TEC variations due to events other than earthquakes						
13	Kwangmyongsong-4, Rocket launch	07-02-2016 00:30		~6 min	~7	Choi and Kil (2017)
14	Total solar eclipse	21-08-2017 17:15		~20 min,	~6–9	Bhaskar et al. (2018)
15	Mine blast, Black Thunder coal mine, eastern Wyoming	Aug-96		~10 min	~0.03	Calais et al. (1998)
16	Eruption of the Calbuco volcano	22-04-2015		~15 min	~0.45	Shults et al. (2016)

T: Thrust; SS: strike-slip; SSS: shallow strike-slip; ST: shallow thrust; SOR: shallow oblique reverse; CID-AT: CID arrival time after the event

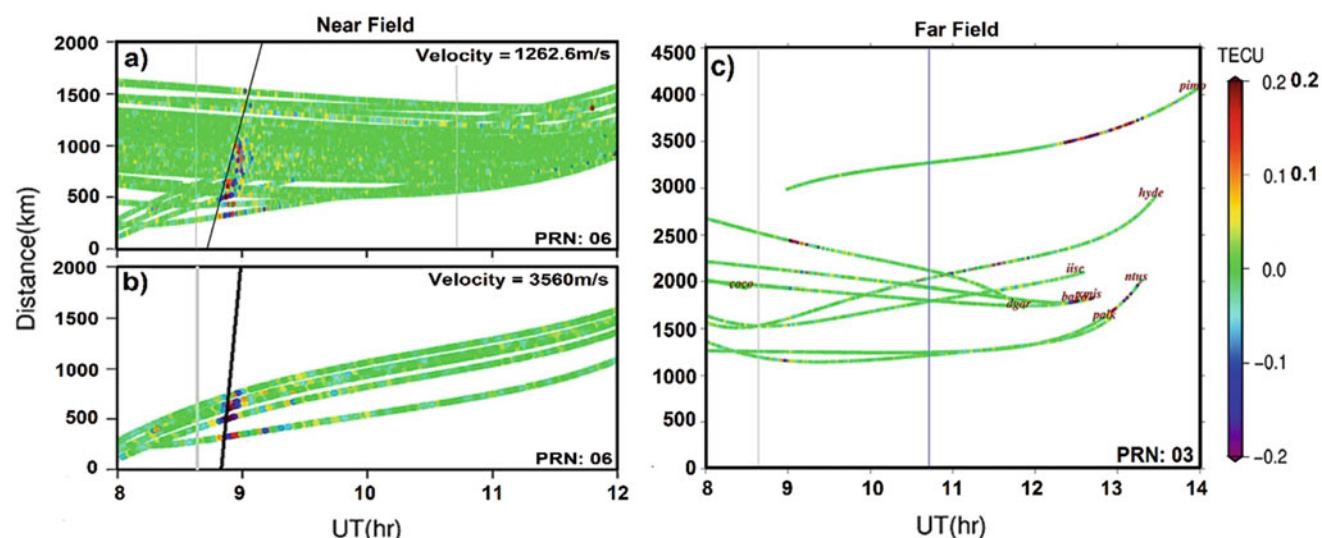
by SAW at 1.2 km/s (Fig. 3a) and Rayleigh wave at 3.5 km/s (Fig. 3b) CID arrivals excited by Rayleigh waves are recorded at the far-field stations (Fig. 3c). The mainshock was followed by an aftershock of Mw 8.2 after 2 hours. The stations HYDE and PIMO seem to illustrate quasi-periodic TEC oscillations that continued for 60–75 min at epicentral distances of ~2000 km and 3500 km, respectively.

The first report on separation of CID into two modes was given by Astafyeva et al. (2009). The initial disturbance propagates with an apparent velocity equal to the speed of the sound at the height of ionosphere's F-layer, up to a distance of ~650 km from the epicenter. But from ~650 to 700 km away from the epicenter, the disturbance appears to split into two separate waves known as fast and slow components with velocities ~3 km/s and 600 m/s, generated by Rayleigh surface waves and vertical motion of the ground, respectively. Later this phenomenon was observed in other studies too (Catherine et al. 2015, 2017). Although it is not possible to discriminate these two components in the near-field, they can be seen in the TEC data after 600–700 km.

Further the fast component waves are clearly seen in the far-field due to their long wavelength and slow component gradually slows down and fades with increasing distance from the epicenter. The possible reasons for the appearance of the slow component could be due to multiple sources like distance from the epicenter as well as the LOS geometry and the azimuthal difference in energy emission of seismic waves (Calais et al. 1998). Two signals with different propagation velocities appear in the TEC time series near the epicenter (Fig. 3a, b) but only long wavelength signal is seen at far field sites which confirms the attenuation of shock acoustic waves at such distances (Fig. 3c). In some cases, Rayleigh wave signatures may not appear at all because of their small amplitudes coming from the prevailing geometric conditions (Cahyadi and Heki 2013) or absence of GPS points farther (>2000 km) from the epicenter.

Directivity of CID Propagation

Operation of dense regional and international GPS networks in a continuous mode enables the accurate mapping of



Total Electron Content in Seismo-ionospheric Studies, Fig. 3 Traveltime diagram of band pass-filtered TEC in the near field (a,b) and in the far-field (c) regions in response to the April 11, 2012, earthquake at the SIPs to the epicenter as recorded by the satellites PRN 6. Vertical thick lines indicate the time of the main shock and the

aftershock at 8:38UT and 10:43UT, respectively. The slanted black line shows the best fit of the CID arrival times denoting the propagation velocities of ionospheric TEC disturbance. The distances are great circle distances between SIPs and the epicenter

directivity of CID propagation. However, paucity of observations over a wide azimuth during the past earthquakes made estimates of CID directionality contentious. As the Lorentz force on the ionospheric charged particles in geomagnetic field hampers their motion perpendicular to geomagnetic field, the ambient magnetic field in the Northern hemisphere attenuates northward propagation of CID and favors southward propagation and allows the northward propagation in Southern hemisphere (Heki and Ping 2005). In the polar region, the ambient magnetic field which is vertical would attenuate the propagation of CID for all directions, whereas in the equatorial regions where the ambient magnetic field is horizontal, the CID would propagate in both north and south directions. Astafyeva et al. (2014) proposed equatorward propagation for much longer distances.

The directivity of CID propagation reported for the Southern Hemisphere 2007 Bengkulu earthquake (Cahyadi and Heki 2013) was in support of the above proposed model. However, azimuthal anisotropy of TEC variations also reported (Otsuka et al. 2006). Catherine et al. (2015) noticed the differences in Rayleigh wave radiation pattern as strong reason for the anisotropy than to the geomagnetic field effect. According to Calais et al. (1998), the ionospheric coupling factor (ICF) controls the ionospheric perturbation directivity. Attenuation of ionospheric coupled wave is less for regions with high ICF (Rolland et al. 2011). At midlatitudes, the CID propagating at angles larger than $\pm 10^\circ$ and $\pm 30^\circ$ with ambient magnetic field lines are attenuated by a factor of 0.5 and 0.25, respectively, and for CID propagating at $\sim \pm 45^\circ$ to the

magnetic field lines the amplitude is close to 0 (Rolland et al. 2013). Calais et al. (1998) suggested two possibilities for the north-south asymmetry in the ionospheric response: (i) the interaction with ambient geomagnetic field and (ii) azimuthal difference in energy emission in terms of seismic wave.

Observations on directivity of CIDs for some of the earthquakes occurring in different geographic regions are tabulated in Table 2. The source directivity-related effects appear to emerge as the potential driving mechanism to explain our observation. From future study of CIDs occurring in different geographical locations, directionality of CID propagation and the influencing factors would be better visualized.

Dip-slip Versus Strike-slip Earthquakes

In dip-slip earthquakes, vertical coseismic crustal displacement is the predominant characteristic that causes uplift or subsidence at the surface depending on reverse or normal motion of the faults, respectively. In strike-slip motion, horizontal motion is dominating with very small vertical motion. Piston like coseismic vertical motion in the earthquake zone plays a crucial role in generating perturbations in the ionosphere. In addition to the TEC fluctuations observed for dip-slip earthquakes ($M > 7$) occurring in subduction zones, strike-slip earthquakes also induce significant CIDs (Astafyeva et al. 2014; Catherine et al. 2015).

Although the CID amplitudes of strike-slip events are less than the dip-slip earthquakes, the characteristics are similar

Total Electron Content in Seismo-ionospheric Studies, Table 2 Details on directivity of CID for earthquakes occurring in different geographic regions

Earthquake	Region	Directionality of CID propagation	Rupture direction	References
Tokachi-Oki earthquake (Mw 8.0) September 26 2003	Northern Hemisphere	Northward CID smaller than the southward CID	NW	Yagi (2004)
Tohoku Earthquake (Mw 9.0) March 11 2011	Northern Hemisphere	Amplitude of northward CID (A_N) is smaller than the southward CID (A_S)	NW	Tsugawa et al. (2011)
Sumatra earthquake (Mw 9.3) Indian Ocean December 26 2004 and 2007 Bengkulu earthquake of Mw 8.5	Equatorial Region	Amplitude of northward CID is more than the southward CID, but could not be confirmed due to the absence of stations south of the epicenter	Northward	Otsuka et al. (2006) Cahyadi and Heki (2013)
Indian Ocean Earthquake (Mw 8.6) April 11 2012	Equatorial Region	CID propagates effectively to the north and south of the epicenter with different velocities A_N being larger than A_S	Rupture geometry is complex and propagated in all directions	Catherine et al. (2015)
Gorkha earthquake (Mw 7.8) April 25 2015	Mid latitude region	No northward propagation. Mostly CID propagated in south south east	South south east	Reddy and Seemala (2015) Catherine et al. (2017)
Illapel earthquake, Chile (Mw 8.3) September 26 2015	Southern Hemisphere	North direction with more north velocity with decreasing velocity toward south	North	Reddy et al. (2016)
Kaikoura earthquake, Newzealand (Mw 7.8) November 14 2016	Southern Hemisphere	North ward propagation of CID	Complex rupture in North East	Bagiya et al. (2018)

(Perevalova et al. 2014). The observed TEC response for dip-slip earthquakes has a precise N-shape and inverted N-shape waves over the regions of uplift and subsidence, respectively, whereas the CID polarities for uplift and subsidence are not clearly noticeable for strike slip events. However, the amplitude of ionospheric disturbance depends on many other factors as described in section “[Amplitudes of CID](#)”. Astafyeva et al. (2014) disclosed that the seismological parameters like magnitude of coseismic horizontal displacements, fault and slip dimensions, or a combination of them may play a role in generating large ionospheric perturbations in large strike-slip events with insignificant vertical ground motion.

Preseismic Ionospheric Anomalies

The study of the ionosphere state prior to the occurrence of large earthquakes has gained attention of the geophysicists, as the precursory studies, on success, may help to minimize the loss during great earthquakes. Availability of thousands of ground-based GPS receivers, on global to regional scale today, opened an opportunity to extensively study and validate the precursory seismo-ionospheric anomalies. It has been found that the ionospheric plasma critical frequency (f_oF_2) diminishes few days to few hours before some large earthquakes in the vicinity of an earthquake epicenter.

Interestingly, investigations of preseismic ionospheric disturbances are also gaining impetus (Liu et al. 2002, 2004; Pulinets et al. 2007; Dautermann et al. 2007; Zhao et al. 2008; Liu et al. 2009, 2010b, 2011b; Heki 2011; Kamogawa and Kakinami 2013; Heki and Enomoto 2015; Masci et al. 2015).

Wyss (1991) proposed that TEC enhancement prior to a large earthquake must satisfy the four requirements to be validated as an earthquake precursor. Le et al. (2011) noticed preseismic TEC anomalies tend to appear around the earthquake epicenter 1–4 days prior, particularly with a higher probability for large and shallow earthquakes. Positive TEC anomalies were noticed 60–40 min before the 2011 Tohoku-Oki, 2004 Sumatra Andaman and the 2010 Maule earthquakes (Heki 2011). The physical mechanism and the appearance of the preseismic TEC anomalies still remains unclear in spite of several new studies.

Besides the positive results on the seismo-ionospheric precursory studies for the short-term prediction, research on coining this information as precursors is underway and still remains as a matter of debate. However, in view of the socio-economical perspective, focused research on seismo-ionospheric precursors could be encouraged to develop accurate regional model that discriminates seismo-ionospheric precursory changes from the other natural and anthropogenic variations.

Influence of Equatorial Ionization Anomaly on CID

EIA (equatorial ionization anomaly) is a spatial variability in ionization density and TEC due to the “fountain effect” setup by the interaction of the diurnal variation of the zonal electric field with the horizontal geomagnetic field at the equatorial region causing uplift of the plasma by $E \times B$ drift (Anderson 1981; Walker et al. 1994). A well-formed EIA must be present in the equatorial region at the local noon time and may influence the spatial variability of TEC amplitude. This may be assessed by computing the percentage of relative TEC perturbation over the background TEC at each SIP during the time of observation.

$$\text{Relative dTEC} = (\text{dTEC}/\text{TEC}) * 100$$

Comparison between dTEC and relative dTEC will help to check the difference in the amplitude of the propagating CID from the background TEC and assess the influence of EIA.

Absence of CIDs

The co-seismic ionospheric perturbations are being captured by the growing GNSS network and its applications are gaining importance to investigate various coupled processes associated with the ionosphere. Nevertheless, Equatorial Plasma Bubble (EPBs), a stable and strong mode of ionospheric irregularities that develop during the postsunset period due to Rayleigh-Taylor (R-T) instability, diffract the signals from navigation and communication systems (Elgowainy and Ashgriz 1997). These EPBs obscure the signatures of CIDs due to the higher electron production rate of EPBs. Therefore, it appears difficult to capture CIDs in the GNSS signals (e.g., 2005, Nias earthquakes, Sumatra; 2017 Mexico earthquake). However, due to the lower production rate of EPBs at the high-latitudinal region, GNSS TEC measurements expose lucid CIDs even during night time (e.g., Iceland, Alaska) (Panda et al. 2019).

Applications of CIDs in Ionospheric Seismology

The various possibilities of GNSS ionospheric imaging are

- (i) As the location of initial TEC increase is associated with the crustal uplift, the first CID arrivals provide the seismic fault location and dimensions within few minutes after the earthquake (Astafyeva et al. 2011; Astafyeva and Shults 2019).
- (ii) The coupling of solid earth-ocean-atmosphere helps to estimate the earthquake source extent using CIDs observed from large GNSS continuous networks and Multi GNSS constellations.

- (iii) CID has a potential to provide information on rupture speed and relative magnitudes of multiple asperities (Heki et al. 2006; Astafyeva et al. 2013b)
- (iv) CID could serve as a new indicator of rupture on the fault with timescale up to 4–5 min in between seismic band (Broadband seismometers that are sensitive to periods longer than a few minutes) and tsunami band (data useful to constrain faulting as slow as a few tens of minutes) (Heki et al. 2006).
- (v) As the validation of tsunami risk in tsunamigenic earthquakes using sparsely located buoys around the ocean takes hours, the CIDs observed approximately ~8–9 min after the onset of earthquake can be utilized in the oceanic monitoring and tsunami warning system (Occhipinti et al. 2013)

Summary

Recent studies on co-seismic ionospheric disturbances are robust for understanding the coupling between lithosphere and ionosphere. Particularly, tremendous increase in the GNSS networks in the last decade, on the global scale, helped to improve the accuracy in estimates of propagation direction and velocity. The ionospheric perturbations imaged by dense GPS arrays act as a proxy to study energy transfer between the solid Earth and ionosphere (lower atmosphere) for distinguishing ionospheric signals of terrestrial versus extra-terrestrial origin. The spatio-temporal and dynamic features of CIDs due to natural and artificial sources and the factors influencing them have been extensively investigated within the last decade by using signals of Global Navigation Satellite Systems. The CID anomalies for different types of earthquake mechanism and tsunamis are analyzed to understand the scaling between anomaly strength and source parameter of the events. Physiognomies of ionospheric response to the earthquake are appropriate not only to decipher earthquake location, fault length, and other source characteristics but also help to assess the tsunamigenic risk. In addition, systematic analyses of pre-seismic ionospheric anomalies, if successful, will help to reduce the socio-economic losses associated with earthquakes.

Cross-References

► GPS, Data Acquisition, and Analysis

Acknowledgments The authors are grateful to Director, CSIR-NGRI for his kind permission. The Ministry of Earth Sciences, Government of India is thankfully acknowledged for sponsoring Tectonic Geodesy Research at CSIR-NGRI.

Bibliography

- Afraimovich EL, Perevalova NP, Plotnikov AV, Uralov AM (2001) The shock-acoustic waves generated by earthquakes. *Ann Geophys* 19:395–409
- Afraimovich EL, Kosogorov EA, Plotnikov AV (2002) Shockacoustic waves generated at rocket launches and earthquakes. *Cosm Res* 40:241–254
- Afraimovich EL, Ding F, Kiryushkin VV, Astafyeva EI, Jin S, Sankov VA (2010) TEC response to the 2008 Wenchuan earthquake in comparison with other strong earthquakes. *Int J Remote Sens* 31(13):3601–3613
- Afraimovich E, Astafyeva E, Demyanov V, Edemskiy I, Gavriluk N et al (2013) A review of GPS/GLONASS studies of the ionospheric response to natural and anthropogenic processes and phenomena. *J Space Weather Space Clim* 3:A27
- Anderson DN (1981) Modeling the ambient, low latitude F-region ionosphere—a review. *J Atmos Sol Terr Phys* 43:753–762
- Artru J, Farges T, Lognonne P (2004) Acoustic waves generated from seismic surface waves: propagation properties determined from Doppler sounding observations and normal-mode modeling. *Geophys J Int* 158:1067–1077
- Artru J, Ducic V, Kanamori H, Lognonne P, Murakami M (2005) Ionospheric detection of gravity waves induced by tsunamis. *Geophys J Int* 160:840–848
- Astafyeva EI, Afraimovich EL (2006) Long-distance propagation of traveling ionospheric disturbances caused by the great Sumatra Andaman earthquake on 26 December 2004. *Earth Planets Space* 58(8):1025–1031
- Astafyeva E, Heki K (2009) Dependence of waveform of near-field coseismic ionospheric disturbances on focal mechanisms. *Earth Planets Space* 61:939–943
- Astafyeva E, Heki K (2011) Vertical TEC over seismically active region during low solar activity. *J Atm Terr Phys* 73:1643–1652
- Astafyeva E, Shults K (2019) Ionospheric GNSS imagery of seismic source: possibilities difficulties and challenges. *J Geophys Res Space Phys* 124:534–543. <https://doi.org/10.1029/2018JA026107>
- Astafyeva E, Heki K, Kiryushkin V, Afraimovich E, Shalimov S (2009) Two-mode long-distance propagation of coseismic ionosphere disturbances. *J Geophys Res* 114:A10307. <https://doi.org/10.1029/2008JA013853>
- Astafyeva E, Lognonné P, Rolland L (2011) First ionosphere images for the seismic slip on the example of the Tohoku-oki earthquake. *Geophys Res Lett* 38:L22104. <https://doi.org/10.1029/2011GL049623>
- Astafyeva E, Shalimov S, Olshanskaya E, Lognonné P (2013a) Ionospheric response to earthquakes of different magnitudes: larger quakes perturb the ionosphere stronger and longer. *Geophys Res Lett* 40(9):1675–1681. <https://doi.org/10.1002/grl.50398>
- Astafyeva E, Rolland L, Lognonné P, Khelifi K, Yahagi T (2013b) Parameters of seismic source as deduced from 1 Hz ionospheric GPS data: case-study of the 2011 Tohoku-oki event. *J Geophys Res Space Phys* 118:5942–5950. <https://doi.org/10.1002/jgra50556>
- Astafyeva E, Rolland L, Sladen A (2014) Strike-slip earthquakes can also be seen in the ionosphere. *Earth Planet Sci Lett* 405:180–193. <https://doi.org/10.1016/j.epsl.2014.08.024>
- Bagiya MS, Sunil PS, Sunil AS, Ramesh DS (2018) Coseismic contortion and coupled nocturnal ionospheric perturbations during 2016 Kaikoura, Mw7.8 New Zealand earthquake. *J Geophys Res Space Phys* 123:1477–1487. <https://doi.org/10.1002/2017JA024584>
- Bhaskar K, Panda D, Gahalaut VK, Catherine JK (2018) The 2017 August 21 American total solar eclipse through the eyes of GPS. *Geophys J Int* 214:651–655. <https://doi.org/10.1093/gji/ggy149>
- Blanc E (1985) Observations in the upper atmosphere of infra-sonic waves from natural or artificial sources: a summary. *Ann Geophys* 3(6):673–688
- Bolt BA (1964) Seismic air waves from the great 1964 Alaskan earthquake. *Nature* 202:1094–1095
- Cahyadi MN, Heki K (2013) Ionospheric disturbances of the 2007 Bengkulu and the 2005 Nias earthquakes, Sumatra, observed with a regional GPS network. *J Geophys Res Space Phys* 118:1777–1787. <https://doi.org/10.1002/jgra.50208>
- Calais E, Minister JB (1995) GPS detection of ionospheric perturbations following the January 17, 1994, Northridge earthquake. *Geophys Res Lett* 22:1045–1048. <https://doi.org/10.1029/95GL00168>
- Calais E, Minister JB (1996) GPS detection of ionospheric perturbations following space shuttle ascent. *Geophys Res Lett* 23:1897–1900
- Calais E, Minister JB (1998) GPS, earthquakes, the ionosphere, and the space shuttle. *Phys Earth Planet Inter* 105:167–181
- Calais E, Minister JB, Hofton MA, Hedlin MAH (1998) Ionospheric signature of surface mine blasts from global positioning system measurements. *Geophys J Int* 132:191–202
- Carrano C, Groves K (2009) Ionospheric data processing and analysis. In: Workshop on Satellite Navigation Science and Technology for Africa, The Abdus Salam ICTP, Trieste
- Catherine JK, Vijayan MSM, SyedaRabiya UB, Shimna K, Gahalaut VK, Ramesh DS (2015) Dichotomy in mode propagation of coseismic ionospheric disturbance: observations from 11 April 2012 Indian Ocean earthquake. *J Geophys Res Space Phys* 120(5):3854–3867. <https://doi.org/10.1002/2014JA020621>
- Catherine JK, Uma Maheshwari D, Gahalaut VK, Roy PNS, Khan PK, Puviarasan N (2017) Ionospheric disturbances triggered by the 25 April 2015 M7.8 Gorkha earthquake, Nepal: constraints from GPS TEC measurements. *J Asian Earth Sci* 133:80–88. <https://doi.org/10.1016/j.jseas.2016.07.014>
- Choi BK, Kil H (2017) Large ionospheric TEC depletion induced by the 2016 North Korea rocket. *Adv Space Res* 59(2):532–541
- Dautermann T, Calais E, Haase J, Garrison J (2007) Investigation of ionospheric electron content variations before earthquakes in southern California 2003–2004. *J Geophys Res* 112:B02106. <https://doi.org/10.1029/2006JB004447>
- Davies K (1990) Ionospheric radio. Peter Peregrinus, London
- Davies K, Baker DM (1965) Ionospheric effects observed around the time of the Alaskan earthquake of March 28, 1964. *J Geophys Res* 70(9):2251–2253
- Ducic V, Artru J, Lognonne P (2003) Ionospheric remote sensing of the Denali earthquake Rayleigh surface waves. *Geophys Res Lett* 30(18):1951. <https://doi.org/10.1029/2003GL017812>
- Elgowainy A, Ashgriz N (1997) The Rayleigh–Taylor instability of viscous fluid layers. *Phys Fluids* 9(6):1635–1649
- Galvan DA, Komjathy A, Hickey MP, Mannucci AJ (2011) The 2009 Samoa and 2010 Chile tsunamis as observed in the ionosphere using GPS total electron content. *J Geophys Res* 116:A06318. <https://doi.org/10.1029/2010JA016204>
- Grawe MA, Makela JJ (2015) The ionospheric responses to the 2011 Tohoku 2012 Haida Gwaii, and 2010 Chile tsunamis: effects of tsunami orientation and observation geometry. *Earth Space Sci* 2:472–483. <https://doi.org/10.1002/2015EA000132>
- Heki K (2006) Explosion energy of the 2004 eruption of the Asama volcano, Central Japan, inferred from ionospheric disturbances. *Geophys Res Lett* 33:L14303. <https://doi.org/10.1029/2006GL026249>
- Heki K (2011) Ionospheric electron enhancement preceding the 2011 Tohoku-Oki earthquake. *Geophys Res Lett* 38:L17312. <https://doi.org/10.1029/2011GL047908>
- Heki K, Enomoto Y (2013) Preseismic ionospheric electron enhancements revisited. *J Geophys Res Space Phys* 118(10):6618–6626
- Heki K, Enomoto Y (2015) Mw dependence of the preseismic ionospheric electron enhancements. *J Geophys Res Space Phys* 120(8):7006–7020. <https://doi.org/10.1002/2015JA021353>

- Heki K, Ping J (2005) Directivity and apparent velocity of the coseismic ionospheric disturbances observed with a dense GPS array. *Earth Planet Sci Lett* 236:845–855
- Heki K, Otsuka Y, Choosakul N, Hemmakorn N, Komolmis T, Maruyama T (2006) Detection of ruptures of Andaman fault segments in the 2004 great Sumatra earthquake with coseismic ionospheric disturbances. *J Geophys Res* 111:B09313. <https://doi.org/10.1029/2005JB004202>
- Jung TK, Liu JY, Tsai HF, Huang BS, Lin CH, Yu SB, Yeh YS (2006) Ionospheric disturbances triggered by the Mw7.6 earthquake off the coast of El Salvador on 13 January 2001. *Terr Atmos Ocean Sci* 17:345–351
- Kamogawa M, Kakinami Y (2013) Is an ionospheric electron enhancement preceding the 2011 Tohoku-Oki earthquake a precursor? *J Geophys Res Space Phys* 118:1751–1754. <https://doi.org/10.1002/jgra.50118>
- Landau LD, Lifshitz EM (1995) Fluid mechanics. Course of theoretical physics, vol 6, 1st edn. Pergamon Press, London. ISBN 978-008-009104-4
- Le H, Liu JY, Liu L (2011) A statistical analysis of ionospheric anomalies before 736 M6.0+ earthquakes during 2002–2010. *J Geophys Res* 116:A02303. <https://doi.org/10.1029/2010JA015781>
- Lee RF, Rolland LM, Mykesell TD (2018) Seismo-ionospheric observations, modeling and backprojection of the 2016 Kaikoura earthquake. *Bull Seismol Soc Am* 108(3B):1794–1806. <https://doi.org/10.1785/0120170299>
- Leonard RS, Barnes RA Jr (1965) Observations of ionospheric disturbances following the Alaska earthquake. *J Geophys Res* 70:1250–1253
- Li JD, Rude CM, Pankratius V (2018) Characterizing the complex two N-wave ionospheric signature of the 2016 Kaikoura earthquake. *J Geophys Res Space Phys* 123(12):10–358
- Liu JY, Chuo YJ, Pulnits SA, Tsai HF, Zeng X (2002) A study on the TEC perturbations prior to the rei-Li, chi-chi and Cgia-Yi earthquakes. In: Hayakawa M, Molchanov OA (eds) *Seismo electromagnetics: lithosphere-atmosphere-ionosphere coupling*. TERRAPUB, Tokyo, pp 297–301
- Liu JY, Chuo YJ, Shan SJ, Tsai YB, Chen YI, Pulnits SA, Yu SB (2004) Pre-earthquake ionospheric anomalies registered by continuous GPS TEC measurement. *Ann Geophys* 22:1585–1593
- Liu JY, Tsai YB, Chen SW, Lee CP, Chen YC, Yen HY, Chang WY, Liu C (2006a) Giant ionospheric disturbances excited by the M9.3 Sumatra earthquake on 26 December 2004. *Geophys Res Lett* 33:L02103. <https://doi.org/10.1029/2005GL023963>
- Liu JY, Tsai YB, Ma KF, Chen YI, Tsai HF, Lin CH, Kamogawa M, Lee CP (2006b) Ionospheric GPS total electron content (TEC) disturbances triggered by the 26 December 2004 Indian Ocean tsunami. *J Geophys Res* 111:A05303. <https://doi.org/10.1029/2005JA011200>
- Liu JY, Chen YI, Chen CH, Liu CY, Chen CY, Nishihasi M, Li JZ, Xia YQ, Oyama KI, Hattori K, Lin CH (2009) Seismoionospheric GPS total electron content anomalies observed before the 12 May 2008 Mw7.9 Wenchuan earthquake. *J Geophys Res* 114:A04320. <https://doi.org/10.1029/2008JA013698>
- Liu JY, Tsai HF, Lin CH, Kamogawa M, Chen YI, Lin CH, Huang BS, Yu SB, Yeh YH (2010a) Coseismic ionospheric disturbances triggered by the Chi-Chi earthquake. *J Geophys Res* 115:A08303. <https://doi.org/10.1029/2009JA014943>
- Liu JY, Chen YI, Chen CH, Hattori K (2010b) Temporal and spatial precursors in the ionospheric global positioning system (GPS) total electron content observed before the 26 December 2004 M9.3 Sumatra–Andaman Earthquake. *J Geophys Res* 115:A09312. <https://doi.org/10.1029/2010JA015313>
- Liu JY, Chen CH, Lin CH, Tsai HF, Chen CH, Kamogawa M (2011a) Ionospheric disturbances triggered by the 11 March 2011 M9.0Tohoku earthquake. *J Geophys Res* 116:A06319. <https://doi.org/10.1029/2011JA016761>
- Liu JY, Le H, Chen YI, Chen CH, Liu L, Wan W, Su YZ, Sun YY, Lin CH, Chen MQ (2011b) Observations and simulations of seismoionospheric GPS total electron content anomalies before the 12 January 2010 M7 Haiti earthquake. *J Geophys Res* 116:A04302. <https://doi.org/10.1029/2010JA015704>
- Mannucci AJ, Wilson BD, Yuan DN, Ho CH, Lindqwister UJ, Runge TF (1998) A global mapping technique for GPS-derived ionospheric TEC measurements. *Radio Sci* 33(8):565–582
- Masci F, Thomas JN, Villani F, Secan JA, Rivera N (2015) On the onset of ionospheric precursors 40 min before strong earthquakes. *J Geophys Res Space Phys* 120. <https://doi.org/10.1002/2014JA020822>
- Matsumura M, Saito A, Iyemori T, Shinagawa H, Tsugawa T, Otsuka Y, Nishioka M, Chen CH (2011) Numerical simulations of atmospheric waves excited by the 2011 off the Pacific coast of Tohoku earthquake. *Earth Planets Space* 63(7):885–889
- Occhipinti G, Lognonné P, Kherani EA, Hébert H (2006) Three-dimensional waveform modeling of ionospheric signature induced by the 2004 Sumatra tsunami. *Geophys Res Lett* 33:L20104. <https://doi.org/10.1029/2006GL026865>
- Occhipinti G, Rolland L, Lognonné P, Watada S (2013) From Sumatra 2004 to Tohoku-Oki 2011: the systematic GPS detection of the ionospheric signature induced by tsunamigenic earthquakes. *J Geophys Res Space Phys* 118:3626–3636. <https://doi.org/10.1002/jgra.50322>
- Otsuka Y, Kotake N, Tsugawa T, Shiokawa K, Ogawa T, Saito S, Kawamura M, Maruyama T, Hemmakorn N, Komolmis T (2006) GPS detection of total electron content variations over Indonesia and Thailand following the 26 December 2004 earthquake. *Earth Planets Space* 58:159–165
- Ozeki M, Heki K (2010) Ionospheric holes made by ballistic missiles from North Korea detected with a Japanese dense GPS array. *J Geophys Res* 115:A09314. <https://doi.org/10.1029/2010JA015531>
- Panda D, Senapati B, Tyagi B, Kundu B (2019) Effects of Rayleigh-Taylor instability and ionospheric plasma bubbles on the global navigation satellite system signal. *J Asian Earth Sci* 170:225–233. <https://doi.org/10.1016/j.jseae.2018.11.006>
- Perevalova NP, Sankov VA, Astafyeva EI, Zhupityaeva AS (2014) Threshold magnitude for ionospheric TEC response to earthquakes. *J Atmos Sol Terr Phys* 108:77–90
- Pulnits SA, Kotsarenko AN, Ciruolo L, Pulnits IA (2007) Special case of ionospheric day-to-day variability associated with earthquake preparation. *Adv Space Res* 39:970–977
- Reddy CD, Seemala GK (2015) Two-mode ionospheric response and Rayleigh wave group velocity distribution reckoned from GPS measurement following Mw 7.8 Nepal earthquake on 25 April 2015. *J Geophys Res* 120(8):7049–7059. <https://doi.org/10.1002/2015JA021502>
- Reddy CD, Shrivastava MN, González G, Baez JC (2016) Ionospheric plasma response to Mw 8.3 Chile Illapel earthquake on September 16 2015. *Pure Appl Geophys* 173:1451–1461. <https://doi.org/10.1007/s00024-016-1282-3>
- Rolland LM, Occhipinti G, Lognonné P, Loevenbruck A (2010) Ionospheric gravity waves detected off shore Hawaii after tsunamis. *Geophys Res Lett* 37:L17101. <https://doi.org/10.1029/2010GL044479>
- Rolland LM, Lognonné P, Munekane H (2011) Detection and modeling of Rayleigh wave induced patterns in the ionosphere. *J Geophys Res* 116:A05320. <https://doi.org/10.1029/2010JA016060>
- Rolland LM, Vergnolle M, Nocquet JM, Sladen A, Dessa JX, Tavakoli F, Nankali HR, Cappa F (2013) Discriminating the tectonic and non-tectonic contributions in the ionospheric signature of the 2011, Mw7.1, dip-slip Van earthquake, Eastern Turkey. *Geophys Res Lett* 40:2518–2522. <https://doi.org/10.1002/grl.50544>

- Row RV (1967) Acoustic-gravity waves in the upper atmosphere due to a nuclear detonation and an earthquake. *J Geophys Res* 72:1599–1610
- Ryu K, Parrot M, Kim SG, Jeong KS, Chae J, Pulinets S, Oyama KI (2014) Suspected seismo-ionospheric coupling observed by satellite measurements and GPS TEC related to the M7.9 Wenchuan earthquake of 12 May 2008. *J Geophys Res Space Phys* 119(10):305–310. <https://doi.org/10.1002/2014JA020613>
- Schaer S, Gurtner W, Feltens J (1998) IONEX: the ionosphere map EXchange format version 1. In: Proceedings of the IGS AC workshop, Darmstadt, Germany, February 9–11, 1998, pp 233–247
- Sekido M, Kondo T, Kawai E, Imae M (2003) Evaluation of GPS-based ionospheric TEC map by comparing with VLBI data. *Radio Sci* 38(4):8–1
- Shults K, Astafyeva E, Adourian S (2016) Ionospheric detection and localization of volcano eruptions on the example of the April 2015 Calbuco events. *J Geophys Res Space Phys* 121(10):10–303. <https://doi.org/10.1002/2016JA023382>
- Song Q, Ding F, Yu T, Wan W, Ning B, Liu L, Zhao B (2015) GPS detection of the coseismic ionospheric disturbances following the 12 May 2008 M7.9 Wenchuan earthquake in China. *Sci China Earth Sci* 58(1):151–158
- Su YC, Liu JY, Chen SP, Tsai HF, Chen MQ (2013) Temporal and spatial precursors in ionospheric total electron content of the 16 October 1999 Mw7.1 Hector Mine earthquake. *J Geophys Res Space Phys* 118:6511–6517. <https://doi.org/10.1002/jgra.50586>
- Tanaka T, Ichinose T, Okuzawa T, Shibata T, Sato Y, Nagasawa C, Ogawa T (1984) HF-Doppler observations of acoustic waves excited by the Urakawa-Oki earthquake on 21 March 1982. *J Atmos Terr Phys* 46:233–245
- Tsugawa T, Saito A, Otsuka Y, Nishioka M, Maruyama T, Kato H, Nagatsuma T, Murata KT (2011) Ionospheric disturbances detected by GPS total electron content observation after the 2011 off the Pacific coast of Tohoku earthquake. *Earth Planets Space* 63:875–879. <https://doi.org/10.5047/eps.2011.06.035>
- Vijayan MS, Varghese MG, Shimna K (2013) IONODETECT: a software to compute Ionospheric TEC perturbations using dual frequency geodetic GPS data, annual report 2012–2013, CSIR 4PI77–79
- Walker GO, Ma JHK, Golton E (1994) The equatorial ionospheric anomaly in electron content from solar minimum to solar maximum for South East Asia. *Ann Geophys* 12:195–209
- Watada S, Kanamori H (1994) Near-source acoustic coupling between atmosphere and solid earth during volcanic eruptions. *EOS Trans* 75:446
- Wyss M (ed) (1991) Evaluation of proposed earthquake precursors. American Geophysical Union, Washington, DC, p 94. <https://doi.org/10.1029/SP032>
- Yagi Y (2004) Source rupture process of the 2003 Tokachi-oki earthquake determined by joint inversion of teleseismic body wave and strong ground motion data. *Earth Planets Space* 56:311–316. <https://doi.org/10.1186/BF03353057>
- Yuen PC, Weaver PF, Suzuki RK, Furumoto AS (1969) Continuous traveling coupling between seismic waves and the ionosphere evident in May 1968 Japan earthquake data. *J Geophys Res* 74:2256–2264
- Zel'Dovich YB, Raizer YP (1966) Physics of shock waves and high-temperature hydrodynamic phenomena, vol 1. Academic, New York/London, 464pp
- Zhang X, Tang L (2015) Detection of ionospheric disturbances driven by the 2014 Chile tsunami using GPS total electron content in New Zealand. *J Geophys Res Space Phys* 120(9):7918–7925. <https://doi.org/10.1002/2014JA020879>
- Zhao B, Wang M, Yu T, Wan W, Lei J, Liu L, Ning B (2008) Is an unusual large enhancement of ionospheric electron density linked with the 2008 great Wenchuan earthquake? *J Geophys Res* 113:1–6

Tracking Earth's Water in Motion from Satellite Gravity Observations

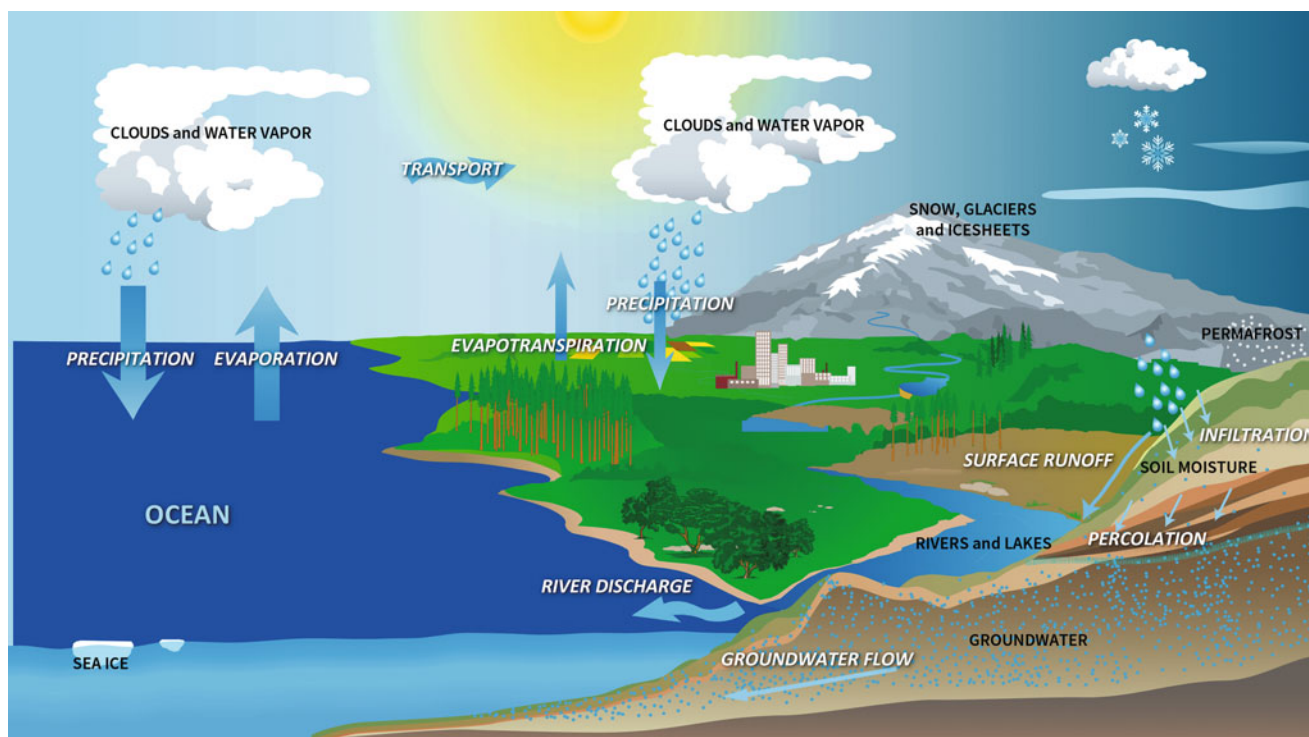
J. L. Chen

Center for Space Research, University of Texas, Austin, TX, USA

Introduction

The global water cycle (see Fig. 1), also known as the hydrological cycle, describes the continuous movement and redistribution of water within one or among different components of the climate system through the physical processes of evaporation, precipitation, surface runoff, subsurface flow, condensation, infiltration, etc., and variations of water storage in different reservoirs, such as soil moisture, groundwater, snow and ice, atmospheric water vapor, and water in the ocean (Dingman 1994). Monitoring and measuring these hydrological variables are critical for understanding the global water cycle and Earth climate and ecosystem. Accurate quantification of water mass variations associated with these hydrological processes at regional to global scales is difficult, due to the lack of adequate in situ observations of related hydrological variables. Predictions from hydrological models, also called land surface models (LSMs) have been a useful data resource for studying the global water cycle. However, large uncertainty is expected for model predicted hydrological variations (also due to the lack of adequate in situ observations to constrain the model predictions).

Since launched in March 2002, the Gravity Recovery and Climate Experiment (GRACE) satellite gravity mission has brought a new era to the studies of large-scale mass transport and redistribution within different components of the Earth system, including the atmosphere, hydrosphere, ocean, cryosphere, and solid Earth. The GRACE (decommissioned in November 2017) has accumulated over 15 years of time-variable satellite gravity measurements with unprecedented accuracy (Tapley et al. 2019). The GRACE Follow-On (GFO) mission was launched in 2018 to continue the endeavor of the GRACE mission. Time-variable gravity measurements from the GRACE and GFO (hereafter noted as GRACE) missions have been used in a wide range of investigations to quantify large-scale mass variations on the surface of the Earth, including ice mass change of polar ice sheets and mountain glaciers, terrestrial water storage change in river basins, regional groundwater depletion, global ocean mass change, and deformations of the solid Earth due to the Glacial Isostatic Adjustment (GIA) and large earthquakes. These studies have significantly improved our understanding of the Earth climate system and the global water cycle (e.g., Cazenave and Chen 2010; Rodell et al. 2018; Tapley et al. 2019).



Tracking Earth's Water in Motion from Satellite Gravity Observations, Fig. 1 Illustration of the global water cycle. (Modified from an original map from NASA)

GRACE Satellite Gravimetry

GRACE is a twin satellites gravity mission, jointly sponsored by US NASA and German Aerospace Center (DLR) (Tapley et al. 2004). The GRACE satellites were launched on 17 March 2002, with an initial altitude of approximately 500 km and a near-polar inclination of 89° . The satellites orbit one behind the other in the same orbit plane and are separated by approximately 220 km along their orbit track. GRACE utilizes a state-of-the-art technique to observe variations of Earth's gravity by tracking the inter-satellite range and range-rate (see Fig. 2) via a K-band ranging (KBR) system (Tapley et al. 2004). In addition, each satellite is equipped with a SuperSTAR Accelerometer, GPS receiver, Star Cameras, and Laser Retro Reflector to complete the complement of science instruments. The GRACE Science Data System uses the KBR range and range rate data, along with ancillary data, to estimate a new gravity field every month in the form of corrections to an a priori background gravity model used in the data processing procedure (Bettadpur 2012). The GFO mission is basically a duplicate of GRACE, with two satellites flying in the same polar orbit (as the GRACE satellites) and carrying the same science instruments with upgraded hardware. GFO also carries an experimental laser ranging interferometer (LRI) instrument that is assessed as a possible improvement for future GRACE-like missions.



Tracking Earth's Water in Motion from Satellite Gravity Observations, Fig. 2 GRACE includes twin satellites orbiting the Earth at an initial altitude of ~ 500 km in a near-polar orbit (89° inclination). GRACE was launched on March 17, 2002, and was decommissioned in November 2017. (Source: US NASA/UT CSR)

Monthly GRACE global gravity solutions are provided by three GRACE data processing centers forming the Science Data System, including the Center for Space Research (CSR) at the University of Texas at Austin, Geoforschungszentrum (GFZ) in Potsdam, and the NASA Jet Propulsion Laboratory (JPL). GRACE solutions are distributed by the NASA Physical Oceanography Distributed Active Archive Center (<https://podaac.jpl.nasa.gov/grace>) and GFZ Information System and Data Center (<http://isdc.gfz-potsdam.de/grace-isdc/>). GRACE monthly solutions are generally expressed in the form of fully normalized spherical harmonic (SH) coefficients, up to certain maximum degree and order (typically up to degree 60 or 96). Atmospheric and high frequency oceanic mass variations, and tidal effects are removed during GRACE de-aliasing process using atmospheric and oceanic model predictions (Bettadpur 2012). GRACE monthly mass concentration (mascon) solutions, as alternative GRACE mass change solutions with less noise and better spatial resolution are also generated by several data processing centers, including NASA JPL, UT CSR, and NASA Goddard Space Flight Center (GSFC) (Watkins et al. 2015; Save et al. 2016; Luthcke et al. 2013).

Quantifying Mass Change Using GRACE Gravity Solutions

When assuming that, at decadal or shorter time scales, GRACE observed time-variable gravity change represents contributions mainly from mass variations on the Earth surface in the Earth geophysical fluids envelope (atmosphere, ocean, hydrosphere, and cryosphere), GRACE monthly gravity SH coefficients C_{lm} and S_{lm} can be readily converted into surface mass load change as (Wahr et al. 1998),

$$\Delta\sigma(\theta, \lambda) = \frac{M_E}{4\pi a^2} \sum_{l=0}^N \sum_{m=0}^l \frac{2l+1}{1+k_l} W_l \tilde{P}_{lm}(\cos \theta) \times [\Delta C_{lm} \cos(m\lambda) + \Delta S_{lm} \sin(m\lambda)] \quad (1)$$

where θ and λ are the colatitude and east longitude, respectively; a is Earth mean radius (6.371×10^6 m); $\tilde{P}_{lm}$ are the fully normalized associated Legendre polynomials of degree l and order m ; M_E is the mass of the Earth (5.97219×10^{24} kg), $\Delta\sigma(\theta, \lambda)$ is mass load changes (in kg/m^2 or equivalent water height in mm) at a given point on the Earth surface, and k_l are load Love numbers of the solid Earth which account for mass redistribution of the Earth due to changing surface loads (see Table 1 in Wahr et al. 1998). W_l is a factor representing normalized Gaussian weighting (with maximum weight of 1), and reduces contributions from high degree and order SH coefficients. The computation of Gaussian weights as

functions of degree l and averaging radius (r) can be found in Wahr et al. (1998). GRACE high degree and order SH coefficients are affected by noise, especially the longitudinal stripes. The longitudinal stripes originate from the correlation between the even and odd degree pairs of GRACE SH coefficients of a given order, and can be removed through the proposed decorrelation filtering (Swenson and Wahr 2006). After decorrelation filtering, a degree of Gaussian smoothing is still needed to further suppress the residual noise, by applying the Gaussian weight W_l to SH coefficients of corresponding degree (l) in Eq. 1.

Monitoring Land Water Cycle Using GRACE

There are two major limitations of the GRACE gravity measurements. One is the coarse spatial resolution (of about a few to several hundred kilometers), which is influenced by several factors, including the truncation of GRACE SH solutions (e.g., at degree 60), and the need of spatial filtering due to the dominant spatial noise at high degrees and orders. Therefore, GRACE gravity measurements are mostly applied to study large regional or basin scale water mass changes. The other is the vertical separation. Over land region, when atmospheric effect is removed, GRACE observed time-variable gravity change represents integrated contribution from mass variations in different reservoirs of the water body, including soil moisture, snow, surface water (in lakes and rivers), and groundwater. GRACE itself cannot separate the different contributions. Independent estimates from other sources are needed when the focus is on a particular hydrological variable.

In any given region (or basin), the hydrological conservation equation can be expressed in terms of water fluxes, e.g., terrestrial water storage (TWS) change rate ($\frac{\Delta TWS}{\Delta t}$), precipitation (P), evapotranspiration (E), and surface and subsurface runoff (R) as,

$$\frac{\Delta TWS}{\Delta t} = P - E - R \quad (2)$$

The hydrological conservation equation can also be expressed in terms of water storage changes in different reservoirs of the water body, including soil moisture (ΔSM), snow water (ΔSN), surface water (ΔSW), and groundwater (ΔGW) as,

$$\Delta TWS = \Delta SM + \Delta SN + \Delta SW + \Delta GW \quad (3)$$

Based on the two hydrological conservation equations (Eqs. 2 and 3), it is theoretically possible to study any related components in the equations, as long as other components can be determined independently. For example, R can be

estimated from GRACE determined $\frac{\Delta TWS}{\Delta t}$ and climate model predicted P and E as,

$$R = \frac{\Delta TWS}{\Delta t} - (P - E) \quad (4)$$

And, groundwater storage change (ΔGW) can be estimated from GRACE determined ΔTWS and land surface model (LSM) predicted soil moisture, snow water, and surface water as

$$\Delta GW = \Delta TWS - (\Delta SM + \Delta SN + \Delta SW) \quad (5)$$

Some previous studies (e.g., Syed et al. 2005, 2009; Seo et al. 2009) combined GRACE observed ΔTWS with model predicted precipitation and evapotranspiration flux to quantify basin-scale river discharge using Eq. (2). Similarly, several other studies (Rodell et al. 2004, 2011; Ramillien et al. 2006; Cesanelli and Guarracino 2011) quantified basin-scale evapotranspiration flux using GRACE TWS change estimates with precipitation and river discharge data from model predictions and in situ observations. GRACE satellite gravity measurements have also been used to study large-scale TWS changes, especially those related to extreme climate events, such as severe floods and droughts (e.g., Papa et al. 2008; Chen et al. 2009, 2010), and groundwater depletions (e.g., Rodell et al. 2009; Famiglietti et al. 2011; Scanlon et al. 2012).

Groundwater plays a key role in connections between the global hydrological cycle and climate change, and is a vital resource in many parts of the world. Accurate quantification of groundwater storage change is difficult, and limited by in situ groundwater observations (Famiglietti et al. 2011). In developed countries (such as the U.S. and Australia) well networks may be relatively dense, but stacked aquifers can result in order of magnitude variations in storage coefficients, increasing uncertainty in converting groundwater level changes to groundwater storage (Rodell et al. 2007). Over the past decade, GRACE satellite gravity measurements have been successfully used to monitor and quantify major groundwater depletions in various regions over the world, including northwest and northern India (e.g., Rodell et al. 2009; Tiwari et al. 2009; Chen et al. 2014), the Middle East (Voss et al. 2013; Joodaki et al. 2014), California's Central Valley (Famiglietti et al. 2011; Scanlon et al. 2012), High Plains Aquifer in the USA (Strassberg et al. 2009), southern Murray Darling Basin in Australia (Leblanc et al. 2009; Chen et al. 2016), and North China Plain (Feng et al. 2013).

Monitoring Ice and Ocean Mass Change Using GRACE

Accurate quantification of polar ice sheet mass balance and its contribution to global sea level change has long been a

challenge, due to the lack of either in situ data or remote sensing measurements. Before the GRACE mission, for a long-time people were not certain whether the entire Antarctic ice sheet was losing or gaining ice. GRACE satellite gravimetry provides a unique and successful means for quantifying ice mass balance of polar ice sheets and mountain glaciers. Compared to other remote sensing techniques and in situ data, GRACE provides the first "direct" measurements of ice mass change on continental and large regional scales, although some uncertainty is expected in GRACE estimates due to the coarse spatial resolution of GRACE solutions, and errors in data calibration and other geophysical signals, e.g., the Global Isostatic Adjustment (GIA) effect, that need to be quantified and removed from GRACE. Over the past 15 years, GRACE satellite gravity measurements have been widely used to study ice mass change of the Antarctic and Greenland ice sheets (e.g., Velicogna and Wahr 2006; Chen et al. 2006), and mountain glaciers (e.g., Luthcke et al. 2008; Lenaerts et al. 2013). GRACE data shows that both Greenland and Antarctica are losing a significant amount of ice over the GRACE time span (April 2002–June 2017), with estimated rates of $\sim 258 \pm 26$ Gt/year and 137 ± 41 Gt/year, respectively (Tapley et al. 2019). The quantification of ice mass change of mountain glaciers is more challenging, limited by GRACE resolution. Optimal data processing methods, such as regional mascons or forward modeling, are needed to appropriately address the leakage issue when using GRACE SH solutions (e.g., Luthcke et al. 2013; Chen et al. 2015). Additional challenge comes from the needed separation of mountain glacier mass signal from surrounding TWS change, which mainly relies on model predicted TWS changes. Large uncertainty of model TWS estimates are expected, especially at interannual and long-term time scales.

Understanding the global sea level change and its driving forces is an important topic in climate change studies, and critical in interpreting and predicting the climate system. Global sea level change is driven by two major forces, water mass exchange between the ocean and land (including ice sheets) and sea water density change due to temperature and salinity change (the so-called steric sea level change). GRACE satellite gravimetry allows, for the first time, direct measure of oceanic mass contribution to global sea level change (e.g., Chambers et al. 2004; Chen et al. 2005). Since around 2005, the study of global sea level change has entered a completely new era, with the integration of observations from three independent modern techniques monitoring the global ocean from three different perspectives. Satellite radar altimeter has been measuring the global sea surface height (SSH) anomalies with unprecedented accuracy since 1992 when the TOPEX/Poseidon, the first of a series of the new generation of satellite altimeter missions, was launched (Fu and Davidson 1995). The ARGO float network is an international joint project designed to provide continuous

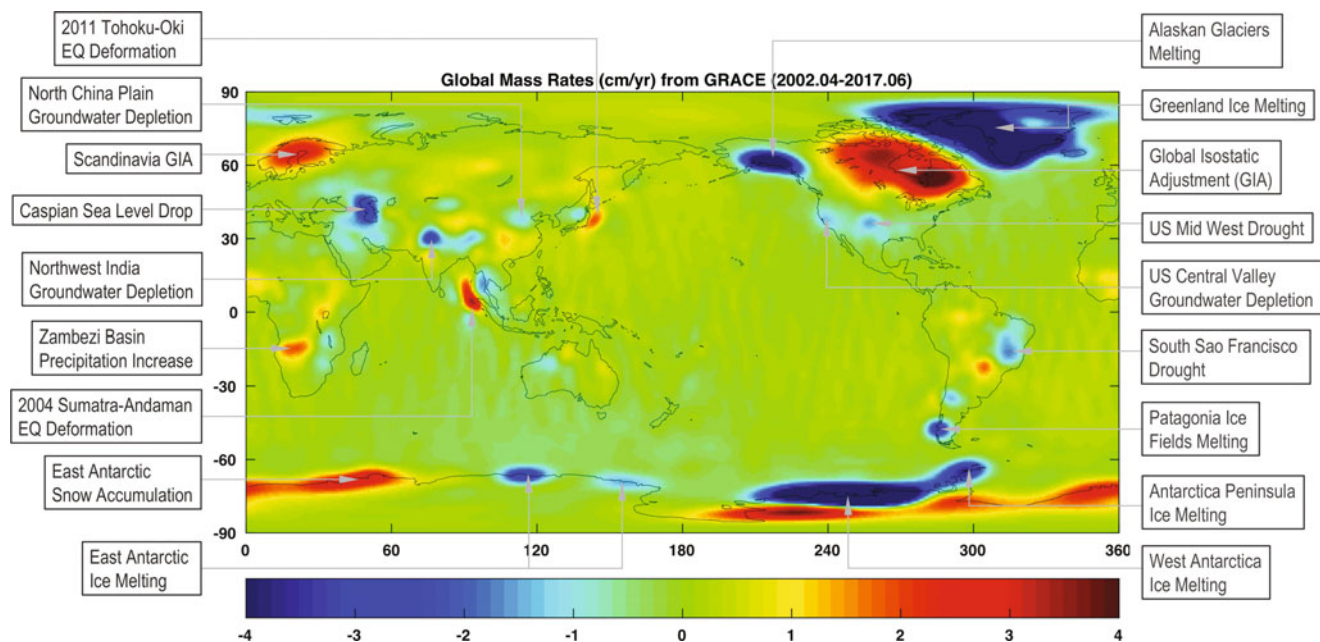
3-dimensional measurements of the temperature, salinity and other geophysical and chemical properties of the global ocean. The deployment of the Argo global floats array started in 2000 (Roemmich and Owens 2000; Argo 2000), and its global coverage has become relatively complete and homogeneous since around 2005. Combining satellite altimeter SSH, Argo ocean temperature and salinity, and GRACE gravity measurements offers the first opportunity to study the global sea level change and its different contributions using global-scale observations (e.g., Cazenave et al. 2009; Leuliette and Willis 2011; Chen et al. 2013; Chambers et al. 2017).

Summary

Over the past 15 plus years, GRACE time-variable gravity measurements have brought a new era to the study of large-scale water mass transport and redistribution within the Earth geophysical fluids system, including the atmosphere, hydrosphere, ocean, and cryosphere, and greatly enhanced our understanding and quantification of the global water cycle. GRACE is inarguably regarded as one of the most successful Earth science missions in the past several decades. GRACE offers the first opportunities for accurately quantifying large-scale TWS changes associated with the global water cycle, and monitor major floods and droughts on global basis. When combined with other independent data (e.g., measurements or model predictions), GRACE

can be used to study different components of the terrestrial water cycle, such as evapotranspiration, river discharge, groundwater, soil moisture, snow water, etc. on regional, basin, continental, and global scales. GRACE satellite gravimetry also provides a new means for quantifying ice mass changes of polar ice sheets and mountain glaciers with unprecedented accuracy. The integration of GRACE gravimetry, satellite altimeter, and Argo in situ ocean temperature and salinity measurements has offered the first opportunities to accurately quantify different driving forces of the global sea level change, which is critical for understanding and predicting the climate system.

GRACE time-variable gravity solutions have helped paint a much clearer picture of the global water cycle, in particular long-term water and ice mass changes in different components of the climate system and different parts of the world (see Fig. 3). After an extended period of data calibration and validation (Cal/Val), the GFO project has recently released its first year of gravity solutions in June 2019, with similar accuracy as the GRACE solutions. The GRACE and GFO missions are expected to provide more than two decades of time-variable gravity measurements, which will be a valuable resource for studying water mass transport in various components of the Earth climate system at broad time scales. Continuous improvements of GRACE data processing methods and background geophysical models are also expected, which will help improve the accuracy and spatial resolution of GRACE measurements and enable wider range of applications of satellite gravity observations.



Tracking Earth's Water in Motion from Satellite Gravity Observations, Fig. 3 Global mass change rates (in cm/year of equivalent water height) observed by GRACE during the period April 2002 and June

2017, based on CSR RL06 gravity solutions with 300 km Gaussian smoothing applied. Some major long-term mass signals from climate change or solid geophysical sources are labeled on the sides

Bibliography

- Argo (2000) Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANO. <https://doi.org/10.17882/42182>
- Bettadpur S (2012) CSR Level-2 processing standards document for product release 05 GRACE 327–742, The GRACE Project, Center for Space Research, University of Texas at Austin
- Cazenave A, Chen J (2010) Time-variable gravity from space and present-day mass redistribution in the Earth system. *Earth Planet Sci Lett* 298:263–274. <https://doi.org/10.1016/j.epsl.2010.07.035>
- Cazenave A, Dominh K, Guinehut S, Berthier E, Llovel W, Ramillien G, Ablain M, Larnicol G (2009) Sea level budget over 2003–2008: a reevaluation from GRACE space gravimetry, satellite altimetry and Argo. *Glob Planet Change* 65:83–88. <https://doi.org/10.1016/j.gloplacha.2008.10.004>
- Cesanelli A, Guarracino L (2011) Estimation of regional evapotranspiration in the extended Salado Basin (Argentina) from satellite gravity measurements. *Hydrogeol J* 19:629–639. <https://doi.org/10.1007/s10040-011-0708-3>
- Chambers DP, Wahr J, Nerem RS (2004) Preliminary observations of global ocean mass variations with GRACE. *Geophys Res Lett* 31: L13310. <https://doi.org/10.1029/2004GL020461>
- Chambers DP, Cazenave A, Champollion N, Dieng H, Llovel W, Forsberg R, von Schuckmann K, Wada Y (2017) Evaluation of the global mean sea level budget between 1993 and 2014. *Surv Geophys* 38:309–327. <https://doi.org/10.1007/s10712-016-9381-3>
- Chen JL, Wilson CR, Tapley BD, Famiglietti JS, Rodell M (2005) Seasonal global mean sea level change from satellite altimeter, GRACE, and geophysical models. *J Geod* 79:532–539. <https://doi.org/10.1007/s00190-005-0005-9>
- Chen JL, Wilson CR, Tapley BD (2006) Satellite gravity measurements confirm accelerated melting of Greenland ice sheet. *Science* 313:1958–1960. <https://doi.org/10.1126/science.1129007>
- Chen JL, Wilson CR, Tapley BD, Yang ZL, Niu GY (2009) 2005 drought event in the Amazon River basin as measured by GRACE and estimated by climate models. *J Geophys Res Solid Earth* 114(B13): B5404. <https://doi.org/10.1029/2008JB006056>
- Chen JL, Wilson CR, Tapley BD (2010) The 2009 exceptional Amazon flood and interannual terrestrial water storage change observed by GRACE. *Water Resour Res* 46:W12526. <https://doi.org/10.1029/2010WR009383>
- Chen JL, Wilson CR, Tapley BD (2013) Contribution of ice sheet and mountain glacier melt to recent sea level rise. *Nat Geosci* 6:549–552. <https://doi.org/10.1038/ngeo1829>
- Chen JL, Li J, Zhang Z, Ni S (2014) Long-term groundwater variations in Northwest India from satellite gravity measurements. *Glob Planet Chang* 116:130–138. <https://doi.org/10.1016/j.gloplacha.2014.02.007>
- Chen JL, Wilson CR, Li J, Zhang Z (2015) Reducing leakage error in GRACE-observed long-term ice mass change: a case study in West Antarctica. *J Geod* 89:925–940. <https://doi.org/10.1007/s00190-015-0824-2>
- Chen JL, Wilson CR, Tapley BD, Scanlon B, Güntner A (2016) Long-term groundwater storage change in Victoria, Australia from satellite gravity and in situ observations. *Glob Planet Chang* 139:56–65. <https://doi.org/10.1016/j.gloplacha.2016.01.002>
- Dingman SL (1994) Physical hydrology. Prentice Hall, Upper Saddle River. ISBN 0-02-329745-X
- Famiglietti JS, Lo M, Ho SL, Bethune J, Anderson KJ, Syed TH, Swenson SC, de Linage CR, Rodell M (2011) Satellites measure recent rates of groundwater depletion in California's Central Valley. *Geophys Res Lett* 38:L3403. <https://doi.org/10.1029/2010GL046442>
- Feng W, Zhong M, Lemoine J-M, Biancale R, Hsu H-T, Xia J (2013) Evaluation of groundwater depletion in North China using the Gravity Recovery and Climate Experiment (GRACE) data and ground-based measurements. *Water Resour Res* 49:2110–2118. <https://doi.org/10.1002/wrcr.20192>
- Fu L-L, Davidson RA (1995) A note on the barotropic response of sea level to time-dependent wind forcing. *J Geophys Res* 100(C12): 24955–24963
- Joodaki G, Wahr J, Swenson S (2014) Estimating the human contribution to groundwater depletion in the Middle East, from GRACE data, land surface models, and well observations. *Water Resour Res* 50:2679–2692. <https://doi.org/10.1002/2013WR014633>
- Leblanc MJ, Tregoning P, Ramillien G, Tweed SO, Fakes A (2009) Basin-scale, integrated observations of the early 21st century multi-year drought in southeast Australia. *Water Resour Res* 45:W4408. <https://doi.org/10.1029/2008WR007333>
- Lenaerts JTM, Angelen JH, Broeke MR, Gardner AS, Wouters B, Meijgaard E (2013) Irreversible mass loss of Canadian Arctic Archipelago glaciers. *Geophys Res Lett* 40:870–874. <https://doi.org/10.1002/grl.50214>
- Leuliette EW, Willis JK (2011) Balancing the sea level budget. *Oceanography* 24(2):122–129. <https://doi.org/10.5670/oceanog.2011.32>
- Luthcke SB, Arendt AA, Rowlands DD, McCarthy JJ, Larsen CF (2008) Recent glacier mass changes in the Gulf of Alaska region from GRACE mascon solutions. *J Glaciol* 54:767–777. <https://doi.org/10.3189/002214308787779933>
- Luthcke SB, Sabaka TJ, Loomis BD, Arendt AA, McCarthy JJ, Camp J (2013) Antarctica, Greenland and Gulf of Alaska land-ice evolution from an iterated GRACE global mascon solution. *J Glaciol* 59:613–631. <https://doi.org/10.3189/2013JoG12J147>
- Papa F, Prigent C, Rossow WB (2008) Monitoring flood and discharge variations in the large Siberian Rivers from a multi-satellite technique. *Surv Geophys* 29:297–317. <https://doi.org/10.1007/s10712-008-9036-0>
- Ramillien G, Frappart F, Güntner A, Ngo-Duc T, Cazenave A, Laval K (2006) Time variations of the regional evapotranspiration rate from Gravity Recovery and Climate Experiment (GRACE) satellite gravimetry. *Water Resour Res* 42:W10403. <https://doi.org/10.1029/2005WR004331>
- Rodell M, Famiglietti JS, Chen JL, Seneviratne SI, Viterbo P, Holl S, Wilson CR (2004) Basin scale estimates of evapotranspiration using GRACE and other observations. *Geophys Res Lett* 31:L20504. <https://doi.org/10.1029/2004GL020873>
- Rodell M, Chen JL, Kato H, Famiglietti JS, Nigro J, Wilson CR (2007) Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE. *Hydrogeol J* 15:159–166. <https://doi.org/10.1007/s10040-006-0103-7>
- Rodell M, Velicogna I, Famiglietti JS (2009) Satellite-based estimates of groundwater depletion in India. *Nature* 460:999–1002. <https://doi.org/10.1038/nature08238>
- Rodell M, McWilliams EB, Famiglietti JS, Beaudoin HK, Nigro J (2011) Estimating evapotranspiration using an observation based terrestrial water budget. *Hydrol Process* 25:4082–4092. <https://doi.org/10.1002/hyp.8369>
- Rodell M, Famiglietti JS, Wiese DN, Reager JT, Beaudoin HK, Landerer FW, Lo M-H (2018) Emerging trends in global freshwater availability. *Nature* 557:651–659. <https://doi.org/10.1038/s41586-018-0123-1>
- Roemmich D, Owens WB (2000) The ARGO project: global ocean observations for understanding for understanding and prediction of climate variability. *Oceanography* 13(2):45–50
- Save H, Bettadpur S, Tapley BD (2016) High-resolution CSR GRACE RL05 mascons. *J Geophys Res Solid Earth* 121:7547–7569. <https://doi.org/10.1002/2016JB013007>
- Scanlon BR, Longuevergne L, Long D (2012) Ground referencing GRACE satellite estimates of groundwater storage changes in the California Central Valley, USA. *Water Resour Res* 48:4520. <https://doi.org/10.1029/2011WR011312>
- Seo K-W, Waliser DE, Tian B, Famiglietti JS, Syed TH (2009) Evaluation of global land-to-ocean fresh water discharge and evapotranspiration using space-based observations. *J Hydrol* 373:508–515
- Strassberg G, Scanlon BR, Chambers D (2009) Evaluation of groundwater storage monitoring with the GRACE satellite: case study of the

- High Plains aquifer, central United States. *Water Resour Res* 45: W5410. <https://doi.org/10.1029/2008WR006892>
- Swenson S, Wahr J (2006) Post-processing removal of correlated errors in GRACE data. *Geophys Res Lett* 33:L8402. <https://doi.org/10.1029/2005GL025285>
- Syed TH, Famiglietti JS, Chen JL, Rodell M, Seneviratne SI, Viterbo P, Wilson CR (2005) Total basin discharge for the Amazon and Mississippi River basins from GRACE and a land-atmosphere water balance. *Geophys Res Lett* 32:L24404. <https://doi.org/10.1029/2005GL024851>
- Syed TH, Famiglietti JS, Chambers DP (2009) GRACE-based estimates of terrestrial freshwater discharge from basin to continental scales. *J Hydrometeorol* 10:22–40. <https://doi.org/10.1175/2008JHM993.1>
- Tapley BD, Bettadpur S, Ries JC, Thompson PF, Watkins MM (2004) GRACE measurements of mass variability in the earth system. *Science* 305:503–506. <https://doi.org/10.1126/science.1099192>
- Tapley BD, Watkins MM, Flechtner F, Reigber C, Bettadpur S et al (2019) Contributions of GRACE to understanding climate change. *Nat Clim Chang* 9:358–369. <https://doi.org/10.1038/s41558-019-0456-2>
- Tiwari VM, Wahr J, Swenson S (2009) Dwindling groundwater resources in northern India, from satellite gravity observations. *Geophys Res Lett* 36:L18401. <https://doi.org/10.1029/2009GL039401>
- Velicogna I, Wahr J (2006) Measurements of time-variable gravity show mass loss in Antarctica. *Science* 311:1754–1756. <https://doi.org/10.1126/science.1123785>
- Voss KA, Famiglietti JS, Lo M, Linage C, Rodell M, Swenson SC (2013) Groundwater depletion in the Middle East from GRACE with implications for transboundary water management in the Tigris-Euphrates-Western Iran region. *Water Resour Res* 49:904–914. <https://doi.org/10.1002/wrcr.20078>
- Wahr J, Molenaar M, Bryan F (1998) Time variability of the Earth's gravity field: hydrological and oceanic effects and their possible detection using GRACE. *J Geophys Res* 103:30205–30230. <https://doi.org/10.1029/98JB02844>
- Watkins MM, Wiese DN, Yuan D-N, Boening C, Landerer FW (2015) Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *J Geophys Res Solid Earth* 120:2648–2671. <https://doi.org/10.1002/2014JB011547>

Trans-European Suture Zone

H. Thybo^{1,2} and I. M. Artemieva^{3,4}

¹Eurasia Institute of Earth Sciences, Istanbul Technical University, Istanbul, Turkey

²Centre for Earth Evolution and Dynamics (CEED), University of Oslo, Oslo, Norway

³Marine Geodynamics, GEOMAR Helmholtz Center for Ocean Research, Kiel, Germany

⁴Department of Geophysics, Stanford University, Palo Alto, CA, USA

Synonyms

Teisseyre-Tornquist Zone; Sorgenfrei-Tornquist Zone; Trans-European Suture Zone; Tornquist Fan

Definition

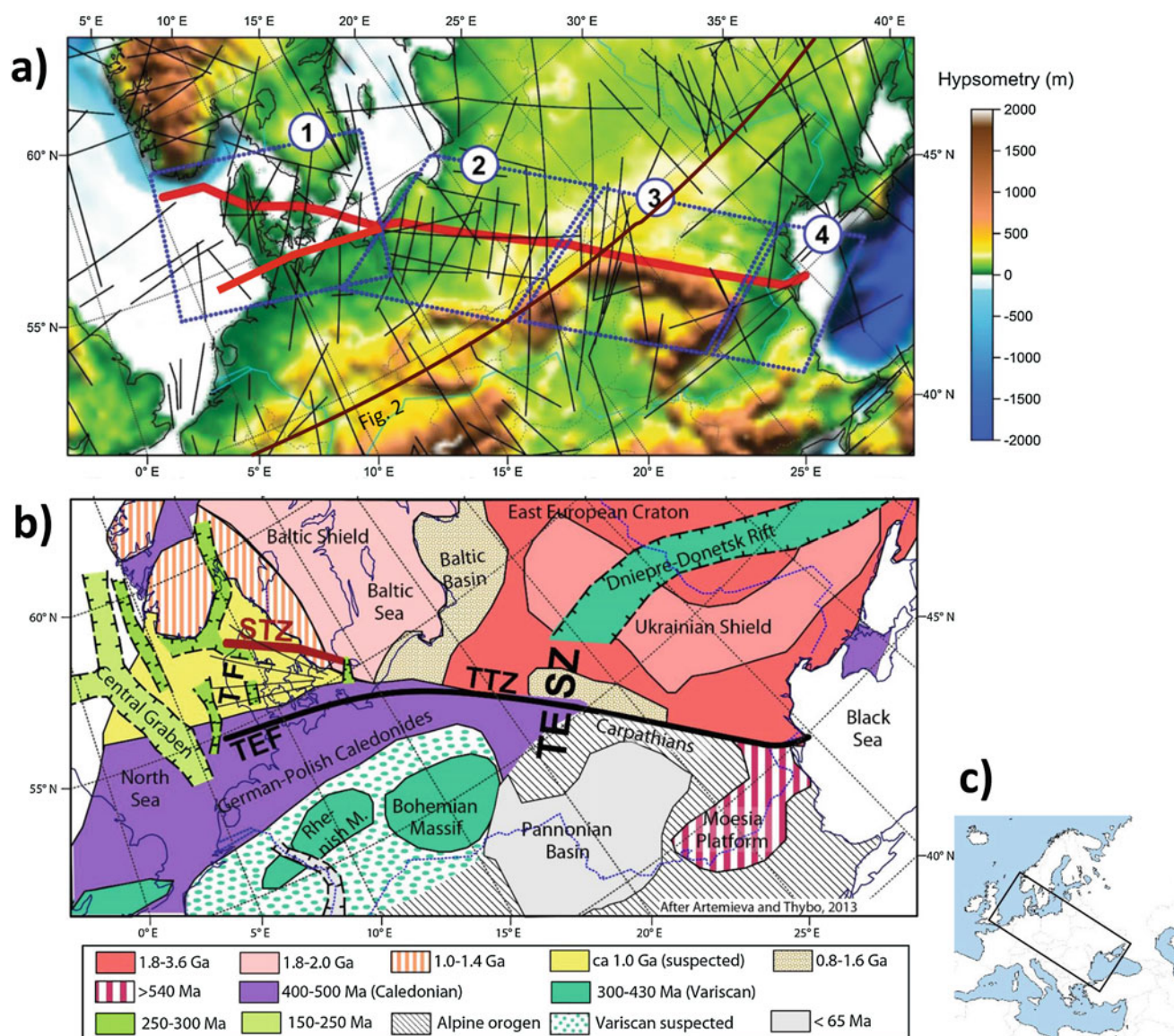
The Trans-European Suture Zone (TESZ) is the transition zone from the Precambrian East European Craton in the north and east to the younger Phanerozoic mobile belts to the south and west. It is the most prominent lithospheric tectonic feature of Europe.

The term Trans-European Suture Zone was only adapted around year 2000 during the Pan-European EUROPROBE program of the European Science Foundation. Until then, parts of the zone were termed Teisseyre-Tornquist Zone, Sorgenfrei-Tornquist Zone, Trans-European Fault, and Tornquist Fan.

Introduction

The Trans-European Suture Zone, TESZ, is associated with major geological differences, abrupt changes in crustal and lithospheric thicknesses, and major anomalies in all geophysical parameters (Fig. 1). Historically, it was discovered as a geological-geophysical lineament around 1900 from geological observations by W. Teisseyre despite the presence of a thick sedimentary cover, and observation by A.J.H. Tornquist of a pronounced, linear magnetic anomaly extending from the Black Sea to the Baltic Sea. The craton to mobile belt transition was, therefore, originally named the Teisseyre-Tornquist Zone (TTZ) in central Europe. The term TESZ was introduced in the late 1990s to emphasize that the transition is represented by a suture zone, rather than a lineament, between Precambrian cratons in the north and east and younger Europe in the south and west, which is composed of a series of terranes that were amalgamated to proto-Europe in the Palaeozoic (Winchester and PACE Network Team 2002).

Another geological lineament across southern Sweden and Denmark into the North Sea, the Sorgenfrei-Tornquist Zone (STZ, Fig. 1b), was for long understood as the NW continuation of the TTZ from a dog-leg offset in the Baltic Sea. It was believed to represent the SW transition between the original Baltica Plate and younger terranes. However, radiometric dating and interpretation of deep crustal seismic data has shown that the STZ is not the edge of the craton but instead a tectonic feature inside the original Baltica Plate in its tectonically reworked SW corner (EUGENO-S Working Group 1988). Interpretation of a large amount of seismic reflection profiles from industry has shown that STZ is a pronounced tectonic inversion zone in the sedimentary cover (Thybo 1997). The TTZ is co-located with a similar inversion zone in Poland, and in this regard, the two zones are similar. Deep seismic profiling has demonstrated that the inversion affected basement and even created the basement horst at the island Bornholm in the transition between TTZ and STZ, where the inversion likely was accompanied by crustal thickening in a



Trans-European Suture Zone, Fig. 1 Location of the Trans-European Suture Zone (TESZ). (a) Hypsometric map around TESZ indicating four characteristic areas. The exact limits of TESZ are undefined, and the zone largely covers the four rhombuses. For orientation the Teisseyre-Tornquist Zone (TTZ) and Sorgenfrei-Tornquist Zone (STZ) are shown by a thick red line, and by a thick black line in all other maps. The map includes the location of seismic profiles in the region (thin

straight lines) and the location of the generalized seismic profile in Fig. 2 (black curved line). (b) Tectonic map of the region around TESZ. Abbreviations: TTZ – Teisseyre-Tornquist Zone, STZ – Sorgenfrei-Tornquist Zone, TF – Tornquist Fan, TEF – Trans-European Fault. (c) Map of Europe showing the TESZ region. (Figure after Artemieva and Thybo (2013))

complex “crocodile tectonic structure” or by “subversion” (BABEL Working Group 1993).

The STZ occurs within the Tornquist Fan, which is defined by a series of transtensional fault zones that were active in the late Palaeozoic, accompanied by intensive volcanic and magmatic activity throughout the area (Thybo 1997). The transtensional displacement was introduced by dextral strike-slip on the TTZ in the triangular area of the TF area of the Baltic Shield. The northeastern termination of the Tornquist Fan largely coincides with the inversion zone STZ and the

southwestern termination has been named the Trans-European Fault in northernmost Germany close to Denmark (Berthelsen 1992).

To the south of Poland, the structures of the TTZ have partly been overprinted by the structures related to the Carpathian Mountains, and its termination in the Black Sea has, so far, not been imaged.

Much new knowledge on TESZ was obtained during the EGT (European GeoTraverse) and EUROPROBE pan-European large-scale geoscientific programs in the 1980s

and 1990s. These programs included projects for integrated geophysical-tectonic-geological studies of the TESZ, usually centered around major acquisition of new deep seismic, wide-angle reflection and refraction, as well as normal-incidence reflection seismic data in the northwestern part of the TESZ, such as in the EUGENO-S, BABEL, and POLONAISE projects, as well as the TOR broadband seismological project. Other projects applied similar approach, including the major CELEBRATION2000 refraction (region 2 and 3 in Fig. 1a, Guterch et al. 2003) and MONA LISA high resolution reflection/refraction (region 1 in Fig. 1a) seismic projects, as well as the PANCAKE and VRANCEA seismic projects in region 3 and the DOBRE 4 and 5 projects in region 4.

Geologic-Tectonic Overview

The East European Craton (EEC) includes the sediment-covered East European Platform and the Baltic and Ukrainian Shields, where basement rocks outcrop. The Trans-European Suture Zone (TESZ) forms the transition zone between cratonic Europe and the tectonically younger western Europe, which formed by amalgamation of a series of terranes onto proto-Europe during the Phanerozoic. The main periods in the evolution of TESZ include the Caledonian (mid Palaeozoic) and Hercynian (late Palaeozoic) orogenies as well as activity in relation to the Africa-Europe collision, affecting the Carpathians and creating inversion tectonics along most of TESZ.

TESZ at the edge of the EEC includes four distinct zones (Fig. 1), which each show unique geophysical properties and structure that attests to a complex evolution and relation between deep and shallow Earth processes topography:

1. The Tornquist Fan around the Danish area, where the craton to non-craton transition splits into a 60° wide fan of fault zones that were active in a dextral transtensional environment in the late Palaeozoic (Thybo 1997). The transtension created a series of deep graben structures. This development was associated with major volcanic and magmatic activity that created huge mafic batholiths in the crust. Their solidification and cooling may have caused the formation of the up-to 5 km deep, late Palaeozoic to early Mesozoic, broad Danish-Norwegian sedimentary basin in the region. The upper mantle shows pronounced variation and contrasts in seismic velocity over a depth interval of >150 km which is not manifested in the gravity signal. The Tornquist Fan shows significant short-wavelength variation in crustal thickness and average seismic velocity. The inversion zone, STZ, formed at the NE edge of the Tornquist Fan, where the strong and thick crust of the Baltic Shield may have acted as a back-stop to Alpine compression (EUGENO-S Working Group 1988).
2. The cratonic margin in central Europe is coincident with the more than 20 km deep Polish basin. The processes that formed this deep basin are not well understood, and probably include transtension as in the Tornquist Fan. TESZ is generally characterized by major crustal thinning from the >45 km thick cratonic crust to the <35 km Caledonian/Variscan crust. However, the crustal thickness and velocity structure within the TTZ part of TESZ is variable along strike between generally thin crust in the NW and thick crust in the SE (Guterch et al. 2003). It was recognized in the 1980s that the TTZ was underlain by a Moho keel in parts of Poland, similar to the basement horst at Bornholm at the transition between TTZ and STZ. Lithosphere thickness changes abruptly across TESZ from <60 km in the SW to >150 km in the NE. Palaeozoic subduction along the entire central European Tornquist Zone has been proposed but not documented. The Alpine-related inversion also created a tectonic inversion zone within the TTZ in Poland. The magnetic lineament of TTZ and the inversion zone generally are co-located, probably because the craton edge also here acts as a backstop to compression.
3. The TESZ has been overprinted by structure related to the formation of the Carpathian Mountains in zone 3. The geometry of the bends of the Carpathian Mountains suggests a strong effect of the craton edge on the deformation processes in the adjacent younger lithosphere. The recent PANCAKE seismic project has documented a strong change in crustal thickness and velocity across TESZ from the ca. 50 km thick crust with high average velocity in the craton, across an up-to 25 km deep basin as in Poland, but here overthrust by the Carpathians, to the young Pannonian Basin with low average velocity crust thinner than 25 km (Starostenko et al. 2013a).
4. The southeastern end of the TESZ around the Danube Delta at the Black Sea may have been significantly affected by Devonian rifting, which also affected large parts of the EEC. The resulting crustal structure shows very little variation in depth to Moho and average velocity between the EEC and Phanerozoic Europe (Starostenko et al. 2013b). This region may have been affected by the closure of the Tethyan Oceans and the formation of the deep Black Sea. One may speculate if the processes associated with closure of the Palaeo-Tethys ocean may have formed background for the rifting that initiated subsidence, which may still be ongoing.

Geophysical Signature of TESZ

TESZ has been covered by a dense network of deep seismic profiles since the international profiling in Eastern Europe in the 1960–1970s and boosted by a series of pan-European programs up to the turn of the millennium (Fig. 1a).

Crustal Profile

The crustal structure displays short-wavelength variation in both seismic velocity and thickness, cf. the generalized seismic profile in Fig. 2. The crust on the western side of TESZ is almost everywhere less than 35 km thick and has very small average velocity in the terranes that were accreted to proto-Europe during the Phanerozoic. On the contrary, the cratonic crust on the eastern side of TESZ is generally more than 45 km thick and has high average velocity. Some localized sedimentary basins (aulacogens that probably formed as rift zones) have slightly thinner crust, although the major aulacogen, the Dniepr-Donetsk Rift, today shows no crustal thinning, but instead is underlain by a thick high-velocity crustal layer. This observation may be explained by magma-compensated crustal thinning as recently observed at many active and extinct rift zones (Thybo and Nielsen 2009).

TESZ is characterized by abrupt change in average velocity and crustal thickness (Artemieva and Thybo 2013). It includes deep sedimentary basins in all subregions. It appears that the abrupt change in lithosphere structure across TESZ has made it a favorable location for formation of extensional and transtensional/strike-slip basins. It is characteristic that the Carpathian Mountains were thrust over the sedimentary basins and that their formation did not cause crustal thickening.

Gravity and Magnetic Signature

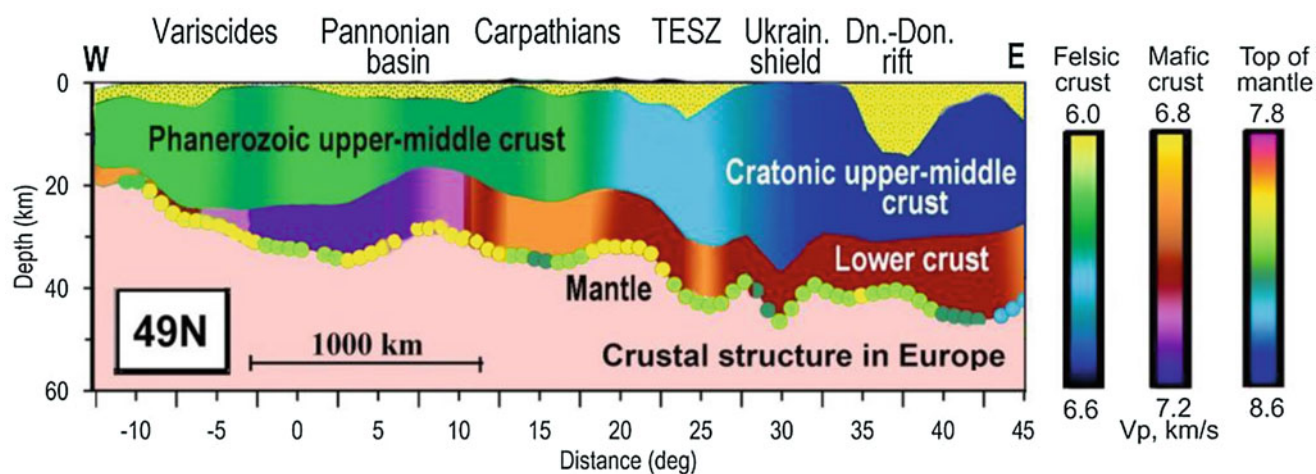
A characteristic lineament in the magnetic anomaly field (Fig. 3a) was one of the original observations of TTZ in central Europe (region 2 and 3). Here, the magnetic anomalies change abruptly from very high amplitude with short-wavelength variation in the cratonic regions to smooth, low-amplitude anomalies in the west. These differences relate to

the crustal composition and cannot be solely explained by the presence of the sedimentary cover in the west, which is evident from the observation that the sharp, abrupt change in several places occurs inside the thick sedimentary basins in TTZ and STZ.

The magnetic transition occurs gradually in the regions at the ends of TESZ (regions 1 and 4). This is particularly clear across the Tornquist Fan, which shows strong amplitudes despite the crust is relatively thin and hot. The high amplitudes have been explained and modeled by the presence of major volcanic and magmatic bodies within the sedimentary sequences and crystalline crust superimposed on the highly magnetic original Baltica crust (Thybo 1997).

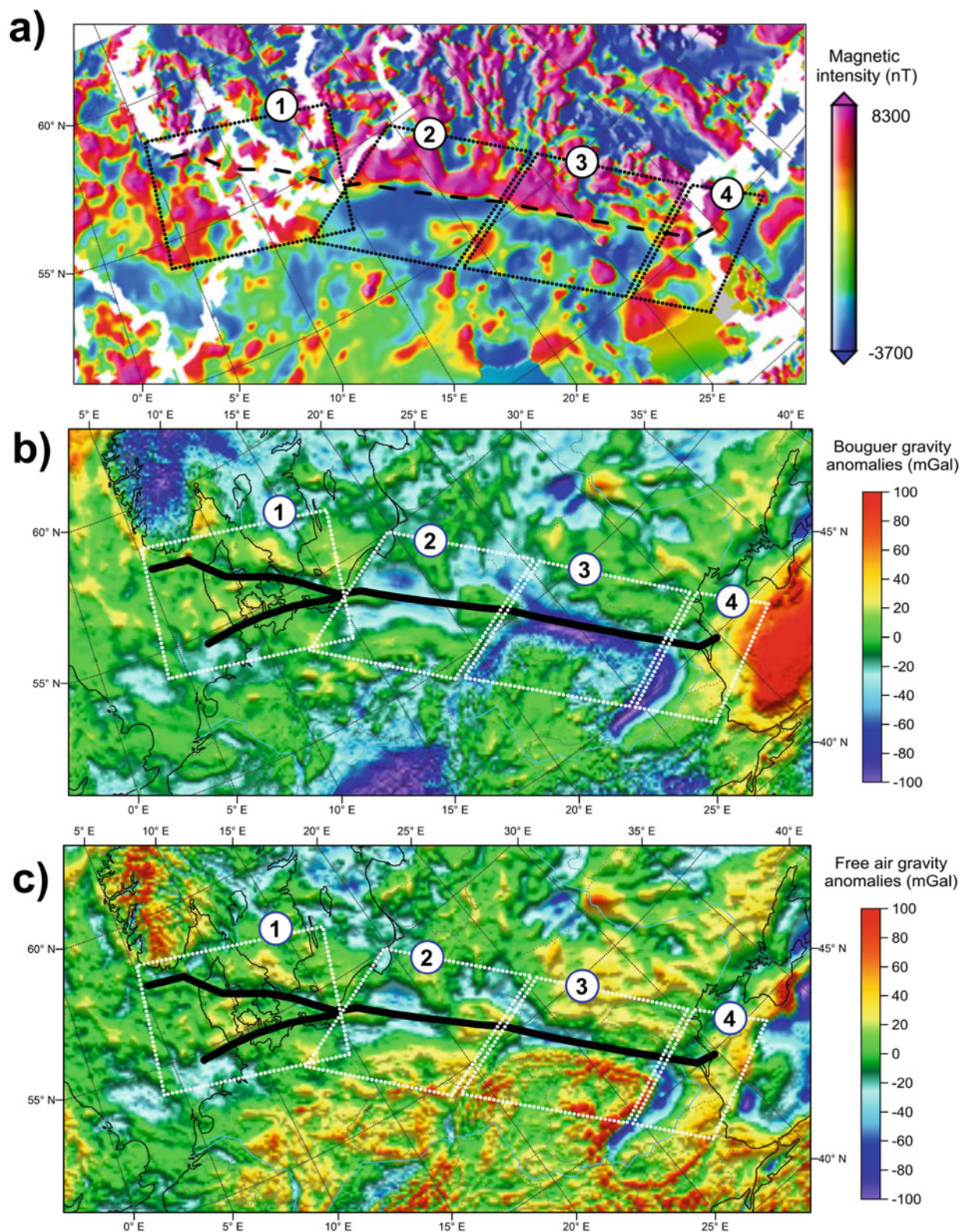
The Bouguer gravity anomaly across the TESZ (Fig. 3b) is surprisingly constant despite the abrupt change in crustal thickness. Likewise, the free air gravity anomaly (Fig. 3c) shows relatively little change across the TESZ from generally 0 mGal in the east to +10 to +20 mGal in the west, apart from the Carpathian and SW Norwegian mountain ranges, indicating general isostatic equilibrium of the lithosphere which, however, may not necessarily indicate crustal isostasy.

The largest negative Bouguer anomalies are observed in a linear belt along TTZ in region 3 to 4, including the Carpathian Mountains and the Moesian platform. Deep seismic profiles indicate that the crust has low velocity in this area, which together with the presence of a thick sedimentary cover may explain the observed negative gravity anomaly, although the thin crust would indicate positive anomalies. Likewise, the deep sedimentary basin in region 2 has negative Bouguer gravity anomalies which also may be explained by the presence of the sedimentary cover together with relatively low density crust as indicated by low seismic velocity.



Trans-European Suture Zone, Fig. 2 Generalized profile across TESZ along 49° (for location see Fig. 1b) based on the EUNASEIS database (Artemieva and Thybo 2013). Overall, the crustal thickness increases from Phanerozoic Europe in the west to cratonic Europe in the

east, accompanied by a general increase in velocity from west to east. Notice the strong short-wavelength variation in crustal thickness and velocity



Trans-European Suture Zone, Fig. 3 Maps of total magnetic field (a), Bouguer gravity (b), and free air gravity (c) anomalies in the TESZ region. Notice the sharp change in the magnetic field from the strongly magnetized crust in cratonic Europe in the east to the weakly magnetized

crust in Phanerozoic Europe in the west. The Bouguer gravity field is relatively stable across the TESZ despite the substantial change in crustal thickness. Free air gravity anomalies within ± 30 mGal indicate that the region is generally in isostatic equilibrium

The Tornquist Fan (region 1) generally includes large positive Bouguer anomalies (>50 mGal). These anomalies are interpreted by the presence of huge mafic magmatic batholiths within the crust (Thybo 1997), as identified from high-seismic velocities.

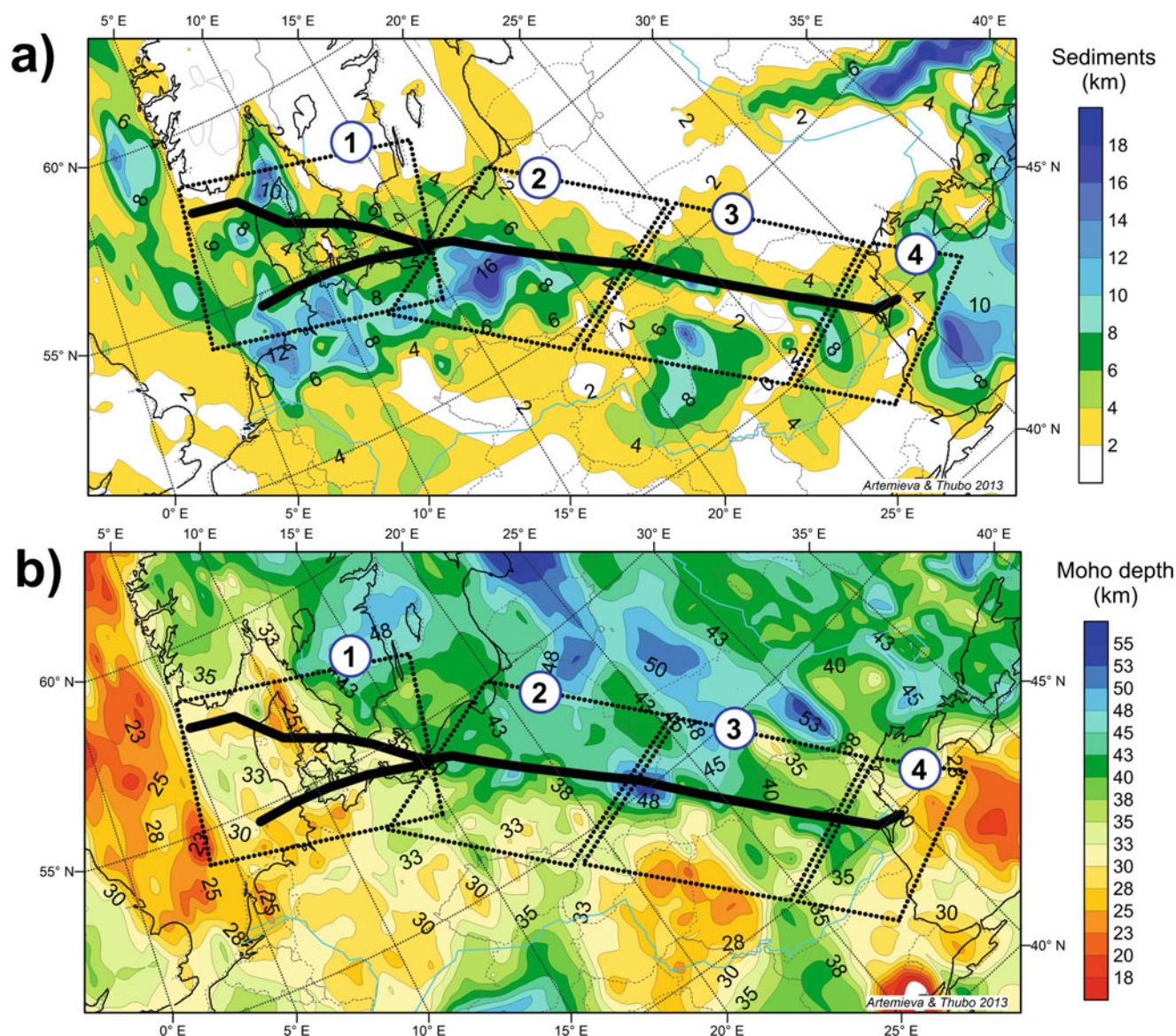
Crustal Structure from Seismic Data

The deepest sedimentary basins around TESZ occur in its northwestern end along the linear TTZ in Poland and within the Tornquist Fan, in particular along the Trans-European Fault Zone (Fig. 4a). Here the basins attain thicknesses of more than 20 km. Similar deep sedimentary basins extend all along the TTZ almost to the Black Sea coast.

Due to the dense coverage by seismic profiles in the whole TESZ region, it is possible to contour a detailed

map of depth to Moho (Fig. 4b; Artemieva and Thybo 2013). Generally, Moho is deep (>45 km) in the cratonic regions and shallow (<35 km) in younger Europe. The almost abrupt transition between these two regions occurs along the TTZ and, more transitional, in the Tornquist Fan region. The Moho depth change is also transitional in the SE end of TESZ. Very thin crust is observed in the rifted North Sea region to the NW of TESZ and in the region around the Black Sea, which may include crust of oceanic origin.

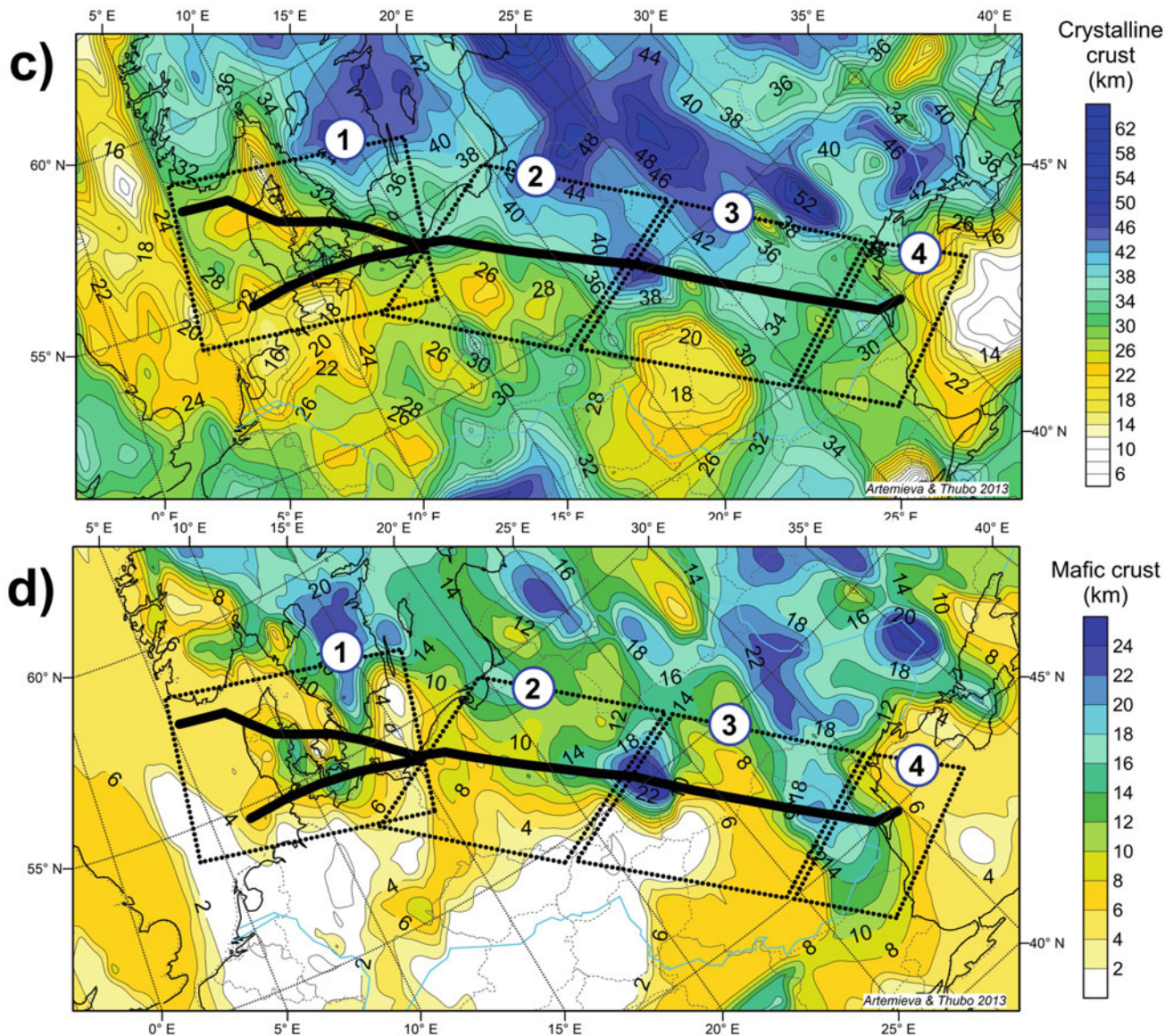
The crystalline crust (Fig. 4c) shows similar thickness variation with cratonic crust being >45 km thick and the crust in younger Europe being <30 km thick. Below the thick sedimentary basins, the crystalline crust is exceptionally thin in most of the Tornquist Fan region.



Trans-European Suture Zone, Fig. 4 (continued)

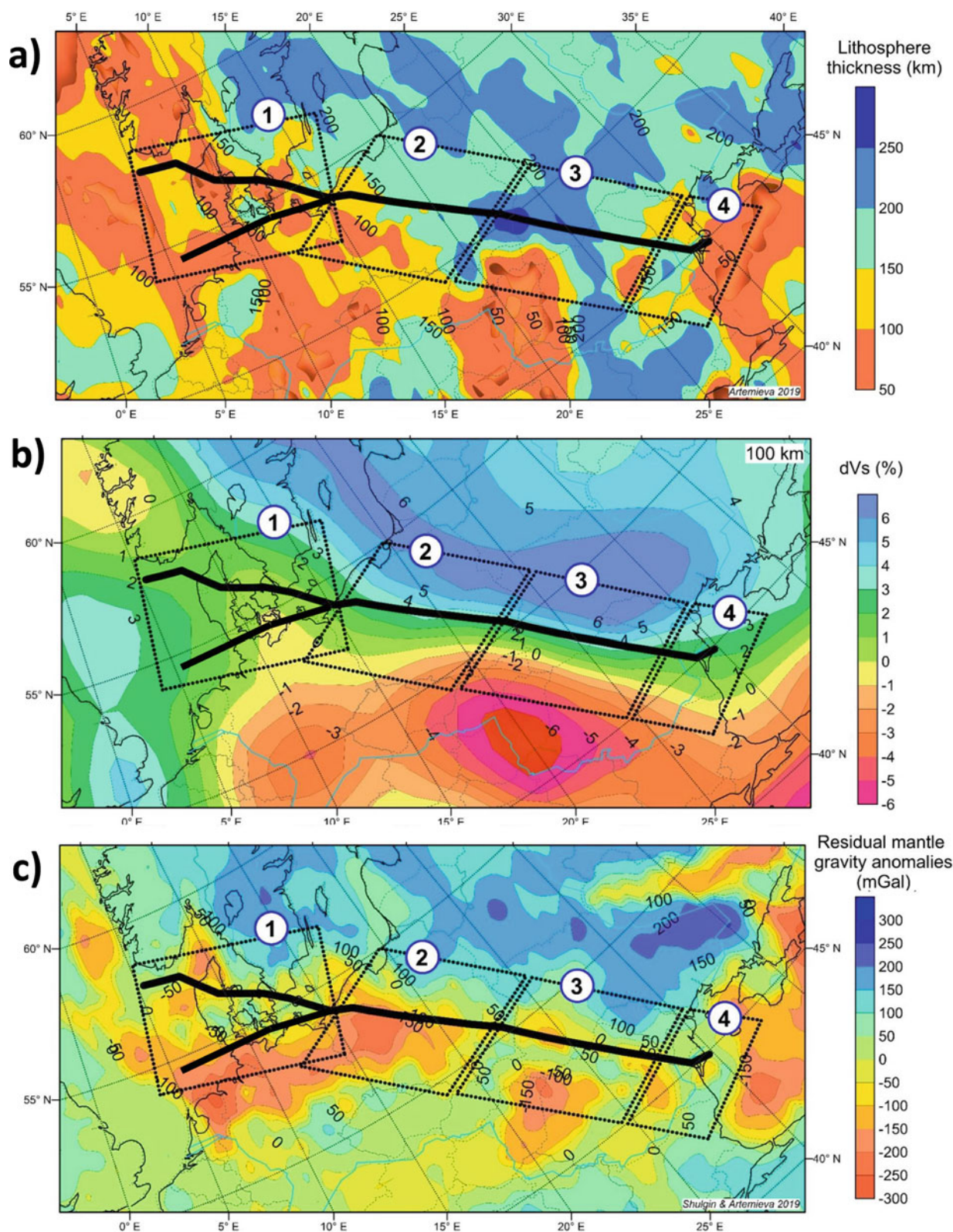
The differences in Moho depth and thickness of the crystalline crust between east and west is mainly due to the presence of a lower-crustal layer with very high seismic velocity (>6.8 km/s) in the eastern cratonic regions, which we interpret as a layer of mafic composition (Fig. 4d). Such layers are usually observed in cratonic regions, and its thickness may have to some degree increased by localized underplating since the initial formation of the cratonic crust. This may in particular be the case in the Tornquist Fan during its extensive magmatic-volcanic evolution in the late Paleozoic.

It is remarkable that no high-velocity lower crustal layer is observed in Caledonian and Variscan Europe. Such layer appears absent throughout this region, despite several of the docking terranes are believed to originate from the Gondwana supercontinent, and therefore expectedly would include such lower crustal layer. It has been proposed that a mafic layer originally was present in these terranes up to the Gondwana-Laurentia collision, and that crustal thickening and the formation of a crustal root during the Variscan orogeny may have led to pressure induced metamorphic reactions that formed eclogite facies rocks from the original mafic lower crust.



Trans-European Suture Zone, Fig. 4 Maps showing seismic properties of the crust around TESZ. (a) Thickness of the sedimentary cover, showing a series of very deep sedimentary basins concentrated in TESZ along the TTZ and around the TF area. (b) Moho depth showing a strong change from thick crust in cratonic Europe to the east to thin crust in Phanerozoic Europe in the west. (c) Thickness of the crystalline crust

showing an even more pronounced change than for the crustal thickness. (d) Thickness of the lower (mafic) crust, showing that the difference in thickness of the crystalline crust mainly is caused by the lack of mafic lower crust in Phanerozoic Europe in the west. (Figure after Artemieva and Thybo (2013))



Trans-European Suture Zone, Fig. 5 Mantle properties in the TESZ region. **(a)** Lithosphere thickness (from Artemieva, 2019), showing the generally very thick lithosphere in cratonic Europe and thin lithosphere in Phanerozoic Europe in the west. **(b)** S-wave velocity anomalies at 100 km depth, showing a change from high velocity in the east to low

velocity in the west for the lithospheric mantle. **(c)** Residual mantle gravity anomalies calculated by subtracting the crustal contribution from the observed gravity field, showing that TESZ marks a transition from high density in the east to low-density in the west for the lithospheric mantle

Some authors have proposed that the formation of such dense lower crustal layer could have led to its delamination. The existence of the presently very thin lithosphere in this region may provide some support to this model, although an alternative interpretation would be that the eclogitic rocks are still present below the seismic Moho with seismic velocities and densities similar to peridotitic upper mantle.

Lithospheric Mantle in TESZ

The lithosphere thickness variation in the TESZ region generally shows similar variation as the crustal thickness. Although the lateral resolution is lower than for the crustal thickness, it appears that the variation occurs over a narrow zone between the very thick (>250 km) lithosphere below the cratons and the very thin (<60 km) lithosphere in Caledonian/Variscan Europe (Fig. 5a; Artemieva et al. 2006). The thin lithosphere in southern Norway (NW part of the map in Fig. 5a) indicates that the extensional tectonics and magmatism within the Tornquist Fan affected the whole lithosphere.

A map of S-wave anomaly at a depth of 100 km in the TESZ region shows a very distinct and abrupt change at the TTZ between very high velocities in the cratonic regions and very small velocities in younger Europe (Fig. 5b). This change in velocity may be partly due to a temperature change between craton and platform, but studies have shown that a substantial part of the change is caused by compositional change. It is remarkable that the Tornquist Fan area has intermediate S-wave velocity at 100 km depth.

This interpretation is supported by calculated residual mantle gravity anomalies (Fig. 5c), which show a similar abrupt change across the TTZ between large positive values in the cratonic regions and negative values in younger Europe. The variation in values (300 to –200 mGal) is pronounced, not least considering that neither Bouguer nor free air anomalies show any significant anomalies (generally less than 40 mGal in total, Fig. 3). Also the mantle gravity values in the Tornquist Fan are similar to the values for young Europe despite the crustal structure and the age of the basement rocks suggest a cratonic origin.

Conclusions

The NW-SE trending Trans-European Suture Zone (TESZ) forms the most pronounced geophysical- tectonic-geological transition zone in Europe affecting the whole lithosphere. It marks the transition between the Precambrian cratonic regions in the northeast and younger Europe, formed by amalgamation of a series of generally Gondwana-derived terranes, in the southwest.

TESZ was originally observed from a clear lineament in magnetic anomalies between high-amplitude, low-wavelength

anomalies in the east, and low-amplitude, long-wavelength anomalies in the west. It is observable by lithospheric anomalies in all geophysical parameters. Seismic velocities and densities are large on the cratonic side and small on the western side, both in the crust and lithospheric mantle. The crust and lithosphere are thick on the cratonic side and thin on the western side. Some of the very deepest sedimentary basins in Europe are located within TESZ and a major late-Palaeozoic volcanic-magmatic region is observed in the Tornquist Fan region in northwestern TESZ.

Bibliography

- Artemieva IM (2019) Lithosphere structure in Europe from thermal isostasy. *Earth Sci Rev* 188:454–468. <https://doi.org/10.1016/j.earscirev.2018.11.004>
- Artemieva IM, Thybo H (2013) EUNaseis: a seismic model for Moho and crustal structure in Europe, Greenland, and the North Atlantic region. *Tectonophysics* 609:97–153. <https://doi.org/10.1016/j.tecto.2013.08.004>
- Artemieva IM, Thybo H, Kaban MK (2006) Deep Europe today: geophysical synthesis of the upper mantle structure and lithospheric processes over 3.5 Ga. *Geol Soc Lond Mem* 32:11–41. <https://doi.org/10.1144/gsl.mem.2006.032.01.02>
- BABEL Working Group (1993) Deep seismic reflection/refraction interpretation of crustal structure along BABEL profiles A and B in the southern Baltic Sea. *Geophys J Int* 112:325–343
- Berthelsen A (1992) Mobile Europe. In: Blundell D, Freeman R, Mueller S (eds) *A continent revealed; the European Geotraverse*. Cambridge University Press, Cambridge, UK, pp 11–32
- EUGENO-S Working Group (1988) Crustal structure and tectonic evolution of the transition between the Baltic Shield and the North German Caledonides (the EUGENO-S Project). *Tectonophysics* 150:253–348
- Guterch A, Grd M, Keller GR, Posgay K, Vozar J, Spicak A, Bruckl E, Hajnal Z, Thybo H, Selvi O (2003) CELEBRATION 2000 seismic experiment. *Stud Geophys Geod* 47:659–669
- Starostenko V, Janik T, Kolomiyets K, Czuba W, Środa P, Grad M, Kovács I, Stephenson R, Lysynchuk D, Thybo H, Artemieva IM, Omelchenko V, Gintov O, Kutas R, Gryn D, Guterch A, Hegedűs E, Komminaho K, Legostaeva O, Tiira T, Tolkunov A (2013a) Seismic velocity model of the crust and upper mantle along profile PAN-CAKE across the Carpathians between the Pannonian Basin and the East European Craton. *Tectonophysics* 608:1049–1072. <https://doi.org/10.1016/j.tecto.2013.07.008>
- Starostenko V, Janik T, Lysynchuk D, Środa P, Czuba W, Kolomiyets K, Aleksandrowski P, Gintov O, Omelchenko V, Komminaho K, Guterch A, Tiira T, Gryn D, Legostaeva O, Thybo H, Tolkunov A (2013b) Mesozoic(?) lithosphere-scale buckling of the East European Craton in southern Ukraine: DOBRE-4 deep seismic profile. *Geophys J Int* 195:740–766. <https://doi.org/10.1093/gji/ggt292>
- Thybo H (1997) Geophysical characteristics of the Tornquist Fan area, northwest Trans-European Suture Zone: indication of late Carboniferous to early Permian dextral transtension. *Geol Mag* 134:597–606. <https://doi.org/10.1017/s0016756897007267>
- Thybo H, Nielsen CA (2009) Magma-compensated crustal thinning in continental rift zones. *Nature* 457:873–876. <https://doi.org/10.1038/nature07688>
- Winchester JA, PACE Network Team (2002) Palaeozoic amalgamation of Central Europe: new results from recent geological and geophysical investigations. *Tectonophysics* 360:5–21. [https://doi.org/10.1016/S0040-1951\(02\)00344-X](https://doi.org/10.1016/S0040-1951(02)00344-X)

Traveltime Tomography Using Controlled-Source Seismic Data

Colin A. Zelt

Department of Earth Science, Rice University, Houston, TX, USA

Synonyms

Active-source data; Traveltime inversion

Definition

Controlled-source seismology	Determination of earth structure using seismic waves generated by artificial sources such as chemical explosives, mechanical vibrators, weight drops, gun shots, etc.
Tomography	Determination of model structure by back projection of the data along a path connecting a source and a receiver.
Traveltime	The total time for a particular wave to travel from a source to a receiver.

Background

Traveltime tomography is the main method by which the Earth's seismic velocity structure is determined on all scales, from the upper few meters to the whole mantle. It was adapted from algorithms used in medical imaging in the 1970s (Dziewonski and Anderson 1984). In seismology, the term tomography refers to the back projection of data along a path connecting a source and receiver using a mathematical inverse method. Ideally, the traveltimes of seismic arrivals corresponding to many criss-crossing rays are used to construct a two-dimensional (2D) or 3D image of the Earth's seismic velocity variations. Tomography is a type of inverse problem, although the terms traveltime inversion and traveltime tomography are used in the seismological community without a well-defined distinction. The latter usually implies a uniform, fine grid model parameterization, in which a smooth model is sought. Any departure from this approach is usually referred to as traveltime inversion, although the distinction is somewhat arbitrary (Levander et al. 2007); this article will discuss both approaches. This article concerns traveltime tomography using controlled-source data and does not discuss other types of tomography that use seismic amplitudes, phases, or waveforms as input data, nor earthquake or surface-wave data.

Data from earthquakes and artificial (controlled) sources are used in traveltime tomography, the former usually for crustal or mantle structure, the latter usually for crustal or near-surface structure. Much of the theory and practice of traveltime tomography is the same or similar for controlled-source and earthquake data. However, earthquake tomography usually uses delay times and includes a determination of the source locations (hypocenters), whereas controlled-source tomography usually uses total times and the source locations are known very accurately. Also, the degree of lateral velocity heterogeneity in the crust, and in particular the near-surface, can be much greater than in the mantle, so there is often no sense of a reference model in controlled-source seismology as there is in earthquake seismology, and this in turn means that a nonlinear tomographic method must be applied to controlled-source data in which the rays must be determined along with the unknown velocity structure as part of the inverse problem. Finally, controlled-source data are often acquired using sources and receivers distributed along a straight or nearly straight line, yielding a so-called 2D dataset. In this case, 2D modeling is used in which lateral homogeneity perpendicular to the line is assumed. Earthquake tomography is almost always 3D given the typical distribution of earthquakes and stations. Today, 3D controlled-source data is not uncommon, but it is expensive to acquire, especially for crustal-scale studies. In some studies, both earthquake and controlled-source data have been inverted simultaneously (e.g., Ramachandran et al. 2005).

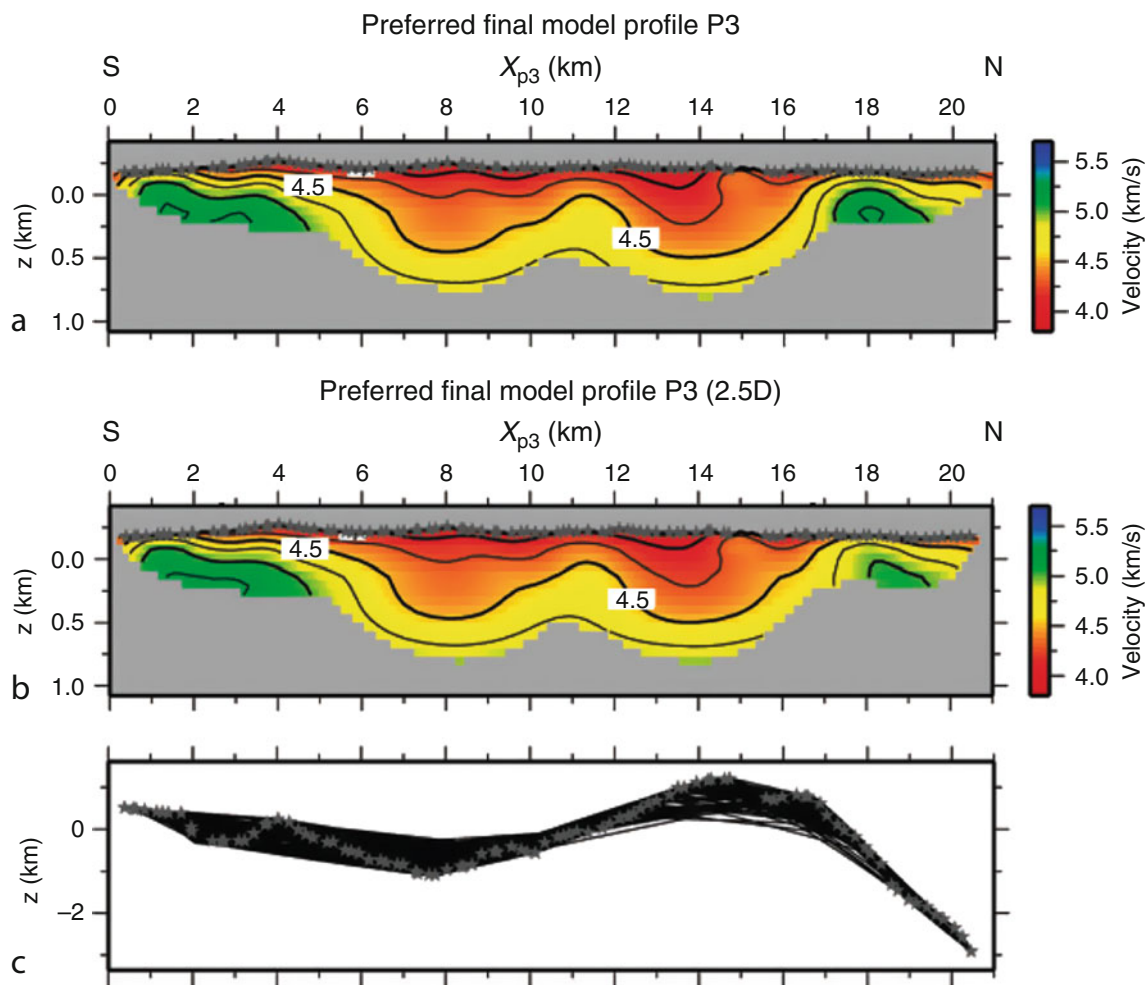
Controlled-source seismology can be divided into two main approaches: (1) near-vertical reflection methods, and (2) refraction/wide-angle reflection methods. Traveltime tomography is applied to data from both experiments, although it is more common for refraction data. The most common applications of traveltime tomography using controlled-source data are (1) tectonic studies of the crust and uppermost mantle, (2) earthquake risk studies for upper crustal fault and basin geometries, (3) petroleum exploration in crosswell studies, model building to facilitate migration, and for refraction statics, and (4) environmental/engineering studies of the near surface (as shallow as a few meters). The traveltimes utilized are often limited to those of the first arrivals corresponding to direct, refracted (turning) or diffracted waves. Sometimes the traveltimes of primary reflections are included, for example, from the basement overlying sediments or the crust-mantle (Moho) boundary, necessitating the need for layers or interfaces in the model parameterization. Only rarely are later refracted arrivals, multiples or conversions utilized. Traveltimes are picked from the seismograms usually either interactively by eye, or using a semi-automated scheme whereby a few picks are made interactively, and the intervening picks are determined automatically using a cross-correlation scheme (Zelt 1999). Usually

only very high quality, spatially-dense data can be picked using fully automated schemes.

Most controlled-source tomography uses only P-wave traveltimes to constrain the compressional wave-velocity structure because P wave are easiest to generate and therefore the P wave arrivals are usually the strongest and easiest to pick. However, S wave experiments are quite common in environmental/engineering studies (e.g., Chambers et al. 2009). Theoretically, traveltime tomography is the same for P- and S-wave data, and if both data are available in the same study, Poisson's ratio can be estimated. The forward modeling component of traveltime tomography is almost always based on ray theory, an infinite-frequency approximation of wave propagation, although recently a finite-frequency approach

for controlled-source data has been proposed (Zelt 2009). Almost all applications of traveltime tomography assume isotropic media because it is rarely necessary to invoke anisotropy to model the observed data. Exceptions to this are mainly for crosswell data and in studies of the oceanic crust (e.g., Pratt and Chapman 1992; Dunn et al. 2000). Traveltime tomography is applied to land and marine data, as well as onshore-offshore data, using the same tomographic methodology. Arrays of airguns are the most common marine source, while on land chemical explosives, mechanical vibrators, weight drops, hammers, shotguns and rifles are common sources.

The ideal dataset for traveltime tomography is one with as dense a spacing of sources and receivers as possible, along a



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 1 Shallow crustal example from the South Portuguese Zone fold-and-thrust belt of first-arrival tomography by Schmelzbach et al. (2008) using the smoothing-regularized Zelt and Barton (1998) algorithm to seek a minimum-structure model. There were 99 shots (grey stars) along the crooked line. (a) Velocity model obtained by projecting all the data onto a straight line and performing a 2D inversion. (b) Velocity model

obtained by performing a 2.5D inversion that preserves the true positions of all shots and receivers. (c) Rays projected into the horizontal plane for the 2.5D modeling to illustrate the extreme crookedness of the experiment. Model regions not sampled by rays are grey. The models in (a) and (b) are very similar, confirming the appropriateness of 2D modeling of traveltime data, even in the case of extreme crooked lines

line for a 2D experiment, or areally distributed for a 3D experiment. As a result, a subsurface point will be sampled by rays at as many different angles as possible, which in turn will yield the best possible spatial model resolution as discussed later. There is usually very little processing of seismic data before traveltime picking, only that required to “clean up” the data to make it easier for picking. This may include bandpass filtering, trace editing, velocity filtering, trace mixing or binning, or deconvolution. Picks should not be interpolated to provide uniform spatial coverage if there are significant trace gaps, since this will provide an incorrect sense of model constraint (Zelt 1999). Uncertainties should be assigned to the pick times to avoid over- or under-fitting the data, and to allow the appropriate up- and down-weighting of less noisy and more noisy data, respectively (Zelt 1999). Elevation or bathymetric corrections should be avoided as they depend on the unknown velocity structure. Instead, the known surface topography or bathymetry should be incorporated into the 2D or 3D model. For a 2D experiment in which the shot and receiver locations deviate significantly from a straight line, it may be necessary to perform 2.5D modeling, i.e., using a 3D model and 3D ray tracing, and the known topographic or bathymetric surfaces, but keeping the velocity model homogeneous in a direction perpendicular to the main trend of the profile (e.g., Van Avendonk et al. 2004). Given the spatial resolution of most velocity models derived from traveltime tomography, the degree of profile “crookedness” must be substantial to warrant 2.5D modeling (Schmelzbach et al. 2008; Fig. 1).

There are several review papers relevant to this article. Nowack and Braile (1993) review traveltime tomography methods in 1D and 2D media. Zelt (1999) focuses on 2D traveltime tomography and inversion, as well as model assessment methods. Rawlinson and Sambridge (2003a) present an overview of traveltime tomography for 2D and 3D structure. Levander et al. (2007) review all aspects of crustal-scale controlled-source seismology, including traveltime tomography.

Forward Modeling

In the 1970s, ray theory, an infinite-frequency approximation of wave propagation, for laterally heterogeneous media was developed (Cerveny et al. 1977; McMechan and Mooney 1980). This allows earth models of arbitrary 2D and 3D complexity to be considered, provided the tenants of ray theory are honored, that is, the velocity field varies slowly with respect to the seismic wavelengths. Sharp velocity changes are modeled as velocity discontinuities (layer boundaries) using Snell’s law. As a result, ray theory models are typically composed of one or a few layers, bounded by smoothly varying interfaces, within which the velocity varies

smoothly. Ray theory forms the basis for the forward modeling used today in almost all traveltime tomography algorithms. There are two main types of infinite-frequency modeling: ray tracing and wavefront tracking.

Two end-member approaches to ray tracing are used: numerical and analytical. In the former case, the velocity field is specified by a regular (e.g., McMechan and Mooney 1980) or irregular (e.g., Zelt and Ellis 1988) grid of nodes, together with a scheme for interpolation between nodes, and a ray is traced by specifying a take-off angle from a point source and solving a set of ordinary differential equations, the ray tracing equations (Cerveny et al. 1977). In the analytic case, the velocity field is specified by a regular (e.g., Chapman and Drummond 1982; White 1989; Rawlinson et al. 2001) or irregular (e.g., Spence et al. 1984) grid of nodes, but with a form of interpolation, for example linear, such that the rays can be calculated analytically within each cell of the grid. In practice, the numerical and analytical techniques may be about equally efficient if the analytical method uses a fine grid of nodes to represent a sufficiently smooth velocity field as required by ray theory, and if the numerical scheme adjusts the step length according to the velocity gradient (Zelt and Ellis 1988).

For an arbitrary 2D velocity model, a non-trivial consideration is the determination of ray take-off angle from the source such that an arbitrary receiver location is reached, or such that a particular layer or interface in the model is reached. Today, most 2D ray tracing algorithms offer an automatic determination of take-off angles by shooting trial rays followed by a bisection scheme to sufficiently refine the take-off angles so that the receivers are hit (e.g., Zelt and Smith 1992). The robustness of the shooting method is crucial because the data from receivers for which it is not possible to find rays will not be used in the inverse step. Also, it is important to be sure that an inability to find rays is because of the velocity model, e.g., a shadow zone, as opposed to a shortcoming of the ray tracing algorithm.

To avoid the limitations of shooting/bisection algorithms, particularly for 3D models, ray bending methods were developed (Um and Thurber 1987). In this approach, a ray connecting a source and receiver is estimated, typically using a 1D reference model, and the ray’s path is iteratively updated using Fermat’s principle until the minimum time path is determined. In practice, shooting methods are more efficient and sufficiently robust for 2D models, whereas bending methods are favored for 3D models.

In the late 1980s, a new infinite-frequency forward modeling approach was introduced in which first-arrival traveltimes are calculated on a fine grid using a finite-difference solution of the eikonal equation (Vidale 1988, 1990; Podvin and Lecomte 1991). Rays are obtained by following the gradient of the time field. These methods are known as wavefront tracking algorithms. Reflected rays can be calculated in a

two-step procedure by calculating downgoing and upgoing wavefronts (Hole and Zelt 1995). Hole and Zelt (1995) also presented a modification to Vidale's approach which otherwise breaks down for velocity contrast of more than ~40%. The advantage of wavefront tracking methods is that they find the ray between any two points with the shortest traveltime, including the diffracted path in the case of a geometrical shadow zone, and they can be very efficient, especially for 3D models. The disadvantages are that they are computationally cumbersome for calculating later arrivals, and they can be inaccurate and/or computationally intensive for high-contrast media. Another type of wavefront tracking scheme that solves the eikonal equation is the fast-marching method, which is computationally efficient and unconditionally stable (Sethian and Popovici 1999).

Nakanishi and Yamaguchi (1986) and Moser (1991) introduced the shortest path ray-tracing method based on Fermat's principle which uses a fine network of nodes and graph theory. It is capable of handling arbitrarily heterogeneous media and calculating first arrivals, reflections, and multiples. Its advantages over other methods are robustness and the fact that the grid does not have to be rectangular or regular. However, its memory and computation requirements are greater. Van Avendonk et al. (2001) developed a hybrid of the shortest path and ray bending methods that uses the shortest path algorithm to find an initial ray, and then refines it using a bending method. The result is an efficient and robust algorithm to calculate accurate traveltimes and rays for refractions and reflections in 2D and 3D media.

Traveltime Tomography and Inversion

The traveltime t between a source and receiver along a ray L is given in integral form for a velocity field $v(\mathbf{r})$ as

$$t = \int_L \frac{1}{v(\mathbf{r})} dL$$

where $\mathbf{r}$ is the position vector in 2D or 3D media. This is a nonlinear inverse problem given the relationship between the measured data (traveltimes) and the unknown model parameters (the velocity field). However, by transforming variables to use slowness $s(\mathbf{r})$, the reciprocal of velocity, instead of velocity as the unknown, a seemingly linear inversion problem is created:

$$t = \int_L s(\mathbf{r}) dL.$$

However, the ray L is also dependent on the velocity (or slowness) model, thus making the inverse problem

nonlinear regardless of what form of model variable or parameterization is used. In global seismology, a 1D (radially-symmetric) earth model is often used to predict the rays given the relatively small velocity heterogeneities in the mantle, and thereby converting the inverse problem into a linear one. In controlled-source seismology, there is no concept of a reference model that is sufficiently accurate to predict the rays in advance given the relatively large velocity heterogeneities that are typically encountered in the crust. This means the controlled-source tomography or inverse problem is always treated as a nonlinear one. This also means the model can be parameterized any number of ways using velocity or slowness, and cells, nodes, or splines, since the problem's nonlinearity must be dealt with regardless of the parameterization. Most often a linearized gradient approach is applied in which a starting model is used and both the model and rays are updated over a series of iterations with the hope that there will be convergence to an acceptable model (the final model).

The model is almost always discretized using cells, nodes, splines, or other interpolating functions; in the latter two cases, the discrete model parameters are the coefficients of the interpolating functions. In the simplest formulation of the tomography problem, the model is parameterized using constant-slowness cells, in which case the equation for the i th data becomes

$$t_i = \sum_j l_{ij} s_j$$

where l_{ij} is the length of the i th ray in the j th model cell and s_j is the slowness in the j th cell. In vector-matrix form this is

$$\mathbf{t} = \mathbf{L}\mathbf{s}.$$

Applying a Taylor series expansion to this equation keeping only the linear term and assuming the rays are independent of slowness yields the perturbation equation

$$\delta\mathbf{t} = \mathbf{G}\delta\mathbf{s}$$

where $\delta\mathbf{t}$ is the data misfit vector equal to the difference between the observed traveltimes and those predicted by a prior model, and $\delta\mathbf{s}$ is the difference between the unknown slowness model and the prior slowness model; the unknown model is also called the estimated model. The partial derivative matrix, $\mathbf{G}$, contains the elements $g_{ij} = \partial t_i / \partial s_j$, and for the constant-slowness cell parameterization, $g_{ij} = l_{ij}$, or $\mathbf{G} = \mathbf{L}$. In the general case, the elements of the partial derivative matrix are $g_{ij} = \partial t_i / \partial m_j$ where m_j is the j th model parameter, which could be the velocity at a node or the coefficient of an interpolating function, or the position of an interface within

the velocity model (e.g., Zelt and Smith 1992). In the general case, the perturbation equation becomes

$$\delta \mathbf{t} = \mathbf{G} \delta \mathbf{m}.$$

The elements of the partial derivative matrix $\mathbf{G}$ are usually calculated analytically to avoid the potential inaccuracy of numerical differencing and the extra computation needed to trace additional rays. The approximations involved in the analytic partial derivatives, e.g., the stationary ray assumption, are typically not a problem since the resultant gradient direction in model space will be improved over a series of iterations.

The perturbation equation is not solved directly because the unknown model parameters are typically under- or mixed-determined (Menke 1989), depending on the model parameterization; with relatively few model parameters, it is possible for the problem to be overdetermined, but even in this case the perturbation equation is not solved directly because it is wise to constrain the magnitude of the model perturbation to avoid violating the linearization assumption. Thus, model constraints, in addition to the observed data, are usually included to stabilize the solution. Also, all observed data contain noise and to avoid over-fitting noisy data, additional model constraints are required to select one model from an infinite number that will statistically predict the observed data at the equivalent desired misfit according to the estimated noise level. This issue points out the non-uniqueness of all inverse problems involving real data, regardless of whether the system of equations is overdetermined or underdetermined.

The additional model constraint is called regularization (e.g., Scales et al. 1990), and in addition to stabilizing a

solution, it is an effective way of steering the solution towards models with desirable pre-defined characteristics. The regularization is most often in the form of the zero-, first- or second-order spatial derivatives of the estimated model parameters, or their perturbation from a background model. This is because it is often desirable to seek an estimated model that is as close as possible to another model, and/or a model that has a minimum structure as measured by its spatial derivatives in keeping with Occam's principle (Constable et al. 1987). Regularization is always arbitrary, and the specific form of regularization that is best for a particular problem depends on the model parameterization, the data, including its coverage and noise level, and the experimental objectives.

The most common forms of regularization in traveltime problems are the minimization of the model's perturbation from a background model, and/or the minimization of the second-order spatial derivative of either the model, or its perturbation from a background model. A model with minimum perturbation is called the smallest model; a model with minimum second-order derivative is called the smoothest model. Any combination of model constraints may be included as a part of the regularization to form the objective function, $\Phi(\mathbf{m})$, where $\mathbf{m}$ is the vector containing the estimated model parameters. The objective function usually measures the square of the data misfit vector and the square of the perturbation or spatial derivative operators comprising the regularization so that minimizing the objective function leads to a linear system of equations that is amenable to efficient algorithms for their exact or approximate solution. One example of an objective function for the traveltime problem is

$$\Phi(\mathbf{m}) = \delta \mathbf{t}^T \mathbf{C}_d^{-1} \delta \mathbf{t} + \lambda \{ \beta [\alpha (\mathbf{v}^T \mathbf{W}_h^T \mathbf{W}_h \mathbf{v} + s_z \mathbf{v}^T \mathbf{W}_v^T \mathbf{W}_v \mathbf{v}) + (1 - \alpha) \Delta \mathbf{v}^T \mathbf{W}_p^T \mathbf{W}_p \Delta \mathbf{v}] + (1 - \beta) [\mathbf{z}^T \mathbf{W}_z^T \mathbf{W}_z \mathbf{z}] \}$$

where $\mathbf{m} = \mathbf{v} + \mathbf{z}$; $\mathbf{v}$ and $\mathbf{z}$ are vectors containing the estimated model parameters defining the velocity (or slowness) field and the interfaces (if any); $\Delta \mathbf{v}$ is the velocity perturbation vector equal to $\mathbf{v} - \mathbf{v}_0$, and $\mathbf{v}_0$ is the background velocity model vector. $\mathbf{C}_d$ is the data covariance matrix containing the estimated pick uncertainties providing the appropriate up-weighting and downweighting of the low-noise and high-noise data, respectively. $\mathbf{W}_h$ and $\mathbf{W}_v$ are the horizontal and vertical roughness matrices containing the second-order spatial finite-difference operators that measure the roughness of the velocity field in the horizontal and vertical directions, respectively; $\mathbf{W}_p$ is the velocity perturbation weighting matrix which is a diagonal matrix containing the weights applied to the perturbation between each estimated and background velocity model parameter; $\mathbf{W}_z$ is the interface roughness

matrix containing the second-order spatial finite-difference operators that measure the roughness of the interfaces. When applying spatial derivative operators to the velocity field, it is common to use different operators to measure the horizontal and vertical structure, and weight these differently, since one would normally expect the earth to contain more heterogeneity vertically compared to horizontally. Any or all of these operators may be normalized by the reference model values so that relative quantities are penalized as opposed to absolute values (e.g., Toomey et al. 1994). In addition, the operators can be weighted according to their spatial position in the model, e.g., penalize model structure more in the deep portion of the model versus the shallow portion since one would generally expect decreasing resolution with depth.

There are four free parameters in this example objective function to control the relative weight of each term. λ determines the overall amount of regularization, that is, the trade-off between fitting the data and constraining the model; β determines the trade-off between velocity and interface regularization; α determines the trade-off between second derivative and perturbation regularization of the velocity parameters; and s_z specifies the relative weight of vertical versus horizontal smoothing regularization of the velocity parameters. In some algorithms, λ is not a free parameter because it is reduced automatically by the algorithm at each iteration from a free-parameter starting value, λ_o (e.g., Zelt and Barton 1998). The reduction of λ stabilizes the inversion by constraining the long-wavelength structure in the initial iterations and allowing progressively finer model structure to enter in later iterations.

Minimizing the objective function with respect to the unknown model parameters ($\mathbf{v}$ and $\mathbf{z}$) leads to a linear system of equations that may be relatively small or very large, depending on the number of data and model parameters. If the linear system is large, it will also typically be very sparse (at least 99.9%), in which case there are efficient algorithms for their solution (e.g., Paige and Saunders 1982; Rawlinson et al. 2001). If the linear system is relatively small, it can be solved directly using standard matrix inversion routines, such as LU decomposition (Press et al. 1992).

The objective function presented is designed to illustrate in a single example the different types of regularization that are most commonly applied. In practice, the objective function will typically be simpler depending on the data, model parameterization, and experimental goal. This objective function can be simplified or modified to correspond to those used in most of the popular traveltime inversion and tomography algorithms used today. For example, the Zelt and Smith (1992) inverse method includes only perturbation constraint on the velocity and interface parameters with respect to the model from the previous iteration; the Zelt and Barton (1998) tomographic method includes only smoothness constraints on the velocity field. Thus, together with their very different model parameterizations to be discussed in the next section, these two algorithms can be viewed as end-members of the same regularized inverse approach, and most algorithms in use today fall somewhere in between. In fact, the main differences between the algorithms discussed in the next section lies not in the details of the objective function, but more so in the form of model parameterization, forward calculation, and the types of arrivals considered.

For layered models, including both velocity and interface parameters in a single inversion scheme, as opposed to solving for each parameter separately, either one after the other, or by alternating between the two, makes the problem significantly more challenging. There will be more model non-uniqueness because of trade-offs between the two parameter

types, and more prior information will be required to deal with the increased ambiguity. Also, more testing will be required to ensure that the tradeoffs are fully understood. But solving for both parameter types simultaneously allows one to explore the trade-offs more thoroughly and is the only way to ensure the overall minimum-structure model is obtained, if that is the objective.

Assuming uncorrelated, Gaussian picking errors, and a model parameterization that does not limit the degree to which the predicted data can match the variability of the observed data, a final model should be chosen that provides a normalized misfit, χ^2 , of one (Bevington 1969).

$$\chi^2 = \frac{1}{N} [\delta \mathbf{t}^T \mathbf{C}_d^{-1} \delta \mathbf{t}]$$

where N is the number of data. A model that provides a χ^2 value less than one means the observed data are being over-fit, and therefore to some extent, the noise in the data is also being fit. Although the χ^2 criterion is a good rule of thumb, there are several instances in which it should not be followed. First, traveltime picking errors are likely not uncorrelated. Second, for spatially sparse data from a region with strong lateral heterogeneities, it may be necessary to use so many independent model parameters in order to achieve $\chi^2 = 1$ that the constraint on some parameters is unacceptably small, in which case fewer parameters may be preferable and a χ^2 value greater than one is allowed (Zelt 1999). An inability to achieve $\chi^2 = 1$ may be because of inconsistent picking, for example for reciprocal pairs, in which case the inconsistencies should either be corrected or used as the basis for assigning the pick uncertainties in the first place (e.g., Zelt et al. 2006a).

Algorithms

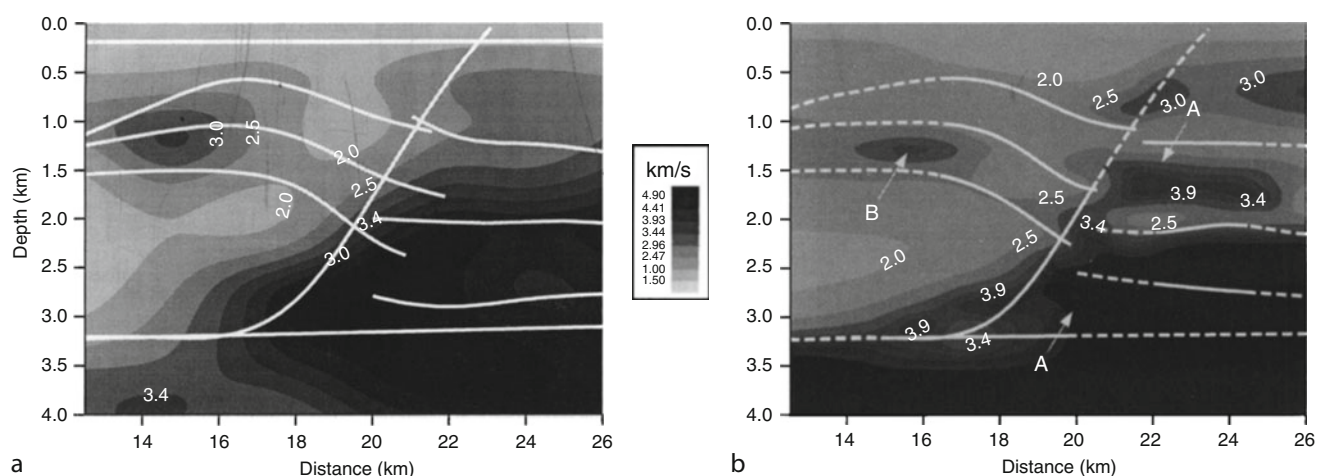
This section describes many of the algorithms available today, each with its own capabilities and limitations. Some of the earliest 2D traveltime tomography and inversion algorithms were presented by Firbas (1981) and White (1989) which used regular parameterizations and only first arrivals, and Spence et al. (1985) which allowed an irregular model grid and later arrivals. Lutter and Nowack (1990) and Lutter et al. (1990) developed a 2D inversion algorithm using a regular grid of nodes and numerical ray tracing that allows for the independent inversion of first arrivals and reflections for velocity and interface geometry, respectively. The Zelt and Smith (1992) 2D algorithm is the opposite in many respects. An irregular grid of velocity and interface nodes can be used, and any type of refracted or reflected arrival can be inverted simultaneously for velocity and interface geometry. In addition, unique to the Zelt and Smith (1992) algorithm, it allows any subset of the total set of model parameters to be inverted

for, holding all others fixed, facilitating an automated forward modeling approach and making it straightforward to incorporate prior information. Clowes et al. (1995) used the Zelt and Smith (1992) algorithm to simultaneously invert refraction and reflection times from a network of 2D profiles such that the model parameters at the intersection points were linked to ensure consistency. This approach can be used to infer 3D structure from 2D data, or develop a starting model for full 3D inversion if offline data is available in addition to inline data (e.g., Zelt et al. 1999).

Hole (1992) developed a 3D first-arrival tomography algorithm using the Vidale (1990) forward modeling scheme and back projection to solve for a smooth velocity field. Hammer et al. (1994) developed a 3D first-arrival tomography algorithm specially suited to sparse data using a spectral, continuous function model parameterization. Toomey et al. (1994) developed a 3D first-arrival tomography algorithm to obtain a smooth velocity field in which the velocity grid is “draped” from an irregular bathymetric surface by vertically shearing the columns of nodes to accurately model seafloor relief. Vertically sheared grids are particularly important for marine data, especially at mid-ocean ridges, where a large velocity discontinuity may exist. Other algorithms with sheared grids have been developed by Van Avendonk et al. (1998) and Korenaga et al. (2000), although they are limited to 2D models, but include reflections. The main difference between these two algorithms is the former uses first arrivals that can turn above or below a reflecting interface at which there is no velocity discontinuity, whereas the latter only uses first arrivals that turn above a reflecting interface.

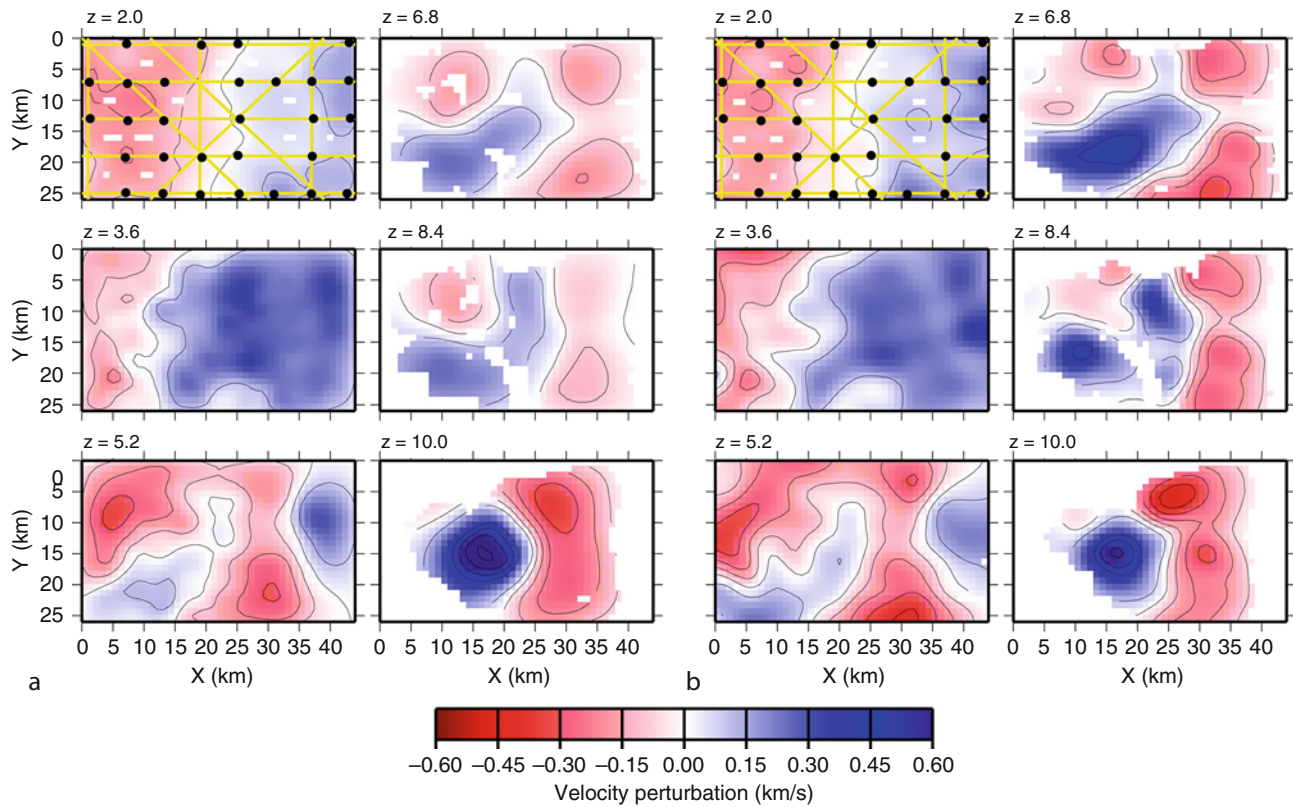
Lailly and Sinoquet (1996) developed a 2D algorithm for inverting reflection times to estimate a smooth velocity model for pre-stack depth migration of reflection data using the wavefront tracking method of Podvin and Lecomte (1991) for the forward calculation (Fig. 2). Zelt and Barton (1998) developed a 3D first-arrival tomography algorithm using smoothing regularization and compared the results with those from the back-projection method of Hole (1992) (Fig. 3). The results showed that for the same level of fit to the data, regularized inversion can provide a simpler model. Zelt et al. (2006b) applied 3D simultaneous refraction and reflection tomography to solve for smooth velocities and multiple interfaces using a sparse crustal-scale 3D dataset (Fig. 4). McCaughey and Singh (1997) and Hobro et al. (2003) developed related 2D and 3D simultaneous refraction and reflection tomography algorithms that solve for smooth velocities and interfaces with the allowance for discontinuities across the layer boundaries (Fig. 5).

Most of the algorithms described above are general purpose in nature and follow more or less from the objective function presented in the previous section. However, more specialized algorithms have been developed and a few are described here to give a sense of what is possible. Hole et al. (1992) developed an algorithm for determining a 3D interface using first arrivals with known velocities above and below the interface. Zhang et al. (1998) inverted traveltime curves instead of points, specifically average slowness and apparent slowness. They claim this balances the contribution from short and long rays, and enhances resolution and convergence speed. Rawlinson and Sambridge (2003b) developed an algorithm for the inversion of refraction and reflection times using



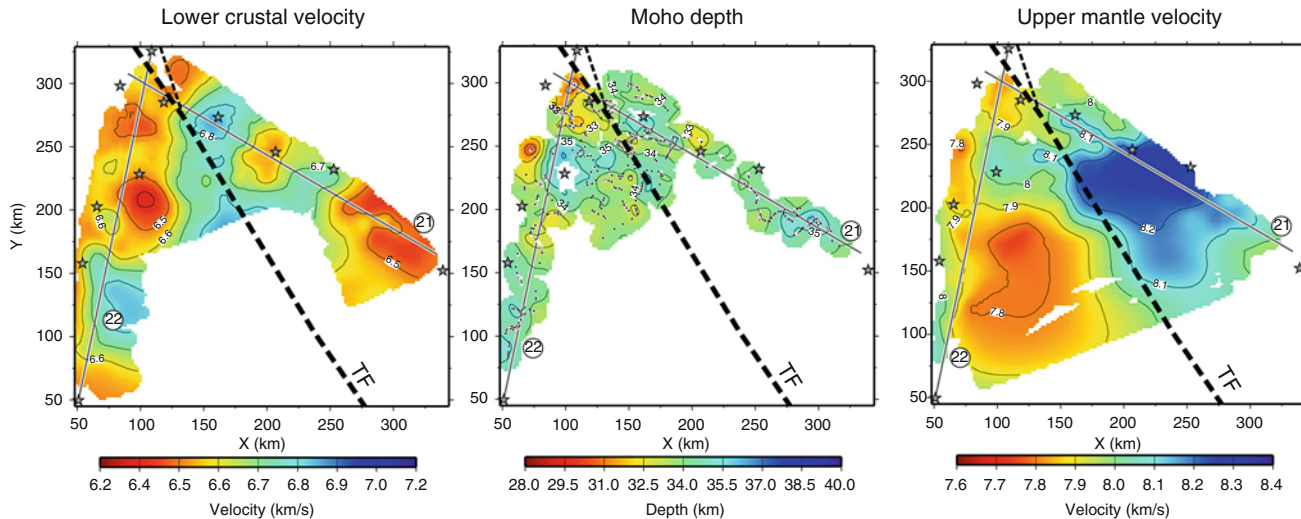
Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 2 Synthetic example of reflection tomography for a smooth velocity model by Lailly and Sinoquet (1996) based on real data from a passive margin with a salt structure and a listric fault. (a) True model. The reflectors (white curves) are embedded in the smooth velocity field

and are allowed to cross each other. (b) Estimated model using reflections from the continuous portions of the reflectors (white curves); the parts of the reflectors represented by dashed lines are not illuminated by rays



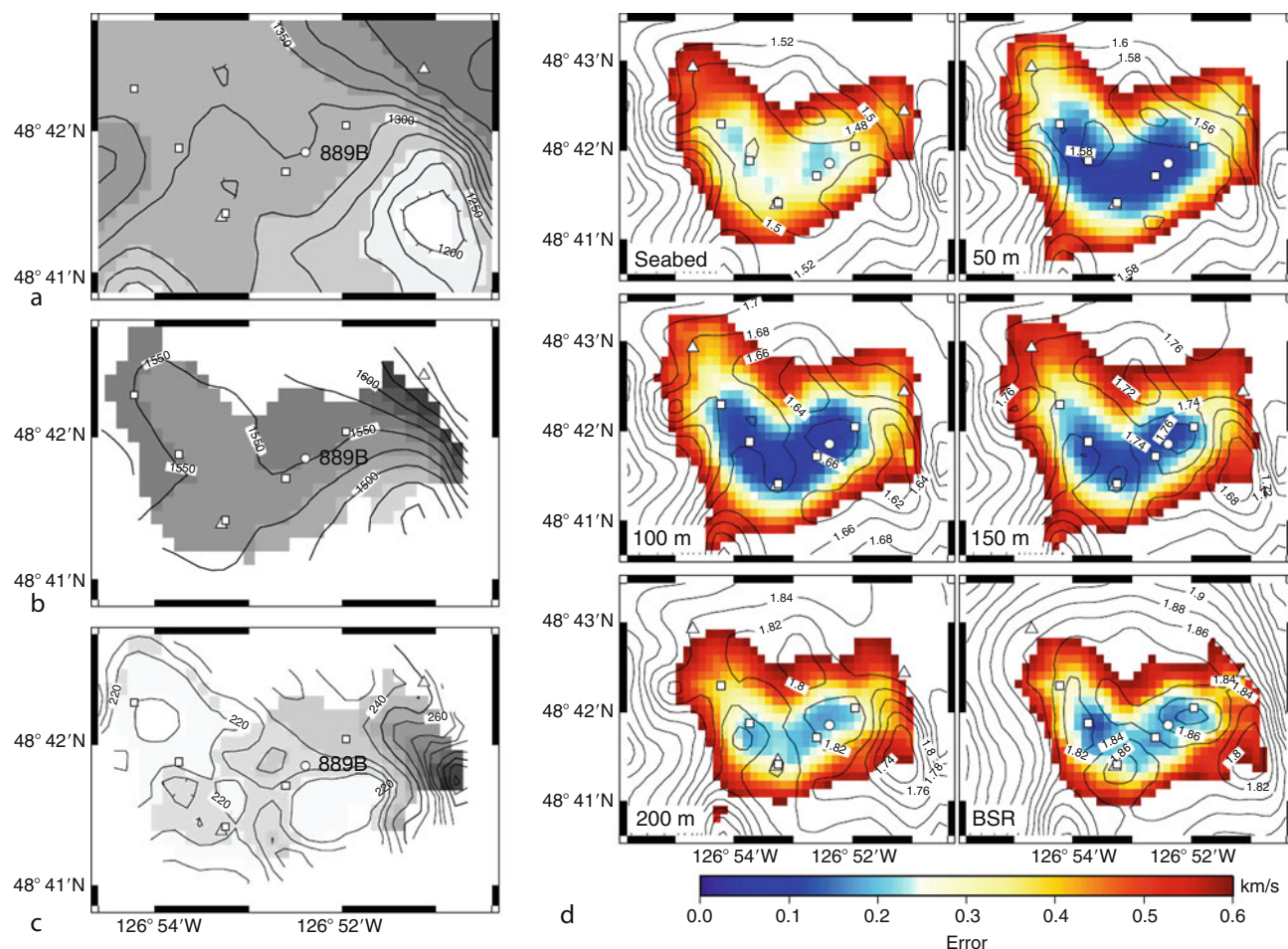
Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 3 Example of 3D first-arrival tomography applied to ocean-bottom-seismometer (OBS) data from the Faeroe Basin to find the minimum-structure model (Zelt and Barton 1998). (a) Model obtained using smoothing-regularized tomography. (b) Model obtained using back-projection tomography method of Hole (1992). Horizontal slices of velocity perturbations with respect to a 1D starting model are shown at

the depths labeled from 2 to 10 km. The OBS locations and shot lines are overlaid on the $z = 2$ km slice. Regions not sampled by rays are blank. Contour interval is 0.1 km/s. Both models provide the same level of fit to the data, but the model from back-projection contains more structure overall. The high-velocity anomaly at $z = 10$ km likely indicates a basement high



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 4 Example of simultaneous 3D first-arrival and reflection tomography of sparse crustal data across the Tintina strike-slip fault in the northern Canadian Cordillera (Zelt et al. 2006b). Left panel shows the lower crustal velocity averaged vertically between 1 and 15 km above the Moho. Center panel shows depth to Moho; dots show reflection

points. Right panel shows the upper mantle velocity averaged vertically between 1 and 15 km below the Moho. Stars are shot locations. The straight lines labeled 21 and 22 indicated the best-fit lines of the crooked profiles containing the receivers locations. Regions not sampled by rays are blank. Dashed line is the surface location of the fault. There is a 0.3–0.4 km/s change in upper mantle velocity across the fault



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 5 Example of simultaneous 3D refraction and reflection tomography of OBS data and coincident single channel reflection profiles over a gas hydrate stability zone offshore Vancouver Island (Hobro et al. 2005). (a) Seafloor depth. (b) Bottom simulating reflector (BSR) depth. (c) BSR depth below the seafloor. (d) Slices through the velocity model at the

seafloor, at 50-m intervals through the hydrate stability zone and immediately above the BSR. Labeled contours indicate velocities (km/s) and color scale marks formal uncertainty estimates. Ten OBS positions and ODP site 889 are marked. Refracted arrivals from above and below the BSR were inverted along with wide-angle and normal-incidence reflections from the BSR

a 3D multi-layered model parameterization. Interfaces are defined by a non-uniform node distribution and velocities vary linearly with depth so rays are calculated analytically as piece-wise circular arcs using a shooting method. Trinks et al. (2005) presented a method for simultaneous refraction and reflection inversion using a 2D layered model parameterization which adapts to non-uniform ray coverage such that the cell size is inversely proportional to the local ray density.

Refraction/wide-angle reflection traveltimes can be jointly inverted with coincident near-vertical reflection times in several ways (Zelt 1999). Arrivals in pre-stack reflection data can be picked and inverted jointly with the wide-angle data, but it is often difficult to pick weak events in pre-stack data. The most common approach is to pick reflections from a stacked reflection section and invert these data simultaneously with the wide-angle data by modeling zero-offset reflections (e.g., McCaughey and Singh 1997; Zelt et al. 2003). One potential

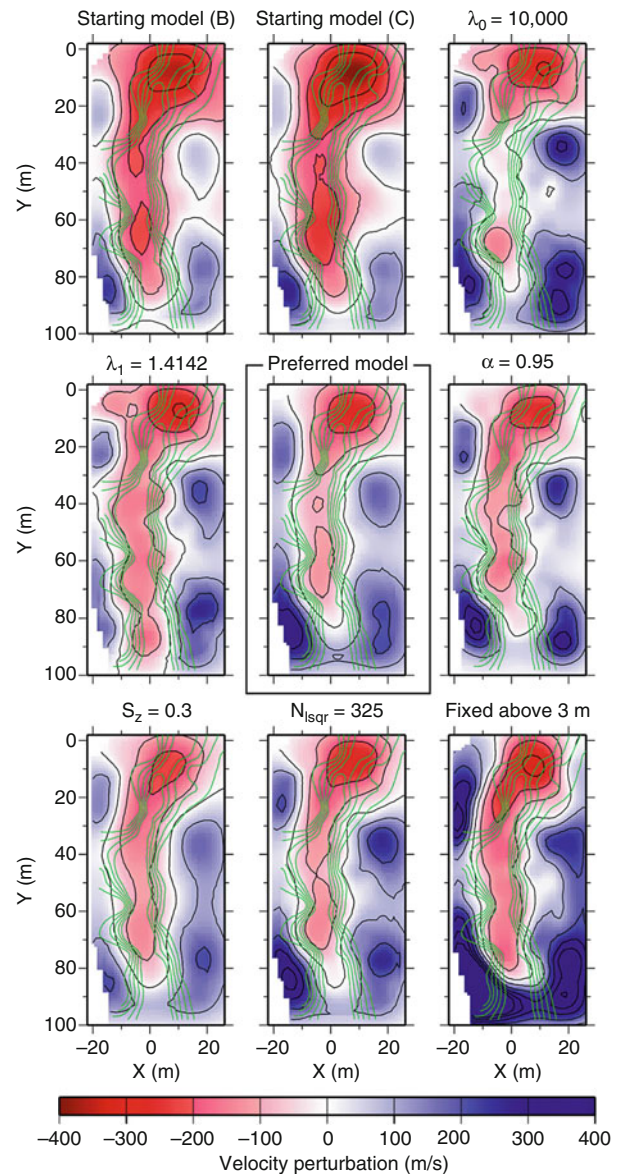
pitfall when using coincident reflection data is incorrectly correlating the near-vertical reflection event with the corresponding wide-angle event or layer boundary; Jaiswal et al. (2006) present a way to avoid this problem. As an alternative to joint inversion, the reflector geometries and interval velocities from stacked data can be used as prior information in the regularization to constrain a layered velocity model (e.g., Bosch et al. 2005).

The close relationship between seismic velocity and density (Brocher 2005) lends gravity data to a joint inversion with traveltime data. Nielsen and Jacobsen (2000) used the Zelt and Smith (1992) algorithm as the basis for a simultaneous inversion of refraction and reflection times with gravity data to derive a layered 2D crustal model. Korenaga et al. (2001) carried out a joint inversion of traveltime and gravity data in which error propagation from the velocity model to the predicted gravity anomalies was taken into account.

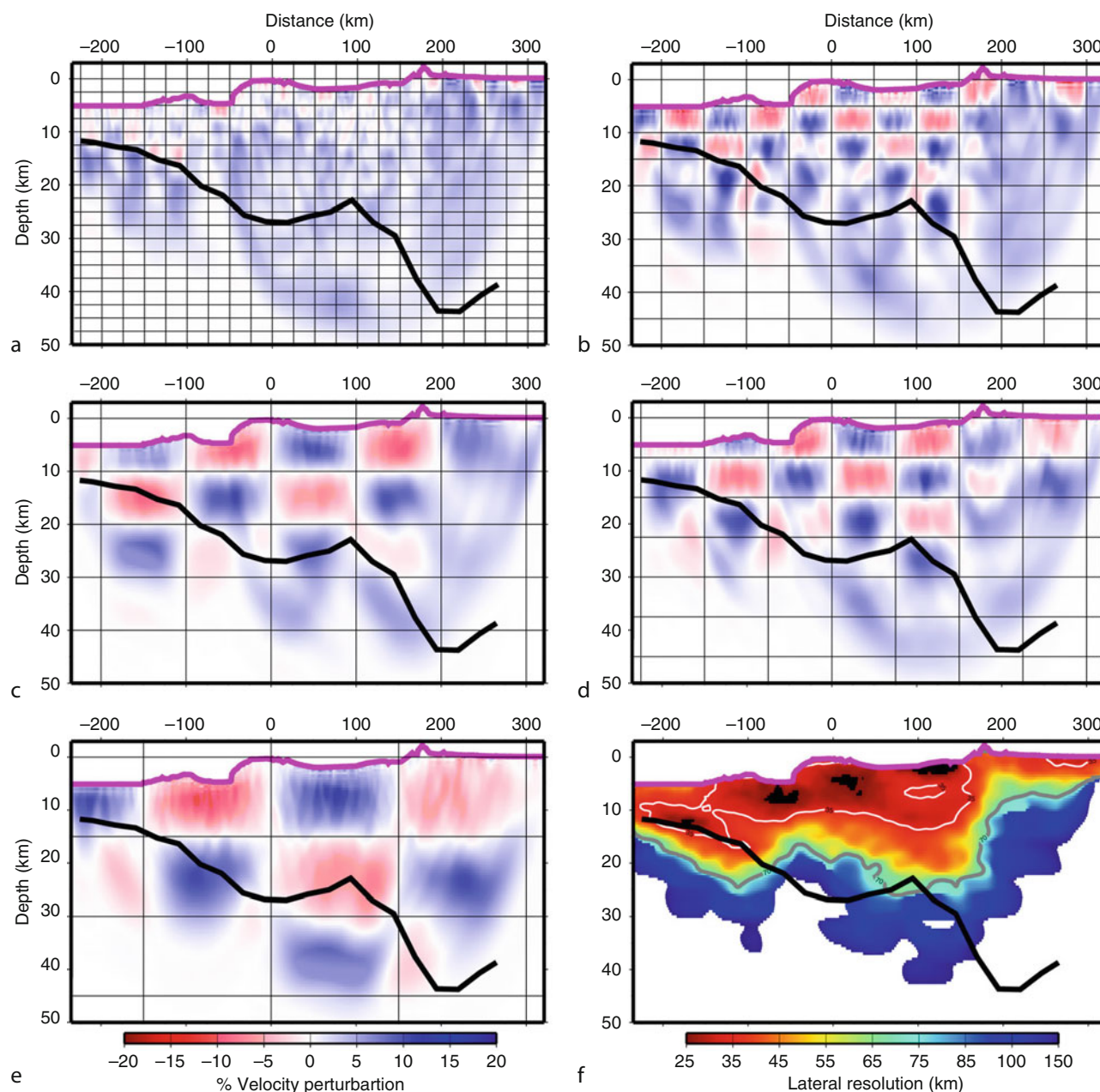
Model Assessment

Once a model to explain a particular set of data is developed, model assessment should be used to explore the non-uniqueness, resolution, and errors associated with the model. This is an attempt to quantify the robustness of the model, a process that is not straightforward for nonlinear inverse problems, and different assessment techniques will be appropriate depending on the data, model parameterization, geologic target, and objectives of the experiment. There are two classes of assessment methods: linear and nonlinear. The former methods assume it is valid to consider only a local neighborhood of the final model within which the model-data relationship is linear. The latter methods involve additional nonlinear inversions of the real data or synthetic data using the true source-receiver geometry of the experiment. Linear methods are quick and easy to apply, although potentially less reliable depending on the nonlinearity of the problem. Nonlinear methods are computational intensive, often more so than what was needed to derive the final model, and may require a specialized inversion algorithm. However, nonlinear methods are the only way to properly estimate the final model's robustness. Zelt (1999) and Rawlinson and Sambridge (2003a) describe many of the assessment techniques in use today, while Zelt et al. (2006a) apply several different techniques to the same 3D dataset (Fig. 6).

The simplest form of model assessment is to examine the ray coverage in the model using ray plots or "hit counts" (the number of rays sampling each model parameter). These can be misleading since a limited distribution of ray angles may provide less independent model constraint than expected from the number of rays alone. For example, rays in the uppermost mantle will tend to be sub-horizontal, and therefore capable of providing little lateral resolution. Toomey et al. (1994) used a measure called derivative weight sum (DWS) that equals the sum of the partial derivatives for each model parameter. This is somewhat more meaningful than hit count since it weights the contribution from each ray according to the parameter's sensitivity to that ray. However, it also fails to account for the degree of independence within the ray set. Other common linear assessment measures include an examination of the diagonals of the posterior resolution and covariance matrices (e.g., Zelt and Smith 1992). These are more precise measures since they account for the independence of the ray set, but they are best used in a relative as opposed to absolute sense, since they do not account for the nonlinearity of the problem. Rows of the resolution matrix, known as resolution kernels, provide a spatial sense of the averaging of the true structure by the model (Zelt 1999). Probably the most precise linear assessment method uses singular value decomposition (SVD) since it is able to quantify the nature and degree of model constraint through construction of an orthogonal set of basis vectors that



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 6 Example of model assessment for near-surface 3D first-arrival tomography at a groundwater contamination site (Zelt et al. 2006a) using the Zelt and Barton (1998) regularized inversion algorithm seeking a minimum-structure model. Target of the surface was a low-velocity paleochannel cut into a clay layer running roughly north-south through the center of the survey area. The depth to the clay layer was known from extensive well data in the area; depth-to-clay contours (green) from 7 to 11 m overlay the depth slices at 10 m for this comparison of nine different models. Models displayed as perturbations relative to a 1D starting model; the starting velocity at this depth is 1150 ms/s. The preferred final model is in the center. The other models were obtained by trying two different starting models, different values for five free parameters that control the inversion, and by fixing the model above 3 m at the starting model values. Overall, the preferred model has minimum structure; the model in the lower left appears simpler in this depth slice, but it is rougher vertically



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 7 Example of 2D first-arrival checkerboard test for crustal data from an onshore-offshore experiment across the South American-Caribbean plate boundary at 67°W (Magnani et al. 2009). There were shots and receivers both onshore and offshore. (a–e) show recovered checkerboard anomalies for sizes from 25×2.5 km to 150×15 km. (f)

Estimated lateral resolution of the velocity model using a 2D version of the method presented by Zelt (1998). Regions with better than 25 km resolution are black; regions with worse than 150 km resolution are white. Pink line indicates the land surface and bathymetry. The coast line is at ~160 km. Black line indicates the Moho from the preferred final model

span model space, each with a specified weight in the model reconstruction (e.g., White 1989; Scales et al. 1990).

A nonlinear assessment of the spatial resolution or uncertainty of a single model parameter, or set of parameters, is possible in which the real data and synthetic data are inverted in the same way that the final model was derived, allowing the full nonlinearity of the problem to be accounted for, including trade-offs between model parameters (Zelt and White 1995;

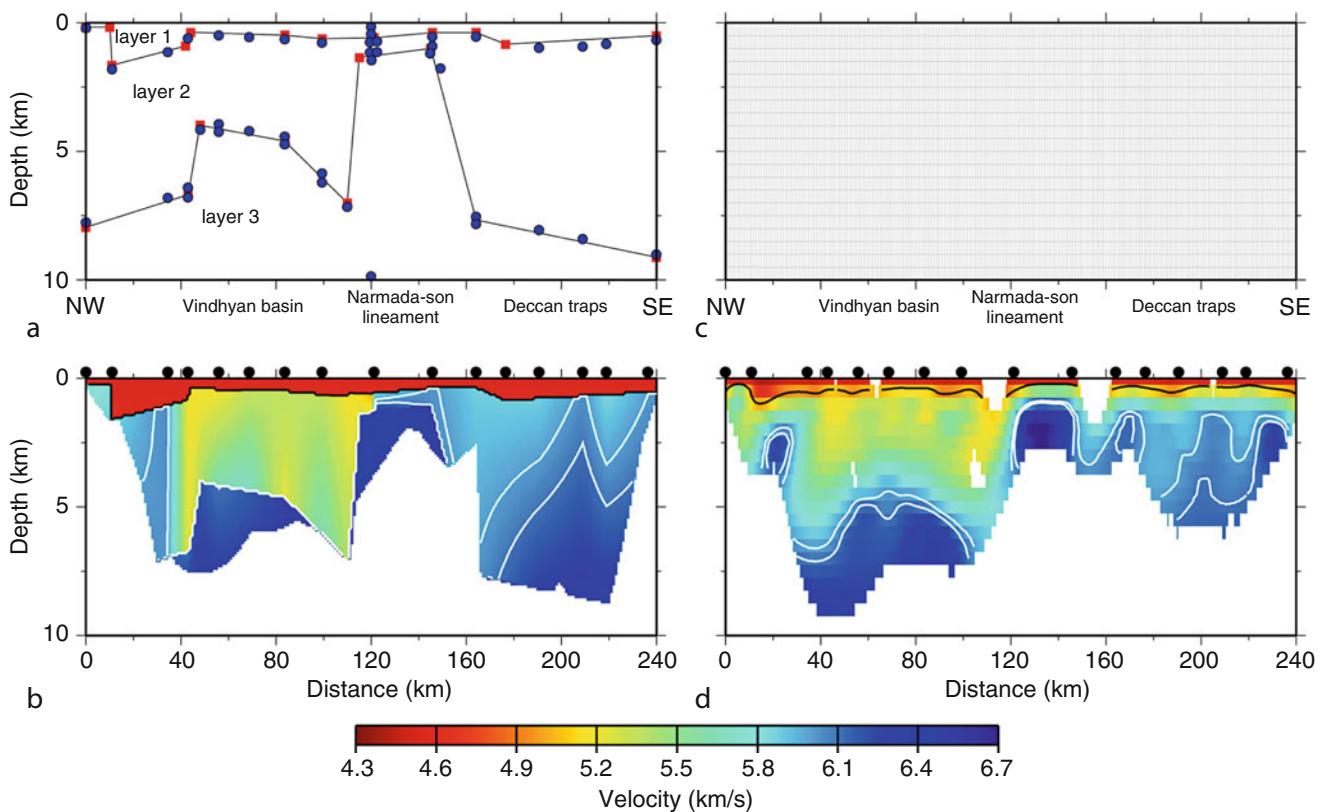
Zelt 1999; Christeson et al. 1999). For spatial resolution, the value of a model parameter is perturbed enough to yield a significant traveltime anomaly with respect to the pick uncertainties. Rays are traced through the perturbed model to calculate a set of perturbed traveltimes. The perturbed data are then inverted using the final model as the starting model. The spatial resolution about the selected parameter will be indicated by the amount that the values of adjacent parameters

differ from their corresponding value in the original final model. If the model is poorly resolved about the selected parameter, then the parameter's perturbation will be smeared into adjacent parameters, perhaps both velocities and interface depths if both parameter types are involved, and the extent of the smearing indicates the spatial resolution. Both positive and negative parameter perturbations should be tested, and it will likely be sufficient to examine only one or two representative velocity and interface parameters for each layer or region of the model.

A parameter-selective algorithm such as the Zelt and Smith (1992) approach allows for a nonlinear estimate of a single parameter's absolute uncertainty (Zelt 1999). The value of the model parameter is slightly perturbed from its value in the final model and held fixed while inverting the observed data involving all other model parameters. The size of the perturbation is increased until the recovered model is unable to fit the observed data as well as the preferred final model based on

an F test comparing the traveltime fits of the preferred and perturbed final models. The maximum parameter perturbation that allows a comparable fit to the observed data is an estimate of its absolute uncertainty. Again, both positive and negative parameter perturbations should be tested, and it will likely be sufficient to examine only one or two representative velocity and interface nodes for each layer or region of the model. A fine grid tomographic approach can also be used to perform tests like these. For example, Zelt and Barton (1998) examined one region of a model where there was significant lateral structure even though the constraint from the ray coverage was known to be low in this area. They added regularization to the inversion to force that region to remain laterally homogeneous and thereby establish the required trade-offs elsewhere in the model.

One of the most common forms of nonlinear model assessment to estimate spatial model resolution is the checkerboard test (e.g., Magnani et al. 2009; Fig. 7). In these tests an



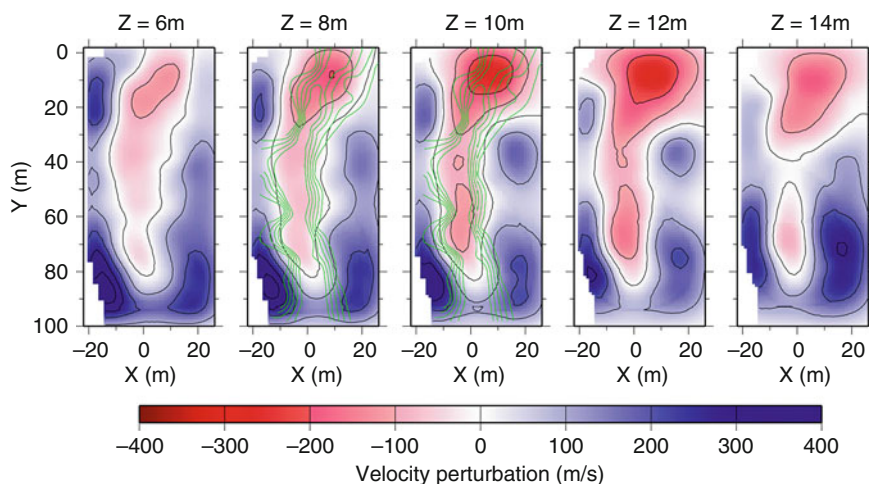
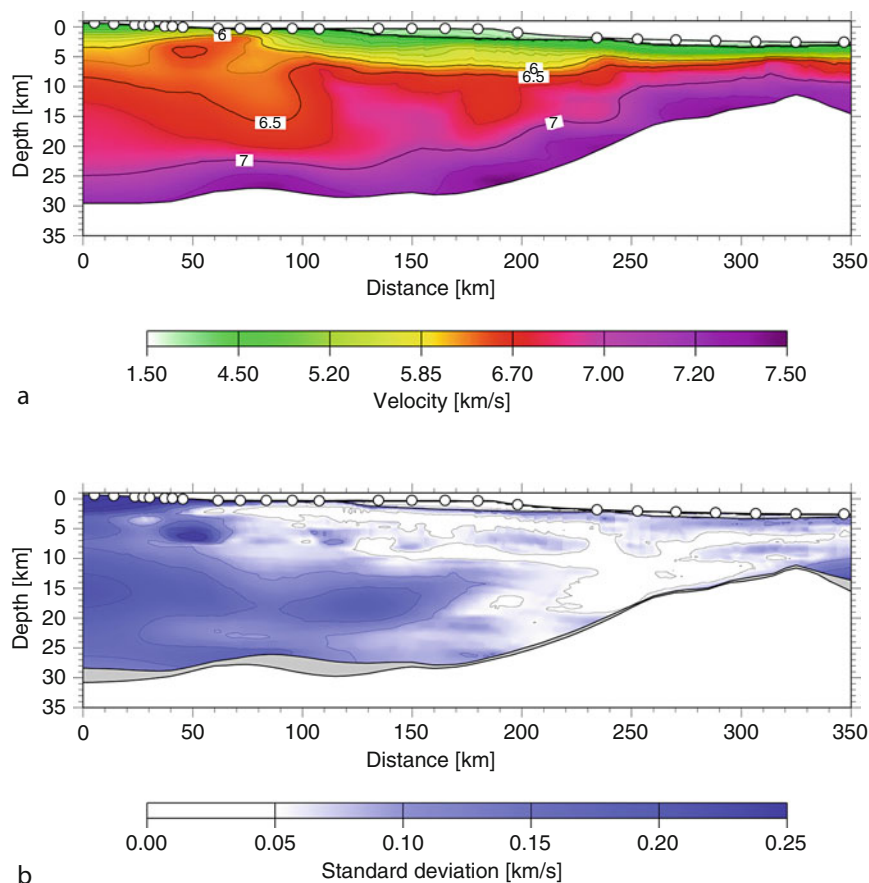
Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 8 Example from central India across the Narmada-Son lineament of applying two different traveltime inversion algorithms to the same refraction/wide-angle reflection data (Zelt et al. 2003) to compare the “geological” model with the minimum-structure model. (a) Layered model parameterization using the Zelt and Smith (1992) algorithm; there are 43 velocity nodes (blue dots) and 21 interface nodes (red squares). (b) Velocity model corresponding to parameterization in (a); both first arrivals and reflections from the top of the third layer were used.

(c) Fine-grid model parameterization using the Zelt and Barton (1998) algorithm to seek a smooth minimum-structure model; there are 7018 slowness cells. (d) Velocity model corresponding to parameterization in (c); only first arrivals were used. Only those portions of the models sampled by rays are shown. Isovelocity contours of 5.0 km/s (black) and 6.0 and 6.1 km/s (white) indicated. Shot point locations (black dots) and geologic features labeled above the models. An interpretation of the main features of the geological model in (b) is supported by the minimum-structure model in (d) since it contains the same features

Traveltime Tomography Using Controlled-Source Seismic Data

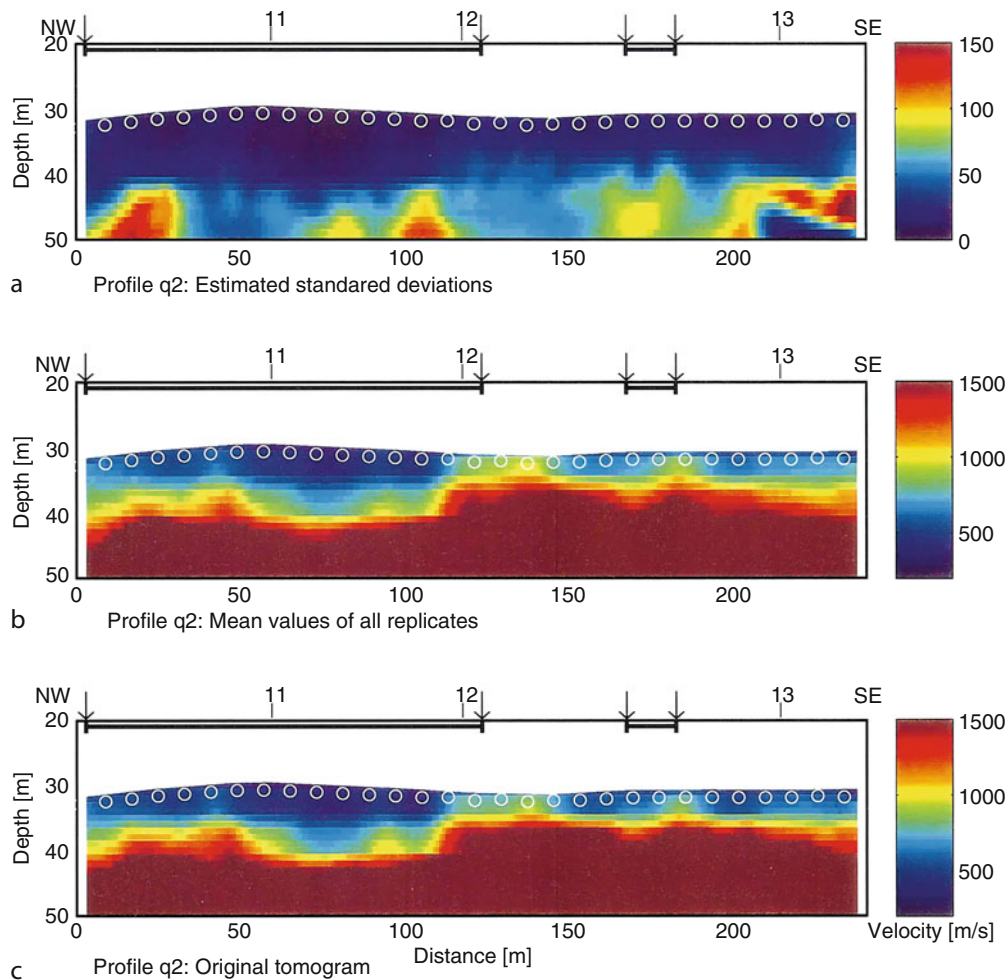
Seismic Data, Fig. 9 Example of model assessment using a nonlinear Monte Carlo procedure to estimate model variance (Korenaga et al. 2000). (a)

Velocity model for the southeast Greenland continental margin obtained by joint traveltime inversion of first arrivals and wide-angle Moho reflections recorded at 18 OBS's and eight land stations (white circles). This model was obtained by averaging 100 Monte Carlo ensembles. (b) Corresponding standard deviation for velocity and depth nodes; the standard deviation of the Moho is indicated by the width of the grey zone at the base of the crust



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 10 Example of near-surface 3D first-arrival tomography at a groundwater contamination site (Zelt et al. 2006a) using the Zelt and Barton (1998) regularized inversion seeking a minimum-structure model. The Horizontal depth slices from 6 to 14 m displayed as perturbations relative to a 1D starting model. Black contour interval is 100 m/s. Target of the surface was a low-velocity paleochannel cut into a clay

layer running roughly north-south through the center of the survey area. The depth to the clay layer was known from extensive well data in the area; depth-to-clay contours (green) from 7 to 11 m overlay the 8 and 10 m depth slices. This experiment consisted of ~600 shots recorded by ~600 receivers yielding ~360,000 traces, providing 187,877 useable picks for the inversion



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 11 Example of near-surface first-arrival tomography over a buried waste disposal site seeking a smooth model (Lanz et al. 1998). (a) Estimated standard deviation of velocities computed using

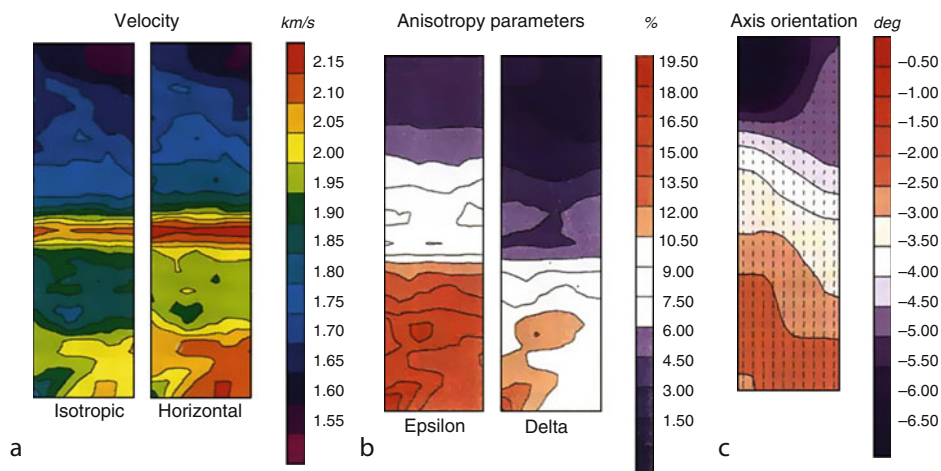
200 randomized datasets. (b) Mean velocity model from 200 randomized datasets. (c) Preferred final model from inversion of original picked data. Open circles are shot locations. Arrows indicate landfill borders defined by several other geophysical techniques

alternating pattern of high and low anomalies is superimposed on the starting model from the inversion of the real data. Synthetic data are calculated for the “checkerboard” model, and then inverted using the same starting model and source-receiver geometry as the real data. The recovered model will closely resemble the checkerboard pattern in regions of good constraint, but will otherwise not resemble the checkerboard model. The resolution at different length scales can be estimated by testing anomaly patterns of different sizes. Anomaly patterns with different polarity, registration, and orientation should be tested to average out the effects of changing ray coverage due to the nonlinearity of the problem (Zelt 1998).

Other nonlinear assessment techniques include trying different starting models, different values of the free parameters in the objective function, and exclusion of subsets of the data considered less reliable (e.g., Zelt et al. 2006a; Fig. 6). Using

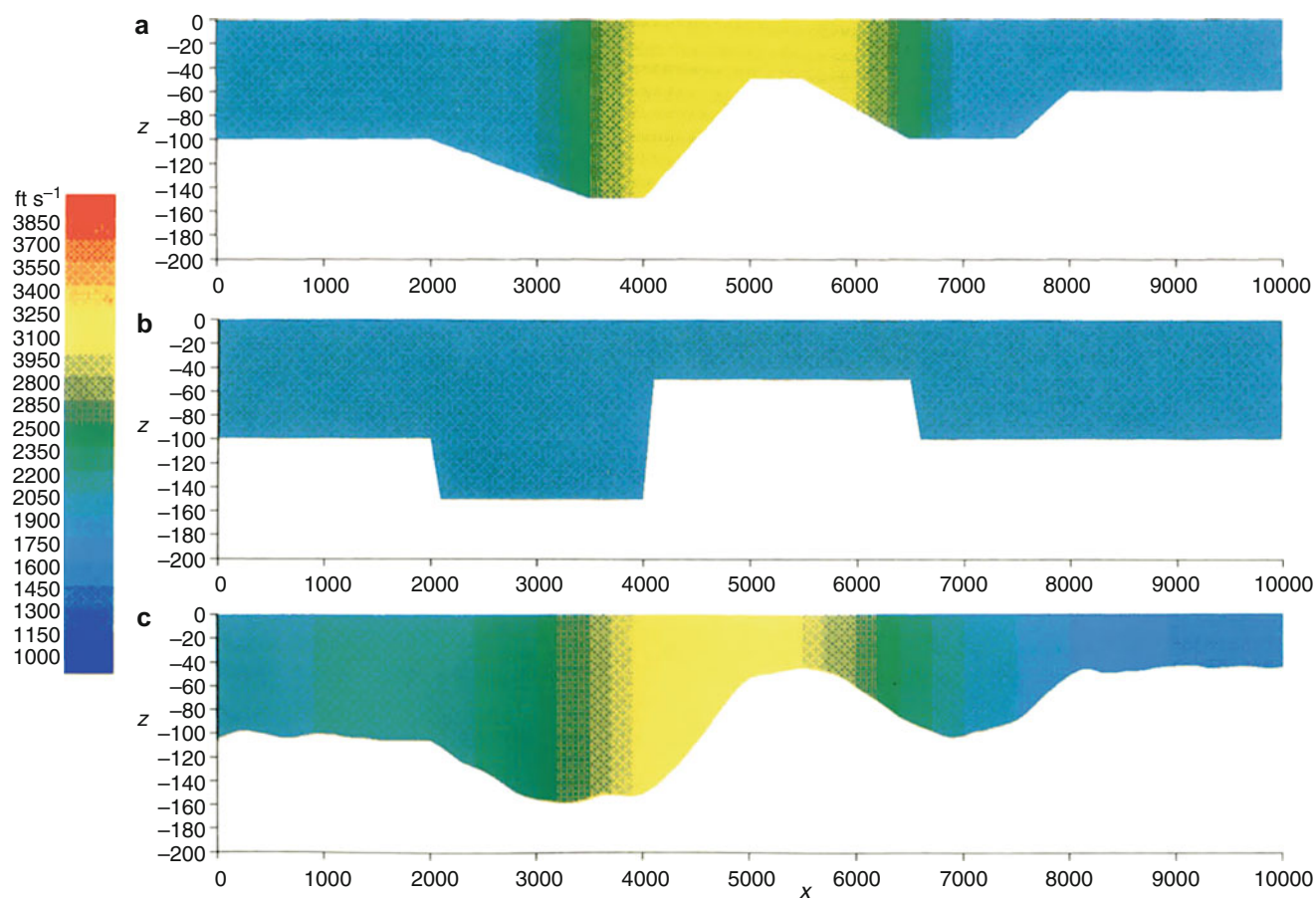
different model parameterizations and different inversion algorithms can be effective, especially when one inversion algorithm seeks a model that satisfies all notions of what is geological reasonable, and one algorithm seeks the minimum-structure model (Fig. 8; Zelt et al. 2003). In this way, it is possible to determine what model structure is consistent with the data to facilitate hypothesis testing, while at the same time establishing what model structure is required by the data.

Perhaps the most complete and computationally intensive form of model assessment involves a nonlinear Monte Carlo procedure. Many inversions are performed in which randomized data and/or randomized starting models are used. The resulting models are used to compute posterior model covariance and resolution estimates. Zhang et al. (1998) and Korenaga et al. (2000) applied this approach to 2D inversions of first arrivals and reflections (Fig. 9).



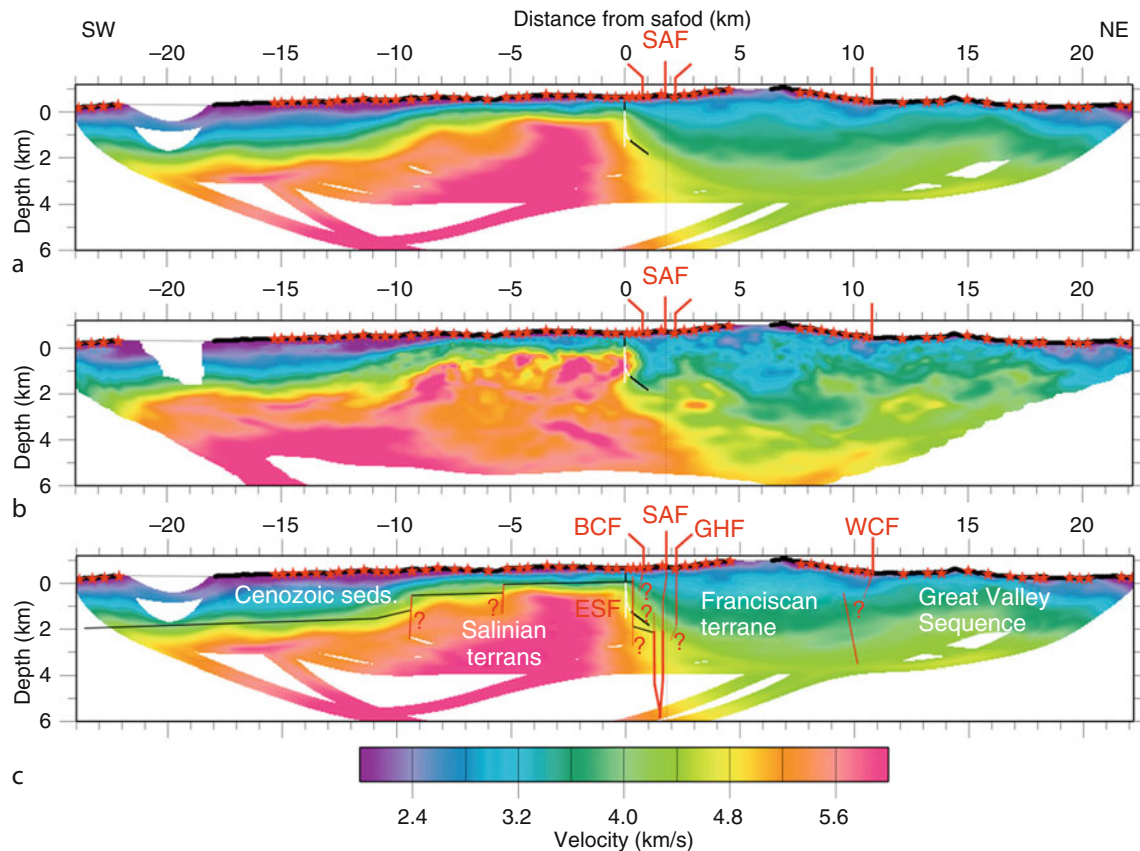
Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 12 Example of crosswell tomography in which transverse isotropy is assumed and regularization was used to seek a smooth model (Pratt and Chapman 1992). The velocity and anisotropic parameters were solved for simultaneously. (a) Comparison of isotropic velocity model

and horizontal velocity model from anisotropic inversion. The model is 20 m across and 60 m deep. (b) Anisotropy parameters. The amount of anisotropy generally increases with depth, reaching nearly 20%. (c) Symmetry axis orientation. The length of the markers is proportional to one of the anisotropy parameters (epsilon)



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 13 Example of imaging the near-surface weathering layer (i.e. the refraction statics problem) using smoothing regularization (Scales et al. 1990). Shots were placed at either end of the model and there were 100 receivers between them. (a) True model. Refractor velocity is

5000 ft/s. (b) Starting model with blocky thickness variations in the weathering layer. Refractor velocity is 4000 ft/s. (c) Final model. Refractor velocity is 5001 ft/s. The inversion solved simultaneously for a laterally-varying weathering-layer velocity and refractor, and a constant refractor velocity



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 14 Example of first-arrival tomography across the San Andreas Fault comparing models from two different algorithms (Hole et al. 2006). There were 63 shots (red stars) and 912 receivers (Thick black line). (a) Velocity model from application of the Hole (1992) back-projection method. (b) Velocity model from application of Zelt and Barton (1998) smoothing-regularized method. Both models provide roughly the same level of fit to the picks. Regions without ray coverage

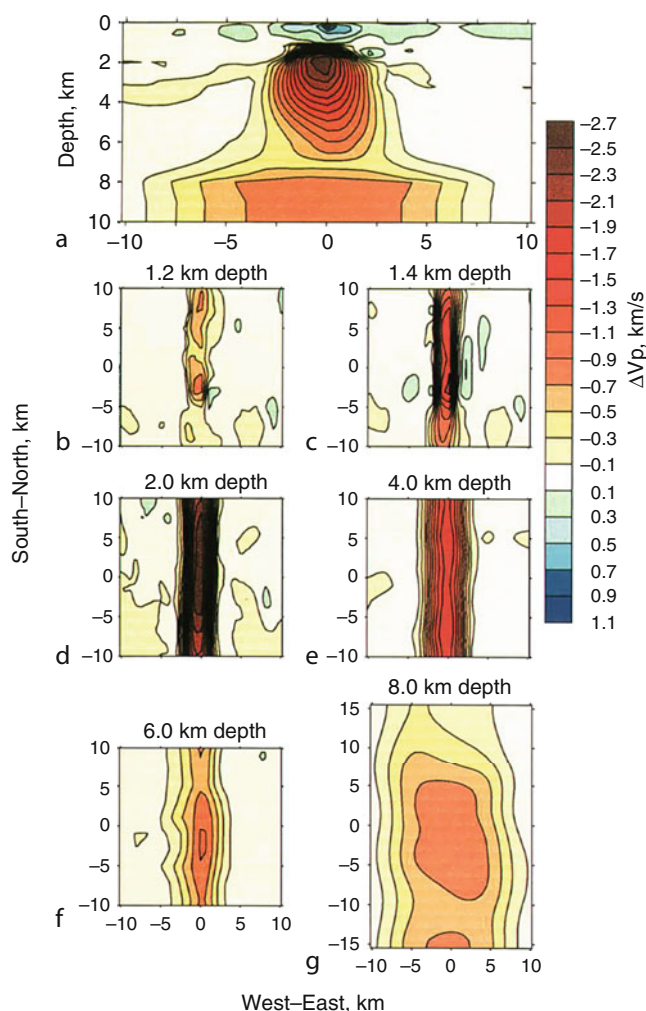
are white. Well-log observations at SAFOD (San Andreas Fault Observatory at Depth) are white for granite and black for sedimentary rock. Note how high velocities (>5 km/s) better match the known position of granite in the well for the model in (b) compared to the model in (a), suggesting that some of the additional structure in the model in (b) may be true. (c) Geology and fault interpretation overlaid on the model in (a). Black lines are top of basement; red lines are faults

More Examples

The examples in Figs. 1, 2, 3, 4, 5, 6, 7, 8, and 9 demonstrate the types of models that can result from a wide range of controlled-source experiments. The remaining examples presented in this section are intended to round out the range of applications to which traveltime tomography and inversion are applied in controlled-source studies, and the types of model structure that can result. The main purpose of the examples is to show what is possible in terms of tomographic imaging of different kinds of geologic features in the near-surface, upper crust, lower crust, and uppermost mantle using traveltime data. All examples shown in this article use only P-wave traveltime data.

Zelt et al. (2006a) applied the regularized 3D first-arrival tomography algorithm of Zelt and Barton (1998) to a dense

3D dataset from a groundwater contamination site (Fig. 10). Lanz et al. (1998) applied 2D first-arrival tomography and a Monte Carlo scheme to estimate a smooth model with standard deviations for a buried waste disposal site (Fig. 11). Pratt and Chapman (1992) applied 2D first-arrival tomography to crosswell data to simultaneously determine velocity and anisotropy parameters (Fig. 12). Scales et al. (1990) imaged a nearsurface weathering layer using 2D first-arrival times, solving simultaneously for lateral variations in the velocity and thickness of the layer (Fig. 13). Hole et al. (2006) applied 2D first-arrival tomography to data recorded across the San Andreas Fault using two different algorithms (Fig. 14). Dunn et al. (2000) applied 3D refraction and reflection tomography incorporating anisotropy to data from the East Pacific Rise (Fig. 15). Van Avendonk et al. (1998) inverted refraction and reflection times for a smooth 2D velocity model and an



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 15 Example of 3D refraction and reflection tomography across the East Pacific Rise at 9°30'N for crust and uppermost mantle structure (Dunn et al. 2000). There were 15 OBS's and 480 explosive sources. Refracted arrivals from the crust and upper mantle and reflections from the Moho were simultaneously inverted. Anisotropy of 4% in the crust and 7% in the upper mantle was included in the modeling. Final velocity model as perturbations relative to a 1D model are shown and are contoured at 0.2 km/s intervals. (a) vertical section through the center of the model. (b–g) Depth slices from 1.2–8.0 km beneath the seafloor. The low-velocity body is interpreted to represent melt distribution within a magmatic system

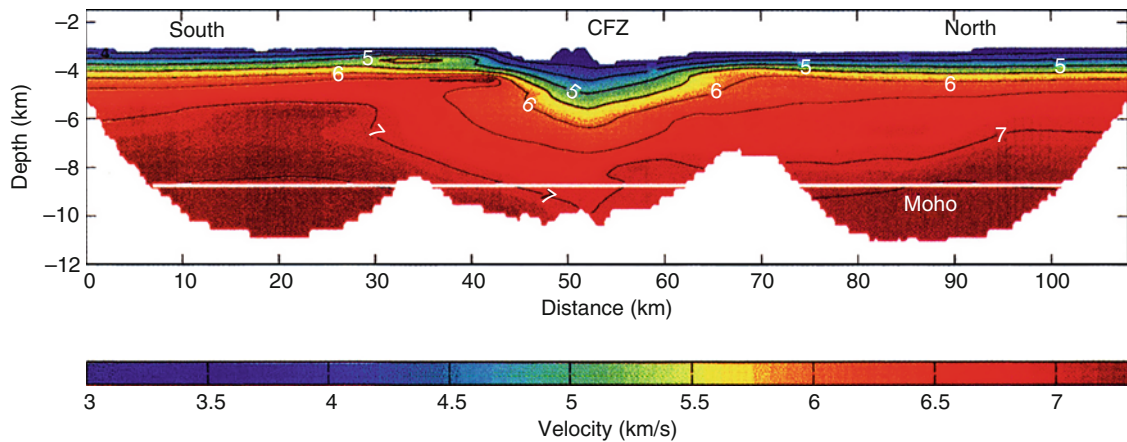
assumed flat Moho across the Clipperton transform fault (Fig. 16). Clark et al. (2008) inverted refractions and reflections from an onshore-offshore survey across the South American-Caribbean plate boundary using a smooth tomography approach for the upper crust, and a layer-based inversion for the lower crust, Moho, and upper mantle (Fig. 17). Zelt et al. (2003) compared layer-based simultaneous refraction and reflection inversion with independent smooth refraction and reflection tomography for three different 2D crustal datasets (Fig. 18).

Future

Full 3D seismic experiments are likely to become more commonplace, necessitating the need for some of the popular 2D tomography and inversion algorithms to be extended to 3D. It is likely that three-component, S-wave, and converted-wave studies will become more widespread, especially for near-surface environmental and engineering studies, necessitating the need for algorithms that can perform coupled P-wave, S-wave, or Poisson's ratio inversions. The desire to incorporate or determine anisotropy will likely become more commonplace in the future, and there are currently only a few traveltime algorithms that include any form of anisotropy (e.g., Dunn et al. 2000).

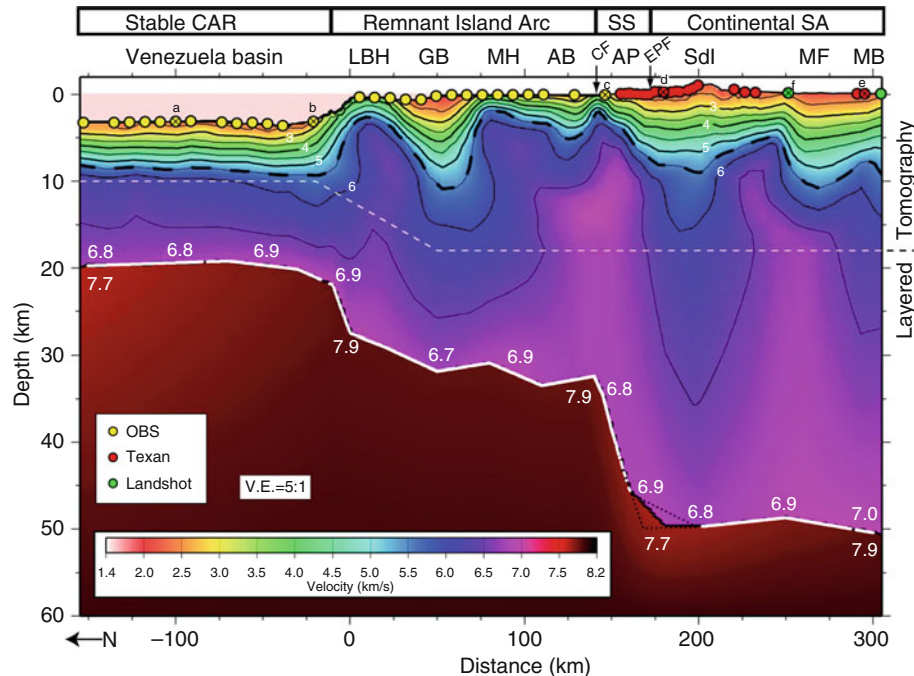
For all scales of seismic experiments, the need for flexible, general-purpose traveltime tomography and inversion algorithms will remain, both as the primary modeling tool and to provide long-wavelength starting models for higher-resolution full waveform inversion techniques (e.g., Brenders and Pratt 2007). The Zelt and Smith (1992) traveltime inversion algorithm has been widely used for many years for several reasons: (1) an irregular, layered model parameterization adaptable to the data coverage and geologic structures, (2) parameter-selective inversion, (3) any type of body-wave data can be modeled, and (4) it is easy to incorporate prior information. As such it stands apart from most other popular algorithms that use a uniform, fine grid parameterization, and consider only first arrivals, and in some cases reflections. However, the Zelt and Smith (1992) algorithm was primarily intended for relatively sparse 2D data. There will likely be a niche for a new version of the Zelt and Smith algorithm if it maintains all of its current advantages, but is extended to 3D and includes regularization to handle dense data, namely second-order derivatives to seek a smooth, minimum-structure model.

About 10 years ago, a theory for traveltime tomography was developed in global seismology to take into account the finite-frequency nature of seismic data (Dahlen et al. 2000). This method is known as finite-frequency traveltime tomography and it yields what are popularly known as banana-doughnut sensitivity kernels. By taking finite-frequency effects into account, this form of traveltime tomography should, theoretically, yield a more accurate estimation of velocity anomalies in terms of their magnitude and spatial resolution because the physics of wave propagation is treated more accurately. The theory developed for global seismology uses as input data delay times relative to a reference Earth model, and as such the theory is generally not applicable to controlled-source data. This is because for controlled-source data there is no requisite reference velocity model known in advance that is capable of yielding realistic synthetic waveforms that are close enough to the recorded seismograms to yield a meaningful delay time through cross correlation of the



Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 16 Example of 2D tomography across the Clipperton transform fault (Van Avendonk et al. 1998). There were only five OBS's. Crust and upper mantle refracted arrivals and Moho reflections were inverted

simultaneously for a smooth velocity model. A flat Moho at 8.6 km depth was assumed. The model shows reduced velocities in the upper crust as far as 10 km south and 18 km north of the transform fault (CFZ)

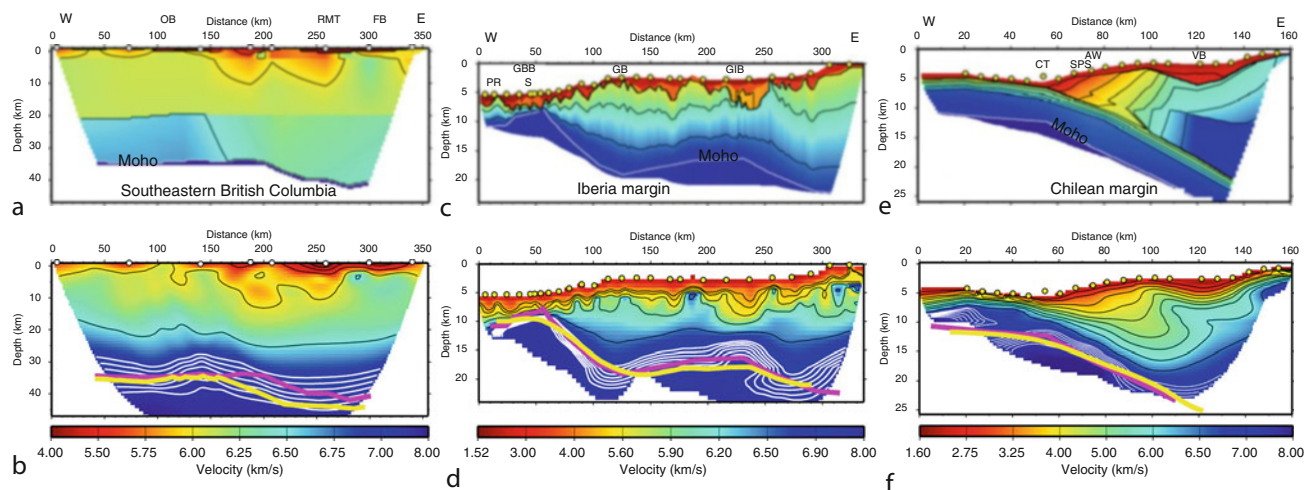


Traveltime Tomography Using Controlled-Source Seismic Data, Fig. 17 Example of crustal velocity model from onshore-offshore survey across the South American-Caribbean plate boundary at 64°W (Clark et al. 2008). There were airguns offshore, two shots on land (green dots), OBS's offshore (yellow dots), and seismometers on land (red dots). The model above the dashed white line was obtained by first-arrival tomography using the Zelt and Barton (1998) smoothing-

regularized tomography algorithm; the model below the dashed line was derived using the Zelt and Smith (1992) layer-based inversion using upper mantle refractions and Moho reflections to determine a sharp velocity contrast for the Moho. The strike-slip system near the coastline at ~150 km accommodates relative motion between the two plates and from this model is interpreted to extend near vertically through the entire crust and offset the Moho

waveforms because the crust and near-surface are much more laterally heterogeneous than the mantle. As a result, a non-linear inverse method must be used, requiring a starting model, an iterative method and recalculated travel paths at each iteration. In addition, the method must be capable of calculating a frequency-dependent traveltime along the total

path, as opposed to a delay time. Zelt (2009) presented a finite-frequency traveltime tomography method specifically designed for controlled-source data, that is, a nonlinear inversion of total traveltimes, with frequency taken into account for both the forward calculation of traveltimes, and in the inverse step by calculating the appropriate frequency-dependent



Traveltime Tomography Using Controlled-Source Seismic Data,

Fig. 18 A comparison of three sets of 2D crustal velocity models obtained using two different inversion algorithms (Zelt et al. 2003). The study areas are southeastern British Columbia, the Iberia margin and the Chilean margin. The top row of “geologic” models were obtained using the Zelt and Smith (1992) layer-based algorithm simultaneously inverting crust and upper mantle refractions and Moho reflections. The bottom row of “minimum-structure” models were obtained using the Zelt and Barton (1998) smoothing-regularized first-arrival tomography algorithm. Open circles are shot locations; yellow circles

are OBS’s. Black contours in the crust have an interval of 0.5 km/s; white contours representing the Moho in the smooth tomographic models have an interval of 0.1 km/s from 7.0–7.6 km/s. Pink lines are the Mohos from the layered models overlain on the tomographic models for comparison. Yellow lines are the Mohos from independent reflection tomography in which the tomographic velocity models were held fixed. Interpretations of the main features of the geological models are supported by the corresponding minimum-structure models since they contain the same features

sensitivity kernels. The results show it is possible to achieve more accurate velocity estimation than in the equivalent infinite-frequency-derived models. And in addition, applications without any regularization are possible, and therefore finite-frequency traveltime tomography has the potential to allow the data alone to determine the model structure in a robust manner. These characteristics of finite-frequency traveltime tomography suggest that it will play an important role in the future. This is especially true for near-surface studies where infinite-frequency (ray) theory is least likely to be valid because the length scales of heterogeneities are often comparable or smaller than the seismic wavelengths.

Cross-References

- Continental Crustal Structure
- Deep Seismic Reflection and Refraction Profiling
- Magnetic Anisotropy
- Ocean Bottom Seismics

Bibliography

- Bevington PR (1969) Data reduction and error analysis for the physical sciences. McGraw-Hill, New York
- Bosch M, Barton P, Singh SC, Trinks I (2005) Inversion of traveltime data under a statistical model for seismic velocities and layer interfaces. *Geophysics* 70:R33–R43

- Brenders AJ, Pratt RG (2007) Full waveform-tomography for lithospheric imaging: results from a blind test in a realistic crustal model. *Geophys J Int* 168:133–151
- Brocher TM (2005) Empirical relations between elastic wavespeeds and density in the Earth’s crust. *Bull Seismol Soc Am* 95: 2081–2092
- Cerveny V, Molotkov I, Psencik I (1977) Ray method in seismology. University of Karlova, Prague
- Chambers G, Levander A, Zelt CA, Dugan B (2009) Seismic waveform tomography with multicomponent data at a groundwater contamination site. *SEG Expanded Abstracts*, 28, 3994
- Chapman CH, Drummond R (1982) Body-wave seismograms in inhomogeneous-media using Maslov asymptotic theory. *Bull Seismol Soc Am* 72:S277–S317
- Christeson GL, McIntosh KD, Shipley TH, Flueh ER, Goedde H (1999) Structure of the Costa Rica convergent margin, offshore Nicoya Peninsula. *J Geophys Res* 104:25443–25468
- Clark SA, Zelt CA, Magnani MB, Levander A (2008) Characterizing the Caribbean-South American plate boundary at 64° W using wide-angle seismic data. *J Geophys Res* 113:B07401. <https://doi.org/10.1029/2007JB005329>
- Clowes RM, Zelt CA, Amor JR, Ellis RM (1995) Lithospheric structure in the southern Canadian Cordillera from a network of seismic refraction lines. *Can J Earth Sci* 32:1485–1513
- Constable SC, Parker RL, Constable CG (1987) Occam’s inversion: a practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics* 52:289–300
- Dahlen FA, Hung SH, Nolet G (2000) Frechet kernels for finite-frequency traveltimes – I. Theory. *Geophys J Int* 141:157–174
- Dunn RA, Toomey DR, Solomon SC (2000) Three-dimensional seismic structure and physical properties of the crust and shallow mantle beneath the East Pacific Rise at 9° 30’N. *J Geophys Res* 105:23537–23555
- Dziewonski AM, Anderson DL (1984) Seismic tomography of the earth’s interior. *Am Sci* 72:483–494

- Firbas P (1981) Inversion of travel-time data for laterally heterogeneous velocity structure – linearization approach. *Geophys J R Astron Soc* 67:189–198
- Hammer PTC, Dorman LM, Hildebrand JA, Cornuelle BD (1994) Jasper Seamount structure: seafloor seismic refraction tomography. *J Geophys Res* 99:6731–6752
- Hobro JWD, Singh SC, Minshull TA (2003) Threedimensional tomographic inversion of combined reflection and refraction seismic traveltime data. *Geophys J Int* 152:79–93
- Hobro JWD, Minshull TA, Singh SC, Chand S (2005) A three-dimensional seismic tomographic study of the gas hydrate stability zone, offshore Vancouver Island. *J Geophys Res* 110:B09102. <https://doi.org/10.1029/2004JB003477>
- Hole JA (1992) Nonlinear high-resolution three-dimensional seismic travel time tomography. *J Geophys Res* 97:6553–6562
- Hole JA, Zelt BC (1995) Three-dimensional finite-difference reflection travel times. *Geophys J Int* 121:427–434
- Hole JA, Clowes RM, Ellis RM (1992) Interface inversion using broad-side seismic refraction data and 3-dimensional travel time calculations. *J Geophys Res* 97:3417–3429
- Hole JA, Ryberg T, Fuis GS, Bleibinhaus F, Sharma AK (2006) Structure of the San Andreas fault zone at SAFOD from a seismic refraction survey. *Geophys Res Lett* 33:L07312. <https://doi.org/10.1029/2005GL025194>
- Jaiswal P, Zelt CA, Pecher IA (2006) Seismic characterization of a gas hydrate system in the Gulf of Mexico using wide-aperture data. *Geophys J Int* 165:108–120
- Korenaga J, Holbrook WS, Kent GM, Kelemen PB, Detrick RS, Larsen HC, Hopper JR, Dahl-Jensen T (2000) Crustal structure of the Southeast Greenland margin from joint refraction and reflection seismic tomography. *J Geophys Res* 105:21591–21614
- Korenaga J, Holbrook WS, Detrick RS, Kelemen PB (2001) Gravity anomalies and crustal structure at the Southeast Greenland margin. *J Geophys Res* 106:8853–8870
- Lailly P, Sinoquet D (1996) Smooth velocity models in reflection tomography for imaging complex geological structures. *Geophys J Int* 124:349–362
- Lanz E, Maurer H, Green AG (1998) Refraction tomography over a buried waste disposal site. *Geophysics* 63:1414–1433
- Levander A, Zelt CA, Symes WW (2007) Active source studies of crust and lithospheric structure. In: Romanowicz B, Dziewonski A (eds) *Treatise on geophysics, seismology and structure of the earth*, vol 1. Elsevier, Amsterdam/Boston, pp 247–288
- Lutter WJ, Nowack RL (1990) Inversion for crustal structure using reflections from the PASSCAL Ouachita experiment. *J Geophys Res* 95:4633–4646
- Lutter WJ, Nowack RL, Braile LW (1990) Seismic imaging of upper crustal structure using travel times from the PASSCAL Ouachita experiment. *J Geophys Res* 95:4621–4631
- Magnani MB, Zelt CA, Levander A, Schmitz M (2009) Crustal structure of the South American-Caribbean plate boundary at 67°W from controlled source seismic data. *J Geophys Res* 114:B02312. <https://doi.org/10.1029/2008JB005817>
- McCaughy M, Singh SC (1997) Simultaneous velocity and interface tomography of normal-incidence and wide-aperture traveltime data. *Geophys J Int* 131:87–99
- McMechan GA, Mooney WD (1980) Asymptotic ray theory and synthetic seismograms for laterally varying structures: theory and application to the Imperial Valley, California. *Bull Seismol Soc Am* 70:2021–2035
- Menke W (1989) *Geophysical data analysis: discrete inverse theory*. Academic, San Diego
- Moser TJ (1991) Shortest path calculation of seismic rays. *Geophysics* 56:59–67
- Nakanishi I, Yamaguchi K (1986) A numerical experiment on nonlinear image reconstruction from first-arrival times for two-dimensional island arc structure. *J Phys Earth* 34:195–201
- Nielsen L, Jacobsen BH (2000) Integrated gravity and wide-angle seismic inversion for 2-D crustal modelling. *Geophys J Int* 140:222–232
- Nowack RL, Braile LW (1993) Refraction and wide-angle reflection tomography: theory and results. In: Iyer HM, Hirahara K (eds) *Seismic tomography: theory and practice*. Chapman and Hall, London, pp 733–765
- Paige CC, Saunders MA (1982) LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Trans Math Softw* 8:43–71
- Podvin P, Lecomte I (1991) Finite difference computation of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *Geophys J Int* 105:271–284
- Pratt RG, Chapman CH (1992) Traveltime tomography in anisotropic media–II. Application. *Geophys J Int* 109:20–37
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical recipes in Fortran: the art of scientific computing*, 2nd edn. Cambridge University Press, Cambridge
- Ramachandran K, Dosso SE, Spence GD, Hyndman RD, Brocher TM (2005) Forearc structure beneath southwestern British Columbia: a three-dimensional tomographic velocity model. *J Geophys Res* 110: B02303. <https://doi.org/10.1029/2004JB003258>
- Rawlinson N, Sambridge M (2003a) Seismic traveltime tomography of the crust and lithosphere. *Adv Geophys* 46:181–198
- Rawlinson N, Sambridge M (2003b) Irregular interface parametrization in 3-D wide-angle seismic traveltime tomography. *Geophys J Int* 155:79–92
- Rawlinson N, Houseman GA, Sambridge M (2001) Inversion of seismic refraction and wide-angle reflection traveltimes for 3-D layered crustal structure. *Geophys J Int* 145:381–401
- Scales JA, Docherty P, Gersztenkorn A (1990) Regularisation of non-linear inverse problems: imaging the near-surface weathering layer. *Inverse Probl* 6:115–131
- Schmelzbach C, Zelt CA, Juhlin C, Carbonell R (2008) P- and Sv-velocity structure of the South Portuguese Zone fold-and-thrust belt, SW Iberia, from traveltime tomography. *Geophys J Int* 175:689–712
- Sethian JA, Popovici AM (1999) 3-D traveltime computation using the fast marching method. *Geophysics* 64:516–523
- Spence GD, Whittall KP, Clowes RM (1984) Practical synthetic seismograms for laterally varying media calculated by asymptotic ray theory. *Bull Seismol Soc Am* 74:1209–1223
- Spence GD, Clowes RM, Ellis RM (1985) Seismic structure across the active subduction zone of western Canada. *J Geophys Res* 90:6754–6772
- Toomey DR, Solomon SC, Purdy GM (1994) Tomographic imaging of the shallow crustal structure of the East Pacific Rise at 9° 30'N. *J Geophys Res* 99:24135–24157
- Trinks I, Singh SC, Chapman CH, Barton PJ, Bosch M, Cherrett A (2005) Adaptive traveltime tomography of densely sampled seismic data. *Geophys J Int* 160:925–938
- Um J, Thurber C (1987) A fast algorithm for two-point seismic ray tracing. *Bull Seismol Soc Am* 77:972–986
- Van Avendonk HJA, Harding AJ, Orcutt JA, McClain JS (1998) A two-dimensional tomographic study of the Clipperton transform, fault. *J Geophys Res* 103:17885–17899
- Van Avendonk HJA, Harding AJ, Orcutt JA, Holbrook WS (2001) Hybrid shortest path and ray bending method for traveltime and raypath calculations. *Geophysics* 66:648–653
- Van Avendonk HJA, Shillington DJ, Holbrook WS, Hornbach MJ (2004) Inferring crustal structure in the Aleutian arc from a sparse wide-angle seismic data set. *Geochem Geophys Geosyst* 5:Q08008. <https://doi.org/10.1029/2003GC000664>
- Vidale JE (1988) Finite-difference calculation of traveltimes. *Bull Seismol Soc Am* 78:2062–2076
- Vidale JE (1990) Finite-difference calculation of traveltimes in three dimensions. *Geophysics* 55:521–526

- White DJ (1989) Two-dimensional seismic refraction tomography. *Geophys J Int* 97:223–245
- Zelt CA (1998) Lateral velocity resolution from three-dimensional seismic refraction data. *Geophys J Int* 135:1101–1112
- Zelt CA (1999) Modeling strategies and model assessment for wide-angle seismic traveltime data. *Geophys J Int* 139:183–204
- Zelt CA (2009) Frequency-dependent traveltime tomography for controlled-source, near-surface seismic data. *EOS Trans AGU*, 90(52), Fall Meeting Suppl, Abstract NS31A–1155
- Zelt CA, Barton PJ (1998) Three-dimensional seismic refraction tomography: a comparison of two methods applied to data from the Faeroe Basin. *J Geophys Res* 103:7187–7210
- Zelt CA, Ellis RM (1988) Practical and efficient ray tracing in two-dimensional media for rapid traveltime and amplitude forward modeling. *Can J Explor Geophys* 24:16–31
- Zelt CA, Smith RB (1992) Seismic traveltime inversion for 2-D crustal velocity structure. *Geophys J Int* 108:16–34
- Zelt CA, White DJ (1995) Crustal structure and tectonics of the south-eastern Canadian Cordillera. *J Geophys Res* 100:24255–24273
- Zelt CA, Hojka AM, Flueh ER, McIntosh KD (1999) 3D simultaneous seismic refraction and reflection tomography of wide-angle data from the central Chilean margin. *Geophys Res Lett* 26:2577–2580
- Zelt CA, Sain K, Naumenko JV, Sawyer DS (2003) Assessment of crustal velocity models using seismic refraction and reflection tomography. *Geophys J Int* 153:609–626
- Zelt CA, Azaria A, Levander A (2006a) 3D seismic refraction traveltime tomography at a groundwater contamination site. *Geophysics* 71: H67–H78
- Zelt CA, Ellis RM, Zelt BC (2006b) 3-D structure across the Tintina strike-slip fault, northern Canadian Cordillera, from seismic refraction and reflection tomography. *Geophys J Int* 167:1292–1308
- Zhang J, ten Brink US, Toksöz MN (1998) Nonlinear refraction and reflection traveltime tomography. *J Geophys Res* 103:29743–29757

Tsunami

Steven N. Ward

Institute of Geophysics and Planetary Physics, University of California at Santa Cruz, Santa Cruz, CA, USA

Definition

Dispersive	Characteristic of waves whose velocity of propagation depends on wave frequency. The shape of a dispersive wave packet changes as it moves along.
Eigenfunction	Functional shape of the horizontal and vertical components of wave motion versus depth in the ocean for a specific wave frequency.
Geometrical spreading	Process of amplitude reduction resulting from the progressive expansion of a wave from its source.
Run-up	Final phase of tsunami life starting when the wave shoals to a size equal to the water depth and begins to break, and ending when

the water runs over land and reaches its highest level.

Shoal

Process of waves coming into shallow water. Shoaling waves slow, shorten their wavelength, and grow in size.

Wavenumber

Wavenumber k equals 2π divided by wavelength λ . Large wavenumbers associate with short waves and small wavenumbers associate with long waves.

Tsunami are gravity-driven water waves. They belong to the same family as common sea waves that we see every day; however, tsunami are distinct in their mode of generation and in their physical traits. Unlike common sea waves that evolve from persistent winds, most tsunami spring from sudden shifts of the ocean floor. These sudden shifts can originate from undersea landslides and volcanoes, but mostly, submarine earthquakes parent tsunami. Compared to wind-driven waves, tsunami waves have periods, velocities, and wavelengths ten or a hundred times larger and present profoundly different shoreline consequences than do their common cousins.

Tsunami = Killer Wave?

In the years since the 2004 Sumatra earthquake, everyone has seen disturbing videos of tsunami-caused destruction. Certainly the prospect of a “killer wave” born from some far off earthquake is frightening. More so is the thought that after traveling with stealth great distances, that wave might suddenly rise up without warning at your own doorstep. Could it be me desperately floating by in the next tsunami video?

Understandably, worst-case scenarios of natural hazards come to mind, but it is important to keep perspective. Tsunami over 2 m high are not common. It takes a submarine earthquake greater than magnitude M8 to source a wave of this size. On a global average, only one M8+ earthquake occurs per year. Of these, maybe one in five strikes under the ocean with an orientation favorable for tsunami excitation. True, tsunami decay in transit and have shorter run-up at distant shores, so wave damage tends to localize within 1,000 km of that one in five quakes. On these accounts, tsunami that induce widespread damage and casualties number only about one per decade. Even at that modest recurrence rate, your personal risk is mitigated increasingly through education and technology. Today, ocean bottom pressure sensors detect stealthy tsunami of a few centimeters height in the open sea. Moreover, with advances in understanding, communication, and implementation since the 2004 Sumatra earthquake, scientists are better, quicker, and more specific at prediction than they were just a decade past. There is a good chance now that you will be warned and take action even if that rare wave does come your way.

Tsunami Characteristics

Tsunami Period, Velocity, and Wavelength

When discussing tsunami, I like to contrast them with something that we all have experienced – waves at the beach. Tsunami belong to the same family as these ordinary ocean waves but with major distinctions: principally tsunami period, tsunami velocity, and tsunami wavelength.

Tsunami period. The period (T) of a wave equals the time elapsed between one passing crest and the next. Common beach waves have a period of about 10 s (darkened column right side Fig. 1). The period of tsunami depends upon what creates it (earthquake, landslide, asteroid impact), but I can tell you that tsunami have far greater periods than beach waves. The “tsunami window” in Fig. 1 covers waves of 70, 200, 500, or even 2,000 s periods. Unlike “splash and dash” beach waves, tsunami arrive and may continue to flow in for several minutes. I like the description – “The ocean turns into a river.” Anyone who has seen those 2004 Sumatra videos can testify that tsunami act more flood-like than wave-like.

Tsunami velocity. Under classical theory, the phase $c(\omega)$, and group $u(\omega)$ velocity of surface gravity waves on a flat ocean of uniform depth h are:

$$c(\omega) = \sqrt{\frac{gh \tanh[k(\omega)h]}{k(\omega)h}} \quad (1)$$

and

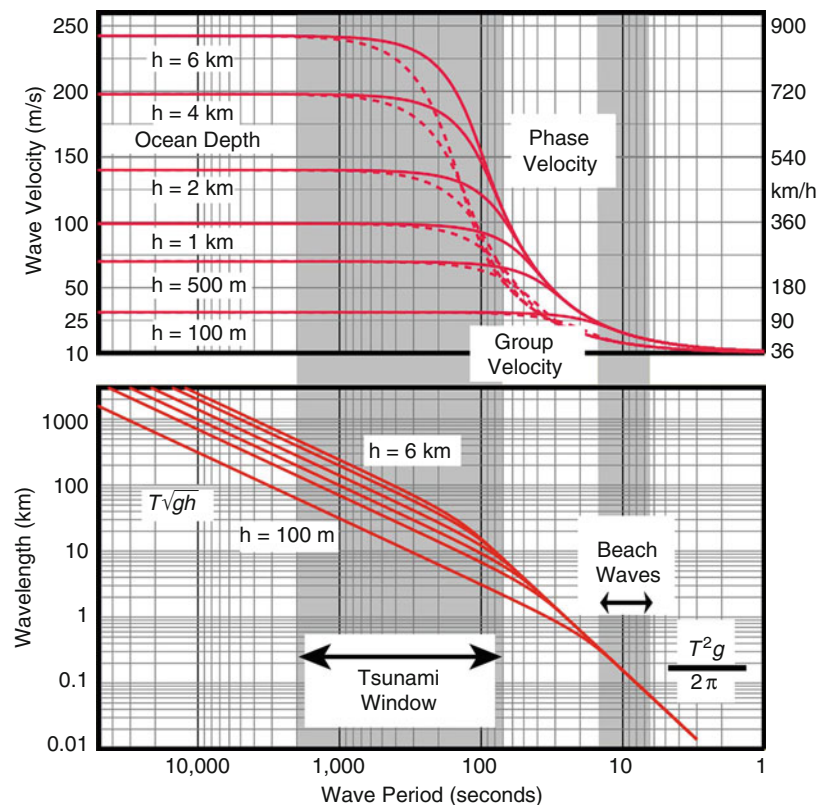
$$u(\omega) = c(\omega) \left[\frac{1}{2} + \frac{k(\omega)h}{\sinh[2k(\omega)h]} \right]. \quad (2)$$

Here, g is the acceleration of gravity (9.8 m/s^2) and $k(\omega)$ is the *wavenumber* associated with a sea wave of frequency $\omega = 2\pi/T$. Wavenumber connects to wavelength as $\lambda(\omega) = 2\pi/k(\omega)$ and to phase velocity as $c(\omega) = \omega/k(\omega)$. Wavenumber also satisfies the relation:

$$\omega^2 = gk(\omega) \tanh[k(\omega)h]. \quad (3)$$

For surface gravity waves spanning 1–50,000 s period, Fig. 1 (*top*) plots $c(\omega)$ and $u(\omega)$. These velocities vary widely and increase for waves of longer periods in deeper water. The speed of ordinary sea waves is about 15 m/s (50 km/h) – the speed of a moped. Because of their longer period, waves in the tsunami window travel much more rapidly, reaching 160–250 m/s (600–900 km/h) in the open ocean. It is said that deep-water tsunami travel at the speed of a jet airliner. While true, be aware that wave energy is the quantity that runs ahead at jet speeds. The physical bits of water in a deep-water tsunami move less than a meter per second (see next section). Waves whose velocity varies with frequency are called *dispersive*. During propagation, dispersion “pulls apart”

Tsunami, Fig. 1 Phase velocity $c(\omega)$ (solid lines) and group velocity $u(\omega)$ (dashed lines) of tsunami waves on a flat earth covered by oceans of 100 m to 6 km depth (*top*). Wavelength associated with each wave period. The “tsunami window” is marked (*bottom*)



originally pulse-like waves into their component frequencies. Dispersion is strongest for waves whose period falls on the steepest slopes of the group velocity curves in Fig. 1.

Tsunami wavelength. Wavelength measures the distance between one wave crest and the next. Wavelength equals the product of wave period times phase velocity $\lambda(\omega) = Tc(\omega)$. Common beach waves have a wavelength of about 100 m – the length of a football field (Fig. 1, *bottom*). Tsunami, with their longer period and higher velocity, have much longer wavelengths than beach waves. In the deep ocean, tsunami span 10, 30, even 100 km between crests. If you could stand on one tsunami crest, the next one might be over the horizon. With wavelengths this large, tsunami slopes are very small even if the wave has large amplitude. For ships at sea, tsunami pass completely unnoticed.

Short wave versus long wave: Discussions of waves of length λ , in oceans of depth h sometimes include two simplifications: a long wave approximation ($\lambda \gg h$, $1/k \gg h$) and a short wave approximation ($\lambda \ll h$, $1/k \ll h$). Under a long wave approximation [$kh \rightarrow 0$, $\tanh(kh) \rightarrow kh$, $\sinh(2kh) \rightarrow 2kh$] Eqs. 1, 2, and 3 predict nondispersive wave propagation with $c(\omega) = u(\omega) = \sqrt{gh}$. Long wave theory holds for the flat part of the curves in Fig. 1 (*top*). Under a short wave approximation, [$kh \rightarrow \infty$, $\tanh(kh) \rightarrow 1$, $\sinh(kh) \rightarrow \infty$] the equations predict dispersive propagation with $c(\omega) = 2u(\omega) = gT/2\pi$. Short wave theory holds to the right in Fig. 1 (*top*) where all the curves lie atop each other. Waves in the tsunami window have intermediate character, behaving like shallow water waves at their longest periods and like deep-water waves at their shortest periods. Neither the long nor the short wave simplification serves

adequately in tsunami studies. A rigorous treatment requires an approach that works for waves of all lengths.

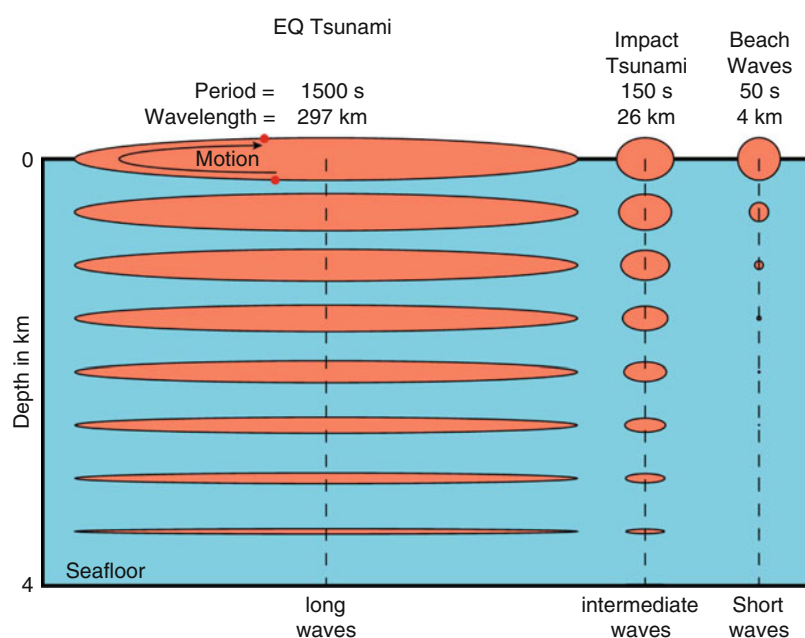
Tsunami Eigenfunctions

Tsunami also differ from waves at the beach in the way they move the ocean. Tsunami *eigenfunctions* describe wave motion in a tsunami mode of a particular frequency. Consider coordinate system (x, y, z) where $\hat{x}$ points north, $\hat{y}$ east, and $\hat{z}$ down. Vertical (u_z) and horizontal (u_x) components of tsunami eigenfunctions normalized to vertical amplitude A_z at the sea surface are:

$$\begin{aligned} u_z(\omega, z) &= A_z \frac{k(\omega)g}{\omega^2} \frac{\sinh[k(\omega)(h-z)]}{\cosh[k(\omega)h]} e^{i[k(\omega)x - \omega t]} \\ u_x(\omega, z) &= A_z \frac{-ik(\omega)g}{\omega^2} \frac{\cosh[k(\omega)(h-z)]}{\cosh[k(\omega)h]} e^{i[k(\omega)x - \omega t]} \end{aligned} \quad (4)$$

Figure 2 plots tsunami eigenfunctions versus depth in a 4-km deep ocean at long (1,500 s), intermediate (150 s), and short (50 s) periods. The ellipses trace the path of a water particle as a wave of frequency ω passes. At 50-s period (Fig. 2, *right*), ordinary ocean waves have deep-water behavior. Water particles move in circles that decay exponentially from the surface. Sensibly, because the eigenfunctions of short waves do not reach to the seafloor, their velocity is independent of ocean depth [$c(\omega) = 2u(\omega) = gT/2\pi$, Fig. 1, *top right*]. The failure of short waves to “feel” the seafloor also means that they cannot be excited by deformations of it. The only means to excite ordinary ocean waves is to disturb the surface.

Tsunami, Fig. 2 Tsunami eigenfunctions in a 4-km deep ocean at periods 1,500, 150, and 50 s. Vertical displacements at the ocean surface has been normalized to $A_z = 1$ m in each case



At 1,500 s period (Fig. 2, *left*), the tsunami has a wavelength of $\lambda = 297$ km and it acts like a long wave. Although vertical displacement peaks at the ocean surface and drops to zero at the seafloor, horizontal displacement persists undiminished through the ocean column. Unlike ordinary waves that are confined near the sea surface, the energy of a tsunami spreads through the entire depth of the sea. You cannot out dive a tsunami. This fact explains why tsunami can be detected by pressure sensors on the seafloor. A second distinction of tsunami versus ordinary sea waves is that their horizontal motion far exceeds their vertical motion. Every meter of up and down in a 1,500-s tsunami involves ≈ 10 m of back and forth. If you were to build a shore-based or space-based “tsunami detector,” the large horizontal motions might make a better target than the vertical motions. Applications of Doppler radar come to mind.

Toward the short period side of the tsunami window at 150 s (Fig. 2, *middle*), λ decreases to 26 km. For these tsunami waves, long wave characteristics begin to break down, and horizontal and vertical motions more closely agree in amplitude.

From (4), it is easy to deduce the peak velocity of surface water in waves of amplitude A_z as:

$$|v_x^{\max}(\omega, z=0)| = \frac{A_z k(\omega) g}{\omega} = \frac{A_z g}{c(\omega)}. \quad (5)$$

For long waves $c(\omega)$ equals $\sqrt{gh}$, and peak surface water velocity in a 1-m amplitude tsunami in 4,000 m of water would be just 4.9 cm/s. Peak water velocity at the sea floor

$$\begin{aligned} |v_x^{\max}(\omega, z=h)| &= \frac{A_z k(\omega) g}{\omega \cosh[k(\omega)h]} \\ &= \frac{A_z g}{c(\omega) \cosh[k(\omega)h]} \end{aligned} \quad (6)$$

is even less. As mentioned, although tsunami travel at several hundred meters per second, the water itself moves at a tiny fraction of this.

Tsunami Excitation

Tsunami get started in many ways. Suppose that the seafloor at positions $\mathbf{r}_0$ uplift instantaneously by an amount $u_z^{\text{bot}}(\mathbf{r}_0)$ at time $\tau(\mathbf{r}_0)$. Under classical tsunami theory in a uniform ocean of depth h , this sea bottom disturbance produces surface tsunami waveforms (vertical component) at observation point $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}}$ and time t of

$$u_z^{\text{surf}}(\mathbf{r}, t) = \text{Re} \int_{\mathbf{k}} d\mathbf{k} \frac{e^{i[\mathbf{k} \cdot \mathbf{r} - \omega(k)t]}}{4\pi^2 \cosh(kh)} F(\mathbf{k}). \quad (7a)$$

with

$$F(\mathbf{k}) = \int_{\mathbf{r}_0} d\mathbf{r}_0 u_z^{\text{bot}}(\mathbf{r}_0) e^{-i[\mathbf{k} \cdot \mathbf{r}_0 - \omega(k)\tau(\mathbf{r}_0)]} \quad (7b)$$

$k = |\mathbf{k}|$, and $\omega^2(k) = gk \tanh(kh)$. The integrals in (7a and 7b) cover all wavenumber space and locations $\mathbf{r}_0$ where the seafloor disturbance $u_z^{\text{bot}}(\mathbf{r}_0) \neq 0$.

Equation 7a looks scary but it has three identifiable pieces:

- The $F(\mathbf{k})$ term is the wavenumber spectrum of the seafloor uplift. This number relates to the amplitude, spatial and time distribution of the uplift. Tsunami trains are dominated by wavenumbers in the span where $F(\mathbf{k})$ is the greatest. The peak of $F(\mathbf{k})$ corresponds to the characteristic dimension of the uplift. Large-dimensioned uplifts produce longer wavelength, lower frequency tsunami than small-dimensioned sources.
- The $1/\cosh(kh)$ term comes from the tsunami eigenfunction shapes (4) and it acts to low-pass filter the source spectrum $F(\mathbf{k})$. Because $1/\cosh(kh) \rightarrow 1$ when $kh \rightarrow 0$, and $1/\cosh(kh) \rightarrow 0$ when $kh \rightarrow \infty$, the filter favors long waves. Due to the low-pass filter effect of the ocean layer, only wavelengths of the uplift that exceed three times the ocean depth (i.e., $kh = 2\pi h/\lambda < \approx 2$) contribute much to tsunami.
- The exponential term in (7a) contains all of the propagation information including travel time, *geometrical spreading*, and frequency dispersion.

By rearranging Eqs. 7a and 7b, vertical tsunami motions at $\mathbf{r}$ can also be written as:

$$u_z^{\text{surf}}(\mathbf{r}, t) = \text{Re} \int_0^\infty \frac{k dk e^{-i\omega(k)t}}{2\pi \cosh(kh)} \sum_{n=-\infty}^\infty J_n(kr) e^{in\theta} F_n(k)$$

with

$$F_n(k) = \int_{\mathbf{r}_0} d\mathbf{r}_0 u_z^{\text{bot}}(\mathbf{r}_0) J_n(kr_0) e^{i(\omega(k)\tau(\mathbf{r}_0) - n\theta_0)} \quad (8)$$

Here θ marks azimuth from north (the $\hat{\mathbf{x}}$ direction) of the observation point $\mathbf{r}$ from the coordinate origin. The $J_n(x)$ are cylindrical Bessel functions. For simply distributed uplift sources, (8) might be easier to evaluate than (7a and 7b).

Tsunami Excitation by Earthquakes

Earthquakes produce 80–90% of tsunami. Not surprisingly, earthquake features determine many sea wave characteristics. Earthquakes result from slip on faults and three primary parameters describe the process – moment, mechanism, and depth.

Moment measures earthquake strength. Moment M_0 is the product of rigidity μ of the source region's rocks, fault area A , and average fault slip Δu . Earthquake moment and earthquake magnitude tie through a number of empirical formulae. One formula defines moment magnitude M_w as $M_w = (2/3)(\log M_0 - 9.05)$. Earthquake moment varies by 2×10^4 within the magnitude range $6.5 \leq M_w \leq 9.5$ (Table 1). Even without a detailed understanding of tsunami generation, it is safe to suppose that the larger the earthquake moment, the larger the tsunami, all else fixed.

Mechanism specifies the orientation of the earthquake fault and the direction of slip on it. Usually, faults are idealized as plane rectangles with normal $\hat{\mathbf{n}}$. Three angles then, summarize earthquake mechanisms – the strike and dip of the fault and the angle of slip vector $\hat{\mathbf{a}}$ measured from the horizontal in the plane of the fault. (Seismologists call this angle the “rake.”) The role of fault mechanism on tsunami production is not as obvious as the influence of moment; however, one might suspect that earthquakes that affect large vertical displacements of the seafloor would be more effective than faults that make large horizontal displacements.

Earthquake *depth* needs no explanation. Because seafloor shifts cause tsunami, the distance of the fault from the seafloor should be important. Presumably, deep earthquakes produce less potent tsunami than similar shallow earthquakes.

Numerical, or synthetic waveforms quantify the roles of earthquake parameters on tsunami generation. For illustration, insert into (7a and 7b) the surface uplift pattern $u_z^{\text{bot}}(\mathbf{r}_0)$ of a small earthquake fault (point source really) placed at depth d in a halfspace. Further assume that the uplift occurs instantly with $\tau(\mathbf{r}_0) = 0$. (Actually, real earthquakes uplift the seafloor over several, or several tens of, seconds. This distinction is not a big issue because tsunami waves have periods of many hundreds of seconds. Uplifts taking a few dozen seconds to develop look “instantaneous” to tsunamis.) Equations 7a and 7b becomes

$$u_z^{\text{surf}}(\mathbf{r}, t) = \int_0^\infty k dk \frac{\cos \omega(k)t}{2\pi \cosh(kh)} [A \Delta u \mathbf{M}_{ij} \epsilon_{ij}], \quad (9)$$

where

$$\begin{aligned} \epsilon_{xx} &= -\frac{1}{4} \left(\frac{\mu}{\lambda + \mu} - kd \right) [J_0(kr) - J_2(kr) \cos 2\theta] e^{-kd} \\ \epsilon_{yy} &= -\frac{1}{4} \left(\frac{\mu}{\lambda + \mu} - kd \right) [J_0(kr) + J_2(kr) \cos 2\theta] e^{-kd} \\ \epsilon_{xy} &= \epsilon_{yx} = \frac{1}{4} \left(\frac{\mu}{\lambda + \mu} - kd \right) [J_2(kr) \sin 2\theta] e^{-kd} \\ \epsilon_{zz} &= -\frac{1}{2} \left(\frac{\mu}{\lambda + \mu} + kd \right) [J_0(kr)] e^{-kd} \\ \epsilon_{xz} &= \epsilon_{zx} = \frac{kd}{2} [J_1(kr) \cos \theta] e^{-kd} \\ \epsilon_{yz} &= \epsilon_{zy} = \frac{kd}{2} [J_1(kr) \sin \theta] e^{-kd} \end{aligned} \quad (10)$$

The six elements of symmetric tensor

$$M_{jk} = (\hat{a}_j \hat{n}_k + \hat{n}_j \hat{a}_k) \quad (11)$$

capsulize the mechanism of the earthquake. In (11), $\hat{\mathbf{n}}$, $\hat{\mathbf{a}}$ are the fault normal and slip vectors introduced above. A pure dip slip earthquake on a vertical north–south trending fault for instance, has $\hat{\mathbf{n}} = \hat{\mathbf{y}}$ and $\hat{\mathbf{a}} = \hat{\mathbf{z}}$, so $M_{yz} = M_{zy} = 1$ and $M_{xx} = M_{yy} = M_{zz} = M_{xz} = M_{zx} = M_{xy} = M_{yx} = 0$.

The bracketed terms in (9) contain all of the relationships between earthquake parameters and tsunami features. Some relationships are easy to spot: tsunami amplitudes from earthquakes are proportional to the product of fault area and average slip ($A \Delta u$); tsunami amplitudes decrease with earthquake depth via the e^{-kd} terms. The ϵ_{ij} provide the dependence of tsunami amplitude and azimuthal radiation pattern on source type. Equation 10 says that tsunami from point sources radiate in azimuthal patterns no more intricate than $\sin 2\theta$ or $\cos 2\theta$.

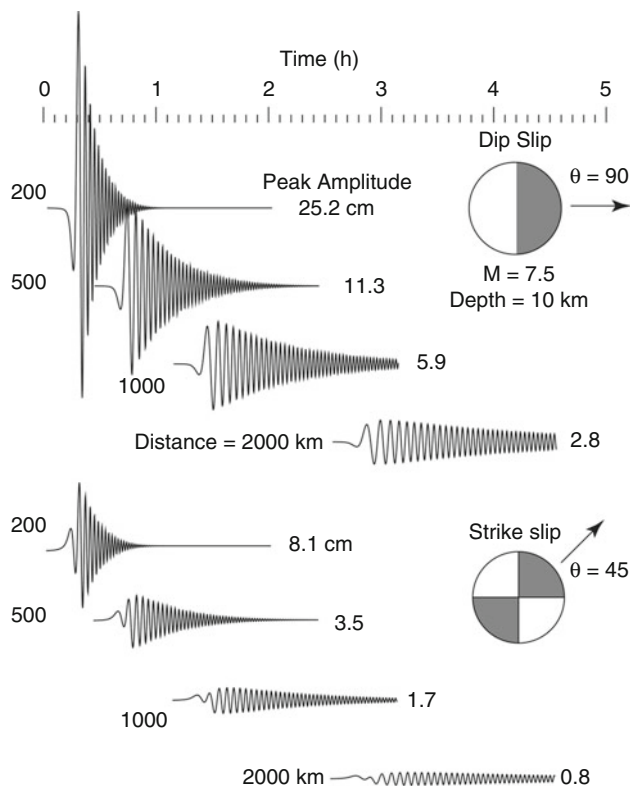
Figure 3 shows 5 h of tsunami waveforms calculated from (9) at distances of $r = 200, 500, 1,000, 2,000$ km from dip slip ($M_{yz} = M_{zy} = 1$) and strike slip ($M_{xy} = M_{yx} = 1$) point sources of magnitude $M_w = 7.5$ ($\Delta u A = 3.98 \times 10^6 \text{ m}^3$. See Table 1) buried at 10 km depth. Sea waves from these sources have

Tsunami, Table 1 Relationship between earthquake magnitude and moment with values of fault area, length, and mean slip for typical tsunami-generating earthquakes. This paper assumes $\log(L) = 0.5 M_w - 1.8$, $\Delta u = 2 \times 10^{-5} L$, and $\lambda = \mu = 5 \times 10^{10} \text{ Pa}$

Magnitude M_w	Moment M_0 (Nm)	Area A (km ²)	Length L (km)	Width W (km)	Slip Δu (m)
6.5	6.3×10^{18}	224	28	8	0.56
7.0	3.5×10^{19}	708	50	14	1.00
7.5	2.0×10^{20}	2,239	89	25	1.78
8.0	1.1×10^{21}	7,079	158	45	3.17
8.5	6.3×10^{21}	22,387	282	79	5.66
9.0	3.5×10^{22}	70,794	501	141	10.0
9.5	2.0×10^{23}	223,872	891	251	17.8

radiation patterns of $\sin\theta$ and $\sin 2\theta$, respectively. I compute the waveforms in Fig. 3 at the azimuth of maximum strength, $\theta = 90^\circ$ and $\theta = 45^\circ$. Frequency dispersion, with the long periods arriving first, is the most conspicuous feature of the waveforms. Tsunami onset rapidly. They reach maximum height in the first few cycles and then decay slowly over an hour or more. Even for this large earthquake, tsunami beyond 500 km distance reach just a few centimeters – hardly killer waves. Note that if the observation direction was west versus east (Fig. 3, *top*), or northwest versus northeast (Fig. 3, *bottom*), the tsunami waveforms would be inverted. Whether tsunami onset in a withdrawal or an inundation at a particular shoreline strictly depends on the style of faulting and the relative positions of the shore and the fault.

Figure 3 demonstrates that for point sources, dip slip earthquakes produce three or four times larger tsunami than strike slip earthquakes of equal moment. The differences in generation efficiency are understood most easily by considering directly the seafloor deformation patterns $u_z^{\text{bot}}(\mathbf{r}_0)$. I find these by setting $t = 0$ and $h = 0$ in (9) so $\cos\omega(k)t/\cosh(kh) = 1$. Figure 4 pictures the uplift patterns for the two faults of Fig. 3



Tsunami, Fig. 3 Synthetic record sections of vertical tsunami motions at distances of 200, 500, 1,000, and 2,000 km from point dip slip (*top*) and strike slip (*bottom*) earthquakes of magnitude $M_w = 7.5$ and depth 10 km. Time runs for 5 h and the peak amplitude of each trace is given in centimeters at the right. The lower half of the focal sphere and azimuth of observation θ are shown toward the right. For other directions, the waveforms should be scaled by $\sin\theta$ and $\sin 2\theta$, respectively

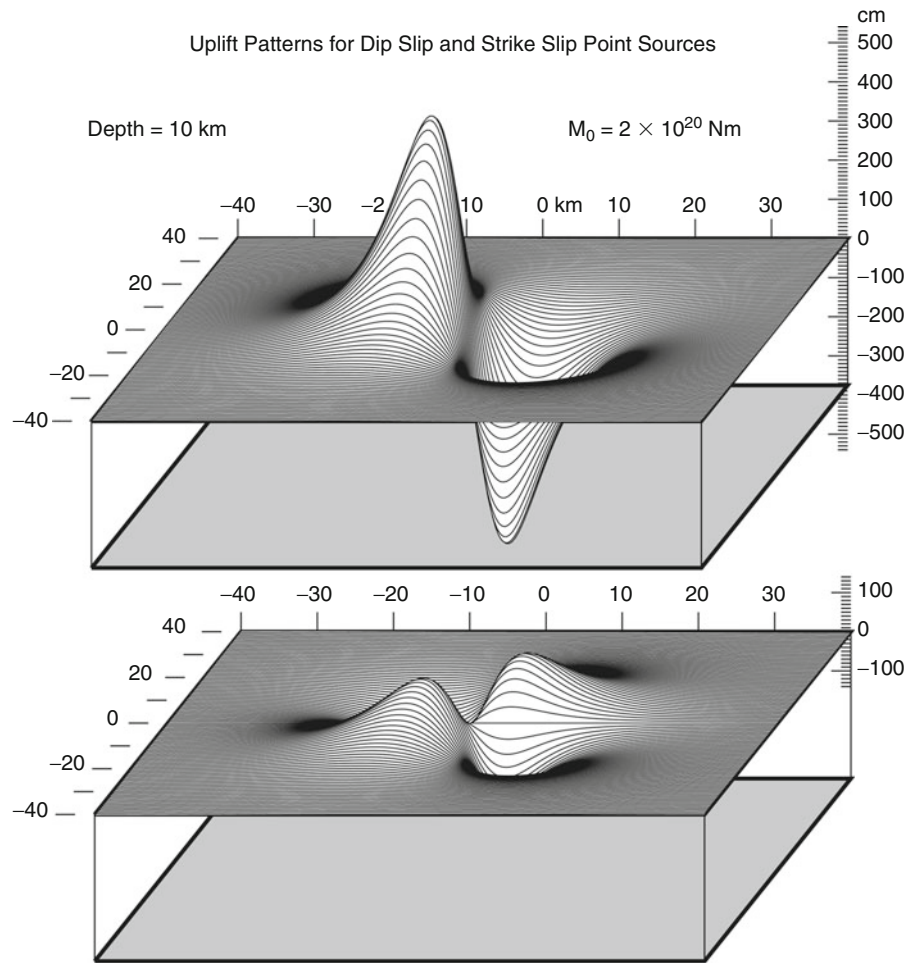
where two ($\sin\theta$) and four-lobed ($\sin 2\theta$) deformations spread over a region 40 km wide. The most striking contrast in the fields is maximum vertical displacement – 1.4 m for the strike slip versus 5.4 m for the dip slip. It is no coincidence that the ratio of maximum uplift for these two faults replicates the ratio of tsunami heights in Fig. 3. After all, vertical seafloor deformation drives tsunami and vertical deformation is controlled largely by the rake of the slip vector $\hat{\mathbf{a}}$. Strike slip faults have a rake of 0° or 180° . Dip slip faults have rake equal $\pm 90^\circ$. Further simulations show that, excepting very shallow, nearly horizontal faults, dip is not a terribly significant factor in tsunami production.

The sea surface cross sections in Fig. 5 (*left* and *middle*) chronicle the birth and early life of a sea wave spawned by M7.5 thrust earthquakes on 45° dipping planes. In these figures, I replace the idealized point sources of Fig. 3 with faults of typical dimension ($L = 89$ km, $W = 25$ km, and $\Delta u = 1.78$ m; see Table 1). In Fig. 5 (*left*) the fault reaches to the sea floor. In Fig. 5 (*middle*) the fault stops 30 km down. Soon after the earthquake, the sea surface forms “dimples” similar to those on the deformed sea floor. The sea surface dimples, however, are smoother and a bit lower in amplitude because of the $1/\cosh(kh)$ low-pass filtering effect of the ocean layer. After a time roughly equal to the dimension of the uplift divided by tsunami speed $\sqrt{gh}$, the leading edges of the wave organize and begin to propagate outward as expanding rings (Fig. 5, *right*). Early on, the wave appears as a single pulse. Characteristic tsunami dispersion begins to be seen only after 10 or 20 min. Consequently, for shorelines close to tsunami sources, seismic sea waves arrive mostly as a single pulse. For distant shorelines, sea waves arrive with many oscillations, dispersion having spread out the initial pulse.

The e^{-kd} terms in the tsunami excitation functions (10) let shallow earthquakes excite higher frequency tsunami than deep earthquakes (compare Fig. 5 *left*, with Fig. 5 *middle*). Higher frequency waves travel more slowly than longer period waves, so high frequency waves contribute to peak tsunami height only while it propagates as a single pulse. After a few hundred kilometers of travel, high frequency waves drift to the back of the wave train (Fig. 5, *left*) and no longer add to the tsunami maximum. At 200 km distance, the shallow earthquake generates a wave about three times larger than the 30-km deep event. If you track the waves out to 2,000 km, however, you would find that the extra high frequencies in the shallow event will have fallen behind and that the maximum wave heights for the two events would be nearly equal. Beyond 2,000 km distance, any earthquake depth less than 30 km appears to be equally efficient in tsunami genesis.

Faults of finite size, like those in Fig. 5, radiate tsunami in distributions more complex than the $\sin\theta$ and $\sin 2\theta$ patterns from point sources. The largest earthquakes have fault lengths of several hundred kilometers. Simulations show that long

Tsunami, Fig. 4 Static vertical displacements of the seafloor for the dip slip (*top*) and strike slip (*bottom*) earthquake point sources that generated the tsunami of Fig. 3. Maximum excursions of the seafloor are 5.4 and 1.4 m, respectively



earthquake faults preferentially emit tsunami in a tight beam perpendicular to the fault strike, regardless of the focal mechanism. This preferential beaming simplifies tsunami forecasting because it tells us which direction to look for the biggest waves.

Tsunami Excitation from Submarine Landslides

Earthquakes parent most tsunami, but other things do too. For instance, earthquake shaking often triggers landslides. If the slide happens under the sea, then tsunami may form. Consider a seafloor landslide confined in a rectangle of length L and width W . Let a constant uplift u_0 start along one width of the rectangle and run down its length ($\hat{x}$ direction, say) at velocity v_r , that is, $\tau(\mathbf{r}_0) = x/v_r$. Placing this $u_z^{\text{bot}}(\mathbf{r}_0)$ and $\tau(\mathbf{r}_0)$ into (7a and 7b), I find the tsunami from this uplift source at observation point $\mathbf{r}$ and time t to be

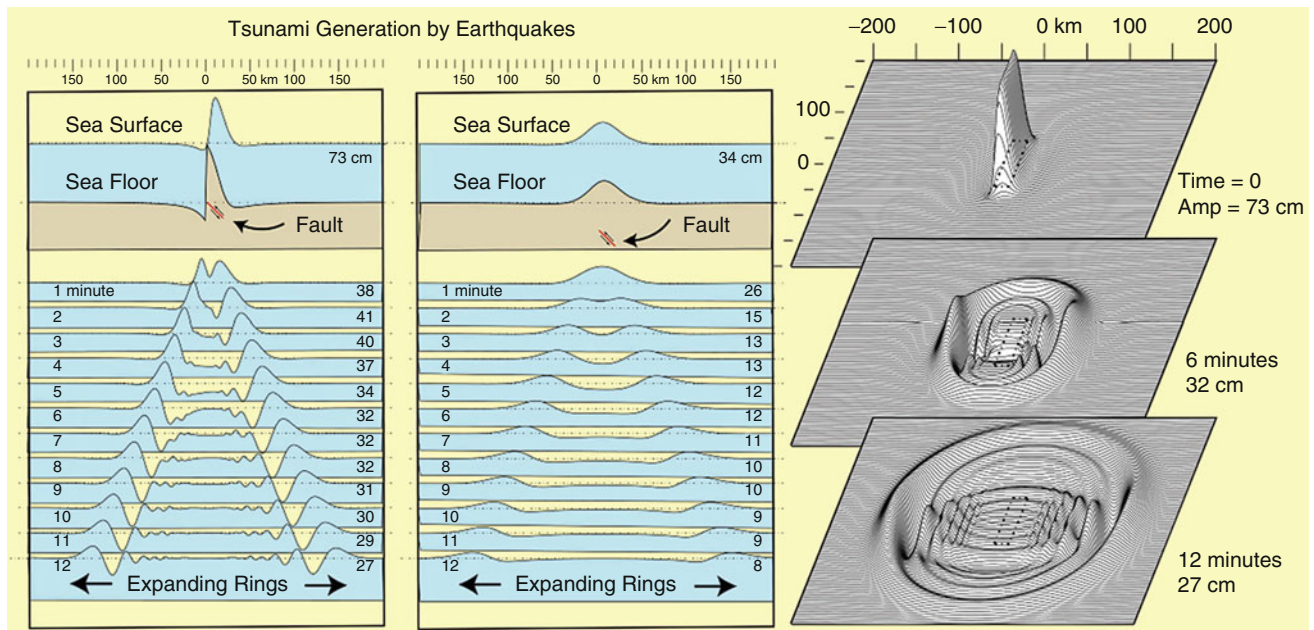
$$u_z^{\text{surf}}(\mathbf{r}, t) = \frac{u_0 L W}{4\pi^2} \operatorname{Re} \int_k \frac{e^{i(\mathbf{k} \cdot \mathbf{r} - \omega(k)t)} e^{-iX(\mathbf{k})}}{\cosh(kh)} \frac{\sin X(\mathbf{k})}{X(\mathbf{k})} \frac{\sin Y(\mathbf{k})}{Y(\mathbf{k})} d\mathbf{k} \quad (12)$$

where

$$X(\mathbf{k}) = \frac{kL}{2} (\hat{\mathbf{k}} \cdot \hat{\mathbf{x}} - c(k)/v_r) \quad \text{and} \quad Y(\mathbf{k}) = \frac{kW}{2} (\hat{\mathbf{k}} \cdot \hat{\mathbf{y}}).$$

The $X(\mathbf{k})$ and $Y(\mathbf{k})$ factors, because they depend on the relative positions of the observation point and the landslide source, instill radiation patterns to the tsunami much like the ε_{ij} do for earthquakes.

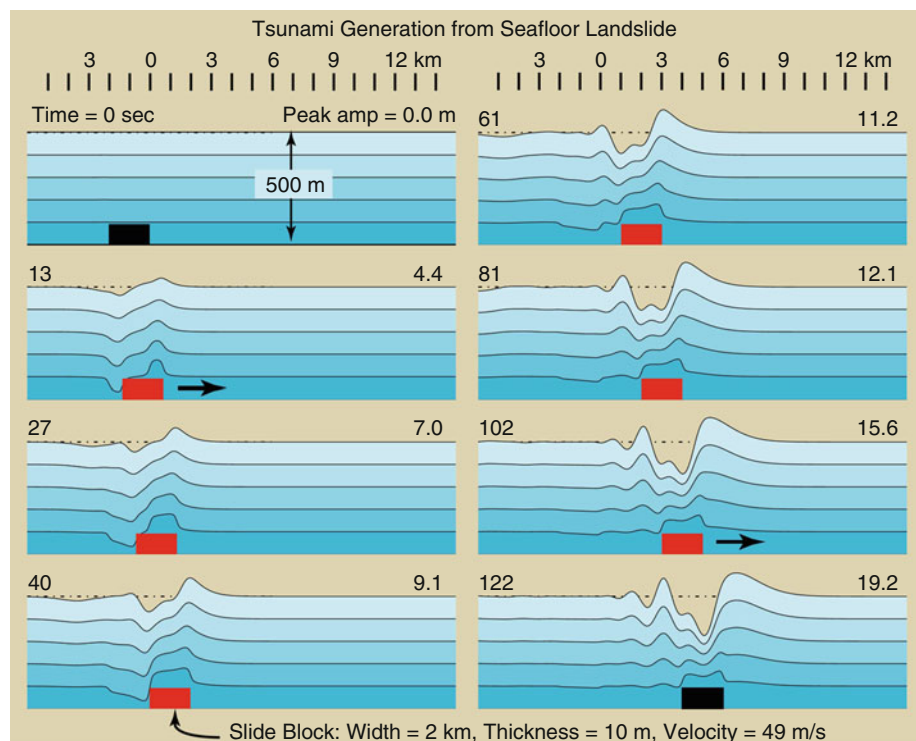
By adding a trailing excavation source to the uplift source (12), I simulate tsunami from a sliding submarine block. Figure 6 pictures in cross section the waves stirred from a 2-km wide block, 10-m thick sliding at 49 m/s for 6 km along the bottom of an ocean 500 m deep. As the block moves, the water must go around the obstruction. Water gets pushed up in front and drawn down behind the slide block. Experiments reveal that, depending on the aspect ratio of the block and the ratio of slide velocity to the tsunami phase velocity, significant beaming and amplification of the tsunami are possible. In particular, when the slide velocity v_r approaches the tsunami speed $c(k) \approx \sqrt{gh}$, $X(\mathbf{k})$ is nearly zero for waves traveling in

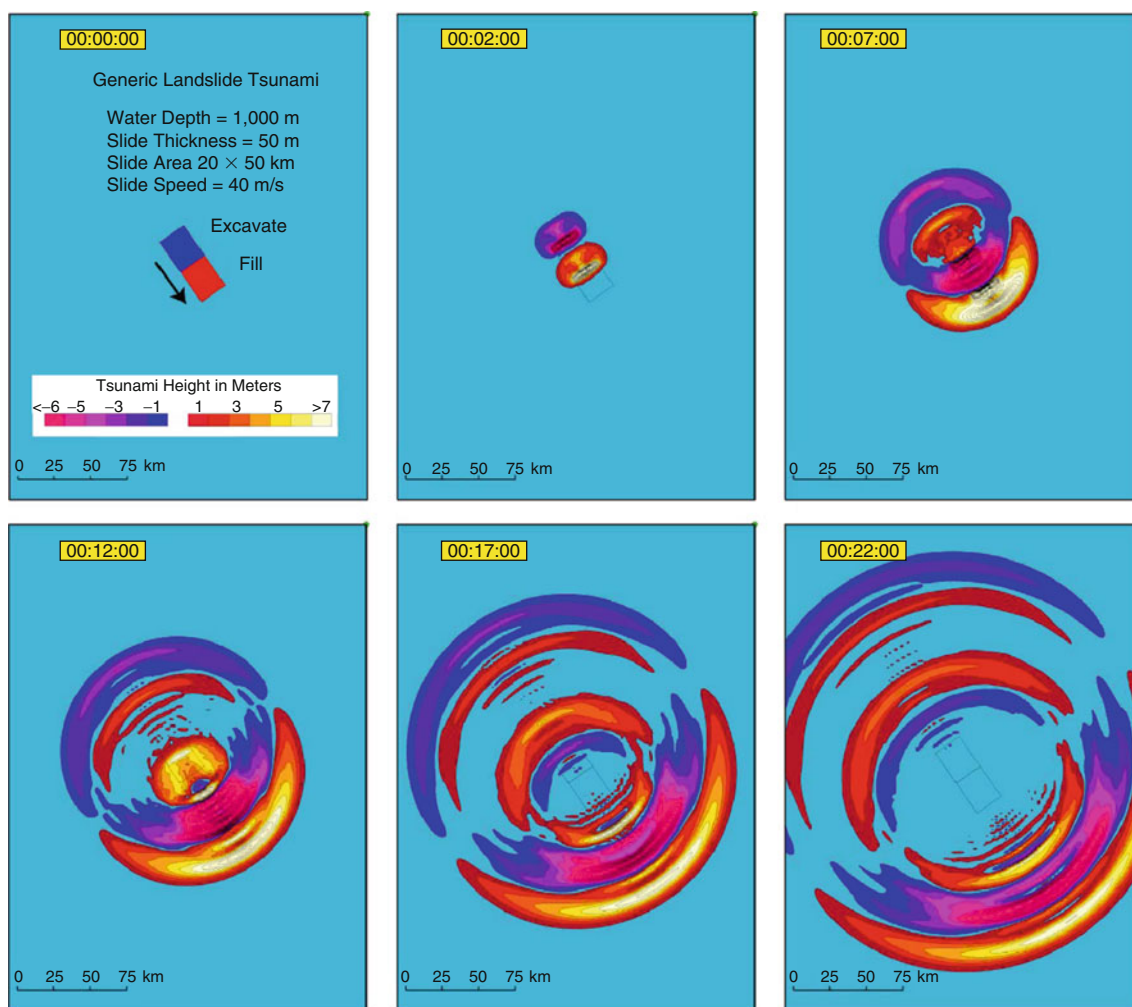


Tsunami, Fig. 5 *Left.* Cross sections of expanding tsunami rings from a M7.5 thrust earthquake. The fault strikes north-south (into the page) and the sections are taken east-west. Elapsed time in minutes and maximum amplitude in centimeters are given at the left and right sides. *Middle.* Cross sections of expanding tsunami rings from the same M7.5 thrust earthquake now buried 30 km. Deeper earthquakes make smaller and

longer wavelength tsunami. *Right.* Views of the expanding tsunami rings from the shallow earthquake case at $t = 6$ and 12 min. The dashed rectangle in the center traces the surface projection of the fault. For large earthquakes, nearly all tsunami energy beams perpendicular to the strike of the fault (toward the left and right in this picture)

Tsunami, Fig. 6 Tsunami produced by a submarine landslide. The slide block is 10 m thick and 2 km wide. The slide starts to the left and runs to the right at 49 m/s. Note the amplification of the wave in the slide direction





Tsunami, Fig. 7 Map view of a tsunami from a landslide block similar to Fig. 6. Note the dipole form of the wave in its early stages and the beaming in the southeast direction of the slide

the slide direction. In this direction, the waves from different parts of the landslide arrive nearly “in phase” and constructively build. Figures 6 and 7 highlight the beaming and amplification effects. In both cases, witness the large tsunami pulse sent off in the direction of the slide. (Quicktime movies of Figs. 6 and 7 and many other Figures in this paper are available for viewing; see Table 2.) Submarine landslides are prime suspects in the creation of “surprise tsunami” from small or distant quakes. Surprise landslide tsunami might initiate well outside of the earthquake uplift area or be far larger than expected given the magnitude of the quake. Historical examples of surprise landslide tsunami include the 1929 Grand Banks and the 1998 Papua New Guinea waves.

Tsunami, Table 2 Quicktime movie links to many of the figures in this article, plus a few others. If you want your own copy: Play the Movie in a Browser, then click “Save to Disk”

Figure 6	http://es.ucsc.edu/~ward/landslide(1_big).mov
Figure 7	http://es.ucsc.edu/~ward/generic-slide.mov
Figure 8	http://es.ucsc.edu/~ward/imp-parab1.mov
Figure 9	http://es.ucsc.edu/~ward/impact-example-3d.mov
Figure 13	http://es.ucsc.edu/~ward/RU(1200s-20m-0.5).mov
Figure 14	http://es.ucsc.edu/~ward/galveston-2m-close.mov
	http://es.ucsc.edu/~ward/galveston-2m.mov
Figure 15	http://es.ucsc.edu/~ward/indo-3D3.mov
	http://es.ucsc.edu/~ward/indo-3D2.mov
Figure 16	http://es.ucsc.edu/~ward/LP-3D.mov
Figure 17	http://es.ucsc.edu/~ward/guinea(270).mov
	http://es.ucsc.edu/~ward/ire-nn.mov
Figure 18	http://es.ucsc.edu/~ward/samoa-92909(c).mov

Tsunami Excitation from Impacts

In addition to tsunami from earthquakes and landslides, another class holds interest – those generated from surface detonations, explosive volcanoes, and asteroid strikes. I call waves from these sources “impact tsunami.” Imagine that the initial stage of cratering by an impact excavates a radially symmetric, parabolic cavity of depth D_C , and radius R_C :

$$u_z^{\text{impact}}(\mathbf{r}) = D_C(1 - r^2/R_C^2) \quad r \leq \sqrt{2}R_C = R_D$$

$$u_z^{\text{impact}}(\mathbf{r}) = 0 \quad r > \sqrt{2}R_C = R_D$$

Based on Eq. 8, the impact tsunami at observation point $\mathbf{r}$ and time t is

$$u_z^{\text{surf}}(\mathbf{r}, t) = \int_0^\infty \frac{k dk}{2\pi} \cos[\omega(k)t] J_0(kr) F_0(k), \quad (13)$$

where

$$\begin{aligned} F_0(k) &= \int_{r_0} r d\mathbf{r}_0 u_z^{\text{impact}}(\mathbf{r}_0) J_0(kr_0) \\ &= \frac{2\pi D_C R_D}{k} J_3(kR_D). \end{aligned}$$

The principal distinction between (8) and (13) is the absence in the latter of the $1/\cosh(kh)$ low-pass ocean filter. Impacts crater the surface of the ocean not the sea floor, so this filter does not come into play.

If we restrict ourselves to asteroid-induced tsunami and suppose that the depth of the impact cavities equals $1/3$ their diameter $d_c = 2R_C$, then D_C relates scales with the density, velocity, and radius of the impacting asteroid as

$$D_C = d_c/3 = (8\varepsilon\rho_1 V_1^2/9\rho_w g)^{1/4} R_1^{3/4} \quad (14)$$

In (14), ε is the fraction of the kinetic energy of the asteroid that goes into the tsunami wave. High velocity impact experiments suggest that $\varepsilon \approx 0.15$. With $\rho_1 = 3 \text{ gm/cm}^3$, $\rho_w = 1 \text{ gm/cm}^3$, and $V_1 = 20 \text{ km/s}$, (14) returns crater depths of 1,195, 2,010, and 4,000 m for asteroids of radius $R_1 = 50, 100$, and 250 m.

Figure 8 plots cross sections of the expanding rings of tsunami waves induced by the impact of a 200-m diameter asteroid at 20 km/s as computed by Eq. 13. Within 100 km of ground zero, tsunami from moderate size (100–250 m) asteroids have heights of many hundreds of meters and dwarf the largest (10–15 m) waves parented by earthquakes. Fortunately, impact tsunami fall on the steep-sloped part of the group velocity curve (Fig. 1) so they decay faster with

distance than earthquake tsunami (more like r^{-1} vs. $r^{-3/4}$ for earthquake tsunami). Figure 9 illustrates how dispersion quickly pulls apart the initial impulse into dozens of waves. At 1,000 and 3,000 km, the tsunami in Fig. 8 would decay to 6 and 2 m amplitude, respectively – still a concern, but not catastrophic. For perspective, asteroids with diameters $>200 \text{ m}$ impact Earth about every 10,000 years, far less frequently than great M9 earthquakes that strike the planet once in 25 years or so.

Tsunami Propagation

In uniform depth oceans, tsunami propagate out from their source in circular rings (e.g., Fig. 9) with ray paths that look like spokes on a wheel. In real oceans, tsunami speeds vary place to place (even at a fixed frequency) so tsunami ray paths refract and become bent. Consequently, in real oceans, both tsunami travel time and amplitude have to be adjusted relative to their values in uniform depth ones. To propagate tsunami in real oceans, I find it best to keep tsunami “mode-like” versus depth but “ray-like” horizontally. I first transform the various integrals over wavenumber (7, 8, 9, etc.) to integrals over frequency because wave frequency, not wave number, is conserved throughout. Using the relations $u(\omega) = d\omega/dk$ and $c(\omega) = \omega/k(\omega)$, I find that tsunami vertical motions from (13) for instance, are to a good approximation:

$$u_z^{\text{surf}}(\mathbf{r}, t) = \int_0^\infty \frac{\cos(\omega t) d\omega}{2\pi k(\omega) u(\omega)} J_0(\omega T(\mathbf{r}, \omega)) F_0(k(\omega)) \cdot \quad (15)$$

$$G(\mathbf{r}) S_L(\omega, \mathbf{r})$$

In (15) the travel time of waves of frequency ω has been changed from $r/c(\omega)$ to

$$T(\mathbf{r}, \omega) = \int_{\text{ray path}} dr/c(\omega, h(\mathbf{r})), \quad (16)$$

where the integral path traces the tsunami ray from the source to the observation point. Equation 15 also incorporates a new shoaling factor $S_L(\omega, \mathbf{r})$ that accounts for wave height changes due to water depth and new ray geometrical spreading factor $G(\mathbf{r}) \leq 1$ that takes into consideration the reduction of wave amplitudes into shadow zones.

Real geometrical rays for tsunami can be very messy (*top left*, Fig. 10). Rather, I employ “network rays” that are simply minimum time paths between source and receiver (*top right*, Fig. 10). Unlike geometrical rays, network rays possess

neither caustics nor shadow zones. To compute the ray geometrical spreading factor $G(\mathbf{r})$, I replace the curvature of the wave front with curvature of the ray in the path integral (Fig. 10, bottom row):

$$G(\mathbf{r}) = \exp \left[-\frac{1}{2} \int_{\text{raypath}} \frac{\partial \alpha(s)}{\partial w} ds \right] \quad (17)$$

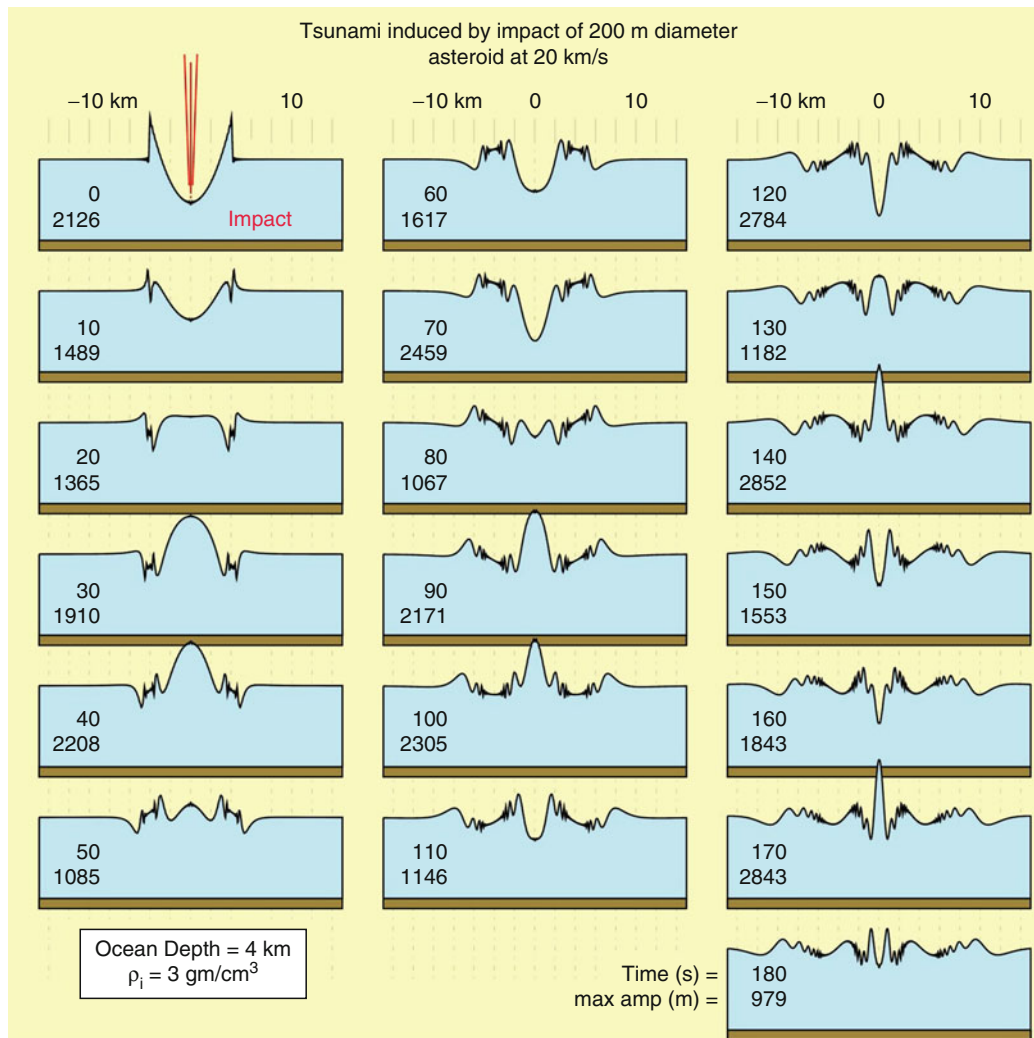
$$\Rightarrow \exp \left[-\frac{1}{2} \int_{\text{raypath}} \left[\frac{\partial \theta(s)}{\partial s} \right] ds \right].$$

In this approach the distinction between wave refraction and diffraction blurs.

Tsunami Shoaling and Run-up

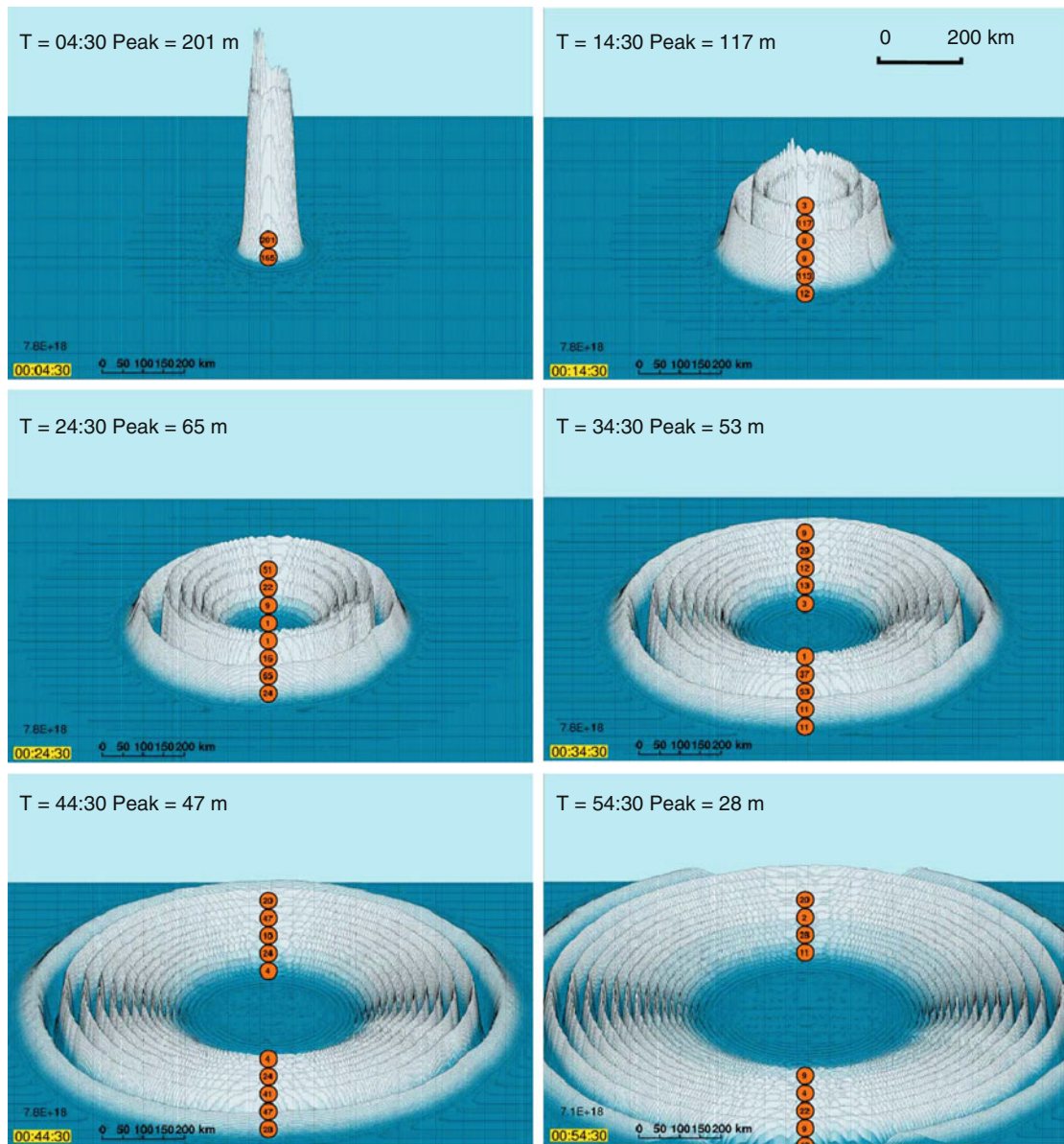
Shoaling. Toward shore, real oceans shallow and the waves carried on them amplify in a process called *shoaling*. Shoaling is easy to understand. The tsunami velocity depends on the ocean depth so as water shallows, tsunami waves slow down. Because their frequency is fixed, the wavelength of a slowing tsunami decreases (Fig. 11, inset) and its energy compresses horizontally. Secondly, because a tsunami occupies the entire water column, as it enters shallow water its energy also gets compressed vertically. The only way for the compressing wave to maintain the same energy flux is for it to grow in amplitude. For the shoaling factor in (15), linear theory gives:

$$S_L(\omega, \mathbf{r}) = \sqrt{\frac{u(\omega, h_0)}{u(\omega, h(\mathbf{r}))}}. \quad (18)$$



Tsunami, Fig. 8 Computed tsunami induced by the impact of a 200-m diameter asteroid at 20 km/s. The waveforms (shown at 10 s intervals) trace the surface of the ocean over a 30 km cross section that cuts rings of

tsunami waves expanding from the impact site at $x = 0$. Maximum amplitude in meters is listed to the left



Tsunami, Fig. 9 Expanding rings of an impact tsunami in a 3D-like style. Note the strong wave spreading due to dispersion

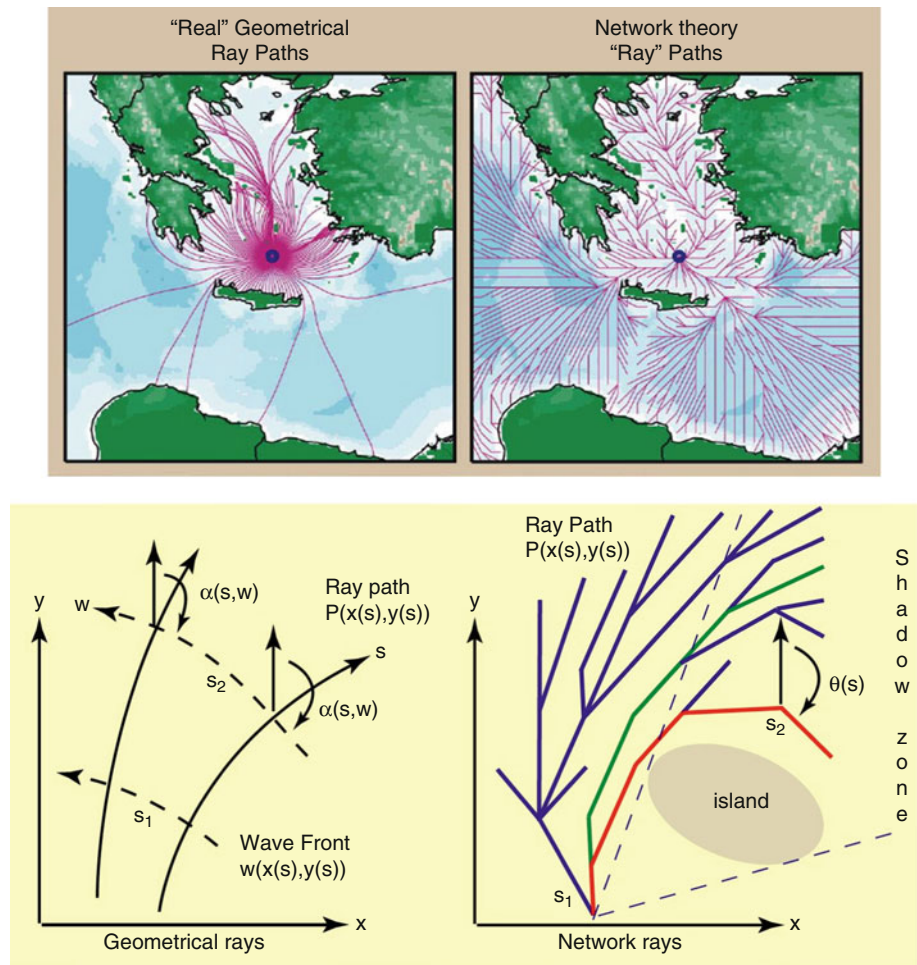
Shoaling amplification depends on the ratio of group velocity at the nucleation-site and the coast-site (ocean depths h_0 and $h(r)$, respectively). As does $G(r)$, S_L reverts to one in oceans of uniform depth.

Figure 11 pictures a shoaling tsunami wave of 150 s period. Initially, a unit height wave comes ashore from 4,000 m of water at the left. As the water shallows, the velocity of the wave decreases and the wave grows in amplitude. By the time it reaches 125 m depth it has slowed from 137 m/s group velocity to 35 m/s and grown in height by a factor of 2. Figure 12 plots (18) as a function of coast-site

depth for sea-wave periods from 10 s, and ocean depths of 2, 4, and 6 km. Because their energy does not occupy the entire water column, beach waves at 10 s period do not amplify much (perhaps 50%) in shoaling. Tsunami waves (100–2,000 s period) experience much stronger shoaling amplification – about 3–6 over a wide range of conditions. For waves of period greater than 250 s, $u(\omega, h_0) = (gh_0)^{1/2}$, $u(\omega, h(r)) = (gh(r))^{1/2}$, and the shoaling factor reduces to Green's Law, $S_L = (h_0/h(r))^{1/4}$ (dashed red line in Fig. 12).

Run-Up. Run-up is the final phase of tsunami life. The run-up phase begins when the approaching tsunami shoals to an

Tsunami, Fig. 10 Real tsunami rays (*top left*) versus network rays (*top, right*). I track the curvature of the ray (*right, bottom*) rather than the curvature of the wave front (*left, bottom*) to account for wave size losses as tsunami propagate around obstacles



amplitude roughly equal to the water depth and the wave begins to break. Run-up also covers the inundation phase where the water runs over land and reaches its maximum excursion above sea level. Run-up is the most complicated phase of tsunami life; nevertheless, “Will I get wet?” is the question everyone wants answered, so tsunami scientists have to take their best shot at carrying the waves to the last mile.

One approach to run-up is wave tank or computer experiments like Figs. 13 and 14, where waves of various sizes and periods are sent onto model beaches. These experiments reveal a wide range of behaviors depending on wave period, wave size, wave direction, number of waves, beach slope, and beach friction. In real-world situations, this information is not known well, so the direct applications of the experimental results are limited. I follow an alternative approach that takes a broad brush view of run-up, but allows for considerable random uncertainty in the outcome. Using the methods above, I propagate tsunami to a position close to shore in shallow water where it can be considered a long wave, but not so shallow that its amplitude $A(h)$ exceeds water depth h there. I estimate run-up height η , on shore by:

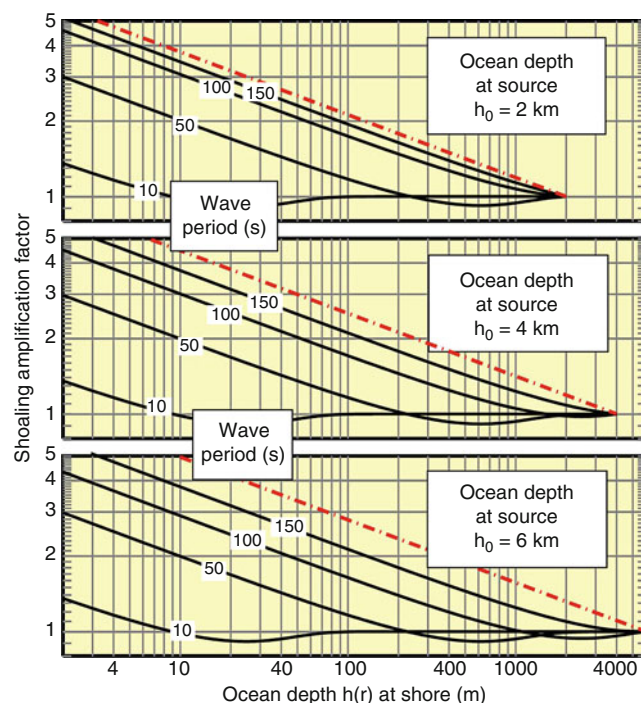
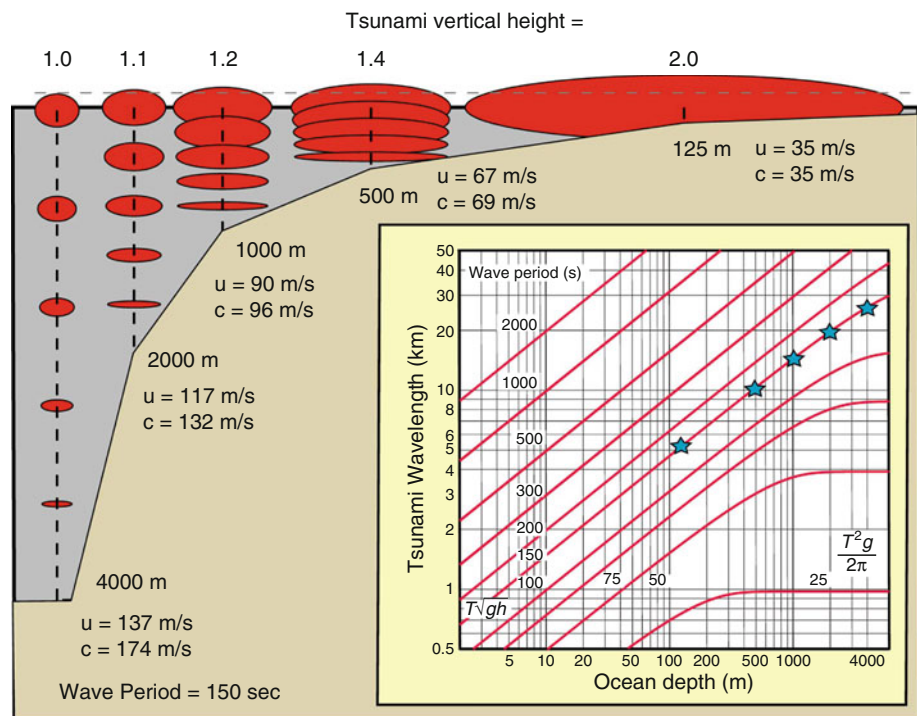
$$\eta = A(h)^{4/5} h^{1/5} \quad (19)$$

Because $A(h) < h$, run-up always exceeds the offshore wave amplitude. Considering the large uncertainties involved, I view run-up at any field site to be a random variable drawn from a distribution function with (19) being its mean and its spread ranging from 1/2 to 2 times the mean. It does not matter at which depth h is used to evaluate (19), provided the water is shallow enough for the tsunami to act like a long wave. According to Green’s Law, a long wave with amplitude $A(h_1)$ in water of depth h_1 , has an amplitude $A(h_2) = A(h_1) (h_1/h_2)^{1/4}$ in water of depth h_2 . Run-up estimate at η_2 at depth h_2

$$\begin{aligned} \eta_2 &= A(h_2)^{4/5} h_2^{1/5} = \left(A(h_1) \left[\frac{h_1}{h_2} \right]^{1/4} \right)^{4/5} h_2^{1/5} \\ &= A^{4/5}(h_1) h_1^{1/5} = \eta_1 \end{aligned}$$

equals estimate η_1 at depth h_1 .

Tsunami, Fig. 11 Effect of shoaling on tsunami eigenfunctions. The shallowing ocean near shore concentrates wave energy into smaller and smaller volumes and tsunami grow in response



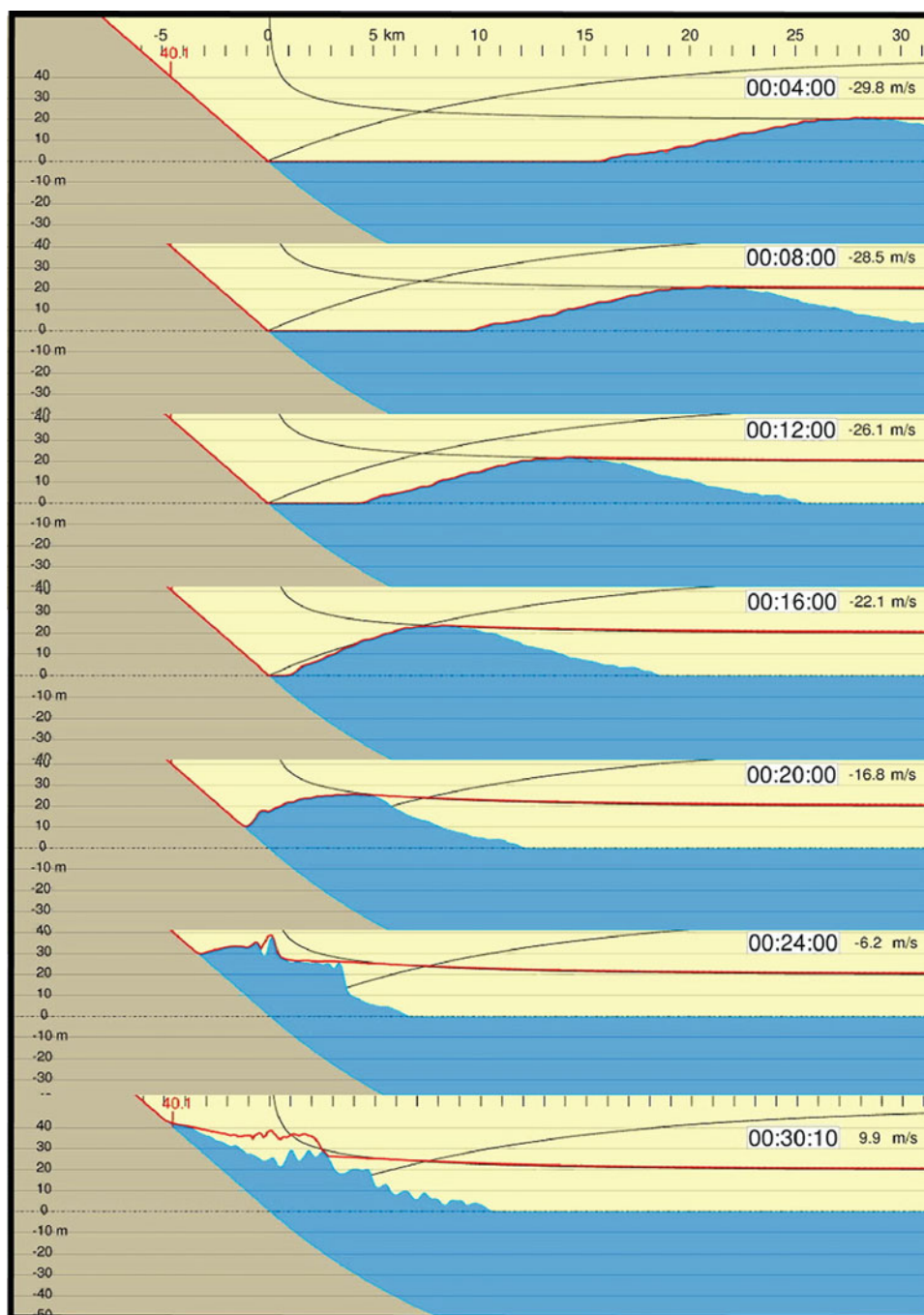
Tsunami, Fig. 12 Shoaling amplification factor for ocean waves of various frequencies and source depths. The dashed red line is Green's Law

Tsunami Samples

Sumatra earthquake 12/26/2004. Figure 15 snapshots a simulation of the 2004 Sumatra tsunami. As modeled by several rectangular patches, the megathrust fault that sourced the wave trends north–south for several 100 km (Fig. 15 top left). As the example in Fig. 5 predicted, nearly all of the wave energy in the 2004 tsunami emits perpendicular to the fault – Sumatra and Thailand toward the east, and India and Sri Lanka toward the west. Not surprisingly, these locations suffered the lion's share of tsunami damage. I believe that the very large run-up near Banda Aceh was due to concentrated slip on a thrust ramp splaying up from the megathrust just offshore at that latitude.

La Palma landslide tsunami. Like its fellows in the Canary Island chain, the volcanic Island of La Palma has experienced cycles of growth and partial collapse. Its flank last failed 550,000 years ago. Since then, new Cumbre Vieja volcano has filled the vacant scar. For the last 125,000 years this volcano grew steadily but in the last 7,000, its structure changed. It now mainly erupts along a north–south trending rift that splits the mountain in half. Moreover, during an eruption in 1949 a fault broke surface along the crest of the volcano and part of its western side slid 5 m down and toward the ocean. The volcano again may be showing initial stages of instability. While certainly collapse is not imminent and it

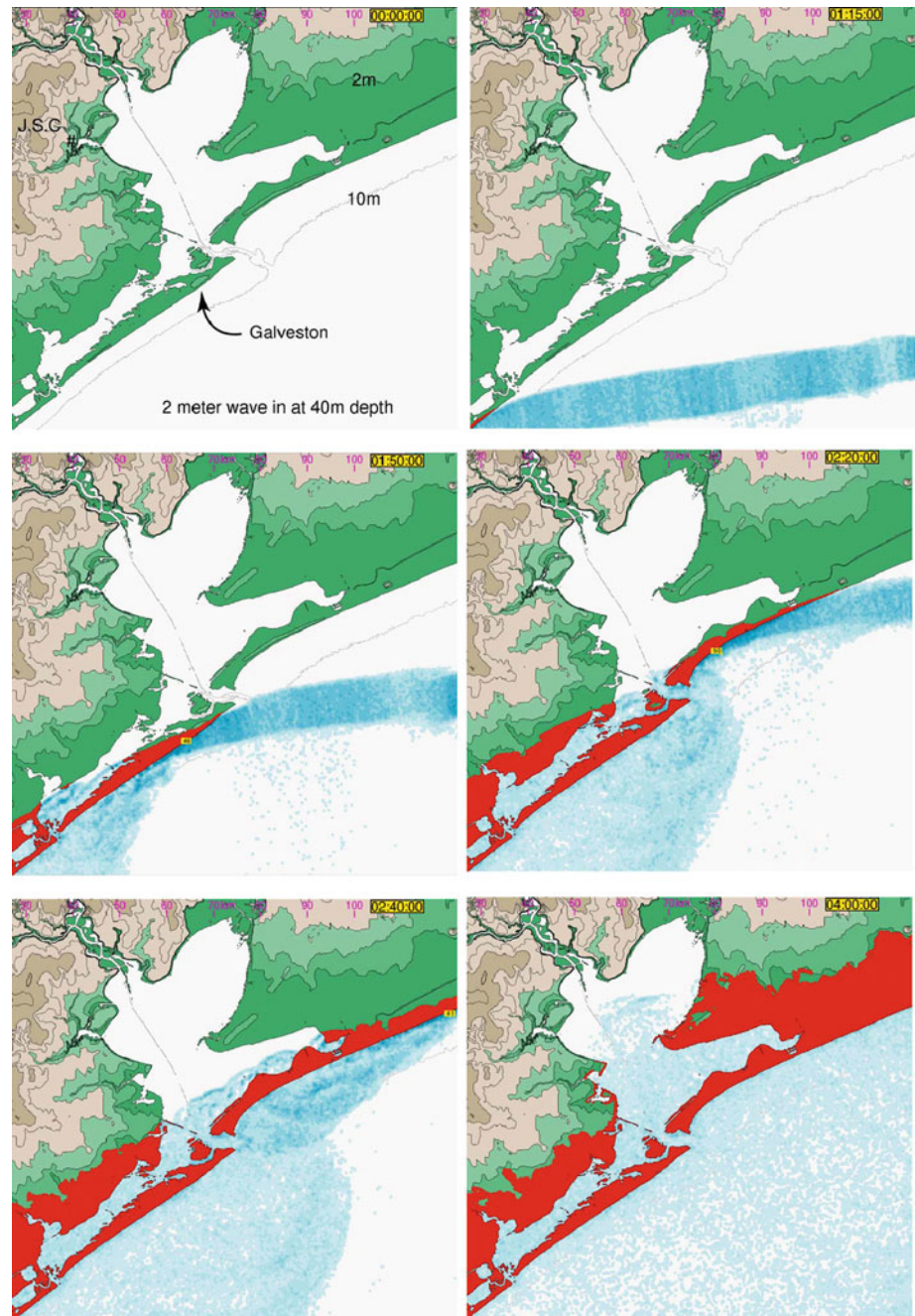
Tsunami, Fig. 13 Run-up and inundation of a 20 m high, 1,200 s duration wave on a steep 0.5° slope. On steep beaches breaking has little time to operate so run-up amplifications are large



may take many eruptive cycles over the next few thousand years to give it that final shove, the volcano will collapse so it is sensible to consider the consequences (Fig. 16). If 500 km³ of its flank falls into the sea like it did last time, a La Palma collapse will spawn a mega tsunami hundreds of meters high locally that will buffet all coasts of the Atlantic with 10–30 m waves.

Gulf of Guinea impact tsunami. In a 2008 training exercise, the US Air Force considered the consequences of, and response to, an impact of a small comet into the Atlantic Ocean off the coast of Guinea. In this scenario, pictured in Fig. 17, tsunami run-ups exceed 10 m along a 1,200 km stretch of coastline. This wave would knock out coastal shipping and oil production facilities for many months and trigger major ripple effects in

Tsunami, Fig. 14 Simulation of 2 m tsunami wave beaching at Galveston Texas. Red area blankets the predicted inundation zone. Note the multiple wave reflections and wave interference that contribute to run-up complexity and randomness

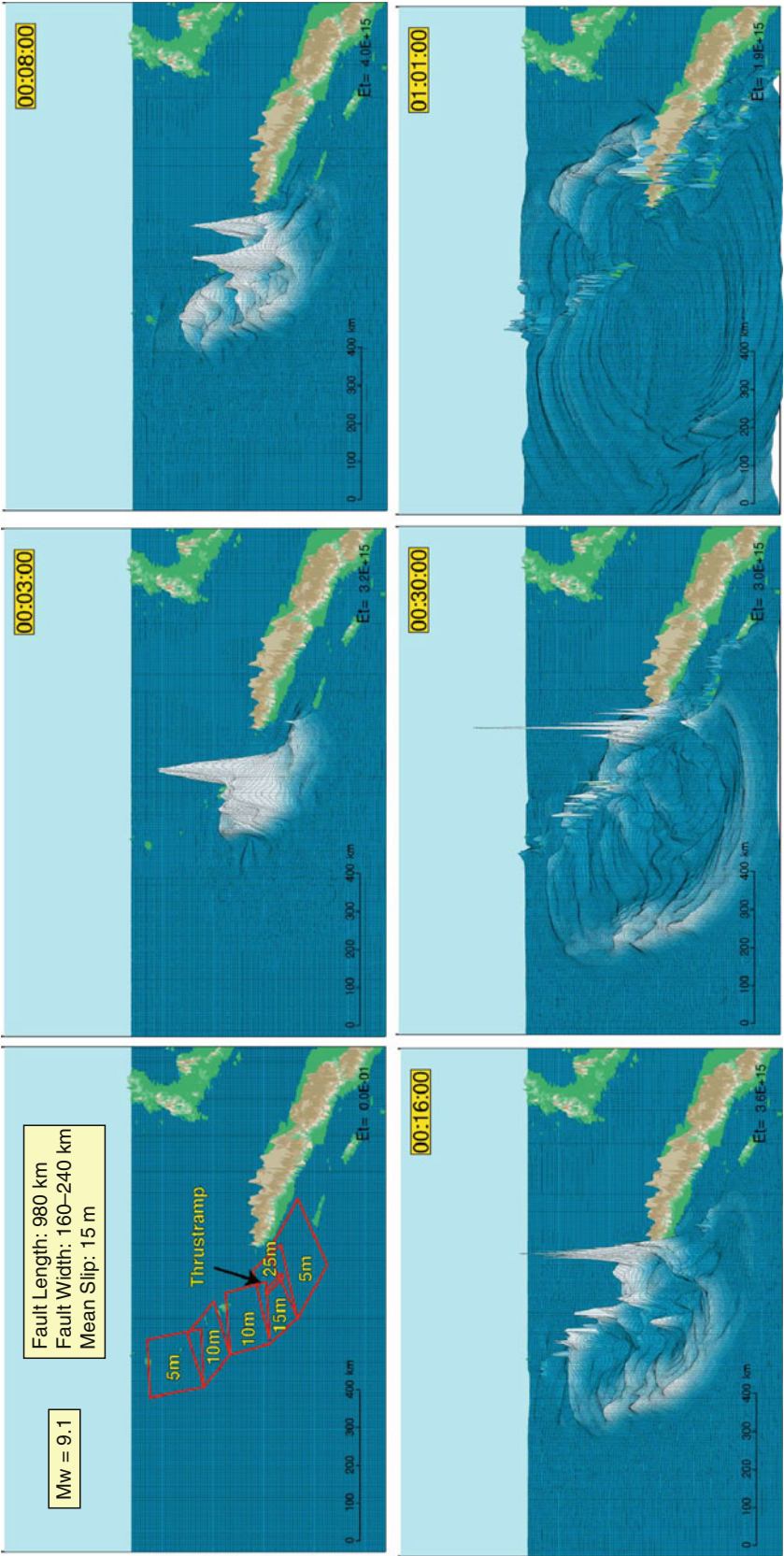


the global economy. Unlike those Hollywood movies, currently there is no defense against an asteroid impact.

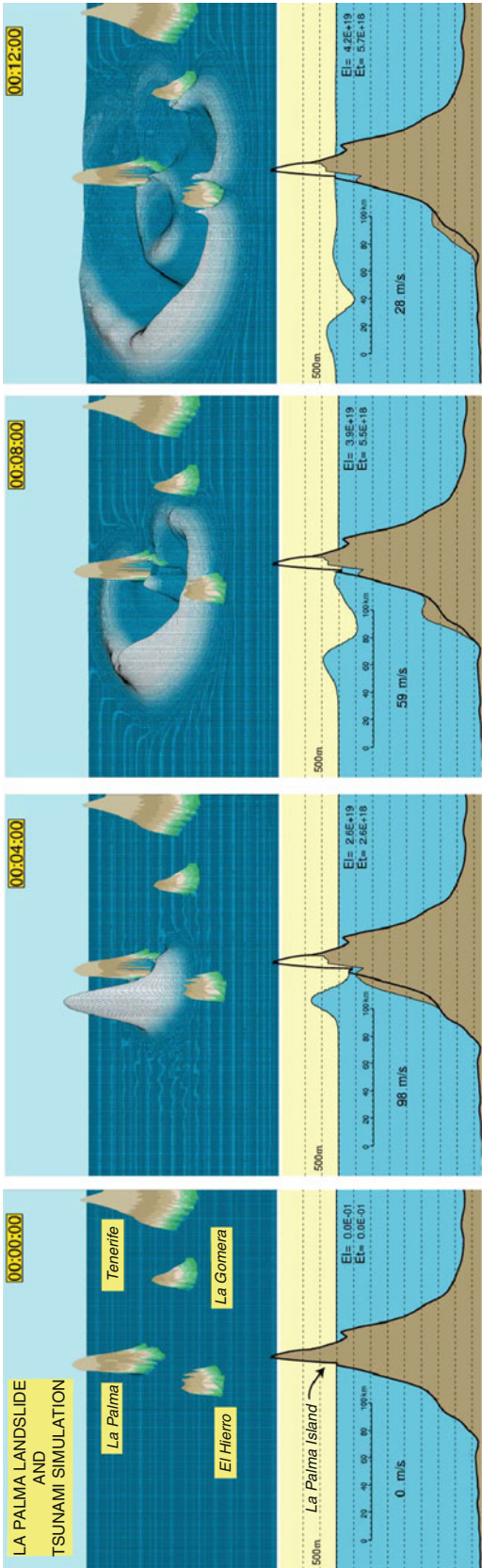
Tsunami Forecasting

Official approach: Perhaps the ultimate goal of tsunami research is forecasting. A forecast predicts the size of a particular tsunami given the knowledge that a potentially dangerous earthquake has occurred already. Tsunami travel

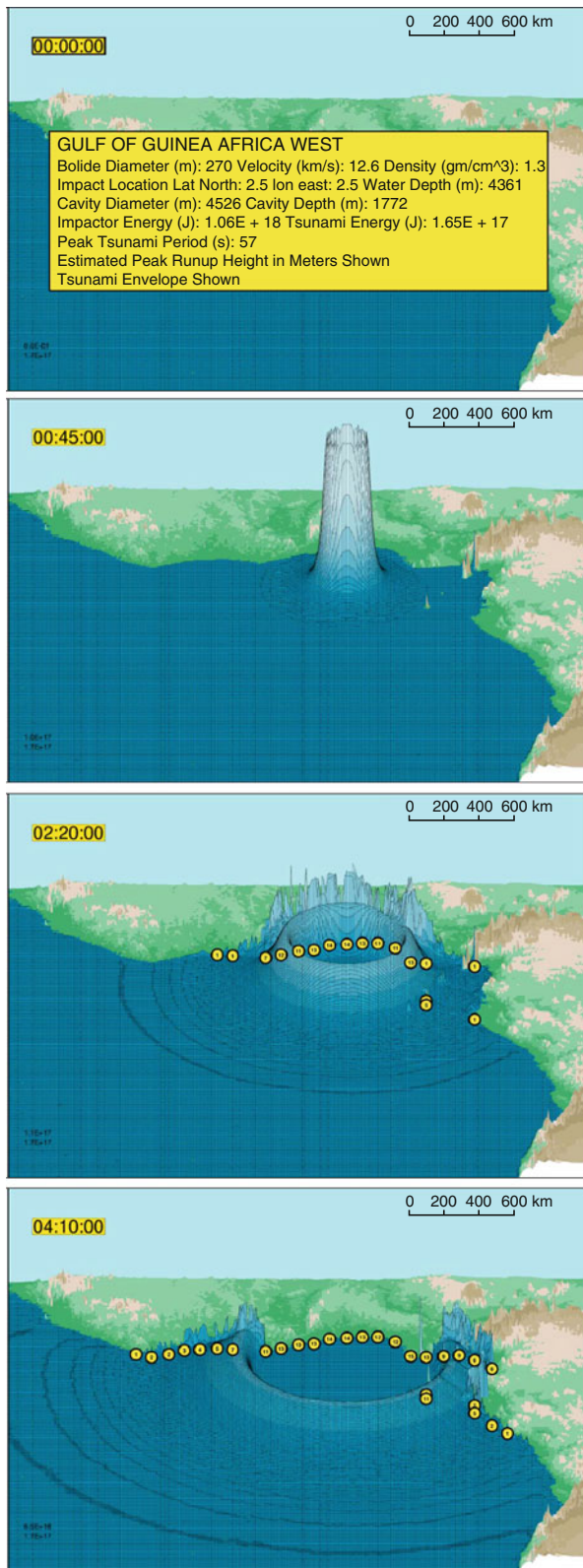
at jet speeds, but for many places there may be several hours between the earthquake and the arrival of the sea waves. This time can be spent analyzing seismograms, estimating earthquake parameters, and forecasting the expected height of the oncoming wave with the aid of computer models and initial field reports. This is precisely the mission of the various warning centers like NOAA's Pacific Tsunami Warning Center (PTWC). The past decade has witnessed considerable progress in tsunami forecasting coupled with faster and more geographically selective warnings. By and large, the



Tsunami, Fig. 15 Simulation of the Sumatra Earthquake Tsunami of December 2004. Note the large wave directed toward the northern tip of Sumatra due to the small thrust ramp offshore there
(upper left)



Tsunami, Fig. 16 (Top row) First 12 min of the La Palma Tsunami. Initial wave heights exceed 500 m. (Bottom row) Landslide and tsunami in cross section. Just like the simple example in Fig. 6, water gets pushed up in front of the slide and drawn down behind



Tsunami, Fig. 17 Guinea Impact Tsunami. Four-hour sequence of waves from the impact of a 270 m diameter icy bolide off the west coast of Africa

current tsunami prediction system functions well. Figure 18 pictures a nearly real-time prediction of tsunami run-up from the M8.1 Samoa Earthquake of September 29, 2009. Based on rapid calculations like these, the PTWC issued accurate advanced warnings and watches for this tsunami – at least for locations greater than 2,000 km distance from the quake where sufficient delay time existed.

The biggest uncertainty in the current procedures lay in fixing earthquake parameters quickly. For instance, in the first hours after a quake, it is not uncommon for estimates of earthquake magnitude to vary 0.3 or 0.4 units as more extensive seismic data arrive. Because earthquake magnitude or earthquake moment is the prime driver for tsunami (Eq. 9), a 0.4 unit uncertainty makes for a factor of 2 or 3 uncertainty in tsunami height all else aside. Partly on this concern, the DART buoy network that provides real-time measurements of passing tsunami was expanded from 6 to 24 sites after the Sumatra earthquake. Information from the DART sensors provides a real-time “reality check” on computer simulations run rapidly on uncertain earthquake parameters.

Do-it-yourself approach: Apart from official predictions, readers of this article have enough information to produce a Do-It-Yourself tsunami forecast. The three step recipe needs: earthquake magnitude M , the ocean depth at source H_o , distance to the quake R , and mean ocean depth $\bar{H}$ between the earthquake and you.

Step 1 – Find Initial Tsunami Amplitude A_0

Let us approximate initial tsunami height by:

$$A_0 = \frac{\alpha \Delta u(M)}{\cosh \left[\frac{4\pi H_o}{W(M)+L(M)} \right]}, \quad (20)$$

where α is the fraction of earthquake slip that transforms into tsunami-making uplift:

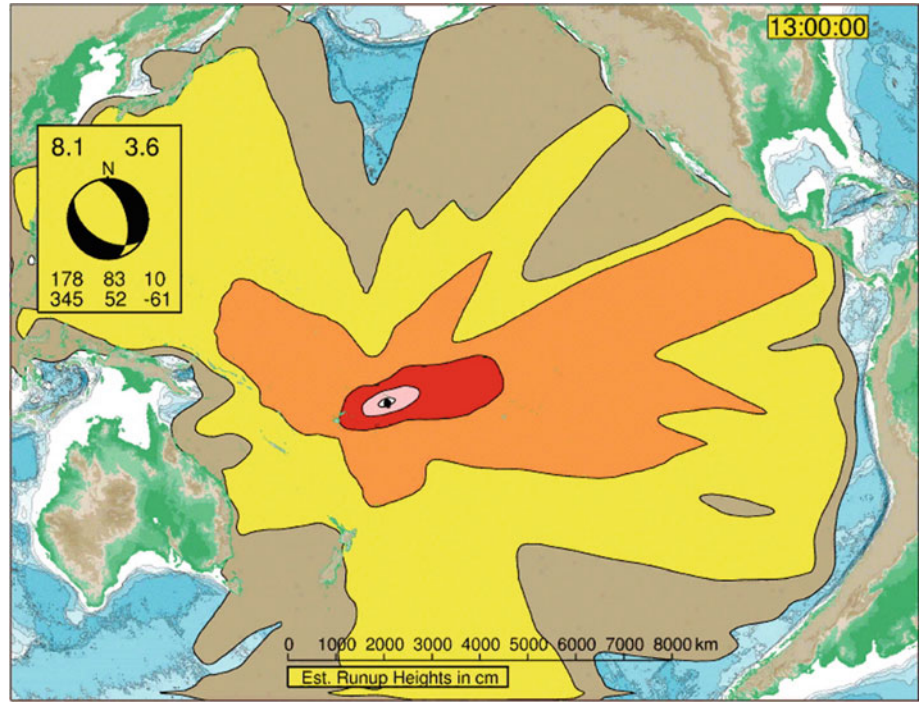
$$\alpha = (1 - \theta/180) \sin \theta \mid \sin \phi \mid,$$

with θ and ϕ being the fault dip and rake angles in degrees. Fault slip $\Delta u(M)$, width $W(M)$, and length $L(M)$ are functions of magnitude and can be read from Table 1 or computed from the formulas in its caption. You recognize the $1/\cosh$ as the low-pass filter due to the ocean layer. Unlike Eqs. 9 and 20 makes no assumption about radiation pattern, so in this sense it represents the worst case. The most efficient mechanism (surface breaking, pure dip slip on planes dipping near 45°) for tsunami generation gives $\alpha = 0.55$.

Step 2 – Correct for Propagation Loss P

Tsunami waves decay as they travel due to geometrical spreading and frequency dispersion. For an ocean of constant depth $\bar{H}$, the propagation loss is roughly.

Tsunami, Fig. 18 Nearly real-time prediction of run-up height from the 9/29/09 M8.1 Samoa Earthquake. Colors correspond to run-up heights: *Brown* 10–25 cm; *Yellow* 25–50 cm; *Orange* 50–100 cm; *Red* 1–2 m; *Pink* > 2 m



$$P = \left(1 + \frac{2R}{L(M)}\right)^{-\psi}, \quad (21)$$

where

$$\psi = 0.5 + 0.575 \exp\left(-0.0175 \frac{L(M)}{\bar{H}}\right). \quad (22)$$

The first term in exponent (22) accounts for square root of distance losses due to geometrical spreading. The second term in (22) accounts for additional losses due to frequency dispersion with larger dimensioned tsunami decaying more slowly. Peak deep-water tsunami amplitude $A_d(R)$ at distance R is

$$\begin{aligned} A_d(R) &= A_0 P \\ &= \frac{\alpha \Delta u(M)}{\cosh[4\pi H_0 / (W(M) + L(M))]} \left(1 + \frac{2R}{L(M)}\right)^{-\psi} \end{aligned} \quad (23)$$

Propagation loss (23) supposes an unobstructed wave path to the observation location. If the waves pass around headlands, squeeze through narrows, or cut into shadow zones, then (23) again overestimates tsunami size.

Step 3 – Correct for Shoaling and Run-Up

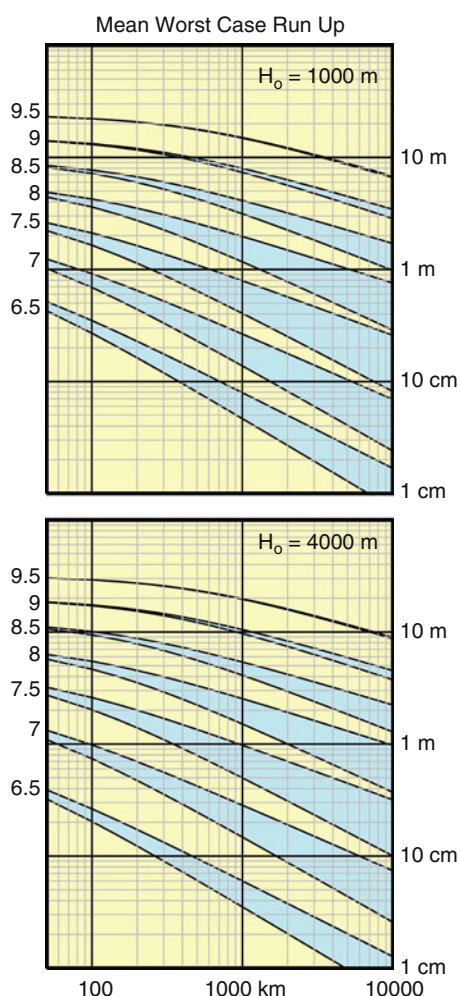
For long waves, the shoaling correction S_L (18) reduces to Green's Law $S_L = (H_0/H)^{1/4}$ and the shoaled tsunami amplitude $A_S(R)$ at distance R is

$$\begin{aligned} A_S(R) &= A_0 P S_L \\ &= \frac{\alpha \Delta u(M)}{\cosh\left[\frac{4\pi H_0}{W(M) + L(M)}\right]} \left(1 + \frac{2R}{L(M)}\right)^{-\psi} \left[\frac{H_0}{H}\right]^{1/4}. \end{aligned} \quad (24)$$

Applying (19) on (24) returns peak wave run-up height $\eta(R)$

$$\begin{aligned} \eta(R) &= [A_0 P S_L]^{4/5} H^{1/5} \\ &= \left[\frac{\alpha \Delta u(M)}{\cosh\left[\frac{4\pi H_0}{W(M) + L(M)}\right]} \left(1 + \frac{2R}{L(M)}\right)^{-\psi} \right]^{4/5} H_0^{1/5}. \end{aligned} \quad (25)$$

Figure 19 plots Do-It-Yourself prediction (25) for earthquake magnitudes 6.5–9.5, at distances R to 10,000 km, for $\bar{H}$ between 1,000 and 4,000 m, and $H_0 = 1,000$ m (*top panel*) and $H_0 = 4,000$ m (*bottom panel*). Because several steps above selected extreme assumptions, these curves represent a mean worst case. The worst worst case might be twice as large still. For $M = 8$, $H_0 = \bar{H} = 4,000$ m, Fig. 19 (*bottom*) predicts run-ups of 3 m at 300 km distance, 2 m at 600 km distance, 1 m at 2,000 km distance, and 0.5 m at 5,000 km distance. These numbers compare well with the more rigorous calculations in Fig. 18.



Tsunami, Fig. 19 Mean worst-case tsunami run-up height versus distance from earthquakes of magnitude 6.5–9.5. The blue areas include an allowance for mean ocean depth between 1,000 (*upper limit*) and 4,000 m (*lower limit*). Top and Bottom panels assume 1,000 and 4,000 m water depth at the source

Summary

Tsunami are gravity-driven water waves caused mostly by undersea earthquakes, but sometimes they are sourced by submarine landslides, surface detonations, explosive volcanoes, and asteroid strikes. Tsunami differ from ordinary beach waves in that they have longer period, larger wavelength, and higher velocity. Because of shoaling effects, these characteristics make tsunami especially damaging because they grow in height proportionally more in approaching shore than do ordinary waves of equal size. This article provides mathematical expressions for tsunami excitation, propagation, and simplified steps to compute wave run-up height from tsunamigenic earthquakes.

Cross-References

- [Earthquake Source Theory](#)
- [Earthquakes and Crustal Deformation](#)
- [Seismic Ray Theory](#)
- [Surface Waves](#)
- [Tsunami Watch and Warning Centers](#)
- [Tsunami: Bay of Bengal](#)

Tsunami Watch and Warning Centers

Shailesh Nayak¹, Srinivasa Kumar Tummala^{2,3} and E. Pattabhi Rama Rao²

¹National Institute of Advanced Studies (NIAS), Bengaluru, India

²Earth System Science Organisation–Indian National Centre for Ocean Information Services (ESSO–INCOIS), Hyderabad, India

³ICG/IOTWMS Secretariat, UNESCO IOC Perth Regional Programme Office, West Perth, WA, Australia

Definition

Tsunami. Japanese word literally meaning harbor wave (tsu + nami). Tsunamis are series of ocean gravity waves caused mainly due to undersea earthquakes and undersea volcanic activity, rarely due to submarine landslides and very infrequently due to meteor impacts.

Tsunamigenic earthquake. Large earthquake occurring mostly in subduction zones capable of generating tsunami.

National Tsunami Warning Center (NTWC). A center officially designated by the government for collating and disseminating tsunami warnings/advisories and other related information within its country according to established national Standard Operating Procedure (SOPs).

Tsunami Service Provider (TSP). A center that monitors seismic and sea level activity and issues timely tsunami forecast information within an Intergovernmental Coordination Group (ICG) framework to NTWCs/TWFPs and other TSPs operating within an ocean basin. TSPs may also act as NTWCs providing tsunami warnings for their own countries.

Tsunami Warning Focal Point (TWFP). A 24 × 7 point of contact (office, operational unit or position, not a person) officially designated by the NTWC or the government to

Srinivasa Kumar Tummala is responsible for the choice and the presentation of the facts contained in the publication and for the opinions expressed therein, which are not necessarily those of UNESCO and do not commit the Organization

receive and disseminate tsunami information from a TSP according to established national SOPs.

Tsunami National Contact (TNC). The person designated by a Member State to an ICG to represent his/her country in the coordination of international tsunami warning and mitigation activities.

Tsunami Bulletin. A real-time situation report of the event.

Service Level. Level of operational capability in accurately advising/warning target populace.

Introduction

The 2004 Sumatra-Andaman earthquake and subsequent tsunami in the Indian Ocean were a grim reminder of nature's fury. The event prompted many countries to set up tsunami warning centers (TWC), which were further integrated into a collaborative network of regional tsunami warning systems in the Indian Ocean and Caribbean and Mediterranean Seas by the Intergovernmental Oceanographic Commission (IOC) of the United Nations Educational, Scientific and Cultural Organization (UNESCO). These regional systems, along with the Pacific Tsunami Warning System (PTWS), form a globally integrated system, built on three pillars: tsunami risk assessment and mitigation; tsunami detection, warning, and dissemination; and tsunami awareness and response.

An important fact that is well known about tsunami is that they travel very fast in open ocean (velocity of the order of several hundreds of kilometers per hour) and cannot be discernible easily by wave heights (of the order of a few centimeters), but slow down as they approach the coast (of the order of several tens of kilometers per hour) with wave heights increasing to disastrous proportions (of the order of several meters). Historically, coastal populations would depend on traditional knowledge and natural signs such as ground shaking or water withdrawing from the beach to identify an impending tsunami. While rapid increase in coastal population has increased risk in terms of loss of lives and economic damages (socioeconomic parameters) many fold, advancement of science has made it possible to mitigate the impacts through forewarning. With appropriate sensor networks, modeling, geospatial technologies, and communication infrastructure, such forewarning systems can detect and monitor tsunamis and provide timely advice to coastal populace at risk. Such a scientific establishment is known as Tsunami Warning Center (TWC). The mission of a TWC is to provide early warnings on potentially destructive tsunamis to emergency/disaster management officials, and as appropriate, directly to the public to prevent loss of life and property. This entry briefly describes TWC, its components, the products that are generated, and level of operational service that these centers are designed for.

A typical TWC comprises observational networks (seismic and sea level) for detection of earthquakes and evaluation of potential for tsunami; modeling for generation, propagation of tsunami wave and inundation; database of bathymetry and coastal topography; computational and communication facilities for data acquisition in real time, analysis, and modeling; decision support systems (DSS) based on standard operating procedure (SOP) for generation and dissemination of tsunami advisories (Nayak and Kumar 2011). Based on the capability and the Area of Service, a TWC can be categorized as a National Tsunami Warning Center (NTWC) or a Tsunami Service Provider (TSP). Within India, the Indian Tsunami Early Warning Centre (ITEWC) was established in 2007 at the Earth System Science Organization (ESSO) – Indian National Centre for Ocean Information Services (INCOIS), Hyderabad (Gupta 2005; Nayak and Kumar 2008a). Along with similar centers in Australia and Indonesia, the ITEWC also serves as one of the tsunami services providers (TSP) in the IOC-coordinated Indian Ocean Tsunami Warning and Mitigation System (IOTWMS).

Types of Tsunamis

Several terms are in use for describing tsunamis: historical tsunami, local tsunami, maremoto, meteorological tsunami (meteotsunami), microtsunami, ocean-wide tsunami, paleotsunami, regional tsunami, and teletsunami (distant tsunami) (UNESCO-IOC 2019). However, from the early warning perspective, three types are considered most relevant, based on the distance from the source and the tsunami travel time. They are the following.

1. Local tsunami originates from a nearby source for which its destructive effects are confined to coasts less than 1 h tsunami travel time, or typically within about 200 km from its source. A local tsunami is usually generated by an earthquake, but can also be caused by a landslide or a pyroclastic flow from a volcanic eruption. Over history, 90% of tsunami casualties have been caused by local tsunamis. These can often be the most dangerous because there is often little warning between the causing event and the arrival of the tsunami.
2. Regional tsunami is capable of creating destruction in a particular geographic region, generally within 1000 km or 1–3 h tsunami travel time from its source.
3. Distant tsunami (also called an ocean-wide, distant, tele, or far-field tsunami) is a tsunami that originates from a far-away source, which is generally more than 1000 km away from the area of interest or more than 3 h tsunami travel time from its source. A distant tsunami is capable of causing widespread destruction, not only in the immediate

region of its generation but across an entire ocean. All known ocean-wide tsunamis have been generated by major earthquakes in the subduction regions.

Regional Tsunami Warning Systems

There are four major tsunami warning systems operating today. They are Pacific Tsunami Warning Mitigating System (PTWS); Indian Ocean Tsunami Early Warning and Mitigation System (IOTWMS); Tsunami Early Warning and Mitigation System in the North-Eastern Atlantic, Mediterranean and Connected Seas (NEAMTWS); and Tsunami and other Coastal Hazards Warning System for the Caribbean and Adjacent Regions (CARIBE-EWS). These regional systems operate as “system-of-systems” through their respective Intergovernmental Coordination Groups (ICGs) that are governed by IOC-UNESCO.

The core regional tsunami warning framework comprises a network of NTWCs or Focal Points that receive tsunami forecast information from one of more TSPs. It is the NTWCs/TWFPs, operating within the legal framework of the sovereign nation in which they reside and serve, that provide warnings, watches, and advisories to their citizens. These warnings are based either on the NTWC’s own analysis of the situation, or on the forecast information received from TSPs, or on a combination of all. The TSPs operating within an ocean basin are interoperable not only within themselves but also with other ocean basin TSPs, by using common and agreed formats and procedures. For instance, the IOTWMS comprises NTWCs/TWFPs in 28 countries bordering the Indian Ocean with three centers acting as the TSPs – Indian Tsunami Early Warning Centre (ITEWC), Indonesian Tsunami Early Warning System (InaTEWS), and Joint Australian Tsunami Warning Centre (JATWC). Other TSPs in the IOC-UNESCO framework operating in other ocean basins include French National Tsunami Warning Centre (CENALT), Institute of Geodynamics, National Observatory of Athens, Kandilli Observatory and Earthquake Research Institute, National Institute of Geophysics and Volcanology (INGV), Instituto Português do Mar e da Atmosfera (IPMA), Pacific Tsunami Warning Center (PTWC), and the Northwest Pacific Tsunami Advisory Center (NWPTAC). Figure 1 shows the area of coverage of the four regional systems and area of service of each TSP.

Key Components of a Tsunami Warning Center

In the following sections, key components and procedures followed by a Tsunami Warning Center are described with reference to the Indian Tsunami Early Warning Centre as an example. It may be noted that while the key components

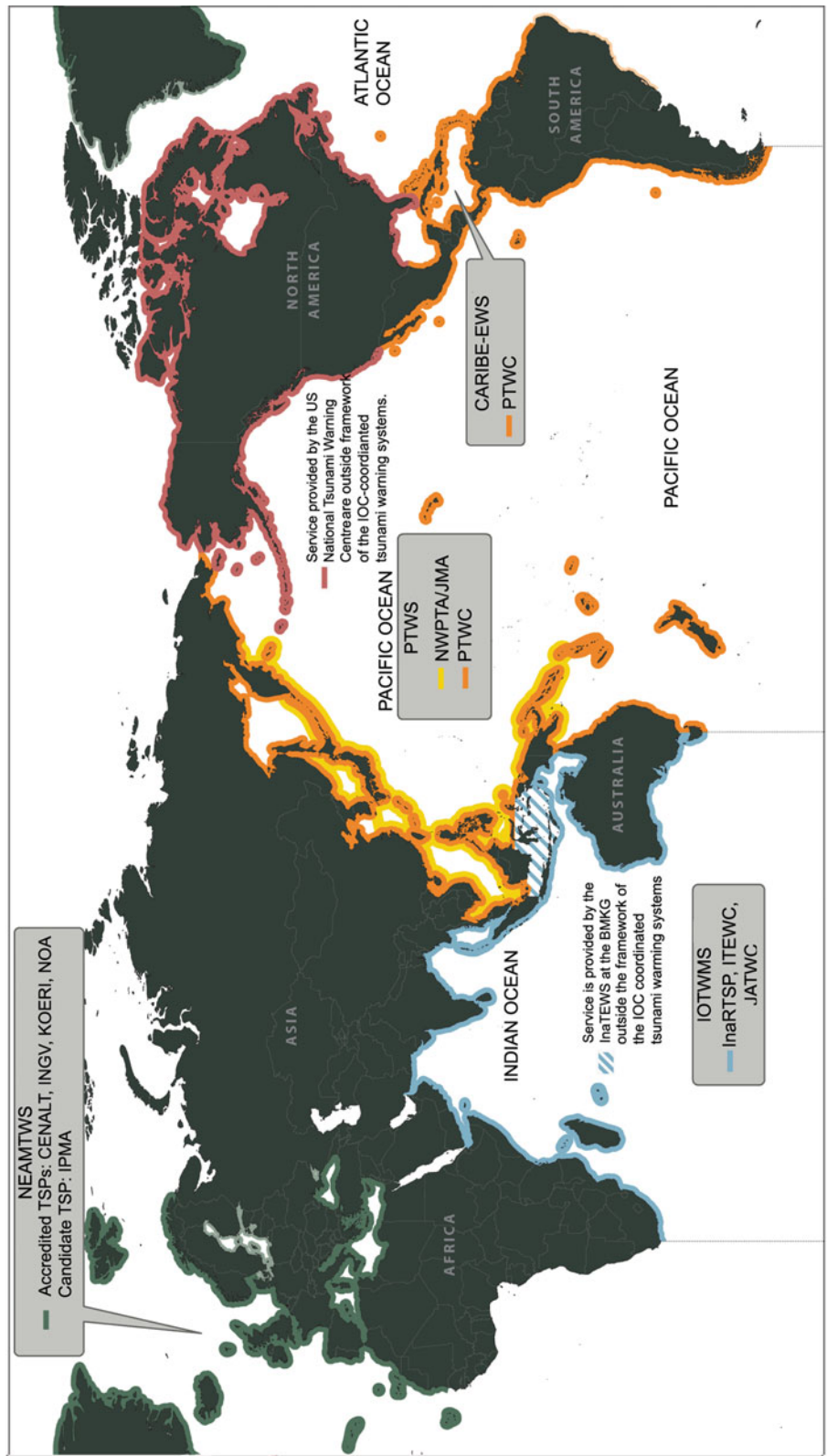
broadly remain the same among different TWCs, procedures and products might vary to suit national or regional requirements. Further, the key elements integrated in a TWC become the essential and minimum requirements in dealing with other ocean-related hazards like: cyclone-related storm surges, and impacts of long-term coastal erosion and sea level rise.

In order to carry out its mission, a TWC must at least consist of five important components: a network of sensors (seismic and sea level) to detect earthquakes and evaluate its potential for tsunamis; tsunami modeling capability (generation, propagation, and inundation models); high-resolution database on bathymetry and coastal topography; communication infrastructure to receive data from observation platforms and issue timely dissemination to stakeholders to facilitate evacuation of coastal areas, and a 24 × 7 operational facility housing the infrastructure, decision support system (DSS) based on standard operating procedures (SOP), and experienced staff (Nayak and Kumar 2008a). A typical outlay of a warning center is shown in Fig. 2.

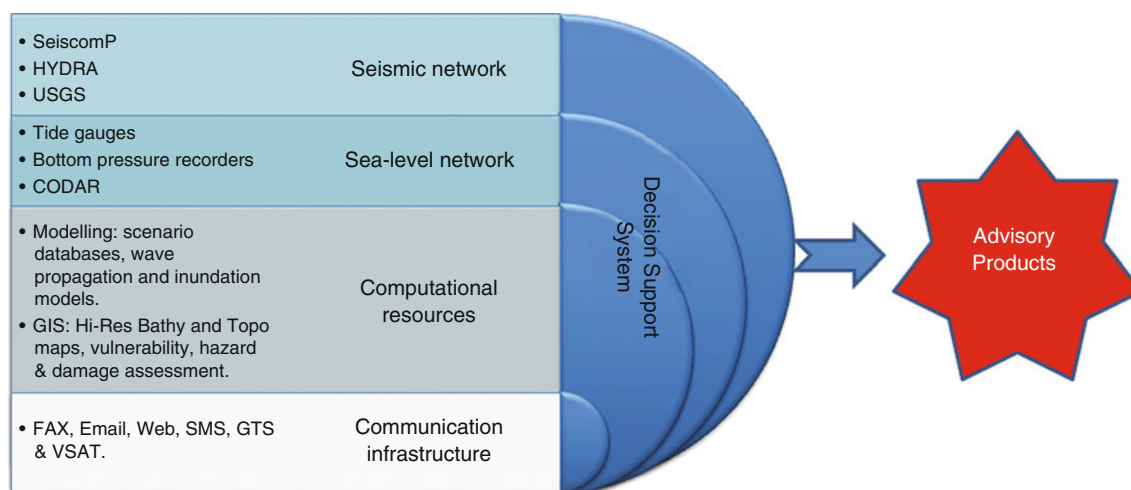
Observation Network

Seismic Network

TWCs use local and global seismographic networks transmitting seismograms in real time to continuously monitor seismicity, in order to locate and estimate the size of potentially tsunamigenic earthquakes. To produce accurate moment magnitude, NTWCs and TSPs require reliable, broad-frequency, low-noise, high dynamic-range, digital seismic data in real time. The timeliness of the data is crucial in issuing an initial bulletin within 5 min of an earthquake. This is especially important for centers with local tsunami sources. In terms of network density, it is recommended that at least eight stations must be operating in a radius of 900 km, with data latency not exceeding 30 s. IRIS Global Seismographic Network (GSN), Comprehensive Nuclear Test Ban Treaty Organization (CTBTO), and the German Research Centre for Geosciences (GFZ) Geofon Extended Virtual Network and stations from other national and regional networks are used in conjunction with national networks for earthquake detection. Focal mechanism of earthquakes that can lead to assessment of its tsunamigenic potential is generally derived through moment tensor inversion of seismic data. Limitations in assessing these parameters accurately enough within the timelines needed to provide tsunami warnings lead to several uncertainties. GNSS data is increasingly being used to estimate source parameters in real time (Blewitt et al. 2006; Sobolev et al. 2007) to reduce uncertainties in tsunami warning. Using high-rate (1 Hz or greater) dense GNSS network, large dynamic and static ground motions within few tens of seconds can be measured accurately. Also, strong motion accelerometer data are being used for estimating magnitude and derive displacements to complement GNSS-derived parameters.



Tsunami Watch and Warning Centers, Fig. 1 Area of coverage of the four regional tsunami warning systems. (Courtesy: IOC-UNESCO)



Tsunami Watch and Warning Centers, Fig. 2 Typical components of an end-to-end tsunami early warning system. (Courtesy: ESSO-INCOIS)

Sea Level Network

In order to confirm whether the earthquake has actually caused a tsunami, it is essential to measure the change in water level as near to the fault zone as possible with high accuracy. TWCs use sea level networks (tide gauge, bottom pressure recorders, and coastal ocean dynamics applications radar (CODAR)) reporting data in real and near real time to verify the generation and evaluate the severity of a tsunami. Numerous international sea level monitoring networks provide essential real-time data that can be accessed by NTWCs and TSPs. Many of these networks are coordinated by the IOC-UNESCO. The most extensive and notable is the Global Sea Level Observing System (GLOSS), conducted under the auspices of the Joint Technical Commission for Oceanography and Marine Meteorology (JCOMM) of the World Meteorological Organization (WMO) and the IOC. A tsunami creates a signal component that is not normally present in the observed sea level in usual times. Fortunately, the signature of tsunami waves at sea level gauges is generally quite distinctive and thus recognizable (see Fig. 3) and can be differentiated from a signal due to storm surge (UNESCO-IOC Manual and Guides 2006).

Techniques to use open ocean sea level data from bottom pressure recorders to estimate the characteristics of source have significantly enhanced the capabilities of TWCs to issue accurate tsunami forecast information for regional and distant tsunamis. In addition to characterizing tsunamis triggered by earthquakes, this methodology can be applied to derive source characteristics of tsunamis caused by atypical sources such as volcanic eruptions and submarine landslides (Patanjali Kumar et al. 2015; Mulia et al. 2018).

Tsunami Forecast Modeling

Guesstimating the generation mechanism of tsunami from the earthquake information, identifying the wave propagation

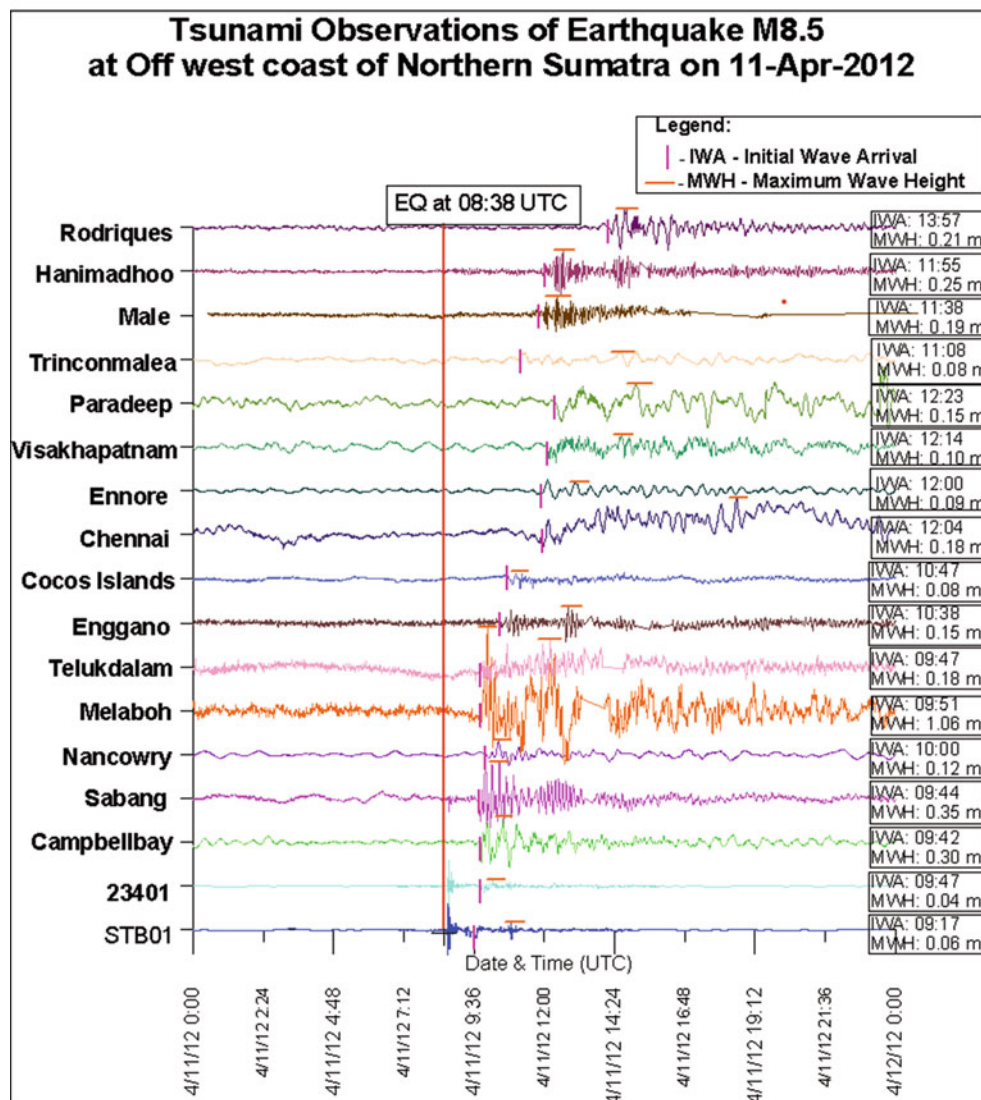
scheme, forecasting the travel time, and mapping inundation zones are critical requirements in tsunami early warning, which can be accomplished using numerical models. Two of the most commonly used tsunami models presently are the Method of Splitting Tsunami (MOST) developed by Titov and Gonzalez (1997) and the Tohoku University's Numerical Analysis Model (TUNAMI) developed by Imamura et al. (1995). Both these models are capable of simulating the three processes of a tsunami: earthquake generation, trans-oceanic propagation, and inundation of land. These models basically take the seismic deformation as input to predict the run-up heights and inundation levels at coastal regions for a given tsunamigenic earthquake (Imamura et al. 1995). The seismic deformation for an earthquake is computed using Mansinha and Smylie (1971) formulation using the earthquake parameters like epicenter, focal depth, strike, dip and rake of the rupture plane, length, width of the rupture, and slip magnitude on the rupture plane.

Pre-run Model Scenario Database

The above models cannot be run in real-time situations because of the computational overheads and resources. Hence, a pre-run scenario database or look-up-table (LUT) approach is most suitable in operational conditions. The forecast strategy is based on a unit-source function methodology, whereby the model runs are individually scaled and combined to produce arbitrary tsunami scenarios. Each unit-source function is equivalent to a tsunami generated by a 7.5 Mw earthquake with a rectangular fault 100 km by 50 km in size and 1 m slip (Gica et al. 2008; Greenslade and Titov 2008; Nayak and Kumar 2008b, c). The faults of the unit functions are placed abutting each other. When the functions are linearly combined, the resultant source function is equivalent to a tsunami generated by a combined fault rupture with assigned slip at each sub-fault. These unit sources are located along the

Tsunami Watch and Warning Centers, Fig. 3

Sea level observations at different tsunami buoys and tide gauges installed in the Indian Ocean region showing water-level variations during the 11 April 2012 tsunami. (Srinivasa Kumar et al. 2012). STB – tsunami buoy deployed by India, 23401 – tsunami buoy deployed by Thailand

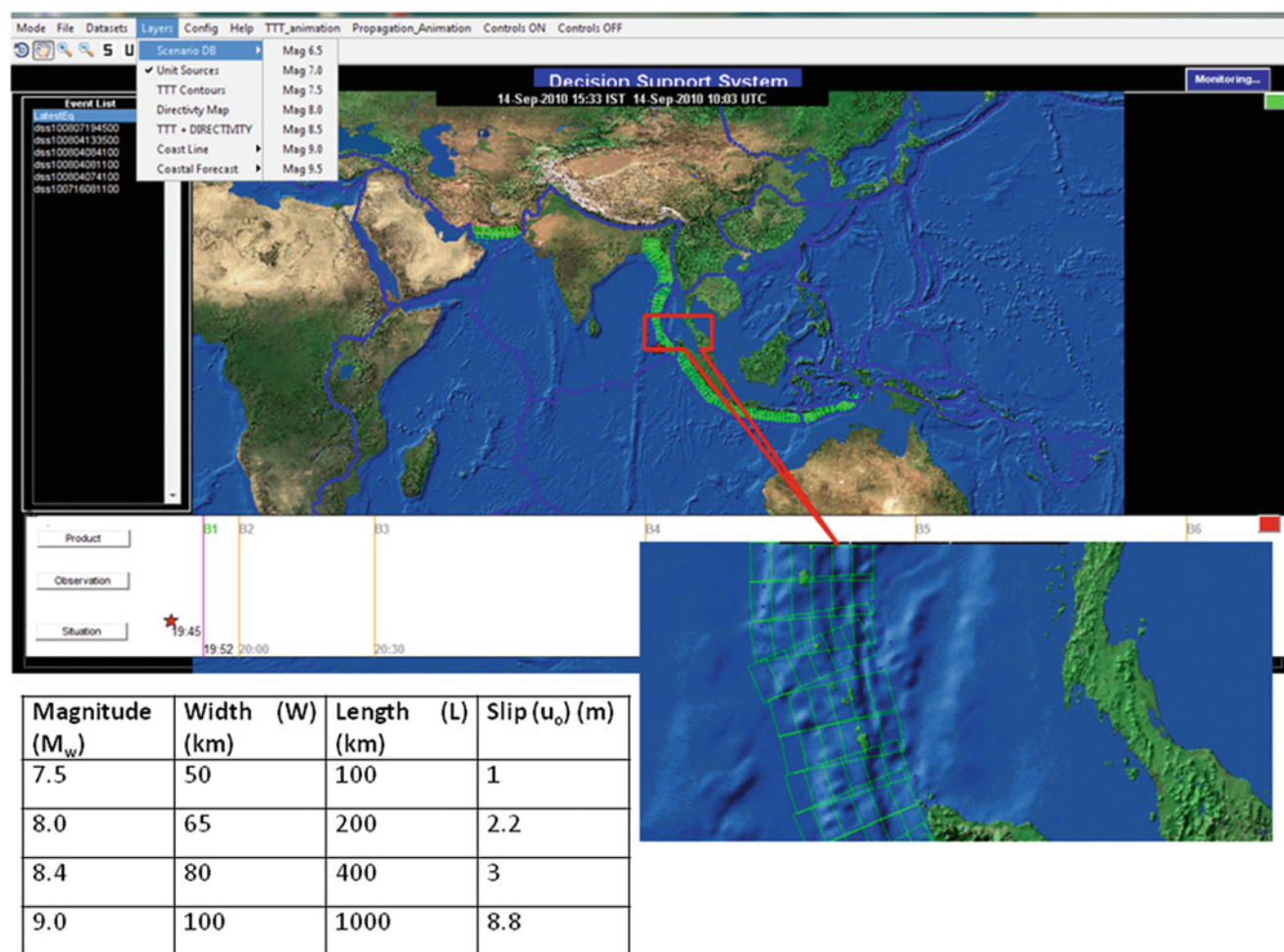


known fault zones for the entire Pacific Basin, Caribbean for the Atlantic region, and Indian Ocean. The key input source parameters are calculated using global relations from magnitude of the earthquake and fed to the model. In worst-case scenario, strike angle is considered as parallel to trench or coast, dip angle 45° , slip angle 90° . Figure 4 shows how the unit sources are set up in the Indian Ocean. A database of pre-run scenarios has been created for 50,000 unit sources covering all the tsunamigenic sources in the Indian Ocean region based on a finite difference code of Tsunami N2 model. Each scenario output contains the expected tsunami wave travel times, run-up heights, and directivity maps. At the time of earthquake occurrence, based on the location and magnitude, the basic unit-source open-ocean propagation scenarios are selected from the pre-run scenario database. These scenarios are then merged and scaled up/down using scaling relations

(Nayak and Kumar 2008c) to reconstruct a scenario representing the actual magnitude. The performance of the model is validated with the observations, and the model results are compared well with real-time observations (Srinivasa Kumar et al. 2010, ISPRS).

Inundation Modeling and Mapping

Most warning centers currently provide information on tsunami wave heights along the coastal regions, but not inundation extent on land. To achieve this, open-ocean wave from the scenario database, described in previous section, has to be extended over to the shallow waters and to the coastal area to simulate inundation over land. Knowing the coastal inundation zone is the first step in disaster management and mitigation, as it directly tells us the extent of population to be evacuated and the estimated economic loss due to an event.



Tsunami Watch and Warning Centers, Fig. 4 Modeling using “unit-source” approach and the resulting scenario database. (Courtesy: ESSO-INCOIS)

To perform accurate modeling of tsunami inundation and run-up, a detailed model of near-shore bathymetry and coastal topography is necessary. This is crucial as the accuracy of model predictions is directly related to the quality of the data used to create the bathymetry and topography of the model area. Coastal bathymetry is the prime determinant of the height of the tsunami wave or storm surge as it approaches the coast. Within the Indian Ocean context, topography of the region at 1:25,000-scale with contours at intervals of 0.5–2 up to 20 m has been acquired. Preliminary maps have been prepared of coastal topography using satellite stereo data (CARTOSAT-1) for the Indian coast. These datasets, along with airborne Light Detection and Ranging (LIDAR) data, were further used to create coastal vulnerability index (CVI) maps and multi-hazard vulnerability maps (MHVM) of coastline of India (Nayak and Kumar 2008c). Similar high-resolution geospatial database is being leveraged in the Atlantic and Pacific Oceans (Gica et al. 2008). With advances in techniques for near-real-time estimation of earthquake source

parameters and availability of high performance computers, it is now becoming feasible to run tsunami inundation models in real time.

Inundation mapping is not the only area where geospatial technologies are used, the inundation levels estimated and the subsequent threat status have to be accurately linked and conveyed to the respective administrative block within the coastal zones or within a province or a country. To implement such mechanism, the concept of coastal forecast points (CFP) and coastal forecast zones (CFZ) was developed (ICG/IOTWS) and implemented within the Indian Ocean region. Another major application of evolving geospatial trends is the application of 3D modeling or “video-grammetric” approach to visualize inundation and to identify evacuation routes and safe houses during inundation.

In India, Multi-hazard Vulnerability Mapping (MHVM) was carried out to identify highly vulnerable coastal zones along the Indian Coast. The multi-hazard vulnerability maps use blended and combined overlay of multiple hazards

(tsunami, storm surge, sea level rise, coastal erosion, etc.) affecting the coastal zone. These maps can aid as critical information during a disaster for the evacuation process and to evolve a management strategy (Mahendra et al. 2011). Further, high-resolution 3D GIS mapping in selected vulnerable areas was developed and integrated on 3D visualization application (3D-VAS). Three-dimensional GIS mapping comprised realistic 3D models of the buildings along with the attributed details of socio-demographic data. The major objectives of 3D GIS mapping are to visualize and effectively communicate the inundation threat at street level and to plan for evacuation routes.

Real-Time Tsunami Modeling for Propagation and Inundation
Most warning centers currently provide information on tsunami wave heights along the coastal regions, but not inundation extent on land. ITEWC is now warning on a finite element based shallow water and circulation model – ADvancedCIRCulation (ADCIRC) – to compute tsunami wave propagation and inundation (Luettich and Westerink 2004). Alternately, coastal high-resolution model grids can be run in real time by forcing them with wave heights along open boundary, extracted from an open ocean propagation simulation. Advances in the computing technology facilitated ITEWC to run the models on experimental basis successfully. The tsunami propagation of 24 h for the Indian Ocean Domain grid is tested for the Sumatra and Makran subduction zones and the simulation takes around 2 min of run time using 480 processors for 0.4 million computational grid. Tsunami open ocean propagation and inundation simulation of 12 h for the 2.8 million grid points is tested at the Sumatra subduction zone for 2004 tsunami, and it takes around 8.5 min (Fig. 5a, b).

Communication Infrastructure

Communication infrastructure is the most crucial and tangible aspect of any warning system for tsunami or any hazard. The ability to report and warn of an event is what makes and defines the warning system. Communication in terms of receiving all scientific data and disseminating information to the target audience within the required time needs a reliable communication infrastructure. During this process, warning centers must respond fast, be as accurate as possible, and be reliable in order to be effective. All tsunami warning systems feature multiple lines of communications (such as SMS, email, fax, radio, texting, and telex, often using hardened dedicated systems), enabling emergency messages to be sent to the emergency services and armed forces, as well as to population alerting systems.

Decision Support System

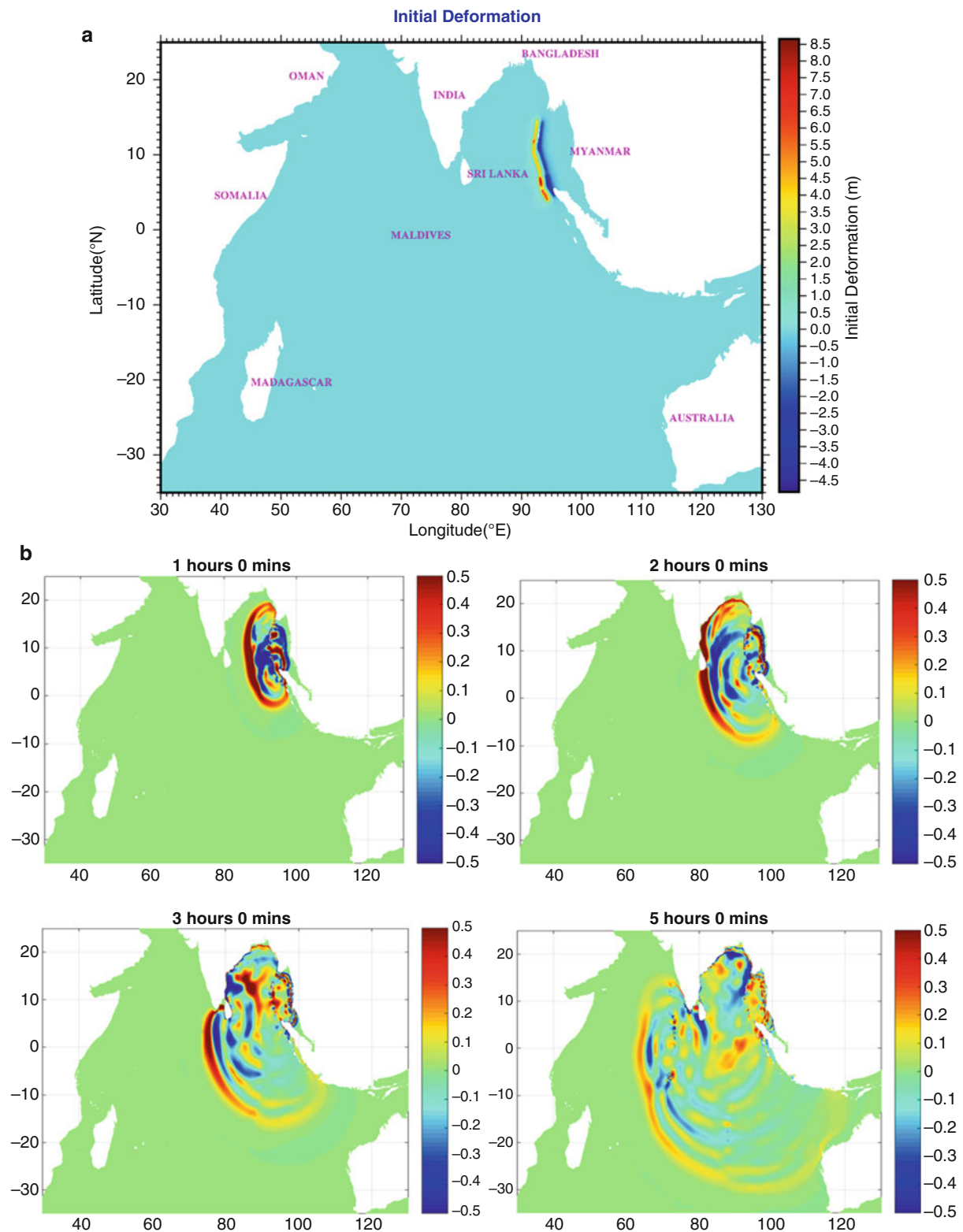
In simple words, when the standard operating procedures are coded into a program that can support organizational decision making, then such a program can be called as a decision

support system. It helps decision makers to compile useful information from a combination of raw real-time data, documents, personal knowledge, and tsunami scenarios models to accurately disseminate warning advisories to the needed area at the right time. The DSS encapsulates and foresees all the operations from the beginning of the event till the passage of threat by issuing timely advisories. Typical bulletins issued by the ITEWC are earthquake information, warning, watch, and alert, which are described in detail in the following sections. All the information gathering, decision-making process, and infrastructure are part of the operational monitoring center.

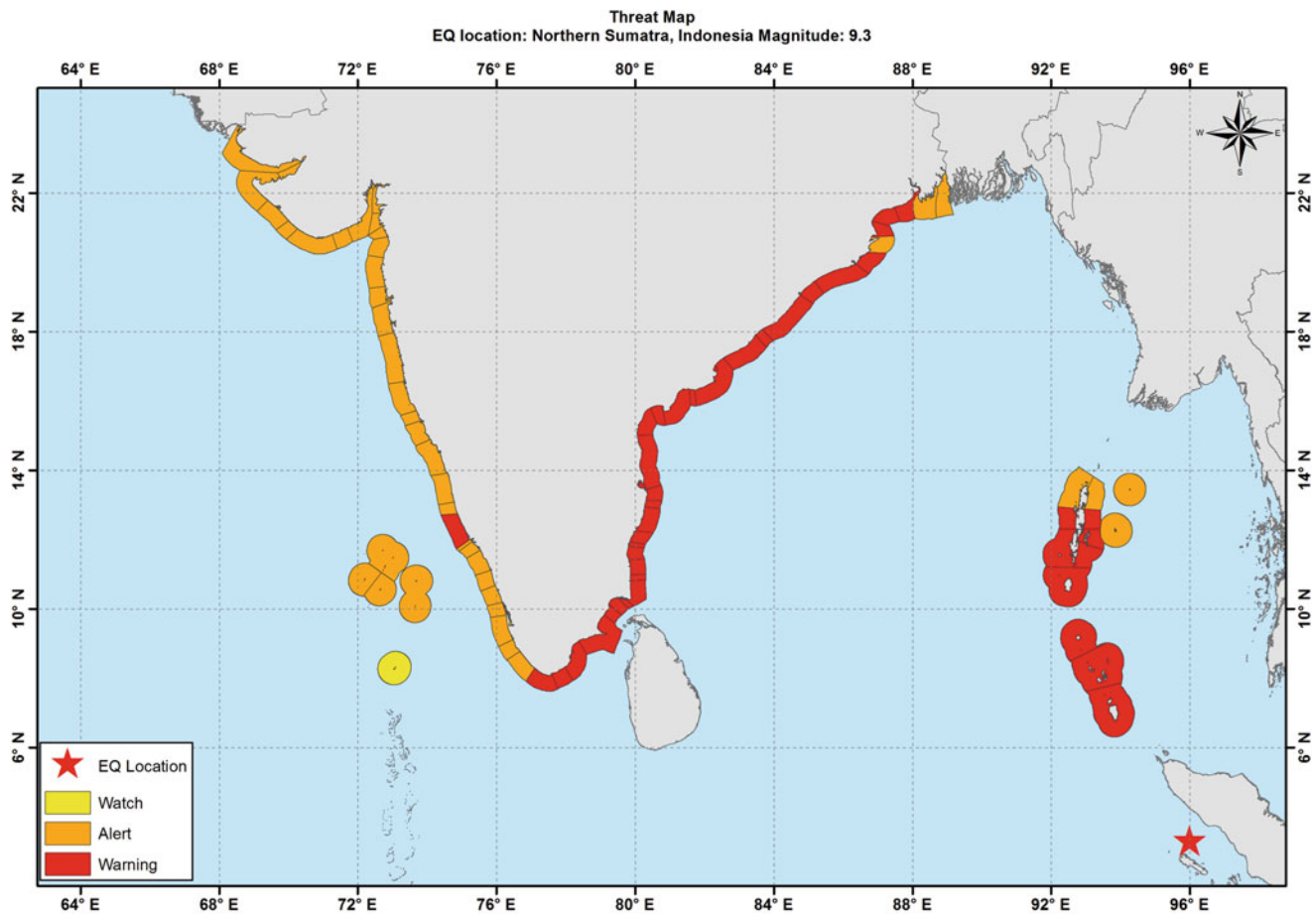
Typical Bulletins Issued by TWCs

Based on the earthquake information, tsunami modeling scenarios, and operational capability, different types of advisory products are issued to the disaster management officials and to the general public. Each bulletin has a distinct meaning and clearly suggests the action to be taken. The earthquake information bulletin issued solely on the earthquake information is a qualitative advisory while all others are quantitative advisories (Figs. 6 and 7) issued taking into consideration the modeling results and real-time sea level observations, along with the earthquake information.

1. Earthquake information bulletin is issued to inform the occurrence of an earthquake and to advise its potential to generate a tsunami based on the magnitude. The bulletin contains information about the earthquake parameters (origin time, latitude, longitude, focal depth, and magnitude) and the name of the geographical area where it has occurred.
2. Warning is the highest level where immediate action has to be taken. This advisory is issued when a potential tsunami with significant widespread inundation is imminent or expected. Warnings alert the public that widespread, dangerous coastal flooding accompanied by powerful currents is possible and may continue for several hours after arrival of the initial wave. Warnings also alert emergency management officials to take action for the entire tsunami hazard zone. Warnings may be updated, adjusted geographically, downgraded, or canceled.
3. Watch is issued to alert emergency management officials and the public of an event which may later impact the watch area. Immediate public evacuation is not required. Local officials should be prepared for evacuation if it is upgraded to warning status. The watch area may be upgraded to a warning or canceled based on updated information and analysis. Watches are normally issued based on seismic information without confirmation that a destructive tsunami is underway.
4. Alert contains information about the earthquake and a qualitative tsunami evaluation message indicating that a tsunami is expected. As conditions warrant, the bulletin will either be continued, upgraded to a watch or warning, or canceled. Alert implies public should avoid beaches since strong



Tsunami Watch and Warning Centers, Fig. 5 (a) Initial deformation of ocean surface due to 26 December 2004 earthquake. (b) Computed tsunami heights due to 26 December 2004 earthquake at different time steps



Tsunami Watch and Warning Centers, Fig. 6 Tsunami threat status map. (Courtesy: ESSO-INCOIS)

currents are expected. Local officials should be prepared for evacuation if it is upgraded to warning status.

5. Cancellation indicates the end of the threat due to tsunami. A cancellation is usually issued after an evaluation of sea level data confirms that a destructive tsunami will not impact the warned area anymore. A cancellation will also be issued following a destructive tsunami when sea level readings indicate that the tsunami is below destructive levels and is subsiding in most locations.

Service Levels

Based on the level of integration of key component into the TWC, each TWC has the capability to deliver products at certain operational level. Each higher service level represents much more detailed threat evaluation and will be directed to more specific agencies/population group for disaster mitigation with increased confidence levels. Presently, the implementation plan

of IOTWS (IOC-UNESCO 2016) classifies these product levels into three groups, based on incremental capacity development:

Service Level I: At this level of operation, the TSP disseminates only the earthquake information and a qualitative tsunami threat evaluation to other national warning centers.

Service Level II: This is a more robust service level that includes all of the elements of Service Level I but adds modeling and forecast elements so that Watches and Advisory products include: estimated wave height(s) (EWH) and estimated time(s) of arrival (ETA) for offshore forecast points, potential threat zones, advanced warning and notification products of threat status and zones in image format. At this service level of operation, real-time observations are included in the decision making and in advisory products.

Service Level III: This is the most advanced level of operational service. At this level, the advisory products include



Tsunami Watch and Warning Centers, Fig. 7 An example of quantitative advisory bulletin product. The product includes graphic of tsunami travel time, along with text indicating the forecast of wave heights

and arrival time along with threat status to each administrative zone in the Indian Ocean. (Courtesy: ESSO-INCOIS)

local risk and hazard assessment using inundation models, and they are more specifically directed to communities at risk. Service Level III operations take into account real-time inundation forecast in evaluating the threat status.

A TWC operating at this service level indicates a mature early warning system capable of detecting and evaluating tsunamigenic earthquakes occurring anywhere in the global oceans.

Conclusion

Tsunami warning systems have made significant progress in recent years, but challenges and uncertainties remain in meeting the information requirements for facilitating effective response of disaster managers and public due to very short timeframes available for tsunami warning. Current systems are capable of providing warning only for earthquake generated tsunamis but cannot provide warnings for tsunamis caused by atypical sources (submarine and coastal landslides, volcanoes, etc.) or near-field tsunamis, as witnessed from the 2018 Palu and Sunda Strait tsunamis in Indonesia. Tsunami warning for near-source regions that requires rapid public response (within few minutes) is extremely challenging considering scientific uncertainties involved in accurate assessment of source characteristics, gaps in warning chains and stakeholder Standard Operating Procedures (SOPs), and lack of public awareness and response.

The most important unknown information about a tsunami in real time is the nature of source. Efforts are underway for use of technologies such as GNSS and BPR data inversion for rapid estimation of source, but current observation networks are far from optimal. Another huge limitation in the accurate modeling and forecasting of tsunamis as they interact with the coast is due to the non-availability of high-resolution topography and bathymetry data. New and emerging ocean sensing capabilities such as increasing number of coastal sea level stations, BPRs on under-sea fiber-optic cables, GNSS networks, advances in seismology modeling, and computational resources present an opportunity to greatly reduce and/or quantify uncertainties associated with tsunami forecasting (Angove et al. 2019).

Strengthening warning capabilities to be able to detect, model, and monitor tsunamis from a wider range of sources is a challenging, but priority task. Further, stronger downstream, including community preparedness, initiatives to enable appropriate response to official warnings and more importantly natural signs of a tsunami are extremely important for saving lives of vulnerable people.

Acknowledgments The authors would like to specially thank Ms. Sunanda Manneela for compiling information in the preparation of the manuscript, and Mr. Ch. Patanjali Kumar, Mr. R. S. Mahendra, Dr. P. L. N. Murty, Mr. Ajay Kumar, and Mr. J. Padmanabham for providing graphical inputs and in developing the manuscript. The authors are grateful to Dr. S. S. C. Shenoi, Director, ESSO-INCOIS for encouragement and cooperation.

Bibliography

Angove M, Arcas D, Bailey R, Carrasco P, Coetzee D, Fry B, Gledhill K, Harada S, von Hillebrandt-Andrade C, Kong L, McCreery C, McCurrach S-J, Miao Y, Sakya AE, Schindelé F (2019) Ocean observations required to minimize uncertainty in global tsunami

- forecasts, warnings, and emergency response. *Front Mar Sci*. <https://doi.org/10.3389/fmars.2019.00350>
- Blewitt G, Kreemer C, Hammond WC (2006) Rapid determination of earthquake magnitude using GPS for tsunami warning systems. *Geophys Res Lett* 33:L11309. <https://doi.org/10.1029/2006GL026145>
- Gica E, Spillane MC, Titov VV, Chamberlin CD, Newman JC (2008) Development of the forecast propagation database for NOAA's short-term inundation forecast for tsunamis (SIFT). NOAA technical memorandum OAR PMEL139. Contribution no. 2937 from NOAA/Pacific Marine Environmental Laboratory
- Greenslade DJM, Titov VV (2008) A comparison study of two numerical tsunami forecasting systems. *Pure Appl Geophys* 165:1991–2001. <https://doi.org/10.1007/s00024-008-0413-x>
- Gupta HK (2005) Mega-tsunami of 26th December, 2004: Indian initiative for early warning system and mitigation of oceanogenic hazards. *Episodes* 28(1):2–5
- Imamura F, Yalciner AC, Ozyurt G (1995) Tsunami modeling manual (TUNAMI model), April. Available at: <http://www.tsunami.civil.tohoku.ac.jp/hokusai3/E/projects/manual-ver-3.1.pdf>
- IOC-UNESCO (2016) Tsunami watch operations – global service definition document, IOC technical series no. 130. Intergovernmental Oceanographic Commission–UNESCO, Paris
- Luetich RA, Westerink JJ (2004) Formulation and numerical implementation of the 2D/3D ADCIRC finite element model version 44.XX. http://adcirc.org/adcirc_theory_2004_12_08.pdf
- Mahendra RS, Mohanty PC, Bisoyi H, Srinivasa Kumar T, Nayak S (2011) Assessment and management of coastal multi-hazard vulnerability along the Cuddalore–Villupuram, east coast of India using geospatial techniques. *Ocean Coast Manag* 54(4):302–311
- Mansinha LA, Smylie DE (1971) Displacement fields of inclined faults. *Bull Seismol Soc Am* 61:1433–1440
- Mulia IE, Gusman AR, Jakir Hossen M, Satake K (2018) Adaptive tsunami source inversion using optimizations and the reciprocity principle. *J Geophys Res Solid Earth* 123:10,749–10,760
- Nayak S, Kumar ST (2008a) The first tsunami early warning center in the Indian Ocean. In: *Risk Wise*. Tudor Rose, Leicester, pp 175–177
- Nayak S, Kumar ST (2008b) Addressing the risk of tsunami in the Indian Ocean. *J South Asian Disaster Stud* 1(1):45–57
- Nayak S, Kumar ST (2008c) Indian tsunami warning system. *Int Arch Photogramm Remote Sens Spat Inf Sci XXXVII:Part B4*
- Nayak S, Kumar ST (2011) Tsunami watch and warning centers. *Encycl Solid Earth Geophys*. https://doi.org/10.1007/978-90-481-8702-7_230
- Patanjali Kumar C, Ajay Kumar B, Uma Devi E, Mahendra RS, Sunanda MV, Pradeep Kumar M, Padmanabham J, Dipankar S, Srinivasa Kumar T (2015) The admissible tsunamigenic source region of 24 September 2013 land-based earthquake – application of backward ray tracing technique. *Curr Sci* 108(9):1712–1716
- Sobolev SV, Babeyko AY, Wang R (2007) Tsunami early warning using GPS-Shield arrays. *J Geophys Res* 112. <https://doi.org/10.1029/2006JB004640>
- Srinivasa Kumar T, Patanjali Kumar C, Nayak S (2010) Performance of the Indian tsunami early warning system. *Int Arch Photogramm Remote Sens Spat Inf Sci XXXVIII(8):271–274*
- Srinivasa Kumar T, Nayak S, Patanjali Kumar C, Yadav RBS, Ajay Kumar B, Sunanda MV, Uma Devi E, Shenoi SSC (2012) Performance of the tsunami forecast system for the Indian Ocean. *Curr Sci* 102(1):110–114
- Titov VV, Gonzalez FI (1997) Implementation and testing of the method of splitting tsunami (MOST) model. NOAA technical memorandum ERL PMEL-112. Contribution no. 1927 from NOAA/Pacific Marine Environmental Laboratory
- UNESCO-IOC (2006) Manuals of sea level interpretation, vol 4. UNESCO, Paris
- UNESCO-IOC (2019) Tsunami glossary, IOC technical series no. 85. UNESCO, Paris

motion, and the Andaman Sea ridge transform fault system which is characterized by the normal and strike slip earthquakes of low magnitude ($M < 6.5$), do not pose tsunami hazard to the Bay of Bengal region. Tsunami hazard in the Bay of Bengal due to seismicity in the diffused plate boundary region between India and Australia in the Indian Ocean needs to be considered as this region produced two great earthquakes on April 11, 2012 (Yadav et al. 2013). Although these earthquakes, and a few more which occurred in the Wharton basin, were characterized by the strike slip motion and did not cause any significant tsunami, due to the complex tectonics involved in the diffused plate boundary region, possibility of a tsunamigenic earthquake may not be completely overruled.

Plate Motion Along the Northern Sunda Arc

In the northern Sunda arc relative motion between the India and Sunda plates is about 3.7 cm/year towards N15° in the northern part (i.e., Indo-Burmese arc region) and about 4.5 cm/year towards N25° in the southern part (i.e., the Sumatra region). The oblique motion between the India and Sunda plates is accommodated along the arc through slip partitioning in the frontal and back arc (Gahalaut and Gahalaut 2007; McCaffrey 2009). In the north, the north-south trending Indo-Burmese arc accommodates slip of about 1.6 cm/year through predominantly dextral motion (Gahalaut et al. 2013) and the remaining motion of about 2.0 cm/year is accommodated across the dextral strike slip Sagaing fault in the Myanmar (Maurin et al. 2010). However, Steckler et al. (2016), based on some additional GPS measurements from Bangladesh, proposed that some of the India-Shan plateau motion is accommodated through thrust motion on the detachment beneath the outer Indo-Burmese wedge. Further south in the Irrawaddy and Andaman region, the motion in the frontal arc becomes oblique. In the back arc, the dextral strike slip motion occurs along the north-south planes and normal motion occurs along the Andaman Sea-ridge transform fault system which is consistent with the Andaman sea opening. In the Sumatra region, the motion in the frontal arc is arc normal, leading to thrust motion on the plate boundary interface. Strike slip motion in the back arc is accommodated through dextral motion along the Sumatra Fault System. The Sumatra Fault System joins the Sagaing fault in the north through the Andaman Sea Ridge transform fault system.

Seismicity of the Northern Sunda Arc Region and Tsunami Hazard

For the convenience of discussion, the seismicity of the Indo-Burmese arc, Irrawaddy, and Andaman-Sumatra region is described separately.

Indo-Burmese Arc

The plate reconstruction models suggest that subduction probably occurred along the Indo-Burmese arc in the geological past (prior to about 50 Ma) when the arc was predominantly southeast-northwest trending. However, after the collision of the India plate with the Eurasian plate, the northern Sunda arc rotated clockwise to become predominantly north-south trending. Several geological and geophysical studies confirm that the subduction occurred across Indo-Burmese arc between India and Burma plates and there is evidence of the subducted Indian slab, but whether the subduction is still active, is debatable (Hall 1997; Guzman-Speziale and Ni 1996; Rao and Kumar 1999; Pesicek et al. 2010; Gahalaut et al. 2013; Steckler et al. 2016).

The historical records of earthquakes and tsunami in the Indo-Burmese arc are generally very poor. Although there are a few unverifiable reports of earthquake occurrence as early as 1548 in Tripura, Assam (two states in the northeastern India), or Bangladesh, there are large uncertainties in the location of historical earthquakes. The catalogue is probably reliable only after 1762. The most notable earthquakes are the April 2, 1762, August 24, 1858, Arakan earthquakes, and the January 10, 1869, Cachar earthquake (Richter 1958). Other than these earthquakes, several instances of damage due to earthquakes are reported from Chittagong, Sylhet, Manipur valley, and Cachar regions. However, it is possible that the high damage in these sediment filled valleys was due to relatively large population and local site effects. The April 2, 1762, Arakan earthquake has been considered as a great tsunamigenic earthquake (Cummins 2007). It caused extensive damage in the Chittagong region through shaking, liquefaction, damming of channels, seiches, etc. (Gulston 1763). The reported uplift of Cheduba Island due to this earthquake was not found to be singularly due to this earthquake (Oldham 1883; Martin and Szeliga 2010). Halsted (1843) reported the effects of the earthquakes which were mostly found to be highly exaggerated by Gupta and Gahalaut (2009). They found that the highly qualitative observations of the coseismic elevation changes are consistent with a predominant strike slip motion on the earthquake rupture which was located under the land and/or under very shallow water, and it was only a major earthquake (Martin and Szeliga 2010) which probably did not cause any major tsunami (Gupta and Gahalaut 2009). The August 24, 1858, Arakan earthquake was felt in many parts of Burma and was severely felt at Kyauk Pyu, Ramree Island, Myanmar. It caused liquefaction, damage to buildings and Pagodas. In this region, the distance between the Sagaing fault and the structurally mapped Arakan trench is less than 200 km, and hence based on the scanty reports of damage it is difficult to judge whether this one and the 1843 and 1848 earthquakes are linked with the Indo-Burmese arc or with the Sagaing fault. But none of these three earthquakes is associated with tsunami. The January

10, 1869, Cachar-Manipur earthquake is referred as the most severe earthquake in the available 2000 years of written historical records of Manipur state of India. However, recent examinations of the historical records suggest that even this earthquake was not a great earthquake and the damage was very much confined in the sediment filled valleys of Imphal and Cachar (Kundu and Gahalaut 2012). As this earthquake occurred on the land, far from the sea, there is no question of tsunami being generated by this earthquake. Based on the observations of recent earthquakes, which also include the 2017 Tripura (M 5.6) and 2016 Manipur earthquakes (Debbarma et al. 2017; Gahalaut et al. 2016), which exhibit strike slip motion at depth more than 25 km, it has been suggested that these earthquakes occur within the Indian plate and not on the interface between the Indian plate and the Indo-Burmese wedge.

Although precise motion on rupture during these historical major earthquakes is not known, the current seismicity and the results from crustal deformation studies suggest that these earthquakes in the Indo-Burmese arc might have occurred through predominant strike slip motion on steep planes within the Indian plate. These observations limit the possibility of a tsunamigenic earthquake in the coastal Indo-Burmese arc region. Lack of paleo-tsunami deposits in the coastal Myanmar region (Aung et al. 2006) further supports this argument.

Irrawaddy Region

Seismicity in the Irrawaddy frontal arc is extremely low as compared to that in the Andaman Nicobar arc in the south and the Indo-Burmese arc in the north. Earthquakes in the Irrawaddy region occur at depths less than 50 km, whereas in the Indo-Burmese arc and the Andaman Nicobar region, earthquakes occur down to a depth of up to 150–200 km (Guzman-Speziale and Ni 1996). From the limited available earthquake focal mechanisms, it is not possible to constrain sense of motion along this part; however, the regional plate motion models suggest that very oblique motion occurs along the frontal arc (Gahalaut and Gahalaut 2007). The reason for very low seismicity of the frontal arc at shallow depth (>50 km) could be that the region is either aseismic or accumulating strain. However, absence of seismicity even at depths greater than 50 km suggests that the subducting slab might not be continuing to that depth at all. In the Andaman and Indo-Burmese arc regions, occurrence of earthquakes at intermediate depths confirms the continuity of the Indian plate slab in those regions. The absence of seismicity at intermediate depth level of the Irrawaddy region is possibly due to the presence of tear in the subducting Indian slab, which has been confirmed from the tomographic studies of the region (Richards et al. 2007). Thus, it is the tear in the slab which is responsible for very low seismicity at shallow depth and no seismicity at deeper depth (Kundu and Gahalaut 2010). Presence of a tear and lack of slab in the region may substantially

reduce the potential of occurrence of a future great earthquake in the region. Accordingly, it may also reduce tsunami hazard in the nearby regions due to major and great earthquakes in this region. It is possible that this tear hindered the northward rupture propagation of the 2004 Sumatra Andaman earthquake. However, it is necessary to further confirm the presence of such a tear using marine deep seismic surveys, detailed tomographic studies, and analysis of earthquakes with the deployment of ocean bottom and land seismometers in the region.

Andaman Sumatra Arc Region

The state of knowledge about the Andaman-Sumatra region was very poor until the occurrence of the December 26, 2004, Sumatra-Andaman (M 9.2), March 28, 2005, Nias (M 8.6) and several other earthquakes since then. These earthquakes have provided immense information about the tectonics and seismogenesis of this region. These earthquakes suggest that Indian plate subducts obliquely beneath the Andaman arc while the subduction becomes arc normal in the Sumatra arc. Precise hypocentral locations of these earthquakes and their aftershocks have helped in mapping the subsurface plate boundary interface in the frontal arc region and its seismogenic width. The 2004 Sumatra Andaman earthquake occurred on a 1400 km long and 100–150 km wide rupture between the northern Sumatra and north Andaman Islands. The large coseismic displacement in the source region (Gahalaut et al. 2006) caused a devastating tsunami which propagated through the Indian Ocean and caused heavy destruction in the countries surrounding the Bay of Bengal, namely, Indonesia, India, Thailand, and Sri Lanka. There were a few remarkable aspects of this earthquake and tsunami caused by it. It is learnt that the rupture in the northern part, i.e., under the Andaman Islands, was slower (Lay et al. 2005) and that it did not reach the surface. Thus, this part of the rupture did not contribute in tsunami generation. It has been suggested that the subducting 90°E ridge probably slowed down the rupture under the Andaman Islands (Gahalaut et al. 2010) and the subducted unconsolidated sediments of Bay of Bengal probably did not allow the rupture to propagate to the surface. These aspects need to be explored further. Earlier, another earthquake occurred here on June 26, 1941, with M 7.7 (Pacheco and Sykes 1992), for which no seismological data exist to suggest whether slow rupture occurred during that earthquake as well. Though the reports are scanty, it appears that even that earthquake did not generate a tsunami (Rajendran et al. 2007). There may be various reasons for that earthquake for not being tsunamigenic, including slow rupture speed and rupture on blind fault. It may be possible that the Andaman segment may not generate major tsunami during future great earthquakes as well. However, it needs to be explored further by analyzing the frontal arc characteristics through deep seismic imaging and modeling.

An earthquake occurred on December 31, 1881 (M 7.9), in the Nicobar region. It caused a tsunami with less than 1 m height in the Andaman Islands and the coastal region of India (Ortiz and Bilham 2003). Unfortunately, the history of the Andaman Nicobar Islands is very poor. The intermittent colonial occupation of the Islands started in the middle of the eighteenth century by the Danish, followed by the British and continuous colonial occupation started only after 1869 until the independence of India in 1947. Prior to the nineteenth century, we do not know much about the history of the Andaman-Nicobar region as the indigenous tribes did not maintain any written records. Thus, prior to the occurrence of earthquakes on 31 October 1847, possibly near Nicobar Islands (Bilham et al. 2005), and the two earthquakes mentioned above, we have no knowledge of historical earthquakes or tsunami in the region. However, it is certain that the great and major earthquakes under the Nicobar and northern Sumatra arc have produced tsunami and are capable of generating tsunami during future great and major earthquakes.

In the northern Sumatra region, after the occurrence of the 2004 Sumatra-Andaman earthquake, another great earthquake occurred to the south on March 28, 2005, which is referred as the Nias earthquake. Despite its large magnitude and normal rupture speed, the earthquake could not generate a major tsunami. Majority of the slip on the rupture occurred under the shallow and no water and hence it could not displace large amount of water, causing only a localized tsunami whose amplitude faltered quickly as it moved in the open ocean. Nevertheless, this region has the potential to generate a tsunami during great earthquakes, though it may not affect the Bay of Bengal region due to rupture directivity. The arc region located further southeast, though capable of causing tsunami during great earthquakes, may not affect the Bay of Bengal region.

The intense earthquake activity in the Sumatra Andaman subduction zone probably triggered earthquakes in the outer rise and also in the diffused plate boundary in the Wharton basin. The largest earthquake in this region occurred on April 11, 2012, with a magnitude of 8.6 which was immediately followed by another large magnitude earthquake (M 8.2). However, both these and other earthquakes in the Wharton basin accompanied strike slip motion and hence were not tsunamigenic and the earthquakes in the outer rise were not so large to cause tsunami. The complex tectonic processes in the diffused plate boundary region do not necessarily warrant that large earthquakes with dip slip motion would not occur in future in that region.

Summary

The Arakan coast of the Indo-Burmese arc can produce a major earthquake. However, it is unlikely that a major

tsunami would be generated by the earthquake in this segment due to predominant strike slip motion on the plate boundary interface which is mostly located on the land and under very shallow water and dips eastward, away from the sea. The Irrawaddy frontal arc region may not produce great or even major earthquakes and hence the tsunami threat from this region is low. The Andaman arc which did not contribute in tsunami generation during the 2004 Sumatra-Andaman earthquake needs to be studied in great detail in view of the slow and blind rupture. The real and definite threat is from the Nicobar and northern Sumatra arc region, which has produced tsunami in the past and is capable of generating tsunami during future great earthquakes. The diffused plate boundary region in Wharton basin may also be considered as a possible region which may produce a tsunamigenic earthquake.

Cross-References

- Earthquakes and Crustal Deformation
- GPS, Data Acquisition, and Analysis
- Plate Tectonics, Precambrian
- Tsunami

Bibliography

- Aung TT, Okamura Y, Satake K, Swe W, Swe TL, Saw H, Tun ST (2006) Paleoseismological field survey along the western coast of Myanmar. *Annu Rep Active Fault Paleoearthq Res Active Fault Res Cent Geol Surv Jpn* 6:171–188
- Bilham RG, Engdahl ER, Feldl N, Satyabala SP (2005) Partial and complete rupture of the Indo-Andaman plate boundary 1847–2004. *Seismol Res Lett* 76:299–311
- Cummins PR (2007) The potential for giant tsunamigenic earthquakes in the northern Bay of Bengal. *Nature* 449:75–78
- Debbarma J, Martin SS, Suresh G, Ahsan A, Gahalaut VK (2017) Preliminary observations from the 3 January 2017, MW 5.6 Manu, Tripura, (India) earthquake. *J Asian Earth Sci* 148:173–180. <https://doi.org/10.1016/j.jseaes.2017.08.030>
- Gahalaut VK, Gahalaut K (2007) Burma plate motion. *J Geophys Res* 112:B10402. <https://doi.org/10.1029/2007JB004928>
- Gahalaut VK, Nagarajan B, Catherine JK, Kumar S (2006) Constraints on 2004 Sumatra–Andaman earthquake rupture from GPS measurements in Andaman-Nicobar Islands. *Earth Planet Sci Lett* 242:365–374. <https://doi.org/10.1016/j.epsl.2005.11.051>
- Gahalaut VK, Subrahmanyam C, Kundu B, Catherine JK, Ambikapathy A (2010) Slow rupture in Andaman during 2004 Sumatra-Andaman earthquake: consequence of subduction of 90°E ridge. *Geophys J Int* 180:1181–1186
- Gahalaut VK, Kundu B, Laishram SS, Catherine J, Kumar A, Singh MD, Tiwari RP, Samanta SK, Ambikapathy A, Mahesh P, Bansal A, Narsaiah M (2013) Aseismic plate boundary in the Indo-Burmese wedge, northwest Sunda Arc. *Geology* 41:235–238
- Gahalaut VK, Martin S, Srinagesh D, Kapil SL, Suresh G, Saikia S, Kumar V, Dadhich H, Prajapati SK, Gautam JL, Baidya P, Mandal S, Jain A (2016) Constraints from seismological, geodetic and macroseismic observations on the 4 January 2016 Tamenglong, Manipur earthquake. *Tectonophysics* 688:36–48

- Gulston E (1763) An account of an earthquake at Chattigaon: translated from the Persian by Mr. Edward Gulston, in the service of the Honourable East India Company, and communicated by him to the reverend Mr. Hirst. *Philos Trans* 53:251–256
- Gupta H, Gahalaut VK (2009) Is the northern Bay of Bengal tsunamigenic. *Bull Seismol Soc Am* 99:3496–3501
- Guzman-Speziale M, Ni JF (1996) Seismicity and active tectonics of the western Sunda Arc. In: Yin A, Harrison TM (eds) *The tectonic evolution of Asia*. Cambridge University Press, New York, pp 63–84
- Halsted EP (1843) Report on the Island of Chedooba, *J Asiatic Soc Bengal* 114:319–446
- Hall R (1997) Cenozoic plate tectonic reconstructions of SE Asia. In: Fraser AJ, Methews SJ, Murphy RW (eds) *Petroleum geology of Southeast Asia*, vol 126. Geological Society of London Special Publication, London, pp 11–23
- Kundu B, Gahalaut VK (2010) An investigation into the seismic potential of the Irrawaddy region, northern Sunda Arc. *Bull Seism Soc Am* 100:891–895
- Kundu B, Gahalaut VK (2012) Earthquake occurrence processes in the Indo-Burmese wedge and Sagaing fault region. *Tectonophysics* 524–525:135–146
- Lay T, Kanamori H, Ammon CJ, Nettles M, Ward SN et al (2005) The great Sumatra-Andaman earthquake of December 26, 2004. *Science* 308:1127–1133
- Martin S, Szeliga W (2010) A catalog of felt intensity data for 570 earthquakes in India from 1636 to 2009. *Bull Seismol Soc Am* 100:562–569
- Maurin T, Masson F, Rangin C, Than Min U, Collard P (2010) First global positioning system results in northern Myanmar: constant and localized slip rate along the Sagaing fault. *Geology* 38:591–594
- McCaffrey R (2009) The tectonic framework of the Sumatran subduction zone. *Annu Rev Earth Planet Sci* 37:345–366
- Okal EA, Synolakis CE (2008) Far-field tsunami hazard from megathrust earthquakes in the Indian Ocean. *Geophys J Int* 172:995–1015
- Oldham T (1883) A catalogue of Indian earthquakes. *Mem Geol Surv India* 19:163–215
- Ortiz M, Bilham R (2003) Source area and rupture parameters of the 31 December 1881 Mw=7.9 Car Nicobar earthquake estimated from tsunamis recorded in the Bay of Bengal. *J Geophys Res* 108:2215. <https://doi.org/10.1029/2002JB001941>
- Pacheco JF, Sykes LR (1992) Seismic moment catalogue of large shallow earthquakes, 1900 to 1989. *Bull Seism Soc Am* 82:1306–1349
- Pesicek JD, Thurber CH, Widiyantooro S, Zhang H, DeShon HR, Engdahl ER (2010) Sharpening the tomographic image of the subducting slab below Sumatra, the Andaman Islands and Burma. *Geophys J Int* 182:433–453
- Rajendran CP, Rajendran K, Anu R, Earnest A, Machado T, Mohan PM, Freymueller J (2007) Crustal deformation and seismic history associated with the 2004 Indian Ocean earthquake: a perspective from the Andaman–Nicobar Islands. *Bull Seism Soc Am* 97:S174–S191
- Rao NP, Kumar MR (1999) Evidences for cessation of Indian plate subduction in the Burmese arc region. *Geophys Res Lett* 26:3149–3152
- Richards S, Lister G, Kennett B (2007) A slab in depth: three dimensional geometry and evolution of the Indo-Australian plate. *Geochim Geophys Res* 12:Q12003. <https://doi.org/10.1029/2007GC001657>
- Richter CF (1958) *Elementary seismology*. Freeman, San Francisco, p 768
- Steckler MS, Mondal DR, Akhter SH, Seeber L, Feng L, Gale J, Hill EM, Howe M (2016) Locked and loading megathrust linked to active subduction beneath the Indo-Burman Ranges. *Nat Geosci* 9:615. <https://doi.org/10.1038/NGEO2760>
- Yadav RK, Kundu B, Gahalaut K, Catherine J, Gahalaut VK, Ambikapathy A, Naidu MS (2013) Coseismic offsets due to the 11 April 2012 Indian Ocean earthquakes (Mw 8.6 and 8.2) derived from GPS measurements. *Geophys Res Lett* 40:3389–3393. <https://doi.org/10.1002/grl.50601>

Uncertainties in Geomodelling Related to Geophysical Inversion

Hans-Jürgen Götze¹ and Florian Wellmann²

¹Institute of Geosciences, Working group Satellite- and Aerogeophysics, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

²Computational Geoscience and Reservoir Engineering, RWTH Aachen University, Aachen, Germany

Synonyms

Ambivalence; Error statistics; Uncertainty quantification; Unpredictability; Vagueness

Definition

In order to be able to make reliable predictions about the quality and validity of geomodelling, the knowledge and quantification of uncertainties in input and output parameters is of increasing importance. For this purpose, the influence of the uncertainties and their extent on the result of the modelling is investigated by statistical methods. The evaluation of these uncertainties is specifically important when geomodelling is combined with geophysical inversion.

Introduction

In all Earth Sciences, scientific practice involves the construction, validation, and application of scientific models. A model is a construct (of equations, relations, imaginary bodies, etc.) that represents one or more (interesting and relevant) aspects of reality. Thus, a model is always a simplification and it is important that the aspect/parameter of reality to be modelled is also represented.

A geoscientific model is an image of a geoscientific process (4-D) or a structure (3-D) that can be verified on the object and is as consistent as possible in itself, which does not necessarily mean that it fully resembles truth. Since geoscientific processes are always four-dimensional and there are always too many equations and too few parameters available for modelling, models must always be simplified, which generally limits their informative value. Therefore, numerical methods are becoming more and more important, which can provide information about how close a model comes to reality and to what extent the input variables influence the modelling result. This leads to the explanation of the terms “uncertainty,” “sensitivity,” and/or “error.”

Uncertainty

The quantification of uncertainty describes an interdisciplinary field of research with the aim of quantitatively assessing input and output data of technical-scientific calculations, which are subject to high variability or are only partially known, and to determine their influence on relevant target quantities, so-called quantities of interest (e.g., material parameters such as velocities, densities, porosities, resistivities, but also the geometry of rock formations). The main focus is on the analysis of mathematical-physical models, which can be described by means of numerical relations (e.g., differential equations). If one concludes directly from the measured input variables (data) on the geometry and the distribution of parameters, one also speaks of inversion – and vice versa of forward modelling. In addition to verification and validation, the determination of uncertainty is an essential quality assurance measure for the evaluation of modelling results.

Error and Accuracy

Unlike uncertainty, which is understood to be the range of possible values within which the true value of a modelling (or measurement) lies, the term error, as it is normally used, means the difference between a “measured” value and the true

value for a measurement. Since the exact or “true” measured value of a quantity/model/inversion is often not known, the error for a measurement/modelling can rarely be determined. Since the true model is usually not known, the accuracy of a model is also usually not known.

Sensitivity

The sensitivity analysis provides general information on how changes in the input conditions influence a result/modeling, i.e., how sensitive, sensible, or fragile a modeling reacts. In principle, the representation as a sensitivity diagram can also be used in technical-physical disciplines. In all these processes, it is characteristic how much physical input variables such as sensitive temperature, pressure, petrological rock composition, or the quality of these variables affect the modelling result.

Uncertainties in the Transformations of Earth Science Problems into a Mathematical Model: Reasoning Behind Modeling and Basic Concepts

As already mentioned above, a distinction is made between forward modeling (see ► [“Gravity Modeling, Theory and Computation”](#)) and inversion. An example of forward modeling is the solution of differential equations under the assumption of constraints and/or initial conditions in a region whose geometry and physical properties are well known. Mathematically, inversion of geophysical data is always an ill-posed problem because it usually suffers from ambiguity. Data inversion is a central step in processing and interpretation of measured data (among many others references, e.g., Clauser 2014).

Geoscientific information for subsurface investigations is typically derived from diverse measurement types with varying quality, combined with aspects of geological knowledge concerning the geological evolution and the current setting (e.g., Wellmann and Caumon 2018). Based on these considerations, several decisions have to be made to derive a mathematical model about subsurface structures and property distributions, which can then be integrated into forward and inverse modeling workflows. We consider here four important decision steps to derive mathematical models of geological structures:

Reduction and Simplicity Concepts in Geosciences

In the past, it has been repeatedly argued that geological generalizations can be traced back to the basic laws of physics and chemistry. In the case of contradictions between geological observations and geological and physical laws, the physical laws are always regarded as the more fundamental. Geological laws therefore arise from the action of physical laws under certain boundary conditions of planet Earth. This

point of view has the great advantage of offering a unified view of nature; it thus also appears as a direct consequence of the principle of simplicity.

Interface Representations as Reduction of Reality

In subsurface investigations, we are often interested in the spatial distribution of specific properties. A logical approach would therefore be to simulate the evolution of this region from initial conditions and first physical principles. But due to the problem of the reductionist concepts described above, this is not (yet) possible on the typical scale of crustal investigations, where we are often interested in resolutions on the scale of meters (or, at least, hundreds of meters). We therefore need to find a meaningful simplification of the entire complexity of the evolution of the Earth.

A joint feature of most of these spatial elements is that they can be represented in a 2-D interface, and this realization has been at the core of geological mapping since the first geological maps. It has logically been extended into the 3-D domain with a wider application of 3-D modeling methods. Here, we will focus in more detail on this concept, describe methods for interface description, the treatment of uncertainties, as well as the link to geophysical forward modeling and inversion.

Scale and Typical Representations

Another important first step in any geoscientific investigation is the decision on the relevant scale of features that should be considered. This decision depends on the aim of the investigation, as well as on data availability and the measurement method with its respective resolution.

The distinction of scales described above is not only relevant for different types of investigations, as they are often investigated by several subdisciplines in geosciences, but also for practical modeling purposes and the specific methods, which developed often within these subdisciplines over many years.

It is important to point out that this distinction of methods to typical scales is not necessarily strict – for example, methods from two-point geostatistics have successfully been used also in large-scale investigations, and interface methods can also be applied to distinguish sedimentary features. Yet, this typical delineation is a useful starting point to select suitable methods for specific scales of investigation.

Considerations of Complexity

Geological interfaces appear in a variety of complexity in different geological settings. The consideration of the relevant level of complexity is important for each modeling study/inversion, as it may have an influence on the selection of suitable mathematical models. This aspect of complexity from the viewpoint of underground modeling is illustrated in Wellmann and Caumon (2018).

Due to this low level of complexity, a wide variety of existing interpolation tools, both conventional deterministic interpolations (e.g., inverse distance, spline surfaces), as well as a wide range of geostatistical methods can be applied to describe these surfaces.

As soon as we consider faults, this situation becomes more complex (Fig. 1). For normal faults and fault networks, “2.5-D”-approaches could still be used, but we obtain domains where no value is assigned. However, for reverse faults and thrust faults, we now obtain points with multiple z-values. For these cases, we now require geometric interpolation methods with more flexibility, and a variety of implicit and explicit interface representations have been developed for this purpose (see below).

The requirement for a more flexible interface description is even more obvious when we consider complex structures, for example, related to multiple deformed systems or dome structures. Here, it is obvious that a simple model is not anymore sufficient due to the complexity of the interface itself, and increasingly due to nontrivial interactions between multiple sequences and unconformities.

Mathematical Models of Interfaces

General Interface Description

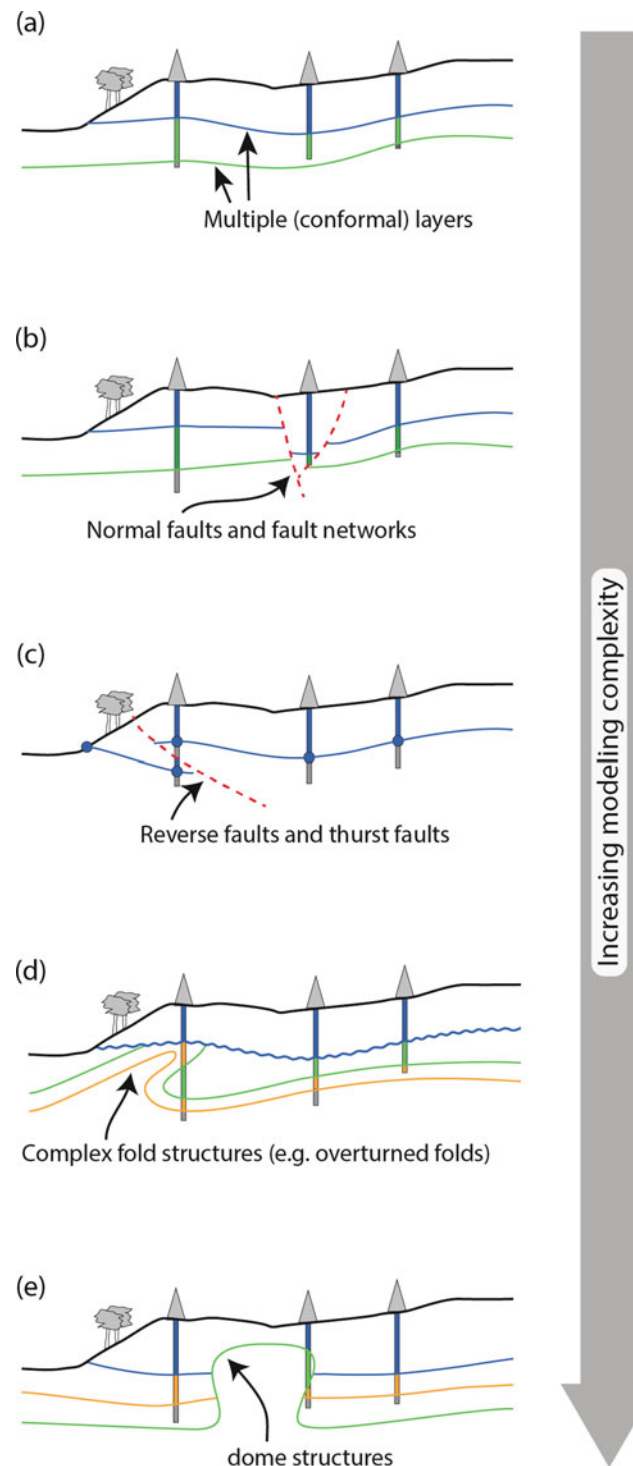
Many commonly used geomodel representations are based on linear interpolation of basis functions, for example, from the class of Radial Basis Functions or Covariance functions (see Wellmann and Caumon 2018). The practical challenge is to decide on a suitable set of basis functions and the combination of multiple functions to describe the interaction in complex settings.

Map-Based Interpolation Approaches

As described above, in map-based “2.5-D” settings, a variety of spatial interpolation approaches can directly be utilized for the purpose of interface description. Here, the basis function is now centered at a location (x,y) in space. A simple choice for the basis function is then the inverse distance function. Other typical choices for the basis functions are Radial Basis Functions (RBF)'s and the related group of geostatistical interpolation approaches.

Explicit and Implicit 3-D Interpolations

An important distinction in the geometric representation of geological interfaces is the use of explicit or implicit methods. Explicit methods describe surfaces directly using either parametric descriptions or local interpolation approaches. The advantage of implicit approaches is that the interface position can directly be evaluated. However, the topological relationship between different geometric elements is fixed in these approaches – an aspect that may be desired in certain



Uncertainties in Geomodelling Related to Geophysical Inversion, Fig. 1 Complexity of interfaces in typical geological settings (modified after Wellmann and Caumon 2018)

circumstances, but a limitation in others. In implicit representations, interfaces are defined as isosurfaces of a scalar field of co-dimension one (i.e.: 2-D interfaces are defined as isosurfaces of a 3-D scalar field). A classic example for these approaches is the Level-Set method, see Wellmann and Caumon (2018).

Uncertainty in the Detection of Model Interfaces by Geological and Geophysical Methods

Modellers have stated the requirement to combine geophysical and geological information. Fullagar et al. (2008) describe geophysical inversion as an extension of geological modeling, and Jessell et al. (2010) express the same view, from the viewpoint of geological modeling. Wellmann and Caumon (2018) discuss similarities between the classical data-driven geophysical inverse workflows and the approaches integrating geological modeling, both centered around the integration of geoscientific knowledge into geophysical inversions.

The most widely used method to consider geological information in the form of geological models is the use of these models to initialize a subsurface parameterization for a geophysical inversion or process simulation. All of these approaches are based on the concept that domains with similar properties can be defined in the subsurface, based on similar geological processes in the formation of a region (Fig. 2). As described above, these geometric domains can be subject to uncertainties. Two general approaches can be distinguished at the moment on the consideration of these uncertainties in subsequent inverse studies: in the form of an

uncertainty identifier, assigned to subdomains (often voxels/cells), which contain the properties that are inverted; or in the form of fully integrated joint geological-geophysical inversions, where the 3-D geological model itself is continuously adjusted.

As Model Describing Initial State

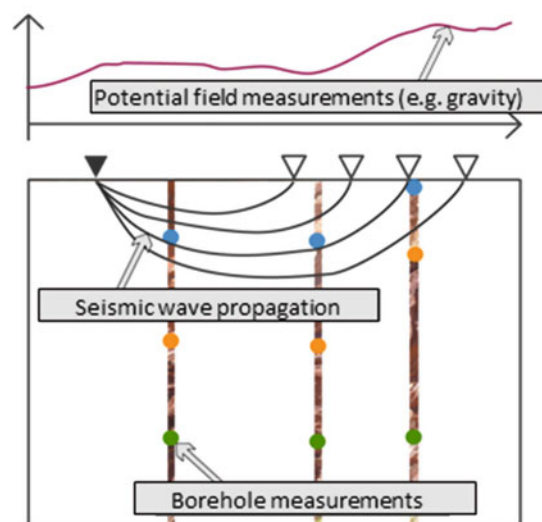
Today's geophysical interpretation requires an interdisciplinary approach, particularly when considering the available amount of state-of-the-art information contained in comprehensive databases employing seismic, gravity, geothermics, and electromagnetic together with geological and petrological studies. The most widely used implementation of 3-D geological models into geophysical inverse approaches is the mapping of properties onto discrete voxels or cells, which contain the properties that are inverted.

An example of interdisciplinary 3-D structural modeling is given by Wehr et al. (2018) presenting a 3-D model of the entire Pyrenees and their forelands by different geophysical observations and data (seismics, borehole, gravity). In particular, basin modeling is part of this, which involves the construction of structures that contain an enormous number of natural deposits (see chapter ► “Sedimentary Basins” by

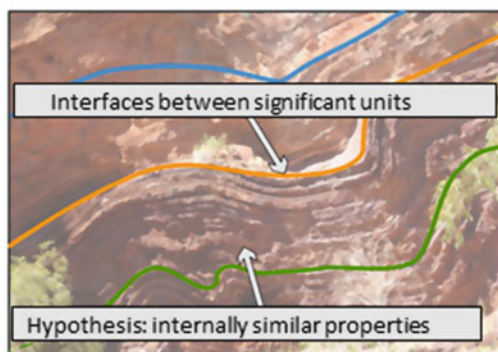
(a) Examples of a rock sequence



(b) Typical measurements



(c) Interpolation of important interfaces in space



Uncertainties in Geomodelling Related to Geophysical Inversion, Fig. 2 Example of the concept of interface description and the estimation on the basis of geological and geophysical observations and measurements

Magdalena Scheck-Wenderoth). In their study, the authors consider the initial structural setting and uncertainties in the form of constraints on the subsequent gravity and seismic tomography inversion steps. The analysis of misfits then allows for an iterative (manual) adjustment of the 3-D geological model.

The integration of human geo-expertise, different techniques, and different geo-disciplines play an important role. This aspect becomes particularly manifest in interactive modeling and automated techniques (Inversion). The interactive 3-D modeling software IGMAS+ (e.g., Alvers et al. 2014) can be viewed as an example. Instead of providing either voxel or grid-based and triangulated facet models, it has become more common now to use hybrid models where all three model types are combined. Single value grids are used to define simple density boundaries like sea bottom or top basement. Voxel-cubes can then be used to describe more complex density/velocity distributions in selected bodies.

Dealing with hybrid models and many different fields requires high-performance algorithms in order to shoot for real-time responses. For fast modelling, it is essential that after model changes the recalculation of the model is done immediately. An important aspect is resampling of model interfaces. By keeping the shape of a highly resolved, e.g., triangle interface, the algorithm should allow (e.g. Menzel 2016) depending on a user-given threshold – for the reduction of the number of triangles by factors of 10 to 100. An example is shown in Fig. 3.

Joint Inverse Frameworks

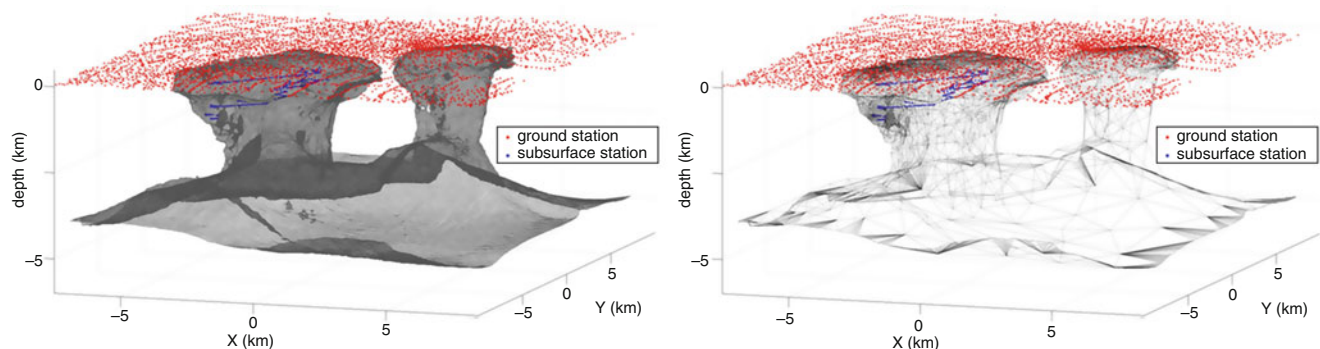
Another approach to combine geological and geophysical information is a full integration in a model-based inverse framework. Here, the geological parameters of the interpolation function (i.e., typically interface points, but orientation

measurements can also be used) are treated as model parameters. In a probabilistic setting, a prior distribution is assigned to these parameters. The geological model (in the form of the 3-D geometric interpolation) then links these parameters through a deterministic forward model to additional observations – both in the form of further geological measurements, which was not included in the interpolation step itself, but also to additional geophysical forward models – in combination with a suitable rock physics model. An example of such an approach is presented in Fig. 4, where a prior 3-D geological model (4a) is parameterized with interface parameters and associated prior distributions (schematically shown in 4b). The model is then linked to further geological observations through likelihood functions (4b, i-iii), as well as geophysical measurements (4c). Finally, the posterior distribution is obtained through a suitable sampling method (here: MCMC with Metropolis-Hastings). The final suite of generated models can then be used to obtain spatial representations of uncertainty, for example, using measures from information theory (Wellmann and Regenauer-Lieb 2012).

Uncertainties in the Output of Material Parameters

Relevance of Rock Physics Models

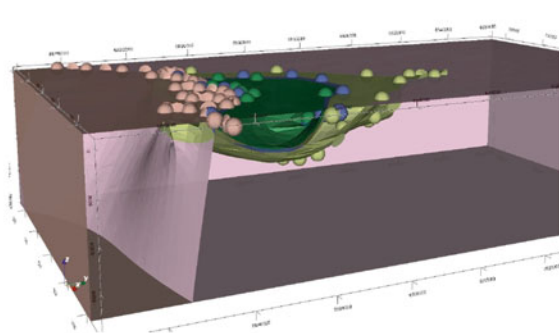
As described above, there is a need for multidisciplinary approaches to more complex issues. Especially in the analysis of lithospheric parameters (densities, velocities, electrical resistivity) an interdisciplinary approach is necessary, since a profound geological and petrological understanding of the near-surface units is needed to analyze and model the geophysical knowledge of the underground units; refer to, e.g.,



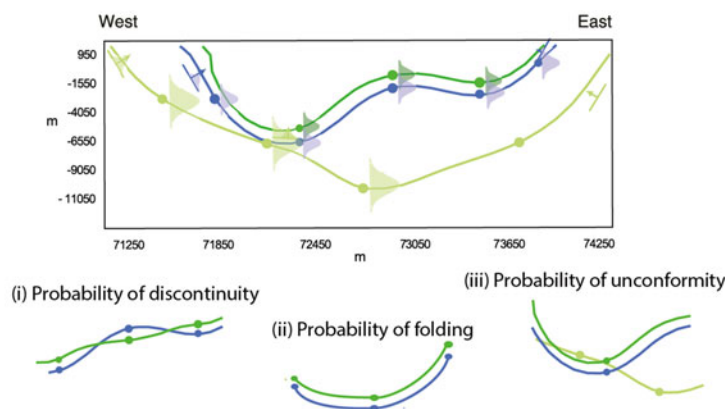
Uncertainties in Geomodelling Related to Geophysical Inversion, Fig. 3 A simplified outer structure of a salt dome in the Gifhorn Trough (Götze et al. 2017). Its geometry built by only 13 900 triangles (right) reduced from originally 380 000 triangles (left) which were obtained by high resolution 3-D-seismic surveys in the area (Menzel 2016). For 3-D gravity modeling 6 000 surface gravity stations (red points, surface stations) and 250 subsurface stations in a salt mine (blue points) were

needed. This structure causes a model gravity field which differs slightly from the gravity field of the original geometry; the differences are smaller than the error of the measurements. Since the distance between measurements and triangulated hull was used as a constraint for the reduction, the model shows a higher geometry resolution near the surface and near the subsurface measurements

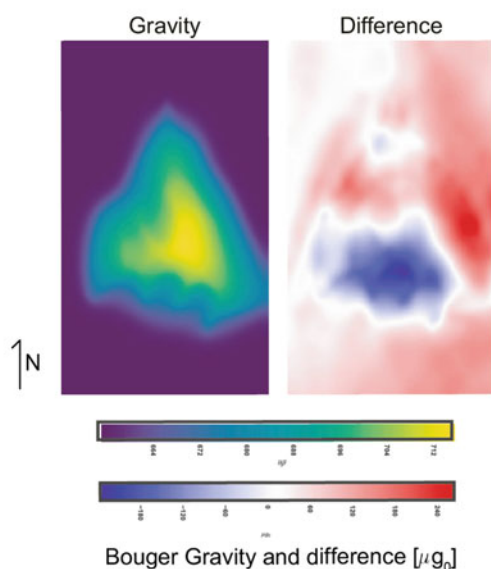
(a) Prior structural geological model



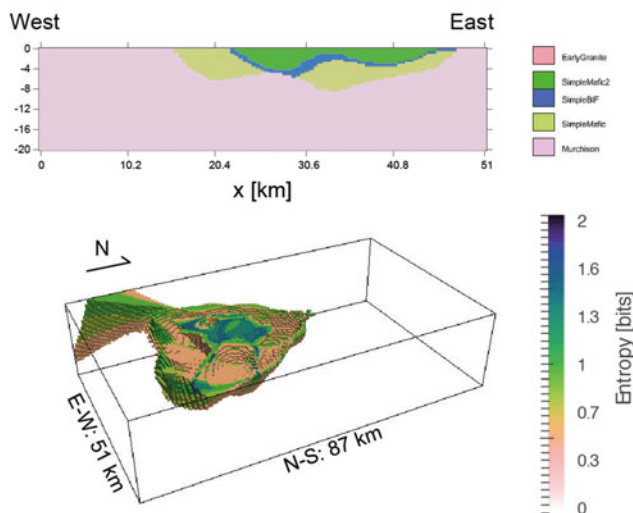
(b) Geological likelihood



(c) Geophysical likelihood



(d) Inverted geometric model and representation of uncertainty



Uncertainties in Geomodelling Related to Geophysical Inversion, Fig. 4 Example of a probabilistic model-based joint geological-geophysical inversion (modified after Wellmann et al. 2018)

Sobolev and Babeyko (1984), Bousquet et al. (2005), Afonso et al. (2008), and Fullea et al. (2009).

When considering changes in physical rock parameters with depth, several factors must be considered, as temperature and pressure changes as well as differences in geochemical and mineralogical composition have a significant influence on rock density (see ► “Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity”).

Advanced computer technology and development of powerful modeling software enable the calculation of new models in which a complex consideration of pressure, temperature, and geochemistry is possible. These programs can also be used to model rock types that are not exposed at the surface. The significance of the parameter change in the lithosphere as

a function of chemistry and the resulting mineral paragenesis due to metamorphic reactions has been shown, for example, in the work of Bousquet et al. (2005), Duisterhoeft et al. (2012), and Semprich et al. (2010).

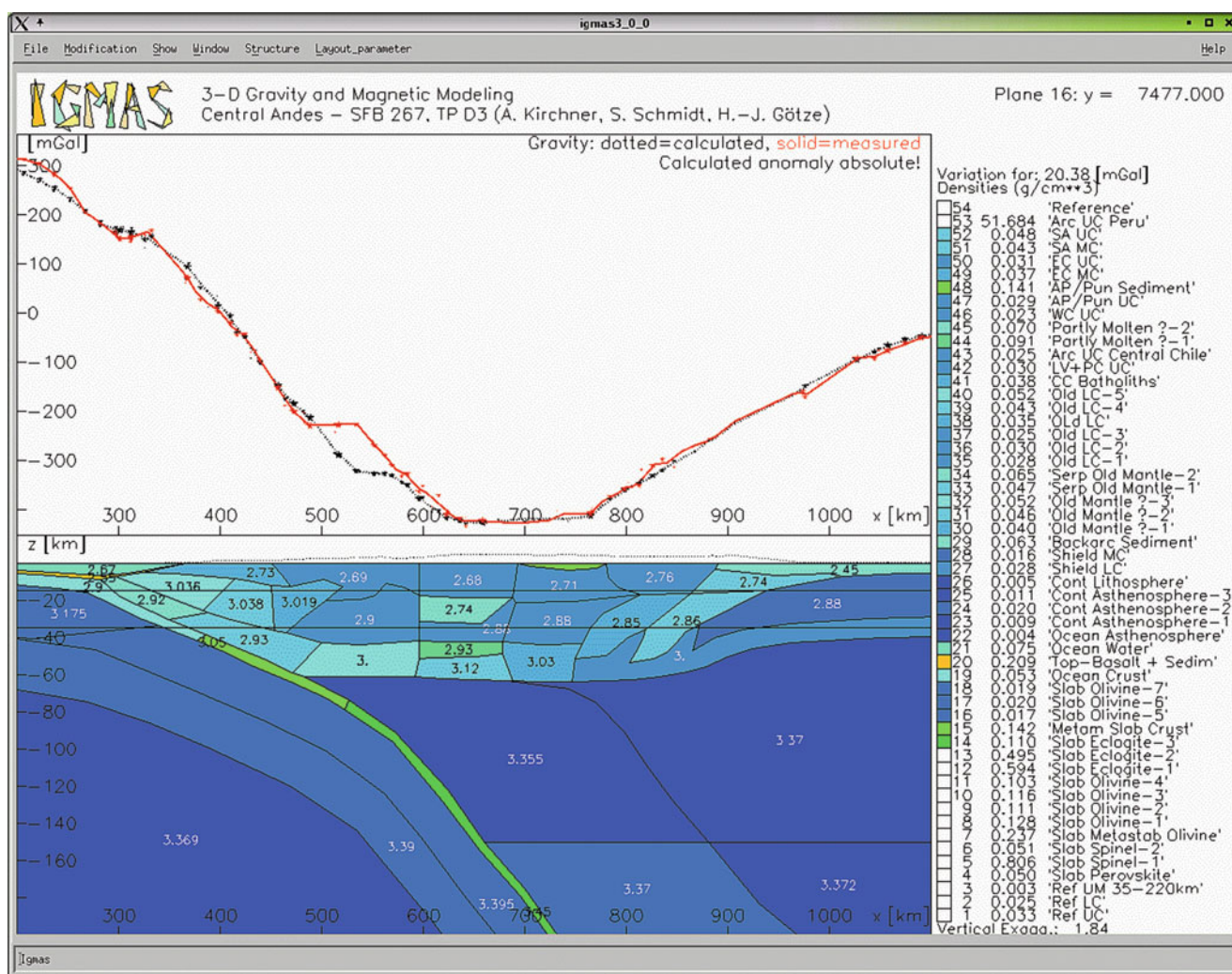
Visualization of Uncertainties as Output of Density Models: An Example from the Central Andes

To visualize model uncertainty of rock parameter density, we used the facilities of our inhouse 3-D-modeling software IGMAS to give an example from the Central Andes. The model is constrained by independent geophysical data (Kirchner 1997). We estimated the uncertainty in the density selected for each body as the tolerable variation in this parameter that would produce a variation in the calculated Bouguer anomaly equal to the estimated error of the observed anomaly,

i.e., $\pm 20 \times 10^{-5} \text{ m/s}^2$. The uncertainties of lithospheric density distribution are presented in Fig. 5. For more detailed information refer also to Tassara et al. (2006). To estimate these uncertainties, we calculated the vertical variation of the vertices defining each discontinuity that produce an average change in the calculated Bouguer anomaly equal to the tolerance limit of $\pm 20 \text{ mGal}$. The uncertainty associated with each discontinuity should be less than that of the vertices. This is because the 3-D effect on the gravity field produced by vertically moving the whole discontinuity is larger than the

effect of moving single vertices. We estimate the expected uncertainty in the depth of each discontinuity to be 20% of the average uncertainty calculated for single vertices.

This analysis gives an indication of the tolerable range of variation in densities for each body. The approach is by no means a tool to evaluate the accuracy of the whole model, and it does not guarantee the resolution of the method. However, this information can be used to analyze the meaning of the final geometry obtained for density discontinuities that are not constrained by independent information.



Uncertainties in Geomodelling Related to Geophysical Inversion, Fig. 5 This screenshot of the Central Andean model by Kirchner (1997) shows a vertical section approximately at the latitude of 21 deg south - from the Pacific to the Chaco of Argentina. The upper box contains the intercept line of the comparison gravity (complete Bouguer anomaly) and the intercept line of the model gravity curve. The lower part shows the density values used for the individual model bodies in white and black notation. The subducted Nazca plate can be clearly identified, the area of the crust and the lithospheric mantle to a depth of approx. 200 km.

The colors symbolize the range of variation possibilities of a density change in the respective rock bodies: dark blue areas show that the density can only be changed slightly in order not to exceed the threshold of $20 \times 10^{-5} \text{ m/s}^2$. The color spectrum then ranges from "light blue" and "green" to orange tones, which symbolize larger variations. The table on the right shows the colors of the lower box together with the "geological" descriptions of the model bodies and the range of variation in g/cm^3 ($1 \text{ g/cm} = 1000 \text{ kg/m}^3$). The smaller the geological body, the greater the possibility to vary the density

Summary

Geophysical inverse methods and geological modeling are increasingly combined to obtain better integrated models of the subsurface – with a consideration of uncertainties. We presented here the reasoning about the use of geological models as well as mathematical methods to represent relevant interfaces, the link to geophysical inverse approaches, as well as to rock physics models. All of these aspects are subject to improvement and addressed in current research.

An important area of future development is calculation and visualization of model uncertainties. Interactive inversion produces thousands of models while exploring the parameter space. Displaying all models with a “similar fit” on top of each other would give an immediate insight into model quality and show where the model is well defined by gravity/magnetic/other geophysical methods and where uncertainty is significant. The challenge is to develop a user-friendly 3-D-overlay technique. First experiments with “fog visualizations” and the possibility of making (low cost) 3-D-plots of even “complicated” underground structures became a promising challenge (e.g., Götte et al. 2017). Alternative methods are the use of transfer functions, for example, plots of probability or entropy (e.g., Wellmann and Regenauer-Lieb 2012). We expect that more interesting insights will come from the active field of uncertainty visualization and communication related to climate models.

Cross-References

- [Earth, Density Distribution](#)
- [Electrical Properties of Rocks](#)
- [Gravity Method, Principles](#)
- [Gravity Modeling, Theory and Computation](#)
- [Gravity, Data to Anomalies](#)
- [Heat Flow, Continental](#)
- [Inverse Theory, Artificial Neural Networks](#)
- [Inverse Theory, Global Optimization](#)
- [Inverse Theory, Linear](#)
- [Inverse Theory, Monte Carlo Method](#)
- [Sedimentary Basins](#)
- [Seismic Properties of Rocks](#)
- [Seismic, Velocity, and Density Relationships](#)
- [Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity](#)

Bibliography

Afonso JC, Fernández M, Ranall G, Griffin WL, Connolly JAD (2008) Integrated geophysical-petrological modeling of the lithosphere and sublithospheric upper mantle: methodology and

- applications. *Geochem Geophys Geosyst* 9:Q05008. <https://doi.org/10.1029/2007GC001834>
- Alvers MR, Götte HJ, Barrio-Alvers L, Plonka C, Schmidt S, Lahmeyer B (2014) A novel warped-space-concept for interactive 3-D-geometry-inversion to improve seismic imaging. *First Break* 32(4):81–87
- Bousquet R, Goffé B, Le Pichon X, de Capitani C, Chopin C, Henry P (2005) Comment on “subduction factory: 1. Theoretical mineralogy, densities, seismic wave speeds, and H₂O contents” by Hacker BR, Abers GA, and Peacock SM. *J Geophys Res Solid Earth* 110: B02206. <https://doi.org/10.1029/2004JB003450>
- Clauser C (2014) Einführung in die Geophysik, vol 407. Springer Spektrum, Berlin/Heidelberg. ISBN: 978-3-642-04495-3
- Duesterhoeft E, Bousquet R, Wichura H, Oberhänsli R (2012) Anorogenic plateau formation: the importance of density changes in the lithosphere. *J Geophys Res Solid Earth* 117., B7:1–13. <https://doi.org/10.1029/2011JB009007>
- Fullagar PK, Pears GA, McMonnies B (2008) Constrained inversion of geologic surfaces – pushing the boundaries. *Lead Edge* 27(1): 98–105
- Fullea J, Afonso JC, Connolly JAD, Fernández M, García-Castellanos D, Zeyen H (2009) LitMod3D: an interactive 3-D software to model the thermal, compositional, density, seismological, and rheological structure of the lithosphere and sublithospheric upper mantle. *Geochem Geophys Geosyst* 10:Q08019. <https://doi.org/10.1029/2009GC002391>
- Götte HJ, Schmidt S, Menzel P (2017) Integrative interpretation of potential field data by 3D-modeling and visualization. *Oil Gas Eur Mag*. <https://doi.org/10.19225/171206>
- Jessell MW, Ailleres L, Kemp AE (2010) Towards an integrated inversion of geoscientific data: what price of geology? *Tectonophysics* 490(3–4):294–306
- Kirchner A (1997) 3D density modeling to adapt the gravity and gravity potential fields of the Central Andes (in German). Dissertation FU Berlin, Berliner Geowissenschaftliche Abhandlungen, Series B, vol 25, Berlin, 98p
- Menzel P (2016) Neue numerische Methoden zur Bearbeitung und 3D-Interpretation von Geodaten und -modellen in interdisziplinärer Forschung. Dissertation, Christian-Albrechts-Universität Kiel, S. 152. http://macau.uni-kiel.de/receive/dissertation_diss_00021287
- Semprich J, Simon N, Podladchikov Y (2010) Density variations in the thickened crust as a function of pressure, temperature, and composition. *Int J Earth Sci* 99(7):1487–1510. <https://doi.org/10.1007/s00531-010-0557-7>
- Sobolev S, Babeyko AY (1984) Modeling of mineral composition, density and elastic wave velocities in anhydrous magmatic rocks. *Surv Geophys* 15:515–544
- Tassara A, Götte HJ, Schmidt S, Hackney R (2006) Three-dimensional density model of the Nazca plate and the Andean continental margin. *J Geophys Res* 111:B09404. <https://doi.org/10.1029/2005JB003976>
- Wehr H, Chevrot S, Courrioux G, Guillen A (2018) A three-dimensional model of the Pyrenees and their foreland basins from geological and gravimetric data. *Tectonophysics* 734–735:16–13. <https://doi.org/10.1016/j.tecto.2018.03.017>
- Wellmann JF, Caumon G (2018) 3-D structural geological models: concepts, methods, and uncertainties. In: Schmeling C (ed) *Academic Press, Advances in Geophysics* 59:1–121
- Wellmann JF, Regenauer-Lieb K (2012) Uncertainties have a meaning: information entropy as a quality measure for 3-D geological models. *Tectonophysics* 526:207–216
- Wellmann JF, De La Varga M, Murdie RE, Gessner K, Jessell M (2018) Uncertainty estimation for a geological model of the sandstone greenstone belt, Western Australia—insights from integrated geological and geophysical inversion in a Bayesian inference framework. *Geol Soc Lond, Spec Publ* 453(1):41–56

Unified Scaling Law for Earthquakes that Generalizes the Fundamental Gutenberg-Richter Relationship

Vladimir Kossobokov

Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Sciences (IEPT RAS), Moscow, Russia

Definition

The hierarchy of movable lithospheric volumes, from grains of rock to tectonic plates, composes a complex dynamical system (Keilis-Borok 1990). A sudden movement of the system volume that originates seismic waves is an *earthquake* which characteristics can be determined from interpretation of seismograms, if any, recorded by seismic instruments. Numerous approaches to sizing earthquakes provided definitions of earthquake *magnitude* (Richter 1935; Bormann 2011).

Naturally, due to hierarchical organization of the lithosphere, the number of earthquakes in a region is scaled by magnitude, in particular, according to the Gutenberg-Richter relationship for the worldwide seismicity (Gutenberg and Richter 1944, 1954):

$$\log_{10} N(M) = a + b \cdot (8 - M), M_- \leq M \leq M^- \quad (1)$$

where the coefficient a characterizes the average level of seismic activity in terms of the annual rate of magnitude 8 earthquakes, the coefficient b is the ratio $N(M)/N(M-1)$, and M_- and M^- are the limits of the magnitude range where the relationship holds.

Introduction

Seismic hazard assessment requires an adequate understanding of the distribution of earthquakes of different size. Until recently the Gutenberg-Richter relationship (1) was the one of common use in the estimation of earthquake hazard and risks, as well as in other researches aimed at investigation of seismic activity in an area of interest. The existing maps of seismic hazard are correlated to the local estimates of the coefficient a obtained from sampling the regional catalogue of earthquakes which data is usually too small for a reliable quantification at fine grid. Note that the probabilistic seismic hazard maps overestimate risks in aseismic areas and underestimate them in seismic areas due to (i) subjective choice of probability model (usually, Poisson distribution in time and Gauss

distribution in space) and (ii) underestimation of the role of standard scaling in proportion to the area, in km^2 .

Unified Scaling Law for Earthquakes

The abovementioned drawbacks of seismic hazard analysis and assessment were clearly recognized in the late 1980s. Kosobokov and Mazhenov (1988, 1994) suggested generalization of the Gutenberg-Richter relationship as follows:

$$\log_{10} N(M, L) = A + B \cdot (5 - M) + C \cdot \log_{10} L,$$

$$M_- \leq M \leq M^- \quad (2)$$

where the coefficient C is the fractal dimension of the set of epicenters that complements to the coefficients A and B which are similar to a and b from Equation (1). The coefficient C shows how the number of earthquakes, N , is changing with linear dimension, L , of an enveloping area. It was shown that C is significantly different from 2 and correlates with the geometry of tectonic structures; high values of C about 1.5 correspond to the regions with dense patterns of faults of different strike and high degree of fracturing, whereas the values of about 1 are related to regions with a predominant major fault zone. Moreover, it was demonstrated that in the case of the Lake Baikal region with the area of 1,500,000 km^2 and $C = 1.25$, the inclusion of aseismic areas leads to underestimation of seismic activity in an area of 1000 km^2 by a factor of 15 and to overestimation by a factor greater than 2 when a characteristic of seismic activity over 1000 km^2 is computed for a grid 10×10 km.

Bak et al. (2002), independently, gave an alternative formulation making use of the inter-event time between earthquakes in Southern California instead of their number. They measured $P_{S,L}(T)$, the distribution of waiting times T , between earthquakes occurring within range L whose magnitudes $M > m = \log(S)$. They found the data collapse onto a single function $f(x)$, which is approximately constant for $x = T \cdot S^{-b} \cdot L^{df}$ in range from below 10^{-2} to above 10^3 , i.e., about constant over 8 orders of magnitude. According to Bak et al. (2002), Equation (3) expresses the unified scaling law for earthquakes, USLE:

$$T^\alpha \cdot P_{S,L}(T) = f(T \cdot S^{-b} \cdot L^{df}) \quad (3)$$

where the Omori law exponent $\alpha \approx 1$, the Gutenberg-Richter value $b \approx 1$, and fractal dimension $df \approx 1.2$. Evidently, $S^{-b} \cdot L^{df}$ is the average number of earthquakes per time unit with magnitude $M > \log(S)$ occurring within the range L , and x is the average number of such earthquakes occurring within

a time interval T . Note that the average of waiting times between earthquakes of magnitude M within $L \times L$ in a unit time is in inverse proportion to $N(M, L)$; b and df are equivalent to B and C of Equation (2). Thus, the USLE, in either formulation (2) or (3), states that the distribution of waiting times depends only on the control parameter $10^{-BM} \cdot L^C$ or $S^{-b} \cdot L^{df}$, correspondingly.

Using the USGS/NEIC Global Hypocenters Data Base, 1964–2001, and a robust box-counting algorithm, Nekrasova and Kossobokov (2002) managed to map the values of A , B , and C in every place on the Earth marked by earthquake occurrence, where the catalog of shallow earthquakes permitted an estimation. Unlike the first studies (Kossobokov and Mazhenkov 1988, 1994; Bak et al. 2002), the values of A , B , and C were determined as averages of numerous estimates achieved in a hundred of randomly rotated box-counting settings at each of the 9276 seismic locations on a regular $\frac{1}{2}^\circ \times \frac{1}{2}^\circ$ grid. Figure 1 shows the global maps of the A , B , and C coefficients at 6753 grid points where reliable estimates based on the USGS Global Hypocenter Data Base system, 1964–2001, were possible; the values of A range from under -1.0 to above 0.5 , which correspond to the rates that differ by a factor of 30 or more; B resides mainly from 0.6 to 1.1 , while C changes from under 1 to 1.4 and larger. The error of the coefficient average estimate is below 0.01 for almost all sites of determination suggesting a high degree of agreement with observations in a wide range of earthquake magnitudes and spatial distances.

Expanding the application of the USLE, Baiesi and Paczuski (2004) introduced the space-time-magnitude proximity η_{ij} between any two earthquakes i and j based on the control parameter $\eta = \tau \times 10^{B \times (S-M)} \times L^C$. The properties of this proximity measure with physical motivations for cluster analysis of seismic phenomena at regional and global scales are discussed at length in Zaliapin et al. (2008) and Zaliapin and Ben-Zion (2016). They suggested for the foreshock and aftershock identification a split of η_{ij} into $R_{ij} \times T_{ij} = (\rho_{ij} \times 10^{-C \times p \times M_i}) \times (\tau_{ij} \times 10^{-B \times q \times M_i})$, where ρ_{ij} and τ_{ij} are the distance and time span between earthquakes i and j , correspondingly, and $p + q = 1$; unlike the Poisson model, the observed seismicity shows a bimodal distribution of $(\log_{10} T, \log_{10} R)$, as has been demonstrated in numerous regions worldwide. Moreover, this 2D presentation permitted to find out quantitative definitions of the burst-like and swarm-like clusters of earthquakes and to relate those with the Earth heat flow production. Thus, the effective viscosity of the lithosphere is proposed to be the main control of the style of earthquake clustering.

Analyzing variability of seismic activity at a regional scale, Nekrasova et al. (2011) have demonstrated a complex distribution of the USLE coefficients A , B , and C in Italy, which do not display any evident general correlation, although following well-organized attractor in the 3D domain of possible values.

Conclusion

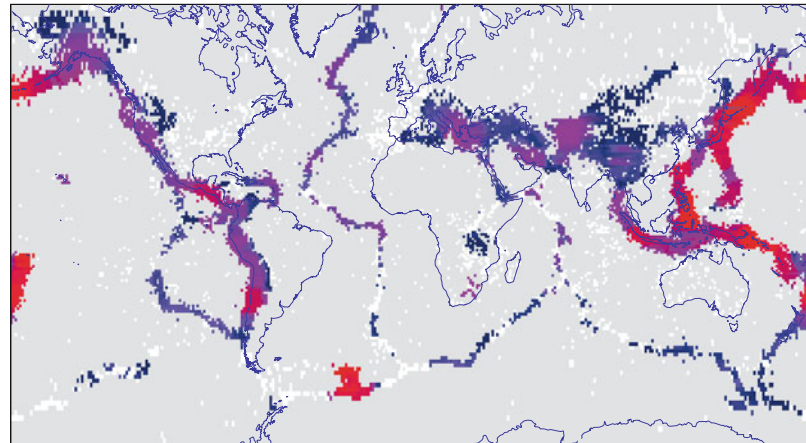
The complexity of using the data from seismic catalogs, where quality differs by location, may affect the results of the USLE-based analysis and interpretation. However, the robust nature of computations confirmed by low values of errors in the bulk of seismic regions allows to conclude general agreement with the assumption of spatial self-similarity and to proceed further investigations into the problem using regional catalogs of earthquakes and separating events into the main-, fore-, and aftershocks. For example, the coefficient C at Los Angeles is about 1.21 both as estimated by Okubo and Aki (1987) for the system of faults in Southern California and as computed by applying the box-counting algorithm to one of the best regional earthquake catalogues; scaling down in proportion to the area from the entire Southern California of about 400 km to the region of Los Angeles, about 40 km in length underestimates the recurrence rate by a factor of $(L_{SCA}/L_{LA})^2 / (L_{SCA}/L_{LA})^{1.21} = 10^{0.79} = 6.2$. By analogy, the underestimation of seismic hazard in a region of about 40 km long at Petropavlovsk-Kamchatsky from the area of Kamchatka about 700 km in length where $C = 1.26$ rises to a factor $(L_K/L_{PK})^2 / (L_K/L_{PK})^{1.26} = 17.5^{0.74} = 8.3$. A comprehensive analysis of seismic hazard at 1140 cities and urban agglomerations in seismic regions worldwide due to scaling in proportion to the area (Nekrasova and Kossobokov 2005) reveals, in addition to abovementioned cases, potential underestimation of a factor of 5.4 for Tokyo (Japan), 7.2 for San Francisco (USA), about 20 for Ulan-Ude (Republic of Buryatia, Russian Federation), etc. Such a discrepancy in the evaluation of seismic hazard propagates nonlinearly into erroneous estimations of risks. An error in rate of a factor of 6 (in assumption of $B = 1$) corresponds to an error about 0.8 in magnitude, which explodes exponentially in the underestimation of seismic energy release by a factor 15 , corresponding to a factor above 4.5 for expected fatalities (Wyss et al. 2012). The confirmed multiplicative scaling of earthquakes changes the traditional view on the recurrence of catastrophic events and helps estimating seismic hazard in a more adequate way.

The analysis of space-time-energy distributions of seismic activity at regional scale confirms the existence of different, intermittent levels of rather long-term steady seismic activity characterized with a near constant value of the USLE control parameter η . The region switches from one level to another one in a shorter-term transition associated with strong, catastrophic events. During such a switch, seismic activity may follow very different scenarios with inter-event scaling of different kinds, such as constant, logarithmic, power law, exponential rise/decay, or a mixture of those as observed in the case of the 2016 strong earthquakes in Italy (Kossobokov and Nekrasova 2017). Evidently, these results do not support the presence of universality in seismic energy release, but the

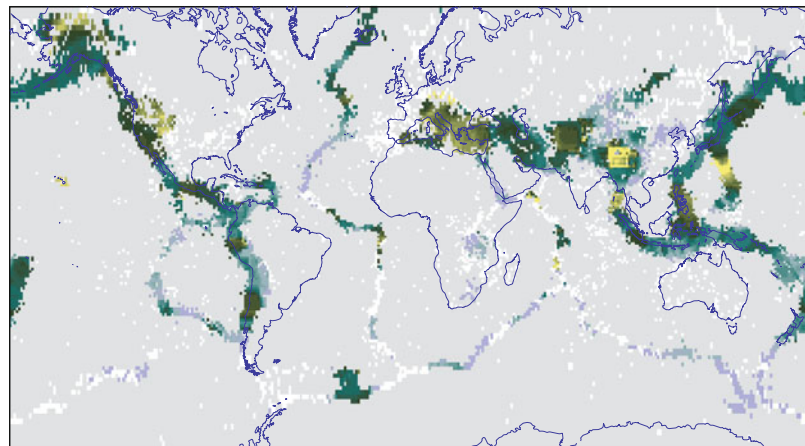
Unified Scaling Law for Earthquakes that Generalizes the Fundamental Gutenberg-Richter Relationship,

Fig. 1 The global maps of the A , B , and C coefficients (Nekrasova and Kossobokov 2002). Cells at $\frac{1}{2}^\circ \times \frac{1}{2}^\circ$ regular grid with epicenters of $M \geq 4.0$ quakes were centers of the five-level hierarchies of square areas with the sides $\frac{1}{2}^\circ$, 1° , 2° , 4° , and 8° of the Earth meridian. The USGS Global Hypocenter Data Base system, 1964–2001, permitted to investigate systematically regions from a wide range of seismic activity. For earthquakes with hypocenters above 100 km, A varies mainly from below -1 to above 0.5 (the corresponding annual rate of magnitude 5 earthquakes differ by a factor up to 30 or more); the magnitude exponent B varies mainly from 0.7 to 1.2 with a sharp maximum of density at 0.9 , while the fractal dimension C changes from under 1 to 1.6 with a maximal density within 1.1 – 1.3 .

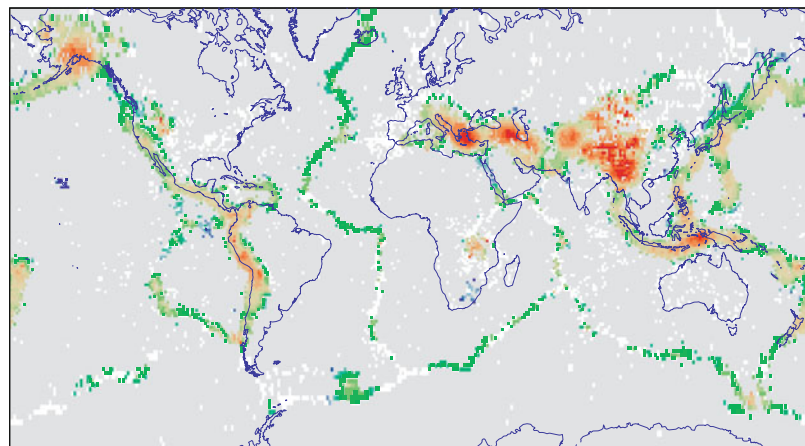
$$\text{Log } N(M, L) = A + B(5 - M) + C \text{Log } L$$



Log of recurrence rate = **A** -1.2 -0.7 -0.2 0.3 0.8 per year



Balance between Magnitude ranges = **B** 0.5 0.7 0.9 1.1 1.3 per unit of magnitude



Fractal dimension = **C** 0.5 0.75 1.0 1.25 1.5 per order of distance in degrees

existence of different statistics of scaling that cannot be uniformly rescaled onto a single, universal curve, even at nearby locations.

Thus, the observed variability of seismic activity provides important constraints on modelling earthquake sequences by geophysicists and can be used to improve local seismic hazard assessments including those of earthquake forecast/prediction techniques. The better understanding of the USLE opens new horizons in the investigation of earthquake occurrence.

Cross-References

- [Earthquake Prediction, M8 Algorithm](#)
- [Earthquake, Aftershocks](#)
- [Earthquake, Foreshocks](#)
- [Earthquake, Magnitude](#)
- [Fractals and Chaos](#)
- [Lithosphere, Mechanical Properties](#)
- [Seismic Hazard](#)
- [Statistical Seismology](#)

Bibliography

- Baiesi M, Paczuski M (2004) Scale-free networks of earthquakes and aftershocks. *Phys Rev E* 69:066106
- Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. *Phys Rev Lett* 88:178501–178504
- Bormann P (2011) Earthquake, magnitude. In: Gupta H (ed) *Encyclopedia of solid earth geophysics*. Springer, Heidelberg, pp 207–217
- Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. *Bull Seismol Soc Am* 34:185–188
- Gutenberg B, Richter CF (1954) *Seismicity of the earth and associated phenomena*, 2nd edn. Princeton University Press, Princeton, 310 pp
- Keilis-Borok VI (1990) The lithosphere of the Earth as a nonlinear system with implications for earthquake prediction. *Rev Geophys* 28:19–34
- Kosobokov VG, Mazhkenov SA (1994) On similarity in the spatial distribution of seismicity. In: Chowdhury DK (ed) *Computational seismology and geodynamics*. American Geophysical Union, Washington, DC, pp 6–15
- Kosobokov VG, Mazhkenov SA (1988) Spatial characteristics of similarity for earthquake sequences: fractality of seismicity. Lecture notes of the workshop on global geophysical informatics with applications to research in earthquake prediction and reduction of seismic risk (15 Nov. 16 Dec., 1988), ICTP, Trieste, 15 pp
- Kosobokov VG, Nekrasova A (2017) Characterizing aftershock sequences of the recent strong earthquakes in Central Italy. *Pure Appl Geophys* 174:3713–3723
- Nekrasova A, Kosobokov V (2002) Generalizing the Gutenberg-Richter scaling law. *EOS Trans AGU* 83(47), Fall Meet Suppl Abstract NG62B-0958
- Nekrasova A, Kosobokov V (2005) Unified Scaling Law for Earthquakes: Mega-cities and urban agglomerations. *Eos Trans AGU* 86(52), Fall Meet Suppl Abstract S23A-0229
- Nekrasova A, Kosobokov V, Aoudia A, Perezan A, Panza GF (2011) A multiscale application of the unified scaling law for earthquakes in the central Mediterranean area and Alpine region. *Pure Appl Geophys* 168:297–327
- Okubo PG, Aki K (1987) Fractal geometry in the San Andreas Fault system. *J Geophys Res* 92:345–356
- Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seismol Soc Am* 25(1):1–32
- Wyss M, Nekrasova A, Kosobokov V (2012) Errors in expected human losses due to incorrect seismic hazard estimates. *Nat Hazards* 62(3):927–935
- Zaliapin I, Ben-Zion Y (2016) A global classification and characterization of earthquake clusters. *Geophys J Int* 207:608–634
- Zaliapin I, Gabrielov A, Keilis-Borok V, Wong H (2008) Clustering analysis of seismicity and aftershock identification. *Phys Rev Lett* 101:018501

Vertical Seismic Profiling

James W. Rector III¹ and Maria-Daphne Mangriotis^{2,3}

¹Department of Civil and Environmental Engineering,
University of California at Berkeley, Berkeley, CA, USA

²School of Geosciences and Grant Institute of Earth Science,
University of Edinburgh, Edinburgh, UK

³Institute of GeoEnergy Engineering, Heriot-Watt University,
Edinburgh, UK

Synonyms

VSP

Definition

Vertical seismic profiling A geophysical technology for measuring the seismic properties in a *profile* of the earth using a set of sources and receivers, either of which is placed along the depth (*vertical*) axis.

Introduction

Vertical seismic profiling (commonly abbreviated to the name VSP) is a major technique in geophysical exploration, which involves measuring downgoing and upgoing seismic wavefields through a stratigraphic sequence. Shortly after the first attempts of developing surface reflection seismology, where both sources and receivers were at the surface, the line of research sets out with the idea of burying either sources or receivers in boreholes. McCollum and LaRue (1931) proposed that local geologic structure could be determined from measuring the arrival of seismic energy from surface sources with buried receivers. This geometry forms the basic principle of determining seismic wave propagation velocities

from VSP measurements even today. The technology of VSP acquisition and processing was initially developed by researchers in the former Soviet Union in the 1950s through the 1970s. It was only later, in the mid-1970s, that non-Soviet geophysicists became systematically engaged in the advance of VSP technology for exploration applications. Today, the fruits of these efforts not only have served in the improvement of velocity analysis but have materialized numerous other applications ranging from structural imaging of complex 3D structures such as faults and salt domes in exploration studies, fracture characterization, and anisotropy measurement to directing drill operations, monitoring secondary recovery processes, CO₂ injection, and microseismic monitoring.

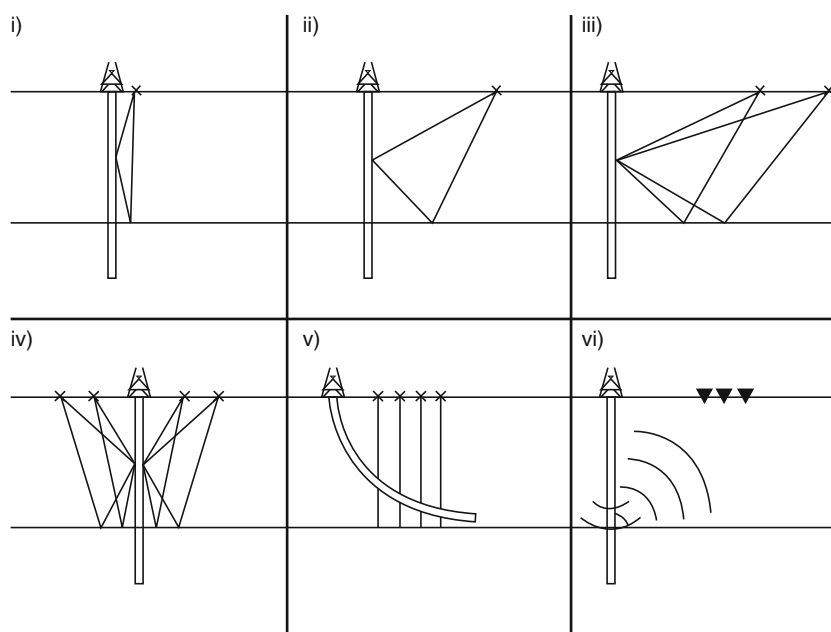
Acquisition Techniques

A variety of data acquisition techniques are currently employed, depending on the survey's application and financial budget. Figure 1 shows the main geometric configurations, which include (i) zero-offset VSP (ZVSP), where the source is near the wellhead above the vertical receiver array; (ii) offset VSP, where the source is located a distance away from the vertical borehole; (iii) multi-offset VSP, for multiple distances of the source to the borehole; (iv) walk-away VSP, where multiple sources are located on a line at the surface and the receivers in a vertical array; (v) walk-above VSP, where the sources are located over receivers buried in a deviated or horizontal well; and (vi) drill-bit SWD (seismic while drilling) VSP, where the drill bit acts as the seismic source and the receivers are on the surface. 3D VSP uses borehole receivers and a 2D surface source geometry (i.e., multi-azimuth, multi-offset, or walk-away VSP).

The common active seismic sources used in VSP include those used in land and marine surface studies (*seismic data acquisition and processing*, *qv*) as well as the noise of the drill bit in the case of SWD-VSP. Alternative sources include naturally occurring or induced microseisms, the latter

Vertical Seismic Profiling,

Fig. 1 Common acquisition geometries for vertical seismic profiling (VSP): (1) zero-offset VSP (ZVSP), (2) offset VSP, (3) multi-offset VSP, (4) walk-away VSP, (5) walk-above VSP, and (6) drill-bit SWD (seismic while drilling VSP)



typically related to hydraulic fracturing or reactivation of fracture networks during hydrocarbon/geothermal production or CO₂ injection. In the case of hydraulic fracturing, perforation shots, which are seismic sources generated from perforation cluster stimulation, can further be used to calibrate the velocity/attenuation model for microseismic monitoring. There is a variety of geophone tools available for use in cased or uncased boreholes, which combine three orthogonal geophones, a hydrophone, and other borehole seismic tools at each recording level. The receiver spacing is set by the depth sampling interval to avoid spatial aliasing. To limit field acquisition time, most commonly multilevel receiver arrays are employed. While conventional 4C sensors are still broadly used, distributed acoustic sensing (DAS) is gaining popularity as a novel seismic acquisition system, which can be deployed for downhole seismic monitoring and imaging. DAS relies on a fiber-optic cable system which records strain changes in response to seismic wave propagation. DAS VSP is especially useful in the case of extreme wellbore conditions and can offer fine sampling along the entire length of the well and permanent monitoring, which makes it ideal for time-lapse VSP which may alternatively suffer from repeatability and higher cost/technical risks using conventional sensors.

The main source of noise in VSP studies are tube waves, which are generated when the fluid in the well is displaced by a seismic source, such as ground roll. Tube waves propagate up and down a well and create secondary tube waves in borehole anomalies, which may interfere with primary

reflections. As a coherent noise mode, tube waves cannot be attenuated by stacking; therefore, both field procedures, such as use of attenuating heavy mud and construction of ditches to attenuate the ground roll, and processing techniques, such as filtering and tube wave modeling, are commonly used to minimize their effect. Interesting applications of tube wave analysis as signal include full-waveform inversion (FWI) for effective hydraulic fracture aperture and compliance (Hunziker et al. 2019), as well as near-surface characterization via Mach waves, which are conical waves which can be generated by tube waves (Wu et al. 2017).

VSP Applications

The primary use of VSP is the integration of VSP and downhole logging to surface reflection surveys. Contrary to surface studies which infer subsurface conditions from measurement of the reflected wavefields with the source and receiver at the surface, a VSP configuration involves closely spaced direct physical measurements of the wavefield between the source position and the reflector. Zero-offset VSP measurements can be correlated with sonic logs and check-shot surveys in the borehole and improve the depth positioning of reflections recorded in surface data. From a velocity perspective, zero-offset VSP measurements are denser than check-shot surveys; hence they provide a finer interval velocity profile. On the other hand, VSP samples a

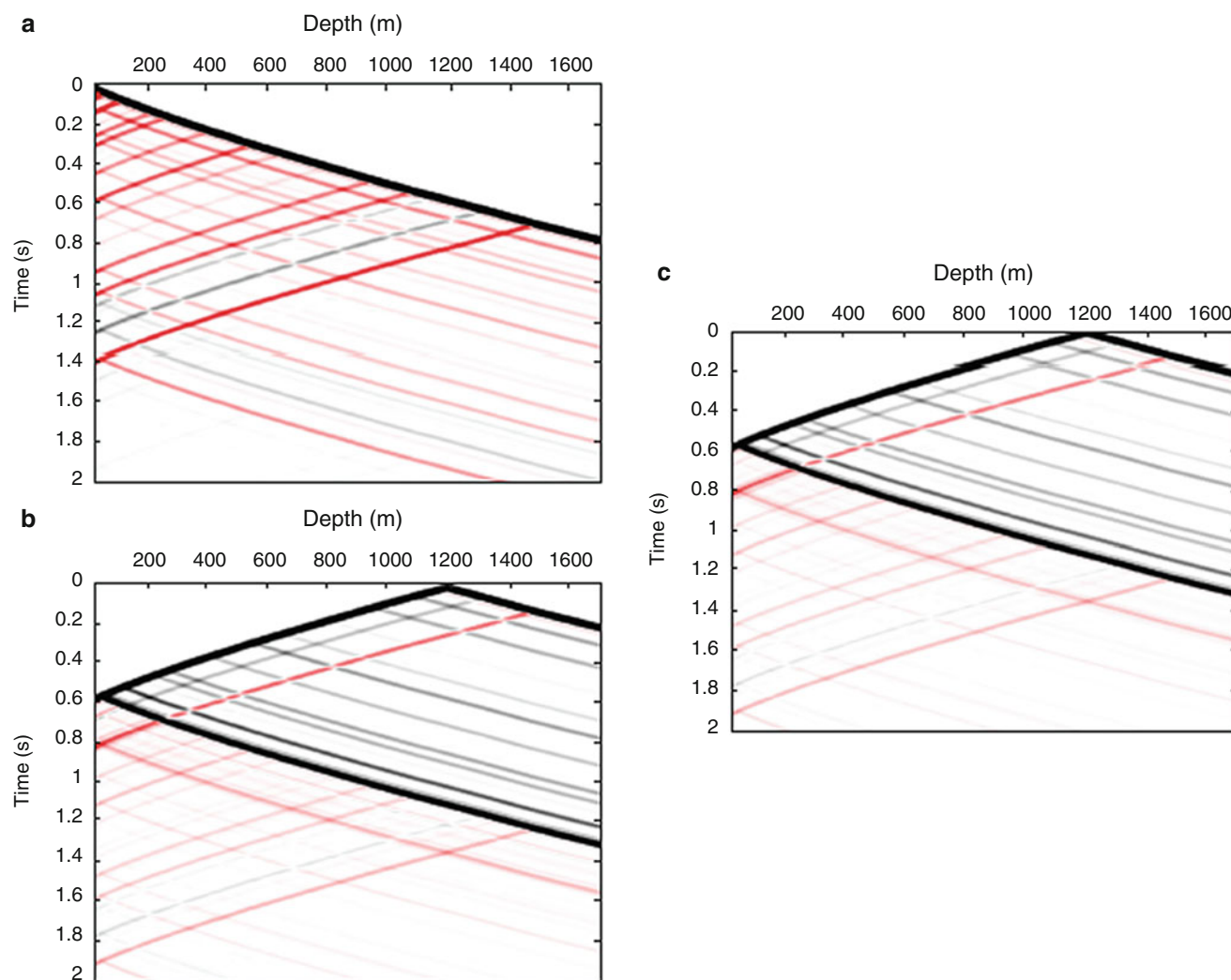
larger velocity body around the well than the high-frequency sonic log, which may be subject to cycle skipping and the effects of the washed-out zones. VSP also contains the signature of the near surface which is often absent in logging. Hence, synthetic seismograms computed from sonic logs may not accurately represent reflections recorded from surface measurements. Reflections from zero-offset VSP, however, can be matched to those from surface measurements via an appropriate static time-shift to account for the differential travel times of the two recording geometries.

Concerning seismic imaging, the recorded wavefield in a buried VSP array has higher signal-to-noise ratio across a larger seismic bandwidth than surface measurements for two reasons. First, surface cultural noise is suppressed in the borehole, and, second, upgoing reflections do not propagate through the near-surface material which is highly *attenuating* (*viscoelastic attenuation*, qv). In addition, the reflections recorded in a VSP suffer less attenuation as the distance to the target is shorter. This shorter distance also implies that the *Fresnel zone* (*waveform modeling and tomography*, qv) is smaller in the case of a VSP, leading to higher resolution compared to surface seismic studies. Given that downgoing and upgoing wave modes are separated in the f - k space, either an f - k filter or a trace by trace arithmetic subtraction of the estimated wave mode can isolate downgoing from upgoing waves. For horizontal wells, separation is achieved through geophone-hydrophone pairing. Estimation of the downgoing wavefield from VSP data is used in the design of deconvolution operators, which can be thereafter applied to surface data to improve the vertical resolution. On the other hand, upgoing reflections can be mapped to their origin position via either a CDP-VSP mapping or *seismic migration* (qv) algorithm or a combination of the two. A migration step is strongly advisable in the case of complex geologies.

In recent years, VSP seismic imaging has transformed from an analysis of acoustic reflections to migrating converted-wave reflections, as well as multiples, through mirror and interferometric migration (Hanafy et al. 2015; Guo et al. 2015). To meet the challenges of complex imaging, velocity estimation is critical. The most developed category of techniques to estimate complex subsurface velocity is full-waveform inversion (FWI) techniques, which have been successfully developed using acoustic/elastic/viscoelastic forward models, including formulations for VSP. Furthermore, FWI has been successfully applied to time-lapse VSP field data for reservoir and CO₂ injection monitoring (e.g., Liang et al. 2013). By integrating velocity estimation and migration, the technique of joint migration inversion (JMI) has been successfully developed for 3D VSP data, whereby both surface and internal multiples may improve illumination and expand imaging capabilities (Marhfoul and Verschuur 2017).

Another promising technique for seismic imaging applications is seismic interferometry, which through the creation of virtual sources (VS) in the location of the receivers eliminates the need for any knowledge of the complex overburden. In addition, unlike traditional migration schemes, interferometric imaging circumvents a priori knowledge of the velocity structure in the vicinity of targets, such as salt flanks, and can be applied to imaging complex structures above or below a receiver array or even by further exploitation of multiples. Further to seismic interferometry of VSP measurements, which can create virtual sources in the subsurface, the technique of Marchenko redatuming can move both sources and receivers in the subsurface, with the potential to reproduce a “virtual VSP” solely from surface seismic data and a background velocity model. Application of this technique to synthetic VSP data shows that the inverted reflectivity response from the Marchenko redatuming and imaging compares very well with true reflectivities and even better than conventionally processed VSP seismic data (Kiraz and Nowack 2018). Figure 2 shows a comparison between forward modeled VSP data (with source and receiver at depth) versus the wavefield from a virtual source at the same depth computed from Marchenko redatuming of synthetic surface seismic, which attests the similarity. Furthermore, Fig. 3 suggests that the inverted reflectivities from Marchenko imaging result in more correct reflectivity model compared to the reflectivity computed from conventional processing of VSP data, due to processing artifacts. Further testing in field data applications is work in progress.

Through the separation of downgoing and upgoing wavefields and their relative amplitudes, it is possible to extract the true *reflection coefficients* (*energy partitioning of seismic waves*, qv) used for lithological interpretation. Moreover, from multi-offset three-component VSP data, it is possible to determine the *AVO response* (*energy partitioning of seismic waves*, qv) with no assumptions to the path of the propagating wavefield, which is required in surface seismic AVO analysis. Another benefit of using VSP measurements is their robustness in estimating *seismic anisotropy* (qv), which improves lithologic interpretations as well as increases the quality of seismic imaging. Given the seismic velocity response to changes in reservoir porosity and fluid content, velocity *anisotropy* (qv) can also be applied in monitoring of fluid-drainage patterns during production. Fracturing will also affect seismic velocity; hence VSP velocity measurements can also be used for monitoring hydraulic fracturing. Byerley et al. (2018) used a downhole DAS array to monitor hydraulic fracturing stages, and observed appearing events, which are potentially converted waves of newly induced fractures. Furthermore, they observed that velocity perturbations, expressed as 4D time-shifts, increased



Vertical Seismic Profiling, Fig. 2 (a) synthetic wavefield for a source at the surface, (b) synthetic wavefield for a physical source located downhole, (c) wavefield using the Marchenko redatuming for a virtual

source located downhole at the same location as source for (b). (Adapted from: Kiraz and Nowack 2018)

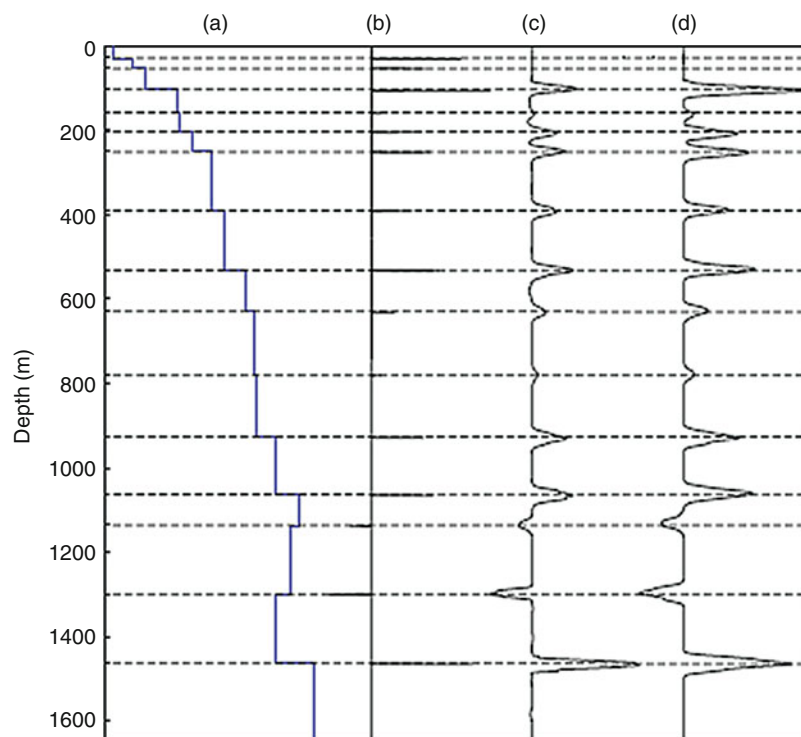
before decreasing again, suggesting potentially that seismic responds to fractures that are fully opened but less so as they start to close. Using shear-wave splitting (SWS) analysis, fracture orientation and density can be determined from the polarization of the fast shear wave and delay time between fast and slow shear waves, respectively. Moreover, by estimating frequency-dependent anisotropy (FDA) linked to scattering and fluid flow in fractures, the size of fractures from micro to cm- and m-scale can be determined (Al-Harrasi et al. 2011).

With *attenuation* (*viscoelastic attenuation*, q_v) gaining increasing importance as a seismic parameter linked to lithological properties, VSP has proven to be a powerful tool for

Q (*viscoelastic attenuation*, q_v) estimation. Aside from true amplitude processing required in stratigraphic interpretation, VSP-derived Q profiles are also used in the design of inverse Q -filters to enhance seismic resolution. Studies also suggest a link between anisotropic attenuation and fracturing; hence VSP azimuthal measurements of attenuation may potentially be applied in monitoring hydraulic fracturing in hydrocarbon and geothermal applications. Using perforation shots, studies have successfully identified low S-wave attenuation zones linked to the growth of stimulated reservoir volume in the Marcellus gas shale (e.g., Tan et al. 2014). Their results suggest that it may be possible to monitor hydraulic stimulation with attenuation analysis.

Vertical Seismic Profiling,

Fig. 3 (a) velocity model, (b) reflectivity model from the velocity model, (c) depth-converted stacked trace from conventionally processed VSP, (d) inverted reflectivity response from the Marchenko redatuming and imaging approach. (Adapted from: Kiraz and Nowack 2018)

**Summary**

VSP surveys have long been used for reducing risk in well placement and improving reservoir monitoring. Intrinsic to the complex geometry of the VSP, which poses several data acquisition and processing challenges, is the ability of the VSP to expand the spatial and time-lapse subsurface characterization potential of conventional downhole logging and surface reflection studies. Henceforth, following the success of VSP surveys and their increased benefits over their financial cost, the VSP technique is a rapidly developing area of geophysical research. In the last years, DAS technology has enabled time-lapse seismic monitoring in challenging environments with unmatched repeatability and fine spacing while reducing costs. Lastly, exciting mathematical and field applications of Marchenko redatuming and interferometry may soon change the acquisition and processing of VSP.

Cross-References

- [Energy Partitioning of Seismic Waves](#)
- [Magnetic Anisotropy](#)
- [Seismic Data Acquisition and Processing](#)
- [Seismic Viscoelastic Attenuation](#)
- [Seismic, Migration](#)
- [Seismic, Waveform Modeling and Tomography](#)

Bibliography

- Al-Harrasi OH, Kendall J-M, Chapman M (2011) Fracture characterization using frequency-dependent shear wave anisotropy analysis. *Geophys J Int* 185:1059–1070
- Byerley G, Monk D, Aaron P, Yates M (2018) Time-lapse seismic monitoring of individual hydraulic frac stages using a downhole DAS array. *Lead Edge* 37(11):802–810
- Guo B, Yu J, Huang Y, Schuster GT (2015) Benefits and limitations of imaging multiples: interferometric and resonant migration. *Lead Edge* 34(7):802–805
- Hanafy SM, Huang Y, Schuster GT (2015) Benefits and limitations of imaging multiples: mirror migration. *Lead Edge* 34(7):796–800
- Hunziker J, Greenwood A, Minato S, Barbosa N, Caspari E, Hollinger K (2019) Bayesian full-waveform tube-wave inversion for effective hydraulic fracture aperture and mechanical fracture compliance. EAGE 81st Conference and Exhibition
- Kiraz MS, Nowack RL (2018) Marchenko redatuming and imaging with application to the Frio carbon sequestration experiment. *Geophys J Int* 215:1633–1643
- Liang L, Li M, Rufino R, Abubakar L, Nutt H, Menkiti H, Dummong S, Tondel R (2013) Application of frequency-domain full-waveform inversion for time-lapse 3D VSP data interpretation. *Soc Explor Geophysicists Expanded Abstr* 32:5107–5112
- Marhfoul BE, Verschuur E (2017) Joint Migration Inversion of 3D full wavefield borehole data. EAGE 79th Conference and Exhibition
- McCollum B, LaRue WW (1931) Utilization of existing wells in seismograph work. *Early Geophys Pap* 1:119–127
- Tan Y, Chai C, Engelder T (2014) Use of S-wave attenuation from perforation shots to map the growth of the stimulated reservoir volume in the Marcellus gas shale. *Lead Edge* 33:1090–1096
- Wu Z, Rector JW, Nihei K, McCallen D (2017) Characterizing seismic soil and unconsolidated sediment using Mach waves. *Soc Explor Geophysicists Expanded Abstr*:2710–2714

Very Long Baseline Interferometry

Helmut Wiesemeyer¹ and Axel Nothnagel²

¹Millimeter and Submillimeter Astronomy, Max-Planck-Institute for Radio Astronomy, Bonn, Germany

²VLBI Research Group, Institute of Geodesy and Geoinformation, University of Bonn, Bonn, Germany

Synonyms

VLBI

Definition

Very Long Baseline Interferometry (VLBI) – correlation interferometry of celestial objects with separately operating radio antennas equipped with frequency standards, off-line data reduction remotely by a digital correlator.

e-VLBI – VLBI with transmission of data by a high-speed network instead of storage media.

Introduction

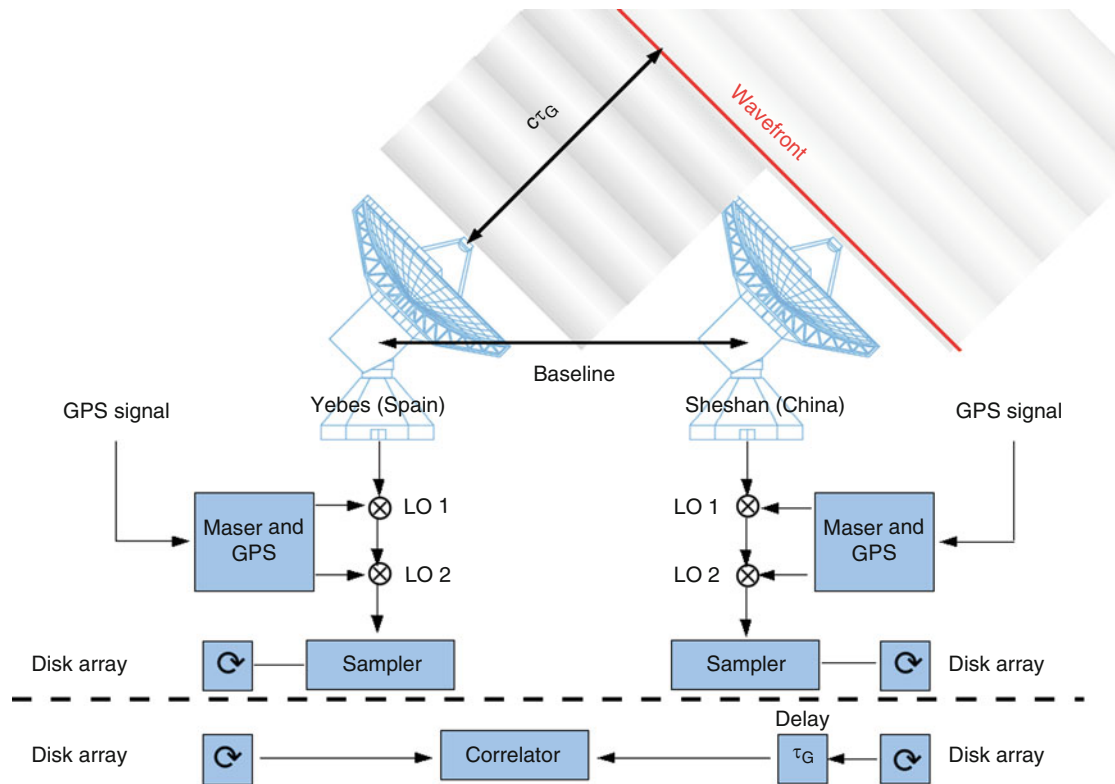
Very long baseline interferometry (VLBI) is a radio astronomical technique to achieve high spatial resolution in astronomical imaging (of the order of milliarcseconds), involving antennas separated by thousands of kilometers (a “VLBI network”). Because the reconstruction of the image depends on the precise knowledge of the vector (hereafter: baseline) connecting the two antennas of a station pair, the a priori knowledge of the image can be used to determine the baseline vector by inverting the measurement equation, with a precision corresponding to a fraction of the wavelength. Such an a priori knowledge is naturally given if the astronomical source on the sky is spatially unresolved and therefore a “point source.” Such sources – usually quasars, that is, the nuclei of bright radio galaxies – are spread over the entire sky and regularly observed in radio astronomy. Whereas VLBI in radio astronomy aims at resolving these active galactic nuclei, improper assumptions about source structure limit the accuracy of the determinations of antenna stations. This can be avoided by using appropriate source structure models (Tornatore and Charlot 2007). The importance of VLBI techniques in Geodesy and Geophysics is evident if the two antennas of a baseline are located on different tectonic plates, whose relative motion and deformation can therefore be determined with high accuracy (qv ► [“Geodesy, Networks, and Reference Systems”](#)).

A VLBI array can be thought as a fictive giant single-dish telescope whose aperture is synthesized while the baseline vector sweeps the sky thanks to the rotation of the Earth. The resulting image therefore depends on the angular momentum vector of the Earth. This dependency implies applications of VLBI in the measurements of the Earth’s spin axis in space and with respect to the solid crust as well as irregularities of the Earth Rotation.

Methods and Techniques in Very Long Baseline Interferometry

The idea of connecting several telescopes to an array in order to synthesize part of the surface of a single, large telescope dates back to 1946, when Ryle and Vonberg (1946) used for the first time a multielement radio telescope for solar observations. In both optical and radio astronomy, the idea turned out to be extremely fruitful for achieving the highest spatial resolutions. For a single telescope, the resolution (in radians) is given by the ratio of wavelength to aperture diameter (up to a factor of order unity). By consequence, a radio telescope working at a wavelength of 21 cm needs an aperture of about 50 km in order to achieve the same angular resolution as, for example, an optical telescope limited by atmospheric seeing. The largest, fully steerable radio telescopes have apertures of the order of 100 m and reach, at that wavelength, an angular resolution of 9 arc minutes. As an alternative to even larger single-dish radio telescopes, several telescopes can be connected by microwave or fiber optics cables in order to form a connected-element interferometer. If the telescopes of such an interferometer are suitably distributed, a sufficiently long observation of a celestial source can be reconstructed to form an image whose angular resolution is of the order of the ratio of the wavelength to the longest among the distance vectors (“baseline,” see Fig. 1) between telescope pairs. The advantage of connected-element interferometers is that the signals can be correlated in real time, such that the phase of the electromagnetic wave when it arrives at the different telescopes is preserved. This is achieved by mixing the incoming signal with that from a local oscillator. In a connected-element interferometer, the local oscillators in all antennas are driven with a common phase reference. The Very Large Array (VLA, New Mexico) and the British e-Merlin array have elements connected by optical fiber cables, with baselines of up to 50 km (VLA) and 200 km (e-Merlin). Despite modern communication technology sometimes used (e.g., satellite links), the physical connection of array elements becomes impractical beyond a certain limit.

The step from a connected-element interferometer to a VLBI experiment consists of abandoning (for practical and economical reasons) the microwave or optical fiber links; registering the data, together with time marks, on a suitable



Very Long Baseline Interferometry, Fig. 1 Schematic representation of an antenna pair on a baseline between Spain and China (distances not to scale). The electromagnetic wave from a distant point source (a quasar) is shown together with the position of the wave front when arriving in China (red line). The signal path corresponding to the time

delay τ_G is shown, as well as the correction for the latter in the data reduction. There is no physical connection between both antennas, each operating with an independent local oscillator (labeled LO). The dashed line separates the measurement from the off-line data processing. (Credits: Greve/IRAM/Rebus)

storage medium; and analyzing the data later with a digital correlator. While local oscillators are necessary to convert the observed frequency band pass to one at a lower frequency (for recording the data), they cannot be locked to a common phase reference because of the large separation between the antennas. Time marks are therefore necessary. This technique became feasible with the availability of reliable frequency standards for the time marks and magnetic tapes of sufficient storage capacity. The first VLBI observation correlated the data of a single baseline between the Hat Creek (California) and Green Bank (West Virginia) observatories (Bare et al. 1967). Since then, the major improvements in VLBI were mainly due to the increase of baseline length (i.e., spatial resolution) and collecting area (i.e., sensitivity) and due to the availability of more reliable clocks and faster digital samplers, leading to larger spectral bandwidth and thus sensitivity. Nowadays, hydrogen masers are used as frequency standards, while arrays of hard disks replace the tapes. The distinction between a connected-element interferometer and a VLBI network is somewhat imprecise. Using network connections of large bandwidth, VLBI data can now be transmitted in real time to the central correlator, while the local oscillators of the antennas participating in the network are

not synchronized. This technique has been demonstrated in 2005 with the European VLBI Network (EVN), involving a transatlantic baseline.

Physical Principles

Interferometric experiments detect both amplitude and phase of electromagnetic waves (conveniently described by complex functions) and combine them in a suitable way to provide images, spectra, and polarization information. Phase-sensitive detection is possible by measuring correlations (or coherence functions) in the incoming electromagnetic field. It has been shown (Sudarshan 1963) that for the case of two-point correlation functions, classical and quantum electrodynamics lead to equivalent results if nonlinear effects are ignored. We therefore limit our discussion to the case of ensembles of incoherent packets of quasi-monochromatic electromagnetic waves from different directions but confined to a small field of view.

The starting point of any interferometric imaging experiment is the van Cittert-Zernike theorem (van Cittert 1934; Zernike 1938), stating that the correlation of the

electromagnetic field at two points (two-point or mutual coherence function) is, up to a complex scaling factor, given by the Fourier transform (also called “visibility function”) of the spatial intensity distribution (e.g., the interference pattern observed on a screen behind an illuminated slit or a grid). The complex scaling factor involves a correction due to the fact that the antenna baselines are not parallel to the plane of the sky (and, for the case of VLBI, not coplanar among them). This so-called delay correction τ_G needs to be applied because of the lag between the arrival times of the front of the electromagnetic wave at the antennas of a given baseline (Fig. 1) and is equivalent to steering the fictive large single-dish telescope mentioned above to the source. The applied delay correction will differ from the theoretical one, owing to the clock error and to the inaccuracy of the assumed baseline length. In the absence of an error-free, absolute time scale, the data streams from an antenna pair rather have to be synchronized by maximizing the correlated signal, applying reasonably guessed delay corrections. A first synchronization is done by using the GPS time signal (qv ► “GPS, Data Acquisition, and Analysis”) to which the local oscillators in the antennas are locked. A refinement needs to be applied later when the data are cross-correlated. There are actually two free parameters per baseline pair, not only the delay correction, but also the frequency of the oscillations, due to the Earth’s rotation, in the interference pattern (“fringes”). This frequency has to be fitted, too, because it depends on the clock rate. For hydrogen masers, whose radio signal at a frequency of 1.4 MHz is now commonly used as a frequency standard for VLBI experiments, the fractional frequency drift (e.g., due to temperature drifts) after 100 s is of the order of 10^{-14} . Such a high precision is required if phase information is to be extracted once the signals are correlated. There is yet another difference between a connected-element interferometer with baselines of up to 10^6 wavelengths and a VLBI experiment: in the former case, the field of view of the interferometer is limited by the central lobe of the antenna reception pattern (“main beam”), whereas in VLBI the field of view is limited by the largest delay τ_G within which the signal from a given direction is coherent. This delay depends on the baseline length with respect to the wavelength, on the spectral bandwidth across which these signals are averaged, and on the accuracy required (e.g., astrometry or precise baseline determinations need a higher degree of coherence). Therefore, the field of view of VLBI observations is much smaller than that of observations with connected-element interferometers. However, nowadays more evolved data analysis techniques exist that permit to overcome this limitation (Garrett et al. 1999).

For each telescope of the VLBI network, the data from the receiving system are recorded on an array of hard disks, together with the time marks from the clock. Later these data are correlated with a digital correlator. Modern

correlators are purely software-based and run on clusters of commodity personal computers (e.g., DiFX, Deller et al. 2007). They allow us to analyze VLBI data for large frequency bands (for high sensitivity to continuum signals) or high-frequency resolution (for observing spectral lines).

Data Calibration and Imaging

After applying the aforementioned delay and fringe rate corrections, the phases of the complex data need to be recovered. The visibility functions are first smoothed to an integration time interval, which depends on the tolerated “visibility smearing”: owing to the rotation of the Earth during this time span, different spatial frequencies are measured at different times, even for a single baseline. Furthermore, at the highest frequencies, where atmospheric fluctuations become important (mm VLBI), the elementary integration times are kept below 20 s, which allows one to apply atmospheric phase corrections. Another upper limit to the integration time interval is given by the stability of the clock. However, the time constant of a hydrogen maser is longer than the two previous limitations.

Unlike connected-element interferometers, where the phase of the front of the incident electromagnetic wave is not too much distorted by atmospheric effects, the phase difference between signals received by different VLBI stations is completely lost. This is because the atmospheric fluctuations at the different antenna sites are uncorrelated and because the local oscillators in the antennas are not locked by a common phase reference. Fortunately, only phase differences with respect to a reference antenna are important. For a triangle of VLBI antennas, the phase errors due to atmospheric and instrumental fluctuations cancel out in the sum of the phase differences between the antennas of the three baselines involved (“phase closure”). Since the number of baselines scales roughly with the square of the number of antennas, whereas the number of phase errors increases only linearly with it, a considerable part of the phase information can be recovered (e.g., 89% for the 18 antennas of the EVN). This recovery of phase information can be improved by a “self-calibration” initially assuming a point source as model, calibrating the measured phases under this assumption, and then creating a new source model using an imaging algorithm. In the underlying theory, complex antenna gains are used, and closure relations also exist for their amplitudes.

Such imaging algorithms are needed because the VLBI stations cannot sample the visibility functions on a regular grid of spatial frequencies. Therefore, the brightness distribution resulting from a Fourier transform is convolved with the “synthesized beam,” which is the Fourier transform of the sampling of spatial frequencies. Iterative algorithms exist to

deconvolve the image such that the result is the brightness distribution as observed with a fictive single-dish telescope whose diameter corresponds to the longest available baseline. The CLEAN algorithm (Högbom 1974) considers the brightness distribution as a collection of spatially unresolved sources. Other deconvolution algorithms have been proposed and applied since then, depending on the nature of the observed brightness distributions. The combination of CLEAN and phase closure is known as hybrid mapping (e.g., Readhead et al. 1980). Global fringe fitting (Schwab and Cotton 1983) extends the closure relations to time delays and fringe rates and is iteratively used together with a deconvolution algorithm like CLEAN.

For high-precision work (e.g., measurements of the Earth rotation parameters, see below), general relativistic effects need to be considered. As an indication, the gravitational time delay due to the Earth's gravity amounts to up to ~ 0.02 ns on a 6000 km long baseline, and the positions of the celestial sources have to be corrected for light bending in the gravitational field of the Sun (4 milliarcseconds for an incidence orthogonal to the direction from Earth to Sun; for a summary and further references, see Heinkelmann and Schuh 2010; Fomalont et al. 2010).

VLBI Networks

VLBI networks now extend all over the world. The most sensitive network is the European VLBI Network (EVN), connecting 18 radio telescopes in Europe, Puerto Rico, and South Africa, operating together for more than 3 months per year. The American Very Large Baseline Array (VLBA) works for the whole year. Other national VLBI facilities exist in the UK (the Merlin array), Japan (Japanese VLBI network JVN), and Australia (long baseline array, LBA). Occasionally, some of these networks work together to form global VLBI networks (e.g., the EVN with the VLBA or Merlin, and the Asia-Pacific telescope APT with observatories in Australia, China, New Zealand, and Japan). Currently, the best angular resolutions (down to 0.1 milliarcsecond) are achieved with global VLBI at 3 mm wavelength. Future improvements of the spatial resolution scale are to be expected from VLBI below 3 mm wavelength and from networks including satellite stations (with the Japanese HALCA antenna, in operation from 1997 to 2005, to be replaced by the VSOP 2 mission). These VLBI networks are mainly used for astrophysical research. However, VLBI is also used extensively for geodetic and geophysical applications (see below). Owing to the importance of the VLBI techniques for these disciplines, the International VLBI Service for Geodesy and Astrometry (IVS, Nothnagel et al. 2017) coordinates the observations of a dedicated network of VLBI stations on all continents (Fig. 2), operated by national research agencies.

Geodetic and Geophysical Applications

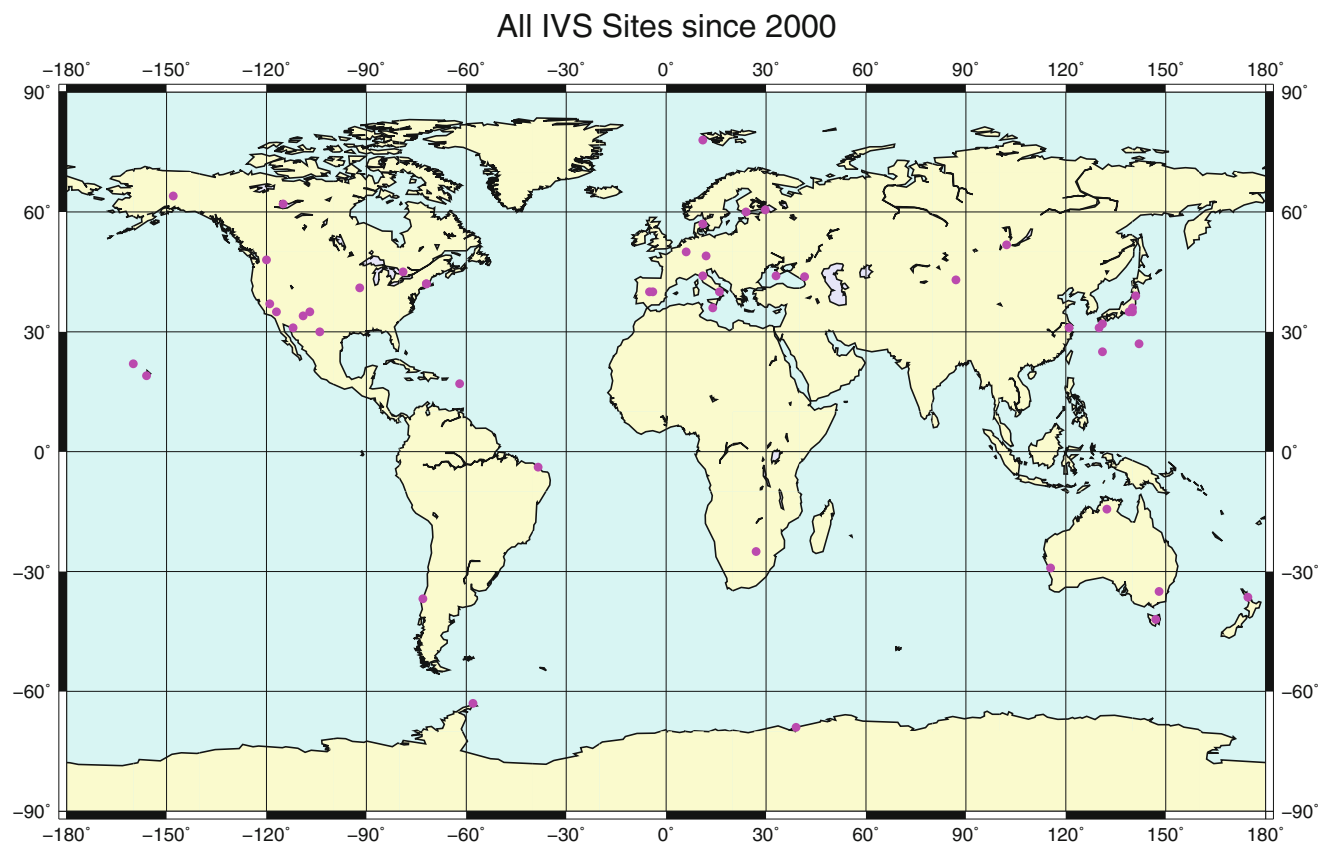
The technique of VLBI is also used for geodetic and geophysical applications. In this context, the positions of compact extragalactic radio sources, like quasars, are considered to be known to sufficient accuracy, and the relative telescope coordinates are inferred from time delay observations which reflect the baseline-source geometry. In most cases, networks of 8 to 12 telescopes (Fig. 2) are employed in series of weekly sessions organized by the IVS for the determination of Earth orientation parameters (qv ► “Earth Rotation”) and radio telescope coordinates. Observing between 50 and 60 radio sources in a repeated sequence of scans of 50–100 s duration over a 24-h period provides enough redundancy and geometric stability to determine telescope coordinates and Earth orientation parameters.

Reference Systems and Tectonics

Geodetic and geophysical applications of VLBI have to do a lot with coordinates of the radio telescopes on Earth and in space. Through sophisticated engineering and constant developments in the theoretical models, the coordinates of radio telescopes around the world can be determined with a relative accuracy of a few millimeters today. These coordinates are always referred to a very specific point of the telescope structure which lies at or close to the intersection of the axes about which the telescope is rotated to aim at the quasars.

The coordinates of geodetically used radio telescopes are determined repeatedly over the years and are fed into the database of the International Earth Rotation and Reference Systems Service (IERS, qv ► “Earth Rotation”, see also <http://www.iers.org>). Together with observations of other space-geodetic techniques like the Global Navigation Satellite Systems (GNSS) or Satellite Laser Ranging (SLR), the VLBI results are merged into the International Terrestrial Reference Frame (ITRF, qv ► “Geodesy, Networks, and Reference Systems”). The last version of the ITRF, ITRF2014 (Altamimi et al. 2016), contains about 1500 points on Earth which are all accurate to sub-centimeter. The importance of VLBI for this endeavor is that VLBI is the technique which is used to determine the scale of the ITRF. Here, VLBI benefits from the fact that the geodetic VLBI measurements link radio telescopes over intercontinental distances with extreme accuracy. The observations of the other techniques are then used to densify the global reference frame.

Closely linked to the coordinates of the radio telescopes are the changes of the positions which are caused by global tectonics better known as continental drift. Geodetic VLBI was the first technique to verify Alfred Wegener's theory of continental drift through real measurements across the North Atlantic between 1979 and 1984 (Carter et al. 1985; Herring



Very Long Baseline Interferometry, Fig. 2 Locations of the VLBI antennas of the IVS network participating in the period of 2000 to 2019

et al. 1986). Today, a full global scenario of several plates and deformation zones exists for which geodetic VLBI has helped to determine the magnitude and direction of the tectonic movements. While the distance between North America and Europe extends by about 20 mm per year, the changes in the Pacific region reach 80 mm per year (Kondo et al. 1987).

Likewise, the positions in space of the Earth and of the quasars observed by VLBI are defined in the International Celestial Reference System (ICRS). The ICRS has been realized through the International Celestial Reference Frame (ICRF, Ma et al. 2009) and its latest successor, the ICRF3 (Charlot et al. 2019), which rely on the VLBI positions of many hundred extragalactic radio sources. Due to the large distances of the latter, the ICRF can be considered as definition of a local inertial system (or quasi-inertial system). The origin of the ICRF is the barycenter of the solar system. Systematic effects and additional random errors of the ICRF2 do not exceed 30–50 microarcseconds.

Earth Rotation

Polar motion and irregularities in the Earth's rotation: Geodetic VLBI is one of the few main techniques for the

determination and monitoring of the Earth's variable rotation. Through the direct link between the terrestrial and the celestial reference frames, VLBI is the only technique which permits to determine all components of Earth rotation without any input from other techniques. While today polar motion is regularly determined precisely with GPS observations (qv ► “GPS, Tectonic Geodesy”) on a routine basis, VLBI uniquely contributes the Earth's phase of rotation represented through the time UT1 (Universal Time 1). UT1 is quite variable, due to movements of the atmosphere and of the oceans, and information on UT1 is needed for a number of applications including orbit transformations of Global Navigation Satellite Systems. For this purpose, the International VLBI Service for Geodesy and Astrometry (IVS) runs daily observations providing UT1 with an accuracy of about 3–8 μ s.

Precession and nutation: Looking at the Earth's axis of rotation in space, one can see that it describes a cone with an opening angle of 23.45° , which needs about 25,700 years to complete. This phenomenon is called precession. Superimposed on the precession cone are periodic movements of the Earth's axis of rotation which have a magnitude of a few arcseconds and are called nutation. For historical reasons, precession and nutation are handled independently although

they are caused by the same phenomenon: The astronomical bodies of the Sun, Moon, and planets exert a torque on the equatorial bulge of the Earth trying to pull it toward the ecliptic. Through its desire to conserve angular momentum, the Earth reacts with a precession effect. Through the periodic rotations of the attracting bodies, superimposed periodicities result, and more than 1000 nutation periods can be identified covering a spectrum from 13.6 years down to a few days.

General relativistic effects in Earth Rotation: Owing to the orbital motion of the Earth, spin-orbit coupling leads to a “geodetic precession” of about 2 arcseconds per century (qv ► “Earth Rotation”, Klioner et al. 2010) that are measured by VLBI, which lends itself also to other tests of general relativity (Fomalont et al. 2010).

Today, geodetic VLBI is the only technique that is used to determine precession and nutation. The accuracy of the nutation offsets determined from 24 h of VLBI observations is on the order of 60–80 microarcseconds which correspond to about 2 mm when transformed to a metric displacement at the Earth’s surface. The agreement between these observations and a model which is derived geophysically using an elastic Earth model is on the order of 1 milliarcsecond. The main cause of the deviations is the effect of the so-called free core nutation (FCN). This phenomenon results from the fact that the Earth’s core has a slightly different ellipticity than the rest of the body resulting in different torques.

Summary

Over several decades, VLBI techniques have been refined up to a point where positions on the Earth’s surface can be measured to a precision of a few millimeters with respect to a system of radio-bright celestial reference sources. As of today, VLBI is indispensable for a range of geodetic applications and services, despite complementary techniques involving Global Navigation Satellite Systems (GNSS) or Satellite Laser Ranging (SLR). While combined results of routinely executed measurements employing VLBI, GNSS, and SLR constitute the foundation for the International Terrestrial Reference Frame (ITRF), VLBI is the only technique allowing for a self-contained description of the components of the Earth’s rotational velocity.

Cross-References

- Earth Rotation
- Geodesy, Networks, and Reference Systems
- GPS, Data Acquisition, and Analysis
- GPS, Tectonic Geodesy
- SAR Interferometry

Bibliography

- Altamimi Z, Rebischung P, Métivier L, Collilieux X (2016) ITRF2014: a new release of the international terrestrial reference frame modeling nonlinear station motions. *JGR Solid Earth* 121:6109–6131
- Bare C, Clark BG, Kellermann KI, Cohen MH, Jauncey DL (1967) Interferometer experiment with independent local oscillators. *Science* 157:189–191
- Carter WE, Robertson DS, Mackay JR (1985) Geodetic radio interferometric surveying: applications and results. *J Geophys Res* 90:4577–4587
- Charlot P et al (2019) The third realization of the international celestial reference frame by very long baseline interferometry. *Astron Astrophys*. In preparation
- Deller AT, Tingay SJ, Bailes M, West C (2007) DiFX: a software correlator for very long baseline interferometry using multiprocessor computing environments. *Publ Astron Soc Pac* 119:318–336
- Fomalont E, Kopeikin S, Jones D, Honma M, Titov O (2010) Recent VLBA/VERA/IVS tests of general relativity. In: Klioner SA, Seidelmann PK, Soffel MH (eds) *Relativity in fundamental astronomy*. Cambridge University Press, Cambridge, pp 291–295
- Garrett MA, Porcas RW, Pedlar A, Muxlow TWB, Garrington ST (1999) Wide-field VLBI imaging. *New Astron Rev* 43:519–522
- Heinkelmann R, Schuh H (2010) Very long baseline interferometry: accuracy and relativistic tests. In: Klioner SA, Seidelmann PK, Soffel MH (eds) *Relativity in fundamental astronomy*. Cambridge University Press, Cambridge, pp 286–290
- Herring TA, Shapiro II, Clark TA, Ma C, Ryan C (1986) Geodesy by radio interferometry: evidence for contemporary plate motion. *J Geophys Res* 91:8341–8347
- Högbom JA (1974) Aperture synthesis with a non-regular distribution of interferometer baselines. *Astron Astrophys Suppl* 15:417–426
- Klioner SA, Gerlach E, Soffel MH (2010) Relativistic aspects of rotational motion of celestial bodies. In: Klioner SA, Seidelmann PK, Soffel MH (eds) *Relativity in fundamental astronomy*. Cambridge University Press, Cambridge, pp 112–123
- Kondo T, Heki K, Takahashi Y (1987) Pacific plate motion detected by the VLBI experiments conducted in 1984–1985. *Radio Res Lab J* 34:1–14
- Ma C, Arias F, Bianco G, Boboltz D, Bolotin S, Charlot P, Engelhardt G, Fey A, Gaume R, Gontier A-M, Heinkelmann R, Jacobs C, Kurdubov S, Lambert S, Malkin Z, Nothnagel A, Petrov L, Skurikhina E, Sokolova J, Souchay J, Sovers O, Tesmer V, Titov O, Wang G, Zharov V, Böckmann S, Collioud A, Gipson J, Gordon D, Lytynen S, MacMillan D, Ojha R (2009) The second realization of the international celestial reference frame by very long baseline interferometry (IERS technical note 35). In: Frankfurt am Main. Verlag des Bundesamtes für Kartographie und Geodäsie, Germany
- Nothnagel A, Artz T, Behrend D, Malkin Z (2017) International VLBI Service for Geodesy and Astrometry Delivering high-quality products and embarking on observations of the next generation. *J Geod* 91(7):711–721. <https://doi.org/10.1007/s00190-016-0950-5>
- Readhead ACS, Walker RC, Pearson TJ, Cohen MH (1980) Mapping radio sources with uncalibrated visibility data. *Nature* 295:137–140
- Ryle M, Vonberg DD (1946) Solar radiation at 175 Mc/s. *Nature* 158:339–340
- Schwab FR, Cotton WD (1983) Global fringe search techniques for VLBI. *Astron J* 88:688–694
- Sudarshan ECG (1963) Equivalence of semiclassical and quantum mechanical descriptions of statistical light beams. *Phys Rev Lett* 10:277–279
- Tornatore V, Charlot P (2007) The impact of radio source structure on European geodetic VLBI measurements. *J Geod* 81:469–478
- van Cittert PH (1934) Die wahrscheinliche Schwingungsverteilung in einer von einer Lichtquelle direkt oder mittels einer Linse beleuchteten Ebene. *Physica* 1:201–210
- Zernike F (1938) The concept of degree of coherence and its application to optical problems. *Physica* 5:785–795

Wavelet Analysis

Mikhail Kulesh

Information Technology, DIS AG, Mahlow, Germany

Synonyms

Time-frequency analysis; Wavelet transform; Wavelets

Definition

Time-frequency representation (TFR)	A view of a time series represented over time and frequency.
Wavelet	A special function that is well localized in both time and frequency domains.
Continuous wavelet transform (CWT)	TFR of the analyzed signal obtained as a set of scalar products of all dilated and translated wavelets with this signal.
Discrete wavelet transform (DWT)	A wavelet transform for which the wavelet is dilated and translated discretely.

Introduction

Spectral analysis methods of the time series based on the Fourier transform are significantly limited in their ability to distinguish the waves with similar frequency content but having the different time of arrival. This limitation can be overcome using different time-frequency analysis techniques that allow us to examine how the frequency content of a given signal changes as a function of time.

A view of a time series represented over both time and frequency is called the time-frequency representation (TFR).

Time-frequency analysis means analysis into a TFR. There are different TFR methods each suitable for different applications:

1. The short-time (alternatively short-term or windowed) Fourier transform that is used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time (Allen and Mills 2004).
2. The Gabor transform is a special case of the short-time Fourier transform that uses a Gaussian function to create a window of time from which the spectrum of the local signal values is computed (Allen and Mills 2004).
3. The Wigner (alternatively Wigner–Ville) distribution (Wahl and Bolton 1993) and the pseudo-Wigner–Ville distribution (Prosser et al. 1999; Andria and Savino 1996) that can be formulated by the multiplicative comparison of a signal with itself, expanded in different directions about each point in time.
4. The S-transform (Stockwell et al. 1996; Pinnegar and Mansinha 2003) that is similar to the short-time Fourier transform, but with a Gaussian window whose width scales inversely, and whose height scales linearly, with the frequency.
5. The continuous wavelet transform (CWT) of a real or complex signal with respect to a real or a complex analyzing wavelet that is a set of scalar products of all dilated and translated wavelets with this signal (Grossmann and Morlet 1984; Heil and Walnut 1989; Chui 1992; Daubechies 1992; Holschneider 1995; Mallat 1999; Antoine 2004). A wavelet is a special function that is well localized in both time and frequency domains.
6. The discrete wavelet transform (DWT) that is a wavelet transform for which the analyzing wavelet is dilated and translated discretely (Heil and Walnut 1989; Holschneider 1995).

The short-time Fourier transform has a fixed resolution dependent on the windowing function width. A wide window gives better frequency resolution (frequency components close to each other can be separated), but poor time resolution. A narrower window gives good time resolution (the time at

which frequencies change), but poor frequency resolution. Among all possible window functions, there is one signal in particular that shows the best performance regarding time and frequency localization: the Gaussian. Thus, the Gabor transform and the S-transform are the short-time Fourier transforms with the best joint time-frequency resolution.

Since the Wigner distribution is a correlation of the signal with a time and frequency translation of itself, there is no limitation in the time-frequency resolution. However, its distribution may show negative energy levels and exhibit interferences that do not have any physical meaning. To minimize the effect of these interferences, the smoothed pseudo-Wigner–Ville distribution can be used. Unfortunately, the use of the smoothing operator introduces time-frequency smearing in the resulting TFR and therefore downgrades its advantage over the Gabor or wavelet transform, namely, the exact localization over time and frequency.

The CWT and DWT are computed by correlating the signal with families of time-frequency atoms (wavelets). Therefore, the time-frequency resolution of this method is limited by the uncertainty relation bound to the simultaneous time-frequency resolution of the analyzing wavelet.

All of these above-mentioned methods are widely used in the geophysical applications. However, the wavelet transform gives a more general framework for the time-frequency analysis of geophysical data; this approach is powerful and elegant and has several remarkable properties that can be directly applied to the polarization and dispersion analysis of seismic waves in the time-frequency domain.

In the next sections, we briefly introduce the direct and inverse wavelet transforms and some of their properties. Then, we will demonstrate the applications of the CWT in the geophysical time-series analysis.

Continuous Wavelet Transform

Let us introduce the following notations:

1. $s(t)$ is a square-integrable (finite-energy) real or complex function over time from $L^2(\mathbf{R})$ space:

$$s(t) \in L^2(\mathbf{R}) \Leftrightarrow \int_{-\infty}^{+\infty} |s(t)|^2 dt < \infty.$$

2. $\omega = 2\pi f$ is the angular frequency (in radian/s) and f is the physical frequency (in Hz).

3. $\hat{s}(\omega) = \int_{-\infty}^{+\infty} s(t)e^{-i\omega t} dt$, $s(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{s}(\omega)e^{i\omega t} d\omega$ are the direct and inverse Fourier transforms.

4. $g(t) \in L^2(\mathbf{R})$ is a real or complex function called analyzing (mother) wavelet (Grossmann and Morlet 1984).

5. $T_t : g(\tau) \mapsto g(\tau - t)$ is a shifting operator for the analyzing wavelet.

6. $D_a : g(\tau) \mapsto g(\tau/a)/a$ is a dilation operator for the analyzing wavelet.

7. $W_g s(t, a)$ are the complex wavelet coefficients dependent on the dimensionless scaling factor $a \in \mathbf{R}$ and dimensional time factor $t \in \mathbf{R}$ (in s).

8. The asterisk superscript means the complex conjugate.

The wavelet transform $W_g s(t, a)$ of the signal $s(t)$ with respect to an analyzing wavelet $g(t)$ is the set of L^2 -scalar products $\langle T_t D_a g, s \rangle$ of the signal with all dilated and translated wavelets:

$$W_g s(t, a) = \langle T_t D_a g, s \rangle = \int_{-\infty}^{+\infty} \frac{1}{a} g^* \left(\frac{\tau - t}{a} \right) s(\tau) d\tau. \quad (1)$$

Note that there are different possibilities to define the dilation operator D_a . We use the factor $1/a$ in this operator. For this factor, the amplitude of the wavelet spectrum is proportional to the amplitude of the signal for pure frequencies. Alternatively, the dilation $D_a : g(\tau) \mapsto g(\tau/a)/\sqrt{a}$ with corresponding adaptation of the transform (1) can be also used (Heil and Walnut 1989), but the relation between the amplitude of the spectrum and the amplitude of the signal will be in this case frequency (scale) dependent.

Using the properties of the Fourier transform, we can also define the shifting and the dilation operators as $T_t : \hat{g}(\omega) \mapsto e^{-i\omega t} \hat{g}(\omega)$ and $D_a : \hat{g}(\omega) \mapsto \hat{g}(a\omega)$ that allows us to rewrite the transform (1) as being used for Fourier-spectra of the analyzed signal and the analyzing wavelet (Holschneider 1995):

$$\begin{aligned} W_g s(t, a) &= \frac{1}{2\pi} \langle T_t D_a \hat{g} \hat{s} \rangle \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{i\omega t} \hat{g}^*(a\omega) \hat{s}(\omega) d\omega. \end{aligned} \quad (2)$$

The last equation can be used for the fast computation methods to calculate the CWT of the discrete signals $s(t_n)$ with the signal length equal to 2^N points. In this way, a fast Fourier transform algorithm can be employed in practice to compute the integral in the Eq. 2.

The Inversion Formula

The wavelet transform preserves all information of the transformed signal since there is an explicit formula that allows us to recover the signal $s(t)$ from its wavelet spectrum (Holschneider 1995):

$$\begin{aligned} s(t) &= M_h W_g s(t, a) \\ &= \frac{1}{C_{g,h}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{a^2} h \left(\frac{t - \tau}{a} \right) W_g s(\tau, a) d\tau da, \end{aligned} \quad (3)$$

where $h(t) \in L^2(\mathbf{R})$ is a reconstruction wavelet used for the inverse transform; it can differ from the analyzing wavelet that has been used to calculate $W_g s(t, a)$. The constant $C_{g,h}$ is the normalization coefficient having the same meaning as the coefficient $1/2\pi$ in the inverse Fourier transform. However, this coefficient depends on which pair of wavelets has been taken for direct and inverse transformation:

$$C_{g,h} = \int_0^{+\infty} \frac{1}{\omega} \left(\hat{g}^*(\omega) \hat{h}(\omega) + \hat{g}^*(-\omega) \hat{h}(-\omega) \right) d\omega. \quad (4)$$

Use of the Physical Frequency

In the practical applications, it is convenient to choose as the scaling parameter the physical frequency f instead of the dimensionless factor $a : f = f_0/a$, where f_0 is the characteristic (central) frequency of the analyzing wavelet. In this case, we can rewrite the direct and inverse transforms (1, 2, and 3) as follows:

$$\begin{aligned} W_g s(t, f) &= \int_{-\infty}^{+\infty} \frac{f}{f_0} g^* \left(\frac{(\tau - t)f}{f_0} \right) s(\tau) d\tau \\ &= \int_{-\infty}^{+\infty} e^{2\pi i t \varphi} \hat{g}^* \left(\frac{f_0 \varphi}{f} \right) \hat{s}(\varphi) d\varphi, \\ s(t) &= M_h W_g s(t, f) \\ &= \frac{1}{C_{g,h}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{f_0} h \left(\frac{(t - \tau)f}{f_0} \right) W_g s(\tau, f) d\tau df. \end{aligned} \quad (5)$$

Wavelets

The analyzing wavelet, that is, wavelet used for direct wavelet transform, must satisfy the following conditions (Antoine 2004):

1. The wavelet must be well localized both in the time and in the frequency domain.

2. The wavelet must satisfy the admissibility condition, which guaranties the invertibility of the wavelet transform:

$$\int_{-\infty}^{+\infty} \frac{|\hat{g}(\omega)|^2}{|\omega|} d\omega < \infty, \text{ see also Eq. 4.}$$

3. In the most cases, the previous condition may be reduced to the requirement that the 0-th moment of the wavelet

$$\text{should vanish: } \int_{-\infty}^{+\infty} g(t) dt = 0.$$

4. The previous property can be generalized by requiring that some consecutive moments vanish, which improves the efficiency of the wavelet at detecting singularities in the

$$\text{signal: } \int_{-\infty}^{+\infty} t^n g(t) dt = 0.$$

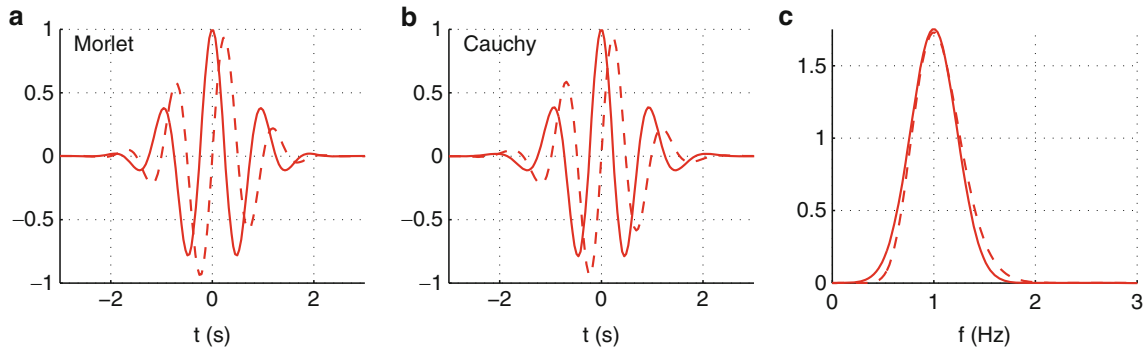
The choice of the analyzing wavelet depends on what typical oscillation modes are present in the signal. For example, to analyze rectangular oscillations, the real Haar wavelet proposed in Haar (1910) or a similar but symmetric French hat (FHAT) wavelet is most suitable. The most convenient for the seismic signals are real wavelets based on different-order derivatives of the Gaussian or complex progressive wavelets. If a two-component signal presented in a complex form is analyzed, or the separation into prograde and retrograde wavelet spectrum is needed, the choice of a complex progressive wavelet is preferable.

A special set of progressive wavelets is the set of functions for which the Fourier transform is real valued. Let us consider two such wavelets:

1. The Morlet wavelet is represented in Fig. 1a. The solid line corresponds to the real part, and the dashed line to the imaginary part of the wavelet. The Fourier spectrum is plotted in Fig. 1c as the solid line. This wavelet can be written with its Fourier transform as:

$$g(t) = e^{i\omega_0 t} e^{-t^2/(2\sigma^2)}, \quad \hat{g}(\omega) = \sigma e^{-(\omega - \omega_0)^2 \sigma^2/2}, \quad (6)$$

where the parameter σ describes the variance of the



Wavelet Analysis, Fig. 1 (a) Morlet wavelet in the time domain, (b) Cauchy wavelet in the time domain, (c) Fourier-spectra of the Morlet wavelet (solid line) and Cauchy wavelet (dashed line)

wavelet, and $\omega_0 = 2\pi f_0$ is the central frequency. Strictly speaking, this function is not a wavelet because it is not of zero mean. However, the function values at negative frequencies are small compared to the values at positive frequencies that allow us to use it as a progressive wavelet with necessary precision in almost all practical applications.

2. The Cauchy wavelet is shown in Fig. 1b and can be written as:

$$\begin{aligned} g(t) &= \left(1 - \frac{2\pi it}{p-1}\right)^{-p}, \\ \hat{g}(\omega) &= \frac{(p-1)^p}{(p-1)!} \left(\frac{\omega}{2\pi}\right)^{p-1} e^{-(p-1)\omega/2\pi}. \end{aligned} \quad (7)$$

This wavelet has the unit central frequency, the parameter p describes the variance of the wavelet, and the Fourier spectrum is nonsymmetrical (Fig. 1c, dashed line). The wavelet transform with respect to this wavelet is closely connected to the analysis of analytical functions over the half plane (Holschneider 1995).

One can choose the translated δ -function $\delta_t = \delta(\tau - t)$ as the reconstruction wavelet $h(t)$ for the inverse wavelet transform, which gives us a rather simple and fast reconstruction formula (Holschneider 1995):

$$s(t) = M_\delta W_{g,s}(t, f) = \frac{1}{C_{g,\delta}} \int_{-\infty}^{+\infty} \frac{1}{f} W_{g,s}(t, f) df. \quad (8)$$

However, the δ -function is not a wavelet and its choice as the analyzing wavelet $g(t)$ for the direct CWT is not possible.

Wavelet Spectrum Amplitude and Phase

Because the wavelet coefficients are complex, when using a complex wavelet, we can write the spectrum $W_{g,s}(t, f)$ as a function of wavelet spectrum modulus $|W_{g,s}(t, f)|$ and phase $\arg W_{g,s}(t, f)$:

$$\begin{aligned} W_{g,s}(t, f) &= |W_{g,s}(t, f)| e^{i \arg W_{g,s}(t, f)}, \\ \arg W_{g,s}(t, f) &= \arctg \left(\frac{\operatorname{Im} W_{g,s}(t, f)}{\operatorname{Re} W_{g,s}(t, f)} \right), \end{aligned} \quad (9)$$

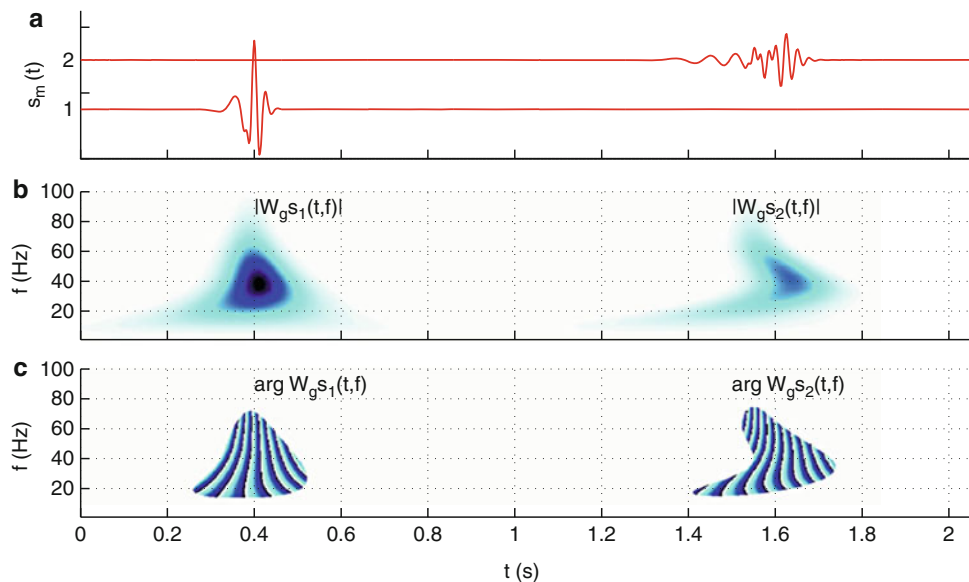
where $\arctg(x) \in [-\pi, \pi]$, which is implemented in most programming languages as $\operatorname{atan2}()$.

Thus, the complex wavelet spectrum contains information about the time-frequency distributions of the signal's amplitude and phase, and this amplitude and phase information can be analyzed separately as shown in the Fig. 2. In this example, we consider a wave propagating in the medium with dispersion and attenuation (see ► “Seismic Viscoelastic Attenuation”). This wave is represented by two signals, $s_1(t)$ and $s_2(t)$, which can be registered from two geophones (see ► “Seismic Instrumentation”) separated by a distance Δx (Fig. 2a).

The wavelet spectrum modulus $|W_{g,s}(t, f)|$ corresponds to the instantaneous amplitude. We can visualize it, for example, as contour line plot, where the horizontal axis corresponds to the time, and the vertical axis represents the frequency (Fig. 2b). The complex wavelets Morlet and Cauchy have a phase shift between the real and imaginary components that prevent here the high-frequency modulus oscillations.

At the same time, a contour line plot of the wavelet spectrum phase $\arg W_{g,s}(t, f)$ represents the instantaneous phase on the different frequencies. Because of practical reasons, the

Wavelet Analysis,
Fig. 2 Wavelet spectrum amplitudes and phases for two signals



phases must be only plotted for time-frequency areas where the modulus $|W_g s(t, f)|$ is not infinitesimal (Fig. 2c).

To analyze the dynamic behavior of multivariate signals using the CWT, one may be interested in investigating a diffeomorphic deformation of the wavelet space $W_{gs1}(t, f) \mapsto W_{gs2}(t, f)$ as shown in Fig. 2. Such a deformation corresponds to a pseudo differential operator acting on signal's wavelet spectrum:

$$D : s(t) \mapsto M_h D W_g s(t, f). \quad (10)$$

In this context, we will consider in the next sections how the wavelet transforms of the source and the propagated signal are related through a transformation operator that explicitly incorporates the phase and group velocities as well as the attenuation factor of the medium.

Prograde and Retrograde Wavelet Spectrum

Both Morlet and Cauchy wavelets are progressive, that is, their Fourier coefficients for negative frequencies are zero. This feature allows us to separate the wavelet spectrum into prograde $W_g^+ s(t, f)$ and retrograde $W_g^- s(t, f)$ components:

$$\begin{aligned} W_g s(t, f) &= W_g^+ s(t, f) + W_g^- s(t, f), \\ W_g^+ s(t, f) &= \begin{cases} W_g s(t, f), & f \geq 0 \\ 0, & f < 0 \end{cases}, \\ W_g^- s(t, f) &= \begin{cases} 0, & f \geq 0 \\ W_g s(t, f), & f < 0. \end{cases} \end{aligned} \quad (11)$$

To perform such a separation for nonprogressive (e.g., real) wavelets, one must first calculate the Hilbert transform of the source signal, which makes the numerical procedure more complex.

An example of the prograde and retrograde wavelet spectrum is shown in Fig. 3, where we consider a two-component synthetic seismogram related to a Rayleigh wave (see

► “Surface Waves”). From two signals, $u_x(t)$ and $u_z(t)$ that correspond to the orthogonal components of the record, we construct a complex signal $z(t) = u_x(t) + iu_z(t)$, as shown in Fig. 3a, where the solid line corresponds to the real part, and the dashed line to the imaginary part of this complex signal.

For this complex signal, we perform CWT using Eq. 5 with respect to Cauchy wavelet (7) and plot only the absolute values of the complex wavelet coefficients as contour line plots for the prograde and the retrograde components (11) separately (Fig. 3b). The analysis of the difference between these two spectra is exactly the idea to find out polarization properties of a signal as introduced in the next sections.

Discrete Wavelet Transform

As the CWT, the discrete transform also has a huge number of applications in science, engineering, mathematics, and computer science. In its basic form, the DWT employs a dyadic grid (integer power of two) in the timescale plane:

$$\begin{aligned} T_{jk} : g(\tau) &\mapsto g(\tau - k/2^j), \quad D_j : g(\tau) \mapsto 2^{j/2} g(2^j \tau), \\ \psi_{jk}(\tau) &= T_{jk} D_j g(\tau), \quad j, k \in \mathbf{N}. \end{aligned}$$

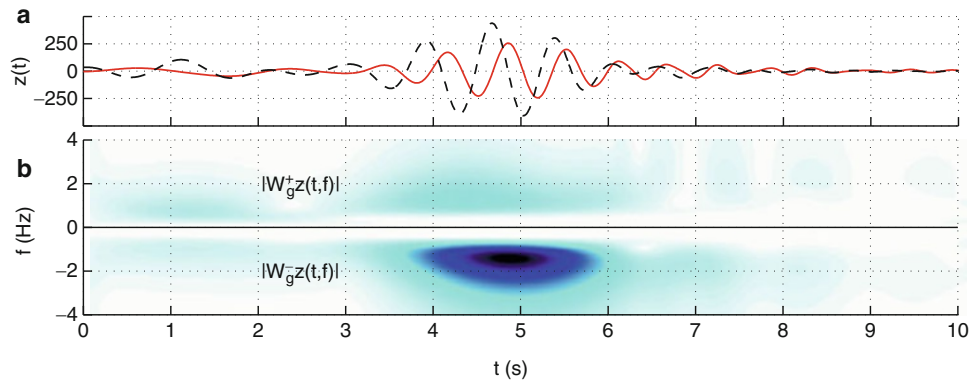
Actually, the transform integral remains continuous for the DWT, but is determined only on a discretised grid of frequencies and time points (Heil and Walnut 1989):

$$W_{g^s j k} = \langle \psi_{jk}, s \rangle = \int_{-\infty}^{+\infty} 2^{j/2} g^*(2^j \tau - k) s(\tau) d\tau. \quad (12)$$

For the DWT, the orthonormal wavelets are usually used. The wavelet $g(t)$ is called orthonormal, when the functions family $\{\psi_{jk}\}$ in the Eq. 12 corresponds to an orthonormal basis in the $L^2(\mathbf{R})$ space, that is, $\langle \psi_{jk}, \psi_{nm} \rangle = \delta_{jn} \delta_{km}$, where δ_{jn} is

Wavelet Analysis,

Fig. 3 Prograde and retrograde wavelet transform of a Rayleigh wave arrival



the Kronecker's delta. In this case, each function $s(t) \in L^2(\mathbf{R})$ can be represented as a sum of a convergent series:

$$s(t) = \sum_{j,k} W_{gsjk} \psi_{jk}(t). \quad (13)$$

The simplest possible orthonormal wavelet is the Haar wavelet:

$$g(t) = \begin{cases} 1, & 0 \leq t < 1/2, \\ -1, & 1/2 \leq t < 1, \\ 0, & t < 0, t \geq 1. \end{cases} \quad (14)$$

This function was proposed by Haar (1910) to give an example of a countable orthonormal system for the space of square-integrable functions on the real line. The technical disadvantage of the Haar wavelet is that it is not continuous, and therefore not differentiable.

Use of Wavelets in Geophysical Time-Series Analysis

Instantaneous Parameters

The concept of instantaneous parameters has been widely used in electric engineering and geophysics. Parameters such as instantaneous amplitude $|W_{gs}(t, f)|$, instantaneous phase $\arg W_{gs}(t, f)$, and instantaneous frequency $\Omega(t, f)$ are directly related to geometry and physical property variations of the medium through which the radar signal propagates. The instantaneous frequency, for example, responds to both wave propagation effects and depositional characteristics, hence it is a physical attribute and can be used as a hydrocarbon indicator by low frequency anomaly, a fracture zone indicator, or a bed thickness indicator. Using the wavelet spectrum, the instantaneous frequency is defined as follows: $\Omega(t, f) = \frac{\partial}{\partial t} \arg W_{gs}(t, f)$.

To avoid the numerical derivatives in the above expression, it is possible to derive an exact analytical formula for the computation of $\Omega(t, f)$ that involves the wavelet transforms of the signal with respect to the analyzing wavelet $g(t)$ and its derivative $g'(t)$. This approach is widely used for instantaneous frequency estimate with Gabor transform (Borum and Jensen 1999) and can be adapted to the case of wavelet transform as follows:

$$\Omega(t, f) = -\text{Im} \left(\frac{W_{g's}(t, f)}{W_{gs}(t, f)} \right). \quad (15)$$

Polarization Analysis

Perhaps the first successful attempt to combine the time-frequency analysis with the polarization analysis of multi-component signals was done by Keilis-Borok et al. (1989).

Afterwards, a great number of the different polarization methods working in the time-frequency domain has been proposed. Some of these methods are based on:

1. The DWT with respect to the orthogonal wavelets and the singular value decomposition (Lilly and Park 1995; Claassen 2001).
2. The S-transform and elliptical motion approximation (Pinnegar 2006; Schimmel and Gallart 2007).
3. The CWT with respect to the orthogonal Mayer's wavelets and the complex covariation method (Soma et al. 2002), the singular value decomposition method (Roueff et al. 2006), the adaptive covariance method (Kulesh et al. 2007), as well as using analytical signals (Burjanek et al. 2009).
4. The wavelet packet decomposition of three-component seismic traces (Lyubushin 2006).

As an example let us consider a simple polarization method based on the complex trace analysis. This method can be applied for the two-component signals. First, we construct a complex signal $z(t) = u_x(t) + iu_z(t)$ from two orthogonal components $u_x(t)$ and $u_z(t)$ of a record to be analyzed. Next, we perform CWT and split the obtained wavelet spectrum into prograde $W_g^+ z(t, f)$ and retrograde $W_g^- z(t, f)$ components using Eq. 11. Then, we can compute the phase difference $\Delta\phi(t, f)$, the major $R(t, f)$ and minor $r(t, f)$ half-axes, and the tilt angle $\theta(t, f)$ of the polarization ellipse in the time-frequency domain as follows

$$\begin{aligned} \Delta\phi(t, f) &= \arg \left(\frac{W_g^+ a(t, f) + W_g^- z(t, f)^*}{W_g^+ z(t, f) - W_g^- z(t, f)^*} \right) + \pi/2, \\ R(t, f) &= |W_g^+ z(t, f)| + |W_g^- z(t, f)|, \\ r(t, f) &= |W_g^+ z(t, f)| - |W_g^- z(t, f)|, \\ \theta(t, f) &= \arg \left(W_g^+ z(t, f) W_g^- z(t, f) \right) / 2. \end{aligned} \quad (16)$$

Thus, the particle motions can be represented by instantaneous attributes that show distinct features for waves of different polarization characteristics. These attributes are explicitly related to the wavelet transform coefficients of the analyzed signal that can be used to perform the complete wave mode analysis (Pacor et al. 2007) and separation/filtering (Diallo et al. 2006) of two-component seismograms in the wavelet domain.

Dispersion Analysis

Surface wave propagation in heterogeneous media can provide a valuable source of information about the subsurface structure and its elastic properties. For example, surface waves can be used to obtain subsurface rigidity through inversion of the shear wave velocity. The processing of

experimental seismic data sets related to the surface waves is computationally expensive and requires sophisticated techniques in order to infer the physical properties and structure of the subsurface from the bulk of available information.

One of these techniques is the calculation and interpretation of the phase and group velocities from the experimental seismograms. The time-frequency representation is today commonly used for the phase and group velocities computation (Levshin et al. 1972; Abbate et al. 1995; Prosser et al. 1999; Niethammer et al. 2001; Pedersen et al. 2003). Most of the methods are essentially based on the study of the square modulus of the representation, which can be interpreted as the energy density in the time-frequency plane. These methods provide only partial information that cannot be directly used to characterize the phase velocities and the different propagation modes.

In order to analyze the dynamical behavior of multi-arrived signals using the model (10), it is also possible to develop a wavelet transformation operator that approximates the wave propagation in the dispersive medium and explicitly incorporates the phase and group velocities as well as the attenuation function (Kulesh et al. 2005):

$$W_g s_m(t, f) = e^{-\alpha(f)D_{mj}} \left| W_g s_j \left(t - \frac{D_{mj}}{C_g(f)}, f \right) \right| \times e^{i \arg W_g s_j(t - D_{mj}/C_p(f), f)}. \quad (17)$$

In this operator, $C_p(f) = f/k(f)$ is the phase velocity, $C_g(f) = 1/k'(f)$ is the group velocity, and $\alpha(f)$ is the attenuation function. The signals $s_m(t)$ and $s_j(t)$ correspond to records obtained from two geophones placed with distance $D_{mj} = D_m - D_j$ apart, and $k(f)$ is the frequency-dependent wave-number function.

The relationship (17) has the following interpretation. The group velocity is a function that “deforms” the image of the absolute value of the source signal’s wavelet spectrum, the phase velocity “deforms” the image of the wavelet spectrum phase, and the attenuation function determines the frequency-dependent real coefficient by which the spectrum is multiplied (see also Fig. 2b, c). Based on this property, one can construct the optimization (Holschneider et al. 2005) and correlation (Kulesh et al. 2008; Waltisberg and Raišutis 2008) methods that allow us to extract both phase and group velocities from the wavelet spectra of multi-arrived signals containing waves with several wave modes.

Summary

In contrast to the Fourier transform, the CWT provides a two-dimensional spectrum of the analyzed signal, where the frequency and the time are treated as independent variables. As a

result, we can investigate the signal in the time domain and in the frequency domain at the same time. The fact that the time-frequency-dependent polarization attributes are explicitly related to the wavelet transform of the signals enables us to perform the wave mode separation in the wavelet domain. It is also possible to construct the wavelet-based dispersion operators that can be used to model wave propagation in a dispersive and attenuating medium.

Cross-References

- [Seismic Instrumentation](#)
- [Seismic Viscoelastic Attenuation](#)
- [Surface Waves](#)

Bibliography

- Abbate A, Frankel J, Das P (1995) Wavelet transform signal processing for dispersion analysis of ultrasonic signals. *IEEE Ultrason Symp* 1:751–755
- Allen RL, Mills DW (2004) *Signal analysis: time, frequency, scale, and structure*. IEEE Press (Wiley-Interscience), Piscataway
- Andria G, Savino M (1996) Interpolated smoothed pseudo Wigner-Ville distribution for accurate spectrum analysis. *IEEE Trans Instrum Meas* 45(4):818–823
- Antoine J-P (2004) Wavelet analysis: a new tool in physics. In: van den Berg JC (ed) *Wavelets in physics*. Agricultural University, Wageningen
- Borum S, Jensen K (1999) Additive analysis/synthesis using analytically derived windows. In: *Proceedings of the 2nd COST G-6 workshop on digital audio effects (DAFx99)*. NTNU, Trondheim, pp 125–128
- Burjanek J, Gassner-Stamm G, Poggi V, Moore JR, Faeh D (2009) Ambient vibration analysis of an unstable mountain slope. *Geophys J Int* 180(2):820–828
- Chui CK (1992) *An introduction to wavelets*. Academic, San Diego
- Claassen J (2001) Robust bearing estimation for three-component stations. *Pure Appl Geophys* 158(1):349–374
- Daubechies I (1992) *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia
- Diallo MS, Kulesh M, Holschneider M, Scherbaum F, Adler F (2006) Characterization of polarization attributes of seismic waves using continuous wavelet transforms. *Geophysics* 71(3):V67–V77
- Grossmann A, Morlet J (1984) Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J Math Anal* 15(4):723–736
- Haar A (1910) *Zur Theorie der orthogonalen Funktionensysteme* (German). *Math Ann* 69:331–371
- Heil C, Walnut DF (1989) Continuous and discrete wavelet transform. *SIAM Rev* 31(4):628–666
- Holschneider M (1995) *Wavelets: an analysis tool*. Oxford University Press, Oxford
- Holschneider M, Diallo MS, Kulesh M, Ohrnberger M, Lück E, Scherbaum F (2005) Characterization of dispersive surface waves using continuous wavelet transforms. *Geophys J Int* 163(2):463–478
- Keilis-Borok VI, Levshin AL, Yanovskaya T (1989) *Seismic surface waves in a laterally inhomogeneous earth*. Kluwer Academic, Boston
- Kulesh M, Holschneider M, Diallo MS, Xie Q, Scherbaum F (2005) Modeling of wave dispersion using continuous wavelet transforms. *Pure Appl Geophys* 162(5):843–855

- Kulesh M, Diallo MS, Holschneider M, Kurennaya K, Krüger F, Ohrnberger M, Scherbaum F (2007) Polarization analysis in the wavelet domain based on the adaptive covariance method. *Geophys J Int* 170(2):667–678
- Kulesh M, Holschneider M, Ohrnberger M, Lück E (2008) Modeling of wave dispersion using continuous wavelet transforms II: wavelet based frequency-velocity analysis. *Pure Appl Geophys* 165(2):255–270
- Levshin AL, Pisarenko VF, Pogrebinsky GA (1972) On a frequency-time analysis of oscillations. *Ann Geophys* 28:211–218
- Lilly JM, Park J (1995) Multiwavelet spectral and polarization analyses of seismic records. *Geophys J Int* 122(3):1001–1021
- Lyubushin AA (2006) Wavelet packet polarization method for an automatic detection of P and S wave arrivals. *Izv Phys Solid Earth* 42(4):30–39
- Mallat S (1999) *A wavelets tour of signal processing*, 2nd edn. Academic, London
- Niethammer M, Jacobs LJ, Qu J, Jarzynski J (2001) Time-frequency representations of Lamb waves. *J Acoust Soc Am* 109(5):1841–1847
- Pacor F, Bindi D, Luzi L, Parolai S, Marzorati S, Monachesi G (2007) Characteristics of strong ground motion data recorded in the Gubbio sedimentary basin (Central Italy). *Bull Earthq Eng* 5(1):27–43
- Pedersen HA, Mars JI, Amblard P-O (2003) Improving surface-wave group velocity measurements by energy reassignment. *Geophysics* 68(2):677–684
- Pinnegar CR (2006) Polarization analysis and polarization filtering of three-component signals with the time-frequency S-transform. *Geophys J Int* 165(2):596–606
- Pinnegar CR, Mansinha L (2003) The S-transform with windows of arbitrary and varying shape. *Geophysics* 68(1):381–385
- Prosser WH, Seale MD, Smith BT (1999) Time-frequency analysis of the dispersion of Lamb modes. *J Acoust Soc Am* 105(5):2669–2676
- Roueff A, Chanussot J, Mars JI (2006) Estimation of polarization parameters using time-frequency representations and its application to waves separation. *Signal Process* 86(12):3714–3731
- Schimmel M, Gallart J (2007) Frequency-dependent phase coherence for noise suppression in seismic array data. *J Geophys Res* 112:B04303
- Soma N, Niitsuma H, Baria R (2002) Reflection technique in time-frequency domain using multicomponent acoustic emission signals and application to geothermal reservoirs. *Geophysics* 67(3):928–938
- Stockwell RG, Mansinha L, Lowe RP (1996) Localization of the complex spectrum: the S-transform. *IEEE Trans Signal Process* 44(4):998–1001
- Wahl TJ, Bolton JS (1993) The application of the Wigner distribution to the identification of structure-borne noise components. *J Sound Vib* 163:101–122
- Waltisberg D, Raišutis R (2008) Group velocity estimation of Lamb waves based on the wavelet transform. *Ultragarasas (Ultrasound)* 63:35–50

Author Index

A

Abers, G.A., 1728
Adam, A., 454
Aerts, W., 547
Agnew, D.C., 1419
Ahmed, I., 1442
Allemmand, T., 1498
Allen, R.M., 282
Anthwar, V., 596
Arora, K., 423, 999
Artemieva, I.M., 1819
Atekwana, E., 33
Aurnou, J.M., 86

B

Babu, E.V.S.S.K., 1256
Bachrach, R., 1280
Baird, R.W., 786
Baker, D.N., 359
Balmino, G., 645
Banerjee, P., 465
Barriot, J.-P., 662
Barton, C.E., 359
Beardsmore, G., 722
Bercovici, D., 1059
Besse, J., 1215
Bhattacharyya, A., 372
Bialas, J., 1155
Biggin, A., 1187
Blanch, J.O., 1137
Bodin, P., 1621
Böhm, S., 149
Bohnhoff, M., 44
Bondár, I., 1644
Bormann, P., 243, 288, 1447
Braitenberg, C., 447
Brown, L., 1319
Bruyninx, C., 547
Bunge, H.-P., 1248
Burov, E., 884
Busse, F.H., 515

C

Campillo, M., 1557
Carlson, D.J., 790
Carlson, R.W., 1
Catherine, J.K., 1802
Cazenave, A., 621
Červený, V., 1472
Chadha, R.K., 1493
Chakravarthi, V., 442
Chambers, J.E., 344, 776
Chambodut, A., 500
Chapman, D.S., 541, 847, 906
Chen, J.L., 621, 1813
Chiozzi, P., 717
Christensen, N.I., 1459
Christensen, U.R., 78
Christophersen, E.A., 1724
Clauser, C., 1304, 1760, 1769
Clay, C.S., 1790
Cloetingh, S., 1794
CoBabe-Ammann, 359
Constable, S., 782
Cooke, M., 1130
Cormier, V.F., 1512
Courtillot, V., 1215
Crampin, S., 1687

D

Das, S., 208, 274
Davis, E.E., 736, 906
Davis, M.G., 541
Davis, P.M., 160
de Mulder, E.F.J., 792
Dentith, M., 993
Derr, J.S., 189
Dewey, J.W., 278
Di Giacomo, D., 288
Dimri, V.P., 399, 405

E

Eder, W., 792
Egbert, G., 1036
Engdahl, E.R., 1644

F

Farnetani, C.G., 1094
Fisher, A.T., 736
Flueh, E.R., 1155
Formaggia, L., 1125, 1133
Forsyth, D.W., 1502
Forte, A.M., 452, 1267
Foss, C., 957, 1310
Fox, P., 359
Freund, F.T., 189
Freymueller, J.T., 558
Fumagalli, A., 1125
Furuya, M., 1335

G

Gahalaut, V.K., 262, 1880
Gaina, C., 1015, 1120
Gallagher, K., 821
Gao, M., 1550
Gamero, E.J., 176
Geddis, E., 861
Ghebreab, W., 1325
Götze, H.-J., 640, 789, 1885
Gray, S.H., 1567
Greff, M., 1215
Grieve, R.A.F., 769
Grünthal, G., 292
Gupta, H.K., 19
Gupta, P.K., 814

H

Hackney, R., 668
Hainzl, S., 1724
Hajra, R., 523
Halgedahl, S.L., 971
Hamilton, D.J., 796
Hanafy, S.M., 1602

Harms, U., 115, 855
 Harris, R.N., 541
 Hase, H., 434
 Hasegawa, A., 1625
 Hasterok, D., 847
 Hearne, M., 308, 316
 Hello, Y., 395
 Henningses, J., 384
 Hermance, J.F., 1042
 Hey, R.N., 1349
 Hinderer, J., 626
 Hinzen, K.-G., 232
 Hofmann, A.W., 1094
 Hrvoic, I., 1029
 Hugill, A., 578

I

Iaffaldano, G., 1248
 Isac, A., 502
 Ismaiel, M., 1085
 Ismail-Zadeh, A., 204
 Iturrarán-Viveros, U., 1283, 1525

J

Jackson, D.D., 53
 Jaiswal, K.S., 308
 James, D.E., 866
 Jaupart, C., 361, 872, 1298
 Jekeli, C., 608, 692
 Jousset, P., 379

K

Kaban, M.K., 585
 Kagan, Y.Y., 53
 Kamogawa, M., 193
 Kato, T., 1705
 Kern, H., 56
 Kijko, A., 1394
 Kim, W.-Y., 1429
 Kind, R., 1580
 Kirby, J.F., 591
 Kobr, M., 527
 Kokolakis, C., 1343
 Korte, M., 514
 Kossobokov, V., 204, 1893
 Kotsiaros, S., 915
 Krawczyk, C.M., 379
 Krishna, K.S., 1085, 1162
 Kück, J., 855
 Kulesh, M., 1909
 Kulhanek, O., 1635
 Kumar, D., 1442, 1487
 Kumar, P.C., 1562
 Kumar, T.V., 1256
 Kundu, B., 262
 Kuras, O., 344, 776

L

La Mura, C., 321
 Laj, C., 507
 Lakhina, G.S., 354, 523
 Lambaré, G., 1498
 Larmat, C., 1790
 Lasocki, S., 8
 Lay, T., 1079

Lee, W.H.K., 1673
 Legrand, J., 547
 Lesur, V., 1015
 Lin, K.-W., 312
 Liu, E., 1385
 Liu, M., 229, 241
 Loke, M.H., 344, 776
 Lowrie, W., 1225

M

Macmillan, S., 488
 Madariaga, R., 224
 Malin, P., 44
 Malin, P.E., 375
 Mallick, K., 596
 Mandal, P., 299
 Mande, M., 502, 1015
 Mangeney, A., 1655
 Manglik, A., 827
 Mangriotis, M.-D., 1897
 Marano, K.D., 308
 Mareschal, J.-C., 361, 872, 1298
 Margerin, L., 1537
 Maruyama, Y., 1425
 Masoudi, A., 384
 McCaffrey, R., 254
 McClain, J.S., 893
 McEnroe, S., 1319
 McGarr, A., 1492
 McNamara, A.K., 176
 Meissner, R., 56
 Mertikas, S.P., 427, 1343
 Meyer, U., 637
 Michael, A.J., 220
 Mishra, D.C., 656
 Mohamed, A.R., 1202
 Montagner, J.-P., 166, 1655
 Morgan, P., 727
 Mulder, W.A., 1149
 Mulyukova, E., 1059
 Musson, R.M.W., 292, 708

N

Nagao, T., 193
 Nagarajan, N., 1007, 1056, 1115
 Nayak, S., 1868
 Neuberg, J.W., 329
 Nolet, G., 183, 395, 1507
 Nothnagel, A., 1902

O

Okal, E.A., 236, 1751
 Orlecka-Sikora, B., 8
 Oro, J., 1661
 Osinski, G.R., 769

P

Padhy, S., 399, 757
 Panda, D., 262
 Panza, G.F., 321
 Parsons, M.A., 359
 Pasquale, V., 112, 717
 Paterson, G.A., 1187
 Patro, P.K., 1050
 Pavlis, N.K., 677, 1714

Pecher, I.A., 1155
 Peltier, W.R., 1107
 Persson, L., 1635
 Peterson, W.K., 359
 Piper, J.D.A., 1202
 Prothero, D.R., 1193
 Pšenčík, I., 1472

Q

Quarteroni, A., 1133
 Quitariano, V., 278

R

Radhakrishna, T., 1202
 Rajesh, R., 1802
 Rama Murthy, V., 1293
 Ramillien, G., 621
 Rani, S., 412
 Ranjbar, H., 1330
 Rao, E.P.R., 1868
 Rao, Y., 1608
 Rao, Y.J.B., 1256
 Rector, J.W., 1897
 Reddy, J.N., 1145
 Reinsch, T., 379
 Rhoades, D.A., 1724
 Richards, P.G., 1429, 1669
 Roberts, P.H., 86
 Robertsson, J.O.A., 1137
 Romanelli, F., 321
 Romanowicz, B., 1738
 Roux, P., 1557
 Roy Chowdhury, K., 127, 1365
 Rucker, D.F., 344, 776
 Rybach, L., 537

S

Sagnotti, L., 923
 Sain, K., 368, 1511, 1562, 1607
 Sambridge, M., 821
 Sánchez-Sesma, F.J., 1283, 1525
 Sandham, W.A., 796
 Sass, J.H., 722
 Satyavani, N., 1567
 Schaff, D.P., 1429
 Scheck-Wenderoth, M., 1353
 Schneider, J.F., 1661
 Schuh, H., 149
 Schuster, G.T., 1407
 Schwartz, M., 486
 Schweitzer, J., 1447
 Scotti, A., 1125
 Selim Saleh, M.M., 37
 Sen, M.K., 807, 1592
 Şengör, A.M.C., 64, 1166
 Shapiro, N.M., 1557
 Sharma, K.K., 596
 Shaw, J., 1187
 Showstack, R., 1685
 Sichoix, L., 662
 Sideris, M.G., 470, 476, 749
 Singh, K.H., 943
 Singh, S.C., 101
 Singh, S.J., 412
 Slater, L., 33

Slosky, D., 312
Smith, A.G., 63, 1274
Smolka, A., 1661
Sobolev, G., 1464
Solomon, S., 1325
Song, X., 144
Srivastava, R.P., 405
Srivastava, V., 1753
Stacey, F.D., 160
Stagpoole, V., 861
Stein, S., 229
St-Laurent, F., 189
Stoffa, P.L., 807
Storchak, D.A., 1447
Stutzmann, E., 1655
Suhadolc, P., 1661
Suttie, N., 1187
Swain, C.J., 591
Szarka, L., 454

T

Tandon, S.K., 1753
Tarling, D.H., 15
Thébault, E., 935
Thériault, R., 189
Thompson, E.M., 316

Thurber, C.H., 301
Thybo, H., 1819
Tiwari, V.M., 626, 656
Tobin, H., 115
Toda, S., 1235
Tóth, T., 1697
Tsurutani, B.T., 354, 523
Tummala, S.K., 1868
Tyburczy, J.A., 176

U

Uyeda, S., 193

V

Vaccari, F., 321
Vaníček, P., 482
Vedanti, N., 405
Verdoya, M., 717
Vinnik, L.P., 1390
von der Osten-Woldenburg, H., 983, 987

W

Wahr, J.M., 156
Wald, D.J., 278, 308, 312, 316
Wang, Y., 1608
Ward, S.N., 1848

Watts, A.B., 831
Wellmann, F., 1885
Wen, L., 164
Whaler, K.A., 350
Whittaker, J., 1120
Wiesemeyer, H., 1902
Wilkinson, P.B., 344, 776
Wood, R., 861
Worden, C.B., 316
Wu, Z., 1429, 1669, 1787

X

Xu, G., 1550

Y

Yamazaki, F., 1425
Yoshino, T., 339
Yu, Y., 1550
Yuan, X., 1580

Z

Zelt, C.A., 1828
Zumberge, M.A., 633

Subject Index

- A**
- Absolute gravity meter (AG), 626, 628
 - Absolute measurements, 1008
 - Absolute palaeointensity, 1187
 - experiments, 1188
 - Absolute positioning, 427
 - Accelerating moment release (AMR), 1725
 - Accelerator mass spectrometry (AMS), 1240
 - Accelerograms, 229
 - Accelerometer-based gradiometers, 705
 - Accretionary subduction zones, 1729
 - Accuracy, 633
 - Acoustic and elastic waves, 762, 766
 - Acoustic and gravity oscillations, 1803
 - Acoustic emission, 24
 - Acoustic gravity waves, 1806
 - Acoustic impedance (AI), 1460, 1489–1491
 - Acoustic phases, 1455
 - Acoustic thermometry, 1752
 - Acoustic wave(s), 1806
 - propagation, 758
 - Active fault(s), 1629
 - zones, 120–121
 - Active remote sensing systems, 1330
 - Active rifting, 75, 1357
 - Active seismic, 1444, 1446
 - Active source seismology. *See* Deep seismic reflection and refraction profiling
 - Active volcano drilling and subduction-induced volcanism, 119
 - Adakites, 1260
 - Adiabatic topographic torques, 90
 - Adjoint methods, 1792
 - ADvancedCIRCulation (ADCIRC), 1875
 - Advanced National Seismic System (ANSS), 1649
 - Advanced spaceborne thermal emission and reflection radiometer (ASTER), 1326
 - Advective heat flux, 736
 - Aeromagnetics, 993
 - Aeromagnetic surveys, 945
 - Afar and Ethiopian rift, 1050
 - AF demagnetization procedure, 1212
 - Africa, 58
 - Aftershocks, 709, 714
 - causes, 230
 - definition, 229
 - earthquake hazard, 230
 - scaling laws, 230
 - Airborne gravity, 638
 - applications and aircraft, 639
 - attitude compensation platforms, 638
 - gravimeter, 579
 - gravity gradient meters, 639
 - principle, 638
 - processing, reductions and corrections, 639
 - scalar gravity meters, 638–639
 - vector gravity meters, 639
 - Airborne magnetic data, 993
 - Airborne magnetic surveys, 937
 - Airborne measurements, 993
 - Air gun, 1700
 - Airy, G.B., 161
 - Airy-Heiskanen isostatic model, 597
 - Airy isostasy, 851
 - Airy method, 161
 - Aleatory uncertainty, 1403
 - Alford rotations, 1693
 - Alfvén's theorem, 80
 - Alfvén velocity, 95
 - Algebraic multigrid (AMG), 1153
 - Alkalic rocks and carbonatites (ARCs), 70
 - All-sky cameras, 1117
 - Alpha decay, 1
 - Alpine uplift and erosion, 1800, 1801
 - Alps, 58
 - Alternating current (AC), 928
 - Altimetry, 836, 1343
 - measurement principles, 1343–1346
 - Ambient magnetic field, 1808
 - Ambient noise, 1443
 - Ambient noise correlation
 - applications in seismology, 1557–1558
 - noise-based monitoring, 1560
 - noise source origin and distribution, 1558
 - seismic imaging, 1558
 - Ambiguity resolution, 552
 - Amplitudes of CID, 1806
 - Amplitude tomography, 1609
 - Amplitude variation with angles (AVA), 1489
 - Amplitude variation with offset (AVO), 802, 1482
 - Amplitude versus offset (AVO), 370
 - Class I anomaly, 370
 - Class II anomaly, 370
 - Class III anomaly, 370
 - gas sand anomalies, 371
 - Analog-to-digital converter (ADC), 778, 785
 - Analogue tectonic modeling, 1800
 - Analytical techniques, 597
 - Analytical theory, 652
 - Analytic signal amplitude, 593
 - Analytic signal peaks, 964
 - Ancient magnetism, 1202
 - Andes, 58, 1251
 - Anelasticity, 173
 - Anelastic structural model, 325
 - Angle-dependent reflection coefficients, 1572
 - Angular unconformity, 1239
 - Angular velocity, 642
 - Anhyseretic remanent magnetization (ARM), 933, 1212, 1229
 - Anisotropic poro-elasticity (APE) model, 1689, 1694
 - Anisotropy, 757, 1463, 1516, 1528
 - media, 1472–1476, 1481–1485
 - of TRM, 1189
 - Anisotropy of anhysteretic remanence (AARM), 933
 - Anisotropy of isothermal remanent magnetization (AIRM), 934
 - Anisotropy of the magnetic remanence (AMR)
 - AARM, 933
 - AIRM, 934
 - HFAMS, 934
 - Anisotropy of the magnetic susceptibility (AMS)
 - analytical techniques, 928
 - bulk low-field, 927
 - Cartesian reference systems, 926
 - diamagnetic and paramagnetic minerals, 926

- Anisotropy of the magnetic susceptibility (AMS) (*cont.*)
 eigenvectors, 926
 ellipsoid shapes, 927
 geology and geophysics, 929–933
 magnetic field, 926
 measurements, 928
 paramagnetic minerals, 928
 petrofabric analysis, 929
 rock-forming minerals, 926
 transient magnetism, 926
- Annual Consultation on the Likelihood of Earthquakes, 1788
- Annual variations, 1009
- Anomalous gravity, 669
 atmospheric correction, 673
 Bouguer correction, 672
 Central Andes in South America, 675
 computing parameters, 669
 computing process, 671
 distant relief effects, 674
 free-air correction, 671
 in geophysics, 670
 normal gravity, 670, 671
 terrain correction, 672–673
- Anorthosites, 1462
- Antarctic Continent, 788
- Antarctic Digital Anomaly Map (ADMAP), 939
- Antarctic Digital Magnetic Anomaly Project (ADMAP), 946
- Antarctic Treaty, 788
- Antenna, 1904
- Antenna-specific errors, 553
- Anthropic seismology, 1655
- Anthropogenic seismicity, 8, 10, 14
- Anticrack, 1633
- Antineutrino energy, 1309
- Anti-Stokes, 388
- Antofagasta (Chile) earthquake, 568
- Apparent crater, 770
- Apparent polar wander paths (APWPs), 1215, 1217, 1231
- Apparent resistivity, 351, 352, 455, 458, 776, 1043, 1044
- Apparent seismic attenuation, 1512
- Aquifer, 717–720
- Arc geochemistry, 1733
- Archaeointensity, 1187
- Archaeological prospecting, 988
- Archaeological surveys, 987
- Archaeology, 803
- Archaeomagnetism
 applications, 18
 archaeological sediments, 18
 chronological dating, 16
 definition, 15
 directional analyses, 16
 errors in dating techniques, 17–18
 features, 15–16
 magnetic dating, 17
 magnetic polarity and polarity excursion dating, 17
 magnetic sourcing, 18
 paleo-intensity analyses, 16
 relative dating, 16
- Archaeoseismites, 233, 235
- Archaeoseismology, 235, 1241
 categories, 233
 quantitative methods in, 235
 schematic work flow, 235
- Archean cratons, 1258
 continental growth curve, 1262
 dome and keel, 1262
 geodynamic regime, 1262
 geological and lithotectonic changes, 1262
 orogenic linear belt, 1262
 vertical and horizontal tectonic regimes, 1262
- Archean-Proterozoic Eons, 1262–1263
- Arc setting, 901
- Argentinean-German Geodetic Observatory (AGGO), 441
- Argentinean satellite SAC-C, 1117
- ARGO float network, 1816
- Array(s), 345
 of seismometers, 1672
- Artificial neural networks (ANNs), 797–802
- Artificial neuron, 799
- Artificial water reservoir triggered earthquakes
 anthropogenic activity, 19
 civil works and engineering projects, 20
 diffusivity, 23
 inhomogeneities, 24
 in situ measurements, 23, 24
 Koyna, 24–28
 Pilot Borehole location, 28
 pore fluid pressure, 23
 reservoir loading, 22
 RTS, 20–23
 triggered/induced seismicity, 19
- Asperities, 1627
- Association, 1432–1433
- ASTER. *See* Advanced spaceborne thermal emission and reflection radiometer (ASTER)
- Asthenosphere, 184, 461, 893, 907, 912, 913, 1257, 1272
- Astronomic azimuth, 429
- Asymmetric rifts, 68
- Asymmetry factor (AF), 1757
- Asymptotic ray tracing methods, 1533
- Atmospheric infrasonic waves, 1803
- Atmospheric pressure, 628–629
- Atomic mass constant, 1762
- Attenuation/internal friction, 1514
- Attenuation, 404, 757, 762, 1596
 model, 1898, 1900
 relations, 323
- Attractors, 410
- Australasian strewn field, 769
- Australia, 852, 1252
- Auto-correlation of wavelets, 1375
- Automatic gain control, 1380
- Automatic PC-controlled monitoring systems, 780
- Avogadro number, 1761
- AVO response, 1899
- Axial gradiometers, 985
- Axial magma chamber (AMC), 101, 104–106, 1164
- Axial melt lens. *See* Axial magma chamber (AMC)
- Axon, 798
- Azimuthally anisotropic reflectivity, 1597
- Azimuthal mode number, 416, 417, 419
- Azimuthal seismic anisotropy, 897
- Azores hotspot, 1091
- B**
- Back-arc, 264, 852, 1174
 spreading centers, 894
- Back-propagation algorithm, 797
- Back-propagation training, 800
- Backscatter coefficient, sigma naught, 1343
- Backus and Gilbert approach, 666
- Backus effect, 1116
- Backus filter, 1376
- Backus–Gilbert method (BGM), 818–820
- Baked contact test, 507, 1230
- Baltic Shield, 58
- Band-limited, 399
 fractal geometry, 404
 power-law, 403
- Bandpass and matched filters, 963
- Bangui anomaly, 919
- Barycentric space-bound celestial reference system, 435
- Basal-shear force, 1270
- Basaltic melt, 352
- Basalts, 1734
- Baseline, 1902
 decorrelation, 1339
- Base loop correction, 658–659
- Basin and range, 59
- Basin elongation ratio, 1757
- Batavia and Gulden Draak microcontinents, 1122–1123
- Batholiths, 1260
- Báth's law, 230, 1724
- Bathymetric measurements, 398
- Bathymetry, 362, 364, 848, 851
 altimetry, 1347
- Bayesian inference, 823
- Bayesian inversion approaches, 565
- Bayesian inversion method, 510
- Beam migration, 1574
- Bedding plane, 929
- Bell Aerospace gradiometer, 702–703
- Benchmarks, 432
- Bering-Kresta Transform Fault, 1177
- Berkner, L.V., 786
- Beta decay, 1
- Bhuj earthquake, 1624
- Bhutan Himalaya, 267
- BIFROST project, 571
- Binary-phase shift keying (BPSK) modulation technique, 548
- Bio-geobatteries, 35
- Bio-geophysics
 cells and biofilms, 33–34
 definition, 33
 disciplinary themes, 34

- metabolic byproducts and mineral weathering, 34
- microbe-mineral transformations, 35
- microbially mediated redox processes, 35
- Biological measurements, 398
- Biot–Savart law, 456
- Biot theory, 1533
 - See also* Poroelasticity
- Bipolar coordinate system, 1532
- Birch, F., 1460
- Birch's law, 1607
- Bispectrum (BS) method, 305
- Bitter pattern method, 975
- Blackett, Patrick, 1202
- Black smoker community, 1353
- Block-diagonal (BD) least squares, 1721–1722
- Block diagonal matrix, 1135
- Block modeling approach, 563
- Blueschist facies, 1259
- Body wave(s), 1607, 1636, 1639–1642
 - experiment types, 1518
 - free oscillations and surface waves, 1518
 - frequency domain, 1517
 - Hilbert transform relation, 1517
 - magnitudes, 243
 - modeling, 43
 - numerical modeling, 1519
 - phases, 897
 - PKnKP waves, 1518
 - reflection-transmission coefficients, 1517
 - relaxation spectrum, 1517
 - viscoelastic attenuation parameters, 1518
- Bolide impacts, 1261
- Boltzmann constant, 1761
- Boninites, 1260
- Boomers, 1700
- Borehole(s), 541–546
 - heat exchanger, 538
 - imaging technologies, 534
 - monitoring, 855
 - and sea-floor instruments, 579
 - temperature systems, 724
 - video, 534
- Borehole seismic networks and arrays, 45–48
 - developments, 44–45
 - GONAF, 49
 - instrumentation and installation, 45
 - noise-suppression and signal-detection advantages, 45
 - SAFOD Main Hole, 49
 - SAFOD Pilot Hole, 48
- Born kernels, 1745
- Bottom-water temperature perturbations, 746–747
- Bouguer anomaly(ies), 643, 663, 673, 751
- Bouguer coherence, 838, 839
- Bouguer correction, 480, 672
- Bouguer gravity anomaly(ies), 587, 597, 1822
- Bouguer gravity map, 601
- Boundary conditions, 644, 1886
- Boundary element method (BEM), 1529
 - advantages of, 1131
 - applications, 1132
 - codes, 1132
 - definition, 1130
 - disadvantages of, 1131
 - displacement discontinuity method, 1132
 - fictitious stress method, 1132
 - vs. finite element method, 1131
- Boundary integral equation method (BIE), 1529
- Boundary-layer cooling theory, 911–912
- Boundary value problem (BVP), 476, 749
 - geodetic BVPs, 477–479
 - multiple vertical datums geodetic BVP, 753–755
 - single (global) vertical datum geodetic BVP, 752–753
- Boundary-value ray tracing, 1476
- Boussinesq approximation, 93
- Box-counting algorithm, 1894
- B-polarization, 458
- Brillouin scattering, 386
- Brittle behavior, 896
- Brittle deformation, 836
- Brittle-ductile transition, 888
- Broadband seismic noise, 1656
- Brownian motion, 400
- Brownian passage time (BPT), 1401, 1788
- Brown model, altimetry, 1347
- Brunhes–Matuyama reversal, 510
- Bruns's equation, 477, 754
- Buchan-type metamorphism, 71
- Building Classification Survey, 1667
- Bulk electrical resistivity, 352
- Bulk modulus, 1762
- Bulk silicate Earth (BSE), 364, 1305
- Bullard plot, 722
- Buoyancy, 1268–1270
- Burger's model of solids, 892
- Burial metamorphism, 71
- Burridge–Knopoff model, 1471
- b*-value of Gutenberg–Richter, 1397–1398
- Byerlee's law, 887, 888
- C**
- Cable coupling, 380
- Caesium vapour sensors, 988
- Calcite precipitation, 35
- Calibration, 505
- Caliper logs, 534
- Calorimeters, 1763
- Candidate models, 502
- Canted-antiferromagnetics, 1320
- Capacitive displacement sensing, 581
- Capacitively coupled instrumentation, 780
- Capacitively coupled systems, 780
- Carbon-14, 4
- Carbon capture and storage (CCS), 9
- Carlsberg Ridge, 1163, 1164
- Carrier phase model, 551–552
- Cartesian coordinate(s), 1473, 1477–1481
 - components, 1268
 - system, 443
- Cartesian distance, 476
- Cascadia subduction zone, 745, 1623
- Cased-hole resistivity logging, 527
- Casualties, 709
- Cauchy's stress and strain, 1284
- Cauchy wavelet, 1911, 1912
- Causality constraint, 211
- Causal linear absorption model, 759
- Cavity decoupling, 1438, 1439
- Celestial intermediate origin (CIO), 151
- Celestial intermediate pole (CIP), 151
- Centennial Catalogue, 714
- Center for Disaster Management and Risk Reduction Technology (CEDIM), 1667
- Central Indian Ridge (CIR), 102
- Central peak structure, 771
- Centrifugal potential, 476
- Chalcophile, 1296
- CHALLENGING Minisatellite Payload (CHAMP), 916, 919, 1117
- Chamoli earthquake, 764
- CHAMP satellite data, 1022
- CHAMP satellite magnetometer measurements, 945
- Chandler oscillation, 152
- Chaos
 - definition, 406
 - theory, 409
- Chaotic rays, 1481, 1483
- Chapman, S., 786
- Characteristic, 712
 - earthquake(s), 54, 1401, 1402, 1405, 1788
- Characteristic remanent magnetization (ChRM), 1208
- Chemical remanences, 15
- Chemical remanent magnetism (CRM), 1204
- Chemical remanent magnetization (CRM), 1321
- Chi-Chi earthquake, 1493
- Chicxulub crater, 769
- Chicxulub structure, 662
- China Digital Seismograph Network (CDSN), 1649
- China's Earthquake Networks Center (CENC), 1650
- China Seismic Intensity Scale (CSIS), 293
- Chinese National Aerogeophysics Surveys, 946
- Chirp signal, 1336
- Choke ring ground plane, 549
- Chondrites, 1305
- Christoffel matrix, 1474, 1475, 1481, 1483
- Citizen science seismology, 280
- Cladding, 384
- Clapeyron slope, 1766
- Classical Brownian motion (cBm), 406
- Classic model of oceanic crust, 1504
- Clausius–Clapeyron equation, 1766
- CLEAN algorithm, 1905
- Climate change, 541–546, 646
- Climate system, 1817
- Climatic loading, 266
- Closed loop error (CE), 658
- CN algorithm, 1788
- CO₂, 194, 195
- Coarse/acquisition code, 548
- Coarse operator, 1135
- Coastal high-resolution model grids, 1875
- Coating, 384
- Cocktail party effect, 1382
- Cocos plate, 745

- Coda attenuation, 761
 Coda-normalization, 763
 CODATA, 644
 Coda wave(s), 1516
 analysis, 760–764
 Coda wave interferometry (CWI), 1547, 1548
 Code division multiplexing (CDM)
 techniques, 549
 Code pseudo-range model, 551
 Coefficient of friction, 9
 Coercivity, 933
 Co-geoid, 480
 Co-geoidal height, 480
 Coherent noise, 1443, 1445
 Cohesion, 9
 Colatitudinal mode number, 416
 Cold atom interferometry gradiometer, 654
 Cold-atom interferometry method, 635
 Collaboratory for the Study of Earthquake
 Predictability (CSEP), 1726
 Collisional orogens, 851
 Collision-controlled orogens, 1175
 Collision plate boundary, 262
 Collision-resistance force, 1270
 Colluvial wedge, 1239
 Coloured inversion, 1489
 Comité Spécial de l'Année Géophysique
 Internationale, 787
 Commission for the Geological Map of the
 World (CGMW), 945
 Commission on the Limits of the Continental
 Shelf (CLCS), 861, 863–865
 Common conversion point (CCP)
 stacking, 1583
 Common image gather (CIG), 1412–1414
 Common-image-point gathers (CIGs), 1571
 Common-midpoint (CMP) stack, 1568, 1569
 Common ray, 1481, 1484
 Common reflection point (CRP), 1572
 Common scatter point (CSP), 1572
 Compass needle, 1000
 Compensated acoustic log, 533
 Completeness magnitude, 13
 Complex propagation velocity, 760
 Complex velocity, 759
 Compositional buoyancy, 519
 Compositional isostasy, 847
 Comprehensive Inversion (CI), 1119
 Comprehensive model (CM), 946
 Comprehensive modeling, 919
 approach, 1119
 Comprehensive Nuclear-Test-Ban Treaty
 (CTBT) monitoring, 1669–1671, 1752
 Comprehensive Test Ban Treaty (CTBT), 48,
 1429–1431, 1435, 1440, 1647
 monitoring system, 1672
 Compressed high intensity radar pulse
 (CHIRP) system, 1699
 Compressibility, 1762
 Compressional waves, 1459, 1580–1581
 velocities, 1460, 1462, 1463
 see also P-waves
 Compressive sampling, 1383
 Compressive stress, 9
 Computational domain, 1133, 1134, 1136
 Computer software packages, 346
 Condition number, 829
 Conduction, 873
 Conductive boundary layer, 893
 Conductive heat flux, 736
 Conductivity, 895
 Conductivity temperature depth (CTD)
 sensor, 397
 Cone penetrometer testing, 1237
 Confidence ellipsoid, 302
 Conglomerate test, 1196, 1229
 Conservation equation, 1815
 Consolidation hypothesis, 1495
 Constitutive law, 758
 Constraints, 644
 Continental crust, dynamics of, 117–118
 Continental crustal structure
 constraints on crustal petrology, 60–61
 cratons, 57–58
 definition, 56
 extensional areas, 58–60
 non-seismic geophysical techniques, 57
 orogens, 58
 seismic methods, 57
 Continental drift, 1905
 definition, 63
 Wegener, A., 63–64
 Continental geotherms, 850
 Continental heat flow, 361
 definition, 722
 global distribution of, 731–734
 heat sources, 729
 and oceanic heat flux, 364
 sampling, thermal conductivity, 726
 statistics, 362
 temperature, 723–724
 thermal conductivity, 724–726
 Continental lithosphere, 728, 836, 838,
 873, 874
 definition, 866
 heat diffusion, 881
 lithosphere–asthenosphere boundary,
 870–871
 mechanical models, 867
 secular transient behavior, 880–882
 seismological/compositional models,
 868–870
 thermal models, 867
 thermal structure, 872–875
 thermal transients, 879–882
 Continental margin, 861–865, 1120
 Continental reference systems, 437
 Continental rifting, 1354
 Continental shelf, 861, 862, 864, 865
 Continents, 850
 Continuous deformation approach, 257
 Continuous Fourier transform, 594
 Continuous free oscillations of the Earth, 420
 Continuous Global Positioning System
 (cGPS), 272
 Continuous linear inverse problems, 818–820
 Continuous monitoring, earth
 environment, 1656
 Continuous observations, 1008
 Continuous wavelet transform (CWT),
 1909, 1910
 inversion formula, 1910–1911
 physical frequency, use of, 1911
 prograde and retrograde wavelet
 spectrum, 1913
 wavelet spectrum amplitude and phase,
 1912–1913
 Continuum fracture models (CFMs), 1126
 Continuum hypothesis, 1532
 Controlled-source EM (CSEM), 782–786
 Controlled-source seismic data
 crosswell tomography, 1842
 crustal velocity model, 1845
 3D first-arrival and reflection
 tomography, 1835
 3D refraction and reflection tomography,
 1836, 1844
 2D tomography, 1845
 geological model, 1839
 minimum-structure model, 1837
 near-surface 3D first-arrival
 tomography, 1840
 nonlinear Monte Carlo procedure, 1840
 ocean-bottom-seismometer, 1835
 onshore-offshore experiment, 1838
 velocity model, 1829
 Controlled-source seismology, 1828.
 See also Deep seismic reflection and
 refraction profiling
 Controlled-source tomography, 1831
 Control points, 428
 Convection, 873, 1248
 mantle (*see* Mantle convection)
 patterns of, 1063
 Convective layer, 893
 Convective plumes, 912
 Conventional hydrocarbon exploitation, 8
 Convergence, 265
 Convergent setting, 1754
 Convolution, 759, 1603
 Convolutional model of seismic trace, 1374
 Coordinated Universal Time, 435
 Coordinate system, 642
 Copernicus program, 1685, 1686
 Cordillera, 851
 Cord rotation, 1679
 Core, 384
 field evolution, 1118
 Core-mantle boundary (CMB), 87–90, 1020,
 1022–1023, 1079–1080, 1258
 Core-mantle coupling
 gravitational torque, 91–93
 length of day time series data, 87
 magnetic torque, 93–98
 topographic torques, 88–90
 viscous torque, 88
 Core-mantle transition zone, 1597
 Coriolis force, 79, 81, 93, 94
 Cornell–McGuire PSHA, 1396–1397
 deaggregation of seismic hazard, 1400
 frequency-magnitude models, 1402–1403
 numerical computation, 1399–1400
 seismic source parameters, estimation of,
 1397–1399
 source-free procedures, 1400–1401
 spatio-temporal models, 1402
 time-dependent models, 1401–1402
 Coronal mass ejections (CMEs), 524

- Corotating interaction regions (CIRs), 524
 Correlation, 1603
 length, 404, 1538
 Correlator, 1903
 Co-seismic crustal deformation, 258–259
 Coseismic crustal displacement, 1808
 Coseismic deformation(s), 468, 564–565
 real-time, 565–566
 velocities from elastic earthquake cycle model, 562
 Coseismic horizontal displacements, 1809
 Coseismic hydrological changes, 1493
 Co-seismic ionospheric disturbances (CIDs), 1803, 1804, 1810
 amplitudes of, 1806
 applications of, 1810
 directivity of CID propagation, 1807–1808
 EIA, 1810
 propagation speed of, 1806–1807
 Co-seismic ionospheric perturbations, 1810
 Cosmogenic isotope palaeointensities, 1187
 Cosmological masses, 642
 Coso Volcanic field in California, 5
 Cosserat elastic theory, 1680
 Costa Rica earthquake, 1639
 Costa Rican subduction zone, 1050–1051
 Cost function, 808
 Coulomb failure criterion, 8, 9
 Coulomb Rate-State model, 1726
 Coulomb's torsion balance, 700
 Counter electrojet, 372
 Coupling, 1472, 1477, 1481, 1484, 1485
 Courant-Friederickson-Lewy stability condition, 1527
 Courant-Friedrich-Levy (CFL) stability condition, 1142
 Covariance, 404
 matrix, 1018, 1019
 Cowling's theorem, 517, 519
 Crack-coalescence, 1695
 Cratering flow-field, 771
 Cratonic root, 185
 Cratons, 57–58, 851
 Cretaceous normal polarity superchron, 1233
 Cretaceous-Paleogene (K-Pg) boundary, 769
 Critical distance, 128
 Crossover distance, 128
 Crust, 168, 894, 1460
 component, 893
 corrections, 1744
 deformation, 630
 density, 849, 851
 evolution, 769
 heat production, 852
 lithospheric field, 1004
 magnetic fields, 517
 profile, 1822
 radioactivity, 852
 rock properties, 60
 roots, 851
 stabilization, 1299, 1302
 thickness, 57, 847, 849, 851, 896
 uplift, 1810
 Crust-mantle coupling, 890
 Crust-mantle decoupling, 890
 Cryogenic gradiometer, 705
 Cryogenic magnetometers, 1194, 1211–1212
 Cryoseismology, 1656
 Cryosphere, 842
 Crystallographic preferred orientation (CPO), 61
 Crystal preferred orientation, 171
 CTBT Organization (CTBTO), 1430, 1440
 Cumulate, 901
 Cumulative distribution function (CDF), 1396, 1399
 Curie depth, 1024
 Curie temperature
 definition, 112
 geomagnetic implications, 113
 and magnetic materials, 112–113
 Curie-Weiss law, 928
 Cycle slips, 553
 Cyclones, 1657
 Cyclostratigraphy, 1233
- D**
- D'Alembert Solution of Wave Equation, 1285
 Damped least square inverse, 817
 Danish satellite Ørsted, 1117
 Darcy's law, 1125
 Darcy's velocities, 1125, 1126
 Data acquisition techniques, 1897–1898
 Data interpretation techniques, 345, 347
 Data inversion problem, 345
 Dataloggers, 1419, 1420, 1424
 Data processing, 997, 998
 Data residual back-propagation, 1612
 Datum, 426
 Davidson-Fletcher-Powell (DFP) method, 1615
 2D crustal velocity models, 1846
 Dead band, 350
 Deaggregation of seismic hazard, 1400
 Death tolls, 710–712
 Debye temperature, 1761, 1762
 Decade of Geopotential Field Research, 1117
 Declination, 1202
 inclination magnetometer (DIM), 1034
 measurements, 1014
 Declinometer, 1034
 Deconvolution, 1374
 and cross correlation imaging conditions, 1573
 De-correlation, 959
 Deep borehole measurements, 738–739
 Deep crustal imaging, 1570
 Deep earthquakes, 1626
 Deep scientific drilling, 115, 122
 Deep seismic(s), 128
 imaging, 1354
 sounding (*see* Deep seismic reflection and refraction profiling)
 zone, 1629
 Deep seismic reflection and refraction profiling
 data acquisition, 129–130
 definition, 127
 in Eastern Europe, 132
 field layouts, 130
 integrated transects, 134
 International Biennial symposia, 132
 (International consortia, 132
 marine-seismic imaging, 133
- multi-channel exploration-seismics, 133
 NMRS recordings, 132
 onshore-offshore, 130
 principles, 128–129
 receivers used in, 129
 recent developments, 137–138
 (re)processing, synthesis and interpretation, 133
 research problems, 138
 sources, 129
 special processing and interpretation tools, 130–132
 time-of-shot recording, 130
 upper-mantle reflectivity-frozen subduction, 134
 WARR recordings, 132
 Deflection-geoid formula, 481
 Deformation, 1250, 1629
 fringe, 1337
 seismometers, 1424
 Deformed alkalic rocks and carbonatites (DARCs), 70
 Deformed rocks, 931
 Deghosting, 1376
 Degree/order, 501
 Degree variances, 1715
 Delta inclination delta declination (DI δ D), 504, 1035
 Demagnetization data, 1208
 Demagnetizing field, 924
 Denali earthquake, 1494
 Denali fault earthquake, 566
 Dendrites, 798
 Density(ies), 177, 633, 847, 849, 1460
 definition, 1607
 rock property, 1607
 Density distribution, Earth
 composition, 163
 density gradient, 162–163
 free oscillation data, 162
 lateral density variations, 164
 mean density, 160–161
 moment of inertia constraint, 161
 self-compression, 161–162
 Depletions, 1816
 Depositional remanent magnetism (DRM), 1204
 Depositional-type remanences, 1187
 Depth-converted stacked trace, 1901
 Depth correlation, 390
 Depth-dependent stretching, 1357
 Depth estimation, 967
 Depth migration, 1570, 1586
 Depth observatory/laboratory, 855
 Depth step, 1575
 De-reverberation, 1376
 Derivative weight sum (DWS), 1837
 Detachment faults, 1165
 Detailed/integrated gravity survey, 660, 661
 Detection, 1431, 1432
 Deterministic deconvolution, 1375–1376, 1702, 1703
 Deterministic seismic hazard analysis (DSHA), 1394, 1395, 1406
 Deterministic seismic hazard assessment, 1788
 Detrital magnetization, 15

- Detrital remanent magnetization (DRM), 1194, 1320–1321
- Deviatoric and explosive components, 12
- 3D first-arrival tomography algorithm, 1834–1837, 1843
- Dharwar craton, 1054
- Diamagnetic minerals, 926
- Diamagnetism, 1203
- Diamond, 7
- Diaplectic glasses, 772
- “Did You Feel It” (DYFI) program, 295
- Dielectric log, 535
- Differential attenuation, 388
- Differential GPS mode (DGPS), 657
- Differential GPS system, 990
- Differential levelling, 432
- Differential rotation of earth’s inner core
definition, 144
seismological observations, 145–146
- Diffraction
Born and Rytov approximation, 1388
definition, 1385
geometrical theory of, 1387
Kirchhoff approximation, 1387–1388
numerical methods, 1388
from single fracture, 1388–1389
theories, 1386
tomography, 1389
wave-field, 1385, 1386
- Diffraction-stack migration method, 1410
- Diffusion, 762
- Digital Earth, 1671
- Digital elevation model (DEM), 470, 1326, 1331
- Digital migration, 1568
- Digital remote sensing data, 1326
- Dilatancy-diffusion model, 1470
- 2D inversion algorithm, 1833
- Diodes, 724
- Dip limits, 1574
- Dipole-dipole array, 780
- Dipole field, 1016
- Dip-slip earthquakes, 324, 1808, 1809
- Direct gravimetric problem, 483
- Directional energy density (DED), 1289
- Directivity of CID propagation, 1807–1808
- Direct methods, 1525
finite difference methods, 1526–1528
variational methods, 1525
- Direct problem of gravity, 644
- Dirichlet boundary conditions, 1135
- Dirichlet boundary-value problem, 617
- Dirichlet’s problem, 476, 1716
- Disaster management, 312
- Disaster risk reduction (DRR), 1667, 1668
- Discontinuous Galerkin method (DG), 1529
- Discrete Fourier transform, 594
- Discrete fracture models (DFMs), 1126–1127
conforming methods, 1127
non-conforming methods, 1127–1128
non-matching methods, 1128
- Discrete inverse problem, 820
- Discrete wavelet transform (DWT), 1909, 1913–1914
- Disk dynamo, 518
- Dislocation creep, 888
- Dislocation theory, 225
- Dispersion, 758–760, 897
analysis, 1914–1915
- Dispersive wave, 1849
- Displacement discontinuity method, 1132
- Displacement seismometers, 1424
- Dissipation, 1268–1270
- Distant Tsunami, 1870
- Distributed acoustic sensing (DAS), 49, 381, 1898
vs. conventional seismic sensing, 381
- Distributed sensing, 386
- Distributed temperature sensor (DTS), 384, 386, 724
- Distributed vibration sensing (DVS), 381
- Distribution error (DE), 658, 659
- Disturbance storm time (Dst) index, 524
- Disturbing potential, 477, 613–614
- Diurnal correction, 958
- Diurnal harmonics, 1009
- Diurnal variations, 997, 1119
- Divergent setting, 1755
- Divided bar, 724
- 2-D model, 346
- Domain decomposition (DD) method
application, 1136–1137
without overlap, 1134–1135
with overlap, 1135–1136
- Doppler effect, 647
- Double-couple, 236–237, 239
- Double couple model, 225
- Double refraction, 1687
- Double square root (DSR) migration
operator, 1572
- Downdip compression (DC), 1629, 1631
- Downdip extension (DE), 1629
- Down-hole oscillator, 724
- Downward continuation, 962
- Downwelling mantle, 1066
- D” layer, 1258
- Draper Lab gradiometers, 702
- Drift correction, 656–658, 660
- Drilling, 528
- Drilling operations, 116
styles of, 117
- Driving forces, 913
- Droughts, 1816
- Dry elastic modulus, 1280
- 4-D time-lapse surveys, 347, 348
- Dulong–Petit law, 1761
- 2.5-D wave modeling, 1618
- DYFI-based Macroseismic Intensity (MI), 278
- DYFI system, 278–281
- Dynamical fault models, 227–229
- Dynamical problem, 649
- Dynamic Composite Elastic Modulus (DYCEM) theory, 1516
- Dynamic earthquake model, 225
- Dynamic height system, 750
- Dynamic poroelasticity, 1282
- Dynamic QMC phase encoding, 1415
- Dynamic ray tracing system, 1479–1482, 1484
- Dynamic topography, 486, 1103, 1795
- Dynamo
advection-diffusion equation, 79
definition, 78
- Ekman number, 80
- equations, 79
- hypothesis motions, 516
- MAC-balance, 81
- magnetic induction equation, 79
- magnetic Prandtl number, 80
- magnetic Reynolds number, 80
- magnetohydrodynamic, 78
- mechanism, 81–82
- models, 79
- non-dimensional control parameters, 80
- Prandtl number, 80
- Rayleigh number, 80
- Reynolds number, 80
- secular variation, 84–85
- self-sustained, 78
- theory, 517
- E**
- Earth, 167, 634
anisotropic and anelastic, 171–173
conductivity, 782
1D-structure of, 168
3D tomographic model, 174
ellipsoid, 483
free oscillations, 629–630
interior characterization with earthquake waves, 167–168
interior structure, 167
layers, 168
mantle, 1248
mantle convection, 169–170
model, 1584
observation, 1343
seismic 3D tomography, 170–171
- Earth gravitational model (EGM), 614
- Earth magnetic anomaly grid (EMAG2), 952
- Earth orientation parameters (EOP), 149, 151–154
- Earthquake(s), 10, 708
aftershocks (*see* Aftershocks)
change in strain energy, 288
clustering, 1894
co-seismic crustal deformation, 258–259
crustal deformation, 630
cycle, 560
death tolls, 712
definition, 254
elastic rebound theory, 255
engineering, 1662
enhanced permeability mechanism, 1495
epicenter distribution, 1633
fault displacement, 710
fault superhighways, 213, 215
features of, 276
on fossil fracture zone, 277
fracture zone, 276
global positioning system, 260
great 1998 Antarctic plate, 277
2018 Gulf of Alaska, 276
Haiti, 712
hazard, 230, 1893
InSAR method, 260
Kanto, 712
LIDAR, 260

- Lisbon, 709, 710
 location techniques, 301–306
 location uncertainty, 302–304
 low-frequency, 331–333
 magnitude, 243–252, 711, 1806, 1893
 mainshock, 229, 230
 multiple-event location, 304–305
 multiple near-simultaneous, 710
 oceanic transform faults, 276, 277
 and plates, 1275
 post-seismic crustal deformation, 259
 precursor, 1809
 pre-seismic crustal deformation, 256–258
 radiated seismic energy, 289–290
 radiation efficiency η_R , 289
 relative location, 305
 response, 308, 312
 risks, 1893
 rupture process, 288
 San Francisco, 709, 710
 scaling, 1893, 1894
 seismic moment, 289
 seismic source, 289
 seismograms, 1893
 sequence, 11
 Shaanxi, 709
 with simultaneous conjugate faulting, 276
 simulators, 1726
 single-event location, 301–302
 stress drop, 289
 strike-slip, 276
 Sumatra, 714
 Syrian, 712
 time reversal and migration location, 305–306
 Tohoku, 712
 tomography, 1828
 triggering, 10
 types, 254
 USLE, 1893–1894
 very-long-period, 335
 volcano-tectonic, 330–331
 waveform cross-correlation, 304–305
 waveform doublets, 146
 2000 Wharton Basin, 277
- Earthquake early warning (EEW)**
 automated mechanical control, 283
 earthquake prediction, 283
 forecast of, 283
 implementations of, 285–287
 onsite warning, 284–285
 personal protection, 283
 P-wave based warning, 284
 rapid detection, 282
 regional, 284
 situation awareness, 283
 strong-motion-based warning, 284
- Earthquake focal mechanism, 238**
 definition, 236
 double-couple, 236–237, 239
 fundamental results, 239
 moment tensor inversion, 239
 non-double-couple, 240
 P-wave first-motion, 237–238
 stress release, 238
 synthetic seismograms, 239
- Earthquake lights (EQLs)**
 classification of, 189–190
 definition, 189
 examples, 190–192
 from Mt. Kimyo, 190
 source and associated electromagnetic phenomenon, 192–193
 from Tagish Lake, 191
- Earthquake precursors and prediction**
 electric, magnetic and EM precursors, 197
 geochemical/hydrological precursors, 196–197
 higher frequency EM emission, 199
 history, 194–195
 lithosphere-atmosphere-ionosphere coupling, 199–200
 macro-anomalies, 196
 pre-seismic EM emissions, 200
 satellite IR images, 197
 telluric current precursors, 197–198
 ultra low frequency precursors, 198–199
- Earthquake prediction, 204, 283, 1788, 1789**
 immediate, 205
 intermediate-term, 204
 long-term, 204
 M8 algorithm, 205–206
 mega-earthquakes, 206–207
 MSc algorithm, 206
 short-term, 205
- Earthquake rupture, 249, 250**
 best-fitting solution, 212
 complexity, 213
 fault, 209
 grid, 212
 inverse problem, 211
 inverse theory, 209
 propagation, 212–213
 seismic moment constraint, 210
 seismograms, 210
 slip rate distribution, 209
 super-shear rupture, 213–215
 tectonic setting, 214, 216–218
 transform faults, 217
- Earthquake sounds**
 applications of sounds, 221–222
 definition, 220
 existence of, 221
 infrasound, 222
 records of sounds, 220–221
 sonification, education and art, 222–223
- Earthquake source theory**
 double couple source, 225
 dynamical fault models, 227–229
 finite source models, 227
 kinematic models, 227
 point source models, 225–227
- Earth rotation, 1906**
 earth orientation parameters, 151–154
 international earth rotation and reference systems service, 154–155
 mathematical formulation, 150–151
- Earth rotation angle (ERA), 151
 Earth rotation parameters (ERP), 149
 Earth's background noise spectrum, 1748
- Earth Science Matters (ESM) Foundation, 795
 Earth's continental crust, 1257
 Earth's core, 164
 potassium radioactivity, 1294–1295
 U and Th radioactivity, 1296
 Earth's mantle, 179
 Earth's rotation, 646
 Earth tide correction, 660
 East-African rift, 60
 Eastern Himalayan Syntaxis (EHS), 265
 East European Craton (EEC), 58, 1821
 East Pacific Rise (EPR), 101, 103–106, 1163–1165, 1503
 East-west hemispheric difference, 165
 Eclogites, 1260, 1462
 Eentec™ model R-1™ rotational seismometer, 1677
- EEPAS model, 1726
 Effective gravitational potential, 89
 Effective normal stress, 9
 Effective plate thickness, 896
 Effective specific heat capacity, 1767
 Effect of Earth ellipticity, 419
 Eigenfrequency, 418, 419
 Eigenfunction, 413–417
 Eikonal equation, 1473–1476, 1485, 1533
 Ekman-Hartmann theory, 96
 Ekman number, 88
 Elastic dislocation approach, 562
 Elastic plate, 896
 Elastic radius, 1438, 1439
 Elastic rebound, 53, 55
 effect, 256
 theory, 1725
 Elastic stiffness, 758
 Elastic wave equation, 759
 Elastic wave propagation, 758–759
 Elastodynamic(s), 1530
 boundary integral equation, 1388
 equation, 1472, 1473
 Electrical anisotropy, 353
 Electrical conductivity, 454–456, 460–462, 464, 1011
 crust, 341–342
 Earth's mantle, 342–343
 liquid-bearing rocks, 341
 of mantle minerals, 342
 oceanic asthenosphere, 343
 Electrical properties of rocks, 339
 electrical conduction, 339–341
 (see also Electrical conductivity)
 electrical conductivity, 339
 Electrical resistivity, 350
 Electrical resistivity surveys, 776
 2-D, 346
 3-D, 347
 4-D time-lapse, 347, 348
 geological parameters, 345
 geometric factor, 345
 profiling, 345
 sounding, 345
 subsurface, 344
 Electrical seismograph model, 1674
 Electrical survey, 344
 Electric field measurement, 782–783

- Electrochemical double layer, 340
 Electrodes, 344, 345, 778, 782, 783
 Electromagnetic (EM), 782
 data acquisition systems, 785
 electric field measurement, 782–783
 feedback, 581
 magnetic field measurement, 783–785
 navigation, 785
 transmitter, 785
 wave impedance, 1043
 wave propagation, 1804
 Electromagnetism, 802
 Electromotive force (EMF), 1210
 Electronic Geophysical Year, 359, 786
 achievements, 360–361
 data preservation and rescue, 360
 electronic data location and access, 360
 history of, 359–360
 operations, 360
 Electronic tiltmeters, 449
 Elevation, 847, 849–851, 853
 correction, 959
 Ellipsoidal approximation, 754
 Ellipsoidal heights, 749, 750, 752, 755, 756
 Ellipticity, 413, 417, 419
Elsasser number, 81, 94, 96
 EMAG2V3, 952
 Embedded discrete fracture model (EDFM), 1128
 Embryonic cleavage, 932
 Emergency management, 316
 Emergency responders, 312
 Emilia-Romagna region, 10
 Empirical equation, 341
 Empirical Green's function method, 325
 Empirical kinematic afterslip models, 566
 Energy flux model (EFM), 761
 Energy function, 808
 Enhanced geothermal response test (EGRT), 392
 Enhanced geothermal systems (EGS), 9
 Enhanced magnetic model (EMM), 953
 Enthalpy, 1762
 Enveloped broadening phenomenon, 1541
 Envelopes, 761
 Environmental magnetism, 1234–1235
 Environmental seismology, 1655
 Eotvos correction, 639
 EOTVOS correction, 656, 659
 Eötvös's original torsion balance, 700
 Epicentral areas, 12
 Epidemic type aftershock sequence (ETAS), 1402, 1725, 1788
 Episodic tremor and slip, 563
 Epistemic uncertainty, 1403
 E-polarization, 458
 Equation of motion, 1473
 Equatorial Electrojet, 1118
 causes, 372
 characteristics of, 372
 definition, 372
 uses, 373
 Equatorial ionization anomaly (EIA), 1809–1810
 Equatorial plasma bubble (EPBs), 1810
 Equatorial shift, 711
 Equipartition principle, 1547
 Equipotential surface, 646
 Equivalent media theories, 1532, 1533
 Error function, 808, 809
 Eruption cycle model, 569
 ETERNA, 658
 Euler deconvolution, 593–594, 967, 968
 Euler frequency, 151
 Euler-Liouville equations, 150
 Euler method, 1134
 Euler pole of rotation., 1231
 Euler's dynamical equations, 150
 Eurasia, 1252
 Euro-Med region, 1648
 European Alps, 1801
 European Association of Remote Sensing Companies, 1686
 European Macroseismic Scale (EMS-98), 293–295
 European Plate Observing System Program (EPOS-ERIC), 14
 Evapotranspiration, 623, 1817
 Exchange anisotropy, 925
 Exothermic and endothermic reactions, 1767
 Explicit finite-difference scheme, 1610
 Exploitation of georesources, 8
 Explosion seismology. *See* Deep seismic reflection and refraction profiling
 Exposure, 1661
 Extended Euler deconvolution, 594
 Extensional areas, 58–60
 Extensometers, 1424
 Extrinsic semiconductors, 340
F
 Faraday, M., 1004
 Faraday's law, 783
 Far-field, 1603, 1806
 Fast Fourier Transform (FFT) techniques, 471–473, 687, 1721
 Fast multipole method (FMM), 1530
 Fatalities, 712–714
 Fault dislocation and static poroelastic strain mechanism, 1495
 Fault displacement, 710
 Fault guided waves, 377
 Fault healing, 1623
 Fault mechanics, 710
 Fault morphology, 255
 Fault resonance, 269
 Fault slip, 55
 Fault zone drilling projects, 121
 Fault zone guided waves (FZGW)
 definition, 375
 geometric spreading, 375
 Fault zone normal and leaky modes, 375
 Fault zone's internal velocity, 377
 Fault zone trapped waves (FZTW), 375
 Feature tracking technique, 1338
 Feldspar, 61
 Fermat's principle, 1507
 Ferromagnetic minerals, 926, 928
 Ferromagnetic objects, 1012
 Ferromagnetism, 1003, 1203
 Ferropericase, 177
 Fiber Bragg gratings (FBG), 385
 Fiber optic distributed strain sensing
 cable coupling, 380
 definition, 379
 localization of sensor position, 381–382
 strain changes along, 380–381
 Fiber optic DTS, 724
 Fiber-optic sensing, 384
 Fictitious stress method, 1132
 Field-aligned currents (FAC), 1002
 Field logistics, 1371
 Field sampling, 1205–1206
 Filtering effect, 1374
 Finite difference, 1610
 approach, 1519
 migration, 1575
 Finite-difference method(s), 1137, 1388
 accuracy and numerical dispersion, 1140–1141
 anisotropy, 1528
 approximations, 1139–1140
 boundary conditions, 1143–1144
 computational resources, 1527
 definition, 1137
 dimensional splitting, 1526
 1-D scalar wave equation, 1526
 implicit method, 1526
 implicit splitting time integration method, 1526
 irregular interfaces, 1527
 Lax–Wendroff correction, 1141–1142
 moment tensor, 1526
 source implementation, 1144
 Taylor series expansions, 1526
 time stepping and stability, 1142–1143
 topography and conformal mapping of grids, 1144
 velocity-stress formulation, 1526
 wave propagation, 1137–1139
 Finite element, 235, 1146
 Finite element concept (FEM), 598
 Finite element method, 346, 1145, 1146, 1525, 1529
 Finite element model(s), 1271'
 of 2-D Navier–Stokes equations, 1147–1148
 of 2-D Poisson equation, 1146–1147
 Finite-frequency kernels, 1509
 Finite-frequency tomography
 cross-correlation, 1509
 Fresnel zone, 1509
 waveform, 1510
 wavefront healing, 1509
 Finite-plate elastic dislocation model, 562
 Finite source models, 227
 First arrival, 1603
 traveltime, 1533
 First motion, 1433, 1437
 Fisher circular distribution, 1209
 Fissure fills, 1239
 Flat earth, 423
 Flat Earth fringe, 1337
 F-layer, 1807
 F2 layer, 1802, 1805
 Flexural lithosphere, 867
 Flexural rigidity, 910
 Flight height, 996

- Floated gradiometer, 702
 Flooding model, 510
 Floods, 1816
 Flow along the well, 392
 Flow direction, 929
 Flowmeters, 535
 Flow outside tangent cylinder, 81–83
 Flow profiling, 390
 Fluid Cube (FC) meta-attribute, 1564–1566
 Fluid density, 1126
 Fluid-extraction-induced seismicity, 12
 Fluid flow, 514
 Fluid-injection-induced seismicity, 12
 Fluid logging, 392
 Fluid saturated rocks, 339
 Fluid substitution, 1281
 Fluxgate gradiometers, 983
 Fluxgate magnetometer(s), 350, 1008, 1029, 1034
 Fluxgate sensors, 988
 Fluxgate-theodolite, 503
 Flux measurement, 1013
 Focal mechanisms, 12
 Focusing/defocusing, 757
 Fold test, 1196
 Football mode, 415
 Force-balance accelerometer, 1423
 Forecasting, tsunami, 1863–1868
 Forensic seismology, 222, 1671–1672
 Foreshock(s), 241
 activation, 1467–1471
 causes of, 242
 characteristics of, 241–242
 definition, 241
 role in prediction of earthquakes, 242
 Formation of rifts and basins, 59
 Formation temperatures, 390
 Forward and inverse modeling techniques, 1132
 Forward-directivity effects, 323
 Forward method, 587
 Forward modeling, 176, 484
 of gravitational gradients, 696
 Fossil magnetism, 1202
 Fossil subduction, 1260
 Fountain effect, 1810
 Four-channel fluxgate gradiometer, 989
 Fourier analysis, 1150
 Fourier power spectrum, 598
 Fourier transform(s), 1285, 1286, 1904
 Fourier transformation, 226
 Fractal(s), 400, 404
 behavior, 407
 definition, 399, 405
 dimension, 400
 porous media, 409
 scaling, 399, 400
 in seismology, 408
 signal analysis, 406, 407
 Fractional Brownian motion (fBm), 406
 Fractional Gaussian noise (fGn), 406
 Fractured porous media
 CFMs, 1126
 computational workflow, 1128–1130
 DFMs, 1126–1128
 Fracture energy (*EG*), 288
 Fracture mechanics, 227
 Fracture zones, 1163
 Fréchet derivatives, 1499
 Fréchet matrix, 1612
 Free-air correction, 671
 Free core nutation (FCN), 152, 1907
 Free oscillations of the Earth, 412, 413, 420
 continuous free oscillations, 420
 Jeans' formula, 419–420
 oscillations of an SNREI Earth model, 415–417
 rotation of the Earth, 417–420
 spheroidal oscillations, 415
 toroidal oscillations, 413–414
 Free-surface boundary condition, 1143
 French GEOSCOPE, 1645
 French hat (FHAT) wavelet, 1911
 Frequency dependence, 1519
 Frequency-dependent anisotropy (FDA), 1900
 Frequency domain, 350
 collocation, 473
 Fresh fault gouge, 1237
 Fresnel zone, 1899
 Frictional strength, 8
 Friction coefficient, 1251
 Fringes, 1904
 FTAN method, 1741
 Fukushima Daiichi nuclear plant, 712
 Full Approximation Scheme (FAS), 1152
 Full multigrid (FMG), 1152
 Full-tensor gradiometer, 700
 Full waveform inversion (FWI), 1384, 1411, 1898, 1899
 Full-waveform systems, 779
 Full width at half maximum (FWHM), 1308
 Fully normalized Associated Legendre functions, 1714
 Fundamental equation of physical geodesy, 477
 Fundamental mode studies (dispersion), 1741–1742
 Fundamental point, 752
 Fused quartz sensors, 579
G
 Gabbros, 900
 Gakkel Ridge, 102
 Galerkin approach, 1153
 Galerkin approximation, 1529
 Galerkin coarse-grid approximation, 1151
 Gamma–gamma or formation density log, 533
 Gamma model, 950–951
 Gardner equation, 1608
 Gardner relation, 1607
 Garlock fault, 563
 Gas-sand detection, 370
 Gassman's equation, 1281
 Gauss coefficient, 1016–1018
 Gaussian distributions, 1019
 Gaussian noise, 400
 Gaussian normal distribution, 1308
 Gauss internal coefficients, 501
 Gauss-Seidel relaxation, 1150
 Gawler Craton, 1054–1055
 Gazdag migration, 1575
 GED4ALL, 1667
 GED4GEM, 1667
 Geiger's method, 301
 Generalized Bruns's formula, 752
 Generalized extreme value (GEV), 1405
 Generalized horizontal spatial gradient method, 455
 General relativity, 1907
 Generated injection air guns, 1701
 Generic mapping tools (GMT), 945
 Genetic algorithms (GA), 797, 810
 Geobarometers, 735
 Geocentric axial dipole (GAD) hypothesis, 1217, 1225–1227
 Geocentric axisymmetric dipole field, 1277
 Geocentric celestial reference system (GCRS), 151
 Geocentric dipole, 1016
 Geocentric Earth-bound terrestrial reference system, 435
 Geocentric octopole, 1016
 Geocentric quadrupole, 1016
 Geochemical fingerprinting, 1260
 Geochemical measurements, 398
 Geochemistry, 1890
 Geochronology, 1259
 Geodesy, 427, 646, 1902
 actual shape of the Earth, 423
 continental reference system, 437
 datum, 426
 definition, 423, 434, 442
 distribution of mass, 445
 disturbing potential, 444
 earth's surface using mean sea level, 424
 'Figure of the Earth', 423
 fundamental station, 440
 geoid and reference ellipsoid, 444
 geoid undulations, 425
 global, 439–442
 global geodetic reference system, 436
 gravity field of earth, 443
 gravity models, 446
 gravity reduction, 446
 historical development, 442
 horizontal reference systems, 439
 in situ measurements, 445
 international services, 439–440
 ITRS, 436–437
 metrology, 435–436
 motional forces, 424
 national reference system, 438–439
 polar motion, 424
 properties of gravitational potential, 443–444
 Pythagorean concept of spherical Earth, 423
 reference system and reference frame, 435
 satellite, 445
 survey, 442
 WGS84, 437
 Geodetic baselines, 428
 Geodetic BVPs
 analytical solutions, 478–479
 Molodensky's BVP, 478
 operational solution, 479
 Stokes's BVP, 477, 478
 Geodetic datum, 431, 436

- Geodetic height system, 750
 Geodetic leveling, 427
 Geodetic precession, 1907
 Geodetic techniques, 1249
 Geodynamic(s)
 computational algorithms, 453
 definition, 452
 forces studied in, 453
 geophysical phenomena studied in, 453
 models, 263
 modern research, 453
 parameters in, 453
 Geodynamo, 144, 517, 519, 916
 definition, 78
 energetics of, 81
 magnetohydrodynamic theory of, 520–522
 model, 511, 1001
 simulations of, 78
 time dependence of, 522
 Geoelectric strike direction, 351
 Geoelectromagnetism
 asthenosphere, 461
 boundaries of mega-units, 461
 crustal conductivity anomalies, 461–463
 definition, 454
 middle crust, 461
 MTS, 457–460
 MVP, 456–457
 MVS, 455
 near-surface, 463
 Geoenergy, 124
 Geographic information system (GIS), 1325, 1328–1329
 Geoid, 465, 589, 646, 896
 anomaly and tectonics, 468
 definition, 466
 and heights, 466
 height variations, 470
 modeling, from gravity data, 466–467
 Geoidal height
 applications, 485–486
 direct gravimetric problem, 483–484
 inverse gravity problem, 484
 truncated geoid, 484–485
 Geoid determination, 470
 fast fourier transform, 471–473
 frequency-domain collocation, 473
 least-squares collocation, 473
 terrain effect, 474
 Geoid undulation, 749, 751–754.
 See also Geoidal height
 Geological Survey of Finland, 945
 Geomagnetically induced currents (GICs), 523, 1004
 on electric power systems, 525
 generation, 525
 severity, 525
 Geomagnetic conditions, 1806
 Geomagnetic core field, 514
 Geomagnetic cyclotron frequency, 1803, 1804
 Geomagnetic depth sounding (GDS), 782, 784, 1004, 1048
 Geomagnetic dipole moment, 515
 Geomagnetic dynamo, theoretical modeling, 1010
 Geomagnetic excursions, 522
 definition, 486
 field configuration, 487
 geographical and temporal extent, 487
 origin, 487
 Geomagnetic field, 194, 503, 915, 920, 1007, 1808, 1810
 changes in, 489–499
 definition, 488
 direction, 1203
 intensity of, 1187
 maps of, 489
 observations, 488–489
 removal of, 998
 spherical harmonic analysis, 489
 Geomagnetic jerks, 515, 522
 Geomagnetic polarity time-scale (GPTS), 507, 945, 1233
 Geomagnetic reversals
 changing frequency over time and the duration of, 508–509
 discovery of, 507
 dynamo mechanisms, 511–512
 future research, 512
 morphology of, 509–511
 polarity timescale, 507
 Geomagnetic storms, 915, 1806
 Geomagnetic survey, 989
 Geomagnetic variations, 1009
 Geomagnetism, 787, 1016, 1017
 Geometrical spreading, 758, 761, 1477, 1480, 1483–1484
 Geometrical theory of diffraction, 1387
 Geometric heights, 749
 Geometric spreading, 1513
 Geomodeling
 complexity, 1886
 deterministic interpolations, 1887
 error and accuracy, 1885
 forward modeling, 1886
 geophysical interpretation, 1888
 geophysical inversion, 1888
 geostatistics, 1886
 human geo-expertise, 1889
 interdisciplinary 3-D structural modeling, 1888
 interface description, 1887
 model, 1885
 relevant scale, 1886
 sensitivity analysis, 1886
 simplicity, 1886
 simplification, 1886
 uncertainty, 1885
 validity, 1885
 Geomorphic markers, 1755
 Geoneutrinos, 1296, 1308–1309
 Geoparks, 796
 Geophone, 1368, 1419, 1422
 Geophysics, 167, 646
 exploration techniques, 345
 fluids, 1815
 indirect effect, 674–675
 inversion, 796–798, 801, 802
 as a remote-sensing method, 958
 time series, 407
 well logging (*see* Well logging)
 Geophysical Observatory in Eastern Sea of Marmara, 122
 Geophysical Observatory on the North Anatolian Fault (GONAF), 49
 Geopotential differences, 753
 Geopotential model (GM), 470, 753
 computation, 471
 Geopotential number, 750
 Geopotential research, 916
 Geo-resources exploitation, 8–14
 Geoscientific insight, 115
 Geoscientific research, 857
 GeoSoft, 951
 Geostatistical interpolation, 1887
 Geostrophic flows, 94
 Geostrophic motion, 485
 Geotherm, 848, 850
 Geothermal energy production, 8
 Geothermal gradients, 729
 Geothermal heat pump
 applications, 540
 closed loop and open loop systems, 539
 components, 538
 definition, 537
 design and installation, 539
 heating and cooling with, 538
 licensing, environmental benefits, 539
 resources, 537
 types of, 538
 Geothermal record of climate change, 541–542
 global geothermal climate-change database, 543–544
 repeat temperature logs, 545
 temperature-depth profiles, 544–545
 temperature reconstruction analysis, 545
 Geothermal wells, 391
 Geothermometers, 735
 German GEOFON global networks, 1645
 Gilbert, William, 516
 Glacial earthquakes, 1655–1657
 in Antarctica, 1656
 Glacial isostatic adjustment (GIA), 558, 630, 1813
 in Fennoscandian region, 572
 mantle viscosity (*see* Mantle viscosity)
 and sea level studies, 570–573
 Glaciers, 788
 Glacier topography, altimetry, 1348
 Global Active Faults, 1667
 Global and regional tomography models, 1796
 Global borehole network stations, 48
 Global cooperation, 788
 Global dynamic reconstruction model, 1797
 Global Earthquake Hazard Map, 1664, 1665
 Global Earthquake Model (GEM)
 activities, 1668
 Building Taxonomy, 1667
 collaboration network, 1664
 Foundation, 1662, 1663
 global databases, 1667
 hazard, 1664
 initiative, 1662
 institutional framework, 1663, 1664
 project, 1400
 resources and technical services, 1668
 risk, 1665

- scientific framework, 1663
 social vulnerability and resilience, 1666
 software and analysis tools, 1666–1667
 training and capacity development program, 1668
- Global Earthquake Risk Map, 1665, 1666
 Global field modeling, 1016–1021
 Global Geodetic Observing System, 439
 Global geodetic reference system, 436
 Global Geoscience Initiative, 796
 Global gravitational model (*GGM*)
 accuracy and resolution, 680
 error propagation, 678
 evaluation and accuracy assessment, 688, 689
 high-resolution, development of, 679
 operational and scientific uses, 678–679
 self-consistency, 678
 signal representation and data characteristics, 680–684
 synthesis, 678
- Global Historical Earthquake Catalogue, 1667
 Global Instrumental Earthquake Catalogue, 1667
 Global isostatic adjustment (GIA), 621, 623, 624
 Global magnetic anomaly map
 aeromagnetic and marine compilations, 945–946
 aeromagnetic surveys, 943
 definition, 943
 early developments, 945
 EMAG2, 952
 EMM, 953
 line-levelling method, 948
 LSC, 948, 949
 merging satellite models, 949
 NGDC-720 model, 952
 preparation, 947, 948
 satellite complications, 946
 satellite data, 943
 short and long wavelength anomalies, 945
 WDMAM2.0, 954, 956
 WDMAM, 943, 949–951
- Global mid-ocean ridge systems, 1087
 Global models, 185
 Global Monitoring for Environment and Security (GMES) program, 1685
 Global Navigation Satellite System (GNSS), 431, 1803, 1804, 1810
 Global optimization
 applications, 811–813
 definition, 807
 genetic algorithms, 810
 neighborhood algorithm, 810
 particle swarm optimization, 811
 simulated annealing, 809–810
 trans-dimensional inversion, 811
- Global paleoreconstruction, 1279
 Global positioning system (GPS), 260, 431, 547, 657, 785, 833, 1112, 1326, 1327, 1424, 1803–1807, 1809, 1810
 antennas, 549
 carrier phase, 551–552
 code pseudo-range, 551
 ground instrumentation, 549–551
 measurement, 264, 551–552
 observations, 265
 receiver, 549–551
 receiver-dependent error, 552–553
 satellite-dependent biases, 553
 satellites, signals and system time, 547–549
 signal propagation biases, 553–555
 single and double differences, 552
- Global Risk Assessment Framework (GRAF), 1668
 Global Seismic Network, 48
 Global Seismographic Network (GSN), 1645
 Global tomographic models, 1081
 Global Vulnerability Database, 1667
 Global water cycle, 1813
 GMPE Strong Motion Modeller's Toolkit, 1666
 Goddard Space Flight Center (GSFC), 945, 1113
 Gondwana, 63
 Goringe Bank, 709
 Göttingen Magnetic Union, 516
 GPS Earth Observation Network System (GEONET), 1706
 GRACE Follow-On (GFO) mission, 1813
 Gradient, 894
 Gradient impedance (GI), 1490
 Gradiometer, 649, 988
 advantages and limitations, 983–984
 areas of application, 984
 conceptual design of, 983
- Gradiometry, 1681
 Granitoids, 29
 Gravimeters, 578
 design principles, 580–581
 external noise sources, 582–583
 intrinsic error sources, 581–582
 performance requirements, 579
 relative, 579
 sensor technology, 579–580
 transportation, 583
- Gravimetric factor, 159
 Gravitational anomalies, 643
 Gravitational constant, 641
 Gravitational potential, 649, 650
 Gravitational potential energy (GPE), 469
 Gravitational torque, 91–93
 Gravity, 896
 acceleration, 641, 642
 anomalies (*see* Anomalous gravity)
 data, 886
 disturbance, 585
 disturbance vectors, 477
 on Earth's surface, 633
 field, 626–628, 630, 631, 648
 gradient meters, 639
 measurement, 633, 1815
 and magnetic signature, 1822
 observed, 669
 potential, 476, 750
 potential data, 483
 solutions, 1815
 vector, 476
 waves, 1806
- Gravity anomalies, 646, 669, 754, 772
 Gravity anomaly(ies), 477, 646, 669, 754, 772
 accuracy, 600
 band-pass filtering, 602
 enhancement, 587
 finite element method, 604
 forward method, 587
 in Geodesy, 597
 for geodetic purposes, 675
 global studies, 586
 inverse method, 588–590
 Klamath Mountains and Cascade Range region, 603
 local studies, 586
 paradox Basin, Utah, 601
 problem formulation, 586
 regional, 597
 regional anomalies, 600
 regional gravity field, 598
 regional studies, 586
 residual, 602
 spectral approach, 600
 types of elements, 599
- Gravity field and steady-state ocean circulation explorer (GOCE) gradiometer, 704
 Gravity field of Earth, 608
 disturbing potential, 613–614
 Green's function, 612–613
 interior gravity field, 619–620
 low-degree spherical harmonics, 615–617
 Newton's law of gravitation, 609–611
 normal reference field, 617–619
 spatial and temporal variation, 611
 spherical harmonics, 611–612
- Gravity gradiometry
 applications, 705–706
 Bell Aerospace, 702–703
 cryogenic gradiometer, 705
 definition, 692
 floated gradiometer, 702
 full-tensor gradiometer, 700
 gravitational gradients, 692–693
 Hughes rotating, 701
 measurement error analysis, 698–699
 modeling gravitational gradients, 695–697
 reference field and power spectral density, 697–698
 static, 692
 superconducting, 703
- Gravity-mapping satellite missions, 684, 685
 Gravity methods, 640
 ambiguity and principle of equivalence, 644
 celestial bodies, 641
 gravitational constant, 644
 gravity units, 644
 inverse problem, 644
 land surveying, 640
 Newton's law and Newtonian potential, 641–643
 objective of, 640
 potential function, 641
 sensor systems, 641
 standard unit, 644
 technical possibilities, 640
- Gravity method, surface
 elevation and position measurements, 657

- Gravity method, surface (*cont.*)
 gravity measurements, 657
 hydrocarbon exploration, 660–661
 land gravity survey, 657–659
 marine gravity survey, 659–660
 minerals, groundwater and engineering sites, surveys for, 661
 regional surveys, for geodynamic studies, 660
- Gravity modeling, 585
 definition, 662
 non-harmonic, 662
 spectral approaches, 667
 theory and computation, 662
- Gravity Recovery and Climate Experiment (GRACE), 272, 1112, 1813
 GOCE space missions, 663
 satellite gravimetry, 1814
 Science Data System, 622, 1814
 twin satellite system, 468
- Gravity Recovery and Climate Experiment (GRACE) mission, 612, 621–622, 624
 co-seismic and post-seismic deformations, 624
 data errors, 622–625
 GIA effect, 624
 hydrology, 623
 land ice loss, 623–624
 modelling temporal geoid variations, 622
 ocean mass change, 624
- Great California ShakeOut earthquake, 223
- Great circle path, 1806
- Great Sumatran Fault, 711
- Great Thermal Divergence in the mantle, 1262
- Green functions, 225, 612–613, 758, 1612, 1617
 homogeneous space, 1286–1287
 reconstruction, 1557
- Green's equivalent layer, 592
- Greenstone belts, 1260
- Grey body law, 882
- Grid-based models, 1889
- Gridding technique, 998, 1468, 1471
- Grid rotated staggered grid, 1527
- Grid search approach, 302
- Grid spacings, 1527
- G Ring laser gyro, 1676
- Grotta Gigante horizontal pendulum, 448
 causes that deform, 451
 free oscillations, 449
 time series of tilt, 449, 451
- Ground-based interferometric radar systems, 430
- Ground motion model (GMM), 1394, 1403
- Ground-motion modeling, 766
- Ground-motion prediction equations (GMPE), 322. *See also* Ground motion model (GMM)
- Ground positioning, 427–431
- Ground surface temperature (GST), 542, 545
- Ground truth (GT) events, 304, 1670
- Ground water flow
 energy equation, 717
 heat advection, 717–719
 heat convection, 720–721
 hydrothermal system, 719–720
- Groundwater storage change (ΔGW), 1816
- Group velocity, 1740, 1915
- GRS80-ellipsoid, 436
- Grüneisen parameter, 1762
- Gulf of Guinea impact tsunami, 1862, 1866
- Gutenberg discontinuity, 184
- Gutenberg–Richter frequency-magnitude relation, 1395–1399, 1401–1403
- Gutenberg–Richter law, 230
- Gutenberg–Richter relation, 13
- Gutenberg–Richter relationship, 13, 1893
- Gutenberg–Richter's law, 1788
- ## H
- Haicheng earthquake, 242
- Haiti earthquake, 712
- Hales discontinuity, 184
- Halite, with low density, 1607
- Hall conductivity, 372
- Hamiltonian, 1475, 1484
- Harmonic splines, 1021–1022
- Haskell's model, 227
- Hawaiian–Emperor hotspot track, 1219
- Hazard curve, 1396, 1397, 1399, 1400, 1403, 1405, 1406
- Hazard Modeller's Toolkit, 1666
- Heat advection, 717–719
- Heat capacity, 1761–1762
 isobaric and isochoric specific heat capacity, 1762–1765
 volumetric heat capacity/thermal capacity, 1764–1766
- Heat conductivity. *See* Thermal conductivity
- Heat convection, 717, 720–721
- Heat diffusion–advection equation, 1760
- Heat flow, 728, 850, 852, 853, 893, 894
 Fourier's law, 728
- Heat flow, seafloor
 bottom-water temperature perturbations, 746–747
 deep borehole measurements, 738–739
 dependence on age and global average, 743–745
 gas hydrate stability, 739–741
 hydrothermal circulation, 741–743
 shallow measurements in marine sediments, 737–738
 subduction, 745–746
- Heat flux, 1298–1301
 at the base of continental crust and lithosphere, 875–879
 from the core, 365
 and heat production, 876–878
 Moho, 877–879
- Heat generation, 850
- Heat pipes, 1257
- Heat producing elements (HPE), 1298
- Heat production, 850, 853
- Heat-pulse method, 392
- Heat transport mechanism, 729–731
- Heavy oil production, 391
- Height anomaly, 478
- Height changes, 635
- Height correction, 751
- Height systems, 749–752
- Helbig analysis, 1313
- Helicity, 522
- Helioseismology, 1557
- Helmert orthometric heights, 751
- Helmert's second condensation reduction, 479
- Helmholtz equation, 1409
- Helmholtz Green's function, 1530
- Hematite, 924, 1320
- Hessian matrix, 1614
- Heterogeneities, 404, 757, 760
- Heterogeneity, 399, 1518
- High-field magnetic anisotropy (HFAMS), 934
- High field strength elements, 1260
- High-frequency seismology, coda wave analysis, 760–764
- High-resolution body wave tomography, 1507
- High-speed streams (HSSs), 524
- High-temperature, 392
- Himalayan earthquakes, 263
- Himalayan mountain range, 1051–1053
- Himalayan Seismic Belt (HSB), 268
- Himalaya/Tibet, 58
- Himalaya–Tibetan orogenic system, 262
- Hokkaido Tokachi-oki earthquake, 565
- Homogeneous dynamo, 518
- Homomorphic deconvolution, 1375
- Hooke's Law, 1284, 1609
- Hopping conduction, 340
- Horizontal angle, 428
- Horizontal gradient amplitude, 593
- Horizontal national reference systems, 439
- Horizontal-to-vertical (H/V) spectral ratio method, 1289, 1427
- Hotine's integral formula, 1716
- Hot-spot(s), 173, 363, 1216, 1257, 1277–1278
 definition, 1086
 global, 1089
 volcanism, 1088
- Huber loss function, 1038
- Hughes rotating gravity gradiometer, 701
- Human-induced earthquake, 271
- Hurricanes, 1657
 typhoon, 1655
- Hurst exponent, 399
- Huttonian uniformitarian approach, 1258
- Huygens' principle, 1530
- Hybrid mapping, 1905
- Hydroacoustic, 1440
- Hydration, 893
- Hydraulic fracturing, 9, 391, 1898, 1899
- Hydrocarbon exploration
 detailed/integrated gravity survey, 660–661
 reconnaissance gravity survey, 660
- Hydrogen darkening, 389
- Hydrogen masers, 1903
- Hydrological cycle, 648
- Hydrology, 623, 630
- Hydrophone, 395
- Hydrostatic equilibrium, 832
- Hydrostatic pressure generating systems, 1461
- Hydrothermal activities, 1165
- Hydrothermal circulation, 737, 741–743, 895, 902
- Hydrothermal metamorphism, 71
- Hydrothermal systems, 719–720

- Hydrothermal vents, 1352
 Hypervelocity impact, 769
 Hypsometric curve and hypsometric integral (HI), 1756
- I**
 IAGA Working Group-V Mod, 502
 Iberian Peninsula, 709
 Ice cap(s), 647
 volume changes, 648
 Iceland deep drilling project (IDDP), 124
 Iceland hotspot, 1092
 Ice mass change, 1816
 ICGEM service, 654
 Identification, 1433–1435
 Igneous rocks, 931
 composition, 1260
 Image co-registration, 1337
 Image registration, 1338
 Impact cratering, 769
 Impact craters, 120
 Impact melt rocks, 772
 Impedance matrix, 1045
 Impedance tensor, 351
 Implicit finite-difference, 1610
 Implosive element, 12
 Inclination, 1115, 1202
 Incorporated Research Institutions for Seismology (IRIS), 1645
 India, 1252
 Indian Ocean, 1252–1253
 India-Shan plateau motion, 1881
 Indirect boundary element method (IBEM), 1529
 Indirect effect on gravity, 480
 Indirect effect on the geoid, 480
 Indirect effect on the potential, 480
 Indo-Gangetic plains, 263
 Induced currents, 1011
 Induced earthquakes, 255
 Induced polarization (IP) surveys, 347, 778
 non-polarizable electrodes, 778
 Induced seismicity, 10, 19
 Induction coil magnetometers, 1014
 Induction coils, 350
 Induction vector(s), 351, 454, 456, 457, 461–463
 Inertial-grade accelerometer, 580
 Inertia tensor, 616
 Infinite-frequency forward modeling
 ray tracing, 1830
 wavefront tracking methods, 1830–1831
 Information Commons, 360
 Infrasound, 222
 Inhomogeneous isotropic media, 1476
 Initial-value ray tracing, 1476
 Inland water, altimetry, 1348
 Inner core boundary (ICB), 87
 Inner core, earth
 anisotropy, 166
 boundary, 165
 deformation of surface, 166
 topography, 166
 Inner core tangent cylinder, 81
 Instrumentation, EM, 635
 data acquisition systems, 785
 electric field measurement, 782–783
 magnetic field measurement, 783–785
 navigation, 785
 transmitter, 785
 Insurance Development Forum (IDF), 1668
 Integral equation methods, 1525, 1530
 Integral (Kirchhoff) migration, 1573
 Integrated electron content (IEC), 1804
 Integrated monitoring, 348
 Integrated Risk Modeling Toolkit (IRMT), 1667
 Intelligent electrodes, 779
 Intensity data fields, 713
 Intensity, earthquakes
 conversion of intensity scales, 295
 history of intensity scales, 293, 294
 intensity attenuation with distance, 295–296
 macroseismic scales and intensity assignment, 292–293
 macroseismic surveys, 295
 magnitudes, 296
 nature of sensors, 293
 strong ground motion, 296–297
 Interactive map viewer, 1666
 Interdisciplinary collaborations, 14
 Interferometer, 1902
 Interferometric synthetic aperture radar (InSAR), 260, 1327
 fundamental principles of, 1336–1338
 limitations, 1340–1342
 processing, 1338–1340
 Interferometry, 1602, 1899
 INTERMAGNET, 505, 1007
 Intermediate-depth earthquakes, 1626, 1633, 1634
 DC-type upper-plane, 1632
 dehydration embrittlement, 1632
 DE-type lower-plane, 1632
 double couple radiation patterns, 1632
 double seismic zone, 1631
 phase transformation, 1632
 seismicity, 1627, 1632
 slab-derived aqueous fluids, 1633
 Internal mass redistributions, 1220
 International Association of Geomagnetism and Aeronomy (IAGA), 945
 International Association of Seismology and Physics of the Earth's Interior (IASPEI), 1663
 Commission on Seismological Observation and Interpretation, 1447
 Magnitude Standard of 2013, 1456
 Seismic Format, 1448
 Standard, 1637
 International Celestial Reference Frame (ICRF), 151, 650, 1906
 International Celestial Reference System (ICRS), 1906
 International Continental Drilling Program (ICDP), 26
 International Continental Scientific Drilling Program (ICDP), 859
 International Council of Scientific Unions, 787
 International Data Centre (IDC), 1647, 1669
 International earth rotation and reference systems service (IERS), 151, 154–155, 1905
 International Geodynamics and Earth Tide Service (IGETS), 627
 International Geomagnetic Reference Field (IGRF), 500, 936, 937, 947, 1115
 correction, 960
 model, 1116
 International Geophysical Cooperation, 786
 International Geophysical Year (IGY), 359
 achievements, 788
 Gold, 786
 definition, 786
 history, 786–787
 operations, 787
 International gravity formula, 789
 International Great Lakes Datum of 1985 (IGLD85), 751
 International Height Reference System, 749
 International Heliophysical Year, 786
 International Lithosphere Program, 855
 International Monitoring System (IMS), 1647, 1669
 International Polar Year (IPY), 786
 International Polar Year 2007–2008, 790
 achievements, 791–792
 challenges, 792
 goals, 791
 history, 791
 International Registry of Seismograph Stations (IRSS), 1654
 International science year, 359
 International Scientific Studies of the Implementation of the CTBT Verification System" (ISS), 1670
 International Seismological Center (ISC), 1507, 1667
 International Terrestrial Reference Frame (ITRF), 151, 426, 650, 1905
 International Terrestrial Reference System (ITRS), 151, 425, 436–437
 International Union of Geodesy and Geophysics (IUGG), 1672
 International Union of Geodesy and Geophysics General (IUGG), 945
 International Union of Geological Sciences (IUGS), 796
 International VLBI Service for Geodesy and Astrometry (IVS), 1905
 International Year of Planet Earth (IYPE), 786, 792–796
 legacy item, 794
 Interplanetary coronal mass ejection (ICME), 524
 Interplanetary magnetic field (IMF), 524
 Interplate coupling, 1626
 Interplate earthquakes, 1625
 seismogenic zone, 1626, 1627
 strong ground motions, 1628
 thermal structures, 1626
 transition zone, 1626
 updip and downdip depth limits, 1626
 Interpolating/shape function, 596
 Interpretation, 585
 Interseismic models, 561

- Interval stationarity, 12
- Intraplate earthquakes, 1622–1623
 classification, 1622
 controversy, 1624
 definition, 299, 1621
 examples, 1623–1624
 generation of, 299
 vs. interplate earthquakes, 1623
 locations and mechanisms of, 300
 occurrences of, 299
 source of driving stresses, 1623
- Intraplate rifts, 73
- Intraplate seismicity, 886
- Intraslab earthquakes, 1625
 characteristics, 1629
 intermediate-deep earthquakes, 1633, 1634
 intermediate-depth, 1631–1633
 mechanisms, 1629
 metastable olivine wedge, 1631
 slab mineralogy, 1629
 subducting oceanic plates/slabs, 1629
- Intrinsic absorption, 761, 762
- Intrinsic attenuation, 758, 761, 1513, 1516
- Intrinsic conduction, 340
- Intrinsic resistivity, 1044
- Intrusion Cube (IC) meta-attribute, 1563–1565
- Intrusive rocks, 931
- Inundation mapping, 1874
- Inundation modeling and mapping,
 1874–1875
- Inventory Data Capture Tools (IDCT), 1667
- Inverse AMS fabrics, 933
- Inverse gravity problem (GIP), 484–485
- Inverse matrix, 1018
- Inverse method, 588–590
 back propagation, 1612
 covariance matrix, 1611
 frequency domain acoustic wave
 equation, 1612
 objective function, 1611
 steps, 1612
 virtual sources, 1612
- Inverse modeling, 797
- Inverse problem, 644, 662, 797
- Inverse Q -filtering, 760
- Inverse Stokes formula, 481
- Inverse theory, global optimization. *See* Global optimization
- Inverse theory, 209
 Bayesian inference, 823
 linear, 827–829
 Markov chain Monte Carlo, 823–826
 Monte Carlo method, 822–823, 826
 nonlinearity and multimodal fitness
 function, 822
- Inverse Vening Meinesz formula, 481
- Inversion technique, 1584
- Inversion theory, 1407–1410
- Inversion using overtones, 1744–1745
- Inverted reflectivity response, 1901
- Ionospheric coupling factor (ICF), 1808
- Ionospheric delay, 1804
- Ionospheric disturbances, 1803
- Ionospheric dynamo, 1002
- Ionospheric magnetic field, 918
- Ionospheric perturbations, 1809, 1810
- Ionospheric pierce point (IPP), 1803, 1806
- Ionospheric plasma, 373
 critical frequency, 1809
- Ionospheric refraction, 554
- Iridium, 396
- Irregular surface, 166
- Irving, Edward, 1203
- Isobaric heat capacity, 1761
- Isochoric heat capacity, 1761
- Isochron diagrams, 3
- Isochron imaging, 1572
- Isosismals, 293, 295, 296
- Isostasy, 586, 646, 831, 832, 843, 844, 847
 Airy model, 832–834, 836, 840
 and landscape evolution, 842–843
 long-wavelength gravity and topography
 anomalies, isostasy and mantle
 dynamics, 843–844
 Pratt model, 832
 time-scales of isostatic adjustment,
 840–842
 Vening-Meinesz model, 832
- Isostatic anomaly, 596
- Isostatic compensation, 896
- Isostatic equilibrium, 831, 840, 847
- Isostatic gravity anomalies, 834
- Isothermal remanent magnetization
 (IRM), 1322
- Isotope geochemistry, 1259
- Isotropic linear poroelastic media, 1282
- Isotropic mechanism, 12
- Isotropic media, 1472–1485
- Iterative gradient method, 1409
- Iterative super-virtual refraction interferometry
 (ISVI), 1606
 advantages, 1606
 limitations, 1606
 noisy data, 1606
- J**
- Jack Hills Hadean detrital zircon, 1262
- Jacobian matrix, 346, 827
- Jacobi relaxation, 1150
- Jan Mayen microcontinent, 1122
- Japanese Meteorological Agency (JMA) scale,
 292, 293, 297
- Japan Seismic Hazard Information Station
 (J-SHIS), 1426
- Japan Trench Fast Drilling Program/Project
 (JFAST), 121, 123
- Jeans' formula, 419–420
- Joint migration inversion (JMI), 1899
- Joint Verification Experiment, 1436, 1437
- Juan de Fuca Ridge (JFR), 101
- K**
- Kalman filters, 1019
- Kamioka Liquid Scintillator Antineutrino
 Detector (KamLAND), 1309
- Kanto, 712
- K-band ranging (KBR) system, 1814
- Kelvin-Christoffel equations, 1463
- Kerguelen plume, 1092
- Key Regions for Enhanced Monitoring and
 Preparedness, 1788
- Killer wave. *See* Tsunami
- Kilaton, 1430, 1431, 1435, 1437, 1439, 1441
- Kinematic dynamo theory, 517–518
- Kinematic earthquake model, 225, 227
- Kinematic GPS, 565
- Kinematic migration, 1499
- Kinematic “turbulent viscosity, 88
- Kirchhoff approximation, 1387–1388
- Kirchhoff migration, 45, 1572, 1573
 method, 1410
 practical considerations in, 1574
- Klamath Mountains and Cascade Range, 602
- Klauder wavelet, 1375
- KMAH index, 1478, 1479, 1483
- Kobe earthquake, 1494
- Koenigsberger ratio, 1310
- Kolarovo anomaly, 485
- Kontinentales Tiefbohrprogramm der
 Bundesrepublik Deutschland
 (KTB)*, 855
- Kopp's law, 1763
- Kostrov's self-similar circular crack, 227
- Koyna, India
 near-source studies, 26, 28
 seismograms, 24
 short-term earthquake, 25, 26, 28
 triggered earthquakes, 24, 25
- Krebs method, 1415, 1416
- Kriging, 948
- KTB depth laboratory (KTB-DL)
 commercial utilization, 859
 data infrastructure, 857
 deep borehole site, 859, 860
 drilling facilities, 856
 establishment, 856
 geoscientific experiments, 859
 multi-phase approach, 856
 objectives, 857
 ongoing monitoring, 855
 operational phase, 856
 outreach and education, 859
 test and calibration, 859
 ultradeep borehole measuring, 856
- Kursk formation, 517
- L**
- Laboratory experiments, 352
- LaCoste suspension, 581
- Lagrangian floats, 395
- Lahars, 331
- LAI coupling. *See* Lithosphere-atmosphere-
 ionosphere coupling (LAI coupling)
- Lamellar magnetism, 1320, 1321
- Landers earthquake, 564
- Land gravity survey, 657–659
- Landsat, 1685
- Landslides, 1657
- Land water cycle, 1815
- La Palma landslide tsunami, 1861, 1862
- Laplace equation, 476, 481, 641, 662,
 918, 1016
- Laplace transform, 1417
- Laplacian operator, 611
- Large igneous province (LIP), 1101
 magmatism, 1102–1103
- Large low shear velocity provinces (LLSVPs),
 180, 1083–1084

- Laschamp excursion, 487
 Laser scanning, 233
 Laser systems, 647
 Last Glacial Maxima/Maximum (LGM), 621, 624, 833, 1108, 1111
 Latent heat, 1766–1767
 Lateral density variations, 164
 Lateral heterogeneities, 1046
 Laterologs, 532
 Lattice conductivity, 882
 pressure dependence, 883
 Lattice-preferred orientation, 1693
 Lava-dike transition, 119
 Lax–Wendroff correction, 1141–1142
 Layer 2A, 901
 4-Layer model, 901
 Lead zirconate titanate (PZT), 1460
 Leaking mode, 1449
 Leap-frog scheme, 1140
 Least-squares adjustment, 1718
 Least-squares collocation (LSC), 473, 479, 686, 947–949, 952
 Least-squares inversion, 1341
 Least-squares method, 346, 1018, 1195, 1526
 Least-squares migration (LSM), 1410–1412
 Least-squares resistivity inversion, 344
 Least-squares solution, 755
 Leeds model, 951
 Legendre polynomial, 622
 Legendre spectrum, 614
 Lehmann discontinuity, 184
 Lehnert number, 96
 Leibniz, 709
 Length of day (LOD), 149
 variations, 86, 87
 Leonhardt and Fabian approach, 512
 Levelling, 427, 432
 Level-Set method, 1887
 Levinson recursion, 1377
 Lid, 896
 LiDAR surveys, 26
 Lifeline operators, 312
 Light detection and ranging (LIDAR), 260
 Light scattering mechanisms, 386
 Limited-memory BFGS (L-BFGS) algorithm, 1613, 1614, 1621
 Limited Test Ban Treaty (LTBT), 1429, 1430
 Lineament mapping, 1325, 1328
 Linear arrivals, 128
 Linear gravimetric inverse problem, 664
 Linear interpolation, 1887
 Linear inverse theory, 827–829
 Backus–Gilbert method (BGM), 818–820
 generalized inverses, 816–818
 linear inverse problems, 815–816
 resolution and error propagation, 820
 Linear transformation, 797
 Linear viscoelastic attenuation, 1512
 Linear viscoelasticity
 anelastic hysteresis, 1513, 1514
 Q and complex velocity, 1514
 $Q^{-1}(\omega)$ relaxation spectrum, 1514
 rheology, 1513
 velocity dispersion, 1515
 Line-levelling, 947, 948, 952
 Line-of-sight (LOS), 1803, 1805–1807
 Line source, 1618
 Line-source technique, 725
 LIP. *See* Large igneous province (LIP)
 Lippmann–Schwinger equation, 1408
 Liquefaction, 233
 phenomenon, 324
 Liquid-bearing rocks, 341
 Lisbon earthquake, 709, 710
 Lithium disposable batteries, 396
 Lithology, 61, 1325
 Lithosphere, 832–836, 838, 840, 842, 849, 850, 893, 896, 1249, 1257, 1258, 1262, 1268–1271, 1462, 1504, 1893
 continental, 836, 838 (*see also* Continental lithosphere)
 defined, 872, 873
 oceanic, 836, 837
 Lithosphere and asthenosphere, 893
 Lithosphere–asthenosphere boundary (LAB), 870–871, 893, 897
 depth, 898
 Lithosphere–asthenosphere system, 1258
 Lithosphere, atmosphere and ionosphere (LAI), 1804
 Lithosphere-atmosphere-ionosphere coupling (LAI coupling), 199–200
 Lithosphere, mechanical properties of
 brittle-plastic properties, 888
 data uncertainties, 890
 elastic behavior, 887
 flexural behavior and lithosphere strength, 886
 horizontal deformation, 890
 oceanic lithosphere vs. continental lithosphere, 888–890
 postglacial rebound timescales, 885
 post-seismic timescales, 885
 quantitative information, sources of, 885–886
 rheological properties, 887
 tectonic deformation, 890–891
 viscous-ductile properties, 888
 Lithospheric anomaly maps, 1118
 Lithospheric field, 916–917, 1003, 1118
 Lithospheric geotherms, 879
 Lithospheric magnetic anomalies
 global models, 919–920
 mathematical background, 917–919
 Lithospheric magnetic field, 1023–1027
 Lithospheric magnetization, 940
 Lithospheric plate, 893
 Lithospheric temperatures, 852
 Lithospheric underplating, 1124
 L1 norm of discrepancies, 664
 L2 norm of discrepancies, 664
 Local energy minimum (LEM) domain states, 975
 experimental evidence, 978
 thermomagnetic treatments, 978–979
 Local gravimetric models (LGMs), 680
 Locally coherent reflected events, 1498
 Local oscillators, 1904
 Local Tsunami, 1869
 Location, 1431, 1433
 of earthquake epicenters, 1670
 LOD. *See* Length of day (LOD)
 Lodestones, 515, 1202
 experiments, 516
 Logic tree, 1404–1405
 method, 1553
 Longitudinal wave, 1448
 Long-term drift, 582
 Long term secular variation, 1008
 Lorentz force, 90, 517–522
 Louisville hotspot, 1219
 Love numbers, 158, 159
 Love waves, 897, 1449, 1739
 Low Earth Orbiting (LEO), 935
 satellites, 1115
 Lower atmosphere, 1810
 Low-frequency earthquakes, 254, 331–333
 Low-seismic-velocity zone (LVZ), 1164
 Low velocity zone (LVZ), 893, 896
 Lyapunov exponents, 1481, 1483
M
 M8 algorithm, 205–207, 1788
 MacCullagh's formulas, 617
 Machine learning (ML), 1490
 Mach-Zehnder interferometer, 381
 Macro-anomalies, 194–196
 Macroscopic falling test mass, 634
 Macroseismic observation, 316
 Mafic basalts, 900
 Magma flow, 931
 Magma mush oceans, 1260
 Magmas, 1734
 Magmatic arcs, 1257
 Magmatism in orogens, 1179
 Magnesiowüstite, 177
 Magnesium-iron silicate mineral, 177
 Magnetic anisotropy
 AMR, 933–934
 AMS, 925–933
 energy, 972
 exchange, 925
 magnetic properties of rocks and minerals, 923
 magnetostriction, 925
 petrostructural phenomenon, 929
 principles, 923–924
 of rocks, 923
 Magnetic anomaly(ies), 772, 967
 crustal and lithospheric fields, 936–937
 definition, 935
 Legendre functions, 935
 planes and ships, 937, 938
 power spectrum, 935, 936
 qualitative interpretation, 939, 940
 quantitative interpretation, 940–941
 satellite measurements, 938
 SH model, 935
 spatial resolution, 936
 WDMAM, 939
 Magnetic coupling paradox, 98
 Magnetic data processing, 957
 Magnetic diffusion layer (MDL), 96
 Magnetic domains
 description, 971
 domain width vs. grain size, 973
 exchange energy, 972

- Magnetic domains (*cont.*)
 experimental studies, 975–981
 grain-size dependence, 977
 in magnetic minerals of paleomagnetic significance, 976–977
 micromagnetic models, 974–975
 near crystal surfaces, 973–974
 structure, classical models of, 972–974
 structure, evolution of, 979–980
 structure in magnetite, 980–981
 structure, temperature dependence of, 974
- Magnetic domain walls
 energy and width of, 972, 973
 near crystal surfaces, 973–974
 widths, 977–978
- Magnetic field, 648, 926
 geometry, 84
 measurement, 783–785
 reversals, 84–85
 sources, 935
- Magnetic field data
 de-culturing of, 959
 definition, 1310
 enhancement of, 961
 expression of magnetization direction in, 1311–1312
 gridding of, 959–960
 image display of, 961–962
 vertical continuation of field, 962
 vertical gradient, 963
- Magnetic flux, 783
- Magnetic foliation, 927
- Magnetic gradiometer, 983
- Magnetic gradiometry, 985, 987, 991
 advantages of, 988
 caesium vapour sensors/fluxgate sensors, 988
 standard gradiometers, 988
- Magnetic isochron, 1249
- Magnetic layer bottom (MLBD)
 determination, 113
- Magnetic lineation, 927
- Magnetic measurements, 398, 503, 1115
- Magnetic methods, 999
- Magnetic modeling
 CMB, core magnetic field at, 1022–1023
 developments in, 1019–1021
 harmonic splines, 1021–1022
 inverse problem and implementation, 1018–1019
 lithospheric magnetic field, Arctic and African regions, 1023–1027
 spatial and temporal description, 1016–1021
 wavelet analysis, 1022
- Magnetic needles, 999
- Magnetic observatories, 503, 915, 917, 920, 1000, 1007
- Magnetic permeability, 455, 457, 517
- Magnetic potential field, 1009
- Magnetic Prandtl numbers, 95
- Magnetic pulsations, 1012
- Magnetic remanence, 972
- Magnetic repeat stations, 1000
- Magnetic Reynolds number, 515, 519
- Magnetic satellite missions, 501
- Magnetic sourcing, 18
- Magnetic storms, 958, 988, 997
- Magnetic stratigraphy
 applications and implications, 1197–1201
 calibration and correlation, 1196–1198
 definition, 1193
 detrital remanent magnetization, 1194
 field and laboratory techniques, 1194–1196
 strengths and weaknesses, 1196–1197
- Magnetic surveys, 1009
- Magnetic susceptibility, 924–926, 1012
 tensor, 926
- Magnetic theodolite, 1014
- Magnetic torque, 93–98
- Magnetic variations
 parametrized models of, 1119
- Magnetisation, 937, 940
- Magnetism, 999
- Magnetite (Fe_3O_4), 112, 923
 bearing rocks, 1207
- Magnetization, 926, 1013
 complex distribution of, 1315–1316
 contrast and anomaly separation, 1311
 direction and source location, 1313
 direction and source shape, 1313–1314
 direction in magnetic field data, 1311–1312
 directions, 1207
 homogeneity and inhomogeneity, 1315–1316
 induced and remanent, 1310
- Magnetized rocks, 515
- Magnetobiostratigraphy, 1198
- Magnetocrystalline anisotropy, 923, 924, 972
- Magnetogram, 988
- Magnetohydrodynamic dynamos, 519
- Magnetohydrodynamics, 1023
- Magnetometers, 983
 definition, 983, 1029
 magnetic field measurement, 1029
 scalar, 1030–1031
 vector, 1033–1035
- Magnetosphere, 1009
- Magnetospheric ring current, 1117
- Magnetostatic energy, 972–973
- Magnetostratigraphy, 1222, 1232, 1233
- Magnetostriction, 925
- Magnetostrophic approximation, 94
- Magnetotactic bacteria, 34
- Magnetotelluric (MT), 782–786
- Magnetotelluric (MT) data processing, 1036
 strike and distortion analysis, 1041–1042
- Magnetotelluric (MT) imaging
 Antarctica, 1054
 Costa Rican subduction zone, 1050–1051
 definition, 1050
 Dharwar craton, 1054–1055
 Ethiopian rift and Afar, 1050
 Gawler Craton, 1054–1055
 Himalayas, 1051–1052
 San Andreas Fault, 1052
- Magnetotelluric (MT) method, 350, 1004
- Magnetotelluric (MT) transfer functions, 1036, 1037
 error estimates, 1039–1041
 least squares transfer function estimation, 1037
- leverage points, 1039
- remote reference estimates, 1039
- robust estimation, 1037–1039
- Magnetotellurics (MT), 782–786
 apparent resistivity, 1043
 case studies, 1048–1049
 characteristic impedance, 1043
 depth of penetration, 1044
 3D models, 1048
 electromagnetic wave impedance, 1043
 fundamental concepts, 1043
 geophysical technique, 1042
 lateral heterogeneities, 1044
 multi-station arrays, 1047
 resistivity, 1044
 surface impedance, 1043
- Magnetotelluric sounding (MTS), 459
 basic formulas, 457–459
 interpretation, 459–460
- Magnetovariational profiling (MVP), 455, 460
 array transfer function, 456
 single site transfer function, 456–457
- Magnetovariational sounding (MVS), 455, 460
- Magnetovariational studies (MVS), 1004, 1056
 conductivity distribution, 1057
 interpretation, 1058
 measurement, 1056
 spectral analysis, 1057
 transfer functions, 1057
- Magnitude, 13, 708, 710, 711, 713, 714
- Magnitude distribution models, 13
- Magnitude of earthquakes
 definition, 243
 magnitude saturation, 249
 non-saturating magnitude scale, 249–251
 relationships between magnitude scales and released seismic energy, 248
 Richter scale, 244–245
 teleseismic magnitude scale, 245–248
- Magnitude saturation, 249
- Magnitude threshold, 1432, 1435, 1441, 1671
- Magsat concept, 1117
- MAGSAT mission, 1116
- Main field, 501
- Main Himalayan Thrust (MHT), 264
- Mainshock, 229, 230
- Man-made earthquake signals, 128
- Mantle, 894, 903
- Mantle component, 893
- Mantle convection
 definition, 1059
 energy sources for, 1064–1065
 and generation of plate tectonics, 1072–1076
 hypothesis, 1267
 jigsaw, 169–170
 layered vs. whole, 1070
 phase transitions on, 1067
 theory of, 1059
 Wegener's theory, 1059
- Mantle D'' layer, 1079
 large low shear velocity provinces, 1083–1084
 post-perovskite, 1082–1083
 seismic velocity anisotropy, 1084–1085

- seismic velocity discontinuities, 1081
 seismic velocity model, 1080–1082
 ultralow velocity zones in, 1084
- Mantle energy budget, 363–367
- Mantle flow, 1736, 1794, 1796
- Mantle geochemistry, 1067–1071
- Mantle-lithosphere interactions, 1797, 1799
- Mantle mixing, 1067–1071
- Mantle plume, 1087, 1357
 activity, 1092
 definition, 1085
 Indian Ocean, 1088
 Kerguelen, 1088
 location of, 1088
 volcanism, 1093
- Mantle plumes, 1120, 1257
 continental flood basalts and breakup, 1099–1100
 definition, 1094, 1095
 existence of, 1097
 genesis and dynamics, 1099–1100
 geochemistry, 1097–1099
 hotspots, 1095–1097
 LIP magmatism and environmental effect, 1102–1103
 lithosphere interaction, 1103
 ridge interaction, 1104
 strength and melting point, 1100
 volcanism, 119
- Mantle reflections, 1260
- Mantle relaxation, 255
- Mantle rheology, 1065–1066
- Mantle transition zone, 343
- Mantle viscosity
 definition, 1107
 depth dependence, 1112–1113
 mantle convection process, 1113–1114
 methods of measurement, 1107–1108
 natural experiments, 1108
 observational data sets, 1108–1112
- Mapping extreme field variations, 965–967
- Mariana-Izu-Bonin system, 1728
- Marine geodesy, 1343
- Marine gravity survey, 659–660
- Marine heat flow, 745
- Marine magnetic surveys, 1012
- Marine seismic data, 1617, 1618
- Marine seismics, 128
- Markov approximation, 1541
- Markov chain Monte Carlo (McMC), 821
 fixed dimension approach, 823–824
 transdimensional approach, 824–826
- Marmousi velocity model, 1408
- Marquês de Pombal Thrust, 710
- Mars Atmosphere and Volatile Evolution (MAVEN), 919
- Mars Global Surveyor (MGS), 919
- Marussi telluroid, 478
- Mascon, 1815
- Mass change, 1815
- Mass continuum, 642
- Mass flux, 179
- Mass motion, 1423
- Mass transport, 1813
- Mass Transport Deposit Cube (MTDC) meta-attribute, 1564–1566
- Mass transport deposits (MTDs), 1565
- Matched filter, 1792
- Matter-wave interferometry, 634
- Maximum angular deviation (MAD), 1209
- Maximum credible earthquake (MCE), 1394, 1398
- Maximum magnitude ($m_{\max}$)
 deterministic procedures, 1398–1399
 probabilistic procedures, 1399
- Maxwell equations, 1016
- Maxwell field equations, 456
- McMC. *See* Markov chain Monte Carlo (McMC)
- Mean dynamic topography (MDT), 752
- Mean errors per degree, 649
- Mean free path, 1539
- Mean lithosphere, 1215
- Mean sea level (MSL), 466, 752
- Mean sea surface, 647
- Measurement accuracy, 389
- Measurement bandwidth, 652
- Measurement noise, 552
- Megacities, 709
- Megathrust, 711
- Melt fraction, 352
- Memento Mori, 223
- Memory tools, 724
- Mercalli-Cancani-Sieberg scale (MCS), 293, 294
- Merging satellite models, 949
- Meridional currents, ionosphere, 1118
- MERMAID, 395, 397
- Meta-attribute
 FC, 1564–1566
 IC, 1563–1565
 MTDC, 1565–1566
 SC, 1564–1565
 TFC, 1563–1564
- Meta-heuristics, 809
- Metal spring-mass sensors, 579
- Metamorphic reactions, 1890
- Metamorphic rocks, 932
- Metamorphism, 1463
 in orogens, 1182
- Metasomatic infiltration, 1099
- Meteors, 190
- Methane-hydrate bottom simulating reflector, 741
- Method of splitting tsunami, 1872
- Metrology, 435
- Mexico City earthquake, 1425
- Mg-perovskite, 1080
- Micro-accelerometer, 648
- Microbe-mineral transformations, 35
- Microbial metabolism, 34
- Microcontinents, 1121
 Batavia and Gulden Draak microcontinents, in Eastern Indian Ocean, 1122–1123
 definition, 1120
 formation mechanisms of, 1122–1124
 Jan Mayen microcontinent, in North Atlantic Ocean, 1122
- Micro-leveling, 959, 998
- Micromagnetic models, 974
- Micropulsation, 997
- Microscopic falling test masses, 634
- Microseismic band, 1656
- Microseismic noise, 583
- Microseisms, 1655
- Microtremor measurements, 1426
- Mid-Atlantic Ridge, 102, 1164, 1166
- Mid-crustal ramp, 269
- Mid oceanic ridge basalts, 365
- Mid-ocean ridge (MOR), 900, 1257, 1258, 1502
 crustal structure, 1503–1505
 internal structure, 1163, 1164
 magmatism, 118
 mantle structure, 1502–1504
 morphology and spreading rates, 1162, 1163
 ridge-axis morphology, 1164, 1165
 seismic structure, 1164, 1165
 spreading rate, 1164, 1165
- Mid-ocean ridge basalts (MORBs), 900, 1068, 1097
- Migration, 1576
 algorithm, 1578
 aperture, 1574
 depth, 1570
 poststack, 1573
 purposes of, 1571
- Migration velocity analysis (MVA), 1411–1415
- Milne's methods, 301
- Mine masking, 1438, 1439
- Mineral exploration, 634
- Mineralogy, 1461, 1462, 1629, 1631
- Mineral physics, 178
- Minimum curvature algorithm, 947
- Minimum norm (MN) inverse, 817
- Mining collapses, 12
- Mining seismicity, 12
- MM-31 scale, 294
- Modal-summation technique, 325
- Model misfit function, 1410
- Models of plate motions, 559
- Mode number
 azimuthal, 416, 417, 419
 colatitudinal, 416
 radial, 417
- Mode summation technique, 420
- Modified seismic gap hypothesis, 54–55
- Modulus defect, 1514
- Mogi model, 569
- Moho boundary, 1092
- Moho depth, 833, 834, 836
- Moho heat flux, 877–879, 1298, 1299, 1302
- Moho refracted phase, 1512
- Mohorovičić discontinuity, 184
- Moho topography, 886
- Molar gas constant, 1761
- Molodensky's BVP, 478
- Moment tensor, 12, 1430, 1435
 inversion, 239
- Momentum conservation, 1250
- Monte Carlo analysis, 304
- Monte Carlo (MC) method, 797, 822–823
- Moore–Penrose inverse (MPI), 818

- MORBs. *See* Mid-ocean-ridge basalts (MORBs)
- Morlet wavelet, 1911
- Mountain belts, 1249
- Mountain front sinuosity (Smf), 1757
- Mountain glaciers, 1817
- MSc algorithm, 206
- Mt. Everest measurement, 467
- Multiaxial perfectly matched layer (M-PML), 1534
- Multi-channel capability, 779
- Multi-channel experiment, 902
- Multi-channel multi-electrode systems, 779
- Multi-channel seismics, 1698
processing considerations for high-resolution, 1701–1704
sources, 1699–1701
- Multidisciplinary approaches, 1889
- Multi-electrode systems, 778
- Multi-fractality, 399
- Multigrid method, 1149
fundamentals of, 1150
generalizations, 1152–1153
geophysical applications, 1153
nonlinear, 1152
- Multigrid solver, 1151
- Multi hazard vulnerability mapping (MHVM), 1875
- Multilayer perceptron (MLP), 797, 799
- Multimodality, 13
- Multi-mode fibers, 380
- Multiple lapse time window analysis (MLTWA), 763
- Multiple near-simultaneous earthquakes, 710
- Multiple prediction through inversion (MPI), 1617
- Multiple scattering, 762–764, 766
- Multi-scale edge detection, 965
- Multi-scale edges, 965
- Multiscale waveform inversion, 1415
- Multisource holography, 1389
- Mushy zone, 165, 166
- Mute, 1605
- N**
- Nankai Trough Seismogenic Zone Experiment (NanTroSEIZE), 121
- Nantou earthquake, 1471
- Nappe complexes, 1259
- National Earthquake Information Center (NEIC), 1520
- National Geophysical Data Center (NGDC), 945
- National seismographic networks, 1648
- National Tsunami Warning Centre (NTWC), 1869, 1870
- Natural gamma ray-log, 532
- Natural gamma spectrometer (NGS), 1306, 1307
- Natural hazards. *See* Tsunami
- Natural prolongation, 861, 862, 864, 865
- Natural remanent magnetization (NRM), 971, 1187, 1195, 1204, 1206, 1320
- Natural seismicity, 12
- Naudy method, 967–969
- Navier's equation, 1285
- Nazca, 1252
plate, 559
- Near-Earth geomagnetic field, 505
- Near-fault ground motions, 323
- Near fields, 1806
- Near-field seismology, 1680
- Near-surface first-arrival tomography, 1841
- Near-vertical reflection methods, 1828
- Neftegorsk earthquake, 1467
- Neighborhood algorithm (NA), 810
- Neo-deterministic seismic hazard assessment (NDSHA), 1788
- NET-VISA approach, 1670
- Network adjustment (leveling), 958
- Network training, 801
- Neumann's problem, 476, 1716
- Neural networks (NN), 800
classification, 797
computational model, 798
generalization, 797
geophysical inversion, 797, 798
pattern recognition, 797
seismics, 802 *see also* Artificial neural networks (ANNs)
- Neurons, 798
- Neutrinos, 1308
- Neutron–neutron log, 533
- New geophysics, 1694
- New Madrid seismicity, 299
- New manual of seismological observatory practice, 1447
- Newton, Isaac, 160
- Newtonian gravitational field theory, 692
- Newtonian potential, 476
- Newton-Krylov-Schwarz methods, 1135
- Newton's integral, 483
- Newton's law of gravitation, 609–611
- Newton's second law, 1283, 1609, 1610
- NGDC-720 lithospheric magnetic model, 952
- Nias earthquake, 1883
- Ninetyeast Ridge (NER), 1088, 1092
- N-MORB, 901
- NMRS recordings, 132
- Noise cross-correlation methodology, 1742
- Nojima Fault Projects, Japan, 121
- Non-accretionary subduction zones, 1729
- Non-arc basaltic rocks, 366
- Non-dipole field, 1016, 1226
- Non-double-couple, 240
- No-net-rotation (NNR), 436
- Non-geological noise sources, 1371
- Nonhomogeneous Poisson process, 12
- Non linearities, 652
- Nonlinear model assessment, 1839
- Nonlinear Monte Carlo procedure, 1841
- Nonlinear multigrid method, 1152
- Non-linear optimization method, 346, 347
- Non-linear optimization problem, 811
- Non-linear statistical data modeling, 798
- Nonlinear transformation, 797
- Non-Newtonian rheology, 1108
- Nonparametric modeling method, 147
- Non-polarizable electrodes, 778
- Non-polarizing electrodes, 350, 783
- Non-rifted ridges, 903
- Non-saturating magnitude scale, 249–251
- Non-seismic geophysical techniques, 57
- Nonshearing components, 12
- Non-tectonic deformation, 272
- Non-tidal gravity changes, 628
- Normal Earth body, 643
- Normal equation system, 652
- Normal gravity, 643, 670, 671
vector, 477
- Normal height, 479
system, 750
- Normalized source strength (NSS), 964–965
- Normal-mode theory, 412, 420
- Normal potential, 477
- Normal stress, 9
- North America, 58
- North Atlantic Magnetic Anomaly Group (NAMAG), 939
- Nuclear magnetic resonance logging, 535
- Nuclear measurement-while-drilling (MWD)
tools, 528
- Nuclear tests, 1672
- Nuclear test, yield estimate of, 1671
- Nucleation, 26–28
- Numerical integration, 650
- Numerical quadrature (NQ) technique, 1718–1719
- Numerical thermo-mechanical models, 890
- Nusselt number, 906, 907, 1062–1064
- NuBelt number, 1760
- Nutation, 149, 152, 1906
- N-wave, 1806
- Nyquist degree, 1721
- Nyquist sampling criterion, 1368
- O**
- Oaxaca earthquake, 1465
- Objective function, 808, 810, 811, 813
- Oblique convergence, 264
- Observational seismology, 1674
- Observatories, 501
- Observed Bouguer gravity, 596, 597
- OBS. *See* Ocean bottom seismometers (OBS)
- Ocean bottom cables, 1156
- Ocean bottom seismics
advantages, 1155
applications, 1157–1160
definition, 1155
future developments, 1160–1161
hydrophones, 1155
nodes, 1157
passive-source seismology, 1160
P-to-S converted waves, 1158–1159
P-waves, 1157–1158
sources, 1157
specification of, 1156
wide-aperture studies, 1159
- Ocean bottom seismometers (OBS), 102, 1502, 1835
experiment, 813
- Ocean Drilling Program, Deep Sea Drilling Project, 738, 741, 747
- Ocean dynamic topography, 647
- Ocean-floor formation, 1013
- Oceanic and continental subduction, 58
- Oceanic Anoxic Events (OAEs), 1103
- Oceanic bathymetry, 910

- Oceanic crust, 1257
- Oceanic crustal structures
- crustal thickness, 109–110
 - fast spreading center, 101, 103–106
 - fault, 110
 - intermediate spreading center, 102
 - lower crust, 107
 - slow spreading center, 102
 - ultraslow spreading center, 102
 - upper mantle, 108–110
- Oceanic geoid, 481–482
- Oceanic geotherms, 850
- Oceanic heat flow, 362, 363
- Oceanic heat flux, 909
- Oceanic heat loss, 367
- Oceanic Island Basalts (OIBs), 1097
- Oceanic lithosphere, 728, 836, 848, 947, 1162, 1257, 1258, 1354
- Oceanic lithospheric thermal structure, 906
- boundary-layer cooling, 911–912
 - creation, 908
 - depth of lithosphere-asthenosphere boundary, 909–910
 - driving force, 913
 - heat flux, 907
 - influence of plumes, 912
 - plate cooling models, 912
 - seafloor depths, 907
- Oceanic Moho, 107–108
- Oceanic monitoring and tsunami warning system, 1810
- Oceanic ridge, 185
- Oceanic topography, 940
- Ocean-Island Basalts (OIB), 1068
- Ocean levelling method, 752
- Oceanography, 646
- Ocean-plate stratigraphy, 1258, 1259
- Ocean surface topography
- altimetry, 1347
- Ocean tides, 157
- Ocean wave-wave interactions, 1657
- Oil, gas and geothermal reservoir monitoring, 630–631
- Okmok volcano, 569, 570
- Olivine, 1250
- Omega-squared model, 226
- Omori law, 230, 259, 404, 1893
- Omori-Utsu law, 1725, 1727
- One-way wave-equation migration (OWEM), 1574
- On-line monitoring, 391
- Onset of modern style plate tectonics, 1257, 1258
- On-site inspection (OSI), 1431, 1669
- Onsite warning, 284–285
- Open-pit mining, 8
- OpenQuake Engine, 1663, 1664, 1666, 1668
- OpenQuake Platform, 1665, 1667
- Open-source Seismic Wave Propagation Code (OpenSWPC), 1527
- Operational earthquake forecast, 1789
- Operator-weighted grid transfer operators, 1153
- Ophiolite, 901, 1259
- model, 900
- Optical bench, 506
- Optical fibers, 384
- cables, 389
- Optically pumped magnetometers, 504, 1014, 1032
- Optical measurements, 398
- Optical time domain reflectometry (OTDR), 380, 386
- Optimal estimation (OE) technique, 1719
- Optimization, global. *See* Global optimization
- Optimization theory, 797
- Orbital sensing parameters, 1117
- Orbit and satellite clock errors, 553
- Organisation for Economic Co-operation and Development (OECD), 1663
- Oriented microcracks, 61
- Orogenic belt, 1279
- architecture, 1170
 - arc massif, 1173
 - back-arc, 1174
 - Bauplan, 1176
 - building materials, 1185
 - carbonatitic magmatism, 1181
 - classes of, 1167
 - classification, 1178
 - coal, 1185
 - collisional systems, 1175
 - collision-controlled orogens, 1176
 - collision magmatism, 1181
 - collision-related metamorphism, 1183
 - definition, 1167
 - economic deposits in, 1184–1185
 - externides, 1172
 - factories of continental crust, 1170
 - features of, 1167
 - fore-arc, 1172
 - forearc magmatism, 1173
 - geomorphology of, 1183–1184
 - hinterland, 1172
 - hydrocarbons, 1184
 - internides, 1172
 - magmatism in orogens, 1179
 - marginal basins, 1174
 - metals, 1185
 - metamorphism in orogens, 1182
 - metamorphism, 1173
 - neutral arcs, 1174
 - non-continental override type, 1179
 - non-override types, 1176
 - orogenic collages, 1179
 - override, 1176
 - plate boundaries, 1167
 - rift-related alkaline, 1181
 - Shuswap metamorphic complex, 1174
 - Sikhote-Alin Orogen, 1179
 - subducting slabs, 1171
 - subduction, 1176
 - subduction-controlled orogens, 1172–1175, 1177
 - subduction magmatism, 1180
 - subduction-related metamorphism, 1182
 - transpressional orogens, 1176
 - trapped ophiolite, 1173
- Orogenic processes, 1257
- Orogens, 58
- Orogeny, 263
- Orthogonal decomposition, 652
- Orthogonal demagnetization plot, 1195
- Orthogonal gradiometer sensors, 700
- Orthogonal transformation, 828, 829
- Orthometric height (H), 466, 479, 749–752, 754–756
- Oscillations of an SNREI Earth model, 415–417
- Oscillations of the first class. *See* Toroidal oscillations
- Oscillations of the second class. *See* Spheroidal oscillations
- Oslo Graben, 60
- Outer Earth's core, 165
- Outer rise region, 1628
- Overhauser magnetometer, 503, 1031, 1117
- Overlapping subdomains, 1134, 1135
- P**
- Pacific Tsunami Warning Center (PTWC), 1863
- Paired metamorphic belts, 1259
- Palaeointensity, 1262
- absolute, 1189
 - measurement, 1187, 1188
- Palaeomagnetic data, 1260
- Palaeomagnetic poles, 1209–1210
- Palaeoseismology, 235
- Paleocene-Eocene Thermal Maximum (PETM), 1102
- Paleoearthquakes, 1236
- Paleo-heat flow, 734–735
- Paleomagnetic recording, 487
- Paleomagnetic techniques, 1319
- Paleomagnetism, 1202, 1203, 1277–1278, 1322
- apparent polar wander path, 1231
 - definition, 1225
 - environmental magnetism, 1234–1235
 - geocentric axial dipole, 1225–1227
 - and magnetic field studies, 1316–1317
 - magnetic polarity reversals, 1231–1234
 - reliability criteria for paleomagnetic data, 1232
 - remanent magnetization, 1228–1230
 - rock magnetism, 1227–1228
 - seventy years of, 1213–1214
 - statistical tests, 1230
 - virtual geomagnetic pole, 1230
- Paleoreconstructions, 1278
- continental separation in ocean-floor, 1278
 - continental shape conservation, 1278
 - paleo-plate reconstruction, 1279
- Paleozoic and Precambrian composites, 1278–1279
- Paleosecular variations, 1226
- Paleoseismology
- definition, 1235
 - evidence for, 1238–1239
 - fault identification and slip rate, 1236–1238
 - hazard assessment, 1245
 - history, 1236
 - issues, 1245–1246
 - magnitude and frequency, 1241–1242
 - off-fault, 1242–1245
 - records, 1236

- Paleoseismology (*cont.*)
 time constraints of, 1239–1241
 PALSAR, 1341
 P and T axes directions, 11
 Pangea, 63
 Parallel computations, 1133, 1134
 Parallelism, 1272
 Parallelization, 1596
 Paramagnetic minerals, 926
 Parameter resolution matrix, 816
 Parameter-selective algorithm, 1839
 Parametric inversion, 967
 Parametrised plate-force models, 1270, 1272
 Parkfield Interventional Earthquake Fieldwork (PIEQF), 223
 Partial anhysteretic remanence (pARM), 933
 Partial differential equation (PDE), 1149, 1150, 1714
 forms of, 1592
 for seismic wave propagation, 1592
 Partially melted asthenosphere, 893
 Partial melting, 1461
 Partial thermal demagnetization, 15
 Partial thermal remanences (PTRMs), 1204
 Particle swarm optimization (PSO), 810–811
 Passive array experiments, 867
 Passive image interferometry, 1548
 Passive rifting, 75, 1356
 Passive seismic, 1444
 Passive seismic tomography, 1557
 Passive-source seismology, 1160
 “Path average” approximation (PAVA), 1745
 Pattern recognition, 797, 798
 PDF. *See* Probability density function (PDF)
 Peak ground acceleration (PGA), 296, 322, 1397, 1403, 1405, 1664
 contour ShakeMap, 318
 Peak-rings, 769
 basin, 771
 Péclelet number, 1760
 Peierls super-plasticity, 888
 Penalty-finite element model, 1147
 Penalty method, 1132
 Pencil-and-paper migrations, 1572
 Pencil structure, 932
 Pendulum, 633
 Penrose model, 900
 Perceptron, 798
 Peregrinus, Petrus, 516
 Perfectly matched layers (PMLs), 1534
 Peridotite, 903, 1180, 1306
 Permanent scatterer InSAR, 1341
 Permanent water level drop, 1494
 Permeability, 1126
 Perturbation method, 1388
 Phase closure, 1904
 Phase delay, 1804
 Phase measurements, 1804
 Phase tensor, 351
 Phase transition, 176, 177, 1766
 Phase velocity, 1474–1476, 1478
 Phenomenological models, 510
 Philippine Sea plate, 1275
 Photogrammetry, 233
 Physical factors, seismic properties, 60
 Physical geodesy, 1717
 Pilot Borehole, 28
 Pixel offset technique, 1338
 Pizzetti-Stokes integral, 614
 Planar deformation features, 772
 Planck’s distribution law, 1775
 Plane waves, 759, 760, 1285, 1286
 characteristics, 1593
 layers of, 1593
 and point source synthetic seismograms, 1594
 reflection coefficient, 1594
 simple synthetics, 1593–1594
 Plasma, 1803, 1810
 frequency, 1804
 medium, 1804
 Plate boundaries, 1248
 Plate cooling models, 912
 Plate driving forces
 boundary-layer model, 1268
 buoyancy-dissipation balance, 1268–1270
 lithospheric models, 1270–1271
 rheological weakening mechanism, 1268
 strain rate, 1267
 viscosity, 1267
 whole-mantle models, 1271–1272
 Plate margins, 1249
 Plates
 compositions, sizes and shapes, 1275
 convergent motions, 1276
 divergent motions, 1276
 earthquakes, 1275
 motion and Euler’s theorem, 1275–1276
 as reference frames, 1277
 sliding motion, 1276
 Plate tectonics, 54, 168, 179, 788, 913, 943, 1248–1250, 1267, 1276, 1626
 generation of, 1072–1076
 (*see also* Precambrian)
 paradigm, 939
 reorganisation, 1122
 theory, 559, 1327
 Platinum resistance temperature detectors, 723
 Plume–lithosphere interaction, 1103
 Plume–ridge interaction, 1104
 Plumes
 in mantle, 1063–1064
 viscosity, 1104
 Plutonic rocks, 931
 POGO satellite series, 1118
 Point masses, 642
 Point source models
 elastodynamics, 225
 inelastic stresses, 226
 moment tensor, 225, 226
 orthogonal linear dipoles, 225
 scalar seismic moment, 226
 seismic radiation, 226, 227
 Poisson differential equation, 641
 Poisson distribution, 12
 Poisson’s field equation, 693
 Poisson’s integral form, 1716
 Poisson’s ratio (σ), 1459, 1462
 Poisson’s relationship, 961
 Polar electrojets, 1119
 Polar ice sheet, 1816
 Polarization analysis, 1914
 Polarization vector, 1474–1476, 1483
 Polar motion, 149, 150, 629, 1906
 Polar Orbiting Geophysical Observatory (POGO), 1116
 Polar sea ice
 altimetry, 1348
 Pole-dipole array, 347
 Pole-pole array, 347
 Pole reduction method, 940
 Polygon/polyhedron modeling of gravity anomalies, 666
 Po phases, 899
 Pore pressure, 9, 271
 diffusion, 23
 Poroelasticity
 definition, 1280
 drained conditions, 1280
 fluid substitution, 1281
 Gassman’s equation, 1281
 linear stress–strain relation, 1280–1281
 Skempton coefficient, 1281–1282
 undrained deformation, 1280
 wave equation, 1282
 Poroelastic stresses, 12
 Porosity, 782, 1461
 Porphyry copper mineralization, 1332
 Portable short-base tiltmeters, 449
 Positioning, 427
 Positive hole theory, 192
 Positivity constraint, 211
 Post-depositional remanent magnetism (PDRM), 1204
 Posteriori probability density function, 808
 Post-glacial rebound, 635, 647
 timescales, 885
 Post-perovskite’ structure, 177
 Post-seismic crustal deformation, 259
 Post-seismic deformations, 566–568
 Post-seismic relaxation, 892
 data, 886
 Post-seismic slip events, 255
 Post-seismic timescales, 885
 Poststack migration, 1569, 1573
 Potassium radioactivity, 1294–1295
 Potential energy, 641
 Potential field inversion, 802
 Potential field transformations, 592–593
 Potential gravitational energy, 1304
 Potsdam gravity value, 790
 Power-law, 400, 404, 406
 scaling, 405
 Precambrian
 cratons, 57
 definition, 1256
 evolution of Earth’s geodynamic regime, 1260–1263
 geodynamic processes, 1258
 juvenile and recycled crustal components, 1257
 onset of plate tectonics, 1260–1263
 plate and plume tectonics, 1258
 plate boundaries, 1257
 plate tectonics, 1258–1260
 supercontinents, 1257
 Precession, 149, 152, 1014, 1906
 Precipitation, 623

- Precision, 633
code, 548
- Preconditioned iterative scheme, 1135
- Preconditioner, 1135
- Precursory scale increase, 1725
- Precursory seismic quiescence (PSQ), 1725
- Predominant method for earthquake location, 301
- Preliminary earth reference model (PREM), 897
- Preliminary reference earth model (PREM), 162, 163, 169, 184, 417–420, 1080, 1515
- PREM. *See* Preliminary reference Earth model (PREM)
- Pre-observational mean (POM), 542, 545
- Preparedness, 712
- Preprocessing, 1617–1618
- Pre-seismic crustal deformation, 256–258
- Pre-seismic ionospheric anomalies, 1809, 1810
- Pre-seismic signals, 1494–1495
- Pre-seismic slips, 1709
- Pre-seismic TEC anomalies, 1809
- Pressure stimulated current (PSC), 200
- Prestack depth migration, 1571
- Prestack migration, 1569
- Principal component analysis (PCA), 1209
- Principal susceptibilities, 926
- Probabilistic seismic hazard analysis (PSHA), 1394–1397, 1405, 1406, 1551
aleatory uncertainty, 1403
deaggregation of seismic hazard, 1400
epistemic uncertainty, 1403
frequency-magnitude models, 1402–1403
numerical computation, 1399–1400
relevant parameters in, 1396
seismic source parameters, estimation of, 1397–1399
source-free procedures, 1400–1401
spatio-temporal models, 1402
time-dependent models, 1401–1402
- Probabilistic seismic hazard assessment (PSHA), 1788
- Probability density function (PDF), 823, 1396, 1399, 1402
- Production logging, 390
- Prolongation, 862–863
- Prompt assessment of global earthquakes for response (PAGER), 320
definition, 308
development, 310
earthquake impact scale, 310–311
impact scale, 309
process, 308
- Propagated error estimates of geoid undulations, 689
- Propagated noise, 1719
- Propagation speed of CIDs, 1806–1807
- Propagation velocity, 1373
- Proto-crust, 1262
- Proto-Earth, 774
- Proton conduction, 340
- Proton magnetometers, 1031
- Proton precession magnetometer, 1014
- Pseudo-gravity transform, 961
- Pseudo-orthogonality condition, 1611
- Pseudo range, 551
- Pseudosection, 346, 347, 351
- Pseudo-single domain (PSD), 973
grains, 1194
- Pseudospectral (PS) methods, 1525, 1529
- PSHA. *See* Probabilistic seismic hazard analysis (PSHA)
- Pulsed neutron logging, 535
- Pulse measurements, 1516
- Pulse transmission technique, 1460, 1461
- P-wave, 37, 41–42, 1475–1478, 1484, 1485, 1607
based warning, 284
energy, 368
first-motion, 237–238
polarization, 45
response, 1489
traveltimes, 1829
- P-wave receiver functions (PRFs), 1391
- Pyroclastic flows, 331
- Pyrolite, 1305, 1391
- Q**
 Q_c , 764
 Q_b , 764
 Q_{scs} , 764
Q Flex QA3000, 580
 Q_{scs} , 764
Qualitative interpretation, 939, 940
Quality factor, 758, 759, 1484
Quantitative interpretation, 940–941
Quantum magnetometers, 983, 1029
Quantum vector magnetometers, 1035
Quartzite, 1250
Quasars, 1905
Quasi-distributed sensor, 386
Quasidynamic circular crack, 228
Quasi-geoid, 466, 479
Quasiperiodic ionosphere disturbances, 1806
Queen Charlotte Transform Fault, 1177
- R**
RA12, 86
Radar altimetry, 836
Radial, 412, 415, 416
eigenfunction, 413
mode number, 417
Radial basis functions (RBF), 697, 1887
Radiated seismic energy (*ES*), 288
Radiating boundary conditions, 1143
Radiation self-absorption, 1307
Radiation theory, 1775
Radiative heat transport, 896
Radiative transfer approach, 1543–1544
Radiative transfer theory (RTT), 761, 763
Radiative transport modeling, 1516
Radioactivity logs, 532
Radio astronomy, 1902
Radio detection and ranging (RADAR), 1332
Radiogenic heat flow, 852
Radiogenic heat production, 363, 1258
Radiogenic heat production, continental crust
crustal differentiation, 1299–1300
crustal temperature and strength, 1300–1302
crustal thickness and heat production, 1298–1299
Radiogenic heat production of rocks, 1304–1305
bulk heat generation rate, 1306
geoneutrinos, 1308–1309
measuring techniques, 1307–1308
tabulated data, 1305–1306
total heat generation rate, 1306
Radiometric dating
applications, 4–7
definition, 1
principle of nuclear transmutation, 1
radioactive elements, 2
Radiometric methods, 233
for age dating, 735
Radionuclide, 1435, 1438, 1440
Radio telescopes, 1905
Radon, 194–197
transform analysis, 1410
Raman scattering, 386
Random error, 389
Random noise, 1443, 1445
Random undersampling, 1383
Random walk, 400
Range, altimeter range, 1343
Rank-deficient matrix, 828
Rao–Mitra inverse (RMI), 817–818
Rapid-response ground-motion maps, 317
Ray-based methods, 1592
Ray-centered coordinates, 1478–1480
Ray coordinates, 1477
Rayleigh–Bénard convection, 1059
Rayleigh–Bénard layer, 1059
Rayleigh number, 170, 907, 1099
Rayleigh scattering, 386, 1532
Rayleigh’s criterion, 45
Rayleigh surface waves, 1502
Rayleigh–Taylor (R-T) instability, 1137, 1810
Rayleigh waves, 897, 1449, 1739, 1806, 1807
radiation pattern, 1808
trains, 1739
Ray parameters, 368, 1477, 1483
Ray perturbation methods, 1481, 1483, 1485
Ray propagator matrix, 1480, 1483
Ray segments, 1499
Ray theory, 1507
based traveltime, 1533
Green function, 1480, 1483
Ray tracing, 1475–1476, 1479–1481, 1483, 1484, 1830
asymptotic methods, 1525
Ray tube, 1477–1478, 1483
Ray velocity, 1476, 1477, 1482
Real-time tsunami modelling, 1875
Receiver clock errors, 552
Receiver function
analysis, 870
common conversion point stacking, 1583–1584
deconvolution, 1581–1582
definition, 1580
examples, 1585–1588
method, 57

- Receiver function (*cont.*)
 moveout correction and summation, 1582–1583
 separation of P and S waves, 1580–1581
 technique, 1580
 waveform modeling, 1584–1585
- Reconnaissance gravity survey, 657, 660
- Redatum, 1603
- Reduced (order) integration, 1148
- Reduction to pole (RTP) transform, 961–962
- Reference Coordinate System, 425
- Reference ellipsoid, 643, 750
- Reference models, 501
- Reference system, 650
- Reflected wavefield amplitude, 1573
- Reflection, 40–41, 902
 coefficients, 1899
 and transmission coefficients, 369
 matrix method, 1592
- Reflectivity method
 applications of, 1597–1600
 compound matrix approach, 1595
 definition, 1592
 extension of, 1597
 methods for computing, 1593
 plane wave, 1594
 theory of, 1594–1596
- Reflectivity model, 1901
- Refraction, 39–40, 1603
 wide-angle reflection methods, 1828
- Regional field modelling, 1020–1022
- Regional gravimetric models (RGMs), 680
- Regional gravity anomaly, 596
- Regional isostatic compensation, 886
- Regional models, 186
- Regional monitoring, 1430
- Regional-residual separation, 597, 964
- Regional seismic signals, 1433
- Regional seismic waves, 1431, 1435
- Regional Tsunami, 1869
- Regional variations, 1520, 1521
- Regional warning, 284
- Regression M-estimate (RME), 1037–1041
- Regularization, 346, 1832, 1833
- Reid's elastic rebound hypothesis, 560
- Relative gravimeters
 functional diagram, 581
 (*see also* Gravimeters)
 sensor technology, 579–580
 types, 579
- Relative IEC (IEC_R), 1804
- Relatively low frequency, 376
- Relative palaeointensities, 1187
- Relative positioning, 427
- Relative satellite positioning, 547
- Relative sea level (RSL), 1109
- Relaxation spectrum, 1515
- Remagnetisation, circle of, 1208
- Remanent magnetization, 1228–1230
 aeromagnetic surveys, 1323
 chemical, 1321
 chemical remanent magnetization, 1228
 definition, 1319
 depositional remanent magnetization, 1228
 detrital, 1320–1321
 hematite-ilmenite series, 1320
 hysteresis behavior, 1319
 isothermal, 1322
 isothermal remanent magnetization, 1229
 magnetic anomalies, 1323
 magnetic minerals, 1319
 natural, 1320
 natural remanent magnetization, 1229
 paleomagnetism, 1322
 permanent magnetization, 1319
 rock magnetism, 1323
 thermoremanent, 1320
 thermoremanent magnetization, 1228
 viscous, 1322
- Remote reference methods, 351, 459
- Remote sensing, 803, 1325–1327, 1343
 applications to geophysics, 1331–1332
 definition, 1330
 images, 1671
- Remove-compute-restore
 technique, 470–471
- Renewal process, 1725
- Repeatability, 389
- Repeating earthquakes, 1670
- Repeat stations, 501
- Representative elementary volume (REV), 1280
- Resampling, 1889
- Research infrastructures, 14
- Reservoir, 1487
 characterization, 1487
 depletion, 11
 geophysics, 802
 properties, 1487
- Reservoir triggered seismicity (RTS), 271
 artificial water reservoirs, 19
 categories, 20
 characteristics, 20, 22
 factors, 20
- Residual gravity anomaly, 596
- Residual mantle anomalies, 587
- Residual statics, 1377
- Residual Stokes kernel, 753–755
- Residual terrain model (RTM) reduction, 481
- Resistivity, 802
 apparent, 776
 automated monitoring systems, 780
 basic instruments, 776
 log, 532
 survey method, 776
- Resolution, 389, 760
 matrices, 829
- Resonance techniques, 1461
- Response spectrum, 322
- Restarted L-BFGS algorithm, 1621
 crosstalk effect, 1613
 DFP, 1615
 features, 1614
 gradient difference, 1614
 image quality, 1615
 implementation, 1614
 quasi-Newton method, 1614
 secant equation, 1615
 velocity slices, 1616, 1617
 wavefield simulations, 1615
- Restriction operator, 1135, 1136
- Reversal test, 1195, 1229
- Reverse time migration (RTM), 1408, 1576, 1792
- Reversible jump Monte Carlo method, 811
- Reynolds number, 80
- Rheologies, 1249
- Rhyolitic melt, 352
- Richter scale, 243, 244
- Ridge internal structure, 1162
- Ridge-push, 1258
 force, 1271
- Ridges, 266
- Rift-dominated processes, 904
- Rifted ridges, 903
- Rifts, 64, 65, 851
 dynamic (genetic) classification of, 75
 geometric classification of, 72–73
 graben and taphrogen, 66
 kinematic classification of, 73–75
 magmatism, 70–71
 metamorphism, 71
 rift chains, 72
 rift clusters, 72
 rift nets, 72
 rift stars, 72
 sedimentation in, 69–70
 solitary, 72
 structure of, 66–69
 valley, 1163
- Rigidity moduli, 1607
- Ring lasers, 1676
- Risk Input Preparation Toolkit, 1666
- Risk Modeller's Toolkit (RMTK), 1666
- River discharge, 1817
- Robust methods, 350
- Rockburst, 8
- Rocket launching, 1806
- Rock-fall signals, 331
- Rocks
 average velocities, 1462
 compressional wave velocities, 1460
 density, 848
 fabric, 61
 high pressures and temperatures, 1460
 magnetism, 923, 937, 1227–1228
 mass, 9
 measurement techniques, 1460–1462
 parameters, 1890
 seismic velocities, 1462–1463
 strain, 931
 types, 1769
 velocity anisotropy, 1463, 1464
- Root mean square (RMS), 833, 834, 839, 844
- Rossby number, 81
- Rotary drilling technique, 116
- Rotated computational star, 1610
- Rotational seismology
 definition, 1673
 explosions to study rotational motions, 1681
 G Ring laser and recording teleseisms, 1676
 large ring laser gyroscopes, 1675–1676
 linear and nonlinear elasticity, 1679
 near-field seismology, 1680
 processing collocated observations of rotations and translations, 1681–1682

- rotational motion measurement, 1675
- strong-motion inertial angular sensors, 1676–1678
- Rotation of the Earth, 413, 417, 420
- Rotatory vibrations, 414
- Rupture propagation, 212–213
- Rytov phase approximation, 1387
- Rytov representation, 1540
- S**
- Saastamoinen model, 554
- Sagnac interferometry, 1675
- Saguenay earthquake sequence, 190
- Sampling error, 1719
- San Andreas, 710
- San Andreas Fault, 1052
- San Andreas Fault Observatory at Depth (SAFOD), 48–50, 121, 304
- San Andreas Fault Observatory at Depth borehole, 121
- Sander Geophysics system, 583
- Sand–shale interfaces, 1159
- San Francisco earthquake, 709, 710
- Satélite de Aplicaciones Científicas-C (SAC-C), 1117
- Satellite(s), 501, 788
 - communication, 396
 - data acquisition, 1118
 - measurements, 938, 1118–1119
 - missions, 943
 - monitoring, 1343
 - radar altimeter, 1816
- Satellite altimetry, 481–482, 486, 648
 - applications, 474–475
 - measures, 1343
- Satellite-based remote sensing, 1326
- Satellite-borne gravity gradiometer, 704
- Satellite gravity gradiometry (SGG), 654, 1722
- Satellite orbit(s), 646
 - perturbation analysis, 647
 - tracking, 158
- Satellite positioning
 - point positioning, 547
 - relative satellite positioning, 547
- Satellite-to-satellite tracking (SST), 654, 1722
- Sawtooth hypothesis, 509
- SBP finite-difference method, 1611
- Scalability, 1135
- Scalar amplitude, 1474, 1478, 1481
- Scalar gravity meters, 638–639
- Scalar magnetometers, 1030, 1116
 - absolute accuracy and precision, 1033
 - optically pumped, 1032
 - overhauser, 1031
 - precession frequency measurement, 1032–1033
 - proton, 1031
- Scalar measurements, 503
- Scale-affine, 399
- Scaled identity matrix, 1019
- Scale-invariance, 403, 404
- Scaling, 399, 400
- Scaling law, 83–84, 406
- ScanSAR, 1335
- Scattered wave techniques, 1588
- Scattering, 757, 760, 761, 766
 - attenuation, 1513
 - coefficient, 762
 - equivalent medium, 1515
 - and intrinsic absorption, 766
 - and intrinsic attenuation, 761
 - multiple, 762–764, 766
 - of seismic waves, 760, 761
 - single, 764, 766
 - SIS, 762
 - stochastic dispersion, 1515, 1516
 - theory, 1745
- Scattering of seismic waves
 - basic concepts, 1537, 1538
 - coherent and incoherent fields, 1538, 1539
 - defined, 1537
 - global scale scattering, 1544–1545
 - Green function retrieval, 1545–1546
 - Markov approximation, 1541–1543
 - propagation regimens, 1539
 - radiative transfer, 1543–1544
 - transmission fluctuation analysis, 1539–1540
- Schmidt seminormalized Associated Legendre functions, 1714
- Science and math education, 788
- Science and technology foresight, 1672
- Scientific drilling, 855, 859, 860
 - definition, 115
 - goals of, 115–116
- Scintrex CG-3 gravity meter, 581
- Sea floor
 - age distribution, 363, 364
 - depth, 894
 - heat loss, 364
 - topography, altimetry, 1344
- Sea-floor spreading, 66, 848, 893, 1122
 - axes, 1352
 - definition, 1349
 - revolutionary, 1349
 - process, 1232
 - step in, 1349
- Sea level, 1343
 - change, 1816
 - variation, 628
- Sea Level Index Points, 1112, 1113
- Sea surface height, 755, 1343
 - anomalies, 1816
- Sea surface topography (SST), 481
- Secular cooling, 365
- Secular variation, 501, 1226
 - geomagnetic, 514
- Sedimentary basins, 1797
 - classification, 1354
 - conceptual models of basin formation, 1356
 - crust beneath, 1358–1360
 - definition, 1353
 - heat flow, 1361
 - mantle lithosphere, 1360
 - mechanisms, 1356
 - sediment fill, 1358
 - types, 1354
- Sedimentation-compaction plane, 932
- Seismic acquisition techniques, 1445
- Seismic activity, 1893, 1894, 1896
- Seismic amplitude calibration, 1487
- Seismic and petrological Moho, 57
- Seismic anisotropy, 61, 173, 1899
- Seismic body waves
 - attenuation, 42
 - modeling, 43
 - P and S waves, 37–39
 - propagation, 39
 - regional and global, 42
 - seismogram, 39
 - velocity, 41–42
- Seismic coupling, 1626
- Seismic data, 1562–1565, 1567
- Seismic data acquisition
 - acquisition geometry, 1368–1370
 - aims, 1366
 - definition, 1365
 - seismic receivers, 1367–1368
 - seismic recorders, 1368
 - seismic sources, 1367
- Seismic data processing
 - absorption, 1373
 - aims, 1366
 - amplitude changes along propagation path, 1373
 - convolution and deconvolution, 1374–1377
 - definition, 1365
 - energy partitioning at interfaces, 1374
 - geometrical spreading, 1373
 - kinematics of the seismic signal, 1371–1373
 - normal moveout, 1371–1372
 - no-stack processing, 1381
 - post-stack processing, 1380–1381
 - pre-processing, 1377
 - pre-stack processing, 1377–1380
 - processing flow, 1377–1382
 - recent developments, 1382–1384
 - signal vs. noise, 1371
 - special processing, 1382
 - velocity, 1373
- Seismic diffraction. *See* Diffraction
- Seismic disaster risk, 1789
- Seismic event, 8
- Seismic gap model, 54
 - assumptions, 54
 - challenges to, 55
 - modified, 54–55
- Seismic hazard, 1661, 1787–1789
 - analysis, 235
 - assessment, 1893
 - curve, 1396, 1397, 1405
 - definition, 1394
 - DSHA, 1394
 - GMM, 1403
 - PSHA (*see* Probabilistic seismic hazard analysis (PSHA))
- Seismic hum, 1655, 1656. *See also* Continuous free oscillations of the Earth
- Seismic imaging, 1567
 - definition, 1407
 - FWI, 1411
 - inversion theory, 1407–1410
 - LSM, 1410–1411
 - migration, 1410
 - multiscale waveform inversion, 1415

- Seismic imaging, 1567 (*cont.*)
 MVA, 1411–1415
 phase-encoded multisource waveform
 inversion, 1415–1417
 wave equation travelttime inversion, 1411
- Seismic instruments, 1419
 requirements for, 1420
 seismometers (*see* Seismometers)
- Seismic intensities, 318
- Seismic interferometry, 1792
- Seismic inversion method, 1608
- Seismicity modulation, 266
- Seismicity rate, 11
- Seismic lightning, 190
- Seismic microzonation
 definition, 1425
 dense borehole data and GIS, 1426
 geomorphological classification
 maps, 1426
 microtremor measurements, 1426
- Seismic migration, 1567, 1609
 algorithm, 1899
- Seismic modeling, 1592
- Seismic moment, 224
- Seismic monitoring, 1669
- Seismic monitoring, of nuclear
 explosions, 1430
 association, 1432–1433
 basic properties of earthquake and
 explosion signals, 1431
 cavity decoupling, 1438, 1439
 detection, 1431, 1432
 event detection capability, 1439–1440
 location, 1433
 methods of identification, 1433–1435
 mine masking, 1438, 1439
 special events, 1437–1438
 yield estimation, 1435–1437
- Seismic monitoring system
 evaluation and design of, 1669–1670
 functional perspective, 1670
 operational perspective, 1670
 physical perspective, 1670
 scientific perspective, 1670
- Seismic noise, 1442, 1489
 example, 1444
 signal enhancement over noise, 1443–1446
 types, 1443–1444
 use of noise as signal, 1446
- Seismic observations, 893
- Seismic phases names, 1447
- Seismic P waves, 396
- Seismic quiescence and activation, 1465
 long-term activation, 1465
 physical mechanisms, 1468–1471
 RTL method, 1466, 1467
 Z-value method, 1466
- Seismic radiation, 225
- Seismic reflection, 1163, 1165
 method, 102
 of deep crust, 1260
 waveform tomography, 1618–1621
- Seismic refraction method, 102, 1164
- Seismic Research Observatory, 48
- Seismic reservoir characterization
 data processing, 1488, 1489
 definition, 1487
 hydrocarbons, 1487
 ingredients, 1492
 inversion, 1489
 ML, 1490
 non-invasive method, 1487
 porosity, 1491
 quantitative interpretation, 1488
 rock physics, 1489
 wave propagation, 1488
 well-tie, 1489
 workflow, 1488, 1490
- Seismic risk, 1662, 1663
- Seismic seiche
 amplitude, 1493
 Assam earthquake, 1492
 definition, 1492
- Seismic signals in well water observations
 consolidation hypothesis, 1495
 coseismic and post seismic signals,
 1493–1494
 definition, 1493
 earthquake enhanced permeability
 mechanism, 1495
 fault dislocation and static poroelastic strain
 mechanism, 1495
 mechanisms, 1495–1496
 preseismic signals, 1494–1495
 transient signals, 1494
- Seismic 3D tomography, 170–171
- Seismic slip, 8
- Seismic sources, 1394–1402, 1404, 1406
- Seismic spectrum, 225
- Seismic tomography, 176, 1220, 1258, 1271,
 1571, 1795
 arrival time, 1507–1508
 definition, 1507
 eigenfrequency, 1508
 finite-frequency tomography, 1509–1510
 history, 1507
 inversion, 1508
 parameterization, 1508
 raypath, 1507
 surface wave, 1509
- Seismic velocity(ies), 772, 900
 definition, 1511, 1607
 vs. density, 1607
 estimation of, 1608
 model, 306, 1080–1081
 types of, 1607
- Seismic viscoelastic attenuation
 amplitude, 1513
 anisotropy, 1516
 geometric spreading, 1513
 ground motion, 1513
 interpretation measurements, 1519–1523
 intrinsic, 1513
 linear, 1513–1515
 measurement and modeling, 1516–1519
 scattering, 1513, 1515, 1516
- Seismic wave(s), 8, 1448–1449
 acoustic phases, 1455
 amplitude measurement, 1456
 core phases, 1454, 1456, 1457
 crustal phases, 1451, 1453–1454
 lowercase letters and signs, 1449
 mantle phases, 1452–1454
 ray path diagrams, 1451
 reflected waves, 1450
 refracted and converted refracted
 wave, 1450
 scattering (*see* Scattering of seismic waves)
 surface waves, 1455
 types of, 1607
- Seismic wave propagation
 asymptotic ray tracing methods, 1533
 boundaries, 1533
 classification, 1526
 definition, 1525
 deterministic approaches, 1525
 DG method, 1529
 direct methods, 1525–1528
 equivalent media theories, 1532, 1533
 FE and SEM method, 1529
 inhomogeneous elastic media, 1525
 integral equation methods, 1530
 non-homogeneous media, 1531, 1532
 poroelastic and thermoelastic media, 1533
 PS method, 1529
- Seismic zonation map
 of China, 1551–1554
 definition, 1551
 probabilistic method, 1551
 seismic hazard analysis, 1551
- Seismogenic permeability, 24
- Seismogenic zone, 1626, 1627, 1629
- Seismogram interpretation
 body waves, 1639–1642
 crustal waves, 1637–1639
 surface waves, 1642–1643
 volcanic earthquakes and unusual seismic
 sources, 1643
- Seismograms, 211, 226, 289, 397
- Seismo-ionospheric, 1802
 anomalies, 1809
 precursors, 1809
- Seismological networks, 1644
- Seismological observation and
 interpretation, 1671
- Seismology, 405, 646, 1662, 1670
 recordings of, 1644
- Seismometers
 calibration, 1423
 deformation, 1424
 displacement, 1424
 electronic design, 1423
 inertial, 1420–1421
 installation, 1423–1424
 mechanical design, 1422–1423
 seismic data recording, 1424–1425
- Seismosonic symphony, 223
- Self-affine, 400, 404
- Self-affinity, 402
- Selfconsistent model, 511
- Self-demagnetization, 924
- Self-Organizing Feature Map (SOM), 801
- Self potential method, 536
- Self-potential (SP) voltage, 778
- Self-similar, 400, 405
- Self-similarity, 399, 400, 402, 1894
- Self-sustained dynamo, 78
- Semblance, 1373

- Semi-analytic methods, 1592
 Sensible heat capacity, 1761
 Sensing methods, 384
 Sensitivity analysis, 351
 Sensor cables, 389
 Sensors and probes, marine sediments, 738
 Sentinel satellites, 1685
 and commercial systems, 1686
 financial challenges, 1686
 free and open access to data, 1686
 and Landsat, 1685
 Serpentine(s), 900, 1462, 1785
 Shaanxi earthquake, 709, 712
 ShakeCast
 applications, 314
 Caltran, 312
 definition, 312
 goals of, 314
 ShakeCast[®], 320
 ShakeMap, 316
 intensity, 317
 systems and uses, 318–320
 Shaking, 12, 316
 Shallow earthquakes, 1625
 Shallow inland intraplate earthquakes, 1628–1629
 Shallow measurements in marine sediments, 737–738
 Shape anisotropy, 924
 Shatter cones, 772
 Shear and implosive components, 12
 Shear failure, 1249
 Shear heating instability of creep, 1633
 Shear (S) waves, 1459
 Shear vs. bulk attenuation, 1519
 Shear wave, 35, 161, 897, 1580–1581
 singularities, 1464
 sleds, 1157 *see also* S-waves
 Shear-wave splitting (SWS), 61, 1463, 1900
 above small earthquakes, 1694
 and anisotropic symmetry, 1689–1692
 definition, 1687
 features in crust, 1689
 in hydrocarbon reservoirs, 1693
 schematic illustration of, 1687
 sources of, 1692–1694
 in upper mantle, 1693
 Shear wave velocities, 897
 Sheeted dikes, 901
 Shillong Plateau, 268
 Shock acoustic waves (SAW), 1803, 1806, 1807
 Shock wave, 771
 Shortest path ray-tracing method, 1831
 Short-term crustal deformation, 257
 Shot-encoded waveform tomography, 1613
 Shur complement matrix, 1135
 Shuttle radar topographic mission (SRTM), 1326, 1337
 Siderophile, 1295
 Siderophile-element enrichments, 772
 Sierra Nevada, 58
 Signal to noise ratio (SNR), 396, 783
 Sill Cube (SC) meta-attribute, 1564–1565
 Sill-like intrusions, 60
 Silver chloride electrodes, 783
 Simple horizontal layer model, 592
 Simulated annealing (SA), 797, 809–810
 Simushir earthquake, 1468
 Single-channel seismics, 1698
 processing considerations for high-resolution, 1701–1704
 sources, 1699–1701
 Single-domain/two-domain transition size, 973
 Single-domain grains of magnetic mineral, 1194
 Single-fold reflection profile, 1568
 Single isotropic scattering (SIS), 762
 Single-look-complex (SLC) image, 1336
 Single mode fibers, 380
 Single scattering, 764, 766
 Single silicate crystals, 1189
 Singular value decomposition (SVD), 346, 828–830, 1341
 Site amplification, 761
 Site effect, 761
 Site-related seismic response spectrum, 1553
 Skempton coefficient, 1281–1282
 Skin depth, 350, 457
 Skin depth effect, 1044
 Slab(s)
 in mantle, 1063–1064
 suction, 1258
 Slab pull, 1258
 force, 1270
 Slab-resistance force, 1270
 Slant TEC, 1803
 Slip deficit, 561
 Slip roughness, 400
 Sliver, 265
 Slopes, 1498
 Slope tomography, 1498
 applications of, 1500
 challenges and perspectives, 1501
 methods, 1499
 Slow earthquake
 definition, 1705
 earlier studies on, 1706
 low-frequency tremor, 1708–1709
 mechanisms of, 1711–1712
 pre-seismic slips, 1709
 scaling relation, 1710
 Slow slip events, 255, 1706
 GPS, 1706
 long-term, 1706–1708
 short-term, 1708–1709 *see also* Slow earthquake
 Small baseline subset (SBS) algorithm, 1341
 Small polaron, 341
 Small-scale heterogeneities, 766
 Smoothness constraint model, 346
 Snell's law, 368
 Snow avalanches, 1655
 Societal applications, 646
 Socio-economic vulnerability and resilience (SVR), 1663
 Solar activity controls, 1001
 Solar eclipses, 1806
 Solar flares, 524
 Solar parameters, 1009
 Solid-solid transitions, 893
 Solid Spherical Harmonic functions, 1714
 Solid-state glasses, 772
 Solitary rifts, 72
 Some earthquakes, 254
 Somigliana–Pizetti formula, 790
 Somigliana's formula, 478
 Somigliana's representation theorem, 1530
 Sonic log, 533
 Sonification, 222–223
 SOPAR database, 290
 So phases, 899
 Sorgenfrei-Tornquist Zone (STZ), 1819
 Sound fixing and ranging (SOFAR), 1751
 Source, 761
 effect, 761
 radiation, 761
 scaling, 1623
 Source parameter imaging ("SPI")
 method, 967
 Source-sensor separation, 994
 South America, 1252
 South-West Indian Ridge, 102, 1164
 Space domain filters, 652
 Space geodetic techniques, 149
 Space, measurements from, 1115–1116
 Space weather, 524, 915
 Sparker, 1700
 Spatial aliasing, 1596
 Spatial and temporal studies, 1007
 Spatial autocorrelation (SPAC) method, 1428
 Spatial coverage, 1116
 Spatial magnetic variation, 1012
 Spatial noise, 1815
 Spatial resolution, 1815
 Spatio-temporally localized events, 1656–1660
 Specific heat capacity, 1760, 1765–1767
 isobaric and isochoric, 1762–1765
 Spectral analysis of variations, 1057
 Spectral element method (SEM), 1529, 1746
 Spectral modeling
 advantages, 591
 applications, 592
 Spectral models of the gravitational
 gradients, 696
 Sphenochasms, 74
 Spherical Cap Harmonic Analysis, 1020
 Spherical float, 395
 Spherical harmonic(s) (SH), 478, 611–612, 935, 1021, 1022, 1815
 coefficients, 649
 series, 649, 1009
 Spherical harmonic analysis (SHA), 501, 918
 block-diagonal (BD) least squares, 1721–1722
 boundary value problems, 1716–1718
 least squares adjustment, 1718
 numerical quadrature (NQ) technique, 1718–1719
 optimal estimation, 1719
 practical aspects, 1719–1721
 statistics of normal matrices, 1722
 Spherically-symmetric, non-rotating,
 elastic, isotropic (SNREI)
 Earth model, 412, 415, 416, 419, 420
 Spherically symmetric model, 167

- Spheroidal oscillations, 412, 414–418
- Spiking filter, 1376
- Spin-canting, 1320
- Spinel lenses, 1633
- Spinner magnetometers, 1194, 1210
- Spontaneous magnetization, 972
- Spontaneous potential (SP) log, 532
- Spreading center. *See* Mid-ocean ridge
- Spreading ridge system, 1086
- Spring-slider model, 1711
- Sputnik, 788
- Stable continental regions (SCRs), 24
- Stacking, 1341
velocity analysis, 1379
- Stagnant-lid convection, 1260
- Standard gradiometers, 988
- Standard kilogram, 637
- Standard linear solid (SLS), 759
- Standard magnetic data processing
corrections, 958
- Standing field hypothesis, 510
- Star trackers, 506
- Static circular shear crack, 228
- Static Quasi-Monte Carlo (QMC) phase
encoding, 1415
- Static spontaneous potential, 532
- Stationary gravimeters, 579
- Stationary Poisson process, 1725
- Statistical analyses, 11
- Statistical deconvolution, 1376–1377
- Statistical seismology
definition, 1724
empirical relations, 1724
global navigation satellite systems, 1727
multiplets, 1724
precursors, 1725
seismic hazard and risk tools, 1727
swarms, 1724
- Steam injection, 391
- Steep angle seismics, 1584
- Steepest-descent algorithm, 799
- Stellar nucleosynthesis, 2
- Stereotomography, 1498
roots of, 1498
- Stick-and-slip process, 265
- Stokes, 388
BVP, 477–479
coefficients, 622
constants, 612
equation, 480
integral, 471–473
integral operator, 478, 479
kernel function, 478
theorem, 476
- Stolt stretch factor, 1575
- Strain, 758
dependence, 1521
- Strainmeter(s), 158, 1424
- Strain-rate tensors, 1268
- Stream length gradient index (SL), 1757
- Stress, 758
evolution, 1726
glut, 226
- Stress-forecasting earthquakes, 1695
- Stress-release model, 1725
- Stress-strain hysteresis, 1514
- Stress-strain relation, 1138
- Strike-slip earthquakes, 1804, 1806, 1808
- Strike-slip fault(s), 710, 711, 1237
- Strong-ground motion
definition, 321–323
earthquake source effects on, 323–324
example of, 325–327
modeling, 325
wave-propagation effects, 324–325
- Strong motion earthquake warning.
See Earthquake early warning (EEW)
- Strong-motion inertial angular sensors,
1676–1678
- Structural geology, 1132
- Structural index, 593
- Structural uplift, 772
- Structural imaging, 1571
- Subaerial rifts, 69
- Subduction, 1250
magmatism, 1180
scars, 1260
tectonics, 1257
- Subduction zone(s), 185, 710, 711, 1163, 1257,
1392, 1808
definition, 1625, 1728
devolatilization and metamorphism,
1733–1734
fault drilling at sea, 123
hazards, 1736
interplate earthquakes, 1626–1628
intraslab earthquakes, 1629–1634
kinematics, 1730–1731
melting and volcanic arc, 1734–1735
morphology and nomenclature, 1728
oceanic plates, 1626
outer rise, trench and forearc, 1728–1730
seismic imaging, 1735–1736
shallow inland intraplate earthquakes,
1628–1629
subduction channel, 1734
thermal structure, plate buoyancy mantle
flow, 1731–1733
thrust zone, 1730
trench-outer rise intraplate earthquakes,
1628
- Sub-grid scales, 88
- Sub-ionospheric point (SIP), 1803,
1806, 1810
- Sub-lithospheric heat flow, 852
- Submarine basaltic glass (SBG), 1189
- Submarine elevations, 862, 864, 865
- Submarine gradiometers, 706
- Submarine ridges, 862, 864, 865
- Submerged prolongation, 861
- Sumatra, 711
Sumatra earthquake, 1471, 1494
- Summation-by-parts (SBP), 1611
- Sunda thrust, 711
- Sun-synchronous components, ionospheric
origin, 1117
- Superconducting gravity gradiometer, 703
- Superconducting gravity meter (SG), 626, 628
- Superconducting gravity sensor, 580
- Supercooled quantum interference devices
(SQUID), 1035
magnetometers, 1211
- Super-dense real-time monitoring of
earthquakes (SUPREME) system, 1426
- Superplumes, 1100
- Super-shear wave rupture, 213–215
- Super-shot gather, 1613
- Super-virtual refraction interferometry (SVI),
1602, 1604–1606
advantages, 1604
ISVI, 1606
limitations, 1605
- Supervised learning, 1490
- Supervised training, 799, 801
- Support vector machine (SVM), 1563
classifiers, 667
- Surface air temperature (SAT), 542, 545
- Surface-consistency, 1377
- Surface gravity anomalies, 686
- Surface heat flow, 850–852
- Surface heat flux, 362
- Surface impedance, 1043
- Surface measurements of gravity, 647
- Surface topography, 1830
- Surface traction, 87
- Surface wave, 897, 1455, 1509, 1607, 1636,
1642–1643, 1740
analysis, 49
attenuation, 1747–1748
definition, 1738
dispersion data, 1743
dispersive and attenuative properties, 1738
generation and propagation, 1741
magnitudes, 245–247
source studies using, 1748
structure studies using, 1741–1742
structure using, 1742
studies, upper mantle anisotropy
from, 1747
velocity of propagation, 1740
- Survey(s), 501
design, 995
lines, 993
specifications, 995
- Susceptibility bridges, 928
- S-velocity model(s), 171
of North Atlantic region, 1798
- Swarm, 505
constellation, 919
mission, 1119
- S wave, 37, 42, 1472, 1475–1478, 1481, 1482,
1484, 1485, 1599, 1607
velocity, 1632, 1633
- S-wave receiver functions (SRFs), 1391
- Sweet-spot depth inversion, 970
- Symmetric rifts, 68
- Symmetry, 1463
- Synapses, 798
- Synthetic aperture radar (SAR)
altimetry, 1348
imaging, 1335–1336
interferometry, 1332
- Synthetic seismograms, 239, 1592.
See also Reflectivity method
- Synthetic wavefield, 1900
- Systematic error(s), 389, 635
- Systematic event mislocation, 146
- System performance evaluation, 1670

- T**
- Tabular cavity collapse, 12
- TAICRUST, 813
- Taiwan Chelungpu Drilling Project, Taiwan, 121
- Taiwan Chelungpu Fault Drilling Project (TCDP), 121
- Tangshan, 712
- Tarantola–Valette approach, 665
- Targeted ('sweet-spot') inversion, 970
- Taylor cylinders, 94
- Tectonic(s)
- definition, 1325
 - digital remote sensing data, 1326
 - earthquakes, 12, 254
 - forces, 465, 468
 - fracture trace analysis, 1328
 - geographic information system, 1328–1329
 - global positioning system, 1327
 - melanges, 1259
 - plates, 1276, 1902
 - plate tectonics theory, 1327
 - remote sensing, 1326–1327
 - seismology, 13
 - stress, 1470
 - studies, 1012
 - transitions, 1795
 - tremors, 254
- Tectonic geodesy, 558
- coseismic deformations, 564–566
 - glacial isostatic adjustment and sea level studies, 570–573
 - history of, 559
 - interseismic deformation, 560–563
 - plate motions, 559–560
 - postseismic deformation, 566–568
 - transient fault movements, 563–564
 - volcanic deformation, 568–570
- Tectonic geomorphology
- definition, 1753
 - and paleoseismological studies of tectonic-landforms, 1758–1759
- Tectonic loading, 10, 893
- of preexisting faults, 10
- Tectonic melanges, 1259
- Tectono-thermal events, 868
- Tectosphere model, 868–870
- TECU, 1802, 1806
- Teisseyre–Tornquist Zone (TTZ), 1819
- Teleseismic earthquakes, 1494
- Teleseismic magnitude scale, 245–248
- Teleseismic monitoring, 1430
- Teleseismic recordings, 289
- Teleseismic signals, 1431, 1433, 1434, 1436
- Teleseismic waves, 268, 1431, 1437
- Telluric current, 194, 197–198
- Telluroid, 478, 675
- Temperature, 848, 852, 894
- dependence, 1763–1765
 - log, 533
 - measurement, 384
- Temperature-depletion compensation, 869
- Temperature-depth, 544
- measurements, 541, 544
 - profiles, 541–546
- Temporal aliasing, 1596
- Temporal decorrelation, 1340
- Temporal evolution, 646
- Temporal gravity changes, 627, 630, 631
- Temporal variation, 1012
- Temps Atomique International (TAI), 435
- Tensile components, 12
- Tensor deconvolution, 594
- Terrain correction, 480, 672–673
- Terrain reductions, 479
- Terrestrial gravitation, 608
- Terrestrial heat flow, 722
- Terrestrial intermediate origin (TIO), 151
- Terrestrial laser scanners, 429
- Terrestrial water cycle, 1817
- Terrestrial water storage (TWS), 623
- Terzaghi's effective stress, 1281
- Test site bias, 1436
- The Hadean Eon, 1260–1262
- Thellier's law(s), 1189
- of linearity, 1189
- Thematic Core Service Anthropogenic Hazards (TCS AH), 14
- Thermal activation, 1519, 1520
- Thermal boundary layers (TBLs), 1097
- Thermal capacity, 1764–1766
- Thermal conduction, 872
- Thermal conductivity, 737, 850, 882
- definition, 1769
 - of different rock types, 883
 - Fourier's law of, 1769
 - histograms of, 1775
 - indirect methods, 1770–1771
 - measurements on, 1773
 - measuring techniques, 1770, 1783
 - metamorphic rocks, 1774, 1778
 - of minerals, 1774
 - petrophysical data, 1769
 - physical properties, 1773
 - plutonic rocks, 1774, 1778
 - pore fluid and solid rock, 1772
 - porosity, 1782
 - radiative, 1775–1777
 - rocks, 1774
 - saturation, 1782
 - sedimentary rocks, 1774, 1778
 - variation of, 1772
 - variation with pressure, 1781–1782, 1784–1785
 - variation with temperature, 1777–1784
 - volcanic rocks, 1774, 1778
 - with phonon, 1770
- Thermal convection, 169
- Thermal demagnetization, 1195, 1213
- Thermal diffusion, 541, 542, 546
- Thermal diffusivity κ , 1769
- Thermal diffusivity, 1783
- Thermal energy (EH), 288
- Thermal expansion, 848
- Thermal/free convection, 1059–1061
- Thermal isostasy, 834, 848
- continental, 849–853
 - definition, 847
 - in oceanic regions, 848–849
 - on terrestrial planets, 853
- Thermal rejuvenation, 1103
- Thermal relaxation of lithosphere, 879
- Thermal remanence, 15
- Thermal remanent magnetization (TRM), 974, 1204
- Thermal stimulation, 391
- Thermal storage properties, 1760–1761
- Thermal tracer/dilution techniques, 392
- Thermal viscous remanent magnetization (TVRM), 1322
- Thermal wind circulation, 83
- 10-Thermistor cable, 740
- Thermistors, 723
- Thermochemical plumes, 1100
- Thermochronometers, 6
- Thermoelastic effects, 12
- Thermoremanent magnetization (TRM), 1187, 1320
- Thin-layer approximation (TLA), 97
- Thinned Fault Cube (TFC) meta-attribute, 1563–1564
- Thomson-Haskell method, 1525
- 4th-order ADER scheme, 1528
- Thorium radioactivity, earth's core, 1296
- Three-dimensional (3D) heterogeneities, 1046
- Three-dimensional (3-D) surveys, 347, 348
- Threshold Test Ban Treaty (TTBT), 1435–1437
- Thrusts, 1259
- Tidal effects, 642
- Tidal force, 156–157
- Tidal gravity changes, 628
- Tidal models, altimetry, 1347
- Tides, 628, 647
- Tie-line leveling, 998
- "Tie" lines, 993
- Tikhonov regularization, 1500
- Tikonov–Arsenine formula, 665
- Tilt-coupling effect, 1676
- Tilted axis of symmetry (TTI), 1572
- Tilt filter, 964
- Tiltmeters, 158
- Time and spatial stamping, 1118
- Time-dependent seismic hazard assessment, 1788
- hybrid/empirical approach, 1788
 - neo-deterministic approach, 1788
 - probabilistic approach, 1788
- Time-domain memory functions, 1519
- Time domain surface wave inversions, 1745
- Time-frequency representation (TFR), 1909, 1910
- Time-lapse seismic monitoring, 1901
- Time-lapse seismics, 1368
- Time migration, 1570
- Time-reversal
- applications, 1790
 - and complexity, 1792–1793
 - definition, 1790
 - location of seismic sources, 1791–1792
 - mirror, 1790
 - reciprocal, 1790
 - seismic interferometry, 1792
- Time sequence measurements of flow logs, 535
- Times of increased probability (TIP), 205–207

- Time-space clustering analysis, 11
Time-variable satellite gravity measurements, 1813
Tipper, 351
Titanomagnetite grains, 931
Titanomagnetite series, 112
T-MORB, 901
Tohoku, 711
 earthquake, 568
Tokachi-oki earthquake, 1494, 1676
Tom McEvilly's seismology group, 376
Tomography, 1608, 1828. *See also* Waveform tomography
Tonalite-Trondhjemite-Granodiorite (TTG), 1260
 complexes, 1260
Tonga-Kermadec subduction system, 1177
Tonga system, 1728
TOPO-EUROPE, 1794
 analogue tectonic modeling, 1800
 deep earth, 1795–1797
 lithosphere structure, 1797
 mantle-lithosphere interactions, 1797
 program design, 1795
 sedimentary basins and geo-resources, 1797–1800
Topography, 479–481, 1251
 fringe, 1337
 lineaments, 1236
 masses, 643
 stresses, 1132
 torques, 88, 89
Tornquist Fan, 1821
Toroidal oscillations, 412–418
Torsional wave, 94
Torsion balance, 700
Torsion magnetometers, 1013
Total electron contents (TEC), 194, 200
 CID (*see* Co-seismic ionospheric disturbances (CID))
 definition, 1802
 fluctuations, 1804
 GPS data, 1804–1805
 pre-seismic ionospheric anomalies, 1809
 slant, 1803
 vertical, 1803
Total gradient/analytic signal, 964
Total magnetic intensity (TMI), 958, 993
Transdanubian conductivity anomaly (TCA), 461
Trans-dimensional inversion, 811
Trans European Suture Zone (TESZ), 1819, 1821
 geophysical signature, 1821–1827
 lithospheric mantle, 1827
 location of, 1820
 mantle properties, 1826
Transfer function(s), 351, 1057
Transformational faulting, 1633
Transform fault(s), 217, 903
 dextral (right-lateral), 274
 earthquakes on oceanic, 276, 277
 mid-oceanic, 274
 right-lateral, 275
 sinistral (left-lateral), 274
 system, 1257
Transform-resistance force, 1270
Transform setting, 1754
Transient electromagnetic (TEM), 350
Transient phenomena, 780
Transient signals, 1494
Transient source methods, 725
Transition zone (TZ)
 definition, 1390
 issue of water in, 1393
 topography and sharpness, 1391–1393
Transport equation, 1473, 1477, 1481
Transversely isotropic medium, 1528
Transversely isotropic solids, 1459, 1463
Transverse wave, 1448
Travelling ionospheric disturbances (TIDs), 1803
Traveltime, 1603
 inversion, 1828
 tomography, 1609, 1620
Travel time-corrected mean field, 1539
Traveltime tomography
 algorithms, 1833–1836
 applications, 1829
 controlled-source data, 1828
 earthquakes, 1828
 earth seismic velocity, 1828
 forward modeling, 1830
 forward modeling component, 1829
 inversion problem, 1831–1833
 model assessment, 1836–1838
 nonlinear assessment, 1838–1841
 reflection tomography, 1834
 seismic data (*see* Controlled-source seismic data)
 seismic experiments, 1844–1846
 spatial model resolution, 1830
Traverse surveys, 428
Triallateral, 428
Triangulated facet models, 1889
Triangulation, 428
 networks, 439
Triggered and Induced Seismicity Working Group (TAIS WG), 14
Triggered earthquake, 10
Triggered seismicity, 19, 24
Triple junction rifts, 75
Tropical cyclone, 1806
Tropospheric refraction, 554–555
True crater, 771
True polar wander (TPW), 1215
 integrated explanation for, 1220–1222
 measure of, 1216–1219
 modeling of, 1220
 physics of, 1216
Tsunami, 249, 251, 708, 709, 711, 712
 Andaman-Sumatra region, 1882–1883
 definition, 1880
 earthquake, 1628, 1706
 eigenfunctions, 1850
 excitation by earthquakes, 1851–1854
 excitation from impacts, 1857
 excitation from submarine landslides, 1854–1856
 forecasting, 1863–1868
 hazard in the Bay of Bengal, 1881
 Indo-Burmese arc, 1881–1882
 Irrawaddy region, 1882
 period, 1849
 propagation, 1857–1859
 regional warning system, 1870
 run-up, 1859–1860
 samples, 1861–1866
 shoaling, 1858–1860
 short vs. long wave, 1850
 types of, 1869–1870
 velocity, 1849
 warning systems, 711
 wavelength, 1850
Tsunamigenic earthquakes, 1879
Tsunami Warning Centre (TWC), 1869
 bulletins issued by, 1875–1877
 communication infrastructure, 1875
 decision support system, 1875
 observation network, 1870–1872
 service levels, 1877–1879
 tsunami forecast modeling, 1872–1875
Tube waves, 1898
T waves
 acoustic thermometry, 1752
 comprehensive nuclear-test ban treaty, 1752
 definition, 1751
 geophysical applications, 1753
 human perception, 1752
 quantitative measurement, 1752
 sources, 1751–1752
Two-dimensional (2-D) survey, 346
Two-phase mixing laws, 352
Two-point ray tracing, 1476, 1480, 1483
Two-way wave-equation
 migration (reverse-time migration), 1576
Typhoons, 1657

U
Uglegorsk earthquake, 1467
Ultra broad band horizontal geodetic pendulum, 447
 Grotta Gigante horizontal pendulum, 448–449
 instrument description, 448
Ultra high pressure (UHP), 1260
Ultra low frequency (ULF), 198–199
Ultra-low velocity zones (ULVZs), 181
Ultramafic peridotites, 901
Ultra-wide band, 430
Uncertainty, 1271
Unconventional hydrocarbon exploitation, 8
Unconventional resources, 124
Undeformed sediments, 929, 931
Underground mining, 8
Underground nuclear explosions, 1430–1432, 1434–1437, 1439
Underground nuclear tests, 1671
Unified scaling law, 403
Unified scaling law for earthquakes (USLE)
 aftershock, 1894
 control parameter, 1894
 foreshock, 1894
 fractal dimension, 1893
 inter-event time, 1893
 seismic activity, 1893

- Uniform California Earthquake Rupture Forecast (UCERF), 1788
- Uniform reduction method, 303
- Uniform stretching model, 1356
- United Nations Educational, Scientific and Cultural Organization (UNESCO), 1664
- United Nations Office for Disaster Risk Reduction (UNDRR), 1664
- United States Agency for International Development (USAID), 1667
- Universal Time 1 (UT1), 149, 1906
- Universal Time (UT) scale, 439
- Unmanned Aerial Vehicles (UAV) technology, 1012
- Unmanned airborne vehicles (UAV), 640
- Unpaired strath terraces, 1755
- Unsupervised training, 799, 801
- Uplifted/tilted terraces and fan surfaces, 1755
- Upper mantle
- asthenosphere, 184
 - definition, 183
 - Gutenberg discontinuity, 184
 - "520 km"-discontinuity, 184
 - Lehmann discontinuity, 184
 - lithosphere, 184
 - plume, 185
 - tomographic techniques, 185–186
- Upward continuation, 962
- Upward fault termination, 1239
- Uralides, 58
- Uranium radioactivity, earth's core, 1296
- USArray, 1650
- U.S. Geological Survey's National Earthquake Information Center (NEIC), 1649
- USGS Global Hypocenter Data Base system, 1894, 1895
- UT1. *See* Universal time 1 (UT1)
- V**
- Valdivia, 708
- Valley floor width-to-height ratio (V_f), 1757
- Valu Fa Ridge (VFR), 101
- Van Allen, J.A., 786
- Van Allen radiation belts, 788
- van Cittert-Zernicke theorem, 1903
- VAN method, 197–198
- Variational equations, 650
- Variometers, 504
- V-cycle, 1151
- Vector and scalar measurements, 1008
- Vector gravity meters, 639
- Vectorial amplitude, 1473–1474, 1478
- coefficients, 1473
- Vector magnetometers, 999, 1116
- absolute measurement, 1034
 - characteristics, 1033
 - fluxgate, 1034
 - quantum, 1034–1035
- Vector measurements, 503, 1008
- Velocity, 41, 177
- analysis, 1698
 - anisotropy, 1463, 1464
 - vs. density, 1607–1608
 - dispersion, 1515
 - estimation, 1571
 - field, 1574
 - measurements, 1460, 1463
 - model, 1498, 1898, 1899, 1901
 - P-waves, 41
 - structure, 58
 - S-waves, 42
- Velocity–pressure formulation, 1147
- Velocity-strengthening, 269
- Velocity-temperature equation, 1511
- Vertical datum, 749, 750
- definition and realization, 752
 - determination of, 755–756
 - multiple vertical datums geodetic BVP, 753–755
 - single (global) vertical datum geodetic BVP, 753
 - unification, 749, 752, 756
- Vertical derivative operator, 479
- Vertical gradient filters, 963
- Vertical pendulums, 449
- Vertical seismic profiling (VSP), 49
- acquisition geometries, 1897–1898
 - applications, 1898–1900
 - definition, 1897
 - geophysical exploration, 1897
 - seismic wave propagation velocities, 1897
 - surveys, 382
- Vertical TEC, 1803
- Very fast simulated annealing (VFSA), 809–811, 813
- Very long baseline interferometry (VLBI)
- data calibration and imaging, 1904–1905
 - definition, 1902
 - geodetic and geophysical applications, 1905–1907
 - methods and techniques, 1902–1903
 - networks, 1905
 - physical principles, 1903–1904
- Very-long-period earthquakes, 329, 335
- Vibroseis, 57
- processing, 1375
- Videogrammetric approach, 1875
- Video record, 712
- Vienna mapping function, 555
- Virieux grid, 1138
- Virtual geomagnetic pole, 509, 1230
- Virtual observatories, 360
- Virtual Observatories in Geosciences (VOiG) conference, 360
- Virtual refraction interferometry, 1603–1604
- advantages, 1604
 - limitations, 1604
- Virtual seismograms, 1560
- Virtual sources (VS), 1899
- Viscoelastic asthenosphere, 896
- Visco-elastic Earth-models, 436
- Visco-elastic media, 759
- Viscoelastic model, 1139
- Viscoplasticity, 1268
- Viscosity of core fluid, 88
- Viscous magnetizations, 15
- Viscous remanent magnetization (VRM), 1195, 1322
- Viscous torque, 88
- Visibility function, 1904
- Voigt notation, 1138
- Volcanic earthquakes, 254
- Volcanic mass changes, 630
- Volcanic systems, 118–120
- Volcanic tremor, 329, 333–334
- Volcanic tremors, 254, 1643
- Volcanism, 1257, 1351, 1732
- in solar system, 4
- Volcano deformation, 568–570
- Volcano-sedimentary assemblages, 1262
- Volcano seismology
- infrasonic signals, seismicity with, 336–337
 - lahars and pyroclastic flows, detection of, 337
 - low-frequency earthquakes, 331–333
 - magma ascent rate, estimation of, 336
 - rock-fall and lahar events, 331
 - stress changes, in volcanic settings, 337
 - very-long-period earthquakes, 335
 - volcanic tremor, 333–334
 - volcano-tectonic earthquakes, 330–331
- Volcano-tectonic earthquakes, 329–331
- Volume expansion coefficient, 1762
- Volume strainmeter, 1493
- Volumetric heat capacity, 1764–1766
- vonKarman, 404
- Von-Karman correlation function, 1538
- Von Karman Sodium (VKS) experiment, 511
- Voxel-cubes, 1889
- V_p/V_s , 1462
- Vredefort structure, 769
- Vulnerability, 1662
- W**
- Wadati–Benioff zone, 1629, 1728
- Waiting time, 403
- Warm-back, 391
- Warped terraces and fan surfaces, 1756
- WARR recordings, 132
- Water fluxes, 1815
- Water storage, 1813
- Wave-equation migration, 1574
- Wave equation traveltimes inversion, 1411
- Waveform cross-correlation, 304–305, 1670
- Waveform modeling
- 3-D seismic modeling, 1610
 - elastic wave equation, 1610
 - finite-difference method, 1611
 - first-order differential equations, 1609
 - fluctuating interface, 1611
 - frequency domain, 1610
 - seismic wave propagation, 1610
 - stress components, 1609
 - structural complexities, 1610
 - wave propagation, 1610, 1611
- Waveform tomography, 1620
- frequency domain, 1609
 - frequency slice, 1620
 - inverse problem, 1609, 1611–1613
 - preprocessing, 1617–1618
 - reflection, 1618–1621
 - restarted L-BFGS, 1613–1617
 - shot-encoded, 1613
 - subsurface interfaces, 1611

- Waveform tomography, 1620 (*cont.*)
 transmission data, 1609
 velocity model, 1619
 waveform modeling, 1609–1611
 Wavefront construction method, 1477
 Wavefront healing, 1509
 Wavefront orthonormal coordinate,
 1478–1480
 Wavefront tracking algorithms, 1830–1831
 Wave heights, 1343
 Wavelet analysis, 594–595, 1022
 applications, 595
 CWT, 1910–1913
 dispersion analysis, 1914–1915
 DWT, 1913–1914
 instantaneous parameters, 1914
 polarization analysis, 1914
 and windowed Fourier transform, 595
 Wavelet processing, 1374–1376
 Wave number (f-k) migration, 1575
 Wave propagation, 226, 404, 757–759, 762,
 1533
 acoustic, 758
 elastic, 758–759
 elastic wave equation, 759
 reflection, 40–41
 refraction, 39–40
 related noise, 1443
 seismic wave attenuation and effects on
 seismic waveform, 759–760
 seismogram, 39
 theory, 1137–1139
 W-cycle, 1151
 Weak-form Galerkin finite element model,
 1146
 Weakly anisotropic media, 1472, 1475, 1477,
 1481, 1483–1485
 Weakly dissipative media, 1483–1484
 Web-based macroseismic surveys, 279
 Wegener, A., 63–64, 1203
 Weighted damped least square inverse, 817
 Well cementing, 391
 Wellhead pressures, 11
 Well logging, 527, 802
 acoustic and electrical applications,
 534–535
 benefits and limitations of, 529
 calibration and standardization of logs, 529
 crossplot technique, 532
 fundamentals of log analysis, 530
 hydrogeologic applications, 535–536
 instrumentation, 530
 logs, 530
 mineral exploration, 536
 in oil industry, 532–534
 principles, 529–530
 wellbore environment, 528–529
 Well-tie, 1487, 1489, 1491
 Well treatments, 391
 Wenchuan earthquake, 1494, 1624
 Wenchuan Earthquake Fault Scientific Drilling
 Project (WFSD), 122
 Wener deconvolution, 967
 Wenner-Schlumberger array, 347
 Werner method, 593, 968
 WGS84 geodetic system, 643
 Wharton Ridge system, 1092
 Wharton spreading ridge migration, 1091
 Whole-mantle models, 1271–1272
 Whole rock isochrons, 6–7
 Wide-angle reflection/refraction profiling.
 See Deep seismic reflection and
 refraction profiling
 Wide-aperture studies, 1159
 Wide Azimuth acquisition, 1445
 Wiener filter, 1377
 Wiener optimum filter, 1376
 Williamson–Adams equation, 161, 162
 Wilson-cycle, 73, 838, 1257
 Windowed Fourier transform, 595
 Wind speed, 1343
 Wireline downhole logging system, 856
 Wood-Anderson (WA) torsion seismograph,
 244
 World Data Centers, 788
 World Days, 788
 World Digital Magnetic Anomaly Map
 (WDMAM), 939
 gamma model, 950–951
 gridded data sets, 947
 leeds model, 951
 marine and airborne magnetic
 compilations, 946
 organizations, 945
 tectonic rifts, 943
 World digital magnetic anomaly map version
 2.0 (WDMAM2.0), 954, 956
 World Geodetic System 1984 (WGS84), 425,
 437, 943
 World Gravity Map, 788
 World Meteorological Organization, 787
 World-Wide Standardized Seismograph
 Network (WWSSN), 1644
 Worms, 965
 Wraparound, 1529
Y
 11-Year solar cycle, 1009
 Yield estimation, 1431, 1435–1437
 Yield stress envelopes (YSEs), 887
 Young sea floor, 362
Z
 Zero-offset VSP (ZVSP), 1898
 Zero-order ray approximation, 1474, 1482,
 1483
 Zero-order WKB approximation, 1575
 Zijderveld plot, 1195
 Zircar aluminum cement, 1194
 Zircon U-Pb geochronology, 1258
 Zoeppritz equation, 369